Contextually Plausible and Diverse 3D Human Motion Prediction

Sadegh Aliakbarian* Microsoft
Fatemeh Saleh ACRV, ANU
Lars Petersson Data61, CSIRO
Stephen Gould ACRV, ANU
Mathieu Salzmann CVLab, EPFL

Abstract

We tackle the task of diverse 3D human motion prediction, that is, forecasting multiple plausible future 3D poses given a sequence of observed 3D poses. In this context, a popular approach consists of using a Conditional Variational Autoencoder (CVAE). However, existing approaches that do so either fail to capture the diversity in human motion, or generate diverse but semantically implausible continuations of the observed motion. In this paper, we address both of these problems by developing a new variational framework that accounts for both diversity and context of the generated future motion. To this end, and in contrast to existing approaches, we condition the sampling of the latent variable that acts as source of diversity on the representation of the past observation, thus encouraging it to carry relevant information. Our experiments demonstrate that our approach yields motions not only of higher quality while retaining diversity, but also that preserve the contextual information contained in the observed motion.

1. Introduction

Human motion prediction is the task of forecasting plausible 3D human motion continuation(s) given a sequence of past 3D human poses. To address this problem, prior work mostly relies on recurrent encoder-decoder architectures, where the encoder processes the observed motion, and the decoder generates a single future trajectory given the encoded representation of the past [5, 10, 16, 22, 23, 24, 28, 21]. While this approach yields valid future motion, it tends to ignore the fact that human motion is stochastic in nature; given one single observation, multiple diverse continuations of the motion are likely and plausible. The lack of stochasticity of these encoder-decoder methods ensues from the fact that both the network operations and the sequences in the training dataset are deterministic[1]. In this

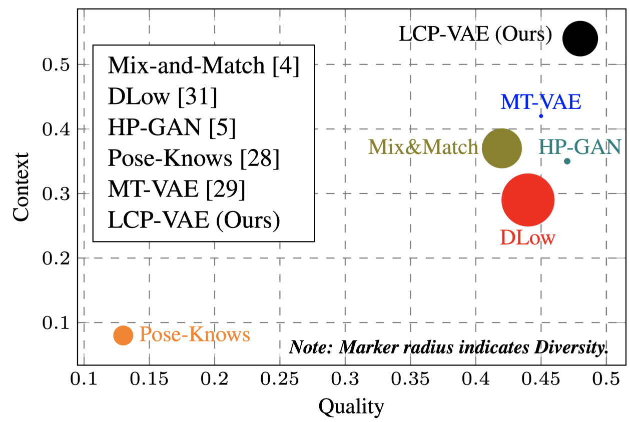


Figure 1. Evaluating the quality, diversity, and context of stochastic motion prediction models. We show each model as a circle in a Context vs. Quality plot, further indicating Diversity with the radius of the marker. These three metrics are defined in Section 5. Our LCP-VAE approach (in black) yields diverse predictions of higher-quality than those of other methods, while better preserving the context of the observations.

paper, we introduce an approach to modeling this stochasticity by learning *multiple modes* of human motion. We focus on generating both diverse *and* contextually and semantically plausible motion predictions. By contextually plausible, we mean motions that are natural continuations of an observed sequence of 3D poses, preserving and continuing the context depicted by the observation. For instance, when the observed poses depict a person walking, we expect the network to predominantly predict future motions corresponding to different walking modes. This is crucial in certain scenarios, such as pedestrian intention forecasting in autonomous driving, user interaction in AR/VR applications, and automatic animation generation, in which context is as important as diversity.

Recent attempts to account for motion stochasticity mostly rely on combining a random vector with an encoding of the observed pose sequence [4, 5, 7, 16, 19, 28, 29, 31]. In particular, the state-of-the-art approaches to diverse human motion prediction [4, 29, 31] use conditional variational autoencoders (CVAEs). Here, we argue that standard CVAEs are ill-suited to generate motions that are di-

*Work done while at the Australian National University.

[1]For complicated tasks such as motion prediction, the training data is typically insufficiently sampled, in that, for any given condition, the dataset contains only a single sample, in effect making the data appear deterministic. For instance, in motion prediction, we never observed twice the same past motion with two different future ones.

verse *and* contextually plausible for the following reasons: First, standard CVAEs struggle at capturing diversity when the conditioning signal is highly informative. In the particular case of human motion prediction, the observed 3D motion (*i.e.*, the condition) contains sufficient signal for an expressive motion decoder to generate a natural continuation given only this condition [4, 29]. A typical conditioning scheme, such as concatenating a random latent vector to the condition, then allows the model to learn to ignore the latent variable and only focus on the condition to minimize the reconstruction loss. Second, while ignoring the latent variable can be prevented by replacing the traditional deterministic conditioning scheme with a stochastic one [4], capturing context in the diverse predictions with CVAEs is impeded by their use of a general prior on the latent variable. Since this prior is independent of the conditioning signal, during inference, nothing encourages the latent variable to be drawn from a region of the latent space corresponding to the observed motion. In other words, the latent variable is sampled independently of the condition and thus may not carry contextual information about the observed motion.

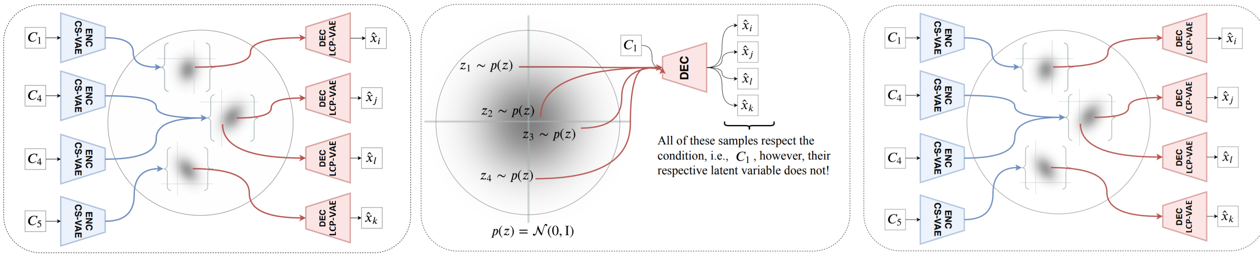
In this paper, we address these weaknesses by explicitly conditioning the sampling of the latent variable on the past observation, thus encouraging the latent variable to encode relevant information. We will show that this further allows us to depart from the traditional deterministic conditioning scheme, and thus facilitate generating both diverse *and* contextually plausible motions. As demonstrated in Fig. 1, our experiments show that our approach not only yields much higher quality of diverse motions compared to the state-of-the-art stochastic motion prediction methods, but also better preserves the contextual and semantic information of the condition, such as the type of action performed by the person, without explicitly exploiting this information.

