HiT-DVAE: Human Motion Generation via Hierarchical Transformer Dynamical VAE

Anonymous authors
Paper under double-blind review

Abstract

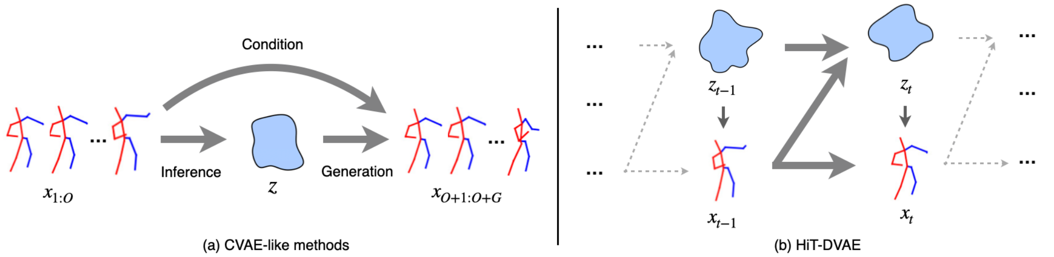
In this paper, we address the problem of 3D human motion generation, which aims at learning a model to generate plausible and diverse future sequences of 3D human poses from an observed one. Current state-of-the-art solutions propose injecting a single random latent vector into a deterministic motion prediction framework. The stochasticity in the generative process is thus modeled at the whole sequence level, which is inconsistent with the inherent time-dependent uncertainty of human motion (*e.g.* people can jump or walk after getting up from a chair). To overcome this limitation we propose Hierarchical Transformer Dynamical Variational Autoencoder (HiT-DVAE), a deep generative model with *sequential latent variables* that can efficiently learn the stochastic dynamics of human motion. The proposed model learns an expressive time-varying latent space that encodes diverse and realistic human motions. A thorough evaluation on HumanEva-I and Human3.6M datasets using various metrics shows that HiT-DVAE performs better than current state-of-the-art methods. Our code will be released upon publication.

1 Introduction

Human motion forecasting has drawn much attention with a wide range of applications such as autonomous driving (Brian et al., 2016), human-robot interaction (Hema & Ashutosh, 2013), and 3D game productions (Van Welbergen et al., 2010). Deterministic human motion prediction, which aims at forecasting the exact future based on a sequence of past observation (Fragkiadaki et al., 2015; Martinez et al., 2017; Mao et al., 2019; Guo et al., 2022a), has a quick development with the availability of large-scale datasets (Ionescu et al., 2014; Mahmood et al., 2019). While this task ignores the time-dependent *uncertainty* of human motion (future motion always has multiple reasonable possibilities, *e.g.* seeing a person lifting his hand, it is hard to tell if he is going to high-five or waving goodbye). Recent works about stochastic human motion generation take this problem into consideration and aim at generating diverse but plausible future motions (Yan et al., 2018; Yuan & Kitani, 2020; Mao et al., 2021; Dang et al., 2022).

Stochastic human motion generation introduces several challenges: i) dynamics: a generation model must be able to learn and mimic the dynamics of the 3D human motion and avoid collapsing to a static motion. ii) diversity: in contrast to deterministic motion prediction, the generation task should not only learn average motion patterns but needs to faithfully reflect the intrinsic intra-class variability. iii) contextual consistency: the generated motions have to be contextually consistent with the observed motion.

Some of these challenges have been partially addressed in previous works. For example, MT-VAE (Yan et al., 2018) combines a motion prediction model based on a recurrent neural network (RNN) with a conditional variational autoencoder (CVAE). The difference between the observed and future poses is encoded into the CVAE latent variable, which was then concatenated with the RNN's hidden state to account for the dynamics. DLow (Yuan & Kitani, 2020) proposes to generate a diverse set of motion sequences by training a large number of different encoders (fifty in their implementation) and a single decoder, obtaining different instances of the latent variable to generate different motions. GSPS (Mao et al., 2021) inherits the diversity loss from DLow but uses a motion prediction framework based on graph convolutional network (GCN) rather than RNN. The diversity of the generated motions is enforced by concatenating random noise to the observed sequences. More recently, Dang et al. (2022) proposes to encode the past observations into a base vector space



(a) CVAE-like methods

(b) HiT-DVAE

Figure 1: Conditional variational autoencoder (CVAE) based methods (left) aim to encode a sequence of observations into a single latent embedding and learn to generate future motions by combining samples from this single latent space with past observations. Our method HiT-DVAE (right) aims to learn a sequential generative model for the joint distribution of data and latent variables. The generation of human motion is conducted using the alternate generations of $\mathbf{z}_t$ and $\mathbf{x}_t$.

and generate diverse future observations from different soft assignments of the base vector where the coefficients of the Gaussian distributions are obtained from a Gumbel-Softmax Sampling (Jang et al., 2017; Maddison et al., 2017).

One common denominator of the above methods is that they all encode the whole sequence of observed human poses into a single time-independent embedding, thus modeling the stochasticity in the generative process at the whole sequence level. However, this is inconsistent with the inherent time-dependent *uncertainty* of human motion. This motivates us to propose HiT-DVAE based on dynamical variational autoencoders (DVAEs) (Girin et al., 2021) which could inherently generate the diversity of sequential outputs. By using a DVAE, the sequence of observations is encoded into a sequence of latent vectors, instead of only using a single latent vector as done in the aforementioned approaches, see Figure 1. This offers larger flexibility to learn and exploit the motion dynamics (Girin et al., 2021). Specifically, we model the generative process with auto-regressive dependencies, in a way that the generation of each frame depends on the previous information, including the previous and the present latent representations and also the previous poses. These auto-regressive dependencies are implemented with a Transformer-based architecture(Vaswani et al., 2017).

Evaluating the quality of the generated data is an open problem in computer vision. For 3D human motion generation, previous works either evaluate the generated data directly on the joint coordinates of the poses (Yuan & Kitani, 2020; Mao et al., 2021) or evaluate on the feature spaces based on a pre-trained feature extractor (Petrovich et al., 2021; Guo et al., 2020). However, both protocols have clear shortcomings: the former just evaluates the best-generated sample and the diversity of all generated data, ignoring the quality of most generated results except for the best one; while the latter depends on the quality of the pre-trained feature extractor. To fully evaluate the data generation quality, we use both evaluation methods and broaden the former by evaluating not only the best-generated sample but also the median result, to consider the overall quality of the generation results and take performance stability into consideration.

We thoroughly test HiT-DVAE on HumanEva-I and Human3.6M datasets, using both explicit (coordinate-based) and implicit metrics (feature-based) to measure the quality of the generated data. Experimental results show that HiT-DVAE consistently achieves state-of-the-art performance on most of the metrics for both datasets, leading to high-quality 3D human pose sequence generation (smaller errors, better features, correct action).

In summary, our contributions are threefold:

- We propose to use a latent-variable model with a *sequence of latent vectors* associated with a sequence of data vectors instead of just using a single latent vector, to better model the inherent time-dependent uncertainty of human motion.
- We design the first combination of a Transformer with a DVAE model for 3D human motion generation and show that this model is able to generate high-quality 3D human motion data by experiments on widely used datasets.
- We discuss the pros and cons of current evaluation protocols for motion generation and broaden the evaluation metrics to better evaluate the quality of generated data.

2 Related Work

Deterministic human motion prediction. Deterministic human motion prediction aims at regressing a single future motion from the past observation which is the most likely to the ground truth. Due to the inherent sequential structure of human motion, 3D human motion prediction has been mostly addressed with RNNs (Fragkiadaki et al., 2015; Jain et al., 2016; Martinez et al., 2017). Meanwhile, recent approaches demonstrate the effectiveness of modeling the spatial connectivity of human joints with GCNs (Mao et al., 2019; Dang et al., 2021; Ma et al., 2022), self-attention (Transformers) (Aksan et al., 2021; Cai et al., 2020) or simply with multi-layer perceptrons (Bouazizi et al., 2022; Guo et al., 2022b). While deterministic methods have achieved promising results, they exhibit strong limitations when it comes to modeling the diversity of plausible human motion forecasts.

