DLow: Diversifying Latent Flows for Diverse Human Motion Prediction
Supplementary Material

Ye Yuan Kris Kitani
Carnegie Mellon University
{yyuan2, kkitani}@cs.cmu.edu

1. Implementation Details

1.1. Network Architectures

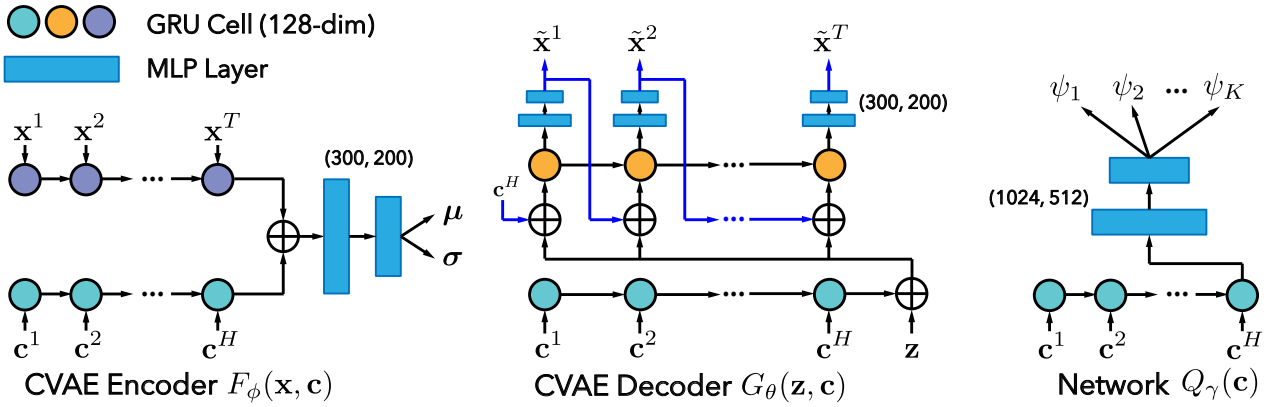

Figure 1. **Network architectures** for the CVAE and DLow. We use GRUs [1] to extract motion features. $\mathbf{x}^t$ and $\mathbf{c}^t$ denotes the t-th pose in $\mathbf{x}$ and $\mathbf{c}$ respectively.

1.2. Training

We use a batch size of 64 and set the latent dimensions n_z to 128 in all experiments. For the CVAE, we sample 5000 training examples every epoch and train the networks for 500 epochs using Adam [2] and a learning rate of 1e-3. The DLow objective in Eq. (9) of the main paper can be rewritten as: $L(\psi) = \beta L_{\text{KL}} + \lambda_d E_d + \lambda_r E_r$. We set $(\beta, \lambda_d, \lambda_r)$ to $(1, 25, 2)$ for Human3.6M and $(1, 50, 2)$ for HumanEva-I. For the mappings T_{ψ_k}, we specify $\mathbf{A}_k$ to be diagonal to reduce the output size of Q_γ. This design is mainly for computational efficiency, as we do find that using a full parametrization of $\mathbf{A}_k$ improves performance. The RBF kernel scale σ_d is set to 100 for Human3.6M and 20 for HumanEva-I. For both datasets, we sample 5000 training examples every epoch and train Q_γ for 500 epochs using Adam with a learning rate of 1e-4.

2. Additional Human3.6M Results

In this section, we show more qualitative results on Human3.6M, including additional comparison with baselines (Fig. 2) and additional examples of DLow (Fig. 3). Please refer to the supplementary video to see the whole motion sequences.

