DLow: Diversifying Latent Flows for Diverse Human Motion Prediction

Ye Yuan and Kris Kitani

Robotics Institute, Carnegie Mellon University
{yyuan2, kkitani}@cs.cmu.edu

Abstract. Deep generative models are often used for human motion prediction as they are able to model multi-modal data distributions and characterize diverse human behavior. While much care has been taken into designing and learning deep generative models, how to efficiently produce diverse samples from a deep generative model *after* it has been trained is still an under-explored problem. To obtain samples from a pretrained generative model, most existing generative human motion prediction methods draw a set of independent Gaussian latent codes and convert them to motion samples. Clearly, this random sampling strategy is not guaranteed to produce diverse samples for two reasons: (1) The independent sampling cannot force the samples to be diverse; (2) The sampling is based solely on likelihood which may only produce samples that correspond to the major modes of the data distribution. To address these problems, we propose a novel sampling method, Diversifying Latent Flows (DLow), to produce a diverse set of samples from a pretrained deep generative model. Unlike random (independent) sampling, the proposed DLow sampling method samples a single random variable and then maps it with a set of learnable mapping functions to a set of correlated latent codes. The correlated latent codes are then decoded into a set of correlated samples. During training, DLow uses a diversity-promoting prior over samples as an objective to optimize the latent mappings to improve sample diversity. The design of the prior is highly flexible and can be customized to generate diverse motions with common features (e.g., similar leg motion but diverse upper-body motion). Our experiments demonstrate that DLow outperforms state-of-the-art baseline methods in terms of sample diversity and accuracy. Our code is released on the project page: https://www.ye-yuan.com/dlow.

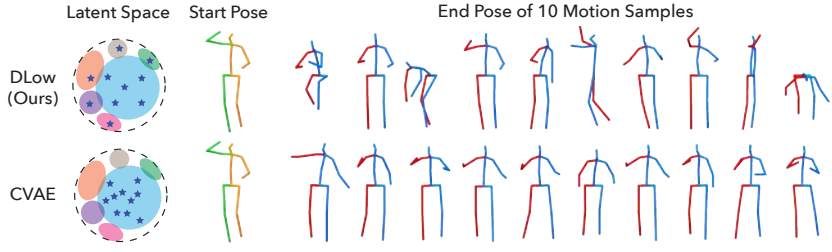

Fig. 1. In the latent space of a conditional variational autoencoder (CVAE), samples (stars) from our method DLow are able to cover more modes (colored ellipses) than the CVAE samples. In the motion space, DLow generates a diverse set of future human motions while the CVAE only produces perturbations of the motion of the major mode.

1 Introduction

Human motion prediction, i.e., predicting the future 3D poses of a person based on past poses, is an important problem in computer vision and has many useful applications in autonomous driving [55], human robot interaction [39] and healthcare [67]. It is a challenging problem because the future motion of a person is potentially diverse and multi-modal due to the complex nature of human behavior. For many safety-critical applications, it is important to predict a diverse set of human motions instead of just the most likely one. For examples, an autonomous vehicle should be aware that a nearby pedestrian can suddenly cross the road even though the pedestrian will most likely remain in place. This diversity requirement calls for a generative approach that can fully characterize the multi-modal distribution of future human motion.

Deep generative models, e.g., variational autoencoders (VAEs) [38], are effective tools to model multi-modal data distributions. Most existing work [68, 46, 6, 61, 42, 71, 3] using deep generative models for human motion prediction is focused on the design of the generative model to allow it to effectively learn the data distribution. After the generative model is learned, little attention has been paid to the sampling method used to produce *motion samples* (predicted future motions) from the *pretrained* generative model (weights kept fixed). Most of prior work predicts a set of motions by randomly sampling a set of latent codes from the latent prior and decoding them with the generator into motion samples. We argue that such a sampling strategy is not guaranteed to produce a diverse set of samples for two reasons: (1) The samples are independently drawn, which makes it difficult to enforce diversity; (2) The samples are drawn based on likelihood only, which means many samples may concentrate around the major modes (which have more observed data) of the data distribution and fail to cover the minor modes (as shown in Fig. 1 (Bottom)). The poor sample efficiency of random sampling means that one needs to draw a large number of samples in order to cover all the modes which is computationally expensive and can lead to high latency, making it unsuitable for real-time applications such as autonomous driving and virtual reality. This prompts us to address an overlooked aspect of diverse human motion prediction — the sampling strategy.

We propose a novel sampling method, Diversifying Latent Flows (DLow), to obtain a diverse set of samples from a pretrained deep generative model. For this work, we use a conditional variational autoencoder (CVAE) as our pretrained generative model but other generative models can also be used with our approach. DLow is inspired by the two previously mentioned problems with random (independent) sampling. To tackle problem (1) where sample independence limits model diversity, we introduce a new random variable and a set of learnable deterministic mapping functions to correlate the motion samples. We first transform the random variable with the mappings functions to generate a set of correlated latent codes which are then decoded into motion samples using the generator. As all motion samples are generated from a common random factor, this formulation allows us to model the joint sample distribution and offers us the opportunity to impose diversity on the samples by optimizing the parameters

of the mapping functions. To address problem (2) where likelihood-based sampling limits diversity, we introduce a diversity-promoting prior (loss function) on the samples during the training of DLow. The prior follows an energy-based formulation using an energy function based on pairwise sample distance. We optimize the mapping functions during training to minimize the cross entropy between the joint sample distribution and diversity-promoting prior to increase sample diversity. To strike a balance between diversity and likelihood, we add a KL term to the optimization to enhance the likelihood of each sample. The relative weights between the prior term and the KL term represent the trade-off between the diversity and likelihood of the generated motion samples. Furthermore, our approach is highly flexible in that by designing different forms of the diversity-promoting prior we can impose a variety of structures on the samples besides diversity. For example, we can design the prior to ask the motion samples to cover the ground truth better to achieve higher sample accuracy. Additionally, other designs of the prior can enable new applications, such as controllable motion prediction, where we generate diverse motion samples that share some common features (e.g., similar leg motion but diverse upper-body motion).

The contributions of this work are the following: (1) We propose a novel perspective for addressing sample diversity in deep generative models — designing sampling methods for a *pretrained* generative model. (2) We propose a principled sampling method, DLow, which formulates diversity sampling as a constrained optimization problem over a set of learnable mapping functions using a diversity-promoting prior on the samples and KL constraints on the latent codes, which allows us to balance between sample diversity and likelihood. (3) Our approach allows for flexible design of the diversity-promoting prior to obtain more accurate samples or enable new applications such as controllable motion prediction. (4) We demonstrate through human motion prediction experiments that our approach outperforms state-of-the-art baseline methods in terms of sample diversity and accuracy.

2 Related Work

Human Motion Prediction. Most previous work takes a deterministic approach to modeling human motion and regress a single future motion from past 3D poses [17, 34, 9, 45, 18, 52, 57, 13, 22, 1, 69, 51] or video frames [10, 77, 75]. While these approaches are able to predict the most likely future motion, they fail to model the multi-modal nature of human motion, which is essential for safety-critical applications. More related to our work, stochastic human motion prediction methods start to gain popularity with the development of deep generative models. These methods [68, 46, 6, 61, 42, 71, 3, 76] often build upon popular generative models such as conditional generative adversarial networks (CGANs; [21]) or conditional variational autoencoders (CVAEs; [38]). The aforementioned methods differ in the design of their generative models, but at test time they follow the same sampling strategy — randomly and independently sampling trajectories from the pretrained generative model without considering the correlation between samples. In this work, we propose a principled sam-

pling method that can produce a diverse set of samples, thus improving sample efficiency compared to the random sampling typically used in prior work.

Diverse Inference. Producing a diverse set of solutions has been investigated in numerous problems in computer vision and machine learning. A branch of these diversity-driven methods stems from the M-Best MAP problem [54, 62], including diverse M-Best solutions [7] and multiple choice learning [28, 44]. Alternatively, submodular function maximization has been applied to select a diverse subset of garments from fashion images [31]. Another type of methods [40, 20, 19, 32, 5, 74, 70] seeks diversity using determinantal point processes (DPPs; [50, 41]) which are efficient probabilistic models that can measure the global diversity and quality within a set. Similarly, Fisher information [60] has been used for diverse feature [23] and data [64] selection. Diversity has also been a key aspect in generative modeling. A vast body of work has tried to alleviate the mode collapse problem in GANs [11, 12, 65, 4, 25, 16, 47, 72] and the posterior collapse problem in VAEs [78, 66, 36, 8, 48, 29]. Normalizing flows [58] have also been used to promote diversity in trajectory forecasting [59, 24]. This line of work aims to improve the diversity of the data distribution learned by deep generative models. We address diversity from a different angle by improving the strategy for producing samples from a pretrained deep generative model.

3 Diversifying Latent Flows (DLow)

For many existing methods on generative vision tasks such as multi-modal human motion prediction, the primary focus is to learn a good generative model that can capture the multi-modal distribution of the data. In contrast, once the generative model is learned, little attention has been paid to devising sampling strategies for producing diverse samples from the *pretrained* generative model.

In this section, we will introduce our method, Diversifying Latent Flows (DLow), as a principled way for drawing a diverse and likely set of samples from a pretrained generative model (weights kept fixed). To provide the proper context, we will first start with a brief review of deep generative models and how traditional methods produce samples from a pretrained generative model.