Stochastic human motion generation. Different from human motion prediction, the task of human motion generation focuses on generating various possibilities of the future to model the uncertainty of motion. While some works alleviate this problem by leveraging the corresponding action labels (Guo et al., 2020; Petrovich et al., 2021) or text information (Tevet et al., 2022; Petrovich et al., 2022), in this work we focus on generating future motion merely from past observations. To solve this problem, two types of approaches have been studied in the recent past: (i) the enhancement of deterministic methods with stochastic variations, *e.g.*, incorporating noise, and (ii) leveraging conditional variational architectures that learn a probability distribution (see Fig. 1 left). In the first category, early works include combining random noise with hidden states either by concatenation (Lin & Amer, 2018; Kundu et al., 2019) or addition (Barsoum et al., 2018). More recently, Mao et al. (2021) further investigated this paradigm with a GCN-based motion prediction model and showed promising results with dedicated losses. In the second category, past observations are encoded to learn a latent space, then a random variable is sampled and combined with observations to predict the future (Walker et al., 2017; Yan et al., 2018; Aliakbarian et al., 2020; Cai et al., 2021; Aliakbarian et al., 2021). Recently, DLow (Yuan & Kitani, 2020) proposed to explicitly generate a large number of samples during training, then use an energy function to promote the diverse generation. This approach was later ameliorated by Dang et al. (2022) with a Gumbel-Softmax sampling from an auxiliary space to get rid of learning massive encoders in DLow. Rather than modeling the whole observation into a single embedding, HuMor (Rempe et al., 2021) exploits an auto-regressive generative model where the current generation will depend on the past prediction. However, HuMor has to use the SMPL (Loper et al., 2015) parameters which are not eligible for many of the motion datasets.

3 Method

The problem of 3D human motion generation can be formalized as follows. Given a sequence of O observed pose vectors $\mathbf{x}_{1:O} = [\mathbf{x}_1, \ldots, \mathbf{x}_O]$, we aim at generating a sequence of the following G frames $\mathbf{x}_{O+1:O+G} = [\mathbf{x}_{O+1}, \ldots \mathbf{x}_{O+G}]$. Each pose vector $\mathbf{x}_t \in \mathbb{R}^{J\times 3}$ represents the 3D Cartesian coordinates of the J joints of a person at time t. For convenience, we use $\mathbf{x}_{1:T}$ ($T = O + G$) to represent the entire sequence in the following. Different from deterministic human motion prediction, we intend to generate multiple plausible future motion sequences. To this end, we propose a new model named Hierarchical Transformer Dynamical Variational AutoEncoder (HiT-DVAE) based on DVAE to deal with the uncertainty of human motion.

3.1 Background: The DVAE family

The proposed HiT-DVAE model is based on the family of dynamical variational autoencoders (DVAEs) (Girin et al., 2021). In a DVAE, the generation of a time series of observed data vector $\mathbf{x}_{1:T}$ involves an associated time series of latent vectors $\mathbf{z}_{1:T}$. This is opposed to a "static" VAE which encodes $\mathbf{x}_{1:T}$ into a single latent variable $\mathbf{z}$. In the most general DVAE formulation, the time dependencies within and across $\mathbf{x}_{1:T}$ and $\mathbf{z}_{1:T}$ are of the autoregressive type and are implemented with RNNs.

In addition to the time-dependent latent variables $\mathbf{z}_{1:T}$, we further add a time-independent latent variable $\mathbf{w}$ to grasp the global context of the whole sequence, inspired by Petrovich et al. (2021) and Li & Mandt (2018). While different from Petrovich et al. (2021), our $\mathbf{w}$ is learned in an unsupervised

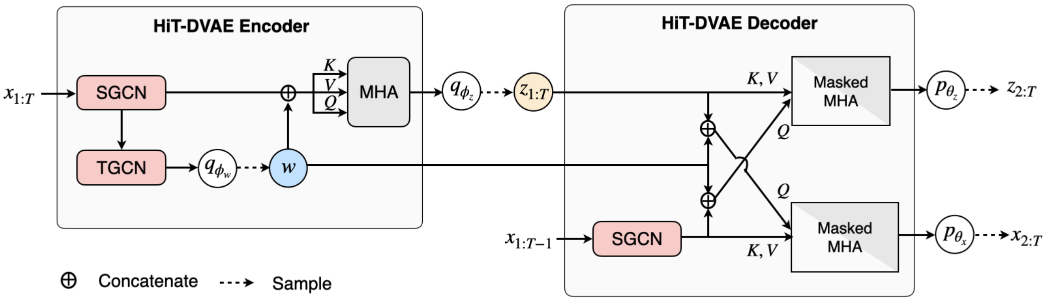


Figure 2: Overview of HiT-DVAE. The Encoder (left) inputs the observed sequence $\mathbf{x}_{1:T}$ to estimate the posterior distribution of the time-dependent latent variables $\mathbf{z}_{1:T}$ and time-independent latent variable $\mathbf{w}$. Then the Decoder (right) reconstructs the motion and the priors of $\mathbf{z}$.

manner within the DVAE methodology, thus not requiring action class labels (Li & Mandt, 2018). The general form of the generative model that we use in this work writes:

$$p_{\boldsymbol{\theta}}(\mathbf{x}_{1:T}, \mathbf{z}_{1:T}, \mathbf{w}) = p_{\boldsymbol{\theta}_{\mathbf{w}}}(\mathbf{w}) \prod_{t=1}^{T} p_{\boldsymbol{\theta}_{\mathbf{x}}}(\mathbf{x}_t | \mathbf{x}_{1:t-1}, \mathbf{z}_{1:t}, \mathbf{w}) p_{\boldsymbol{\theta}_{\mathbf{z}}}(\mathbf{z}_t | \mathbf{x}_{1:t-1}, \mathbf{z}_{1:t-1}, \mathbf{w}), \tag{1}$$

where $p_{\boldsymbol{\theta}_{\mathbf{w}}}(\mathbf{w})$ is a standard normal distribution. $p_{\boldsymbol{\theta}_{\mathbf{x}}}(\mathbf{x}_t | \mathbf{x}_{1:t-1}, \mathbf{z}_{1:t}, \mathbf{w}) = \mathcal{N}(\mathbf{x}_t; \boldsymbol{\mu}_{\boldsymbol{\theta}_{\mathbf{x}},t}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}_{\mathbf{x}},t})$ and $p_{\boldsymbol{\theta}_{\mathbf{z}}}(\mathbf{z}_t | \mathbf{x}_{1:t-1}, \mathbf{z}_{1:t-1}, \mathbf{w}) = \mathcal{N}(\mathbf{z}_t; \boldsymbol{\mu}_{\boldsymbol{\theta}_{\mathbf{z}},t}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}_{\mathbf{z}},t})$, where $\mathcal{N}(\cdot;\ \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the Gaussian distribution of mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. $\boldsymbol{\theta}_{\mathbf{z}}$ and $\boldsymbol{\theta}_{\mathbf{x}}$ denote the parameters of the neural network which generates the parameters of the corresponding density functions, and $\boldsymbol{\theta} = \boldsymbol{\theta}_{\mathbf{z}} \cup \boldsymbol{\theta}_{\mathbf{x}}$.

In order to learn this generative model, an inference model with parameters $\boldsymbol{\phi} = \boldsymbol{\phi}_{\mathbf{w}} \cup \boldsymbol{\phi}_{\mathbf{z}}$ is introduced. In the present work, we consider the following inference model:

$$q_{\boldsymbol{\phi}}(\mathbf{z}_{1:T}, \mathbf{w} | \mathbf{x}_{1:T}) = q_{\boldsymbol{\phi}_{\mathbf{w}}}(\mathbf{w} | \mathbf{x}_{1:T}) \prod_{t=1}^{T} q_{\boldsymbol{\phi}_{\mathbf{z}}}(\mathbf{z}_t | \mathbf{x}_{1:T}, \mathbf{w}), \tag{2}$$

where $q_{\boldsymbol{\phi}_{\mathbf{w}}}(\mathbf{w} | \mathbf{x}_{1:T}) = \mathcal{N}(\mathbf{w}; \boldsymbol{\mu}_{\boldsymbol{\phi}_{\mathbf{w}}}, \boldsymbol{\Sigma}_{\boldsymbol{\phi}_{\mathbf{w}}})$ and $q_{\boldsymbol{\phi}_{\mathbf{z}}}(\mathbf{z}_t | \mathbf{x}_{1:T}, \mathbf{w}) = \mathcal{N}(\mathbf{z}_t; \boldsymbol{\mu}_{\boldsymbol{\phi}_{\mathbf{z}},t}, \boldsymbol{\Sigma}_{\boldsymbol{\phi}_{\mathbf{z}},t})$.