2. Related Work

Most motion prediction methods are based on *deterministic* models [8, 9, 10, 11, 13, 22, 23, 24, 21], casting motion prediction as a regression task where only one outcome is possible given the observation. While this may produce accurate predictions, it fails to reflect the stochastic nature of human motion, where multiple futures can be highly likely for a single series of past observations. Modeling stochasticity is the topic of this paper, and we therefore focus the discussion on the methods that have attempted to do so.

The general trend to incorporate variations in the predicted motions consists of combining information about the observed pose sequence with a random vector. In this context, two types of approaches have been studied: The techniques that directly incorporate the random vector into the motion decoder (typically an RNN) and those that make use of an additional CVAE. In the first class of methods, [19] samples a random vector $z_t \sim \mathcal{N}(0, I)$ at each time step and concatenates it to the pose input to the RNN decoder. By relying on different and independent random vectors at each time step, however, this strategy is prone to generating discontinuous motions. To overcome this, [16] makes use of a single random vector to generate the entire sequence. As we will show in our experiments, by relying on concatenation, these two methods contain parameters that are specific to the random vector, and thus give the model the flexibility to ignore this information. In [5], instead of using concatenation, the random vector is added to the hidden state produced by the RNN encoder. While addition prevents having parameters that are specific to the random vector, this vector is first transformed by multiplication with a learnable parameter matrix, and thus can again be zeroed out so as to remove the source of diversity, as observed in our experiments.

The second category of stochastic methods introduce an additional CVAE between the RNN encoder and decoder. In this context, [28] proposes to directly use the pose as conditioning variable. As will be shown in our experiments, while this approach is able to maintain some degree of diversity, albeit less than ours, it yields motions of lower quality because of its use of independent random vectors at each time step. Instead of perturbing the pose, [29] uses the RNN encoder's hidden state as conditioning variable in the CVAE, concatenating it with the latent variable. While this approach generates high-quality motions, it suffers from the fact that the CVAE decoder gives the model the flexibility to ignore the random vector, which therefore yields low-diversity outputs. To overcome this, [4] perturbs the hidden states via a stochastic Mix-and-Match operation instead of concatenation. Such a perturbation prevents the decoder from decoupling the noise and the condition. However, since the perturbation is not learned and is a non-parametric operation, the quality and the context of the generated motion are inferior to those obtained with our approach. Similarly to Mix-and-Match [4], DLow [31] aims to modify the sampling process of a fixed, pretrained CVAE so as to ensure diversity in the output space. While this approach is successful at generating diverse motions, it relies on a two-stage training process, making it impossible to *learn* diversity in motions in an end-to-end manner. Furthermore, diversity is achieved by learning a fixed number of transformations, thus preventing this method to generalize to arbitrary numbers of motion modes. Finally, [31] does not account for the fact that the diverse continuations of an observation should preserve its context and semantics. More importantly, all of the above-mentioned CVAE-based approaches use priors that are independent of the condition. We will show in our experiments that such designs are ill-suited for human motion prediction. By contrast, our approach *learns* a conditional prior and is thus able to generate diverse motions of higher quality, carrying the contextual



(a) CVAE at training (b) CVAE at inference (c) LCP-VAE (ours) at inference

Figure 2. Training and inference for a standard CVAE versus inference for our of LCP-VAE. (a) During the training of a standard CVAE, the encoder takes as input the combination of the data and the corresponding condition and compresses it into the latent space. The decoder then samples a latent variable from the approximate posterior computed by the encoder, combines the latent variable with the conditioning signal, and reconstructs the data. (b) During inference in a standard CVAE, the decoder samples different latent variables from the prior distribution, combines them with the condition, and generates samples that all respect the conditioning signal. However, there is no guarantee that the latent variable is sampled from the region of the prior that corresponds to the given condition. (c) In LCP-VAE, at inference time, we utilize an additional encoder acting on the conditioning signal to approximate the posterior of each condition. To generate a sample given a condition, our decoder then samples a latent variable from the posterior of its condition, instead of using a general prior distribution as in CVAEs, and generates a sample.

information of the conditioning signal.

3. Motivation

In essence, VAEs utilize neural networks to learn the distribution of the data. To this end, VAEs first learn to generate a latent variable z given the data x, *i.e.*, approximate the posterior distribution $q_\phi(z|x)$, where ϕ are the parameters of a neural network, the encoder, whose goal is to model the variation of the data. From this latent random variable z, VAEs then generate a new sample x by learning $p_\theta(x|z)$, where θ denotes the parameters of another neural network, the decoder, whose goal is to maximize the log likelihood of the data. These two networks, *i.e.*, the encoder and the decoder, are trained jointly, using a prior over the latent variable. By using a variational approximation of the posterior, training translates to maximizing the variational lower bound of the log likelihood with respect to the parameters ϕ and θ, given by

$$\log p_\theta(x) \geq \mathbb{E}_{q_\phi(z|x)}\Big[\log p_\theta(x|z)\Big] - KL\Big(q_\phi(z|x)\,||\,p(z)\Big)\ , \tag{1}$$

where the second term on the right hand side encodes the KL divergence between the posterior $q_\phi(z|x)$ and a chosen prior distribution $p(z)$. As an extension to VAEs, CVAEs use auxiliary information, *i.e.*, the conditioning variable or observation, to generate the data x. In the standard setting, both the encoder and the decoder are conditioned on the conditioning variable c. That is, the encoder becomes $q_\phi(z|x, c)$ and the decoder $p_\theta(x|z, c)$. Then, in theory, the objective of the model should become