Background: Deep Generative Models. Let $\mathbf{x} \in \mathcal{X}$ denote data (e.g., human motion) drawn from a data distribution $p(\mathbf{x}|\mathbf{c})$ where $\mathbf{c}$ is some conditional information (e.g., past motion). One can reparameterize the data distribution by introducing a latent variable $\mathbf{z} \in \mathcal{Z}$ such that $p(\mathbf{x}|\mathbf{c}) = \int_{\mathbf{z}} p(\mathbf{x}|\mathbf{z},\mathbf{c})p(\mathbf{z})d\mathbf{z}$, where $p(\mathbf{z})$ is a Gaussian prior distribution. Deep generative models learn $p(\mathbf{x}|\mathbf{c})$ by modeling the conditional distribution $p(\mathbf{x}|\mathbf{z},\mathbf{c})$, and the generative process can be described as sampling $\mathbf{z}$ and mapping them to data samples $\mathbf{x}$ using a deterministic *generator* function $G_\theta : \mathcal{Z} \to \mathcal{X}$ as

$$\mathbf{z} \sim p(\mathbf{z}), \tag{1}$$

$$\mathbf{x} = G_\theta(\mathbf{z},\mathbf{c}), \tag{2}$$

where the generator G_θ is instantiated as a deep neural network parametrized by θ. This generative process produces samples from the implicit sample distribution $p_\theta(\mathbf{x}|\mathbf{c})$ of the generative model, and the goal of generative modeling is to

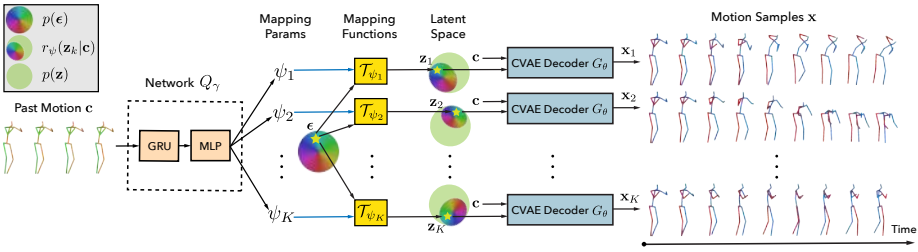

Fig. 2. Overview of our DLow framework applied to diverse human motion prediction. The network Q_γ takes past motion $\mathbf{c}$ as input and outputs the parameters of the mapping functions $\mathcal{T}_{\psi_1}, \ldots, \mathcal{T}_{\psi_K}$. Each mapping $\mathcal{T}_{\psi_k}$ transforms the random variable $\boldsymbol{\epsilon}$ to a different latent code $\mathbf{z}_k$ and also warps the density $p(\boldsymbol{\epsilon})$ to the latent code density $r_\psi(\mathbf{z}_k|\mathbf{c})$. Each latent code $\mathbf{z}_k$ is decoded by the CVAE decoder into a motion sample $\mathbf{x}_k$.

learn a generator G_θ such that $p_\theta(\mathbf{x}|\mathbf{c}) \approx p(\mathbf{x}|\mathbf{c})$. There are various approaches for learning the generator function G_θ, which yield different types of deep generative models such as variational autoencoders (VAEs; [38]), normalizing flows (NFs; [58]), and generative adversarial networks (GANs; [21]). Note that even though the discussion in this work is focused on conditional generative models, our method can be readily applied to the unconditional case.

Random Sampling. Once the generator function G_θ is learned, traditional approaches produce samples from the learned data distribution $p_\theta(\mathbf{x}|\mathbf{c})$ by first randomly sampling a set of latent codes $Z = \{\mathbf{z}_1, \ldots, \mathbf{z}_K\}$ from the latent prior $p(\mathbf{z})$ (Eq. (1)) and decode Z with the generator G_θ into a set of data samples $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_K\}$ (Eq. (2)). We argue that such a sampling strategy may result in a less diverse sample set for two reasons: (1) Independent sampling cannot model the repulsion between samples within a diverse set; (2) The sampling is only based on the data likelihood and many samples can concentrate around a small number of modes that have more training data. As a result, random sampling can lead to low sample efficiency because many samples are similar to one another and fail to cover other modes in the data distribution.

DLow Sampling. To address the above issues with the random sampling approach, we propose an alternative sampling method, Diversifying Latent Flows (DLow), that can generate a diverse and likely set of samples from a pretrained deep generative model. Again, we stress that the weights of the generative model are kept fixed for DLow. We later apply DLow to the task of human motion prediction in Sec. 4 to demonstrate DLow's ability to improve sample diversity.

Instead of sampling each latent code $\mathbf{z}_k \in Z$ independently according to $p(\mathbf{z})$, we introduce a random variable $\boldsymbol{\epsilon}$ and conditionally generate the latent codes Z and data samples X as follows:

$$\boldsymbol{\epsilon} \sim p(\boldsymbol{\epsilon})\,, \tag{3}$$

$$\mathbf{z}_k = \mathcal{T}_{\psi_k}(\boldsymbol{\epsilon})\,, \qquad 1 \le k \le K\,, \tag{4}$$

$$\mathbf{x}_k = G_\theta(\mathbf{z}_k, \mathbf{c})\,, \qquad 1 \le k \le K\,, \tag{5}$$

where $p(\boldsymbol{\epsilon})$ is a Gaussian distribution, $\mathcal{T}_{\psi_1}, \ldots, \mathcal{T}_{\psi_K}$ are latent mapping functions with parameters $\psi = \{\psi_1, \ldots, \psi_K\}$, and each $\mathcal{T}_{\psi_k}$ maps $\boldsymbol{\epsilon}$ to a different latent code $\mathbf{z}_k$. The above generative process defines a joint distribution $r_\psi(X, Z|\mathbf{c}) = p_\theta(X|Z, \mathbf{c}) r_\psi(Z|\mathbf{c})$ over the samples X and latent codes Z, where $p_\theta(X|Z, \mathbf{c})$ is the conditional distribution induced by the generator $G_\theta(\mathbf{z}, \mathbf{c})$. Notice that in our setup, $r_\psi(X, Z|\mathbf{c})$ depends only on ψ as the generator parameters θ are learned in advance and are kept fixed. The data samples X can be viewed as a sample from the joint sample distribution $r_\psi(X|\mathbf{c}) = \int r_\psi(X, Z|\mathbf{c}) dZ$ and the latent codes Z can be regarded as a sample from the joint latent distribution $r_\psi(Z|\mathbf{c})$ induced by warping $p(\boldsymbol{\epsilon})$ through $\mathcal{T}_{\psi_1}, \ldots, \mathcal{T}_{\psi_K}$. If we further marginalize out all variables except for $\mathbf{x}_k$ from $r_\psi(X|\mathbf{c})$, we obtain the marginal sample distribution $r_\psi(\mathbf{x}_k|\mathbf{c})$ from which each sample $\mathbf{x}_k$ is drawn. Similarly, each latent code $\mathbf{z}_k \in Z$ can be viewed as a latent sample from the marginal latent distribution $r_\psi(\mathbf{z}_k|\mathbf{c})$.

The above distribution reparametrizations are illustrated in Fig. 2. We can see that all latent codes Z and data samples X are correlated as they are uniquely determined by $\boldsymbol{\epsilon}$, and by sampling $\boldsymbol{\epsilon}$ one can easily produce Z and X from the joint latent distribution $r_\psi(Z|\mathbf{c})$ and joint sample distribution $r_\psi(X|\mathbf{c})$. Because $r_\psi(Z|\mathbf{c})$ and $r_\psi(X|\mathbf{c})$ are controlled by the latent mapping functions $\mathcal{T}_{\psi_1}, \ldots, \mathcal{T}_{\psi_K}$, we can impose structural constraints on $r_\psi(Z|\mathbf{c})$ and $r_\psi(X|\mathbf{c})$ by optimizing the parameters ψ of the latent mapping functions.

To encourage the diversity of samples X, we introduce a diversity-promoting prior $p(X)$ (specific form defined later) and formulate a constrained optimization problem:

$$\min_{\psi} \quad -\mathbb{E}_{X \sim r_\psi(X|\mathbf{c})}[\log p(X)]\,, \tag{6}$$

$$\text{s.t.} \quad \mathrm{KL}(r_\psi(\mathbf{z}_k|\mathbf{c}) \| p(\mathbf{z}_k)) = 0\,, \quad 1 \le k \le K\,, \tag{7}$$

where we minimize the cross entropy between the sample distribution $r_\psi(X|\mathbf{c})$ and the diversity-promoting prior $p(X)$. However, the objective in Eq. (6) alone can result in very low-likelihood samples $\mathbf{x}_k$ corresponding to latent codes $\mathbf{z}_k$ that are far away from the Gaussian prior $p(\mathbf{z}_k)$. To ensure that each sample $\mathbf{x}_k$ also has high likelihood under the generative model $p_\theta(\mathbf{x}|\mathbf{c})$, we add constraints in Eq. (7) on the KL divergence between $r_\psi(\mathbf{z}_k|\mathbf{c})$ and the Gaussian prior $p(\mathbf{z}_k)$ (same as $p(\mathbf{z})$) to make $r_\psi(\mathbf{z}_k|\mathbf{c}) = p(\mathbf{z}_k)$ and thus $r_\psi(\mathbf{x}_k|\mathbf{c}) = p_\theta(\mathbf{x}_k|\mathbf{c})$ where $r_\psi(\mathbf{x}_k|\mathbf{c}) = \int p_\theta(\mathbf{x}_k|\mathbf{z}_k, \mathbf{c}) r_\psi(\mathbf{z}_k|\mathbf{c}) d\mathbf{z}_k$ and $p_\theta(\mathbf{x}_k|\mathbf{c}) = \int p_\theta(\mathbf{x}_k|\mathbf{z}_k, \mathbf{c}) p(\mathbf{z}_k) d\mathbf{z}_k$. To optimize this constrained objective, we soften the constraints with the Lagrangian function:

