Multi-Objective Diverse Human Motion Prediction with Knowledge Distillation

Hengbo Ma[1,2†] Jiachen Li[3] Ramtin Hosseini[1,4†] Masayoshi Tomizuka[2] Chiho Choi[1]
[1]Honda Research Institute, USA [2]University of California, Berkeley
[3]Stanford University [4]Tufts University
{hengbo_ma, tomizuka}@berkeley.edu jiachen_li@stanford.edu ramtin.hosseini@tufts.edu
cchoi@honda-ri.com

Abstract

Obtaining accurate and diverse human motion prediction is essential to many industrial applications, especially robotics and autonomous driving. Recent research has explored several techniques to enhance diversity and maintain the accuracy of human motion prediction at the same time. However, most of them need to define a combined loss, such as the weighted sum of accuracy loss and diversity loss, and then decide their weights as hyperparameters before training. In this work, we aim to design a prediction framework that can balance the accuracy sampling and diversity sampling during the testing phase. In order to achieve this target, we propose a multi-objective conditional variational inference prediction model. We also propose a short-term oracle to encourage the prediction framework to explore more diverse future motions. We evaluate the performance of our proposed approach on two standard human motion datasets. The experiment results show that our approach is effective and on a par with state-of-the-art performance in terms of accuracy and diversity.

1. Introduction

Human motion prediction plays a significant role in several applications such as human-robot interaction [3, 4], autonomous driving [13, 14, 33, 39], and animation [45]. For instance, an autonomous driving system can make a safe planning strategy given an accurate motion prediction of pedestrians. Moreover, robots can cooperate reasonably with people when they have a good understanding of human beings' future plans. However, since diversity and uncertainty are human future motion's intrinsic properties, it becomes a challenging problem in the computer science community. Unlike the vehicle trajectory prediction scenarios where we can get prior knowledge such as the traffic rules and routing information [40, 41] to constrain the different modes of trajectories, we can hardly get any prior knowledge about what humans will do in the future. Thus, we can only leverage the information from the given dataset, which increases the difficulty of diverse human motion prediction.

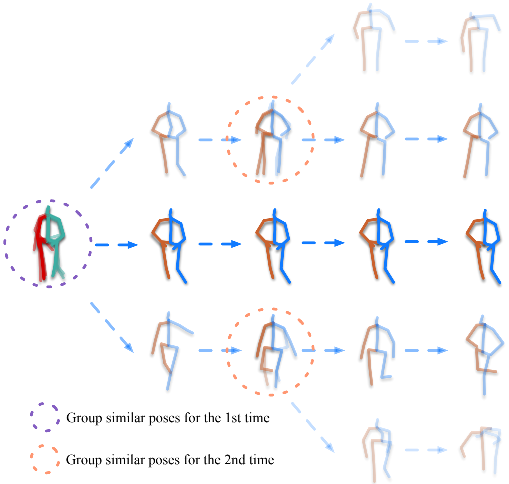

Group similar poses for the 1st time
Group similar poses for the 2nd time

Figure 1. Illustration of obtaining multi-modal pseudo future motions in a dataset. We can cluster the similar initial poses (purple dashed circle) and share their future poses as the common ground truth. The solid poses are the ground truth and the transparent ones are the augmented poses. We argue that such an approach can be applied recursively (orange dashed circle), which will lead to discovering more different and realistic modes of motion in the data.

There are two lines of research in this area. First, several works attempt to get an accurate human motion prediction without considering the diversity, such as [36] based on graph neural network and [56] based on recurrent neural network. On the other line, some research investigates how to increase the diversity of human motion prediction based on deep generative models [2, 47, 60, 63] or diverse sampling techniques [61]. Deep generative models such as variational autoencoder and generative adversarial network naturally capture the stochastic behaviors, while they may suffer from mode collapse problems. Otherwise, even if we assume that the generative models can capture the actual

[†] Work done during Hengbo and Ramtin's internship at Honda Research Institute.

data distribution, the data distribution can still be very imbalanced and skewed, which makes that sampling the minor modes is challenging within a limited number of samples. Several works [42, 61, 62] propose new losses to increase diversity while keeping the prediction natural and accurate. In [62], a multiple sampling function is designed to explicitly capture the different modes of the distribution based on a pre-trained conditional variational autoencoder. By using this pre-trained variational autoencoder, such methods can control the likelihood of predicted motion with a training hyperparameter. In [5, 32, 42], they proposed generative models to learn the distribution implicitly. However, these works still have to choose hyperparameters before training to balance the likelihood and diversity sampling. It implies that such approaches cannot be adjusted and controlled during the testing phase. Considering the real-world application such as pedestrian motion prediction in autonomous driving, we not only need to know most of the different possible modes of motion but also need to know which modes will most likely happen. It will be more practical if we can decide the balance of accuracy sampling and diversity sampling during the testing phase for the purpose of designing the risk-averse or risk-seeking planner of autonomous vehicles. Hence, we introduce a multi-objective variational inference framework with two different priors. The proposed structure makes it possible to adjust the ratio between accuracy and diversity sampling during the testing time.

Meanwhile, since there is only one ground-truth future motion poses given a historical observation, several works [49, 59] propose to use a similarity cluster-based technique to get the multi-modal pseudo-ground-truth future motions. Similar initial poses are grouped, and their corresponding future poses can be viewed as the pseudo possible future motions for each initial pose in the group. We argue that such logic can also be applied recursively. We can group similar poses again at certain steps and get the shared futures. A demonstration is shown in Figure 1. This strategy can boost the diversity of future motions. However, the sampling number will exponentially increase due to the recursive queries during training and make such direct implementation intractable. In order to solve this issue, we introduce an oracle that provides several possible future motions with a short-term horizon to instruct the predictor repeatedly. To summarize, our contributions are three folds:

- We propose a unified multi-objective conditional variational autoencoder based human motion prediction framework, which can adjust the ratio of sample numbers of accuracy and diversity sampling during testing.
- We propose to learn a short-term oracle system and distill the oracle's knowledge into the prediction framework to increase the diversity of human future motions. In order to achieve this goal, we propose a novel sample-based loss to supervise the predictor during the training phase.
- We evaluate the performance of our proposed approach on two human motion datasets. The experiments results show that our methods can achieve state-of-the-art performance.

2. Related Work

Human motion prediction. Human motion prediction has been investigated with many different approaches in the computer vision community. At the early stage, several methods [1,10,37,43,50,54,57] without deep learning techniques are proposed such as Gaussian process [58], hidden Markov model [10], and latent variable models [54]. Such methods can achieve good performance for recurrent human motion data. However, they may not be suitable for more complicated irregular human motions. As several promising deep learning models such as recurrent neural network (RNN) [6, 12, 18] and graph neural network (GNN) [11, 30, 34, 35, 48] are proposed recently, there are several research focusing on how to incorporate the models above to enhance the deterministic human motion prediction accuracy. Several works such as [21, 27, 44, 46, 64] are based on RNN, and [36, 42] utilize graph neural network (GNN) to capture both the temporal and spatial information. In order to get more diverse human motion prediction, several probabilistic models [2, 5, 7, 28, 29, 32, 38, 47, 63] are applied to capture the uncertainty of human motion. Deep generative models can be used to estimate the data distribution. There are several approaches based on variational autoencoders [2,7,28,29], generative adversarial networks [5, 16, 32, 38] and normalizing flows [20, 47, 63].

