Diverse Human Motion Prediction via Gumbel-Softmax Sampling from an Auxiliary Space

Lingwei Dang
South China University of Technology
Guangzhou, Guangdong, China
csdanglw@mail.scut.edu.cn

Yongwei Nie*
South China University of Technology
Guangzhou, Guangdong, China
nieyongwei@scut.edu.cn

Chengjiang Long
Meta Reality Lab
Burlingame, CA, USA
clong1@fb.com

Qing Zhang
Sun Yat-sen University
Guangzhou, Guangdong, China
zhangqing.whu.cs@gmail.com

Guiqing Li
South China University of Technology
Guangzhou, Guangdong, China
ligq@scut.edu.cn

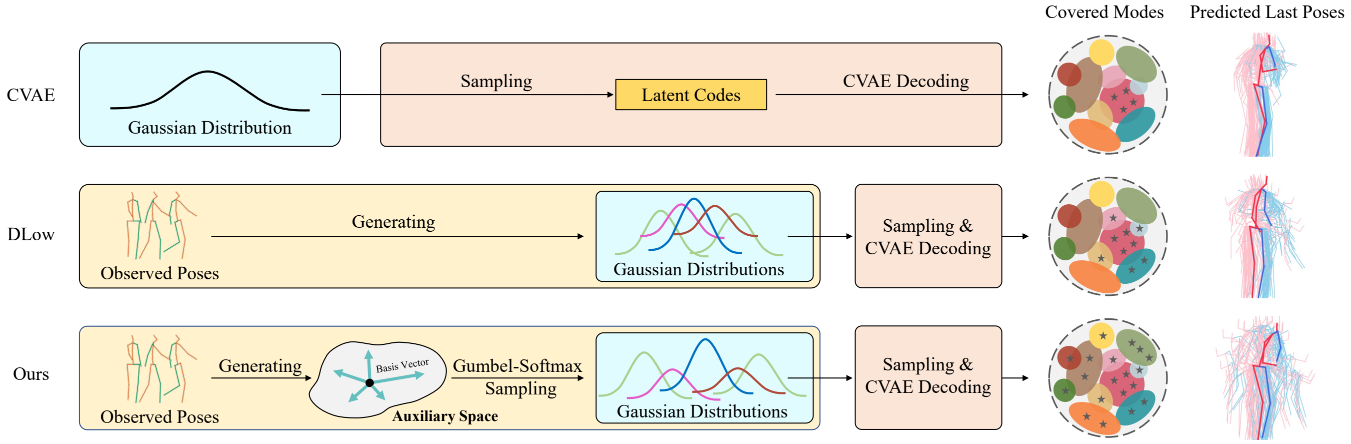

Figure 1: Different strategies for sampling diverse results from an imbalanced multimodal distribution. The vanilla CVAE model randomly samples latent codes from a prior distribution which are then decoded into results that only reside in the major mode of the target distribution. DLow [52] first generates multiple Gaussian distributions, and then samples latent codes from different Gaussian priors. The Gaussian priors can be viewed as corresponding to different modes of the target distribution, therefore this method can cover more modes than random sampling. Our method generates multiple Gaussian distributions by sampling points from an auxiliary space. Due to the high flexibility and capacity of the space, our method is able to cover even more modes than DLow. The rightmost are the last poses of future pose sequences predicted from a given input, all stacked together to visually show that our results are more diverse than the others.

ABSTRACT

Diverse human motion prediction aims at predicting multiple possible future pose sequences from a sequence of observed poses. Previous approaches usually employ deep generative networks to model the conditional distribution of data, and then randomly sample outcomes from the distribution. While different results can be obtained, they are usually the most likely ones which are not diverse enough. Recent work explicitly learns multiple modes of the conditional distribution via a deterministic network, which however can only cover a fixed number of modes within a limited range. In this paper, we propose a novel sampling strategy for sampling very diverse results from an imbalanced multimodal distribution learned by a deep generative model. Our method works by generating an auxiliary space and smartly making randomly sampling from the auxiliary space equivalent to the diverse sampling from the target distribution. We propose a simple yet effective network architecture that implements this novel sampling strategy, which incorporates a Gumbel-Softmax coefficient matrix sampling method and an aggressive diversity promoting hinge loss function. Extensive experiments demonstrate that our method significantly improves both the diversity and accuracy of the samplings compared with previous state-of-the-art sampling approaches. Code and pre-trained models are available at https://github.com/Droliven/diverse_sampling.

*Corresponding author.

MM '22, October 10–14, 2022, Lisboa, Portugal
© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-9203-7/22/10…$15.00
https://doi.org/10.1145/3503161.3547956

CCS CONCEPTS

• **Computing methodologies** → **Computer vision**; **Activity recognition and understanding**.

KEYWORDS

Human motion prediction, stochastic prediction, diverse prediction

ACM Reference Format:

Lingwei Dang, Yongwei Nie, Chengjiang Long, Qing Zhang, and Guiqing Li. 2022. Diverse Human Motion Prediction via Gumbel-Softmax Sampling from an Auxiliary Space. In *Proceedings of the 30th ACM International Conference on Multimedia (MM '22), October 10–14, 2022, Lisboa, Portugal.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3503161.3547956

1 INTRODUCTION

Human Motion Prediction (HMP) has a wide range of applications in autonomous driving, human-robot interaction, and animation creation. Most previous works [1–3, 9, 11–15, 19, 25–28, 31, 32, 34, 35, 37, 39–41, 45, 46] perform deterministic HMP that only generates one result in the future. Recently, many diverse HMP approaches can predict multiple possible future motions. Due to the stochasticity of human motion, multiple solutions naturally exist, and forecasting them is of great importance in practice. For example, it would be better for a vehicle to know that a pedestrian in front of it may not only walk ahead but also turn left suddenly.

Diverse HMP approaches like [6, 24, 50] adopt deep generative networks, such as GAN [20] or CVAE [23], to learn a conditional distribution of future poses given previous ones. Taking CVAE as an example (see Figure 1 top), after training a CVAE, one can randomly sample latent codes (noises) from a prior distribution (*e.g.*, a Gaussian prior), and then decode the random noises to future sequences by the CVAE decoder. However, since the CVAE model is obtained by maximizing the likelihood of the training data that is often highly imbalanced, it usually learns an imbalanced multimodal conditional distribution. Latent codes drawn at random from the prior distribution most probably correspond to the most likely results that fall in the dominant mode of the distribution of data, while ignoring other results of low probability but high fidelity.

Recently, Yuan *et al.* [52] proposed a method called DLow sampling. As shown in the second row of Figure 1, given observed poses, DLow uses a neural network to generate multiple Gaussian distributions, and then samples latent codes from all the generated Gaussian distributions. They optimize the network to diversify the Gaussian distributions, making them corresponding to different modes of the target distribution. However, directly generating Gaussian distributions has two limitations. Firstly, a network can only generate a fixed number of Gaussian distributions, while there may exist much more modes in the target data distribution. Secondly, it entangles the performance of diverse prediction with the learning of the network parameters, requiring the latter to consider all training data and make tradeoffs between them, thus in turn limiting the diverse prediction performance.

In this paper, we propose a sampling strategy that disentangles the above direct dependency between a network and the intermediate Gaussian distributions. As shown in the third row of Figure 1, instead of Gaussian distributions, we learn a set of basis vectors from the observed poses. We assume that the basis vectors determine an auxiliary space, and any linear combination of the basis vectors corresponds to a point in the auxiliary space. We randomly sample a set of points from the auxiliary space by the Gumbel-Softmax sampling strategy, and then map them to Gaussian distributions which finally correspond to different modes of the target distribution. In other words, we use a network to learn an auxiliary space, and build the following connection between the auxiliary space and the target distribution: *randomly sampling from the auxiliary space corresponds to diverse sampling from the target distribution.* The diverse prediction is now tied to the structure of the auxiliary space rather than directly to the parameters of a network. Since the auxiliary space can be flexibly deformed in terms of both size and shape, our sampling method can cover all modes of the target distribution in theory, supporting very diverse human motion prediction. Note that at the training stage, we sample a fixed number of points from the auxiliary space, which facilitates the training of the auxiliary space. After training, since the shape of the auxiliary space has already been constructed, we can sample any number of points from it.

