Multi-Objective Diverse Human Motion Prediction with Knowledge Distillation

Hengbo Ma\textsuperscript{1,2†} Jiachen Li\textsuperscript{3} Ramtin Hosseini\textsuperscript{1,4†} Masayoshi Tomizuka\textsuperscript{2} Chiho Choi\textsuperscript{1}
\textsuperscript{1}Honda Research Institute, USA \textsuperscript{2}University of California, Berkeley
\textsuperscript{3}Stanford University \textsuperscript{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 \cite{3, 4}, autonomous driving \cite{13, 14, 33, 39}, and animation \cite{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 \cite{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.

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 \cite{36} based on graph neural network and \cite{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 \cite{2, 47, 60, 63} or diverse sampling techniques \cite{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
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.
3. Problem Formulation

Our goal is to predict the possible future human motions given a dataset \( D \). We denote the human motion with time horizon \( T = T_h + T_f \) as \( X_{t-T_h+1:t+T_f} = [X_{t-T_h+1}, \ldots, X_{t+T_f}] \), where \( X_t \in \mathbb{R}^d \) 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 \( C = X_{t-T_h+1:t} \), we intend to get the future motion distribution \( P(X_{t+1:t+T_f} | C, \rho) \). 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 \( X_{t+1:T_f} \sim P(X_{t+1:T_f} | C, \rho), \) \( \forall \rho \). The larger \( \rho \) is, the more diverse samples will be generated and focuses on the rare cases, and the smaller \( \rho \) is, the prediction will focus more on the most likely modes. For simplicity, we use \( X \) represent \( X_{t+1:T_f} \) 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:

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

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

\[
\begin{align*}
    e_h(C) &= [\text{MLP} \circ \text{RNN}](C) \\
    e_f(X) &= [\text{MLP} \circ \text{RNN}](X),
\end{align*}
\]

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 \( e_h(C) \) and latent variable \( Z \), We denote the decoder function \( d_0(X | C, Z) \) as:

\[
d_0(X | C, Z) = [\text{MLP} \circ \text{RNN}](e_h(C)||Z),
\]

where \( \theta \) 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 \( Z \). “||” 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 \( \phi \) are in the supplementary material.
4.1.1 Accuracy Sampler

The first objective is to infer the accuracy prior distribution \( Q_{\text{acc}}(Z|C) \). 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:

\[
L_{\text{ELBO}} = \mathbb{E}_{Q_{\psi}(Z|X,C)}[\log P_{\theta}(X|Z,C)] - D_{KL}[Q_{\psi}(Z|X,C)||Q_{\text{acc}}(Z|C)],
\]

where \( Q_{\psi}(Z|X,C) \) is the posterior distribution of latent variable \( Z \) 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 \( \mathcal{N}(0, I) \) which is independent of \( C \), we model \( Q_{\text{acc}}(Z|C) \) as a Gaussian distribution \( \mathcal{N}(\mu_{\text{acc}}(C), \Sigma_{\text{acc}}(C)) \). The \( D_{KL}[Q_{\psi}||Q_{\text{acc}}] \) is:

\[
\frac{1}{2} \log \left( \frac{\Sigma_{\text{acc}}}{\Sigma_{\psi}} \right) - n_z + Tr(\Sigma_{\text{acc}}^{-1} \Sigma_{\psi}) + ||\mu_{\text{acc}} - \mu_{\psi}||_{\Sigma_{\text{acc}}^{-1}}^2.
\]

which can be calculated analytically. Since we have no control of the distribution \( Q_{\text{acc}}(Z|C) \), 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:

\[
R_{\text{acc}} = \min_{i} \| \hat{X}^i - X \|^2
\]

\[
z^i \sim Q(Z|C)
\]

\[
\hat{X}^i = d_{\theta}(X|C, z^i), \quad i = 1, \ldots, n_{\text{acc}},
\]