Diverse forecasting. In [60], the authors propose an approach that can learn a representation for motion reconstruction and transformation together. Also, GAN-like models are utilized in [5, 32] to capture the diverse human motion prediction. There are also some research using a different representation to improve the diversity [64]. In [61], a diversity sampling function which is formulated as a determinantal point process [22, 23, 31] is proposed. Especially in [62], the authors argue that even though the existed likelihood-based methods can have a good estimation of the data, they can still be challenging to sample some minor modes given a fixed number of samples. Hence, they propose to learn another diversity sampling function that can generate diverse motions based on one pre-trained variational autoencoder model. However, the proposed model needs to choose hyperparameters to balance the likelihood and diversity before training. We investigate the diverse human motion prediction in an orthogonal direction with the related work. We aim to get a unified model that can adjust the sample number ratio between accuracy and diversity samples during the testing phase. Besides, we attempt to explore more diverse and natural modes by utilizing pseudo future motions with a short-term oracle, and any models mentioned above can be integrated.

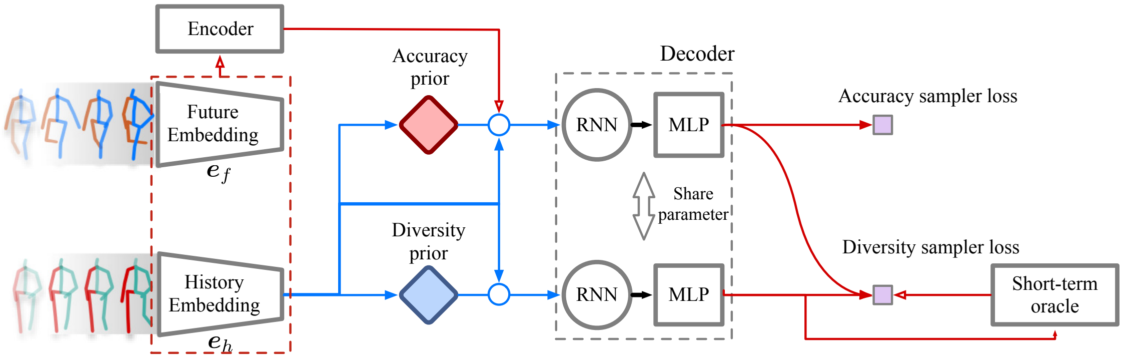


Figure 2. Overview of the proposed framework. Red lines indicate the pipeline, which are only used during training. Blue lines indicate the pipeline used in both the training and testing phase. During training, several samples are generated from both accuracy prior function (red diamond, Section 4.1.1) and diversity prior function (blue diamond, Section 4.1.2). The accuracy prior function will be only updated by accuracy sampler loss defined in Section 4.1.1. The diversity prior function will be updated by the diversity sampler loss, which depends on all the samples. The short-term oracle function is introduced in Section 4.2.

3. Problem Formulation

Our goal is to predict the possible future human motions given a dataset $\mathcal{D}$. We denote the human motion with time horizon $T = T_h + T_f$ as $\boldsymbol{X}_{t-T_h+1:t+T_f} = [\boldsymbol{X}_{t-T_h+1}, \dots, \boldsymbol{X}_{t+T_f}]$, where $\boldsymbol{X}_t \in \mathbb{R}^d$ is the human joints Cartesian coordinates at time step t. T_h and T_f are the historical horizon and future horizon respectively. Given an observation $\boldsymbol{C} = \boldsymbol{X}_{t-T_h+1:t}$, we intend to get the future motion distribution $P(\boldsymbol{X}_{t+1:t+T_f}|\boldsymbol{C}, \rho)$. Since such conditional probabilistic distribution may have several dominant modes, it is difficult to sample the other modes given a fixed sampling number. In contrast, if we focus on increasing the diversity of the samples, the prediction accuracy will be undermined. In this work, we introduce a variable $\rho \in [0, 1]$ to control the degree of diversity of prediction, i.e., we intend to get M samples $X^i_{t+1:t+T_f} \sim P(\boldsymbol{X}_{t+1:t+T_f}|\boldsymbol{C}, \rho), i = 1, \dots, M$. The larger ρ is, the more diverse samples will be generated and focuses on the rare cases, and the smaller ρ is, the prediction will focus more on the most likely modes. For simplicity, we use $\boldsymbol{X}$ represent $\boldsymbol{X}_{t+1:t+T_f}$ in the case that the time step index is not necessary.

4. Methodology

We first introduce the multi-objective generative prediction framework based on conditional variational inference. Then we introduce the proposed short-term oracle, which provides multi-modal supervision to the prediction framework. Finally, we introduce our proposed approach's training strategy and testing procedure. The overall framework is illustrated in Figure 2.

4.1. Multi-Objective Predictor

In general, we can represent a probabilistic distribution via a latent variable model:

$$P(\boldsymbol{X}|\boldsymbol{C}; Q) = \mathbb{E}_{\boldsymbol{Z}\sim Q(\boldsymbol{Z}|\boldsymbol{C})}[P(\boldsymbol{X}|\boldsymbol{C}, \boldsymbol{Z})], \tag{1}$$

where $Q(\boldsymbol{Z}|\boldsymbol{C})$ is the conditional prior distribution of latent variable $\boldsymbol{Z} \in \mathbb{R}^{d_z}$ whose dimension is d_z. $P(\boldsymbol{X}|\boldsymbol{C}, \boldsymbol{Z})$ is defined as the conditional likelihood given the observation information $\boldsymbol{C}$ and latent variable $\boldsymbol{Z}$. We can vary the prior distribution Q to achieve different distributions of $\boldsymbol{X}$ given the same observation $\boldsymbol{C}$. In our proposed approach, we introduce two different prior distributions $Q_{\text{acc}}(\boldsymbol{Z}|\boldsymbol{C})$ and $Q_{\text{div}}(\boldsymbol{Z}|\boldsymbol{C})$. We intend to estimate the data distribution $P_{\mathcal{D}}$ using $P(\boldsymbol{X}|\boldsymbol{C}; Q)$ with prior $Q_{\text{acc}}(\boldsymbol{Z}|\boldsymbol{C})$, and get the most diverse distribution which mainly focuses on the minor modes by sampling from $Q_{\text{div}}(\boldsymbol{Z}|\boldsymbol{C})$. The overall framework is illustrated in Figure 2. Similar to [62], we define the historical observation encoder $\mathrm{e}_{\boldsymbol{h}}(\boldsymbol{C})$ and future information encoder $\mathrm{e}_{\boldsymbol{f}}(\boldsymbol{X})$ as

$$\begin{aligned} \mathrm{e}_{\boldsymbol{h}}(\boldsymbol{C}) &= [\text{MLP} \circ \text{RNN}](\boldsymbol{C}) \\ \mathrm{e}_{\boldsymbol{f}}(\boldsymbol{X}) &= [\text{MLP} \circ \text{RNN}](\boldsymbol{X}), \end{aligned} \tag{2}$$

where we first encode the temporal information of trajectories by using a recurrent neural network (RNN) and then use a forward neural network to map the states of RNN to the feature embedding space. Based on the historical embedding $\mathrm{e}_{\boldsymbol{h}}(\boldsymbol{C})$ and latent variable $\boldsymbol{Z}$, We denote the decoder function $\mathrm{d}_\theta(\boldsymbol{X}|\boldsymbol{C}, \boldsymbol{Z})$ as:

$$\mathrm{d}_\theta(\boldsymbol{X}|\boldsymbol{C}, \boldsymbol{Z}) = [\text{MLP} \circ \text{RNN}](\mathrm{e}_{\boldsymbol{h}}(\boldsymbol{C})||\boldsymbol{Z}), \tag{3}$$

where θ is the parameter of the decoder. In general, the outputs of the decoder are the parameters of a probabilistic distribution, e.g., the mean and variance of a Gaussian distribution. In this work, we use a deterministic decoder, and the output of the decoder is the predicted poses. For convenience, the output of the decoder is also denoted by Equation 3. The randomness of the decoder is only dependent on $\boldsymbol{Z}$. "$||$" represents the concatenate operator of two vectors. We use a similar neural network structure for the decoder with the encoders. The details of the operator $\circ$ are in the supplementary material.