2.1. Additional Comparison with Baselines on Human3.6M

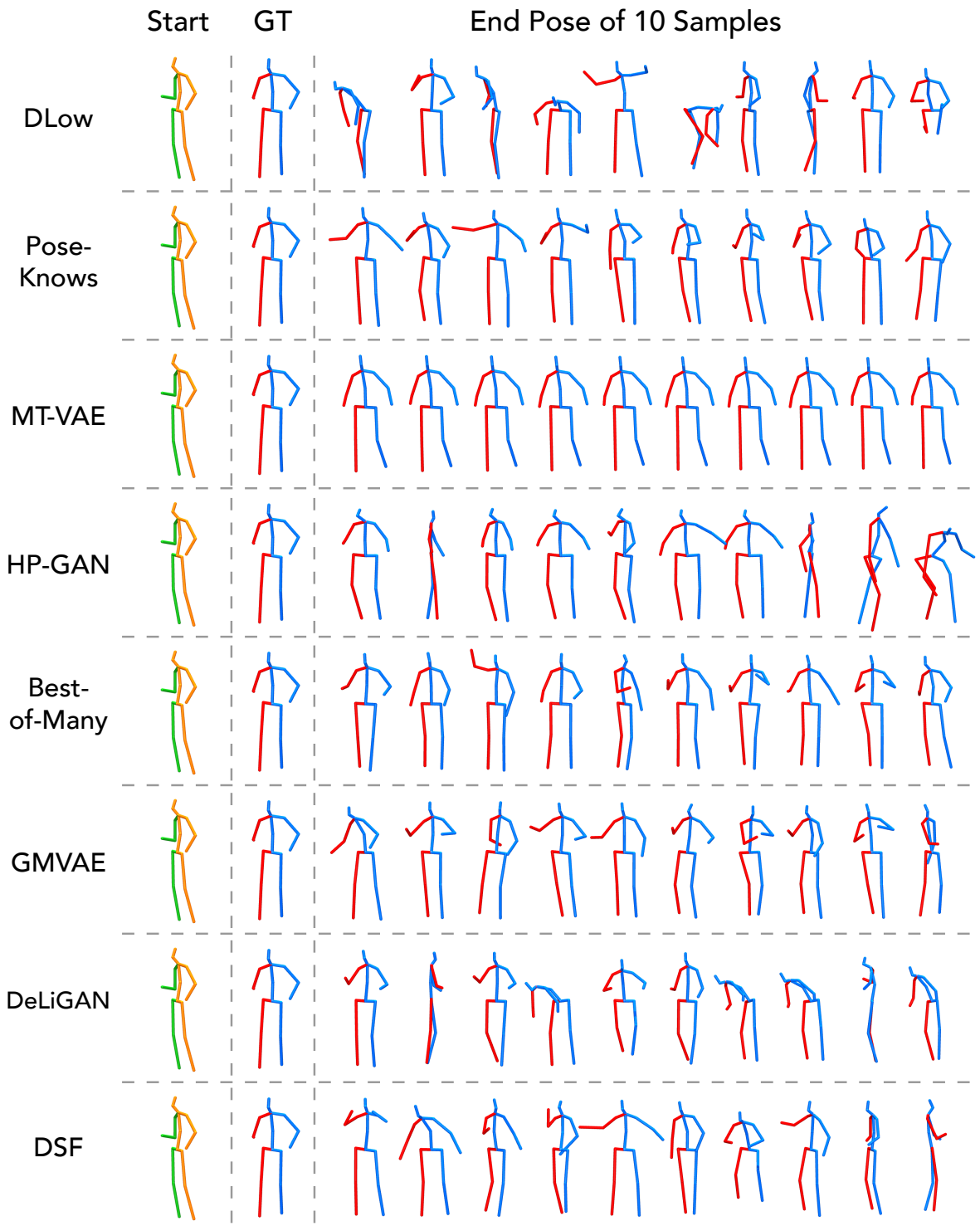

Figure 2. **Additional comparison with the baselines on Human3.6M.** We show the start pose, the end pose of the ground truth future motion, and the end pose of 10 motion samples by each method.