The Gumbel-Softmax sampling method samples points from the auxiliary space by generating a coefficient matrix that linearly combines the basis vectors. There exist other sampling strategies such as Uniform-Softmax sampling and Gaussian-Softmax sampling. We compare with them and find that the Gumbel-Softmax sampling is more effective as it is more aggressive in assigning larger weights to relatively fewer basis vectors. Training/testing with these weights can make better use of each basis vector to sample more distinctive points from the auxiliary space that correspond to more diverse modes of the target distribution. Finally, In order to train our model, we propose a hinge-diversity loss function which explicitly requires the distance between any pair of predictions to be greater than a user-specified threshold. The hinge-diversity loss further strengthens the diversity of predictions while less affecting their accuracy.

In summary, the contributions of this work are three-fold:

- We propose a novel sampling method that is highly capable and convenient for diverse and accurate sampling from a complex imbalanced multimodal distribution, by converting sampling from the distribution into randomly sampling of points from an auxiliary space.
- We propose a Gumbel-Softmax sampling method to sample points from the auxiliary space, and a hinge-diversity loss to train our framework, both of which further improve the performance of our method.
- Extensive comparisons and ablation experimental results conducted on Human3.6M [22] and HumanEva-I [44] demonstrate the effectiveness of our approach.

2 RELATED WORK

Deterministic Human Motion Prediction. Most previous approaches target deterministic HMP, by which only one output is produced per sequence of historical poses. Considering the ability of Recurrent Neural Networks (RNNs) in modeling temporal dependencies of sequential data, many approaches [3, 11, 12, 19, 28, 32, 39, 41, 45] use RNNs to tackle the sequence-to-sequence HMP problem, which however usually suffer from problems of discontinuity and error accumulation. Instead of RNNs, recent works

[13–15, 25–27, 31, 34, 35, 37, 38] employ Graph Convolutional Networks (GCNs) [18, 42, 43] for this task, as GCNs are effective in discovering spatial and temporal relations between pairs of human joints. Similar to GCNs, Transformer [17, 48] can capture long-term dependencies between human joints, and has been adapted to handle the deterministic HMP problem [1, 2, 9, 40]. Different from the above methods, this paper attempts to tackle stochastic HMP which outputs multiple possible results given one input.

Diverse Human Motion Prediction. Many efforts have been paid to the diverse HMP problem [4–6, 10, 24, 29, 33, 36, 47, 50, 52]. For example, Barsoum *et al.* [6] proposed HP-GAN which is a generative adversarial framework that models the probability density function of future human poses conditioned on given poses. At test time, a random vector **z** controls the generation of different future poses. Yan *et al.* [50] proposed MT-VAE using VAE [23, 30] to model the conditional distribution of data. In MT-VAE, a random variable **z** encodes a latent transformation that transforms the observed poses to specific future poses. Both GANs and VAEs randomly sample latent vectors **z** from a prior distribution which however are usually decoded into similar results. To alleviate the problem, Kundu *et al.* [24] proposed BiHMP-GAN in which a discriminator is used to regress the random vector **z** originally fed into the generator, enforcing one-to-one mapping between the latent vector **z** and the corresponding motion prediction. Aliakbarian *et al.* [5] believed that the generative models tend to ignore the random vectors. To prevent such ignoring, they proposed a Mix-and-Match perturbation mechanism to sufficiently mix random noises and conditional poses in [5]. In their later work [4], a random noise is generated directly conditioned on the input poses. Instead of randomly sampling **z**, Yuan *et al.* [52] proposed a sampling strategy called DLow by which different random vectors that correspond to diverse predictions are explicitly inferred, achieving impressive results in sampling from minor modes. However, DLow is limited by its design of inferring random vectors directly from a network. Our method disentangles this dependency and obtains more diverse results with higher accuracy. Recently, the method of [36] directly maps a random vector together with the observed poses to a future sequence, without relying on generative models. For results of different random vectors but the same input poses, it applies a diversity loss to enlarge the differences between them, and meanwhile uses many prior constraints to guarantee their plausibility. We compare with this very different method and show that our method outperforms it on diversity and accuracy metrics.

3 METHODOLOGY

We use CVAE to model the distribution of data, then propose a post-hoc sampling strategy to sample diverse results from the distribution. We therefore first introduce the background of CVAE-based stochastic HMP. Since our method is based on DLow [52], we also briefly introduce DLow, and finally describe our method in detail.

3.1 Background

3.1.1 CVAE-based Stochastic Prediction. Let $p(\mathbf{y}|\mathbf{x})$ denote the distribution of $\mathbf{y}$ given $\mathbf{x}$, where $\mathbf{x}$ is an observed pose sequence and $\mathbf{y}$ is a possible future pose sequence that may appear after $\mathbf{x}$. To sample $\mathbf{y}$ from $p(\mathbf{y}|\mathbf{x})$, one usually introduces a latent variable $\mathbf{z}$ and reparameterizes $p(\mathbf{y}|\mathbf{x})$ as $p(\mathbf{y}|\mathbf{x}) = \int p(\mathbf{y}|\mathbf{x},\mathbf{z})p(\mathbf{z})d\mathbf{z}$. Then, $\mathbf{y}$ can be generated in two steps:

$$\begin{aligned} &\mathbf{z} \sim p(\mathbf{z}), \\ &\mathbf{y} = \mathcal{G}_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{z}), \end{aligned} \tag{1}$$

where a random vector $\mathbf{z}$ is sampled from a prior distribution $p(\mathbf{z})$ (*e.g.*, Gaussian) at first, then $\mathbf{y}$ is generated by a deterministic function $\mathcal{G}_{\boldsymbol{\theta}}$ parameterized by $\boldsymbol{\theta}$ taking $\mathbf{x}$ and $\mathbf{z}$ as input. To learn $\mathcal{G}_{\boldsymbol{\theta}}$, a popular way is to use a CVAE which maximizes the log-likelihood of data $\mathbf{y}$ given $\mathbf{x}$, by introducing an approximate posterior $q(\mathbf{z}|\mathbf{x},\mathbf{y})$ and maximizing the following evidence lower bound (ELBO):

$$\begin{aligned} &\log p(\mathbf{y}|\mathbf{x}) = \log \int p(\mathbf{y}|\mathbf{x},\mathbf{z})p(\mathbf{z})d\mathbf{z} \\ &= \log \int \frac{p(\mathbf{y}|\mathbf{x},\mathbf{z})p(\mathbf{z})}{q(\mathbf{z}|\mathbf{x},\mathbf{y})} q(\mathbf{z}|\mathbf{x},\mathbf{y})d\mathbf{z} \geq \mathbb{E}_q \log \frac{p(\mathbf{y}|\mathbf{x},\mathbf{z})p(\mathbf{z})}{q(\mathbf{z}|\mathbf{x},\mathbf{y})}. \end{aligned} \tag{2}$$

CVAE models the two distributions of $q(\mathbf{z}|\mathbf{x},\mathbf{y})$ and $p(\mathbf{y}|\mathbf{x},\mathbf{z})$ by two neural networks $\mathcal{F}_{\boldsymbol{\phi}}$ and $\mathcal{G}_{\boldsymbol{\theta}}$, and estimates their parameters by optimizing the following loss function:

$$\mathcal{L}(\mathbf{y},\boldsymbol{\phi},\boldsymbol{\theta}) = -\mathcal{KL}\left(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x},\mathbf{y})\|p(\mathbf{z})\right) + \mathbb{E}_{q_{\boldsymbol{\phi}}} \log p_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{x},\mathbf{z}). \tag{3}$$

During training, the encoder $\mathcal{F}_{\boldsymbol{\phi}}$ first generates $\mathbf{z}$ given $\mathbf{x}$ and $\mathbf{y}$, and then the decoder $\mathcal{G}_{\boldsymbol{\theta}}$ reconstructs the input $\mathbf{y}$ given $\mathbf{z}$ and $\mathbf{x}$. At test time, one can sample a $\mathbf{z}$ from the prior distribution $p(\mathbf{z})$, and then predict a $\tilde{\mathbf{y}}$ by $\mathcal{G}_{\boldsymbol{\theta}}$ given $\mathbf{z}$ and $\mathbf{x}$. For multiple predictions, one needs to sample $\mathbf{z}_1, \cdots, \mathbf{z}_K$ independently, and predict $\tilde{\mathbf{y}}_1, \cdots, \tilde{\mathbf{y}}_K$ accordingly using the same $\mathbf{x}$. However, extensive experiments demonstrate that the diversity of $\tilde{\mathbf{y}}_1, \cdots, \tilde{\mathbf{y}}_K$ is not satisfactory.