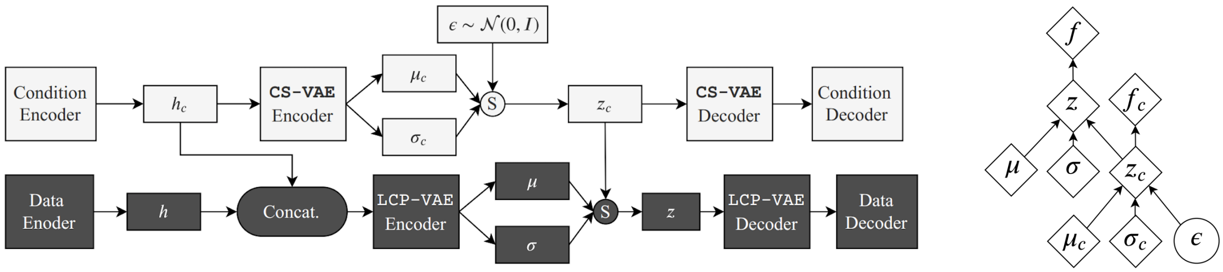
$$\log p_\theta(x|c) \geq \mathbb{E}_{q_\phi(z|x,c)}\Big[\log p_\theta(x|z, c)\Big] - KL\Big(q_\phi(z|x, c)\,||\,p(z|c)\Big)\ . \tag{2}$$

In practice, however, *the prior distribution of the latent variable is still assumed to be independent of c, i.e.,* $p(z \mid c) = p(z)$. As illustrated by Fig. 2, at test time, this translates to sampling a latent variable from a region of the prior that may not be correlated with the observed condition.

In this paper, we overcome this limitation by explicitly making the sampling of the latent variable depend on the condition. In other words, instead of using $p(z)$ as prior distribution, we truly use $p(z|c)$. This not only respects the theory behind the design of CVAEs, but, as we empirically demonstrate, leads to generating motions of higher quality, that preserve the context of the conditioning signal, *i.e.*, the observed past motion. To achieve this, we develop a CVAE architecture that learns a distribution not only of the latent variable but also of the conditioning one. We then use this distribution as a prior over the latent variable, making its sampling explicitly dependent on the condition.

4. Our Approach: LCP-VAE

In this section, we describe our approach to generating diverse and plausible continuations of 3D human pose sequences, where the latent variables are sampled from an appropriate region of the prior distribution. We dub it LCP-VAE for VAEs with Learned Conditional Priors. In essence, our framework consists of two autoencoders, one acting on the past observed sequence of 3D poses (*i.e.*, the conditioning signal) and the other on the future sequence of



(a) The LCP-VAE framework (b) Extended Reparameterization Trick

Figure 3. In an LCP-VAE (a), the sampling of z is conditioned on the CVAE condition via the latent variable z_c. Specifically, the posterior distribution of the condition acts as prior on the data posterior. This corresponds to the *extended reparameterization trick* (b) illustrated on the right. Note that this approximate posterior is not normally distributed anymore.

3D poses (*i.e.*, the continuation of the motion, also referred to as the data) which we wish to model. The latent representation of the condition then serves as a conditioning variable to generate motions from a learned distribution.

As discussed in Section 3, our approach forces the sampling of the random latent variable to depend on the conditioning one. By making this dependency explicit, we (1) sample an informative latent variable *given* the condition, and thus generate motions of higher quality that preserve the context of the observed motion, and (2) prevent the network from ignoring the latent variable in the presence of a strong condition, thus enabling it to generate diverse outputs.

Note that conditioning the VAE *encoder* via standard strategies, *e.g.*, concatenation, is perfectly fine, since the two inputs to the encoder, *i.e.*, the data and the condition, are deterministic and useful to compress the sample into the latent space. However, conditioning the VAE *decoder* requires special care, which is what we focus on below.

4.1. Stochastically Conditioning the Decoder

We propose to make the sampling of the latent variable explicitly depend on the condition instead of treating these two variables as independent. To this end, we first learn the distribution of the condition via a simple VAE, which we refer to as CS-VAE because it acts on the conditioning signal. The goal of CS-VAE is to reconstruct the condition, *i.e.*, the observed past motion, given its latent representation. We take the prior of CS-VAE to be a standard Normal distribution $\mathcal{N}(0, I)$. Following [14], this allows us to approximate the CS-VAE posterior with another Normal distribution via the reparametrization trick expressed as

$$z_c = \mu_c + \sigma_c \odot \epsilon \,, \tag{3}$$

where $\epsilon \sim \mathcal{N}(0, I)$, and μ_c and σ_c are the parameter vectors of the posterior distribution generated by the VAE encoder, and thus $z_c \sim \mathcal{N}(\mu_c, \text{diag}(\sigma_c)^2)$.

Following the same strategy for the data VAE, that acts on the future motion, translates to assuming independence of the conditioning and the data, which we seek to avoid. Therefore, as illustrated in Fig. 3 (Bottom), we instead define the LCP-VAE posterior not as directly normally distributed but conditioned on the posterior of CS-VAE. To this end, we extend the standard reparameterization trick as

$$\begin{aligned} z = \mu + \sigma \odot z_c =& \mu + \sigma \odot (\mu_c + \sigma_c \odot \epsilon) \\ =& \underbrace{(\mu + \sigma \odot \mu_c)}_{\text{LCP-VAE's mean}} + \underbrace{(\sigma \odot \sigma_c)}_{\text{LCP-VAE's std.}} \odot \epsilon \,, \end{aligned} \tag{4}$$

where z_c comes from Eq. 3, and μ and σ are the parameter vectors generated by the LCP-VAE encoder. In fact, z_c in Eq. 3 is a sample from the scaled and translated version of $\mathcal{N}(0, I)$ given μ_c and σ_c, and z in Eq. 4 is a sample from the scaled and translated version of $\mathcal{N}(\mu_c, \text{diag}(\sigma_c)^2)$ given μ and σ. Since we have access to the observations during both training and testing, we always sample z_c from the condition posterior. As z is sampled given z_c, one expects the latent variable z to carry information about the strong condition, and thus a sample generated from z to correspond to a plausible sample given the condition. This extended reparameterization trick lets us sample one single informative latent variable that contains information about both the data and the conditioning signal. This further allows us to avoid conditioning the LCP-VAE decoder by concatenating the latent variable with a deterministic representation of the condition. Note that our sampling strategy changes the variational family of the LCP-VAE posterior. In fact, the posterior is no longer $\mathcal{N}(\mu, \text{diag}(\sigma)^2)$, but a Gaussian distribution with mean $\mu + \sigma \odot \mu_c$ and covariance matrix $\text{diag}(\sigma \odot \sigma_c)^2$. This will be accounted for when designing the KL divergence loss discussed below.