The training objective is to maximize the Evidence Lower Bound (ELBO):

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathbf{x}_{1:T}) = \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}_{1:T}, \mathbf{w} | \mathbf{x}_{1:T})} \left[\ln p_{\boldsymbol{\theta}}(\mathbf{x}_{1:T}, \mathbf{z}_{1:T}, \mathbf{w}) - \ln q_{\boldsymbol{\phi}}(\mathbf{z}_{1:T}, \mathbf{w} | \mathbf{x}_{1:T}) \right]. \tag{3}$$

3.2 HiT-DVAE

Although the above equations define the probabilistic dependencies between the different random variables, there are plenty of ways to implement these dependencies. In previous literature, this is always realized by RNNs (Girin et al., 2021). In this paper, we propose a hierarchical DVAE architecture based on the Transformer of Pavlakos et al. (2017). The pipeline of HiT-DVAE is shown in Figure 2, and the pseudo-code for training and generation could be found in Appendix D.

3.2.1 Generative Model (HiT-DVAE Decoder)

The generation of both $\mathbf{x}_{1:T}$ and $\mathbf{z}_{1:T}$ is performed via the multi-head attention (MHA) mechanisms of the original Transformer architecture (Vaswani et al., 2017). The outputs of the two cross MHA modules are the parameters of the respective probability distributions defined in Equation 1. Note that in the present implementation, we set $\boldsymbol{\Sigma}_{\boldsymbol{\theta}_{\mathbf{x}},t} = \boldsymbol{I}$ (the identity matrix), as in previous works (Aliakbarian et al., 2021; Rempe et al., 2021; Guo et al., 2020; Petrovich et al., 2021; Yuan & Kitani, 2020). The covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\theta}_{\mathbf{z}},t}$ is a diagonal matrix (the diagonal entries being in the vector $\boldsymbol{v}_{\boldsymbol{\theta}_{\mathbf{z}},t}$). The output of the first MHA module is thus $\{\boldsymbol{\mu}_{\boldsymbol{\theta}_{\mathbf{z}},t}, \boldsymbol{v}_{\boldsymbol{\theta}_{\mathbf{z}},t}\}$, and the output of the second MHA module is $\boldsymbol{\mu}_{\boldsymbol{\theta}_{\mathbf{x}},t}$. As for the inputs of the MHA modules, it is noted that the past observations are first processed with a Spatial Graph Convolutional Network (SGCN) that extracts pose features from the raw poses $\mathbf{x}_t$. We denote this pose feature extraction operation as f_{D} and we detail it below. Then, the generative processes of $\mathbf{x}_t$ and $\mathbf{z}_t$ differ on what variables are used as queries, keys, and values in the attention mechanism. In the present work, the decoder is given by:

$$\boldsymbol{\mu}_{\boldsymbol{\theta}_{\mathbf{x}},t} = \text{MHA}\left(Q_{\boldsymbol{\theta}_{\mathbf{x}},t}, K_{\boldsymbol{\theta}_{\mathbf{x}}}, V_{\boldsymbol{\theta}_{\mathbf{x}}}\right) \tag{4}$$

$$Q_{\boldsymbol{\theta}_{\mathbf{x}},t} = \begin{bmatrix} \mathbf{z}_t \\ \mathbf{w} \end{bmatrix}, K_{\boldsymbol{\theta}_{\mathbf{x}}} = V_{\boldsymbol{\theta}_{\mathbf{x}}} = [\, f_{\text{D}}(\mathbf{x}_1), \ \ldots, \ f_{\text{D}}(\mathbf{x}_T)\,], \tag{5}$$

$$\begin{bmatrix} \boldsymbol{\mu}_{\boldsymbol{\theta}_{\mathbf{z}},t} \\ \boldsymbol{v}_{\boldsymbol{\theta}_{\mathbf{z}},t} \end{bmatrix} = \text{MHA}\left(Q_{\boldsymbol{\theta}_{\mathbf{z}},t}, K_{\boldsymbol{\theta}_{\mathbf{z}}}, V_{\boldsymbol{\theta}_{\mathbf{z}}}\right), \tag{6}$$

$$Q_{\boldsymbol{\theta}_{\mathbf{z}},t} = \begin{bmatrix} f_{\text{D}}(\mathbf{x}_{t-1}) \\ \mathbf{w} \end{bmatrix}, K_{\boldsymbol{\theta}_{\mathbf{z}}} = V_{\boldsymbol{\theta}_{\mathbf{z}}} = [\, \mathbf{z}_1, \ \ldots, \ \mathbf{z}_T \,], \tag{7}$$

where a mask is used to prevent $\mathbf{z}_t$ and $\mathbf{x}_t$ from being generated from future latent and observed variables. More discussions could be found in Appendix A.

In the conventional Transformer architecture, the past observations (*e.g.* $\mathbf{x}_{1:t-1}$) are used as queries (Q) for the generation of the current state (*e.g.* $\mathbf{x}_t$). However, in our implementation, we use the past observations as keys (K) and values (V). This is motivated by the fact that this led to notably better results than with the "conventional" variable ordering. An intuitive explanation is that directly exploiting $\mathbf{x}_{1:t-1}$ to predict $\mathbf{x}_t$ leads to poor information in $\mathbf{z}_{1:T}$ and a poor decoder capacity.

3.2.2 Inference Model (HiT-DVAE Encoder)

The inference of the latent variables $\mathbf{w}$ and $\mathbf{z}_{1:T}$ from $\mathbf{x}_{1:T}$ at the HiT-DVAE encoder is performed via a multi-head self-attention mechanism (MHSA). Pose features are first extracted with an SGCN similar to the one of the decoder and denoted here f_{E}. The series of pose features are then fed into a temporal GCN with T nodes, where each node indicates a time frame, and then into a fully connected (FC) layer to output the posterior mean and covariance matrix of $\mathbf{w}$. A sample of $\mathbf{w}$ is drawn from the corresponding posterior, concatenated to the pose features extracted from $\mathbf{x}_{1:T}$, and then fed into the Transformer encoder such that:

$$\begin{bmatrix} \boldsymbol{\mu}_{\boldsymbol{\phi}_{\mathbf{z}},t} \\ \boldsymbol{v}_{\boldsymbol{\phi}_{\mathbf{z}},t} \end{bmatrix} = \text{MHSA}\left(Q_{\boldsymbol{\phi}_{\mathbf{z}},t}, K_{\boldsymbol{\phi}_{\mathbf{z}}}, V_{\boldsymbol{\phi}_{\mathbf{z}}}\right), \tag{8}$$

$$Q_{\boldsymbol{\phi}_{\mathbf{z}},t} = \begin{bmatrix} f_{\text{E}}(\mathbf{x}_t) \\ \mathbf{w}, \end{bmatrix}, \tag{9}$$

$$K_{\boldsymbol{\phi}_{\mathbf{z}}} = V_{\boldsymbol{\phi}_{\mathbf{z}}} = \begin{bmatrix} f_{\text{E}}(\mathbf{x}_1), & \ldots, & f_{\text{E}}(\mathbf{x}_T) \\ \mathbf{w}, & \ldots, & \mathbf{w} \end{bmatrix}, \tag{10}$$

where each of the output at index t indicates the distribution parameters of the latent variable $\mathbf{z}_t$, namely $\boldsymbol{\mu}_{\boldsymbol{\phi}_{\mathbf{z}}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\phi}_{\mathbf{z}}}$ (with diagonal entries $\boldsymbol{v}_{\boldsymbol{\phi}_{\mathbf{z}},t}$).

3.3 Training and Testing

In the case of the proposed HiT-DVAE, optimizing ELBO (see Equation 3) boils down to (i) minimizing the L_2 loss on the reconstructed poses while (ii) minimizing the KL divergence between the posterior and prior distributions over the latent variables ($\mathcal{L}_{\text{ELBO}} = -\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathbf{x}_{1:T})$). Following Yuan & Kitani (2020) and Mao et al. (2021), we explicitly generate K motion sequences and compute the reconstruction loss with respect to the ground-truth as well as the pseudo-multi-modal ground-truth. And we also promote the diverse generation ($\mathcal{L}_{\text{DIV}}$) and penalize unrealistic samples ($\mathcal{L}_{\text{REAL}}$). Details and more explanations could be found in Appendix C. Altogether, our final training loss writes:

$$\mathcal{L} = \mathcal{L}_{\text{ELBO}} + \mathcal{L}_{\text{DIV}} + \mathcal{L}_{\text{REAL}}. \tag{11}$$

The above losses allow training the proposed HiT-DVAE model to reconstruct full sequences $\mathbf{x}_{1:T}$. At training time, the model is trained from ground-truth (GT) input sequences. At test time, between O+1 and O+G, the previously generated data vectors are used at input, but not the GT. There is thus a mismatch between the training and testing condition, which degrades the quality of the generated sequences. To avoid this mismatch, we use scheduled sampling during training (Bengio et al., 2015): we progressively introduce the past generated vectors at the input during the training iterations which improves the quality of generated data, see Section 4.4 for more discussion.