4.1.1 Accuracy Sampler

The first objective is to infer the accuracy prior distribution $Q_{\text{acc}}(\boldsymbol{Z}|\boldsymbol{C})$. We intend to approximate the data distribution by sampling from the accuracy prior distribution. Hence, we apply the variational inference to maximize the evidence lower bound (ELBO) of the log-likelihood:

$$\begin{aligned}\mathcal{L}_{\text{ELBO}} &= \mathbb{E}_{Q_\psi(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{C})}[\log P_\theta(\boldsymbol{X}|\boldsymbol{Z},\boldsymbol{C})] \\ &- D_{KL}[Q_\psi(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{C})||Q_{\text{acc}}(\boldsymbol{Z}|\boldsymbol{C})],\end{aligned} \tag{4}$$

where $Q_\psi(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{C})$ is the posterior distribution of latent variable $\boldsymbol{Z}$ given the historical observation and future information. There are some works [8, 52, 55, 65] investigating the collapse problems for conditional variational inference. Those works argue that using a universal prior distribution, i.e., an independent isotropic Gaussian distribution, may not be a good choice for conditional distribution estimation [8, 52]. It is difficult to capture complex conditional multi-modal data and introduce strong model bias resulting in missing modes [53, 55, 65]. Hence, instead of using an isotropic Gaussian distribution $\mathcal{N}(0, \mathbf{I})$ which is independent of $\boldsymbol{C}$, we model $Q_{\text{acc}}(\boldsymbol{Z}|\boldsymbol{C})$ as a Gaussian distribution $\mathcal{N}(\mu_{\phi_{\text{acc}}}(\boldsymbol{C}), \Sigma_{\phi_{\text{acc}}}(\boldsymbol{C}))$. The $D_{KL}[Q_\psi||Q_{\text{acc}}]$ is:

$$\frac{1}{2}[\log\frac{|\Sigma_{\phi_{\text{acc}}}|}{|\Sigma_\psi|} - n_z + Tr(\Sigma^{-1}_{\phi_{\text{acc}}}\Sigma_\psi) + ||\mu_{\phi_{\text{acc}}} - \mu_\psi||^2_{\Sigma^{-1}_{\phi_{\text{acc}}}}], \tag{5}$$

which can be calculated analytically. Since we have no control of the distribution $Q_{\text{acc}}(\boldsymbol{Z}|\boldsymbol{C})$, it could be arbitrarily distribution and it will increase the difficulty of training. In order to constrain the prior distribution, we use the best-of-many loss as the regularization of the prior model:

$$\begin{aligned}\mathcal{R}_{\text{acc}} &= \min_i \|\hat{\boldsymbol{X}}^i - \boldsymbol{X}\|^2 \\ &\boldsymbol{z}^i \sim Q(\boldsymbol{Z}|\boldsymbol{C}) \\ &\hat{\boldsymbol{X}}^i = \text{d}_\theta(\boldsymbol{X}|\boldsymbol{C}, \boldsymbol{z}^i), i = 1, \ldots, n_{\text{acc}},\end{aligned} \tag{6}$$

where n_{acc} is the number of samples. Then the overall loss for the accuracy sampler is:

$$\mathcal{L}_{\text{A}}(\theta, \psi) = -\lambda_{\text{ELBO}}\mathcal{L}_{\text{ELBO}} + \lambda_{\text{acc}}\mathcal{R}_{\text{acc}}, \tag{7}$$

where λ_{elbo} and λ_{acc} are used to balance two losses.

4.1.2 Diversity Sampler

In order to explore the different modes of possible future poses, we propose to learn another prior distribution $Q_{\text{div}}(\boldsymbol{Z}|\boldsymbol{C})$ with parameter ϕ_{div}. We utilize a common diversity loss definition:

$$\begin{aligned}&\text{DIV}(\mathcal{X}, \mathcal{Y}) = \frac{1}{N_x N_y}\sum_{i,j} e^{-d(\boldsymbol{X}^i, \boldsymbol{Y}^j)} \\ &\boldsymbol{X}^i, \boldsymbol{Y}^j \in \mathcal{X}, \mathcal{Y}, i = 1, \ldots, N_x, j = 1, \ldots, N_y,\end{aligned} \tag{8}$$

where $\mathcal{X}$ and $\mathcal{Y}$ represent two sets of samples with size N_x and N_y. $d(\cdot, \cdot)$ is a metric defined in the Euclidean space. We define the metric as $d(x, y) = \eta||x - y||_2$, where η is a parameter to determine the sensitivity of the distance between two samples. We denote the set of the samples which are generated by the accuracy sampler as $\mathcal{X}_{\text{acc}}$ and the set of samples generated by the diversity sampler as $\mathcal{X}_{\text{div}}$. Then we define the diversity loss as:

$$\mathcal{L}_{\text{div}} = \alpha_{\text{div}}\text{DIV}(\mathcal{X}_{\text{div}}, \mathcal{X}_{\text{div}}) + (1 - \alpha_{\text{div}})\text{DIV}(\mathcal{X}_{\text{div}}, \mathcal{X}_{\text{acc}}), \tag{9}$$

where $\text{DIV}(\mathcal{X}_{\text{div}}, \mathcal{X}_{\text{div}})$ represents the diversity of samples generated by the diversity sampler. $\text{DIV}(\mathcal{X}_{\text{div}}, \mathcal{X}_{\text{acc}})$ represents the average pairwise distance between the samples from accuracy and diversity sampler. In the previous works, when the weight of diversity loss is large, it will have a negative influence on the accuracy sampler to approximate the data distribution. Since we intend to disentangle the accuracy objective and diversity objective, we only increase the pairwise distances between samples from the diversity sampler by using the first term in Equation 9 , and we make the samples from the diversity sampler dissimilar to the samples from the accuracy sampler by using the second term in Equation 9. We can determine the relative importance of the two items in 9 by a weight α_{div}. A larger α_{div} means that we focus on making the samples from Q_{div} more different.

Only using the diversity loss is not enough to get a realistic prediction since it is possible to increase diversity in the wrong way. For instance, one model can generate random noises or arbitrary invalid poses. Hence, we need to use human motion in the data to constrain the prediction. In order to constrain each generated poses from the diversity sampler, we assume that there exists an oracle:

$$\tilde{\boldsymbol{X}}_{t+1:t+\tau} \sim \mathcal{O}(\boldsymbol{X}_t, \tau), \tag{10}$$

where $\mathcal{O}(\boldsymbol{X}_t, \tau)$ is the probabilistic distribution of future poses with horizon τ given the current initial pose $\boldsymbol{X}_t$. The oracle $\mathcal{O}$ can be seen as a teacher to distill the "knowledge" of the future poses into the predictor. Based on the oracle, we define a sample-based loss:

$$\begin{aligned}&\mathcal{L}_{\text{ref}}(\tau) = \frac{1}{n_{\text{div}}}\sum_{i,s}\min_j \|\hat{\boldsymbol{X}}^i_{s\tau+1:(s+1)\tau} - \tilde{\boldsymbol{X}}^j_{s\tau+1:(s+1)\tau)}\|^2, \\ &\quad s.t. \boldsymbol{z}^i \sim Q_{\text{div}}(\boldsymbol{Z}|\boldsymbol{C}), \hat{\boldsymbol{X}}^i_{1:T} = \text{d}_\theta(\boldsymbol{X}|\boldsymbol{C}, \boldsymbol{z}^i), \\ &\quad \tilde{\boldsymbol{X}}^j_{s\tau+1:(s+1)\tau} \sim \mathcal{O}(\hat{\boldsymbol{X}}_{s\tau}, \tau), \\ &\quad i = 1, \ldots, n_{\text{div}}, j = 1, \ldots, n_{\text{o}}, s = 0, \ldots, T/\tau - 1,\end{aligned} \tag{11}$$

where τ represents the time interval of predicted poses from the oracle. n_{div} is the number of samples generated from diversity prior, and n_{o} is the number of samples which the oracle provides. W.l.o.g, we assume that the current time step is 0, and the prediction horizon is T. Given one sample $\hat{\boldsymbol{X}}^i_{1:T}$, the oracle provides several possible short-term