4.2. Learning

To learn the parameters of our model, we rely on the availability of a dataset $D = \{X_1, X_2, ..., X_N\}$ containing N training motion sequences X_i, $1 \leq i \leq N$. Each motion, $X_i = \{x_i^1, x_i^2, ..., x_i^T\}$, consists of a sequence of T poses, and each pose, $x_i^t = \{x_{i,1}^t, x_{i,2}^t, ..., x_{i,J}^t\}$, comprises

J joints forming a skeleton. The pose of each joint is represented as a 4D quaternion. Each training sample X_i contains a past observed motion, or condition, $x_i^{1:t}$, and a future motion, $x_i^{t+1:T}$. For `CS-VAE`, which learns the distribution of the condition, we define the loss as the KL divergence between its posterior and the standard Gaussian prior, *i.e.*,

$$\begin{aligned}\mathcal{L}_{prior}^{\texttt{CS-VAE}} &= KL\Big(\mathcal{N}(\mu_c, \text{diag}(\sigma_c)^2)\Big\|\mathcal{N}(0, I)\Big) \\ &= -\frac{1}{2}\sum_{j=1}^{d}\Big(1 + \log(\sigma_{c_j}^2) - \mu_{c_j}^2 - \sigma_{c_j}^2\Big)\,, \end{aligned} \tag{5}$$

where d is the dimension of the latent variable z_c. By contrast, for `LCP-VAE`, we define the loss as the KL divergence between the posterior of `LCP-VAE` and the posterior of `CS-VAE`, *i.e.*, of the condition. To this end, at each training iteration, we freeze the weights of `CS-VAE` before computing the KL divergence, since we do not want to move the posterior of the condition but that of the data. The KL divergence is then computed as the divergence between two multivariate Normal distributions, which yields

$$\mathcal{L}_{prior}^{\texttt{LCP-VAE}} = \\ KL\Big(\mathcal{N}(\mu + \sigma \odot \mu_c, \text{diag}(\sigma \odot \sigma_c)^2)\Big\|\mathcal{N}(\mu_c, \text{diag}(\sigma_c)^2)\Big)\,. \tag{6}$$

Let $\Sigma = \text{diag}(\sigma)^2$, $\Sigma_c = \text{diag}(\sigma_c)^2$, d be the dimensionality of the latent space and $tr\{\cdot\}$ the trace of a square matrix. The loss in Eq. 6 can then be written as[2]

$$\begin{aligned}\mathcal{L}_{prior}^{\texttt{LCP-VAE}} = -\frac{1}{2}\Big[&\log\frac{1}{|\Sigma|} - d + tr\{\Sigma\} \\ &+ (\mu_c - (\mu + \Sigma\mu_c))^T\Sigma_c^{-1}(\mu_c - (\mu + \Sigma\mu_c))\Big]\,.\end{aligned} \tag{7}$$

After computing the loss in Eq. 7, we unfreeze `CS-VAE` and update it with the gradient of the loss in Eq. 5. Trying to match the posterior of `LCP-VAE` to that of `CS-VAE` allows us to effectively use our extended reparameterization trick in Eq. 4. Furthermore, we use the standard reconstruction loss for both `CS-VAE` and `LCP-VAE`, thus minimizing the mean squared error (MSE)

$$\mathcal{L}_{rec} = -\sum_{k=t'}^{T'}\sum_{j=1}^{J}\left\|\hat{x}_{i,j}^k - x_{i,j}^k\right\|^2\,, \tag{8}$$

in which for $\mathcal{L}_{rec}^{\texttt{CS-VAE}}$, $t' = 1$ and $T' = t$ (*i.e.*, the observation), and for $\mathcal{L}_{rec}^{\texttt{LCP-VAE}}$, $t' = t + 1$ and $T' = T$ (*i.e.*, the future motion). Thus, our complete loss is

$$\mathcal{L} = \lambda(\mathcal{L}_{prior}^{\texttt{CS-VAE}} + \mathcal{L}_{prior}^{\texttt{LCP-VAE}}) + \mathcal{L}_{rec}^{\texttt{CS-VAE}} + \mathcal{L}_{rec}^{\texttt{LCP-VAE}}\,. \tag{9}$$

In practice, since the nature of our data is sequential, we use a recurrent model for the `LCP-VAE` encoder. Thus, following the standard practice in diverse human motion prediction [4, 28, 29], we weigh the KL divergence terms by a function λ corresponding to the KL annealing weight of [6]. We start from $\lambda = 0$, forcing the model to encode as much information in z as possible, and gradually increase it to $\lambda = 1$ during training, following a logistic curve. We then continue training with $\lambda = 1$.

In short, our method can be interpreted as a simple yet effective CVAE framework for altering the variational family of the posterior such that a latent variable from this posterior distribution is explicitly sampled given the condition, both during training and inference (as illustrated in Fig. 2). We also empirically observed that `LCP-VAE` is less likely to suffer from posterior collapse due to the fact that the latent variable is sampled given the condition, and therefore, the decoder is not able to decompose the latent variable from the condition and thus to ignore it and only rely solely on the condition, the phenomena that happens in standard CVAEs.

5. Experiments

In this section, we compare our approach to the state of the art and evaluate the influence of the different components of our method. We provide the implementation details of our approach in the supplementary material.