Once our model is trained, we could use it to generate various future motion sequences of arbitrary length. Given O observations $\mathbf{x}_{1:O}$, we sample $\mathbf{z}_{1:O}$ and $\mathbf{w}$ from the inference model. Then, we generate the next G frames $\hat{\mathbf{x}}_{O+1:O+G}$ by recursively sampling from the generative distributions $p_{\boldsymbol{\theta}_{\mathbf{x}}}(\mathbf{x}_t|\mathbf{x}_{1:t-1}, \mathbf{z}_t, \mathbf{w})$ and $p_{\boldsymbol{\theta}_{\mathbf{z}}}(\mathbf{z}_t|\mathbf{x}_{t-1}, \mathbf{z}_{1:t-1}, \mathbf{w})$. The diversity of generated data comes from the different samples of $\mathbf{z}_{O+1:O+G}$ and $\mathbf{w}$.

4 Experiments

4.1 Datasets

Following Mao et al. (2021) and Yuan & Kitani (2020), we train and evaluate the proposed method on the Human3.6M (Ionescu et al., 2013) and HumanEva-I (Sigal & Black, 2006) datasets, and remove global translations of the poses for both datasets.

Human3.6M is the most commonly used dataset for human motion-related tasks. It contains 7 actors (S1,5,6,7,8,9,11) performing 15 annotated actions recorded at 50 Hz. The human pose is represented by 32 joints, while we follow Mao et al. (2021) and only use 17 of the joints in the training and all testing implementations of HiT-DVAE. We use S1,5,6,7,8 as the training set and the other two subjects as the test set. To be comparable with previous methods(Mao et al., 2021; Yuan & Kitani, 2020), we observe the first 25 frames and predict the next 100 frames ($O = 25$, $G = 100$).

HumanEva-I contains 5 actions (Box, Gesture, Jog, ThrowCatch, Walking) performed by 3 actors, recorded at 60 Hz. Each pose is represented by 15 joints. Following previous literature(Mao et al., 2021; Yuan & Kitani, 2020), We observe the first 15 frames and predict the next 60 frames ($O = 15$, $G = 60$).

4.2 Evaluation Protocols

Explicit evaluation metrics. Following Mao et al. (2021) and Yuan & Kitani (2020), we evaluate the error and diversity of our results with the following metrics, calculating directly on the joint locations of poses: (i) Average Pairwise Distance (APD): average $L2$ distance between all pairs of generated sequences: $\frac{1}{K(K-1)}\sum_{i=1}^{K}\sum_{j=1,j\neq i}^{K}\|\hat{\mathbf{x}}^{i}_{O+1:O+G} - \hat{\mathbf{x}}^{j}_{O+1:O+G}\|_2$, where K is the total number of generated sequences. APD measures the capacity of the model to generate diverse samples without considering their quality. (ii) Average Displacement Error (ADE): $L2$ distance between the ground truth and the *best* generated sample with minimum error among all, averaged over all frames of the sequence:$\frac{1}{G}\min_k \|\hat{\mathbf{x}}^{k}_{O+1:O+G} - \mathbf{x}_{O+1:O+G}\|_2$. ADE evaluates the upper bound of the generation quality of a model but not its usual performance. (iii) Final Displacement Error (FDE): Similar to ADE, FDE evaluates the distance between the ground truth and the best sample, but just on the final frame instead of the whole sequence: $\min_k \|\hat{\mathbf{x}}^{k}_{O+G} - \mathbf{x}_{O+G}\|$. (iv) Multi-Modal ADE (MMADE) and Multi-Modal FDE (MMFDE): ADE and FDE on pseudo-multi-modal ground-truth.

These coordinate-based metrics are widely used for evaluating the quality of generated motion (Ionescu et al., 2013), in terms of explicit diversity on Cartesian coordinates and the upper bound of the generated samples (*i.e.* the best sample). However, when merely using these metrics, we could not avoid the corner case where the model tends to generate a few samples with high quality and the others which totally fail. In this case the model could still achieve high scores both in APD and ADE/FDE, whereas it lost our original intention to generate plausible and diverse future sequences.

To this end, we propose two solutions: (1) instead of just evaluating (ii-iv) on the *best* sample, we also evaluate these criteria on the *median* example, *i.e.* the generated motion with the median error instead of minimum error among all generated examples; (2) in addition to these explicit measurements based on pose coordinate, we also consider implicit measurements based on a pre-trained action classifier to evaluate on feature space, as described below.

Implicit evaluation metrics. Following Petrovich et al. (2021) and Guo et al. (2020), we use a GRU-based action classifier pre-trained on real data to evaluate the quality of generated data by: (i) calculating Recognition Accuracy(Acc) of the classifier on generated data to evaluate if they could be recognized as the correct action class; (ii) extracting features of the generated data and real data respectively by the same action classifier, and calculating the Frechet Inception Distance (FID) of

Table 1: **Results on HumanEva-I.** "Real data" means real motion in the testing set, showing the theoretical upper bounds of accuracy (Acc) on generation methods. The suffix "b" or "m" represents the best/median metrics. ↑ (↓) means higher (lower) is better. † indicates results taken from DLow, ⋆ indicates results obtained by using the official code repository.

	Acc	FID	APD	ADEb	FDEb	MMADEb	MMFDEb	ADEm	FDEm	MMADEm	MMFDEm
	(%) ↑	↓	(m) ↑	(m) ↓	(m) ↓	(m) ↓	(m) ↓	(m) ↓	(m) ↓	(m) ↓	(m) ↓
Real data	88.3	-	-	-	-	-	-	-	-	-	-
ERD† Fragkiadaki et al. (2015)	-	-	0	0.382	0.461	0.521	0.595	-	-	-	-
acLSTM† Li et al. (2017)	-	-	0	0.429	0.541	0.530	0.608	-	-	-	-
Pose-Knows† Walker et al. (2017)	-	-	2.308	0.269	0.296	0.384	0.375	-	-	-	-
MT-VAE† Yan et al. (2018)	-	-	0.021	0.345	0.403	0.518	0.577	-	-	-	-
HP-GAN† Barsoum et al. (2018)	-	-	1.139	0.772	0.749	0.776	0.769	-	-	-	-
BoM† Bhattacharyya et al. (2018)	-	-	2.846	0.271	0.279	0.373	0.351	-	-	-	-
GMVAE† Dilokthanakul et al. (2016)	-	-	2.443	0.305	0.345	0.408	0.410	-	-	-	-
DeLiGAN† Gurumurthy et al. (2017)	-	-	2.177	0.306	0.322	0.385	0.371	-	-	-	-
DSF† Yuan & Kitani (2019)	-	-	4.538	0.273	0.290	0.364	0.340	-	-	-	-
DLow⋆ Yuan & Kitani (2020)	52.7	3.472	4.853	0.248	0.262	0.361	0.337	0.577	0.717	0.646	0.753
GSPS⋆ Mao et al. (2021)	51.6	1.604	5.825	0.233	0.244	0.343	0.331	0.686	0.794	0.735	0.825
GumbelSample⋆ Dang et al. (2022)	62.4	1.457	**6.109**	**0.220**	**0.234**	0.342	0.316	0.698	0.758	0.744	0.783
HiT-DVAE	**72.6**	**0.089**	4.721	0.282	0.261	**0.335**	**0.290**	**0.579**	**0.665**	**0.610**	**0.683**

Table 2: **Results on Human3.6M.** "Real data" means real motion in the testing set, showing the theoretical upper bounds of accuracy (Acc) on generation methods. ↑ (↓) means higher (lower) is better. † indicates results taken from DLow, and ⋆ indicates results obtained by using the official code repository.