3.1.2 DLow Sampling. To enable diverse prediction, Yuan *et al.* [52] proposed DLow. Given $\mathbf{x}$, they used a network $Q_{\boldsymbol{\psi}}$ parameterized by $\boldsymbol{\psi}$ to generate K Gaussian distributions:

$$\{(\mathbf{A}_k, \mathbf{b}_k)\}_{k=1}^{K} = Q_{\boldsymbol{\psi}}(\mathbf{x}), \tag{4}$$

where $\mathbf{A}_k \in \mathbb{R}^{n_z \times n_z}$, $\mathbf{b}_k \in \mathbb{R}^{n_z}$ are variance and mean of the k^{th} Gaussian distribution, and n_z is the dimension size. Then, they predicted K results by:

$$\begin{aligned} &\boldsymbol{\epsilon} \sim \mathcal{N}(0,1), \\ &\mathbf{z}_k = \mathbf{A}_k\boldsymbol{\epsilon} + \mathbf{b}_k, \quad 1 \leq k \leq K, \\ &\mathbf{y}_k = \mathcal{G}_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{z}_k), \quad 1 \leq k \leq K. \end{aligned} \tag{5}$$

Compared with Eq. 1, the above DLow sampling method learns K Gaussian distributions and uses the reparameterization trick to sample latent variables from these distributions: $\mathbf{z}_k \sim \mathcal{N}(\mathbf{b}_k, \mathbf{A}_k)$, and finally maps $\mathbf{z}_k$ and the input poses $\mathbf{x}$ to future poses using $\mathcal{G}_{\boldsymbol{\theta}}$ that has already been learned by the CVAE model.

More formally, DLow samples a result $\mathbf{y}_k$ from the distribution of $r_{\boldsymbol{\psi}}(\mathbf{y}_k|\mathbf{x}) = \int p_{\boldsymbol{\theta}}(\mathbf{y}_k|\mathbf{x},\mathbf{z}_k)r_{\boldsymbol{\psi}}(\mathbf{z}_k|\mathbf{x})d\mathbf{z}_k$ where $p_{\boldsymbol{\theta}}(\mathbf{y}_k|\mathbf{x},\mathbf{z}_k)$ is the conditional distribution modeled by $\mathcal{G}_{\boldsymbol{\theta}}$, and $r_{\boldsymbol{\psi}}(\mathbf{z}_k|\mathbf{x})$, *i.e.*, $\mathcal{N}(\mathbf{b}_k, \mathbf{A}_k)$, is the latent distribution modeled by the network $Q_{\boldsymbol{\psi}}$.

To train $Q_{\boldsymbol{\psi}}$, DLow minimizes the following diversity loss to enlarge the distances between pairs of results predicted from the same input $\mathbf{x}$:

$$\mathcal{L}_{div} = \frac{1}{K(K-1)} \sum_{i=1}^{K} \sum_{j \neq i}^{K} \exp\left(-\frac{\mathcal{D}^2(\tilde{\mathbf{y}}_i, \tilde{\mathbf{y}}_j)}{\sigma}\right), \tag{6}$$

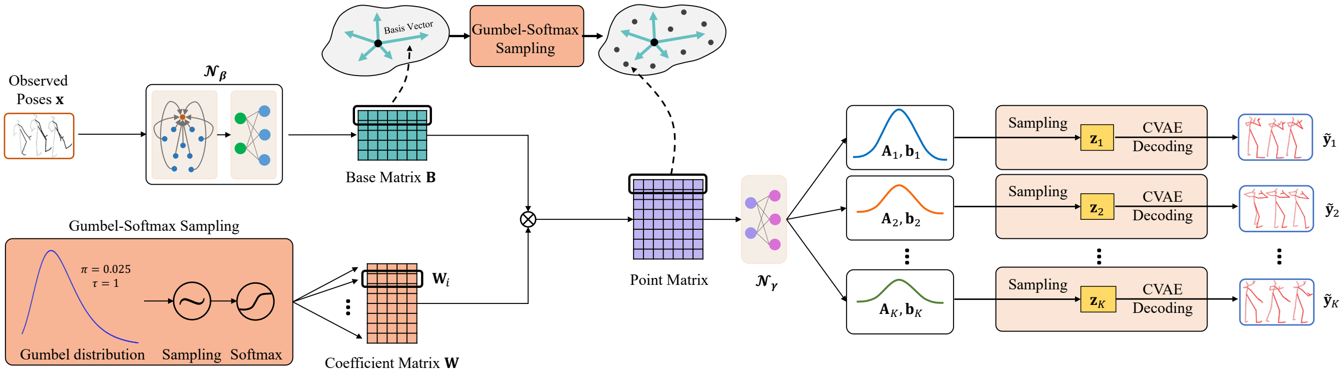

Figure 2: On one hand, we use a network $\mathcal{N}_\beta$ to generate a base matrix from the observed poses. On the other hand, we employ the Gumbel-Softmax sampling method to generate a coefficient matrix. The multiplication of the two matrices samples multiple points from the auxiliary space determined by the base matrix. We then employ another network $\mathcal{N}_\gamma$ to map these points to a set of Gaussian distributions from which latent codes are drawn and finally decoded into future pose sequences.

where $\tilde{\mathbf{y}}_\mathbf{i}$ or $\tilde{\mathbf{y}}_\mathbf{j}$ denotes a predicted pose sequence, and $\mathcal{D}(\cdot,\cdot)$ calculates the Euclidean distance between two predictions. Besides, the following accuracy loss is minimized:

$$\mathcal{L}_{acc} = \min_k \|\mathbf{y} - \tilde{\mathbf{y}}_k\|_2, k \in [1, K]. \tag{7}$$

This loss computes $\mathcal{L}_2$ distance between every prediction $\tilde{\mathbf{y}}_k$ and the ground truth $\mathbf{y}$ and returns the minimum one. Minimizing this loss makes at least one of the predictions similar to the ground truth. Finally, a Kullback-Leibler divergence loss is imposed:

$$\mathcal{L}_{KL} = \mathcal{KL}\left(r_\psi(\mathbf{z}_k|\mathbf{x})||p(\mathbf{z})\right), k \in [1, K], \tag{8}$$

where $p(\mathbf{z})$ is the prior distribution used to train the CVAE. This loss makes $\mathbf{z}_k$ correspond to a high-likelihood sample $\mathbf{y}_k$ under the generative model $p_\theta(\mathbf{y}|\mathbf{x},\mathbf{z})$, guaranteeing the plausibility of the predicted poses.

3.2 Our method

While DLow improves sampling diversity compared to the random sampling method, we observe that its effectiveness is limited in two ways. (1) Firstly, DLow entangles its prediction performance with the learning of the network Q_ψ. However, the network is trained on all the training data, hence its performance is inevitably averaged over all the data, reducing its ability to make extreme predictions existing at minor modes. (2) Secondly, due to the entanglement, DLow can only sample K predictions at a time. However, it is more preferable that a sampling method can sample any number of samples at test time.