Datasets. For our experiments, we use the Human3.6M [12], CMU MoCap[3] and Penn Action [32] datasets. Human3.6M comprises more than 800 long indoor motion sequences performed by 11 subjects, leading to 3.6M frames. Each frame contains a person annotated with 3D joint positions and rotation matrices for all 32 joints. In our experiments, for our approach and the replicated VAE-based baselines, we represent each joint in 4D quaternion space. We follow the standard preprocessing and evaluation settings used in [4, 10, 13, 22, 24]. The CMU MoCap dataset is another large-scale motion capture dataset covering diverse human activities, such as jumping, running, walking, and playing basketball. Each frame contains a person annotated with 3D joint rotation matrices for all 38 joints. As for Human3.6M, and following standard practice [17, 21], we represent each joint in 4D quaternion space. The real-world Penn Action dataset [32] contains 2326 sequences of 15 different actions, where for each person, 13 joints are annotated in 2D space. For this dataset, we therefore tackle the task of 2D motion prediction. The results on Penn Action are provided in the supp. material.

Evaluation Metrics. To quantitatively evaluate our approach and the state-of-the-art stochastic motion prediction methods [4, 5, 28, 29, 31], we report the test reconstruction error, along with the KL-divergence on the held-out test set.

[2]See the supplementary material for more detail on the KL divergence between two multivariate Gaussians and the derivation of Eq. 7.

[3]Available at `http://mocap.cs.cmu.edu/`.

Table 1. Comparison of `LCP-VAE` with the state-of-the-art stochastic motion prediction methods.

Results on Human3.6M

Method	Test MSE (KL) (Reconstructed)	Diversity (Sampled)	Quality (Sampled)	Context (Sampled)
MT-VAE [29]	0.51 (0.06)	0.26	0.45	0.42
Pose-Knows [28]	2.08 (N/A)	1.70	0.13	0.08
HP-GAN [5]	0.61 (N/A)	0.48	0.47	0.35
Mix-and-Match [4]	0.55 (2.03)	3.52	0.42	0.37
DLow [31]	0.42 (0.61)	**4.71**	0.44	0.29
`LCP-VAE`	**0.41 (8.07)**	3.12	**0.48**	**0.54**

Results on CMU MoCap

Method	Test MSE (KL) (Reconstructed)	Diversity (Sampled)	Quality (Sampled)	Context (Sampled)
MT-VAE [29]	0.25 (0.08)	0.41	0.46	0.80
Pose-Knows [28]	1.93 (N/A)	3.00	0.18	0.27
HP-GAN [5]	0.24 (N/A)	0.43	0.45	0.73
Mix-and-Match [4]	0.25 (2.92)	2.63	0.46	0.78
DLow [31]	0.26 (0.25)	**2.90**	0.44	0.34
`LCP-VAE`	**0.23 (4.13)**	2.36	**0.48**	**0.88**

Table 2. Comparison of the generated motions with the ground-truth future motions in terms of context. The gap between the performance of the state-of-the-art pose-based action classifier [18] with and without true future motions is 0.22 / 0.54 for Human3.6M / CMU MoCap. Using our predictions, this gap decreases to 0.06 / 0.08, showing that our predictions reflect the correct class label.

Setting	Obs.	Future Motion	Context (H3.6M / CMU)
Lower bound	GT	Zero velocity	0.38 / 0.42
Upper bound	GT	GT	0.60 / 0.96
Ours	GT	Sampled from `LCP-VAE`	0.54 / 0.88

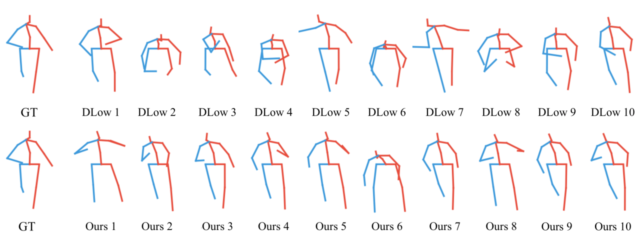


Figure 4. Qualitative comparison of diversity with the state-of-the-art DLow method [31]. The first column shows the last time step of the ground-truth motion, and the next 10 columns represent the last time step of 10 different samples from DLow [31] (top row) and from our approach (bottom row). While DLow generates highly diverse motions, it loses the context of the observed motion (*walking* in this example). By contrast, our approach generates diverse motions that preserve context in most cases, with a few exceptions, such as Sample 6.

Additionally, we report quality [4] and diversity [4, 30, 31] metrics, which should be considered together. Specifically, to measure the diversity of the motions generated by a stochastic model, we make use of the average distance between all pairs of the K motions generated from the same observation. To measure quality, following the quality metric of [4], we train a binary classifier to discriminate real (ground-truth) samples from fake (generated) ones. The accuracy of this classifier on the test set is inversely proportional to the quality of the generated motions. Note that the results of such a classifier were shown in [4] to match those of human evaluation. Furthermore, we report a context metric measured as the performance of a strong action classifier [18] trained on ground-truth motions. Specifically, the classifier is tested on each of the K motions generated from each observation. For N observations and K continuations per observation, the accuracy is measured by computing the argmax over each prediction's probability vector, and we report context as the mean class accuracy on the $K \times N$ motions. Unless otherwise stated, for all metrics, we use $K = 50$ motions per test observation. For all experiments, we follow [4, 29] and use 16 frames (*i.e.*, 640ms) as observation to generate the next 60 frames (*i.e.*, 2.4sec).