	Acc	FID	APD	ADE	FDE	MMADE	MMFDE	ADE	FDE	MMADE	MMFDE
	(%) ↑	↓	(m) ↑	Best sample (m) ↓				Median sample (m) ↓			
Real data	85.5	-	-	-	-	-	-	-	-	-	-
ERD† Fragkiadaki et al. (2015)	-	-	0	0.722	0.969	0.776	0.995	-	-	-	-
acLSTM† Li et al. (2017)	-	-	0	0.789	1.126	0.849	1.139	-	-	-	-
Pose-Knows† Walker et al. (2017)	-	-	6.723	0.461	0.560	0.522	0.569	-	-	-	-
MT-VAE† Yan et al. (2018)	-	-	0.403	0.457	0.595	0.716	0.883	-	-	-	-
HP-GAN† Barsoum et al. (2018)	-	-	7.214	0.858	0.867	0.847	0.858	-	-	-	-
BoM† Bhattacharyya et al. (2018)	-	-	6.265	0.448	0.533	0.514	0.544	-	-	-	-
GMVAE† Dilokthanakul et al. (2016)	-	-	6.769	0.461	0.555	0.524	0.566	-	-	-	-
DeLiGAN† Gurumurthy et al. (2017)	-	-	6.509	0.483	0.534	0.520	0.545	-	-	-	-
DSF† Yuan & Kitani (2019)	-	-	9.330	0.493	0.592	0.550	0.599	-	-	-	-
DLow⋆ Yuan & Kitani (2020)	65.9	**1.412**	11.741	0.425	0.518	0.495	0.531	0.896	1.284	0.948	1.289
GSPS⋆ Mao et al. (2021)	65.0	2.030	14.757	0.389	0.496	0.476	0.525	1.013	1.372	1.065	1.381
GumbelSample⋆ Dang et al. (2022)	66.5	2.659	**15.310**	**0.370**	**0.485**	**0.475**	0.516	0.922	1.345	1.005	1.361
HiT-DVAE	**70.0**	1.708	8.942	0.472	0.505	0.497	**0.514**	**0.804**	**1.034**	**0.812**	**1.028**

these two distributions to evaluate the overall quality of the generated data. We train a classifier for each of the datasets separately on their training splits.

4.3 QUANTITATIVE RESULTS

We evaluate HiT-DVAE on HumanEva-I and Human3.6m dataset using the explicit and implicit metrics described in Sec 4.2 and observe that the proposed method outperforms the state-of-the-art methods on most of the evaluation metrics. Note that considering the large number of metrics used (some of them are even incompatible), we aim to find a balance of all the metrics to get overall good performance.

HumanEva-I As shown in Table 1, HiT-DVAE achieves significantly better results than other state-of-the-art methods on ACC and FID. This means that the feature distributions of the motions generated by HiT-DVAE are more similar to the corresponding ground-truth, and the action patterns of these generated motions could be better recognized. Besides, we observe that HiT-DVAE achieves comparable results with state-of-the-art on explicit evaluation of diversity (APD) and errors of the *best* sample (ADE$_b$, FDE$_b$, MMADE$_b$, MMFDE$_b$). As discussed in Sec 4.2, just considering errors of the *best* sample along with the diversity is not reliable, thus it is important to consider the error of *median* samples. We could find that HiT-DVAE is also better on *median* errors (ADE$_m$, FDE$_m$, MMADE$_m$, MMFDE$_m$), indicating better overall generation quality. Note that APD is not always

Table 3: **Ablation study on different architecture designs** on HumanEva-I and Human3.6M. "w/o SS" means without scheduled sampling, "w/o Att." means using an LSTM instead of transformer, "w/o $\mathbf{w}$" means without using the time-independent latent variable $\mathbf{w}$. "w/o Att. & $\mathbf{w}$ " means without attention and no use of $\mathbf{w}$.

Architecture	ACC (%) ↑	FID ↓	APD (m) ↑	ADEb (m) ↓	FDEb (m) ↓	MMADEb (m) ↓	MMFDEb (m) ↓	ADEm (m) ↓	FDEm (m) ↓	MMADEm (m) ↓	MMFDEm (m) ↓
HumanEva-I											
HiT-DVAE	72.6	**0.089**	4.721	0.282	0.261	**0.335**	**0.290**	0.579	0.665	0.610	0.683
w/o SS	69.6	0.359	**4.777**	0.314	0.300	0.358	0.315	0.596	0.708	0.624	0.727
w/o Att.	72.9	0.264	3.921	**0.265**	**0.243**	0.348	0.295	**0.510**	**0.604**	**0.569**	**0.650**
w/o $\mathbf{w}$	74.0	0.306	4.244	0.287	0.263	0.360	0.303	0.535	0.662	0.591	0.689
w/o Att. & $\mathbf{w}$	**76.5**	1.262	0.023	0.538	0.621	0.594	0.660	0.538	0.622	0.595	0.662
Human3.6M											
HiT-DVAE	70.0	1.708	8.942	**0.472**	**0.505**	**0.497**	**0.514**	0.804	1.034	0.812	1.028
w/o SS	65.0	1.751	**10.339**	0.477	0.513	0.501	0.519	0.867	1.108	0.877	1.101
w/o Att.	70.5	1.475	7.189	0.537	0.569	0.555	0.567	0.805	1.060	0.826	1.056
w/o $\mathbf{w}$	**71.1**	1.565	6.249	0.528	0.582	0.557	0.590	**0.735**	**0.942**	**0.753**	**0.942**
w/o Att. & $\mathbf{w}$	64.2	**0.659**	2.688	0.752	0.892	0.787	0.896	0.897	1.184	0.934	1.186

Table 4: **Ablation study on different loss terms** on HumanEva-I and Human3.6M. The suffix "b" or "m" represents the best/median metrics. Using all three losses results in the best performance.

$\mathcal{L}_{\text{ELBO}}$	$\mathcal{L}_{\text{DIV}}$	$\mathcal{L}_{\text{REAL}}$	ACC (%) ↑	FID ↓	APD (m) ↑	ADEb (m) ↓	FDEb (m) ↓	MMADEb (m) ↓	MMFDEb (m) ↓	ADEm (m) ↓	FDEm (m) ↓	MMADEm (m) ↓	MMFDEm (m) ↓
HumanEva-I													
✓	✓	✓	**72.8**	**0.080**	4.823	**0.280**	**0.251**	**0.333**	**0.282**	**0.583**	0.690	**0.615**	0.705
✓	✓		62.7	0.621	**137.207**	0.306	0.279	0.356	0.315	0.636	**0.677**	0.669	0.698
✓			70.5	0.271	4.848	0.287	0.258	0.335	0.289	0.591	0.687	0.616	**0.693**
Human3.6M													
✓	✓	✓	**70.0**	1.708	8.942	0.472	0.505	0.497	0.514	**0.804**	**1.034**	**0.812**	**1.028**
✓	✓		68.7	1.778	**9.604**	**0.460**	**0.495**	**0.485**	**0.501**	0.865	1.107	0.875	1.102
✓			69.4	**1.690**	9.398	0.461	0.507	0.487	0.507	0.838	1.098	0.849	1.092

better for larger values, because extremely large diversity usually comes with large joint errors for most generation results ("median" metrics), and also low recognition accuracy. This represents that some of the generated samples might totally fail and the quality of generation is not guaranteed.

Human3.6M Similar conclusions can be drawn from the results on the Human3.6M dataset, as shown in Table 2. When training the action classifier for Human3.6M dataset, we group the 15 actions into 5 groups instead of training on all the 15 action labels (see details in Appendix E). This is because Human3.6M dataset is not designed for action classification and some actions in this dataset are quite alike. For example, we could not see many differences between "eating" and "smoking" by looking at the skeletons of a person. The re-grouping is based on the confusion matrix of an initial classifier trained on real data with all class labels. After re-grouping, the average classification accuracy on real data increases from 48.1% to 85.5%. We report results using this 5-group classifier because we believe that a better classifier is more reliable for calculating accuracy and extracting features for FID. Note that even on the 15-action classifier with low real-data accuracy, HiT-DVAE still performs better than other state-of-the-art methods, see details in Appendix E.