To solve these problems, we propose a new sampling method which disentangles the direct correlation between the tasks of diverse prediction and the network parameter learning. Figure 2 illustrates the sampling process of our method. On one hand, we design a network $\mathcal{N}_\beta$ parameterized by $\boldsymbol{\beta}$ that takes $\mathbf{x}$ as input and outputs a base matrix $\mathbf{B} \in \mathbb{R}^{M \times n_b}$:

$$\mathbf{B} = \mathcal{N}_\beta(\mathbf{x}), \tag{9}$$

where each row of $\mathbf{B}$ is a basis vector of dimension n_b and there are M basis vectors in total. One can imagine that the basis vectors together form a space which we call “auxiliary space” in this paper. A point in the space can be obtained by linearly combining the basis vectors. On the other hand, we use the Gumbel-Softmax sampling strategy (introduced later) to sample a coefficient matrix $\mathbf{W} \in \mathbb{R}^{K \times M}$ in which each row $\mathbf{W}_i$ $(i \in [1, K])$ contains M weights used to combine the basis vectors. These weights should satisfy $\sum_{j=1}^{M} \mathbf{W}_{i,j} = 1$. Then, we multiply $\mathbf{W}$ and $\mathbf{B}$ together to obtain a point matrix $\mathbf{WB} \in \mathbb{R}^{K \times n_b}$ where each row represents a point sampled from the auxiliary space. This operator samples K points from the auxiliary space in total. Finally, we use another network $\mathcal{N}_\gamma : \mathbb{R}^{K \times n_b} \to \mathbb{R}^{K \times n_z}$ parameterized by $\boldsymbol{\gamma}$ to further transform the K points to $\mathbf{A}_\mathbf{k}$ and $\mathbf{b}_k$:

$$\{\mathbf{A}_k, \mathbf{b}_k\}_{k=1}^{K} = \mathcal{N}_\gamma(\mathbf{WB}). \tag{10}$$

Based on the above preparations and incorporating with the sampling process defined in Eq. 5, our sampling process is (Alg. 1):

$$\begin{cases} \boldsymbol{\epsilon} \sim \mathcal{N}(0, 1), \\ \mathbf{B} = \mathcal{N}_\beta(\mathbf{x}), \\ \mathbf{W} \leftarrow \text{Gumbel-Softmax sampling}, \\ \{\mathbf{A}_k, \mathbf{b}_k\}_{k=1}^{K} = \mathcal{N}_\gamma(\mathbf{WB}), \\ \mathbf{z}_k = \mathbf{A}_k \boldsymbol{\epsilon} + \mathbf{b}_k, & 1 \le k \le K, \\ \mathbf{y}_\mathbf{k} = \mathcal{G}_\theta(\mathbf{x}, \mathbf{z}_k), & 1 \le k \le K. \end{cases} \tag{11}$$

Compared with DLow which relies on a network to directly output different Gaussian distributions, our method samples the Gaussian distributions from the auxiliary space characterized by $\mathbf{B}$. At the training stage, the sampling number K is fixed to train the structure of the auxiliary space and make it match with the sampling strategy (*e.g.*, Gumbel-Softmax random sampling) such that the points sampled from the space by the sampling strategy can yield predictions of high diversity. At test time, since the auxiliary space and the relationship between the space and the sampling strategy has already been established, we can sample any number of points as needed. We stress that although our model is trained on all the training data, these data are used to form the shape of the auxiliary space which is flexible and adjustable to accommodate all the data.

Algorithm 1 Diverse sampling from a complex distribution by randomly Gumbel-Softmax sampling from an auxiliary space

```
Input: Observed pose sequence $\mathbf{x}$, number of samples $K$, auxiliary space generation network $\mathcal{N}_{\boldsymbol{\beta}}$, Gaussian distribution generation network $\mathcal{N}_{\boldsymbol{\gamma}}$, CVAE decoder network $\mathcal{G}_{\boldsymbol{\theta}}$
Output: A set of samples $\{\tilde{\mathbf{y}}_k\}_{k=1}^{K}$
$\mathbf{B} = \mathcal{N}_{\boldsymbol{\beta}}(\mathbf{x})$ // generate an auxiliary space given input poses
$\mathbf{W} \longleftarrow$ Gumbel-Softmax sampling // see Algorithm 2
$\mathbf{P} = \mathbf{WB}$ // Multiply $\mathbf{W}$ and $\mathbf{B}$ to obtain a point matrix $\mathbf{P}$
$\{\mathrm{A}_k, \mathbf{b}_k\}_{k=1}^{K} = \mathcal{N}_{\boldsymbol{\gamma}}(\mathbf{P})$ // convert points into means and variances
$\boldsymbol{\epsilon} \sim \mathcal{N}(0,1)$ // sampling an $\boldsymbol{\epsilon}$ from the normal distribution
for $k = 1$ to $K$ do
    $\mathbf{z}_k = \mathrm{A}_k \boldsymbol{\epsilon} + \mathbf{b}_k$ // reparameterization trick
    $\tilde{\mathbf{y}}_k = \mathcal{G}_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z}_k)$ // decode $\mathbf{z}_k$ and $\mathbf{x}$ into a result $\tilde{\mathbf{y}}_k$
end for
```

In the following, we detail components of our model that help shape the auxiliary space.

3.2.1 Network Architectures. We have two sub-networks $N_{\boldsymbol{\beta}}$ and $N_{\boldsymbol{\gamma}}$. For $N_{\boldsymbol{\beta}}$, the input is $\mathbf{x} \in \mathbb{R}^{J \times C \times H}$ where H is the length of the input sequence, J is the number of joints of a pose, and each joint has C coordinates. The output is $\mathbf{B} \in \mathbb{R}^{M \times n_b}$. Firstly, we use a GCN [37] to extract features in $\mathbb{R}^{J \times F}$ from $\mathbf{x}$ where F is the dimension size of the features. Then, we use an MLP to map the feature map in $\mathbb{R}^{J \times F}$ to $\mathbf{B}$ in $\mathbb{R}^{M \times n_b}$. For $N_{\boldsymbol{\gamma}}$, we employ another MLP that maps a feature map in $\mathbb{R}^{K \times n_b}$ to a feature map in $\mathbb{R}^{K \times n_z}$. Please refer to the supplemental material for more details of the network designs.

3.2.2 Gumbel-Softmax Sampling. We randomly sample a coefficient matrix $\mathbf{W}$ by the Gumbel-Softmax sampling method (Alg. 2) by which each row $\mathbf{W}_i$ ($i \in [1, K]$) of $\mathbf{W}$ is calculated as:

$$\begin{cases} u_{ij} \sim \mathrm{U}(0,1), j \in [1, M], \\ g_{ij} = -\log(-\log(u_{ij})), j \in [1, M], \\ \mathbf{W}_{ij} = \dfrac{\pi + g_{ij}}{\tau}, j \in [1, M], \\ \mathbf{W}_i = \mathrm{Softmax}(\mathbf{W}_i) \end{cases} \tag{12}$$

where $\mathrm{U}(0,1)$ is the uniform distribution, π and τ are parameters of the Gumbel distribution which are set to $1/M$ and 1, respectively.

Besides the Gumbel distribution, we can also sample from a uniform or Gaussian distribution at first and then apply the Softmax normalization to obtain a coefficient matrix. However, Gumbel-Softmax sampling is more aggressive than Uniform-Softmax and Gaussian-Softmax sampling in assigning larger weights for a few basis vectors while making other basis vectors sharing just a small portion of the weight. In other words, the Gumbel-Softmax sampling strategy can samples points more near to the basis vectors. This benefits the learning of the basis vectors, because the network only needs to diversify the basis vectors to obtain diverse Gaussian distributions and eventually generate diverse poses. Please see our ablation study of comparisons among them.

3.2.3 Training Losses. Let $\{\tilde{\mathbf{y}}_k\}_{k=1}^{K}$ be the K results predicted from an input, we impose three kinds of loss functions on $\{\tilde{\mathbf{y}}_k\}_{k=1}^{K}$ to train the proposed sampling framework.

Algorithm 2 Gumbel-Softmax coefficient matrix generation

```
Input: Number of coefficient vectors $K$, dimension size $M$ of a coefficient vector, Gumbel distribution parameters $\pi$ and $\tau$
Output: A coefficient matrix $\mathbf{W} \in \mathbb{R}^{K \times M}$
Declare a matrix $\mathbf{W} \in \mathbb{R}^{K \times M}$
for $i = 1$ to $K$ do
    for $j = 1$ to $M$ do
        $u \sim \mathrm{U}(0,1)$ // sample a value from uniform distribution
        $g = -\log(-\log(u))$
        $\mathbf{W}_{ij} = \frac{\pi + g}{\tau}$
    end for
    $\mathbf{W}_i = \mathrm{Softmax}(\mathbf{W}_i)$ // normalize the $i^{th}$ row of $\mathbf{W}$
end for
```

(1) *Hinge-diversity loss.* In order to enhance the diversity of the results, we propose the following hinge-diversity loss:

$$\mathcal{L}_{hdiv} = \frac{1}{K(K-1)} \sum_{i=1}^{K} \sum_{j \neq i}^{K} \max\left(0, \eta - \left\|\tilde{\mathbf{y}}_i - \tilde{\mathbf{y}}_j\right\|_2\right), \tag{13}$$

where η is a user-defined threshold. By the hinge-diversity loss, we explicitly enforce the distance between any pair of generated predictions to be no less than η. Compared with the diversity loss defined in Eq. 6, the hinge-diversity loss is more aggressive in enforcing the diversity of the predictions while less affecting the accuracy of the results (see ablation studies).