5.1. Comparison to the State of the Art

In Table 1, we compare our approach (whose detailed architecture is described in the supplementary material) with the state-of-the-art stochastic motion prediction models [4, 5, 28, 29, 31]. Note that, when evaluating stochastic motion prediction models, one should consider the reported metrics jointly to truly evaluate the performance of a stochastic model, as we do in Fig. 1. For instance, while MT-VAE [29] and HP-GAN [5] generate high-quality motions, they are not diverse. Conversely, while Pose-Knows [28] generates diverse motions, they are of low quality. By contrast, our approach generates both high quality and diverse motions. This is also the case of Mix-and-Match [4] and DLow [31], which, however, preserve much less contextual information and semantics of the observed motions. Fig. 4 compares motions generated by our approach and the most diverse model DLow. Note that DLow [31] is unable to preserve the contextual information of the observation, whereas our approach does. In fact, none of the baselines effectively conveys the semantics of the observation to the generated motions. As shown in Table 2, the upper bounds for context on Human3.6M / CMU are 0.60 / 0.96 (*i.e.*, the performance of [18] given the ground-truth motions), which our method approaches closely, yielding context scores of 0.54 / 0.88 when given only the observed motion. The fact that our method better preserves context is further evidenced by the t-SNE [20] plots of Fig. 6, where different samples of various actions are better separated for our approach than for MT-VAE [29], the method that best retains context among the baselines. Altogether, as also supported by the qualitative results of Fig. 5 and in the supplementary material, our approach yields diverse, high-quality and context-preserving predictions.

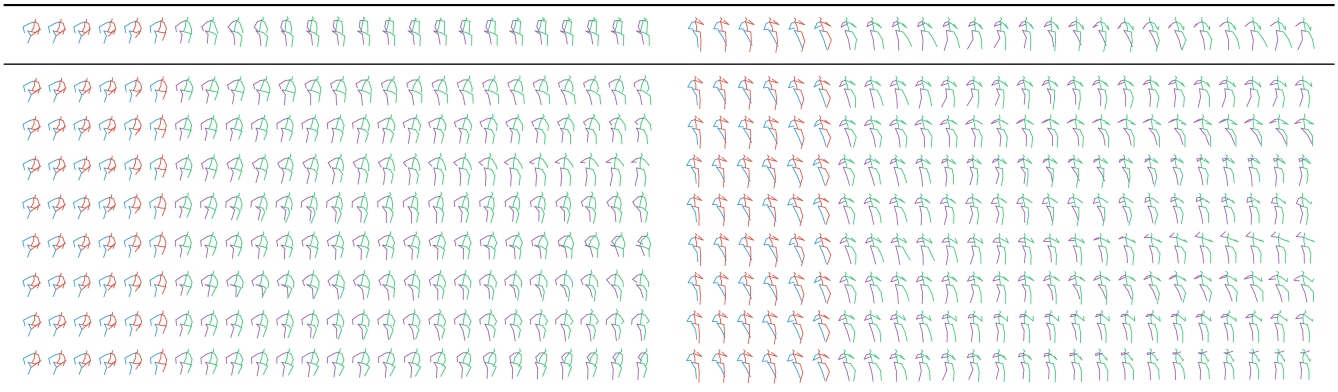


Figure 5. Qualitative examples of the diversity in human motion. The first row depicts the ground-truth motion. The first six poses of each row correspond to the observation (the condition) and the remaining ones are sampled from our model. Each row is a randomly sampled motion (not cherry-picked). Note that all motions are natural, with a smooth transition from the observed poses to the generated ones.

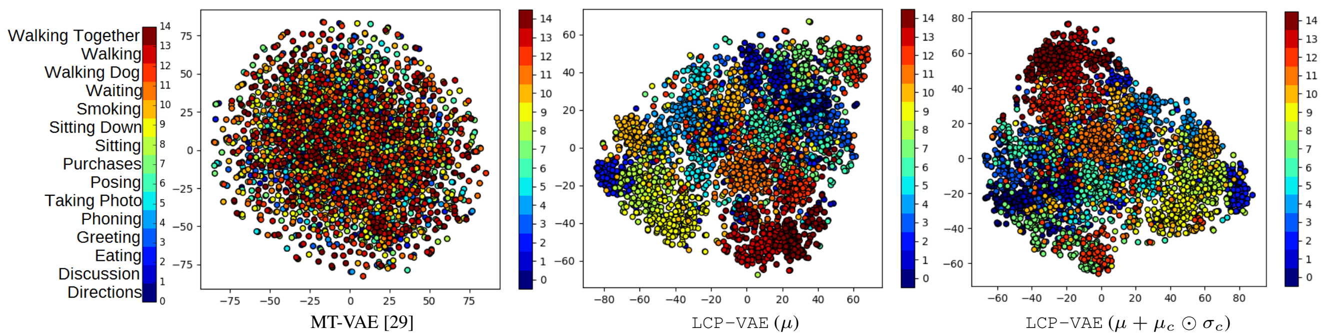


Figure 6. t-SNE plots of the posterior mean for 3750 test motions. With MT-VAE [29], all classes are mixed, suggesting that the latent variable carries little information about the motion type. By contrast, our condition-dependent sampling allows `LCP-VAE` to better preserve context. Note that some actions, such as "Discussion" and "Directions", are very hard to identify and are thus spread over other actions. Others, such as "Walking", "Walking with dog", and "Walking together", or "Sitting" and "Sitting down", overlap due to their similarity.

Evaluating Sampling Quality. To further evaluate the sampling quality, we evaluate stochastic baselines using the standard mean angle error (MAE) metric in Euler space. To this end, we use the best of the $K = 50$ generated motions for each observation (referred to as S-MSE in [29]). A model that generates more diverse motions has higher chances of producing a motion close to the ground-truth one. As shown in Table 3, this is the case with our approach, Mix-and-Match [4], and DLow [31][4], which all yield high diversity. Note, however, that Mix-and-Match [4] and DLow [31] focus mainly on encouraging motion diversity, either by introducing a stochastic conditioning scheme that ignores part of the conditioning signal [4] or via a fixed set of latent transformations that spread the latent variable across different regions of the latent space [31]. As such, they inevitably lose some context in the generated motions. By contrast, we learn a context-preserving latent representation, which helps our approach to generate higher quality motions. In the supplementary material, we additionally compare our approach with the state-of-the-art deterministic motion prediction models.

5.2. Ablation Studies and Analysis

In this section, we analyze the performance of different stochastic motion prediction models. We further provide insights on different means of conditioning for the encoder and the decoder of our model.