4.4 Ablation Study

Table 3 shows ablation studies on HiT-DVAE with different architecture designs. We put the best results in bold and underline the second. We could find that without schedule sampling ("w/o SS"), HiT-DVAE tends to generate more diverse results but with worse quality either on explicit metrics or implicit metrics. Compared with using LSTM ("w/o Att."), we found that the use of the attention mechanism brings higher overall generation quality with better performance on explicit metrics and multi-modal metrics. Besides, the global time-independent variable $\mathbf{w}$ brings more diversity for both datasets ("w/o $\mathbf{w}$"). When we consider a vanilla DVAE model ("w/o Att. & $\mathbf{w}$") without the hierarchical transformer (HIT) architecture, it is very likely to collapse to a static state on the sequential latent space, which leads to moderate generation quality, and much worse diversity (APD =

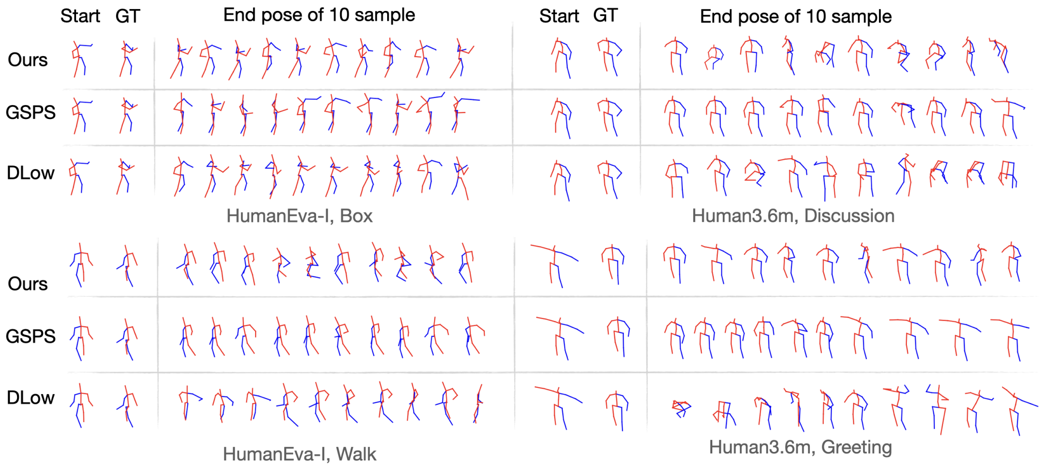


Figure 3: **Qualitative visualization** on four different actions of the HumanEva and Human3.6M datasets. “Start”/“GT” means the last observed frame / the ground truth last frame respectively. The subsequent frames correspond to the last frames of 10 different generated sequences. Note that GSPS (Mao et al., 2021) and DLow (Yuan & Kitani, 2020) highly diverge from that last GT frame or tend to have generations with less diversity. While our approach generates different alternatives and keeps the essence of the particular action.

0.023 for HumanEva-I and APD = 2.688 for Human3.6M). The final setting of HiT-DVAE performs well on almost all the metrics and balances different evaluations.

Moreover, to further understand the impact of each loss term on the performance, we conduct an ablation study on the three loss terms in Table 4. We observe that $\mathcal{L}_{\text{DIV}}$ increases the diversity and facilitates the model to generate different motion sequences, and $\mathcal{L}_{\text{REAL}}$ can help compensating for excessive diversity which results in non-realistic and non-feasible poses (*i.e.* APD = 137.207 in HumanEva-I).

4.5 Qualitative results

To qualitatively evaluate our generated results, we visualize various generating samples of our methods in Figure 3 compared with other state-of-the-art methods. we can see that other methods either generate very similar samples for all the generations or result in some weird motions, while our method performs well on all the generations with diverse but reasonable results. More visualizations in video with a simple user study could be found in Appendix F.

5 Conclusions

In this paper, we investigate the temporal probabilistic model combined with attention mechanisms for human motion generation. In particular, we proposed a deep generative model, HiT-DVAE, modeling temporal dependencies between the observations and the latent variables. The use of our method enables the motion dynamics to be learned from a sequence of latent variables instead of a single latent variable, which gives richer representation information for human motion which has inherent time-dependent uncertainty and thus results in better performance. We evaluated our method on two widely used datasets, HumanEva and Human3.6M, and reported state-of-the-art results with various evaluation metrics.

Ethical Statement

The proposed research raises few direct ethical concerns since the contributions of the paper did not require the collection/sharing of new data or experiments on humans. Possible ethical concerns could arise, though, from the datasets used in our experiments, and more precisely from the potential biases, these datasets could induce. Regarding Human3.6M, the gender of actors was balanced, and a certain diversity of body mass index was also sought. However, there could still be biases *e.g.* related to cultural background, since we do not have any information about that. The information available on the HumanEva dataset is scarce, and it is difficult to understand the potential biases it can contain.

Reproducibility Statement

As mentioned in the abstract, the code and pretrained models will be publicly released upon acceptance. This includes the code for the data loaders, the model, the training, and the evaluation codes. For a primary test or proof of concept, please refer to Appendix B for the implementation details and Appendix D for the pseudo-code.

References

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

Sadegh Aliakbarian, Fatemeh Sadat Saleh, Mathieu Salzmann, Lars Petersson, and Stephen Gould. A stochastic conditioning scheme for diverse human motion prediction. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

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

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*, pp. 1418–1427, 2018.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. *Advances in Neural Information Processing Systems (NeurIPS)*, 2015.

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*, pp. 8485–8493, 2018.

Arij Bouazizi, Adrian Holzbock, Ulrich Kressel, Klaus Dietmayer, and Vasileios Belagiannis. Motionmixer: Mlp-based 3d human body pose forecasting. In *Proceedings of the AAAI International Joint Conference on Artificial Intelligence (IJCAI)*, 2022.

Paden Brian, Cap Michal, Yong Sze, Zheng, Yershov Dmitry, and Frazzoli Emilio. A survey of motion planning and control techniques for self-driving urban vehicles. In *IEEE Transactions on intelligent vehicles*, 2016.

Yujun Cai, Lin Huang, Yiwei Wang, Tat-Jen Cham, Jianfei Cai, Junsong Yuan, Jun Liu, Xu Yang, Yiheng Zhu, Xiaohui Shen, et al. Learning progressive joint propagation for human motion prediction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

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

Lingwei Dang, Yongwei Nie, Chengjiang Long, Qing Zhang, and Guiqing Li. Msr-gcn: Multi-scale residual graph convolution networks for human motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11467–11476, October 2021.

Lingwei Dang, Yongwei Nie, Chengjiang Long, Qing Zhang, and Guiqing Li. Diverse human motion prediction via gumbel-softmax sampling from an auxiliary space. In *Proceedings of the ACM International Conference on Multimedia (ACMMM)*, October 2022.

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.

Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2017.

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*, pp. 4346–4354, 2015.

Laurent Girin, Simon Leglaive, Xiaoyu Bie, Julien Diard, Thomas Hueber, and Xavier Alameda-Pineda. Dynamical variational autoencoders: A comprehensive review. *Foundations and Trends in Machine Learning*, 15(1-2):1–175, 2021.

Chuan Guo, Xinxin Zuo, Sen Wang, Shihao Zou, Qingyao Sun, Annan Deng, Minglun Gong, and Li Cheng. Action2motion: Conditioned generation of 3d human motions. In *Proceedings of the ACM International Conference on Multimedia (ACMMM)*, pp. 2021–2029, 2020.

Wen Guo, Xiaoyu Bie, Xavier Alameda-Pineda, and Francesc Moreno-Noguer. Multi-person extreme motion prediction. *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022a.

Wen Guo, Yuming Du, Xi Shen, Vincent Lepetit, Xavier Alameda-Pineda, and Francesc Moreno-Noguer. Back to mlp: A simple baseline for human motion prediction. *arXiv e-prints*, pp. arXiv–2207, 2022b.

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*, pp. 166–174, 2017.

Koppula Hema, Swetha and Saxena Ashutosh. Anticipating human activities for reactive robotic response. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2013.

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.

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, jul 2014.

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*, pp. 5308–5317, 2016.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2017.

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 (AAAI)*, 2019.

Yingzhen Li and Stephan Mandt. Disentangled sequential autoencoder. In *Proceedings of the International Conference on Machine Learning (ICML)*, Stockholm, Sweden, 2018.

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

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

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, October 2015.

Tiezheng Ma, Yongwei Nie, Chengjiang Long, Qing Zhang, and Guiqing Li. Progressively generating better initial guesses towards next stages for high-quality human motion prediction. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6437–6446, 2022.

Chris J Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2017.

Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5442–5451, 2019.

Wei Mao, Miaomiao Liu, Mathieu Salzmann, and Hongdong Li. Learning trajectory dependencies for human motion prediction. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 9489–9497, 2019.

Wei Mao, Miaomiao Liu, and Mathieu Salzmann. Generating smooth pose sequences for diverse human motion prediction. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 13309–13318, 2021.

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*, pp. 2891–2900, 2017.

Georgios Pavlakos, Xiaowei Zhou, Konstantinos G Derpanis, and Kostas Daniilidis. Coarse-to-fine volumetric prediction for single-image 3d human pose. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7025–7034, 2017.