$$\min_{\psi} \; -\mathbb{E}_{X \sim r_\psi(X|\mathbf{c})}[\log p(X)] + \beta \sum_{k=1}^{K} \mathrm{KL}(r_\psi(\mathbf{z}_k|\mathbf{c}) \| p(\mathbf{z}_k))\,, \tag{8}$$

where we use the same Lagrangian multiplier β for all constraints. Despite having similar form, the above objective is very *different* from the objective function of β-VAE [30] in many ways: (1) our goal is to learn a diverse sampling distribution $r_\psi(X|\mathbf{c})$ for a pretrained generative model rather than learning the generative model itself; (2) The first part in our objective is a diversifying term instead of a

reconstruction term; (3) Our objective function applies to most deep generative models, not just VAEs. In this objective, the softening of the hard KL constraints allows for the trade-off between the diversity and likelihood of the samples X. For small β, $r_\psi(\mathbf{z}_k|\mathbf{c})$ is allowed to deviate from $p(\mathbf{z}_k)$ so that $r_\psi(\mathbf{z}_1|\mathbf{c}), \dots, r_\psi(\mathbf{z}_K|\mathbf{c})$ can potentially attend to different regions in the latent space as shown in Fig. 2 (latent space) to further improve sample diversity. For large β, the objective will focus on minimizing the KL term so that $r_\psi(\mathbf{z}_k|\mathbf{c}) \approx p(\mathbf{z}_k)$ and $r_\psi(\mathbf{x}_k|\mathbf{c}) \approx p_\theta(\mathbf{x}_k|\mathbf{c})$, and thus the sample $\mathbf{x}_k$ will have high likelihood under $p_\theta(\mathbf{x}_k|\mathbf{c})$.

The overall DLow objective is defined as:

$$L_{\text{DLow}} = L_{\text{prior}} + \beta L_{\text{KL}}\,, \tag{9}$$

where L_{prior} and L_{KL} are the first and second term in Eq. (8) respectively. In the following, we will discuss in detail how we design the latent mapping functions $\mathcal{T}_{\psi_1}, \dots, \mathcal{T}_{\psi_K}$ and the diversity-promoting prior $p(X)$.

Latent Mapping Functions. Each latent mapping $\mathcal{T}_{\psi_k}$ transforms the Gaussian distribution $p(\boldsymbol{\epsilon})$ to the marginal latent distribution $r_\psi(\mathbf{z}_k|\mathbf{c})$ for latent code $\mathbf{z}_k$ where $\mathcal{T}_{\psi_k}$ is also conditioned on $\mathbf{c}$. As $r_\psi(\mathbf{z}_k|\mathbf{c})$ should stay close to the Gaussian latent prior $p(\mathbf{z}_k)$, it would be ideal if the mapping $\mathcal{T}_{\psi_k}$ makes $r_\psi(\mathbf{z}_k|\mathbf{c})$ also a Gaussian. Thus, we design $\mathcal{T}_{\psi_k}$ to be an invertible affine transformation:

$$\mathcal{T}_{\psi_k}(\boldsymbol{\epsilon}) = \mathbf{A}_k(\mathbf{c})\boldsymbol{\epsilon} + \mathbf{b}_k(\mathbf{c})\,, \tag{10}$$

where the mapping parameters $\psi_k = \{\mathbf{A}_k(\mathbf{c}), \mathbf{b}_k(\mathbf{c})\}$, $\mathbf{A}_k \in \mathbb{R}^{n_z \times n_z}$ is a non-singular matrix, $\mathbf{b}_k \in \mathbb{R}^{n_z}$ is a vector, and n_z is the number of dimensions for $\mathbf{z}_k$ and $\boldsymbol{\epsilon}$. As shown in Fig. 2 and Fig. 3 (Right), we use a K-head network $Q_\gamma(\mathbf{c})$ to output $\psi_1, \dots, \psi_K$, and the parameters γ of the network $Q_\gamma(\mathbf{c})$ are the parameters to be optimized with the DLow objective in Eq. (9).

Under the invertible affine transformation $\mathcal{T}_{\psi_k}$, $r_\psi(\mathbf{z}_k|\mathbf{c})$ becomes a Gaussian distribution $\mathcal{N}(\mathbf{b}_k, \mathbf{A}_k\mathbf{A}_k^T)$. This allows us to compute the KL divergence terms in L_{KL} analytically:

$$\text{KL}(r_\psi(\mathbf{z}_k|\mathbf{c})\|p(\mathbf{z}_k)) = \frac{1}{2}\left(\text{tr}\left(\mathbf{A}_k\mathbf{A}_k^T\right) + \mathbf{b}_k^T\mathbf{b}_k - n_z - \log\det\left(\mathbf{A}_k\mathbf{A}_k^T\right)\right). \tag{11}$$

The KL divergence is minimized when $r_\psi(\mathbf{z}_k|\mathbf{c}) = p(\mathbf{z}_k)$ which implies that $\mathbf{A}_k\mathbf{A}_k^T = \mathbf{I}$ and $\mathbf{b}_k = \mathbf{0}$. Geometrically, this means that $\mathbf{A}_k$ is in the orthogonal group $O(n_z)$, which includes all rotations and reflections in an n_z-dimensional space. This means any mapping $\mathcal{T}_{\psi_k}$ that is a rotation or reflection operation will minimize the KL divergence. As mentioned before, there is a trade-off between diversity and likelihood in Eq. (9). To improve sample diversity (minimize L_{prior}) without compromising likelihood (KL divergence), we can optimize $\mathcal{T}_{\psi_1}, \dots, \mathcal{T}_{\psi_K}$ to be different rotations or reflections to map $\boldsymbol{\epsilon}$ to different feasible points $\mathbf{z}_1, \dots, \mathbf{z}_k$ in the latent space. This geometric understanding sheds light on the mapping space admitted by the hard KL constraints. In practice, we use soft KL constraints in the DLow objective to further enlarge the feasible mapping space which allows us to achieve lower L_{prior} and better sample diversity.

Diversity-Promoting Prior. In the DLow objective, a diversity-promoting prior $p(X)$ on the joint sample distribution is used to guide the optimization of the latent mapping functions $\mathcal{T}_{\psi_1}, \ldots, \mathcal{T}_{\psi_K}$. With an energy-based formulation, the prior $p(X)$ can be defined using an energy function $E(X)$:

$$p(X) = \exp(-E(X))/\mathcal{S}\,, \tag{12}$$

where $\mathcal{S}$ is a normalizing constant. Dropping the constant $\mathcal{S}$, the first term in Eq. (8) can be rewritten as

$$L_{\text{prior}} = \mathbb{E}_{X \sim r_\psi(X|\mathbf{c})}[E(X)]\,. \tag{13}$$

To promote sample diversity of X, we design an energy function $E := E_d$ based on a pairwise distance metric $\mathcal{D}$:

$$E_d(X) = \frac{1}{K(K-1)} \sum_{i=1}^{K} \sum_{j \neq i}^{K} \exp\left(-\frac{\mathcal{D}^2(\mathbf{x}_i, \mathbf{x}_j)}{\sigma_d}\right), \tag{14}$$

where we use the Euclidean distance for $\mathcal{D}$ and an RBF kernel with scale σ_d. Minimizing L_{prior} moves the samples towards a lower-energy (diverse) configuration. L_{prior} can be evaluated efficiently with the reparametrization trick [38].

Up to this point, we have described the proposed sampling method, DLow, for generating a diverse set of samples from a pretrained generative model $p_\theta(\mathbf{x}|\mathbf{c})$. By introducing a common random variable $\boldsymbol{\epsilon}$, DLow allows us to generate correlated samples X. Moreover, by introducing learnable mapping functions $\mathcal{T}_{\psi_k}$, we can model the joint sample distribution $r_\psi(X|\mathbf{c})$ and impose structural constraints, such as diversity, on the sample set X which cannot be modeled by random sampling from the generative model.

4 Diverse Human Motion Prediction

Equipped with a method to generate diverse samples from a pretrained deep generative model, we now turn our attention to the task of diverse human motion prediction. Suppose the pose of a person is a V-dimensional vector consisting of 3D joint positions, we use $\mathbf{c} \in \mathbb{R}^{H \times V}$ to denote the past motion of H time steps and $\mathbf{x} \in \mathbb{R}^{T \times V}$ to denote the future motion over a future time horizon of T. Given a past motion $\mathbf{c}$, the goal of diverse human motion prediction is to generate a diverse set of future motions $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_K\}$.