Why existing methods fail to generate diverse *and* plausible motions? Here, we further analyze the behavior of different baselines under different evaluation metrics, reported in Fig. 1 and Table 1. These results show that MT-VAE [29] tends to ignore the random variable z, thus ignoring the root of stochasticity. As a consequence, it yields a low diversity, much lower than ours, but produces samples of high quality, albeit almost identical. We empirically traced this back to the magnitude of the weights acting on z being orders of magnitude smaller than that of those

[4]We used the official implementation of DLow [31] (`https://github.com/Khrylx/DLow`) and trained it to generate 50 diverse predictions via training 50 latent transformations.

Table 3. Comparison with the state-of-the-art stochastic motion prediction models in terms of sampling quality for 4 actions of Human3.6M (all methods use the best of $K = 50$ sampled motions).

	Walking						Eating					
Method	80	160	320	400	560	1000	80	160	320	400	560	1000
MT-VAE [29]	0.73	0.79	0.90	0.93	0.95	1.05	0.68	0.74	0.95	1.00	1.03	1.38
HP-GAN [5]	0.61	0.62	0.71	0.79	0.83	1.07	0.53	0.67	0.79	0.88	0.97	1.12
Pose-Knows [28]	0.56	0.66	0.98	1.05	1.28	1.60	0.44	0.60	0.71	0.84	1.05	1.54
Mix&Match [4]	0.33	0.48	0.56	0.58	0.64	**0.68**	0.23	0.34	0.41	**0.50**	0.61	0.91
DLow [31]	0.31	0.42	0.53	0.75	0.83	0.96	0.24	0.32	0.44	0.55	0.77	0.97
`LCP-VAE`	**0.22**	**0.36**	**0.47**	**0.52**	**0.58**	0.69	**0.19**	**0.28**	**0.40**	0.51	**0.58**	**0.90**
	Smoking						Discussion					
Method	80	160	320	400	560	1000	80	160	320	400	560	1000
MT-VAE [29]	1.00	1.14	1.43	1.44	1.68	1.99	0.80	1.01	1.22	1.35	1.56	1.69
HP-GAN [5]	0.64	0.78	1.05	1.12	1.64	1.84	0.79	1.00	1.12	1.29	1.43	1.71
Pose-Knows [28]	0.59	0.83	1.25	1.36	1.67	2.03	0.73	1.10	1.33	1.34	1.45	1.85
Mix&Match [4]	0.23	**0.42**	0.79	0.77	0.82	1.25	0.25	0.60	0.83	0.89	1.12	1.30
DLow [31]	**0.21**	0.43	0.80	0.79	0.97	1.65	0.31	0.55	**0.80**	0.88	1.15	1.33
`LCP-VAE`	0.23	0.43	**0.77**	**0.75**	**0.78**	**1.23**	**0.21**	**0.52**	0.81	**0.84**	**1.04**	**1.28**

acting on the condition, suggesting that MT-VAE [29] has learned to ignore the latent variable. Similarly, our experiments show that HP-GAN [5] also yields limited diversity despite its use of random noise during inference. This is due to the fact[5] that in [5], z is linearly transformed by a learnable weight matrix W before being added to the motion representation. We then observed that the magnitude of the parameters in W is almost zero, allowing HP-GAN [5] to ignore z. Unlike [5, 29], Pose-Knows [28] produces more diverse motions, but their quality is very low. By inspecting their code[6], we observed that the random vectors that are concatenated to the poses at each time-step are sampled independently of each other, which translates to discontinuities in the generated motions. The Mix-and-Match approach [4] yields generated motions with reasonably high diversity and quality. Its architecture is very close to that of MT-VAE [29], but the deterministic concatenation operation is replaced with a stochastic perturbation of the hidden state with the noise. While this prevents the model from ignoring the random noise, the perturbation is not learnt, which translates to lower-quality predictions than ours and that of other baselines. DLow [31], which learns to diversify the sampled latent variables of a fixed, pre-trained CVAE through multiple transformations, also generates highly diverse motions. However, the predictions tend to ignore the context of the observed motion as the transformations are trained so as to maximize diversity. While this indeed makes the DLow predictions diverse, the resulting model relies on a fixed number of learned transformations, and thus cannot generate an arbitrary number of diverse motions at test time. That is, a DLow model with K transformations can generate only K different modes of the data; generating $K+1$ modes would require retraining the model.

[5] We noticed this by studying and running the authors' publicly available code at `https://github.com/ebarsoum/hpgan`.

[6] `https://github.com/puffin444/poseknows`

Table 4. Evaluation of various architecture designs for a CVAE.

Encoder Conditioning	Decoder Conditioning	Diversity	Context
Concat (z_c)	Reparam. (z_c)	3.35	0.51
Concat (h_t)	Concat (h_t)	0.24	0.43
Concat (z_c)	Concat (z_c)	1.28	0.40
Concat (h_t)	Reparam (z_c)	3.12	0.54

Ablation Study on Different Means of Conditioning. Finally, we study various designs to condition the VAE encoder and decoder. To this end, we employ either a deterministic representation of the observation, h_t, or a stochastic one, z_c, as conditioning variable. We then evaluate the conditioning of the VAE encoder on either of these variables, which, as discussed before, can be done safely via concatenation as both the condition and the data carry useful information to compress the future motion into the latent space. By contrast, for the decoder, we compare the use of h_t or z_c via either concatenation or our extended reparameterization trick of Eq. 4. As evidenced by the first and fourth rows of Table 4, the use of a stochastic condition and of our extended reparameterization trick to condition the decoder are both critical to successfully achieve diversity while maintaining context. Furthermore, as shown in the fourth row, for the encoder, a deterministic condition works better than a stochastic one. Note that, depending on the application, one may prefer a deterministic condition (h_t) over a stochastic one if contextual information in the generated motion matters more than diversity, or a stochastic one (z_c) in the opposite situation. In this work, as we aim to generate contextually plausible and diverse motions, we use a deterministic condition for the encoder and a stochastic one for the decoder, which yields a good trade-off between diversity and context, as depicted by the fourth row in Table 4.