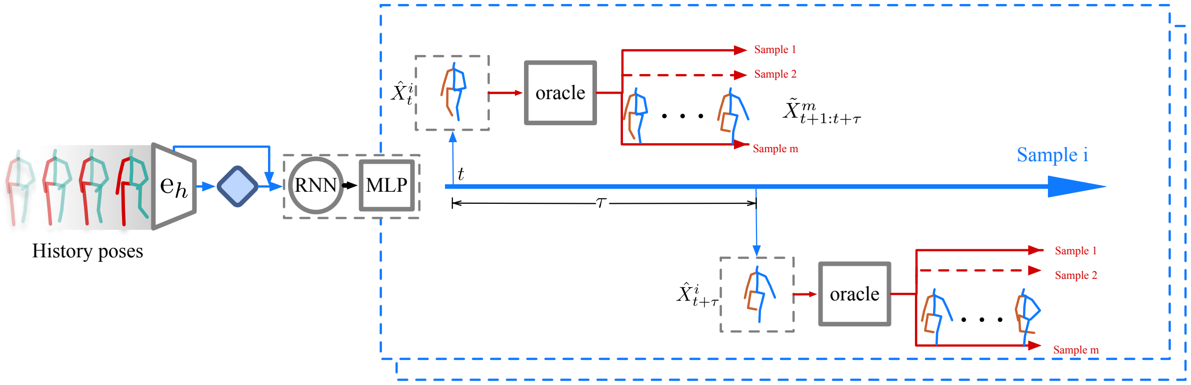


Figure 3. The procedure of short-term oracle supervision. During training, we can get several predicted human motions. For each sample (indicated by the blue arrow), the poses will be fed to the oracle after each τ time steps. The oracle will provide several possible future poses as options. The predicted human motions only need to be similar to one of the options in each short time horizon.

futures $\tilde{\boldsymbol{X}}^j_{s\tau+1:(s+1)\tau}$ given the current predicted pose $\hat{\boldsymbol{X}}^i_{s\tau}$ recursively. We enforce the short-term predicted sequence $\hat{\boldsymbol{X}}^i_{s\tau+1:(s+1)\tau}$ to be similar with one of the provided futures $\tilde{\boldsymbol{X}}^j_{s\tau+1:(s+1)\tau}$. Notice that the diversity loss $\mathcal{L}_{\text{div}}$ defined in Equation 9 will encourage the predictor to choose one of the provided future human motions which is useful to increase the diversity. The illustration of the oracle supervision procedure is shown in Figure 3.

We also adopt several widely-used physical feasibility losses [42, 62, 63] such as the limbs' constraint $\mathcal{L}_{\text{limb}}$ and the velocity constraint $\mathcal{L}_{\text{vel}}$ as $\mathcal{L}_{\text{phy}}$:

$$\mathcal{L}_{\text{phy}} = \lambda_{\text{vel}}\mathcal{L}_{\text{vel}} + \mathcal{L}_{\text{limb}}. \tag{12}$$

The details of each item in Equation 12 are provided in the supplementary material. Therefore, the overall loss for the diversity sampler is:

$$\mathcal{L}_{\text{D}} = \lambda_{\text{ref}}\mathcal{L}_{\text{ref}} + \lambda_{\text{div}}\mathcal{L}_{\text{div}} + \mathcal{L}_{\text{phy}}, \tag{13}$$

where λ_{ref} and λ_{div} decide the importance of losses. Besides, we use a low-pass filter to smooth the predicted poses generated by the diversity sampler after training. Please see the details in the supplementary material.

4.2. Short-term Oracle Design

We introduce an oracle to supervise the predictor in Section 4.1.2. In this section, we discuss how to obtain the oracle. We propose to learn a short-term oracle $\mathcal{O}(\boldsymbol{X}, \tau)$ by using another conditional variational autoencoder to capture the pseudo-ground-truth multi-modality. In order to achieve such goal, several works utilize the similarity search techniques [59]. This method is also used in [61, 62] as the multi-modality evaluation metrics. In our work, we define:

$$\begin{aligned}
&\Omega(\boldsymbol{X}_t) = \mathbf{S}(\mathcal{X}_{\text{o}}; \tau, K)\\
&\mathcal{X}_{\text{o}} = \{\boldsymbol{X}^1_{t+1:t+\tau} \dots \boldsymbol{X}^{|\mathcal{X}_{\text{o}}|}_{t+1:t+\tau}\}\\
&d(\boldsymbol{X}^j_t, \boldsymbol{X}_t) \le \delta, \forall j = 1, \dots, |\mathcal{X}_{\text{o}}|,
\end{aligned} \tag{14}$$

where $\mathcal{X}_o$ represents the set of all the future poses whose corresponding initial poses $\boldsymbol{X}^j_t$ are in a ball with radius δ which centered at the given initial pose $\boldsymbol{X}_t$. The ball is defined by metric $d(\cdot, \cdot)$. $\Omega(\boldsymbol{X}_t)$ represents the set of K selected future poses which has time horizon τ given the initial pose $\boldsymbol{X}_t$. Since there can be many similar poses to the given initial poses and most of the corresponding future poses are very similar, we need to select a proper fixed number of future poses in $\mathcal{X}_{\text{o}}$ in order to capture the different modes. Here we use the k-determinantal point process (k-DPP) as the selection strategy $\mathbf{S}$ to choose the future poses.

4.2.1 k-Determinantal Point Process

k-determinantal point process [31] is widely used to sample the diverse points given a fixed number of samples. Given a set $\mathcal{X} = \{X_1, X_2, \dots, X_n\}$, a k-determinantal point process defined on $\mathcal{X}$ is a probability measure on $2^{\mathcal{X}}$:

$$\mathbf{Pr}(\mathcal{S}) = \frac{\det(L_{\mathcal{S}})\mathbf{1}(|\mathcal{S}| = k)}{\sum_{\mathcal{S}\subset[n], |\mathcal{S}|=k} \det(L_{\mathcal{S}})}, \tag{15}$$

where we denote $\mathcal{S}$ as a subset of $\mathcal{X}$ and $L_{\mathcal{S}} \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}$ as the similarity matrix:

$$\{L_{\mathcal{S}}\}_{ij} = e^{-d(X^i_{t+1:t+\tau}, X^j_{t+1:t+\tau})}. \tag{16}$$

We preprocess the training data to augment each case with K futures poses. Several sampling algorithms [17, 25] for the determinantal point process can be used directly.