2.2. Additional Examples of DLow on Human3.6M

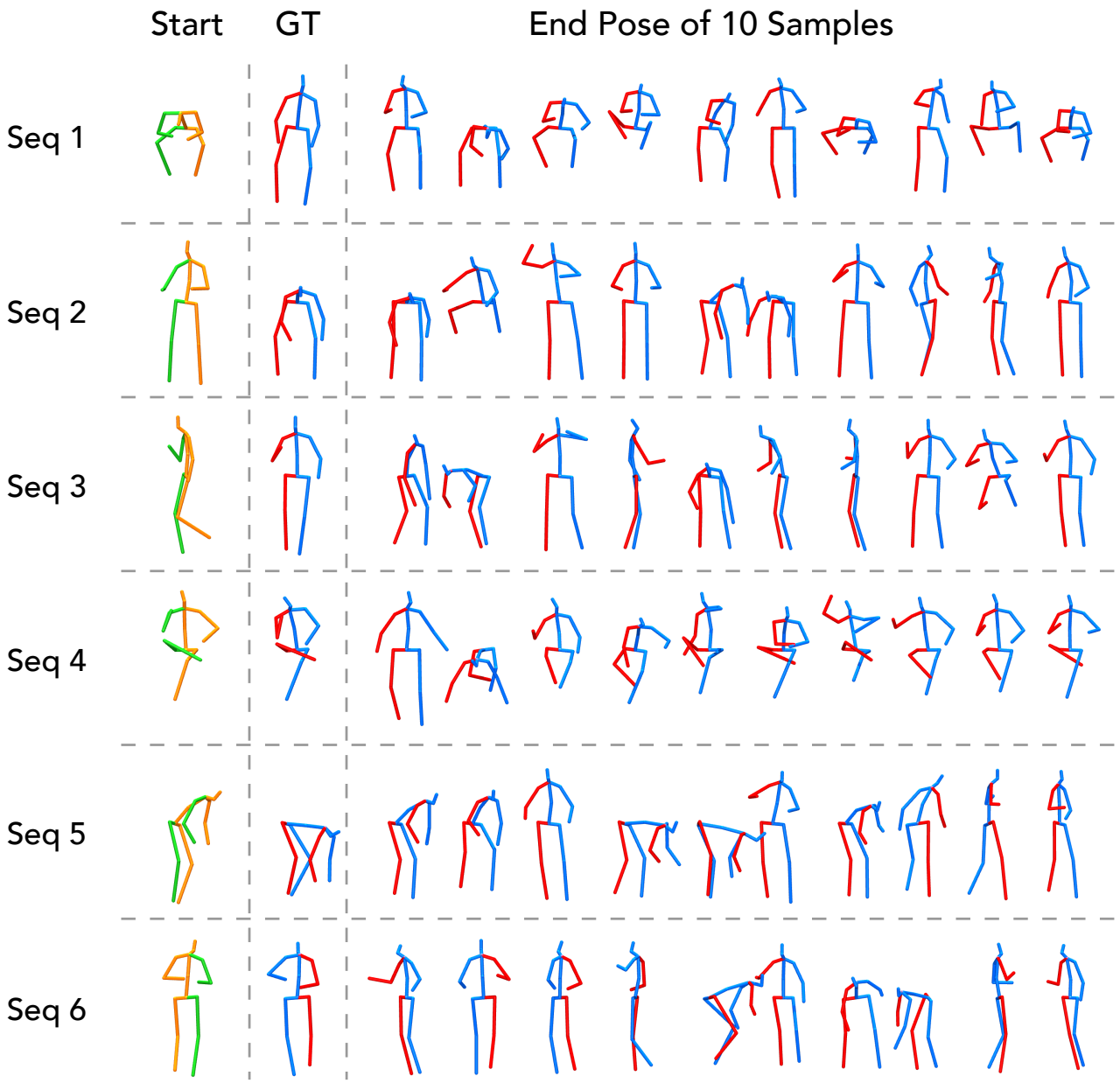

Figure 3. **Additional examples of DLow on Human3.6M.** Each row corresponds to a different sequence, where we show the start pose, the end pose of the ground truth future motion, and the end pose of 10 motion samples.

3. Additional Controllable Motion Prediction Results

In Fig. 4, we show additional results on controllable motion prediction using Human3.6M, where we use DLow to constrain the motion samples to have similar leg motion to the reference motion but diverse upper-body motion. Notice that DLow is able to produce samples with similar leg motion, while CVAE (random) samples cannot enforce similar leg motion. We further show some quantitative results in Table 1, where we compute the average leg motion distance from motion samples to the reference motion and the APD for upper-body motion.

Implementation Details. We use the same networks in Fig. 3 of the main paper and the same hyperparmeters and training procedure given in the implementation details of the main paper. The main modification is that we use Eq. 24 in the paper for the energy function E of the prior $p(X)$, and the DLow objective in Eq. 12 can be rewritten as: $L(\psi) = \beta L_{\text{KL}} + \lambda_d E_d + \lambda_s E_s + \lambda_r E_r$. We set $(\beta, \lambda_d, \lambda_s, \lambda_r)$ to $(1, 50, 10, 0)$. We also use a full parametrization of $\mathbf{A}_k$ instead of a diagonal one.