To capture the multi-modal distribution of the future trajectory $\mathbf{x}$, we take a generative approach and use a conditional variational autoencoder (CVAE) to learn the future trajectory distribution $p_\theta(\mathbf{x}|\mathbf{c})$. Here we use the CVAE for its stability over other popular approaches such as CGANs, but other suitable deep generative models could also be used. The CVAE uses a varitional lower bound [35] as a surrogate for the intractable true data log-likelihood:

$$\mathcal{L}(\mathbf{x}; \theta, \phi) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x},\mathbf{c})}\left[\log p_\theta(\mathbf{x}|\mathbf{z}, \mathbf{c})\right] - \mathrm{KL}\left(q_\phi(\mathbf{z}|\mathbf{x}, \mathbf{c}) \| p(\mathbf{z})\right), \tag{15}$$

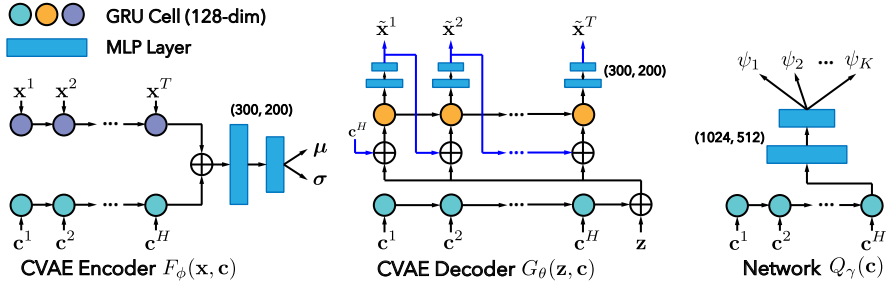

Fig. 3. Network architectures for the CVAE and DLow. We use GRUs [14] to extract motion features. $\mathbf{x}^t$ and $\mathbf{c}^t$ denotes the t-th pose in $\mathbf{x}$ and $\mathbf{c}$ respectively.

where $q_\phi(\mathbf{z}|\mathbf{x}, \mathbf{c})$ is an ϕ-parametrized approximate posterior distribution. We use multivariate Gaussians for the prior, posterior (encoder distribution) and likelihood (decoder distribution): $p(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$, $q_\phi(\mathbf{z}|\mathbf{x}, \mathbf{c}) = \mathcal{N}(\boldsymbol{\mu}, \mathrm{Diag}(\boldsymbol{\sigma}^2))$, and $p_\theta(\mathbf{x}|\mathbf{z}, \mathbf{c}) = \mathcal{N}(\tilde{\mathbf{x}}, \alpha\mathbf{I})$ where α is a hyperparameter. Both the encoder and decoder are implemented as recurrent neural networks (RNNs). As shown in Fig. 3, the encoder network F_ϕ outputs the parameters of the posterior distribution: $(\boldsymbol{\mu}, \boldsymbol{\sigma}) = F_\phi(\mathbf{x}, \mathbf{c})$; the decoder network G_θ outputs the reconstructed future trajectory $\tilde{\mathbf{x}} = G_\theta(\mathbf{z}, \mathbf{c})$. The CVAE is learned via jointly optimizing the encoder and decoder with Eq. (15).

4.1 Diversity Sampling with DLow

Once the CVAE is learned, we follow the DLow framework proposed in Sec. 3 to optimize the network Q_γ (Fig. 3 (Right)) and learn the latent mapping functions $\mathcal{T}_{\psi_1}, \ldots, \mathcal{T}_{\psi_K}$. Before doing this, to fully leverage the DLow framework, we will look at one of DLow's key feature, i.e., the design of the diversity-promoting prior $p(X)$ in L_{prior} can be flexibly changed by modifying the underlying energy function $E(X)$. This allows us to impose various structural constraints besides diversity on the sample set X. Below, we will provide two examples of such prior designs that (1) improve sample accuracy or (2) enable new applications such as controllable motion prediction.

Reconstruction Energy. To ensure that the sample set X is both diverse and accurate, i.e., the ground truth future motion $\hat{\mathbf{x}}$ is close to one of the samples in X, we can modify the prior's energy function E in Eq. (12) by adding a reconstruction term E_r:

$$E(X) = E_d(X) + \lambda_r E_r(X)\,, \tag{16}$$

$$E_r(X) = \min_k \mathcal{D}^2(\mathbf{x}_k, \hat{\mathbf{x}})\,, \tag{17}$$

where λ_r is a weighting factor and we use Euclidean distance as the distance metric $\mathcal{D}$. As DLow produces a correlated set of samples X instead of independent samples, the network Q_γ can learn to distribute samples in a way that are both diverse and accurate, covering the ground truth better. We use this prior design for our main experiments.

Controllable Motion Prediction. Another possible design of the diversity-promoting prior $p(X)$ is one that promotes diversity in a certain subspace of the sample space. In the context of human motion prediction, we may want certain body parts to move similarly but other parts to move differently. For example, we may want leg motion to be similar but upper-body motion to be diverse across motion samples. We call this task controllable motion prediction, i.e., finding a set of diverse samples that share some common features, which can allow users or down-stream systems to explore variations of a certain type of samples.

Formally, we divide the human joints into two sets, J_s and J_d, and ask samples in X to have similar motions for joints J_s but diverse motions for joints J_d. We can slice a motion sample $\mathbf{x}_k$ into two parts: $\mathbf{x}_k = \left(\mathbf{x}_k^s, \mathbf{x}_k^d\right)$ where $\mathbf{x}_k^s$ and $\mathbf{x}_k^d$ correspond to J_s and J_d respectively. Similarly, we can slice the sample set X into two sets: $X_s = \{\mathbf{x}_1^s, \dots, \mathbf{x}_K^s\}$ and $X_d = \{\mathbf{x}_1^d, \dots, \mathbf{x}_K^d\}$. We then define a new energy function E for the prior $p(X)$:

$$E(X) = E_d(X_d) + \lambda_s E_s(X_s) + \lambda_r E_r(X)\,, \tag{18}$$

$$E_s(X_s) = \frac{1}{K(K-1)} \sum_{i=1}^{K} \sum_{j \neq i}^{K} \mathcal{D}^2(\mathbf{x}_i^s, \mathbf{x}_j^s)\,, \tag{19}$$

where we add another energy term E_s weighted by λ_s to minimize the motion distance between samples for joints J_s, and we only compute the diversity-promoting term E_d using motions of joints J_d. After optimizing Q_γ using the DLow objective with the new energy E, we can produce diverse samples X that have similar motions for joints J_s.

Furthermore, we may also want to use a reference motion sample $\mathbf{x}_{\text{ref}}$ to provide the desired features. To achieve this, we can treat $\mathbf{x}_{\text{ref}}$ as the first sample $\mathbf{x}_1$ in X. We first find its corresponding latent code $\mathbf{z}_1 := \mathbf{z}_{\text{ref}}$ using the CVAE encoder: $\mathbf{z}_{\text{ref}} = F_\phi^{\boldsymbol{\mu}}(\mathbf{x}_{\text{ref}}, \mathbf{c})$. We can then find the common variable $\boldsymbol{\epsilon}_{\text{ref}}$ for generating X using the inverse mapping $\mathcal{T}_{\psi_1}^{-1}$:

$$\boldsymbol{\epsilon}_{\text{ref}} = \mathcal{T}_{\psi_1}^{-1}(\mathbf{z}_{\text{ref}}) = \mathbf{A}_1^{-1}(\mathbf{z}_{\text{ref}} - \mathbf{b}_1)\,. \tag{20}$$

With $\boldsymbol{\epsilon}_{\text{ref}}$ known, we can generate X that includes $\mathbf{x}_{\text{ref}}$. In practice, we force $\mathcal{T}_{\psi_1}$ to be an identity mapping to enforce $r_\psi(\mathbf{z}_1|\mathbf{c}) = p(\mathbf{z}_1)$ so that $r_\psi(\mathbf{z}_1|\mathbf{c})$ covers the posterior distribution of $\mathbf{z}_{\text{ref}}$. Otherwise, if $\mathbf{z}_{\text{ref}}$ lies outside of the high density region of $r_\psi(\mathbf{z}_1|\mathbf{c})$, it may lead to low-likelihood $\boldsymbol{\epsilon}_{\text{ref}}$ after the inverse mapping.

5 Experiments

Datasets. We perform evaluation on two public motion capture datasets: Human3.6M [33] and HumanEva-I [63]. Human3.6M is a large-scale dataset with 11 subjects (7 with ground truth) and 3.6 million video frames in total. Each subject performs 15 actions and the human motion is recorded at 50 Hz. Following previous work [53, 49, 73, 56], we adopt a 17-joint skeleton and train on five subjects (S1, S5, S6, S7, S8) and test on two subjects (S9 and S11). HumanEva-I is a relatively small dataset, containing only three subjects recorded at 60 Hz.

We adopt a 15-joint skeleton [56] and use the same train/test split provided in the dataset. By using both a large dataset with more variation in motion and a small dataset with less variation, we can better evaluate the generalization of our method to different types of data. For Human3.6M, we predict future motion for 2 seconds based on observed motion of 0.5 seconds. For HumanEva-I, we forecast future motion for 1 second given observed motion of 0.25 seconds.

Baselines. To fully evaluate our method, we consider three types of baselines: (1) Deterministic motion prediction methods, including **ERD** [17] and **acLSTM** [45]; (2) Stochastic motion prediction methods, including CVAE based methods, **Pose-Knows** [68] and **MT-VAE** [71], as well as a CGAN based method, **HP-GAN** [6]; (3) Diversity-promoting methods for generative models, including **Best-of-Many** [8], **GMVAE** [15], **DeLiGAN** [27], and **DSF** [74].