(2) *Accuracy loss.* To ensure the accuracy of results, we also adopt the accuracy loss $\mathcal{L}_{acc}$ defined in Eq. 7 that enforces at least one of the predictions to be similar to the ground truth.

(3) *KL loss.* Finally the KL loss defined in Eq. 8 is a very important loss which ensures the model to produce realistic and plausible results instead of those with high diversity but are physically invalid. Our KL loss is now defined as:

$$\mathcal{L}'_{KL} = \mathcal{KL}\left(r_{\boldsymbol{\beta},\boldsymbol{\gamma}}(\mathbf{z}_k|\mathbf{x}) || p(\mathbf{z})\right), k \in [1, K], \tag{14}$$

where $r_{\boldsymbol{\beta},\boldsymbol{\gamma}}(\mathbf{z}_k|\mathbf{x})$ is the latent distribution of $\mathbf{z}_k$ encoded by networks $N_{\boldsymbol{\beta}}$ and $N_{\boldsymbol{\gamma}}$ with parameters of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$.

Altogether, our training loss is:

$$\mathcal{L} = \lambda_{hdiv}\mathcal{L}_{hdiv} + \lambda_{acc}\mathcal{L}_{acc} + \lambda_{KL}\mathcal{L}'_{KL}, \tag{15}$$

where λs are hyper-parameters used to balance the three terms.

4 EXPERIMENTS

4.1 Experimental Settings

Datasets. Following [36, 52], we evaluate our method on two public motion capture datasets: Human3.6M[1] [22] and HumanEva-I[2] [44]. (1) **Human3.6M** contains 7 subjects each performing 15 action categories. We use the data of five subjects (S1, S5, S6, S7, S8) for training, and the other two (S9, S11) for testing. After removing redundant joints, each pose has 17 joints. We input 25 frames, *i.e.*, 0.5s (50fps), to forecast 100 frames (2s) in the future. (2) **HumanEva-I** comprises 3 subjects each performing 5 action categories. Each

[1] The authors Lingwei Dang and Yongwei Nie signed the license and produced all the experimental results in this paper. Meta did not have access to the Human3.6M dataset.

[2] We strictly comply with the agreement of using this dataset for non-commercial research purpose only.

Table 1: Quantitative comparisons. All the results are calculated by sampling 50 times for each input historical pose sequence. The best results are marked in bold.

	Method	Human3.6M [22]					HumanEva-I [44]				
		APD ↑	ADE ↓	FDE ↓	MMADE ↓	MMFDE ↓	APD ↑	ADE ↓	FDE ↓	MMADE ↓	MMFDE ↓
deterministic	LTD [37]	0.000	0.516	0.756	0.627	0.795	0.000	0.415	0.555	0.509	0.613
	MSR [15]	0.000	0.508	0.742	0.621	0.791	0.000	0.371	0.493	0.472	0.548
stochastic	Pose-Knows [49]	6.723	0.461	0.560	0.522	0.569	2.308	0.269	0.296	0.384	0.375
	MT-VAE [50]	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
	BoM [7]	6.265	0.448	0.533	0.514	0.544	2.846	0.271	0.279	0.373	0.351
	GMVAE [16]	6.769	0.461	0.555	0.524	0.566	2.443	0.305	0.345	0.408	0.410
	DeLiGAN [21]	6.509	0.483	0.534	0.520	0.545	2.177	0.306	0.322	0.385	0.371
	DSF [51]	9.330	0.493	0.592	0.550	0.599	4.538	0.273	0.290	0.364	0.340
	DLow [52]	11.741	0.425	0.518	0.495	0.531	4.855	0.251	0.268	0.362	0.339
	GSPS [36]	14.757	0.389	0.496	0.476	0.525	5.825	0.233	0.244	0.343	0.331
	Ours	**15.310**	**0.370**	**0.485**	**0.475**	**0.516**	**6.109**	**0.220**	**0.234**	**0.342**	**0.316**

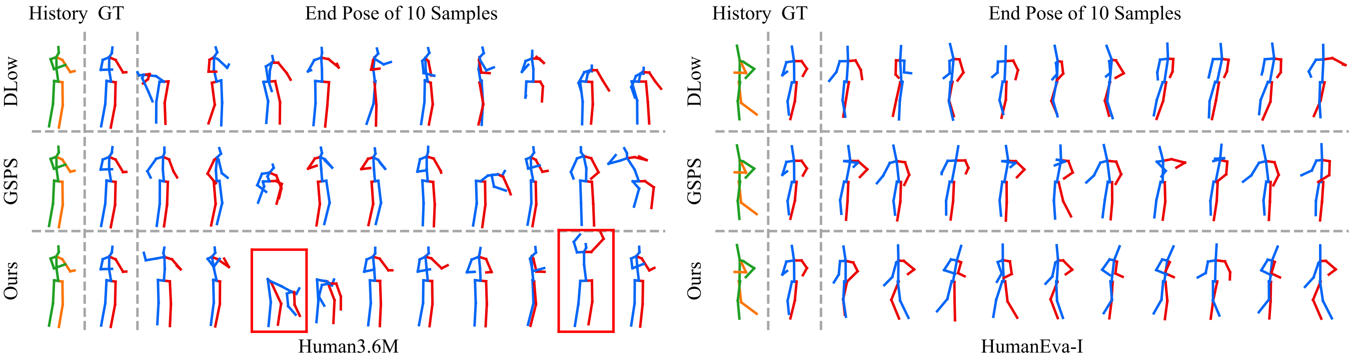

Figure 3: Qualitative comparisons. For the same input, we show end poses of 10 predicted results. Please see actions of the poses.

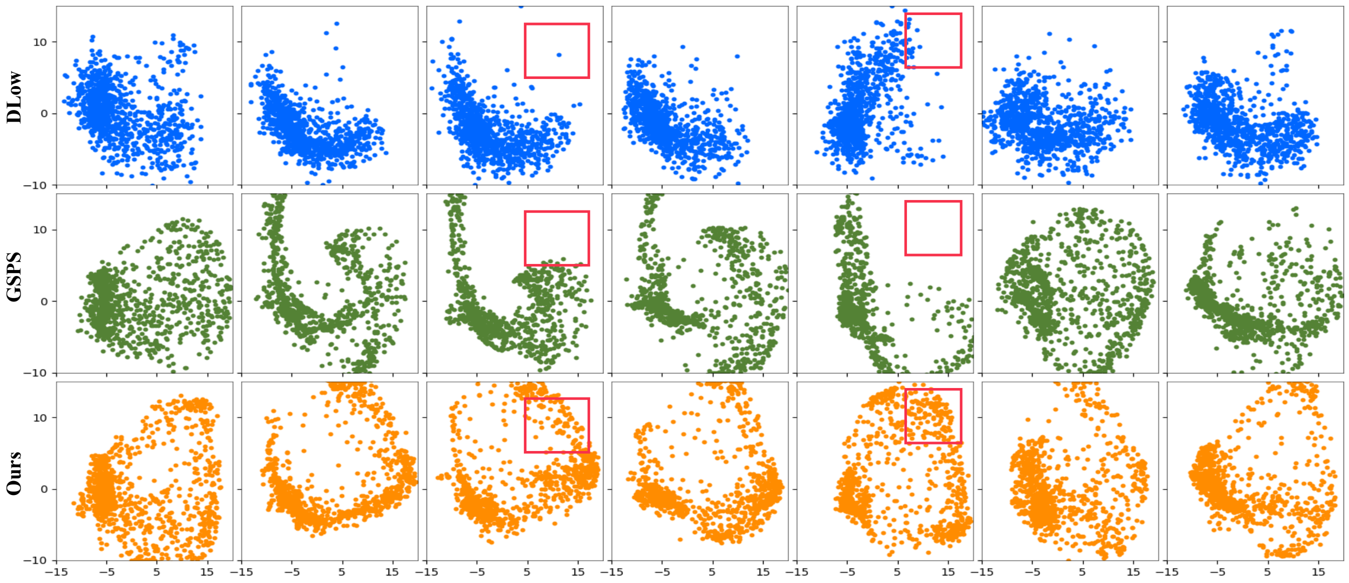

Figure 4: Holistic views of results. 1000 pose sequences are predicted and projected to 2D points. Note the regions marked with red boxes where our method can sample points from while DLow and GSPS fail to.

pose has 15 joints. We forecast 60 future poses (1s, 60fps) given 15 (0.25s) frames.