Method	Leg Dist ↓	Upper-body APD ↑
DLow	**1.071**	**12.741**
CVAE	2.958	6.051

Table 1. **Quantitative results** for controllable motion prediction.

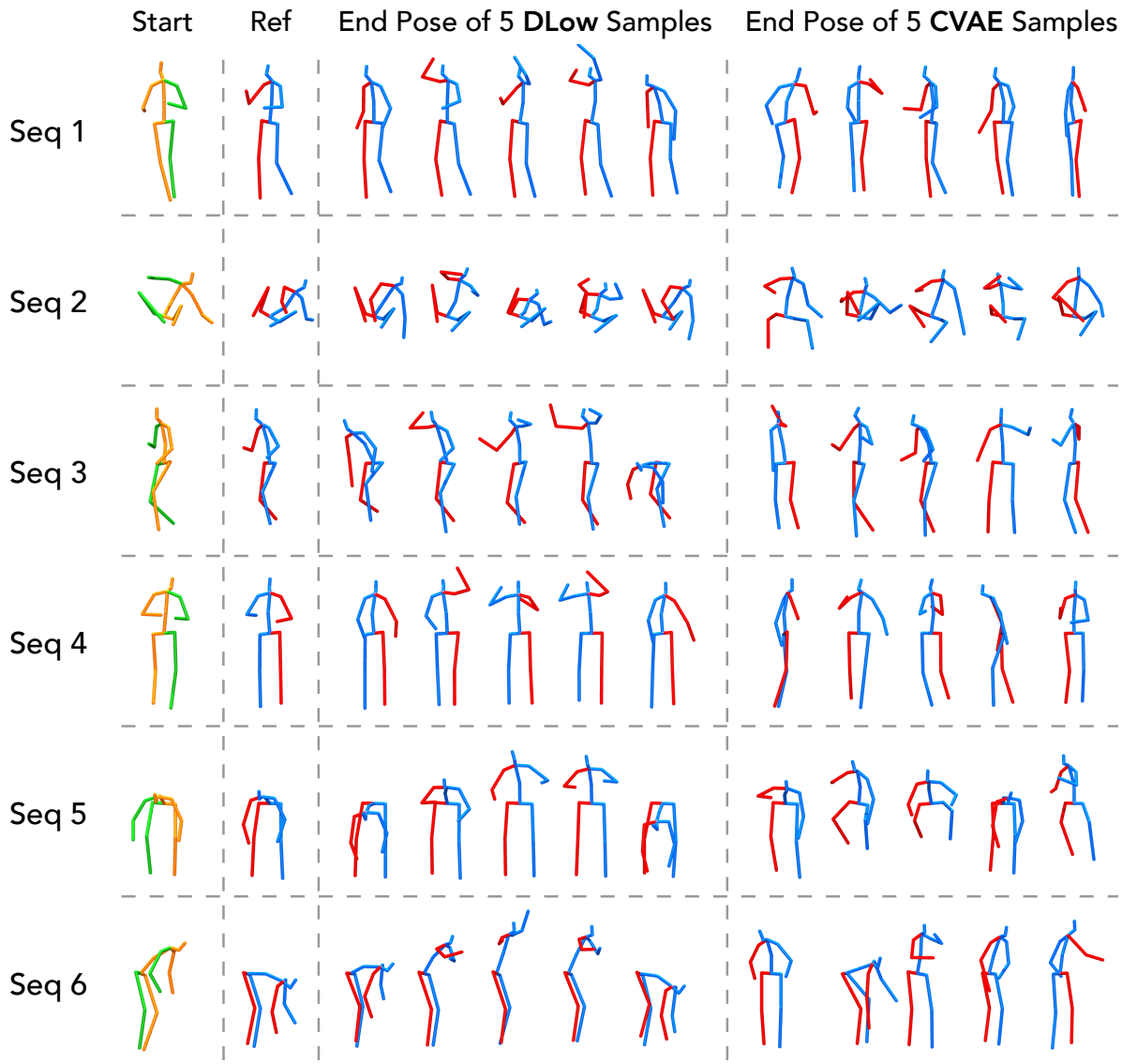

Figure 4. **Additional results on controllable motion prediction.** Notice that DLow can produce motion samples that have similar leg motion to the reference (Ref) yet diverse upper-body motion, while CVAE (random) samples cannot enforce similar leg motion.