Metrics. We use the following metrics to measure both sample *diversity* and *accuracy*. (1) **Average Pairwise Distance (APD)**: average $L2$ distance between all pairs of motion samples to measure diversity within samples, which is computed as $\frac{1}{K(K-1)} \sum_{i=1}^{K} \sum_{j \neq i}^{K} \|\mathbf{x}_i - \mathbf{x}_j\|$. (2) **Average Displacement Error (ADE)**: average $L2$ distance over all time steps between the ground truth motion $\hat{\mathbf{x}}$ and the closest sample, which is computed as $\frac{1}{T} \min_{\mathbf{x} \in X} \|\hat{\mathbf{x}} - \mathbf{x}\|$. (3) **Final Displacement Error (FDE)**: $L2$ distance between the final ground truth pose $\mathbf{x}^T$ and the closest sample's final pose, which is computed as $\min_{\mathbf{x} \in X} \|\hat{\mathbf{x}}^T - \mathbf{x}^T\|$. (4) **Multi-Modal ADE (MMADE)**: the multi-modal version of ADE that obtains multi-modal ground truth future motions by grouping similar past motions. (5) **Multi-Modal FDE (MMFDE)**: the multi-modal version of FDE.

In these metrics, APD has been used to measure sample diversity [3]. ADE and FDE are common metrics for evaluating sample accuracy in trajectory forecasting literature [2, 43, 26]. MMADE and MMFDE [74] are metrics used to measure a method's ability to produce multi-modal predictions.

Implementation Details. 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 [37] and a learning rate of 1e-3. The DLow objective in Eq. (9) 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.

5.1 Quantitative Results

We summarize the quantitative results on Human3.6M and HumanEva-I in Table 1. The metrics are computed with the sample set size $K = 50$. For both datasets, we can see that our method, DLow, outperforms all baselines in terms of both sample diversity (APD) and accuracy (ADE, FDE) as well as covering multi-modal ground truth (MMADE, MMFDE). Determinstic methods like

Method	Human3.6M [33]					HumanEva-I [63]				
	APD ↑	ADE ↓	FDE ↓	MMADE ↓	MMFDE ↓	APD ↑	ADE ↓	FDE ↓	MMADE ↓	MMFDE ↓
DLow (Ours)	**11.741**	**0.425**	**0.518**	**0.495**	**0.531**	**4.855**	**0.251**	**0.268**	**0.362**	**0.339**
ERD [17]	0	0.722	0.969	0.776	0.995	0	0.382	0.461	0.521	0.595
acLSTM [45]	0	0.789	1.126	0.849	1.139	0	0.429	0.541	0.530	0.608
Pose-Knows [68]	6.723	0.461	0.560	0.522	0.569	2.308	0.269	0.296	0.384	0.375
MT-VAE [71]	0.403	0.457	0.595	0.716	0.883	0.021	0.345	0.403	0.518	0.577
HP-GAN [6]	7.214	0.858	0.867	0.847	0.858	1.139	0.772	0.749	0.776	0.769
Best-of-Many [8]	6.265	0.448	0.533	0.514	0.544	2.846	0.271	0.279	0.373	0.351
GMVAE [15]	6.769	0.461	0.555	0.524	0.566	2.443	0.305	0.345	0.408	0.410
DeLiGAN [27]	6.509	0.483	0.534	0.520	0.545	2.177	0.306	0.322	0.385	0.371
DSF [74]	9.330	0.493	0.592	0.550	0.599	4.538	0.273	0.290	0.364	0.340

Table 1. Quantitative results on Human3.6M and HumanEva-I.

Energy		Human3.6M [33]					HumanEva-I [63]				
E_d	E_r	APD ↑	ADE ↓	FDE ↓	MMADE ↓	MMFDE ↓	APD ↑	ADE ↓	FDE ↓	MMADE ↓	MMFDE ↓
✓	✓	11.741	**0.425**	**0.518**	**0.495**	**0.531**	4.855	**0.251**	**0.268**	**0.362**	**0.339**
✓	✗	**13.091**	0.546	0.663	0.599	0.669	**4.927**	0.263	0.281	0.368	0.347
✗	✓	6.844	0.432	0.525	0.500	0.539	2.355	0.252	0.277	0.376	0.366
✗	✗	6.383	0.520	0.629	0.577	0.638	2.247	0.281	0.317	0.395	0.393

Table 2. Ablation study on Human3.6M and HumanEva-I.

ERD [17] and acLSTM [45] do not perform well because they only predict one future trajectory which can lead to mode averaging. Methods like MT-VAE [71] produce trajectories samples that lack diversity so they fail to cover the multi-modal ground-truth (indicated by high MMADE and MMFDE) despite having decently low ADE and FDE. We would also like to point out the closest competitor DSF [74] can only generate one deterministic set of samples, while our method can produce multiple diverse sets by sampling $\boldsymbol{\epsilon}$. We also show how each metric changes against various K in Appendix C.

Ablation Study. We further perform an ablation study (Table 2) to analyze the effects of the two energy terms E_d and E_r in Eq. (16). First, without the reconstruction term E_r, the DLow variant is able to achieve higher diversity (APD) at the cost of sample accuracy (ADE, FDE, MMADE, MMFDE). This is expected because the network only optimizes the diversity term E_d and focuses solely on diversity. Second, for the variant without E_d, both sample diversity and accuracy decrease. It is intuitive to see why the diversity (APD) decreases. To see why the sample accuracy (ADE, FDE, MMADE, MMFDE) also decreases, we should consider the fact that a more diverse set of samples have a better chance at covering the ground truth. Finally, when we remove both E_d and E_r (i.e., only optimize L_{KL}), the results are the worst, which is expected.

5.2 Qualitative Results

To visually evaluate the diversity and accuracy of each method, we present a qualitative comparison in Fig. 4 where we render the start pose, the end pose of the ground truth future motion, and the end pose of 10 motion samples. Note that we do not model the global translation of the person, which is why some sitting motions appear to be floating. For Human3.6M, we can see that our

Fig. 4. Qualitative Results on Human3.6M and HumanEva-I.

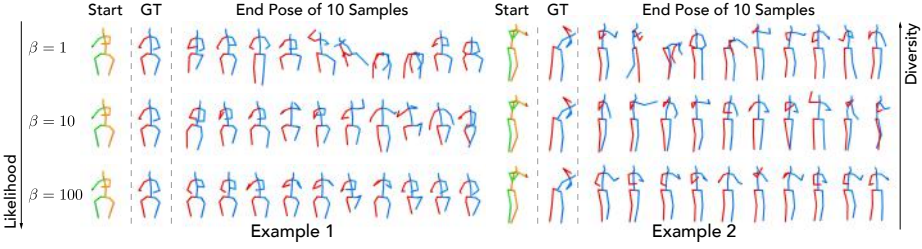

Fig. 5. Varying β in DLow allows us to balance between diversity and likelihood.

method DLow can predict a wide array of future motions, including standing, sitting, bending, crouching, and turning, which cover the ground truth bending motion. In contrast, the baseline methods mostly produce perturbations of a single motion — standing. For HumanEva-I, we can see that DLow produces interesting variations of the fighting motion, while the baselines produce almost identical future motions.

Diversity vs. Likelihood. As discussed in the approach section, the β in Eq. (8) represents the trade-off between sample diversity and likelihood. To verify this, we trained three DLow models with different β (1, 10, 100) and visualize the motion samples generated by each model in Fig. 5. We can see that a larger β leads to less diverse samples which correspond to the major mode of the generator distribution, while a smaller β can produce more diverse motion samples covering other plausible yet less likely future motions.

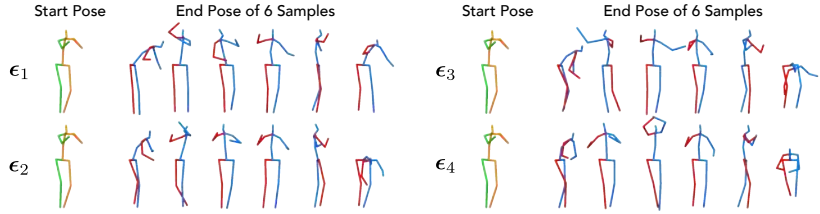

Fig. 6. Effect of varying ϵ on motion samples.

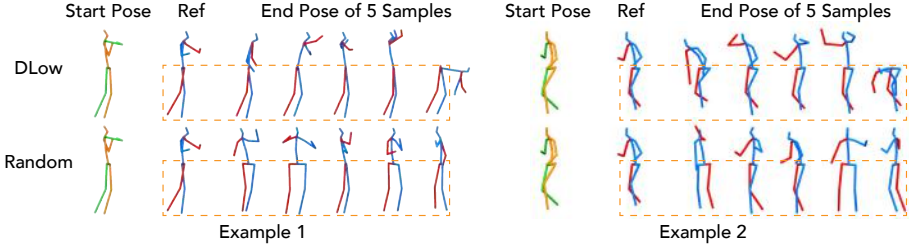

Fig. 7. Controllable Motion Prediction. DLow enables samples to have more similar leg motion to the reference.

Effect of varying ϵ. A key difference between our method and DSF [74] is that we can generate multiple diverse sets of samples while DSF can only produce a fixed diverse set. To demonstrate this, we show in Fig. 6 how the motion samples of DLow change with different ϵ. By comparing the four sets of motion samples, one can conclude that changing ϵ varies each set of samples but preserves the main structure of each motion.