Evaluation Metrics. We use five metrics to evaluate our method. (1) **APD**: the Average Pairwise Distance of results predicted from an input [5]. This metric measures the diversity of the results. (2) **ADE** and **FDE**: ADE computes the Average Displacement Error between the ground truth and the result most similar to the ground truth. FDE, which stands for Final Displacement Error, only calculates the distance between the last pose of GT and the last pose of the most similar result to GT. (3) **MMADE** and **MMFDE** are the multi-modal versions of ADE and FDE which were introduced in [52]. To compute them, for each training sample $(\mathbf{x}, \mathbf{y})$, one needs to search the whole dataset for a set of $\{(\mathbf{x}_p, \mathbf{y}_p)\}_{p=1}^{P}$ whose past motion $\mathbf{x}_p$ is similar enough to $\mathbf{x}$, and take their future motion $\{\mathbf{y}_p\}_{p=1}^{P}$ as the pseudo ground truths of $\mathbf{x}$. MMADE is then computed as: $\frac{1}{P}\sum_{p=1}^{P}\min_i \left\|\tilde{\mathbf{y}}_i - \mathbf{y}_p\right\|_2, i \in \{1, \cdots, K\}$. Similar to FDE, MMFDE only calculates the error of end poses. By ADE, FDE, MMADE, and MMFDE, we can know the accuracy of results.

We employ additional metrics, *e.g.*, ADE-m, FDE-m, ACC, FID suggested by [8] for further evaluation. please refer to the supplementary material for more details.

Implementation Details. In default, we use $M = 40$ basis vectors. At training time, we sample $K = 50$ points from the auxiliary space. At test time, we set $K = 50$ to compare with previous approaches, and set K to numbers from 2 to 1000 in ablation studies. We set $n_b = 128$, and $n_z = 64$. For Human3.6M, we set $\lambda_{hdiv} = 20$, $\lambda_{acc} = 40$, and $\lambda_{KL} = 0.5$, and η in Eq. 13 to 25. These numbers for the HumanEva-I dataset are 100, 25, 0.1 and 20, respectively.

We implement our method in PyTorch, training it by the Adam optimizer with a learning rate of $1e-3$ for the first 100 training epochs. Then the learning rate starts to decrease, eventually becoming $7e-4$ after a total of 500 epochs of training. Following [36, 52], for each epoch, we randomly sample 5000 samples from Human3.6M or 2000 samples from HumanEva-I for training. The batchsize is set to 16 for both datasets.

4.2 Comparison with Previous Approaches

We compare our method with both kinds of prediction methods: (1) The most recent deterministic methods including LTD [37] and MSR [15]. (2) Stochastic methods including HP-GAN [6], Pose-Knows [49], MT-VAE [50], BoM [7], GMVAE [16], DeLiGAN [21], DSF [51], DLow [52], and GSPS [36]. For each input historical pose sequence, all stochastic methods predict 50 different future sequences.

Table 1 shows comparisons among all the compared methods. On both datasets, our method outperforms all the other approaches on all the evaluation metrics. Since deterministic approaches can only generate one output, the diversity of their results is 0.000. In terms of prediction accuracy, deterministic methods are inferior to stochastic methods too. This may be due to two reasons. Firstly, deterministic approaches are not good at long-term prediction (*e.g.*, more than 1 second). Secondly, stochastic methods can predict multiple results among which there may be a very good one. Since our method is based on DLow, let us focus on the comparisons between them. On human3.6M, DLow achieves a diversity of 11.741, while that of our method is 15.310, which is a very significant improvement of about 30%. Our method is also better than DLow in terms of prediction accuracy. For ADE, our accuracy is improved by 14.9% (0.370 *v.s.* 0.425). For FDE, MMADE, MMFDE, the improvements are: 6.8%, 4.2%, and 2.9%, respectively. The comparisons on the HumanEva-I dataset show similar trends: our method improves DLow on all metrics. GSPS is one of the latest stochastic prediction methods which directly predicts very diverse results as long as they are reasonable under many prior constraints. Our method outperforms GSPS, reaching a new state-of-the-art.

To show the quality of the predicted poses, we visualize end poses of pose sequences predicted by DLow [52], GSPS [36] and our method. The examples in Figure 3 show that our method produces more diverse results than DLow and GSPS. For example, on the left, our method can predict actions of "raising hands" and "picking up things" (marked by red boxes). Please refer to the supplemental video for how our method smoothly transitions the action from the input "normal standing" to the two very different actions.

Figure 4 illustrates the holistic views of results. Given an input, we generate 1000 results and project them into 2D space. Note that DLow can only sample 50 results at a time. To generate 1000 results for DLow, we repeatedly run DLow 20 times. As can be seen, our results occupy the 2D space more evenly. Please compare regions marked by red boxes where our method can sample points from while DLow and GSPS cannot.

4.3 Ablation Study

Table 2 shows five groups of ablation studies that validate the design components of our method. All models are trained on the HumanEva-I dataset for 200 epochs.

(1) **Number of basis vectors.** We set M to 20, 30, 40, 50, and 60. Overall, M has little effect on the results. For example, APD (*i.e.*, diversity) varies within a narrow range of [5.929, 5.993], with the best diversity obtained when $M = 40$. MMADE and MMFDE steadily improve as M increases. We finally choose 40 as the default value of M. (2) **Dimension size of auxiliary space.** We set n_b to 32, 64, 128, 256, and 512. The best APD, ADE, and FDE are obtained when $n_b = 128$. For MMADE and MMFDE, the best values are obtained when $n_b = 256$. We finally choose 128 as the default value of n_b. (3) **Gumbel *v.s.* Gaussian and Uniform.** The experimental results validate that Gumbel-Softmax sampling strategy is better than Uniform-Softmax and Gaussian-Softmax sampling methods when used in our sampling process. The fact that the differences between different sampling methods is not evident can be ascribed to the high capability of the proposed auxiliary-space-based resampling method. The space itself can be flexibly adjusted to match the three sampling methods to output good results. (4) **Using $\mathcal{N}_y$ or not.** In this ablation study, we remove the second MLP network $\mathcal{N}_y$, and directly use each row of the point matrix as a pair of $\mathbf{A}_k$ and $\mathbf{b}_k$. The diversity drops significantly (expected but undesired), while the accuracy increases. This is reasonable as diversity and accuracy are two conflict objectives: the increase of the diversity inevitably decreases the accuracy. We ultimately choose to integrate the network into our framework to achieve higher diversity with acceptable accuracy. (5) **Diversity loss *v.s.* hinge-diversity loss.** We replace our hinge-diversity loss with the diversity loss defined in Eq. 6. Note that the default weight of our hinge-diversity loss is $\lambda_{hdiv} = 25$. For the diversity loss, we first use a weight of

Table 2: We perform 5 groups of ablation studies. ∗ indicates default choices. Please refer to the main text for details.