6. Conclusion

In this paper, we have studied the problem of conditionally generating diverse and contextually plausible 3D human motions. We have studied the reasons why existing CVAE-based methods fail to achieve these two goals jointly, and have addressed their weaknesses by forcing the sampling of the latent variable to explicitly depend on the observed motions. We have demonstrated that our approach can generate high quality diverse motions that carry out the semantics of the observed motion, such as the type of action performed by the person, without explicitly exploiting this information. These properties are highly beneficial to practical applications that can exploit motion predictions, such as action anticipation [2, 25, 26], pedestrian intention forecasting [3], and human tracking [1, 15, 27].

References

[1] Alexandre Alahi, Kratarth Goel, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, and Silvio Savarese. Social lstm: Human trajectory prediction in crowded spaces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 961–971, 2016.

[2] Mohammad Sadegh Aliakbarian, Fatemehsadat Saleh, Basura Fernando, Mathieu Salzmann, Lars Petersson, and Lars Andersson. Deep action-and context-aware sequence learning for activity recognition and anticipation. *arXiv preprint arXiv:1611.05520*, 2016.

[3] Mohammad Sadegh Aliakbarian, Fatemeh Sadat Saleh, Mathieu Salzmann, Basura Fernando, Lars Petersson, and Lars Andersson. Viena^2: A driving anticipation dataset. In *Asian Conference on Computer Vision*, pages 449–466. Springer, 2018.

[4] Sadegh Aliakbarian, Fatemeh Sadat Saleh, Mathieu Salzmann, Lars Petersson, and Stephen Gould. A stochastic conditioning scheme for diverse human motion prediction. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[5] Emad Barsoum, John Kender, and Zicheng Liu. Hp-gan: Probabilistic 3d human motion prediction via gan. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1418–1427, 2018.

[6] Samuel R Bowman, Luke Vilnis, Oriol Vinyals, Andrew M Dai, Rafal Jozefowicz, and Samy Bengio. Generating sentences from a continuous space. *arXiv preprint arXiv:1511.06349*, 2015.

[7] Judith Bütepage, Hedvig Kjellström, and Danica Kragic. Anticipating many futures: Online human motion prediction and generation for human-robot interaction. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1–9. IEEE, 2018.

[8] Katerina Fragkiadaki, Sergey Levine, Panna Felsen, and Jitendra Malik. Recurrent network models for human dynamics. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4346–4354, 2015.

[9] Partha Ghosh, Jie Song, Emre Aksan, and Otmar Hilliges. Learning human motion models for long-term predictions. In *2017 International Conference on 3D Vision (3DV)*, pages 458–466. IEEE, 2017.

[10] Liang-Yan Gui, Yu-Xiong Wang, Xiaodan Liang, and José MF Moura. Adversarial geometry-aware human motion prediction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 786–803, 2018.

[11] Liang-Yan Gui, Yu-Xiong Wang, Deva Ramanan, and José MF Moura. Few-shot human motion prediction via meta-learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 432–450, 2018.

[12] 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.

[13] Ashesh Jain, Amir R Zamir, Silvio Savarese, and Ashutosh Saxena. Structural-rnn: Deep learning on spatio-temporal graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5308–5317, 2016.

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

[15] Vineet Kosaraju, Amir Sadeghian, Roberto Martín-Martín, Ian Reid, Hamid Rezatofighi, and Silvio Savarese. Social-bigat: Multimodal trajectory forecasting using bicycle-gan and graph attention networks. In *Advances in Neural Information Processing Systems*, pages 137–146, 2019.

[16] Jogendra Nath Kundu, Maharshi Gor, and R Venkatesh Babu. Bihmp-gan: Bidirectional 3d human motion prediction gan. *arXiv preprint arXiv:1812.02591*, 2018.

[17] Chen Li, Zhen Zhang, Wee Sun Lee, and Gim Hee Lee. Convolutional sequence to sequence model for human dynamics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5226–5234, 2018.

[18] Chao Li, Qiaoyong Zhong, Di Xie, and Shiliang Pu. Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation. *arXiv preprint arXiv:1804.06055*, 2018.

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

[20] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[21] Wei Mao, Miaomiao Liu, Mathieu Salzmann, and Hongdong Li. Learning trajectory dependencies for human motion prediction. In *ICCV*, 2019.

[22] Julieta Martinez, Michael J Black, and Javier Romero. On human motion prediction using recurrent neural networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4674–4683. IEEE, 2017.

[23] Dario Pavllo, Christoph Feichtenhofer, Michael Auli, and David Grangier. Modeling human motion with quaternion-based neural networks. *arXiv preprint arXiv:1901.07677*, 2019.

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

[25] Cristian Rodriguez, Basura Fernando, and Hongdong Li. Action anticipation by predicting future dynamic images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0, 2018.

[26] Mohammad Sadegh Aliakbarian, Fatemeh Sadat Saleh, Mathieu Salzmann, Basura Fernando, Lars Petersson, and Lars Andersson. Encouraging lstms to anticipate actions very early. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[27] Fatemeh Saleh, Sadegh Aliakbarian, Mathieu Salzmann, and Stephen Gould. Artist: Autoregressive trajectory inpainting and scoring for tracking. *arXiv preprint arXiv:2004.07482*, 2020.

[28] Jacob Walker, Kenneth Marino, Abhinav Gupta, and Martial Hebert. The pose knows: Video forecasting by generating pose futures. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pages 3352–3361. IEEE, 2017.

[29] 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 *European Conference on Computer Vision*, pages 276–293. Springer, 2018.

[30] Dingdong Yang, Seunghoon Hong, Yunseok Jang, Tiangchen Zhao, and Honglak Lee. Diversity-sensitive conditional generative adversarial networks. In *International Conference on Learning Representations*, 2019.

[31] Ye Yuan and Kris Kitani. Dlow: Diversifying latent flows for diverse human motion prediction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[32] Weiyu Zhang, Menglong Zhu, and Konstantinos G Derpanis. From actemes to action: A strongly-supervised representation for detailed action understanding. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2248–2255, 2013.