Controllable Motion Prediction. As highlighted before, the flexible design of the diversity-promoting prior enables a new application, controllable motion prediction, where we predict diverse motions that share some common features. We showcase this application by conducting an experiment using the energy function defined in Eq. (18). The network is trained so that the leg motion of the motion samples is similar while the upper-body motion is diverse. The results are shown in Fig. 7. We can see that given a reference motion, our method can generate diverse upper-body motion and preserve similar leg motion, while random samples from the CVAE cannot enforce similar leg motion. Please refer to Appendix B for more results.

6 Conclusion

We have proposed a novel sampling strategy, DLow, for deep generative models to obtain a diverse set of future human motions. We introduced learnable latent mapping functions which allowed us to generate a set of correlated samples, whose diversity can be optimized by a diversity-promoting prior. Experiments demonstrated superior performance in generating diverse motion samples. Moreover, we showed that the flexible design of the diversity-promoting prior further enables new applications, such as controllable human motion prediction. We hope that our exploration of deep generative models through the lens of diversity will

encourage more work towards understanding the complex nature of modeling and predicting future human behavior.

References

1. Aksan, E., Kaufmann, M., Hilliges, O.: Structured prediction helps 3d human motion modelling. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 7144–7153 (2019)
2. Alahi, A., Goel, K., Ramanathan, V., Robicquet, A., Fei-Fei, L., Savarese, S.: Social lstm: Human trajectory prediction in crowded spaces. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 961–971 (2016)
3. Aliakbarian, S., Saleh, F.S., Salzmann, M., Petersson, L., Gould, S.: A stochastic conditioning scheme for diverse human motion prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5223–5232 (2020)
4. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein gan. arXiv preprint arXiv:1701.07875 (2017)
5. Azadi, S., Feng, J., Darrell, T.: Learning detection with diverse proposals. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7149–7157 (2017)
6. Barsoum, E., Kender, J., Liu, Z.: Hp-gan: Probabilistic 3d human motion prediction via gan. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 1418–1427 (2018)
7. Batra, D., Yadollahpour, P., Guzman-Rivera, A., Shakhnarovich, G.: Diverse m-best solutions in markov random fields. In: European Conference on Computer Vision. pp. 1–16. Springer (2012)
8. Bhattacharyya, A., Schiele, B., Fritz, M.: Accurate and diverse sampling of sequences based on a best of many sample objective. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8485–8493 (2018)
9. Butepage, J., Black, M.J., Kragic, D., Kjellstrom, H.: Deep representation learning for human motion prediction and classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6158–6166 (2017)
10. Chao, Y.W., Yang, J., Price, B., Cohen, S., Deng, J.: Forecasting human dynamics from static images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 548–556 (2017)
11. Che, T., Li, Y., Jacob, A.P., Bengio, Y., Li, W.: Mode regularized generative adversarial networks. arXiv preprint arXiv:1612.02136 (2016)
12. Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., Abbeel, P.: Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In: Advances in neural information processing systems. pp. 2172–2180 (2016)
13. Chiu, H.k., Adeli, E., Wang, B., Huang, D.A., Niebles, J.C.: Action-agnostic human pose forecasting. In: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 1423–1432. IEEE (2019)
14. Chung, J., Gulcehre, C., Cho, K., Bengio, Y.: Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555 (2014)
15. Dilokthanakul, N., Mediano, P.A., Garnelo, M., Lee, M.C., Salimbeni, H., Arulkumaran, K., Shanahan, M.: Deep unsupervised clustering with gaussian mixture variational autoencoders. arXiv preprint arXiv:1611.02648 (2016)

16. Elfeki, M., Couprie, C., Riviere, M., Elhoseiny, M.: Gdpp: Learning diverse generations using determinantal point process. arXiv preprint arXiv:1812.00068 (2018)
17. Fragkiadaki, K., Levine, S., Felsen, P., Malik, J.: Recurrent network models for human dynamics. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 4346–4354 (2015)
18. Ghosh, P., Song, J., Aksan, E., Hilliges, O.: Learning human motion models for long-term predictions. In: 2017 International Conference on 3D Vision (3DV). pp. 458–466. IEEE (2017)
19. Gillenwater, J.A., Kulesza, A., Fox, E., Taskar, B.: Expectation-maximization for learning determinantal point processes. In: Advances in Neural Information Processing Systems. pp. 3149–3157 (2014)
20. Gong, B., Chao, W.L., Grauman, K., Sha, F.: Diverse sequential subset selection for supervised video summarization. In: Advances in Neural Information Processing Systems. pp. 2069–2077 (2014)
21. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in neural information processing systems. pp. 2672–2680 (2014)
22. Gopalakrishnan, A., Mali, A., Kifer, D., Giles, L., Ororbia, A.G.: A neural temporal model for human motion prediction. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 12116–12125 (2019)
23. Gu, Q., Li, Z., Han, J.: Generalized fisher score for feature selection. arXiv preprint arXiv:1202.3725 (2012)
24. Guan, J., Yuan, Y., Kitani, K.M., Rhinehart, N.: Generative hybrid representations for activity forecasting with no-regret learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2020)
25. Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A.C.: Improved training of wasserstein gans. In: Advances in neural information processing systems. pp. 5767–5777 (2017)
26. Gupta, A., Johnson, J., Fei-Fei, L., Savarese, S., Alahi, A.: Social gan: Socially acceptable trajectories with generative adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2255–2264 (2018)
27. Gurumurthy, S., Kiran Sarvadevabhatla, R., Venkatesh Babu, R.: Deligan: Generative adversarial networks for diverse and limited data. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 166–174 (2017)
28. Guzman-Rivera, A., Batra, D., Kohli, P.: Multiple choice learning: Learning to produce multiple structured outputs. In: Advances in Neural Information Processing Systems. pp. 1799–1807 (2012)
29. He, J., Spokoyny, D., Neubig, G., Berg-Kirkpatrick, T.: Lagging inference networks and posterior collapse in variational autoencoders. arXiv preprint arXiv:1901.05534 (2019)
30. Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., Lerchner, A.: beta-vae: Learning basic visual concepts with a constrained variational framework. Iclr **2**(5), 6 (2017)
31. Hsiao, W.L., Grauman, K.: Creating capsule wardrobes from fashion images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7161–7170 (2018)
32. Huang, D.A., Ma, M., Ma, W.C., Kitani, K.M.: How do we use our hands? discovering a diverse set of common grasps. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 666–675 (2015)

33. 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)
34. Jain, A., Zamir, A.R., Savarese, S., Saxena, A.: Structural-rnn: Deep learning on spatio-temporal graphs. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5308–5317 (2016)
35. Jordan, M.I., Ghahramani, Z., Jaakkola, T.S., Saul, L.K.: An introduction to variational methods for graphical models. Machine learning **37**(2), 183–233 (1999)
36. Kim, Y., Wiseman, S., Miller, A.C., Sontag, D., Rush, A.M.: Semi-amortized variational autoencoders. arXiv preprint arXiv:1802.02550 (2018)
37. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
38. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
39. Koppula, H.S., Saxena, A.: Anticipating human activities for reactive robotic response. In: IROS. p. 2071. Tokyo (2013)
40. Kulesza, A., Taskar, B.: k-dpps: Fixed-size determinantal point processes. In: Proceedings of the 28th International Conference on Machine Learning (ICML-11). pp. 1193–1200 (2011)
41. Kulesza, A., Taskar, B., et al.: Determinantal point processes for machine learning. Foundations and Trends® in Machine Learning **5**(2–3), 123–286 (2012)
42. Kundu, J.N., Gor, M., Babu, R.V.: Bihmp-gan: bidirectional 3d human motion prediction gan. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 8553–8560 (2019)
43. Lee, N., Choi, W., Vernaza, P., Choy, C.B., Torr, P.H., Chandraker, M.: Desire: Distant future prediction in dynamic scenes with interacting agents. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 336–345 (2017)
44. Lee, S., Prakash, S.P.S., Cogswell, M., Ranjan, V., Crandall, D., Batra, D.: Stochastic multiple choice learning for training diverse deep ensembles. In: Advances in Neural Information Processing Systems. pp. 2119–2127 (2016)
45. Li, Z., Zhou, Y., Xiao, S., He, C., Huang, Z., Li, H.: Auto-conditioned recurrent networks for extended complex human motion synthesis. arXiv preprint arXiv:1707.05363 (2017)
46. Lin, X., Amer, M.R.: Human motion modeling using dvgans. arXiv preprint arXiv:1804.10652 (2018)
47. Lin, Z., Khetan, A., Fanti, G., Oh, S.: Pacgan: The power of two samples in generative adversarial networks. In: Advances in Neural Information Processing Systems. pp. 1498–1507 (2018)
48. Liu, X., Gao, J., Celikyilmaz, A., Carin, L., et al.: Cyclical annealing schedule: A simple approach to mitigating kl vanishing. arXiv preprint arXiv:1903.10145 (2019)
49. Luvizon, D.C., Picard, D., Tabia, H.: 2d/3d pose estimation and action recognition using multitask deep learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5137–5146 (2018)
50. Macchi, O.: The coincidence approach to stochastic point processes. Advances in Applied Probability **7**(1), 83–122 (1975)
51. Mao, W., Liu, M., Salzmann, M., Li, H.: Learning trajectory dependencies for human motion prediction. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 9489–9497 (2019)