	① Number of basis vectors					② Dimension of auxiliary space					③ Sampling method			④ N_γ		⑤ $\mathcal{L}_{hdiv}$ *v.s.* $\mathcal{L}_{div}$		
	20	30	40*	50	60	32	64	128*	256	512	Gumbel*	Gaussian	Uniform	w/N_γ*	w/o N_γ	$\mathcal{L}_{hdiv}$(25)*	$\mathcal{L}_{div}$ (25)	$\mathcal{L}_{div}$ (1300)
APD ↑	5.929	5.946	**5.993**	5.969	5.951	5.885	5.931	**5.993**	5.963	5.957	**5.993**	5.847	5.730	**5.993**	5.182	**5.993**	3.843	**5.993**
ADE ↓	0.234	0.234	0.231	**0.229**	0.233	0.236	0.233	**0.231**	0.233	0.236	**0.231**	0.240	0.233	0.231	**0.229**	0.231	**0.211**	0.235
FDE ↓	0.240	0.243	0.240	**0.239**	0.241	0.245	0.243	**0.240**	0.241	0.243	**0.240**	0.246	0.241	0.240	**0.236**	0.240	**0.218**	0.242
MMADE ↓	0.345	0.345	0.340	0.339	**0.338**	0.337	0.342	0.340	**0.336**	0.342	**0.340**	0.343	0.344	0.340	**0.322**	0.340	**0.309**	0.343
MMFDE ↓	0.321	0.319	0.313	0.315	**0.312**	0.312	0.318	0.313	**0.310**	0.316	**0.313**	0.320	0.323	0.313	**0.298**	0.313	**0.287**	0.321

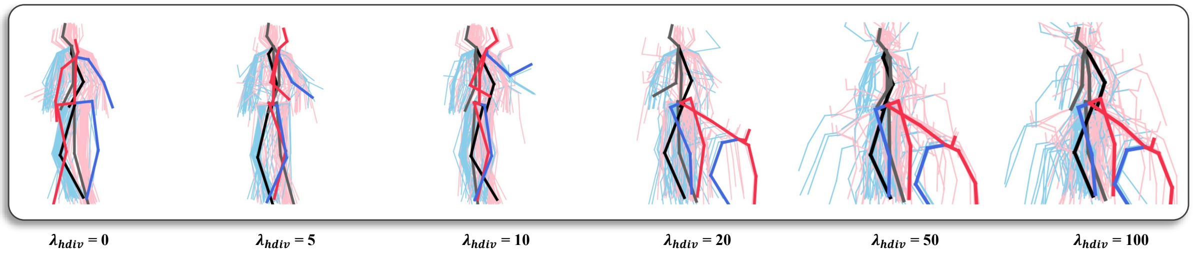

Figure 5: Increase λ_{hdiv} from 0 to 100. The last poses of 50 pose sequences predicted from an input are stacked together to illustrate the holistic view of results. Poses most similar to and different from the ground truth are highlighted.

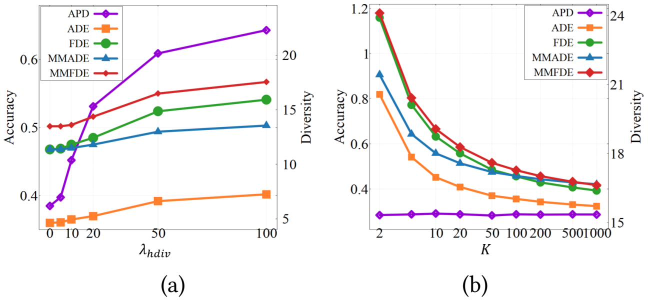

Figure 6: (a) As λ_{hdiv} increases, APD (diversity) drastically increases from 6.174 to 22.300, while accuracy decreases. (b) As K increases, accuracy becomes better while diversity nearly does not change.

25. However, the diversity is much lower: 3.843 *v.s.* our 5.993. We increase the weight of the diversity loss to 1300 until it produces the same diversity as ours, but now its accuracy is lower than that of our hinge-diversity loss. These studies show our hinge-diversity loss better prompts diversity while less affecting the accuracy.

In Figure 5, we perform an ablation study on the weight of the hinge-diversity loss. We set λ_{hdiv} to 0, 5, 10, 20, 50, and 100. All models are trained on Human3.6M for 200 epochs. We stack the end poses of all the results predicted from an input. The pose most similar to (in black and gray) and different from (red and blue) the ground truth are highlighted. Visually, as λ_{hdiv} increases, more diverse results are obtained. However, Figure 6 (a) shows that larger λ_{hdiv} leads to lower accuracy. We finally choose 20 as the default value of λ_{hdiv}, obtaining both satisfactory diversity and accuracy.

In Figure 6 (b), we increase K from 2 to 1000. As K increases, ADE, FDE, MMADE and MMFDE all decrease, meaning more accurate results are obtained. This indicates that we can obtain more accurate results by sampling more of them. The diversity nearly does not change as K increases, which is a good property as we can obtain diverse results with just a few samplings.

4.4 Limitations and Future Work

One limitation is that we have to adjust the weights of the loss functions to make a tradeoff between diversity and accuracy, though we note that DLow and GSPS suffer from this limitation too. Another limitation is that similar to DLow and GSPS our method occasionally generates odd poses with such as slightly long bones or unnatural actions. In the future, we can add more regularization terms, such as the bone and angle constraints adopted by GSPS, into our model to prevent these failures.

5 CONCLUSION

We have presented a diverse pose prediction algorithm. Our method first generates an auxiliary space from the input, then samples points from the auxiliary space by the Gumbel-Softmax sampling strategy, and finally maps the points to Gaussian distributions from which we sample latent codes and finally decode them into target predictions. We have demonstrated the influence of the dimension size of the auxiliary space and the number of basis vectors that characterize the auxiliary space on the performance of our method. We have also illustrated the effectiveness of the proposed hinge-diversity loss in promoting diversity while persisting accuracy. Although we only apply this method to tackle the task of stochastic human motion prediction, we believe it can also be used to handle many other stochastic prediction/generation problems.

ACKNOWLEDGMENTS

This work is sponsored by Prof. Yongwei Nie's and Prof. Guiqing Li's Natural Science Foundation of China projects (62072191, 61972160), and their Natural Science Foundation of Guangdong Province projects (2019A1515010860, 2021A1515012301).

REFERENCES

[1] Emre Aksan, Peng Cao, Manuel Kaufmann, and Otmar Hilliges. 2020. Attention, please: A spatio-temporal transformer for 3d human motion prediction. *arXiv preprint arXiv:2004.08692* 2, 3 (2020), 5.

[2] Emre Aksan, Manuel Kaufmann, Peng Cao, and Otmar Hilliges. 2021. A spatio-temporal transformer for 3D human motion prediction. In *2021 International Conference on 3D Vision (3DV)*. IEEE, 565–574.

[3] Emre Aksan, Manuel Kaufmann, and Otmar Hilliges. 2019. Structured prediction helps 3d human motion modelling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 7144–7153.

[4] Sadegh Aliakbarian, Fatemeh Saleh, Lars Petersson, Stephen Gould, and Mathieu Salzmann. 2021. Contextually plausible and diverse 3d human motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 11333–11342.

[5] Sadegh Aliakbarian, Fatemeh Sadat Saleh, Mathieu Salzmann, Lars Petersson, and Stephen Gould. 2020. A stochastic conditioning scheme for diverse human motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5223–5232.

[6] Emad Barsoum, John Kender, and Zicheng Liu. 2018. Hp-gan: Probabilistic 3d human motion prediction via gan. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 1418–1427.

[7] Apratim Bhattacharyya, Bernt Schiele, and Mario Fritz. 2018. 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*. 8485–8493.

[8] Xiaoyu Bie, Wen Guo, Simon Leglaive, Lauren Girin, Francesc Moreno-Noguer, and Xavier Alameda-Pineda. 2022. HiT-DVAE: Human Motion Generation via Hierarchical Transformer Dynamical VAE. *arXiv preprint arXiv:2204.01565* (2022).

[9] Yujun Cai, Lin Huang, Yiwei Wang, Tat-Jen Cham, Jianfei Cai, Junsong Yuan, Jun Liu, Xu Yang, Yiheng Zhu, Xiaohui Shen, et al. 2020. Learning progressive joint propagation for human motion prediction. In *European Conference on Computer Vision*. Springer, 226–242.

[10] Yujun Cai, Yiwei Wang, Yiheng Zhu, Tat-Jen Cham, Jianfei Cai, Junsong Yuan, Jun Liu, Chuanxia Zheng, Sijie Yan, Henghui Ding, et al. 2021. A unified 3d human motion synthesis model via conditional variational auto-encoder. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 11645–11655.

[11] Hsu-kuang Chiu, Ehsan Adeli, Borui Wang, De-An Huang, and Juan Carlos Niebles. 2019. Action-agnostic human pose forecasting. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 1423–1432.

[12] Enric Corona, Albert Pumarola, Guillem Alenya, and Francesc Moreno-Noguer. 2020. Context-aware human motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6992–7001.

[13] Qiongjie Cui and Huaijiang Sun. 2021. Towards accurate 3d human motion prediction from incomplete observations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4801–4810.

[14] Qiongjie Cui, Huaijiang Sun, and Fei Yang. 2020. Learning dynamic relationships for 3d human motion prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 6519–6527.