3.1. Metrics vs. Number of Samples K

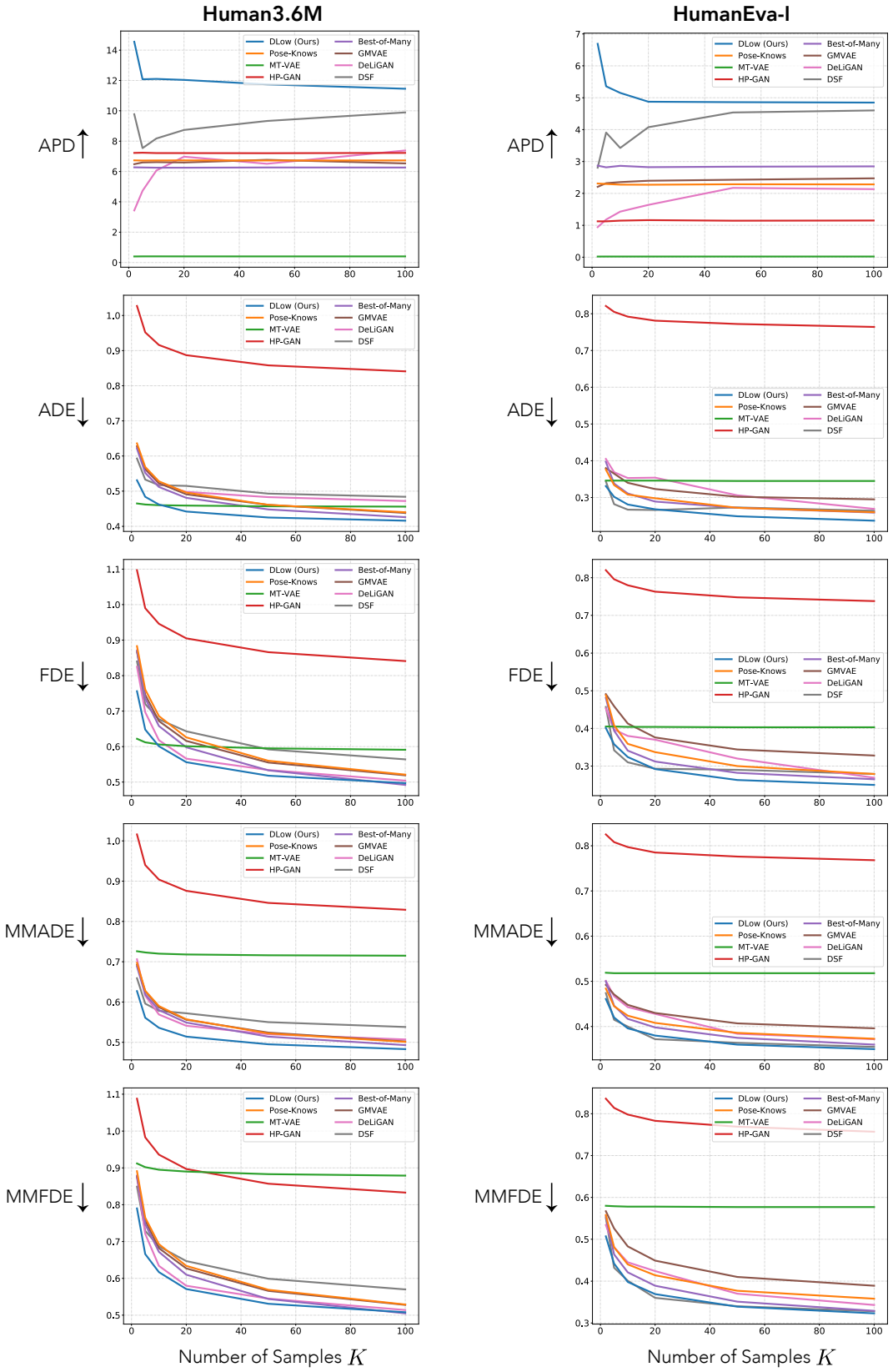

Figure 5. **Metrics vs. Number of Samples** K on both Human3.6M (Left) and HumanEva-I (Right).