52. Martinez, J., Black, M.J., Romero, J.: On human motion prediction using recurrent neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2891–2900 (2017)
53. Martinez, J., Hossain, R., Romero, J., Little, J.J.: A simple yet effective baseline for 3d human pose estimation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2640–2649 (2017)
54. Nilsson, D.: An efficient algorithm for finding the m most probable configurationsin probabilistic expert systems. Statistics and computing **8**(2), 159–173 (1998)
55. Paden, B., Čáp, M., Yong, S.Z., Yershov, D., Frazzoli, E.: A survey of motion planning and control techniques for self-driving urban vehicles. IEEE Transactions on intelligent vehicles **1**(1), 33–55 (2016)
56. Pavllo, D., Feichtenhofer, C., Grangier, D., Auli, M.: 3d human pose estimation in video with temporal convolutions and semi-supervised training. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7753–7762 (2019)
57. Pavllo, D., Grangier, D., Auli, M.: Quaternet: A quaternion-based recurrent model for human motion. arXiv preprint arXiv:1805.06485 (2018)
58. Rezende, D.J., Mohamed, S.: Variational inference with normalizing flows. arXiv preprint arXiv:1505.05770 (2015)
59. Rhinehart, N., Kitani, K.M., Vernaza, P.: R2p2: A reparameterized pushforward policy for diverse, precise generative path forecasting. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 772–788 (2018)
60. Rissanen, J.J.: Fisher information and stochastic complexity. IEEE transactions on information theory **42**(1), 40–47 (1996)
61. Ruiz, A.H., Gall, J., Moreno-Noguer, F.: Human motion prediction via spatio-temporal inpainting. arXiv preprint arXiv:1812.05478 (2018)
62. Seroussi, B., Golmard, J.L.: An algorithm directly finding the k most probable configurations in bayesian networks. International Journal of Approximate Reasoning **11**(3), 205–233 (1994)
63. 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 (2010)
64. Sourati, J., Akcakaya, M., Erdogmus, D., Leen, T.K., Dy, J.G.: A probabilistic active learning algorithm based on fisher information ratio. IEEE transactions on pattern analysis and machine intelligence **40**(8), 2023–2029 (2017)
65. Srivastava, A., Valkov, L., Russell, C., Gutmann, M.U., Sutton, C.: Veegan: Reducing mode collapse in gans using implicit variational learning. In: Advances in Neural Information Processing Systems. pp. 3308–3318 (2017)
66. Tolstikhin, I., Bousquet, O., Gelly, S., Schoelkopf, B.: Wasserstein auto-encoders. arXiv preprint arXiv: 1711.01558 (2017)
67. Troje, N.F.: Decomposing biological motion: A framework for analysis and synthesis of human gait patterns. Journal of vision **2**(5), 2–2 (2002)
68. Walker, J., Marino, K., Gupta, A., Hebert, M.: The pose knows: Video forecasting by generating pose futures. In: Proceedings of the IEEE international conference on computer vision. pp. 3332–3341 (2017)
69. Wang, B., Adeli, E., Chiu, H.k., Huang, D.A., Niebles, J.C.: Imitation learning for human pose prediction. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 7124–7133 (2019)
70. Weng, X., Yuan, Y., Kitani, K.: Joint 3d tracking and forecasting with graph neural network and diversity sampling. arXiv:2003.07847 (2020)

71. Yan, X., Rastogi, A., Villegas, R., Sunkavalli, K., Shechtman, E., Hadap, S., Yumer, E., Lee, H.: Mt-vae: Learning motion transformations to generate multimodal human dynamics. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 265–281 (2018)
72. Yang, D., Hong, S., Jang, Y., Zhao, T., Lee, H.: Diversity-sensitive conditional generative adversarial networks. arXiv preprint arXiv:1901.09024 (2019)
73. Yang, W., Ouyang, W., Wang, X., Ren, J., Li, H., Wang, X.: 3d human pose estimation in the wild by adversarial learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5255–5264 (2018)
74. Yuan, Y., Kitani, K.: Diverse trajectory forecasting with determinantal point processes. arXiv preprint arXiv:1907.04967 (2019)
75. Yuan, Y., Kitani, K.: Ego-pose estimation and forecasting as real-time pd control. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 10082–10092 (2019)
76. Yuan, Y., Kitani, K.: Residual force control for agile human behavior imitation and extended motion synthesis. arXiv preprint arXiv:2006.07364 (2020)
77. Zhang, J.Y., Felsen, P., Kanazawa, A., Malik, J.: Predicting 3d human dynamics from video. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 7114–7123 (2019)
78. Zhao, S., Song, J., Ermon, S.: Infovae: Information maximizing variational autoencoders. arXiv preprint arXiv:1706.02262 (2017)

A Additional Human3.6M Results

In this section, we show more qualitative results on Human3.6M, including additional comparison with baselines (Fig. 8) and additional examples of DLow (Fig. 9). Please refer to the video to see the whole motion sequences.

A.1 Additional Comparison with Baselines on Human3.6M

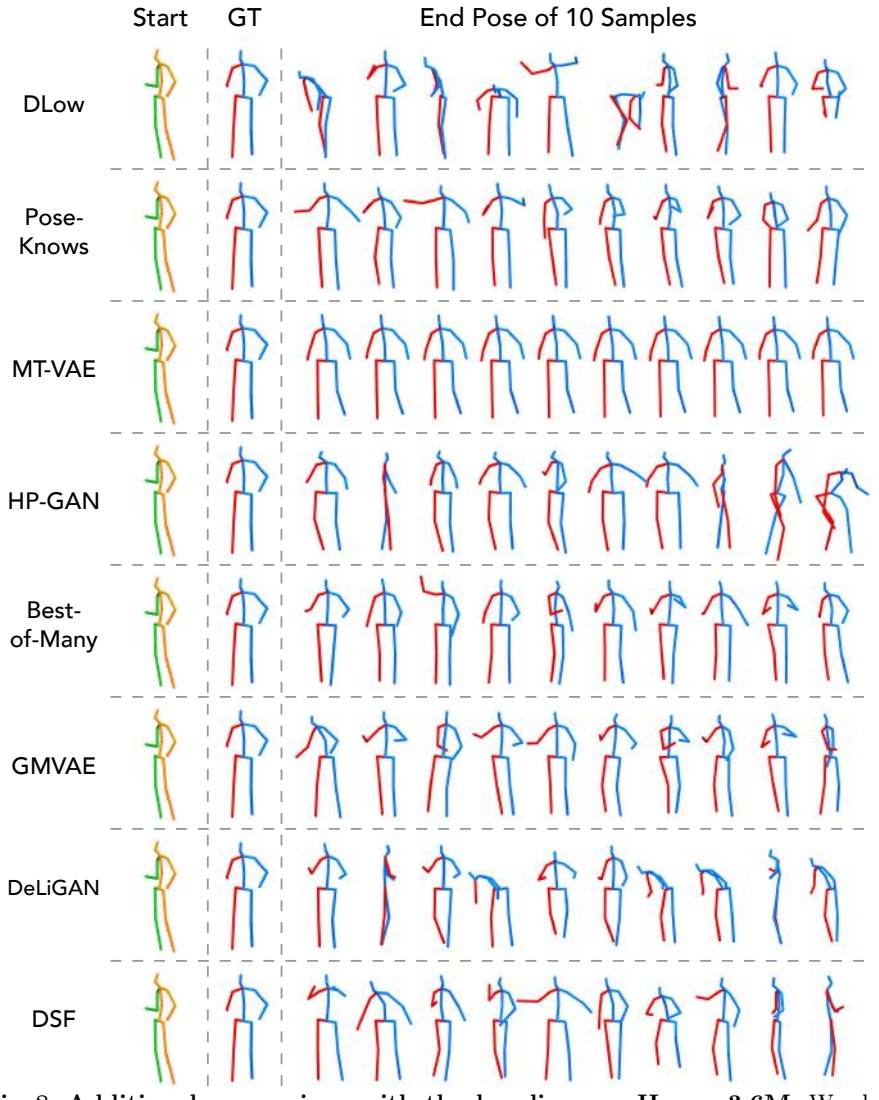

Fig. 8. Additional comparison with the baselines on Human3.6M. We show the start pose, the end pose of the ground truth future motion, and the end pose of 10 motion samples by each method.