Mathis Petrovich, Michael J Black, and Gül Varol. Action-conditioned 3d human motion synthesis with transformer vae. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 10985–10995, 2021.

Mathis Petrovich, Michael J. Black, and Gül Varol. TEMOS: Generating diverse human motions from textual descriptions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022.

Davis Rempe, Tolga Birdal, Aaron Hertzmann, Jimei Yang, Srinath Sridhar, and Leonidas J Guibas. Humor: 3d human motion model for robust pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 11488–11499, 2021.

Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2015.

Leonid Sigal and Michael J Black. Humaneva: Synchronized video and motion capture dataset for evaluation of articulated human motion. *Brown Univertsity TR*, 120(2), 2006.

Casper Kaae Sønderby, Tapani Raiko, Lars Maaløe, Søren Kaae Sønderby, and Ole Winther. Ladder variational autoencoders. *Advances in Neural Information Processing Systems (NeurIPS)*, 29, 2016.

Guy Tevet, Brian Gordon, Amir Hertz, Amit H Bermano, and Daniel Cohen-Or. Motionclip: Exposing human motion generation to clip space. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022.

Herwin Van Welbergen, Ben JH Van Basten, Arjan Egges, Zs M Ruttkay, and Mark H Overmars. Real time animation of virtual humans: a trade-off between naturalness and control. In *Computer Graphics Forum*, 2010.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems (NeurIPS)*, 30, 2017.

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*, pp. 3332–3341, 2017.

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)*, pp. 265–281, 2018.

Ye Yuan and Kris Kitani. Diverse trajectory forecasting with determinantal point processes. *arXiv preprint arXiv:1907.04967*, 2019.

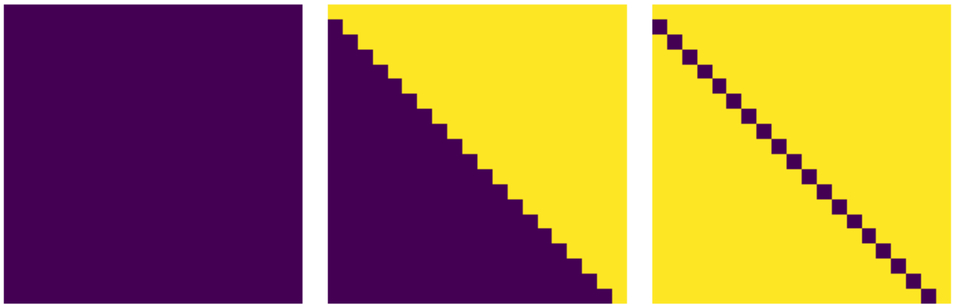
Ye Yuan and Kris Kitani. Dlow: Diversifying latent flows for diverse human motion prediction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 346–364. Springer, 2020.

A Probabilistic Dependencies via Masked MHA

The temporal dependencies are implemented via the mask of the attention modules of the transformer decoder and encoder. The attention matrix in a Transformer layer is computed as follows:

$$\text{Att}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{Softmax}\left(\boldsymbol{\mathcal{M}} \circ \frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d_k}}\right)\boldsymbol{V}, \tag{12}$$

where $\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}$ represent the query, key, and value, and d_k represents the input feature dimension of the query and key. $\boldsymbol{\mathcal{M}}$ is the attention mask and $\circ$ denotes the element-wise multiplication. Obviously, an upper triangular mask without a diagonal can prevent the model to see the future input. In this case, we can generate the entire sequence of $\mathbf{x}_{1:T}$ or $\mathbf{z}_{1:T}$ simultaneously. In practice, given an observed sequence with length T, we only generate $\mathbf{x}_{2:T}$ and $\mathbf{z}_{2:T}$ to bypass the estimation of initial state $\mathbf{x}_0$ and $\mathbf{z}_0$.



(a) $p_{\boldsymbol{\theta}_{\mathbf{x}}}(\mathbf{x}_t|\mathbf{x}_{1:T}, \mathbf{z}_t, \mathbf{w})$ (b) $p_{\boldsymbol{\theta}_{\mathbf{x}}}(\mathbf{x}_t|\mathbf{x}_{1:t-1}, \mathbf{z}_t, \mathbf{w})$ (c) $p_{\boldsymbol{\theta}_{\mathbf{x}}}(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{z}_t, \mathbf{w})$

Figure 4: Probabilistic dependencies on the generation of $\mathbf{x}_t$ with different mask designs, the yellow blocks indicate the elements that will be masked in the attention computation.

Figure 4 shows three cases of probabilistic dependencies when using different masks in the Transformer layer. Note that Figure 4 (a) is a non-causal situation, thus we can not generate future motion via these dependencies. The mask in Figure 4 (c) will make the attention computed only on one element, thus the attention mask is meaningless in this case. We choose the mask shown in Figure 4 (b) in our proposed HiT-DVAE.

B Implementation Details

We set the dimension of $\mathbf{z}_t$ to 16 and $\mathbf{w}$ to 32, and employ the same GCN architecture described in (Mao et al., 2019). We use 1 GCN block with a hidden size of 8 for spatial GCN and 4 GCN blocks with a hidden size of 64 for temporal GCN. For the Transformer encoder and the decoder for generating $\mathbf{z}_t$, we set the input feature dimension to 64, with 4 multi-head, followed by a FC layer with a dimension of 256, whereas for the Transformer decoder to generate $\mathbf{x}_t$, we set those parameters to 256, 4, 1024 respectively.

We generate $K = 50$ samples for each observation. We train the model for 500 epochs with 1000 training samples per epoch, using the Adam optimizer, and set the learning rate to 0.001, batch size to 64 for HumanEva, and 32 for H3.6M. We applied a linear KL annealing (Sønderby et al., 2016) for the first 20 epochs to warm up the latent space, then we take 80 epochs to increase the probability of schedule sampling from 0 to 1. For HumanEva, we train with a sequence length of 75, where the inference of $\mathbf{w}$ only takes 15 frames with a random start point. The weights of different loss terms $(\lambda_R, \lambda_{MM}, \lambda_{DIV-l}, \lambda_{DIV-u}, \lambda_L, \lambda_A, \lambda_{NF}, \lambda_{KL-Z}, \lambda_{KL-W})$ and the normalizing factors (α_l, α_u) are set to (10, 5, 0.1, 0.2, 100, 1, 0.001, 0.5, 0.1) and (15, 50). For H3.6M, we train with a sequence length of 125, where $\mathbf{w}$ is inferred from 25 frames. The weights of different loss terms and the normalizing factors are set to (20, 10, 0.1, 0.2, 100, 1, 0.01, 0.5, 0.1) and (100, 300) respectively.

C Training losses

C.1 Pseudo-multi-modal ground truth

Following Yuan & Kitani (2020) and Mao et al. (2021), we select pseudo-multi-modal ground truth from the training dataset $\mathcal{T}$ using the distance between the last pose of the history. That is, for a training sample $\mathbf{x}_{1:T} \in \mathcal{T}$, we have its pseudo ground truth $\mathbf{x}^m_{1:T}$ when:

$$||\mathbf{x}_O - \mathbf{x}^m_O||_2 < \tau, \quad \forall \mathbf{x}^m_{1:T} \in \mathcal{T}, \tag{13}$$

where O is the last frame of history and τ is the threshold. We set $\tau = 0.5$ for both Human3.6M and HumanEva-I.

C.2 ELBO.