[15] Lingwei Dang, Yongwei Nie, Chengjiang Long, Qing Zhang, and Guiqing Li. 2021. MSR-GCN: Multi-Scale Residual Graph Convolution Networks for Human Motion Prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 11467–11476.

[16] Nat Dilokthanakul, Pedro AM Mediano, Marta Garnelo, Matthew CH Lee, Hugh Salimbeni, Kai Arulkumaran, and Murray Shanahan. 2016. Deep unsupervised clustering with gaussian mixture variational autoencoders. *arXiv preprint arXiv:1611.02648* (2016).

[17] Xinzhi Dong, Chengjiang Long, Wenju Xu, and Chunxia Xiao. 2021. Dual Graph Convolutional Networks with Transformer and Curriculum Learning for Image Captioning. In *Proceedings of the ACM International Conference on Multimedia*. 2615–2624.

[18] Jinghai Duan, Le Wang, Chengjiang Long, Sanping Zhou, Fang Zheng, Liushuai Shi, and Gang Hua. 2022. Complementary Attention Gated Network for Pedestrian Trajectory Prediction. In *AAAI Conference on Artificial Intelligence*.

[19] Katerina Fragkiadaki, Sergey Levine, Panna Felsen, and Jitendra Malik. 2015. Recurrent network models for human dynamics. In *Proceedings of the IEEE international conference on computer vision*. 4346–4354.

[20] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *Advances in neural information processing systems* 27 (2014).

[21] Swaminathan Gurumurthy, Ravi Kiran Sarvadevabhatla, and R Venkatesh Babu. 2017. Deligan: Generative adversarial networks for diverse and limited data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 166–174.

[22] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. 2013. 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 (2013), 1325–1339.

[23] Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).

[24] Jogendra Nath Kundu, Maharshi Gor, and R Venkatesh Babu. 2019. Bihmp-gan: Bidirectional 3d human motion prediction gan. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33. 8553–8560.

[25] Bin Li, Jian Tian, Zhongfei Zhang, Hailin Feng, and Xi Li. 2020. Multitask non-autoregressive model for human motion prediction. *IEEE Transactions on Image Processing* 30 (2020), 2562–2574.

[26] Maosen Li, Siheng Chen, Xu Chen, Ya Zhang, Yanfeng Wang, and Qi Tian. 2021. Symbiotic graph neural networks for 3d skeleton-based human action recognition and motion prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021).

[27] Maosen Li, Siheng Chen, Yangheng Zhao, Ya Zhang, Yanfeng Wang, and Qi Tian. 2020. Dynamic multiscale graph neural networks for 3d skeleton based human motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 214–223.

[28] Zimo Li, Yi Zhou, Shuangjiu Xiao, Chong He, Zeng Huang, and Hao Li. 2017. Auto-conditioned recurrent networks for extended complex human motion synthesis. *arXiv preprint arXiv:1707.05363* (2017).

[29] Zhenguang Liu, Kedi Lyu, Shuang Wu, Haipeng Chen, Yanbin Hao, and Shouling Ji. 2021. Aggregated multi-gans for controlled 3d human motion prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 2225–2232.

[30] Zhian Liu, Yongwei Nie, Chengjiang Long, Qing Zhang, and Guiqing Li. 2021. A Hybrid Video Anomaly Detection Framework via Memory-Augmented Flow Reconstruction and Flow-Guided Frame Prediction. In *Proceedings of the IEEE International Conference on Computer Vision*. 13588–13597.

[31] Zhenguang Liu, Pengxiang Su, Shuang Wu, Xuanjing Shen, Haipeng Chen, Yanbin Hao, and Meng Wang. 2021. Motion prediction using trajectory cues. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 13299–13308.

[32] Zhenguang Liu, Shuang Wu, Shuyuan Jin, Qi Liu, Shouling Ji, Shijian Lu, and Li Cheng. 2022. Investigating Pose Representations and Motion Contexts Modeling for 3D Motion Prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2022).

[33] Kedi Lyu, Zhenguang Liu, Shuang Wu, Haipeng Chen, Xuhong Zhang, and Yuyu Yin. 2021. Learning Human Motion Prediction via Stochastic Differential Equations. In *Proceedings of the 29th ACM International Conference on Multimedia*. 4976–4984.

[34] Tiezheng Ma, Yongwei Nie, Chengjiang Long, Qing Zhang, and Guiqing Li. 2022. Progressively Generating Better Initial Guesses Towards Next Stages for High-Quality Human Motion Prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 6437–6446.

[35] Wei Mao, Miaomiao Liu, and Mathieu Salzmann. 2020. History repeats itself: Human motion prediction via motion attention. In *European Conference on Computer Vision*. Springer, 474–489.

[36] Wei Mao, Miaomiao Liu, and Mathieu Salzmann. 2021. Generating Smooth Pose Sequences for Diverse Human Motion Prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 13309–13318.

[37] Wei Mao, Miaomiao Liu, Mathieu Salzmann, and Hongdong Li. 2019. Learning trajectory dependencies for human motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 9489–9497.

[38] Wei Mao, Miaomiao Liu, Mathieu Salzmann, and Hongdong Li. 2021. Multi-level motion attention for human motion prediction. *International Journal of Computer Vision* 129, 9 (2021), 2513–2535.

[39] Julieta Martinez, Michael J Black, and Javier Romero. 2017. On human motion prediction using recurrent neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2891–2900.

[40] Angel Martínez-González, Michael Villamizar, and Jean-Marc Odobez. 2021. Pose Transformers (POTR): Human Motion Prediction with Non-Autoregressive Transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2276–2284.

[41] Hai-Feng Sang, Zi-Zhen Chen, and Da-Kuo He. 2020. Human motion prediction based on attention mechanism. *Multimedia Tools and Applications* 79, 9 (2020), 5529–5544.

[42] Liushuai Shi, Le Wang, Chengjiang Long, Sanping Zhou, Fang Zheng, Nanning Zheng, and Gang Hua. 2022. Social Interpretable Tree for Pedestrian Trajectory Prediction. In *AAAI Conference on Artificial Intelligence*.

[43] Liushuai Shi, Le Wang, Chengjiang Long, Sanping Zhou, Mo Zhou, Zhenxing Niu, and Gang Hua. 2021. SGCN: Sparse Graph Convolution for Pedestrian Trajectory Prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 8994–9003.

[44] Leonid Sigal, Alexandru O Balan, and Michael J Black. 2010. Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *International journal of computer vision* 87, 1 (2010), 4–27.

[45] Sijie Song, Cuiling Lan, Junliang Xing, Wenjun Zeng, and Jiaying Liu. 2017. An end-to-end spatio-temporal attention model for human action recognition from skeleton data. In *Proceedings of the AAAI conference on artificial intelligence*,

Vol. 31.

[46] Pengxiang Su, Zhenguang Liu, Shuang Wu, Lei Zhu, Yifang Yin, and Xuanjing Shen. 2021. Motion Prediction via Joint Dependency Modeling in Phase Space. In *Proceedings of the 29th ACM International Conference on Multimedia*. 713–721.

[47] Julian Tanke, Chintan Zaveri, and Juergen Gall. 2021. Intention-based Long-Term Human Motion Anticipation. In *2021 International Conference on 3D Vision (3DV)*. IEEE, 596–605.

[48] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[49] Jacob Walker, Kenneth Marino, Abhinav Gupta, and Martial Hebert. 2017. The pose knows: Video forecasting by generating pose futures. In *Proceedings of the IEEE international conference on computer vision*. 3332–3341.

[50] Xinchen Yan, Akash Rastogi, Ruben Villegas, Kalyan Sunkavalli, Eli Shechtman, Sunil Hadap, Ersin Yumer, and Honglak Lee. 2018. Mt-vae: Learning motion transformations to generate multimodal human dynamics. In *Proceedings of the European conference on computer vision (ECCV)*. 265–281.

[51] Ye Yuan and Kris Kitani. 2019. Diverse trajectory forecasting with determinantal point processes. *arXiv preprint arXiv:1907.04967* (2019).

[52] Ye Yuan and Kris Kitani. 2020. Dlow: Diversifying latent flows for diverse human motion prediction. In *European Conference on Computer Vision*. Springer, 346–364.