A.2 Additional Examples of DLow on Human3.6M

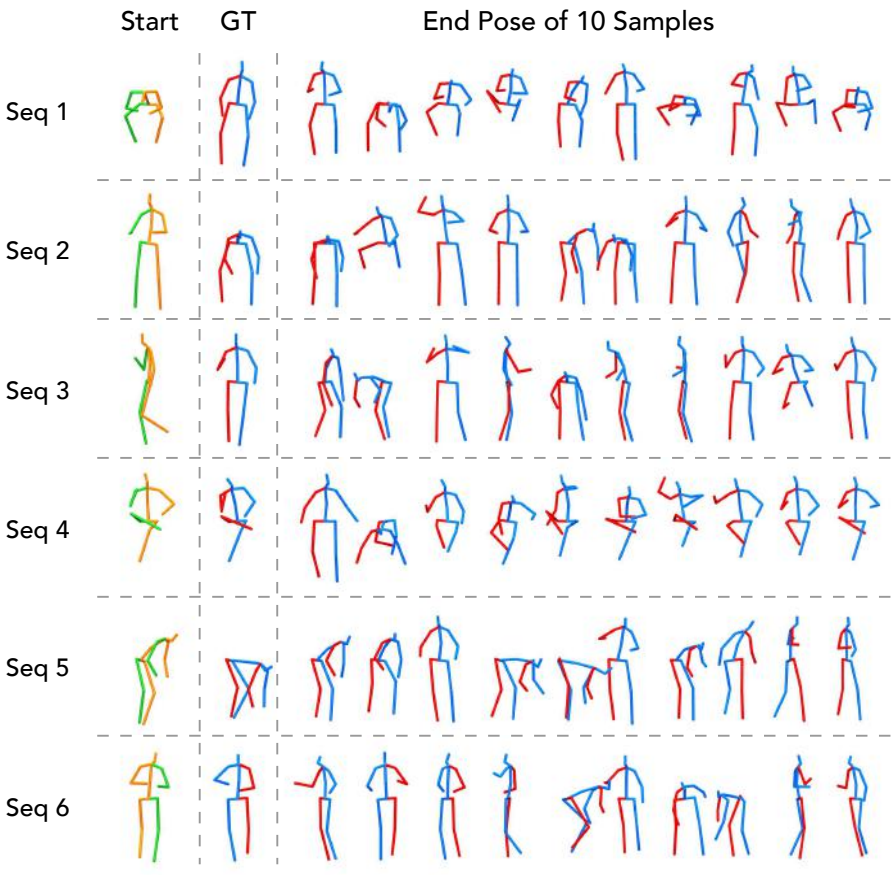

Fig. 9. Additional examples of DLow on Human3.6M. Each row corresponds to a different sequence, where we show the start pose, the end pose of the ground truth future motion, and the end pose of 10 motion samples.

B Additional Controllable Motion Prediction Results

In Fig. 10, 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 3, 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: $L(\psi) = \beta L_{\mathrm{KL}} + \lambda_d E_d + \lambda_s E_s + \lambda_r E_r$. 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 3. Quantitative results for controllable motion prediction.

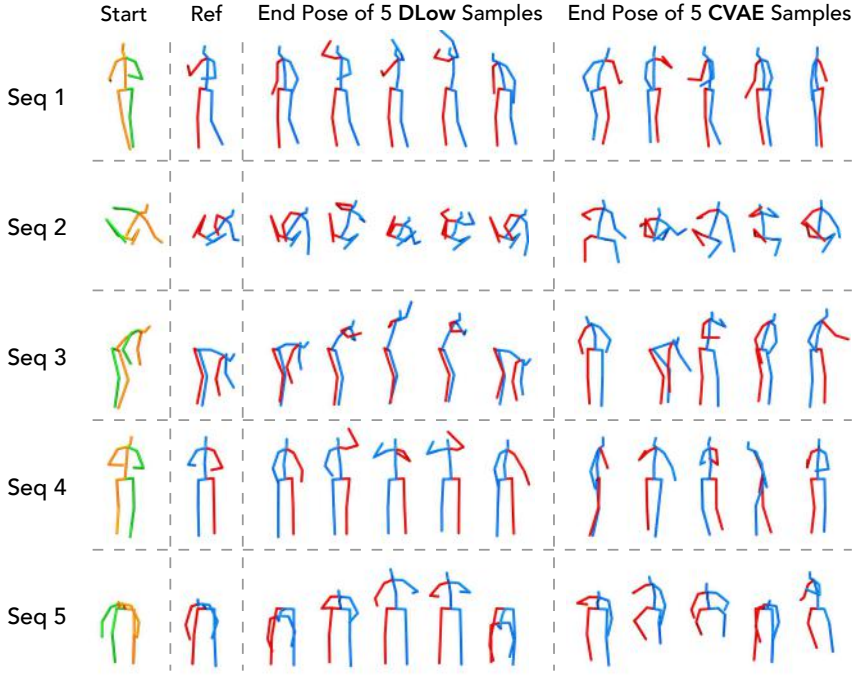

Fig. 10. Additional results on controllable motion prediction. DLow can produce motion samples that have similar leg motion to the reference (Ref) yet diverse upper-body motion, while CVAE (random) samples cannot enforce similar leg motion.

C Metrics vs. Number of Samples K

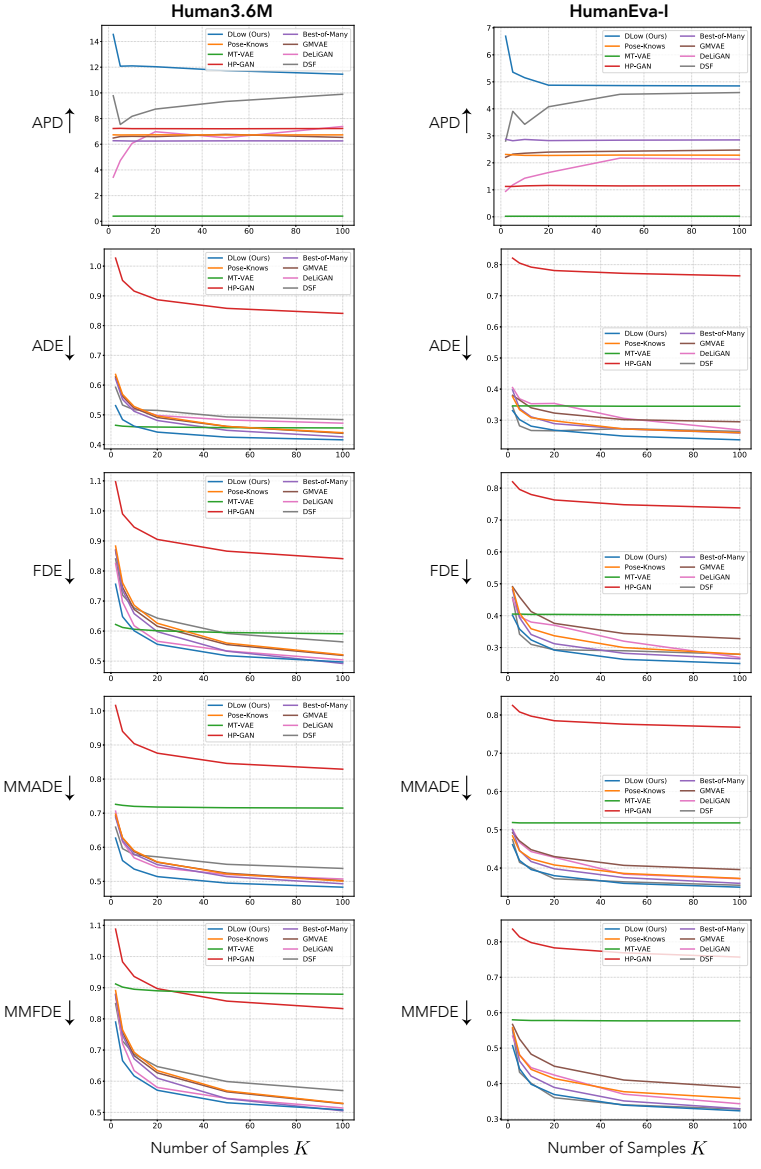

Fig. 11. Metrics vs. Number of Samples K on both Human3.6M (Left) and HumanEva-I (Right).

D 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. 12) and additional examples of DLow (Fig. 13).

D.1 Additional Comparison with Baselines on HumanEva-I

Fig. 12. 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.

D.2 Additional Examples of DLow on HumanEva-I

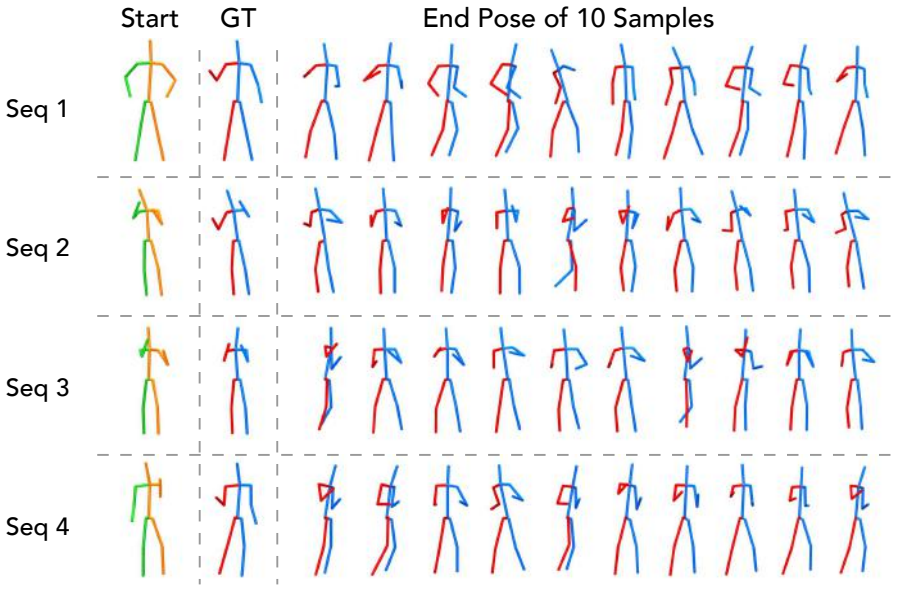

Fig. 13. Additional examples of DLow on HumanEva-I. Each row corresponds to a different sequence, where we show the start pose, the end pose of the ground truth future motion, and the end pose of 10 motion samples.