4.2.2 Short-term Oracle Model

The short-term oracle can be trained with any approach proposed in Section 2. In our experiments, we use a conditional variational autoencoder similar to the likelihood sampler defined above after getting the augmented data with the prediction horizon τ. Now, we can provide more diverse futures given the exact same historical observation. Since the augmented data is balanced by the k-determinantal point process, there will be fewer extremely minor modes and hence mitigate the trouble of rare-case sampling. The details of the short-term oracle neural network structure are provided in the supplementary material.

5. Training and Testing Process

The training procedure is summarized in Algorithm 1. We generate the same number of samples from both accuracy prior and diversity prior for training. Notice that the diversity loss $\mathcal{L}_{\text{div}}$ does not backpropagate to the accuracy prior $Q_{\text{acc}}(\boldsymbol{Z}|\boldsymbol{C})$ since we do not want the diversity loss influence the accuracy prior. After we get the optimized model, we can decide the ratio of diverse samples, which mainly focus on the most different modes compared with the major modes by adjusting the ratio number ρ. The testing procedure is summarized in Algorithm 2.

```
Algorithm 1: Training Procedure
  Input: $N$: number of epoches. $n_{\text{acc}}$: number of
         samples for accuracy sampler $Q_{\text{acc}}$. $n_{\text{div}}$:
         number of samples for diversity sampler
         $Q_{\text{div}}$. $n_o$: number of samples generated from
         oracle $\mathcal{O}$.
  Output: $\theta$, $\phi_{\text{acc}}$, $\phi_{\text{div}}$
  Data: Training dataset $\mathcal{D}_{train}$
  while epoch $\leq N$ do
      Sample $\mathcal{B} = \{\boldsymbol{X}^i, \boldsymbol{C}^i\}_i \sim \mathcal{D}_{train}$
      foreach $\boldsymbol{X}, \boldsymbol{C} \in \mathcal{B}$ do
          Generate $n_{\text{acc}}$ samples:
          $\hat{\boldsymbol{X}}^i_{\text{acc}} = \text{d}(\boldsymbol{X}|\boldsymbol{C}, \boldsymbol{z}^i), \boldsymbol{z}^i \sim Q_{\text{acc}}(\boldsymbol{Z}|\boldsymbol{C})$
          Generate $n_{\text{div}}$ samples:
          $\hat{\boldsymbol{X}}^i_{\text{div}} = \text{d}(\boldsymbol{X}|\boldsymbol{C}, \boldsymbol{z}^i), \boldsymbol{z}^i \sim Q_{\text{div}}(\boldsymbol{Z}|\boldsymbol{C})$
          for $s = 0, \ldots, T_f/\tau - 1$ do
              Generate $n_o$ samples:
              $\tilde{\boldsymbol{X}}^j_{t+s\tau+1:t+(s+1)\tau} \sim \mathcal{O}(\hat{\boldsymbol{X}}_{\text{div},t+s\tau}, \tau)$
      Update $\theta, \psi, \phi_{\text{acc}}$ with $\mathcal{L}_{\text{A}}$
      Update $\theta, \phi_{\text{div}}$ with $\mathcal{L}_{\text{D}}$
```

```
Algorithm 2: Testing Procedure
  Input: $\rho$: The proportion of samples from $Q_{\text{div}}$ in
         the total samples , $M$: the total number of
         samples
  Output: $\hat{\boldsymbol{X}}$, The predicted poses
  Data: Testing Dataset $\mathcal{D}_{test}$
  foreach $\boldsymbol{X}, \boldsymbol{C} \in \mathcal{D}_{test}$ do
      Generate $(1-\rho)M$ samples from $Q_{\text{acc}}$
      Generate $\rho M$ samples from $Q_{\text{div}}$
```

6. Experiments

In this section, we introduce the datasets and evaluation metrics first. Then the quantitative, qualitative analysis, and ablation analysis are provided. Implementation details, additional results, limitations, and future work are provided in the supplementary material.

6.1. Datasets

We evaluate our method on Human3.6M [26] and HumanEva-I dataset [51] and use identical settings with the other baselines. Human3.6M dataset consists of 11 subjects and 3.6 million video frames. There are 15 actions for each subject. The human motion is recorded at 50Hz. We adopt a 17-joint skeleton representation in our work. We use five subjects (S1, S5, S6, S7, S8) for training and testing with the other two subjects (S9 and S11). The predicted future motion horizon is 2 seconds (100 time steps), and the historical motion horizon is 0.5 seconds (25 time steps). HumanEva-I dataset includes three subjects. The record rate of human motion is 60Hz. We choose to use the 15-joint skeleton representation. We use the same training and testing datasets which are provided by the official website. We predict future motion for 1 second (60 time steps) with 0.25 seconds (15 time steps) observation.

6.2. Evaluation Metrics

The following metrics are used to evaluate the performance of methods. For accuracy, we use Average Displacement Error (ADE) which is defined as the average Euclidean distance over the prediction time steps between the ground truth motion $\boldsymbol{X}_{t+1:t+T_f}$ and the closest sample [62], and Final Displacement Error (FDE), which is the Euclidean distance between the final ground truth pose and the final predicted pose, i.e., $\min_i \|\hat{\boldsymbol{X}}^i_{t+T_f} - \boldsymbol{X}_{t+T_f}\|$. For diversity, we use Average Pairwise Distance (APD), which is the L2 distance between all pairs of motion samples, which is computed as $\frac{1}{K(K-1)} \sum_{i \neq j} \|\hat{\boldsymbol{X}}^i_{t+1:t+T_f} - \hat{\boldsymbol{X}}^j_{t+1:t+T_f}\|$.

6.3. Quantitative Analysis

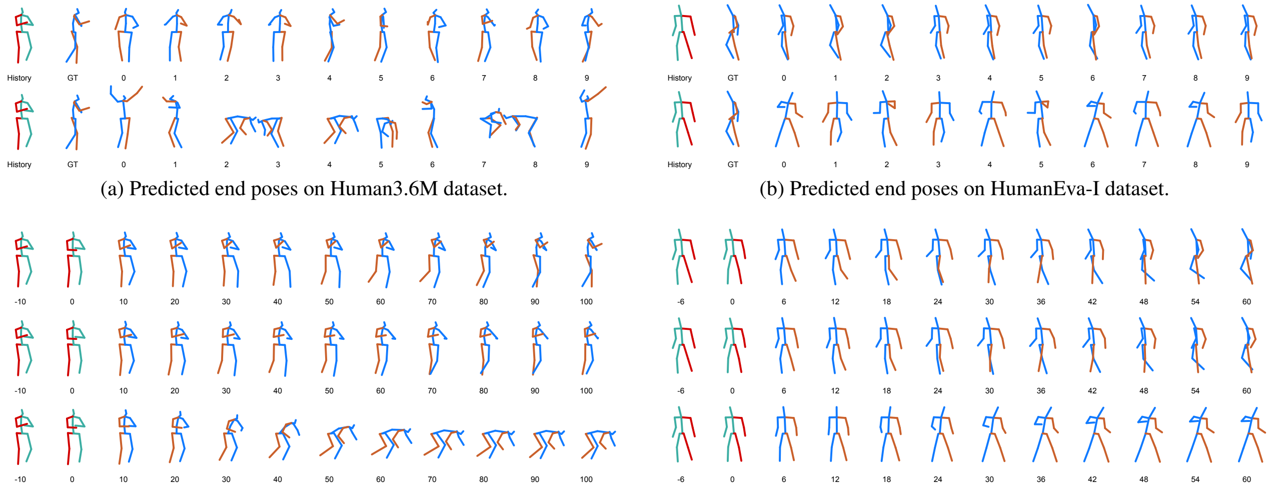
We compare our approach with several baselines in Table 1. The baselines include deterministic methods such as acLSTM [37] and ERD [21], probabilistic approaches such as MT-VAE [60] and Dlow [62], etc. We use 50 samples to evaluate the prediction performance for all methods. We directly use the results of baselines from [62] and [64].