We generate K motion sequences $\{\hat{\mathbf{x}}^k_{1:T}\}^K_{k=1}$ to compute the reconstruction loss:

$$\mathcal{L}_{\text{R}} = \min_k ||\hat{\mathbf{x}}^k_{1:T} - \mathbf{x}_{1:T}||^2, \tag{14}$$

$$\mathcal{L}_{\text{MM}} = \frac{1}{M} \sum_{m=1}^{M} \min_k ||\hat{\mathbf{x}}^k_{1:T} - \mathbf{x}^m_{1:T}||^2, \tag{15}$$

where $\mathbf{x}_{1:T}$ is the ground-truth, and $\mathbf{x}^m_{1:T}$ are the pseudo-ground truth sequences. We select M pseudo ground-truth sequences for each candidate. If the total number of its pseudo ground-truth is less than M, we In addition to the reconstruction losses, we need to minimize the KL divergence:

$$\mathcal{L}_{\text{KL-Z}} = \frac{1}{T} \sum_{t=1}^{T} D_{KL}(q_{\boldsymbol{\phi}_{\mathbf{z}}}(\mathbf{z}_t|\mathbf{x}_{1:T}, \mathbf{w})||p_{\boldsymbol{\theta}_{\mathbf{z}}}(\mathbf{z}_t|\mathbf{x}_{t-1}, \mathbf{z}_{1:t-1}, \mathbf{w})) \tag{16}$$

$$\mathcal{L}_{\text{KL-W}} = D_{KL}(q_{\boldsymbol{\phi}_{\mathbf{w}}}(\mathbf{w}|\mathbf{x}_{1:T})||p_{\boldsymbol{\theta}_{\mathbf{w}}}(\mathbf{w})). \tag{17}$$

The final evidence lower bound (ELBO) writes:

$$\mathcal{L}_{\text{ELBO}} = \lambda_{\text{R}}\mathcal{L}_{\text{R}} + \lambda_{\text{MM}}\mathcal{L}_{\text{MM}} + \lambda_{\text{KL-Z}}\mathcal{L}_{\text{KL-Z}} + \lambda_{\text{KL-W}}\mathcal{L}_{\text{KL-W}}. \tag{18}$$

C.3 Diversity loss.

As suggested by Yuan & Kitani (2020); Mao et al. (2021), we add diversity promoting losses on the upper body and lower body:

$$\mathcal{L}_{\text{DIV}} = \sum_{p\in\{l,u\}} \lambda_{\text{DIV-}p} \frac{2}{K(K-1)} \sum_{k=1}^{K} \sum_{k'=k+1}^{K} \exp\left(-\frac{||\hat{\mathbf{x}}^{k,p}_{1:T} - \hat{\mathbf{x}}^{k',p}_{1:T}||_1}{\alpha_p}\right), \tag{19}$$

where l (u) indicates the lower (upper) body part and α_p is a normalizing factor.

C.4 Realistic pose loss.

Follow Mao et al. (2021), we employ three extra losses to penalize for unrealistic poses, $\mathcal{L}_{\text{L}}$ for shifting limb length, $\mathcal{L}_{\text{A}}$ for aberrant joint angles and $\mathcal{L}_{\text{NF}}$ for negative prior pose probability from a pre-trained pose prior model based on normalizing flow Rezende & Mohamed (2015); Dinh et al. (2017). The realistic loss can be written as $\mathcal{L}_{\text{REAL}} = \lambda_{\text{L}}\mathcal{L}_{\text{L}} + \lambda_{\text{A}}\mathcal{L}_{\text{A}} + \lambda_{\text{NF}}\mathcal{L}_{\text{NF}}$

Altogether, our final training loss writes:

$$\mathcal{L} = \mathcal{L}_{\text{ELBO}} + \mathcal{L}_{\text{DIV}} + \mathcal{L}_{\text{REAL}}. \tag{20}$$

D PSEUDO-CODE FOR HIT-DVAE

Here, we provide the pseudo-code for HiT-DVAE in training and generation:

Algorithm 1 HiT-DVAE in training

```
Inputs:
▷ Observation on human sequence $\mathbf{x}_{1:T}$
for epo in epochs do
    Inference:
    ▷ Compute posterior of $\mathbf{w}$ and sample $\mathbf{w} \sim q_{\boldsymbol{\phi}_{\mathbf{w}}}(\mathbf{w}|\mathbf{x}_{1:T}) = \mathcal{N}(\mathbf{w};\ \boldsymbol{\mu}_{\boldsymbol{\phi}_{\mathbf{w}}}, \boldsymbol{\Sigma}_{\boldsymbol{\phi}_{\mathbf{w}}})$
    ▷ Compute posterior $\mathbf{z}_{1:T}$ and sample $\mathbf{z}_t \sim q_{\boldsymbol{\phi}_{\mathbf{z}}}(\mathbf{z}_t|\mathbf{x}_{1:T}, \mathbf{w}) = \mathcal{N}(\mathbf{z}_t;\ \boldsymbol{\mu}_{\boldsymbol{\phi}_{\mathbf{z}},t}, \boldsymbol{\Sigma}_{\boldsymbol{\phi}_{\mathbf{z}},t})$ for $t = 1, ..., T$
    Generation:
    ▷ Compute the distribution of $\mathbf{x}_{2:T}$ via $p_{\boldsymbol{\theta}_{\mathbf{x}}}(\mathbf{x}_t|\mathbf{x}_{1:t-1}, \mathbf{z}_{1:t}, \mathbf{w}) = \mathcal{N}(\mathbf{x}_t;\ \boldsymbol{\mu}_{\boldsymbol{\theta}_{\mathbf{x}},t}, \boldsymbol{I})$ for $t = 2, ..., T$
    ▷ Compute the prior of $\mathbf{z}_{2:T}$ via $p_{\boldsymbol{\theta}_{\mathbf{z}}}(\mathbf{z}_t|\mathbf{x}_{1:t-1}, \mathbf{z}_{1:t-1}, \mathbf{w}) = \mathcal{N}(\mathbf{z}_t;\ \boldsymbol{\mu}_{\boldsymbol{\theta}_{\mathbf{z}},t}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}_{\mathbf{z}},t})$ for $t = 2, ..., T$
    Compute loss and optimize via Adam
end for
```

Algorithm 2 HiT-DVAE in generation

```
Inputs:
▷ Observation on human sequence $\mathbf{x}_{1:O}$
Initialization:
▷ Compute posterior of $\mathbf{w}$ and $\mathbf{z}_{1:O}$
for t in range(O+1, O+G) do
    ▷ Generate $\hat{\mathbf{z}}_t$ via $\mathbf{z}_t \sim p_{\boldsymbol{\theta}_{\mathbf{z}}}(\mathbf{z}_t|\mathbf{x}_{1:O}, \hat{\mathbf{x}}_{O+1:t-1}, \mathbf{z}_{1:t-1}, \mathbf{w})$
    ▷ Generate $\hat{\mathbf{x}}_t$ via $\mathbf{x}_t \sim p_{\boldsymbol{\theta}_{\mathbf{x}}}(\mathbf{x}_t|\mathbf{x}_{1:O}, \hat{\mathbf{x}}_{O+1:t-1}, \hat{\mathbf{z}}_{1:t}, \mathbf{w})$
end for
Output:
▷ Generated human motion sequence $\hat{\mathbf{x}}_{O+1:O+G}$
```

E ACTION-CLASSIFIER

As explained in the main paper, we trained an RNN-based classifier to calculate the implicit evaluation metrics ACC and FID following Guo et al. (2020) and Petrovich et al. (2021). The classifier we use is built upon 2 simple GRU layers with a hidden size of 128. When training on Human3.6M dataset, we found that some action classes do not differ much from each other, which makes it difficult to train a good classifier. As our goal is to have a classifier that offers good features, we believe the classifier with low accuracy on real data is not reliable enough, so we group the 5 similar actions, and trained the classifier on these 5 groups instead of the 15 original classes. The groups of actions are detailed in Table 5.
Note that even on the classifier trained on the 15 original classes, our method still performs better than others, as shown in Table 6.

Table 5: Groups of actions of Human3.6m dataset, for training the action classifier.

group number	original classes
0	Directions, Discussion, Greeting, Photo, Posing, Purchases, WalkDog, Waiting
1	Eating, Phoning, Sitting, Smoking
2	SittingDown
3	Walking
4	WalkTogether

Table 6: Implicit evaluations by different classification models on Human3.6m dataset. Our method always performs better.

	5 groups		15 classes	
	Acc (%) ↑	FID ↓	Acc (%) ↑	FID ↓
Real data	85.5	-	48.1	-
DLow Yuan & Kitani (2020)	65.9	**1.412**	22.7	1.566
GSPS Mao et al. (2021)	65.0	2.030	22.2	1.915
GumbelSample Dang et al. (2022)	66.5	2.659	24.4	2.471
HiT-DVAE	**70.0**	1.708	**28.1**	**1.466**

F More Visualisation Results in Video

As shown in the main paper, we qualitatively compare our methods with the state-of-the-art methods (Mao et al., 2021; Yuan & Kitani, 2020) by showing the end poses of 10 generated motion sequences. In the attached video, we further compare the results by visualizing 10 generated samples for each method. For better visualization, we highlight the "unrealistic" generated motions which evidently do not correspond to the action label and are not similar to the real motion, based on a simple user study: We sent a survey to 5 people to select the most unrealistic generation results and marked the results with the agreement of at least 3 people. We can see that our method generates more realistic motion sequences and keep the diversity, while others are less stable.