4. Additional HumanEva-I Results

We also show more qualitative results on HumanEva-I which is a much smaller dataset with less motion variation. We present additional comparison with baselines (Fig. 6) and additional examples of DLow (Fig. 7).

4.1. Additional Comparison with Baselines on HumanEva-I

Figure 6. **Additional comparison with the baselines on HumanEva-I.** We show the start pose, the end pose of the ground truth future motion, and the end pose of 10 motion samples by each method.

4.2. Additional Examples of DLow on HumanEva-I

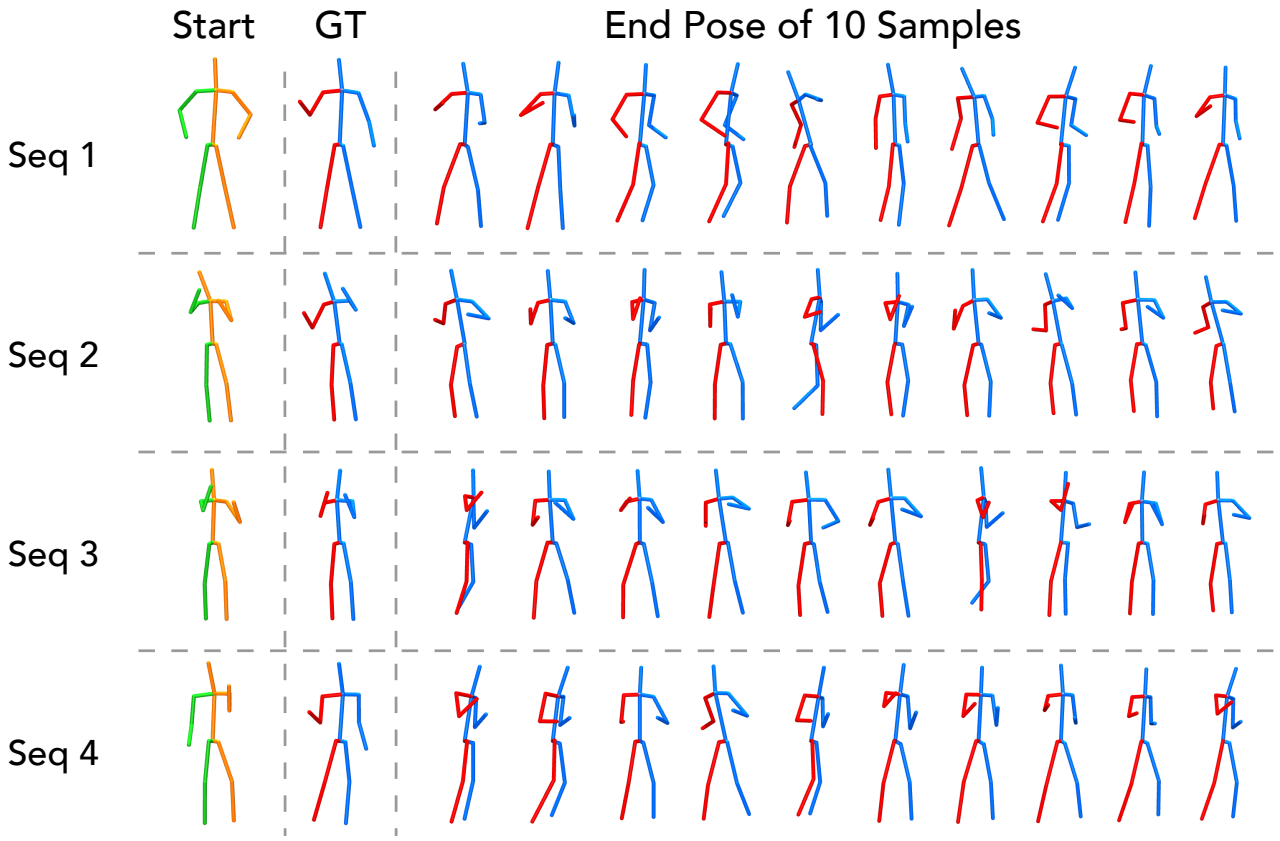

Figure 7. **Additional examples of DLow on HumanEva-I.** Each row corresponds to a different sequence, where we show the start pose, the end pose of the ground truth future motion, and the end pose of 10 motion samples.

References

[1] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014. 1

[2] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 1