In Table 1 we can conclude that our method with $\rho = 0.46$ can achieve a better performance compared with the other baselines in terms of all the metrics. In general, probabilistic methods such as Best-of-Many and GMVAE can achieve better accuracy and diversity than deterministic ones such as acLSTM and ERD. We can observe that most of the methods will have worse ADE and FDE if the APD

Human3.6M											
	ERD [21]	acLSTM [37]	Pose-Knows [56]	MT-VAE [60]	HP-GAN [5]	BoM [9]	GMVAE [19]	DeLiGAN [24]	DSF [61]	Dlow [62]	DCT5/DCT20 [64]	Ours
APD ↑	0	0	6.723	0.403	7.214	6.265	6.769	6.509	9.330	11.74	12.579/**15.920**	**14.24**
ADE ↓	0.722	0.789	0.461	0.457	0.858	0.448	0.461	0.483	0.493	0.425	**0.412**/0.416	**0.414**
FDE ↓	0.969	1.126	0.560	0.595	0.867	0.533	0.555	0.534	0.592	0.518	**0.514**/0.522	**0.516**

HumanEva-I												
	ERD [21]	acLSTM [37]	Pose-Knows [56]	MT-VAE [60]	HP-GAN [5]	BoM [9]	GMVAE [19]	DeLiGAN [24]	DSF [61]	Dlow [62]	DCT5/DCT20 [64]	Ours
APD ↑	0	0	2.308	0.021	1.139	2.846	2.443	2.177	4.538	4.855	4.181/**6.266**	**5.786**
ADE ↓	0.382	0.429	0.269	0.345	0.772	0.271	0.305	0.306	0.273	0.251	**0.234**/0.239	**0.228**
FDE ↓	0.461	0.541	0.296	0.403	0.749	0.279	0.345	0.322	0.290	0.268	**0.244**/0.253	**0.236**

Table 1. Quantitative results on Human3.6M and HumanEva-I dataset. Our results and the best results of baselines are highlighted.



(a) Predicted end poses on Human3.6M dataset.

(b) Predicted end poses on HumanEva-I dataset.

(c) Predicted human motion on Human3.6M dataset.

(d) Predicted human motion on HumanEva-I dataset.

Figure 4. Visualization of prediction results on Human3.6M and HumanEva-I dataset. Figure 4a and Figure 4b illustrate ten predicted end poses generated from both accuracy prior (first row) and diversity prior (second row). Figure 4c and Figure 4d show the predicted time sequences. The sequence in the first row is the ground truth motion. The sequence in the second row is one of the samples generated from accuracy prior and the sequence in the third row is one of the samples generated from diversity prior.

is larger in general. It is because there exists a trade-off between diversity and accuracy. Compared with DLow, our approach improve the performance on both Human3.6M and HumanEva-I datasets. We also compare our results with DCT5 and DCT20 in [64], which use the frequency representation with the CVAE framework. Our results are on a par with their performance on both datasets.

6.4. Qualitative Analysis

We illustrate 10 end poses of random samples generated from both accuracy prior function and diversity function in Figure 4a and 4b. The first row shows the samples from the accuracy prior function. We notice that most samples are similar to the ground truth, which represents that the accuracy sampler can generate the predicted future human motions with high accuracy. The second row shows the samples from the diversity sampler. We notice that the predicted poses from the diversity sampler have more different modes and are not similar to the samples generated from the accuracy sampler. It can be attributed to the second item in the diversity loss $\mathcal{L}_{\text{div}}$ in Equation 9, where we encourage our diversity prior to generating dissimilar samples to the ones generated from the accuracy sampler. We also illustrate two samples of predicted human motion from both accuracy and diversity sampler for both datasets in Figure 4c and Figure 4d. We notice that the predicted time sequences are smooth. The samples from the accuracy sampler can be very accurate compared with the ground truth. More visualization results are provided in the supplementary material.

6.5. Different Sampling Ratio

In Figure 5, we illustrate the different metrics values with respect to the number of samples n_{acc} generated from the accuracy sampler during testing. When n_{acc} equals 0, it means that we only sample from the diversity prior distribution. We can see that ADE and FDE increase since the diversity sampler is designed to focus on exploring more different possible modes instead of matching the likelihood of data. Hence, we observe that APD can achieve around 18 when $n_{\text{acc}} = 0$. When n_{acc} increases, we observe that both the accuracy metrics (ADE and FDE) and diversity metric (APD) decrease. The accuracy metrics decrease slowly when the n_{acc} is large enough. When $n_{\text{acc}} = 50$, which means that all the samples are generated from the accuracy sampler, we observe that APD decreases to around 6 and the accuracy metrics achieve the best performance.

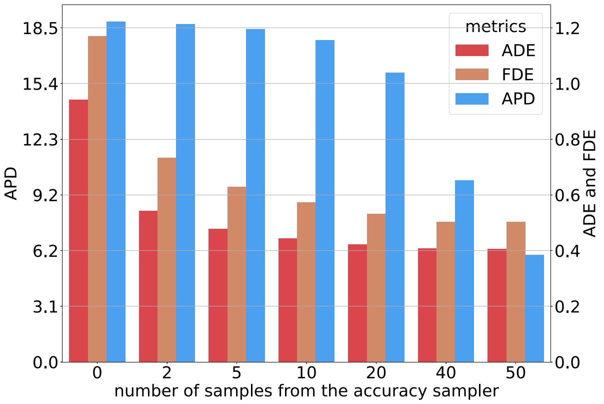
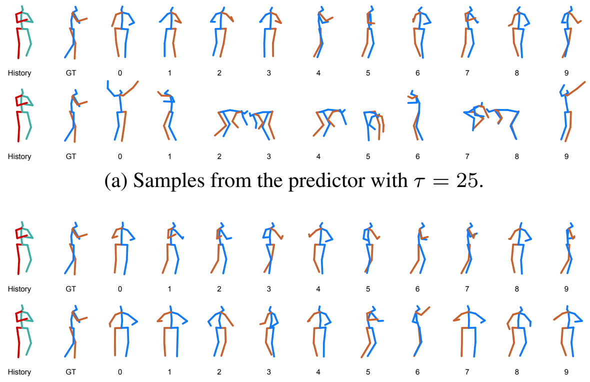


Figure 5. APD, ADE and FDE with respect to n_{acc} on Human3.6M dataset. Red and brown bars indicate the accuracy metrics ADE and FDE. The blue bar indicates the APD.



(a) Samples from the predictor with $\tau = 25$.

(b) Samples from the predictor with $\tau = 100$.

Figure 6. Visualization of predicted end poses of motions on Human3.6M dataset with different oracles. Figure 6a illustrates the performance of the predictor with oracle ($\tau = 25$). Figure 6b illustrates the performance of the predictor with oracle ($\tau = 100$). In each figure, the first row are the samples generated from accuracy prior function. The second row are the samples generated from the diversity prior function.