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

\[
L_{\lambda}(\theta, \psi) = -\lambda_{\text{ELBO}} L_{\text{ELBO}} + \lambda_{\text{acc}} R_{\text{acc}},
\]

where \( \lambda_{\text{elbo}} \) and \( \lambda_{\text{acc}} \) 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 \( Q_{\text{div}}(Z|C) \) with parameter \( \phi_{\text{div}} \). We utilize a common diversity loss definition:

\[
\text{DIV}(\mathcal{X}, \mathcal{Y}) = \frac{1}{N_x N_y} \sum_{i,j} e^{-d(X^i, Y^j)}
\]

\[
X^i, Y^j \in \mathcal{X}, \mathcal{Y}, \quad i = 1, \ldots, N_x, \quad j = 1, \ldots, N_y,
\]

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 \( \eta \) 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:

\[
L_{\text{div}} = \alpha_{\text{div}} \text{DIV}(\mathcal{X}_{\text{div}}, \mathcal{X}_{\text{div}}) + (1 - \alpha_{\text{div}}) \text{DIV}(\mathcal{X}_{\text{div}}, \mathcal{X}_{\text{acc}}),
\]

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 \( \alpha_{\text{div}} \). A larger \( \alpha_{\text{div}} \) means that we focus on making the samples from \( Q_{\text{div}} \) 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:

\[
\hat{X}_{i+1:T} \sim O(X_i, \tau),
\]

where \( O(X_i, \tau) \) is the probabilistic distribution of future poses with horizon \( \tau \) given the current initial pose \( X_i \). The oracle \( O \) 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:

\[
L_{\text{ref}}(\tau) = \frac{1}{n_{\text{div}}} \sum_{i,s,t} \min_j \| \hat{X}^i_{s+1:(s+1)\tau} - \hat{X}^j_{s+1:(s+1)\tau} \|_2^2,
\]

\[
s.t. z^i \sim Q_{\text{div}}(Z|C), \quad \hat{X}^i_{1:T} = d_{\theta}(X|C, z^i),
\]

\[
\hat{X}^j_{s+1:(s+1)\tau} \sim O(X_{s\tau}, \tau),
\]

\[
i = 1, \ldots, n_{\text{div}}, \quad j = 1, \ldots, n_{\text{div}}, \quad s = 0, \ldots, T/\tau - 1,
\]

where \( \tau \) represents the time interval of predicted poses from the oracle. \( n_{\text{div}} \) is the number of samples generated from diversity prior, and \( n_{\text{div}} \) 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 \( X_{1:T} \), the oracle provides several possible short-term
futures $\tilde{X}_{\tau+1:(s+1)\tau}^j$ given the current predicted pose $\tilde{X}_s^j$ recursively. We enforce the short-term predicted sequence $\tilde{X}_{\tau+1:(s+1)\tau}^j$ to be similar with one of the provided futures $\tilde{X}_{\tau+1:(s+1)\tau}^j$. Notice that the diversity loss $L_{\text{div}}$ 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 $L_{\text{limb}}$ and the velocity constraint $L_{\text{vel}}$ as $L_{\text{phy}}$:

$$L_{\text{phy}} = \lambda_{\text{vel}} L_{\text{vel}} + L_{\text{limb}}.$$  \hspace{1cm} (12)

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

$$L_{D} = \lambda_{\text{ref}} L_{\text{ref}} + \lambda_{\text{div}} L_{\text{div}} + L_{\text{phy}},$$  \hspace{1cm} (13)

where $\lambda_{\text{ref}}$ and $\lambda_{\text{div}}$ 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 $O(\mathcal{X}, \tau)$ 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 [39]. This method is also used in [61, 62] as the multi-modality evaluation metrics. In our work, we define:

$$\Omega(\mathcal{X}_t) = S(\mathcal{X}_o; \tau, K)$$

where $\mathcal{X}_o$ represents the set of all the future poses whose corresponding initial poses $\mathcal{X}_i^j$ are in a ball with radius $\delta$ which centered at the given initial pose $\mathcal{X}_t$. The ball is defined by metric $d(\cdot, \cdot)$. $\Omega(\mathcal{X}_t)$ represents the set of $K$ selected future poses which has time horizon $\tau$ given the initial pose $\mathcal{X}_t$. 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 $\mathcal{X}_o$ in order to capture the different modes. Here we use the k-determinantal point process (k-DPP) as the selection strategy $S$ 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 $\mathcal{X} = \{X_1, X_2, \ldots, X_n\}$, a k-determinantal point process defined on $\mathcal{X}$ is a probability measure on $2^\mathcal{X}$:

$$\Pr(S) = \frac{\det(L_S)1(|S| = k)}{\sum_{S \subset [n], |S| = k} \det(L_S)},$$  \hspace{1cm} (15)