6.6. Ablation Analysis

Using Short-term Oracle Prediction Horizon τ In order to investigate whether dividing the prediction horizon into several short-term ones can help the predictor discover more possible modes, we evaluate our models with the oracles which have different prediction horizon length τ. We compare our framework supervised by a short-term oracle with $\tau = 25$ and our framework supervised by the one with the full-length of prediction horizon, i.e., the prediction is not divided into short-term subsequences. We show the results of Human3.6M dataset in Figure 6. We can see that when using the oracle with $\tau = 100$, i.e., the prediction horizon of the oracle is not short-term and the horizon is the same as the target prediction horizon, and the diversity is lower than the one which has $\tau = 25$. It shows that the oracle with a short prediction horizon indeed increases the diversity. We also compare the different metrics of two models with different τ, and the results are summarized in Table 2. We notice that ADE and FDE with $n_{\text{acc}} = 50$ of both models with $\tau = 100$ and $\tau = 25$ are similar since all the samples are from the accuracy samplers. However, when n_{acc} decreases, we observe that the diversity of the model supervised by the oracle with $\tau = 100$ does not increase so much. We also observed that ADE and FDE of the model supervised with oracle (τ=100) does not change too much when n_{acc} is greater than 28 and APD does not change too much when n_{acc} is smaller than 14. It is reasonable since the model supervised by the oracle with $\tau = 100$ only explores limited and less possible diverse modes than the one supervised by the oracle with $\tau = 25$. It also supports our suggestion that the short-term oracle indeed helps the predictor discover more possible future motions meanwhile the accuracy of prediction is maintained.

	$\tau = 25$, short-term							
n_{acc}	0	7	14	21	28	35	42	50
ADE ↓	0.941	0.459	0.433	0.421	0.413	0.411	0.407	0.404
FDE ↓	1.170	0.598	0.551	0.529	0.515	0.510	0.504	0.501
APD ↑	18.79	18.16	17.14	15.73	14.04	12.02	9.265	5.927
	$\tau = 100$, non-short-term							
n_{acc}	0	7	14	21	28	35	42	50
ADE ↓	0.504	0.431	0.417	0.409	0.406	0.402	0.401	0.402
FDE ↓	0.580	0.523	0.505	0.495	0.491	0.486	0.488	0.497
APD ↑	7.346	7.397	7.360	7.234	6.997	6.683	6.255	5.651

Table 2. The comparison with different τ on Human3.6M dataset.

7. Conclusion

In this work, we propose a multi-objective diverse human motion prediction framework, which can enable adjustable sampling during the testing time. In order to enhance the diversity of predicted poses, we introduce a short-term oracle to instruct the predictor to discover more diverse possible modes of future poses. Such a framework overcomes the trade-off between likelihood sampling and diversity sampling. Thanks to both the multi-objective structure and short-term oracle, Our proposed approach achieves state-of-the-art performance in terms of accuracy and diversity. The experiment results and ablation studies demonstrate the effectiveness of the proposed method. Several future directions could be investigated. First, since our proposed approach is a general framework, more complicated structures such as graph neural network and transformer can be incorporated. Second, we currently assume that the horizon of short-term oracle is fixed. How to dynamically decide the short-term horizon will be the future work.

References

[1] Emre Aksan, Manuel Kaufmann, and Otmar Hilliges. Structured prediction helps 3d human motion modelling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7144–7153, 2019. 2

[2] Mohammad Sadegh Aliakbarian, Fatemeh Sadat Saleh, Mathieu Salzmann, Lars Petersson, Stephen Gould, and Amirhossein Habibian. Learning variations in human motion via mix-and-match perturbation. *arXiv preprint arXiv:1908.00733*, 2019. 1, 2

[3] Andrea Bajcsy, Somil Bansal, Ellis Ratner, Claire J Tomlin, and Anca D Dragan. A robust control framework for human motion prediction. *IEEE Robotics and Automation Letters*, 6(1):24–31, 2020. 1

[4] Andrea Bajcsy, Anand Siththaranjan, Claire J Tomlin, and Anca D Dragan. Analyzing human models that adapt online. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2754–2760. IEEE, 2021. 1

[5] Emad Barsoum, John Kender, and Zicheng Liu. Hp-gan: Probabilistic 3d human motion prediction via gan. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 1418–1427, 2018. 2, 7

[6] Justin Bayer, Daan Wierstra, Julian Togelius, and Jürgen Schmidhuber. Evolving memory cell structures for sequence learning. In *International conference on artificial neural networks*, pages 755–764. Springer, 2009. 2

[7] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013. 2

[8] Apratim Bhattacharyya, Michael Hanselmann, Mario Fritz, Bernt Schiele, and Christoph-Nikolas Straehle. Conditional flow variational autoencoders for structured sequence prediction. *arXiv preprint arXiv:1908.09008*, 2019. 4

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

[10] Matthew Brand and Aaron Hertzmann. Style machines. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 183–192, 2000. 2

[11] Defu Cao, Jiachen Li, Hengbo Ma, and Masayoshi Tomizuka. Spectral temporal graph neural network for trajectory prediction. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1839–1845. IEEE, 2021. 2

[12] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014. 2

[13] Chiho Choi, Joon Hee Choi, Jiachen Li, and Srikanth Malla. Shared cross-modal trajectory prediction for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 244–253, 2021. 1

[14] Chiho Choi and Behzad Dariush. Looking to relations for future trajectory forecast. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 921–930, 2019. 1

[15] Enric Corona, Albert Pumarola, Guillem Alenya, and Francesc Moreno-Noguer. Context-aware human motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6992–7001, 2020. 14

[16] Antonia Creswell, Tom White, Vincent Dumoulin, Kai Arulkumaran, Biswa Sengupta, and Anil A Bharath. Generative adversarial networks: An overview. *IEEE Signal Processing Magazine*, 35(1):53–65, 2018. 2

[17] Michal Derezinski, Daniele Calandriello, and Michal Valko. Exact sampling of determinantal point processes with sublinear time preprocessing. *Advances in neural information processing systems*, 32, 2019. 5

[18] Rahul Dey and Fathi M Salem. Gate-variants of gated recurrent unit (gru) neural networks. In *2017 IEEE 60th international midwest symposium on circuits and systems (MWSCAS)*, pages 1597–1600. IEEE, 2017. 2

[19] Nat Dilokthanakul, Pedro AM Mediano, Marta Garnelo, Matthew CH Lee, Hugh Salimbeni, Kai Arulkumaran, and Murray Shanahan. Deep unsupervised clustering with gaussian mixture variational autoencoders. *arXiv preprint arXiv:1611.02648*, 2016. 7

[20] Conor Durkan, Artur Bekasov, Iain Murray, and George Papamakarios. Neural spline flows. *Advances in neural information processing systems*, 32, 2019. 2, 13

[21] Katerina Fragkiadaki, Sergey Levine, Panna Felsen, and Jitendra Malik. Recurrent network models for human dynamics. In *Proceedings of the IEEE international conference on computer vision*, pages 4346–4354, 2015. 2, 6, 7

[22] Jennifer A Gillenwater, Alex Kulesza, Emily Fox, and Ben Taskar. Expectation-maximization for learning determinantal point processes. *Advances in Neural Information Processing Systems*, 27, 2014. 2

[23] Boqing Gong, Wei-Lun Chao, Kristen Grauman, and Fei Sha. Diverse sequential subset selection for supervised video summarization. *Advances in neural information processing systems*, 27, 2014. 2

[24] Swaminathan Gurumurthy, Ravi Kiran Sarvadevabhatla, and R Venkatesh Babu. Deligan: Generative adversarial networks for diverse and limited data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 166–174, 2017. 7

[25] J Ben Hough, Manjunath Krishnapur, Yuval Peres, and Bálint Virág. Determinantal processes and independence. *Probability surveys*, 3:206–229, 2006. 5

[26] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. 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. 6

[27] Ashesh Jain, Amir R Zamir, Silvio Savarese, and Ashutosh Saxena. Structural-rnn: Deep learning on spatio-temporal

graphs. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 5308–5317, 2016. 2

[28] Durk P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. Semi-supervised learning with deep generative models. *Advances in neural information processing systems*, 27, 2014. 2

[29] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 2

[30] Thomas Kipf, Ethan Fetaya, Kuan-Chieh Wang, Max Welling, and Richard Zemel. Neural relational inference for interacting systems. In *International Conference on Machine Learning*, pages 2688–2697. PMLR, 2018. 2

[31] Alex Kulesza and Ben Taskar. k-dpps: Fixed-size determinantal point processes. In *ICML*, 2011. 2, 5

[32] Jogendra Nath Kundu, Maharshi Gor, and R Venkatesh Babu. Bihmp-gan: Bidirectional 3d human motion prediction gan. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 8553–8560, 2019. 2

[33] Jiachen Li, Hengbo Ma, and Masayoshi Tomizuka. Conditional generative neural system for probabilistic trajectory prediction. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 6150–6156. IEEE, 2019. 1

[34] Jiachen Li, Fan Yang, Hengbo Ma, Srikanth Malla, Masayoshi Tomizuka, and Chiho Choi. Rain: Reinforced hybrid attention inference network for motion forecasting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16096–16106, 2021. 2

[35] Jiachen Li, Fan Yang, Masayoshi Tomizuka, and Chiho Choi. Evolvegraph: Multi-agent trajectory prediction with dynamic relational reasoning. *Advances in neural information processing systems*, 33:19783–19794, 2020. 2

[36] Maosen Li, Siheng Chen, Yangheng Zhao, Ya Zhang, Yanfeng Wang, and Qi Tian. 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*, pages 214–223, 2020. 1, 2

[37] Xiu Li, Hongdong Li, Hanbyul Joo, Yebin Liu, and Yaser Sheikh. Structure from recurrent motion: From rigidity to recurrency. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3032–3040, 2018. 2, 6, 7

[38] Xiao Lin and Mohamed R Amer. Human motion modeling using dvgans. *arXiv preprint arXiv:1804.10652*, 2018. 2

[39] Hengbo Ma, Jiachen Li, Wei Zhan, and Masayoshi Tomizuka. Wasserstein generative learning with kinematic constraints for probabilistic interactive driving behavior prediction. In *2019 IEEE Intelligent Vehicles Symposium (IV)*, pages 2477–2483. IEEE, 2019. 1

[40] Hengbo Ma, Yaofeng Sun, Jiachen Li, and Masayoshi Tomizuka. Multi-agent driving behavior prediction across different scenarios with self-supervised domain knowledge. In *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*, pages 3122–3129. IEEE, 2021. 1

[41] Hengbo Ma, Yaofeng Sun, Jiachen Li, Masayoshi Tomizuka, and Chiho Choi. Continual multi-agent interaction behavior prediction with conditional generative memory. *IEEE Robotics and Automation Letters*, 6(4):8410–8417, 2021. 1

[42] Wei Mao, Miaomiao Liu, and Mathieu Salzmann. Generating smooth pose sequences for diverse human motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13309–13318, 2021. 2, 5, 13, 14

[43] Wei Mao, Miaomiao Liu, Mathieu Salzmann, and Hongdong Li. Learning trajectory dependencies for human motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9489–9497, 2019. 2

[44] Julieta Martinez, Michael J Black, and Javier Romero. On human motion prediction using recurrent neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2891–2900, 2017. 2

[45] Soohwan Park, Hoseok Ryu, Seyoung Lee, Sunmin Lee, and Jehee Lee. Learning predict-and-simulate policies from unorganized human motion data. *ACM Transactions on Graphics (TOG)*, 38(6):1–11, 2019. 1

[46] Dario Pavllo, David Grangier, and Michael Auli. Quaternet: A quaternion-based recurrent model for human motion. *arXiv preprint arXiv:1805.06485*, 2018. 2

[47] Nicholas Rhinehart, Kris M Kitani, and Paul Vernaza. R2p2: A reparameterized pushforward policy for diverse, precise generative path forecasting. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 772–788, 2018. 1, 2

[48] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE transactions on neural networks*, 20(1):61–80, 2008. 2

[49] Jan Sedmidubsky and Pavel Zezula. Similarity search in 3d human motion data. In *Proceedings of the 2019 on International Conference on Multimedia Retrieval*, pages 5–6, 2019. 2

[50] Hedvig Sidenbladh, Michael J Black, and Leonid Sigal. Implicit probabilistic models of human motion for synthesis and tracking. In *European conference on computer vision*, pages 784–800. Springer, 2002. 2

[51] Leonid Sigal, Alexandru O Balan, and Michael J Black. Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *International journal of computer vision*, 87(1):4–27, 2010. 6

[52] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems*, 28, 2015. 4

[53] Chen Tang, Wei Zhan, and Masayoshi Tomizuka. Exploring social posterior collapse in variational autoencoder for interaction modeling. *Advances in Neural Information Processing Systems*, 34, 2021. 4

[54] Graham W Taylor, Geoffrey E Hinton, and Sam Roweis. Modeling human motion using binary latent variables. *Advances in neural information processing systems*, 19, 2006. 2

[55] Jakub Tomczak and Max Welling. Vae with a vampprior. In *International Conference on Artificial Intelligence and Statistics*, pages 1214–1223. PMLR, 2018. 4

[56] Jacob Walker, Kenneth Marino, Abhinav Gupta, and Martial Hebert. The pose knows: Video forecasting by generating pose futures. In *Proceedings of the IEEE international conference on computer vision*, pages 3332–3341, 2017. 1, 7

[57] Borui Wang, Ehsan Adeli, Hsu-kuang Chiu, De-An Huang, and Juan Carlos Niebles. Imitation learning for human pose prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7124–7133, 2019. 2

[58] Jack M Wang, David J Fleet, and Aaron Hertzmann. Gaussian process dynamical models for human motion. *IEEE transactions on pattern analysis and machine intelligence*, 30(2):283–298, 2007. 2

[59] Jingwei Xu, Huazhe Xu, Bingbing Ni, Xiaokang Yang, and Trevor Darrell. Video prediction via example guidance. In *International Conference on Machine Learning*, pages 10628–10637. PMLR, 2020. 2, 5

[60] Xinchen Yan, Akash Rastogi, Ruben Villegas, Kalyan Sunkavalli, Eli Shechtman, Sunil Hadap, Ersin Yumer, and Honglak Lee. Mt-vae: Learning motion transformations to generate multimodal human dynamics. In *Proceedings of the European conference on computer vision (ECCV)*, pages 265–281, 2018. 1, 2, 6, 7

[61] Ye Yuan and Kris Kitani. Diverse trajectory forecasting with determinantal point processes. *arXiv preprint arXiv:1907.04967*, 2019. 1, 2, 5, 7, 14

[62] Ye Yuan and Kris Kitani. Dlow: Diversifying latent flows for diverse human motion prediction. In *European Conference on Computer Vision*, pages 346–364. Springer, 2020. 2, 3, 5, 6, 7, 14

[63] Andrei Zanfir, Eduard Gabriel Bazavan, Hongyi Xu, William T Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Weakly supervised 3d human pose and shape reconstruction with normalizing flows. In *European Conference on Computer Vision*, pages 465–481. Springer, 2020. 1, 2, 5, 13

[64] Yan Zhang, Michael J Black, and Siyu Tang. We are more than our joints: Predicting how 3d bodies move. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3372–3382, 2021. 2, 6, 7, 14

[65] Zachary Ziegler and Alexander Rush. Latent normalizing flows for discrete sequences. In *International Conference on Machine Learning*, pages 7673–7682. PMLR, 2019. 4