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

$$\{L_S\}_{ij} = e^{-d(X_{t+1:t+\tau}, X_{t+1:t+\tau}^j)},$$  \hspace{1cm} (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 $\tau$. 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 $L_{\text{div}}$ does not backpropagate to the accuracy prior $Q_{\text{acc}}(Z|C)$ 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 $\rho$. 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 $Q_{\text{div}}$. $n_{\text{oracle}}$: number of samples generated from oracle $\mathcal{O}$.

**Output:** $\theta, \phi_{\text{acc}}, \phi_{\text{div}}$.

**Data:** Training dataset $\mathcal{D}_{\text{train}}$.

1. while epoch $\leq N$ do
2. Sample $\mathcal{B} = \{X^i, C^i\}_{i=1}^\infty \sim \mathcal{D}_{\text{train}}$.
3. foreach $X, C \in \mathcal{B}$ do
4. Generate $n_{\text{acc}}$ samples: $X_{\text{acc}}^i = d(X|C, z^i)^i \sim Q_{\text{acc}}(Z|C)$.
5. Generate $n_{\text{div}}$ samples: $X_{\text{div}}^i = d(X|C, z^i)^i \sim Q_{\text{div}}(Z|C)$.
6. for $s = 0, \ldots, T_f/\tau - 1$ do
7. Generate $n_s$ samples: $X_{f+s\tau+1:T_f+1}^i \sim Q(\hat{X}_{\text{div}}, t+s\tau, \tau)$.
8. Update $\theta, \psi, \phi_{\text{acc}}$ with $L_A$.
9. Update $\theta, \phi_{\text{div}}$ with $L_D$.

### 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:** $\hat{X}$: The predicted poses.

**Data:** Testing Dataset $\mathcal{D}_{\text{test}}$.

1. foreach $X, C \in \mathcal{D}_{\text{test}}$ do
2. Generate $(1-\rho)M$ samples from $Q_{\text{acc}}$.
3. 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 $X_{f+1:T_f}$ 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., $\text{min}_t \|X_{f+1:T_f} - X_{t+T_f}\|$. For diversity, we use Average Pairwise Distance (APD), which is the L2 distance between all pairs of motion samples, which is computed as $\frac{1}{K(K-1)} \sum_{i \neq j} \|X_{f+1:T_f} \sim X_{f+1:T_f}\|$.  

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
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 $L_{div}$ 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 $n_{acc} = 0$. 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 $n_{acc} = 50$, 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.

```
<table>
<thead>
<tr>
<th>n_acc</th>
<th>0</th>
<th>7</th>
<th>14</th>
<th>21</th>
<th>28</th>
<th>35</th>
<th>42</th>
<th>50</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADE ↓</td>
<td>0.941</td>
<td>0.459</td>
<td>0.433</td>
<td>0.421</td>
<td>0.413</td>
<td>0.411</td>
<td>0.407</td>
<td>0.404</td>
</tr>
<tr>
<td>FDE ↓</td>
<td>1.170</td>
<td>0.598</td>
<td>0.551</td>
<td>0.529</td>
<td>0.515</td>
<td>0.510</td>
<td>0.504</td>
<td>0.501</td>
</tr>
<tr>
<td>APD ↑</td>
<td>18.79</td>
<td>18.16</td>
<td>17.14</td>
<td>15.73</td>
<td>14.04</td>
<td>12.02</td>
<td>9.265</td>
<td>5.927</td>
</tr>
</tbody>
</table>
```

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

using the oracle with τ = 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 τ = 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_acc = 50 of both models with τ = 100 and τ = 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 τ = 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 τ = 100 only explores limited and less possible diverse modes than the one supervised by the oracle with τ = 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.

### 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


[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