BeLFusion: Latent Diffusion for Behavior-Driven Human Motion Prediction

German Barquero Sergio Escalera Cristina Palmero
Universitat de Barcelona and Computer Vision Center, Spain
{germanbarquero, sescalera}@ub.edu, crpalmec7@alumnes.ub.edu
https://barquerogerman.github.io/BeLFusion/

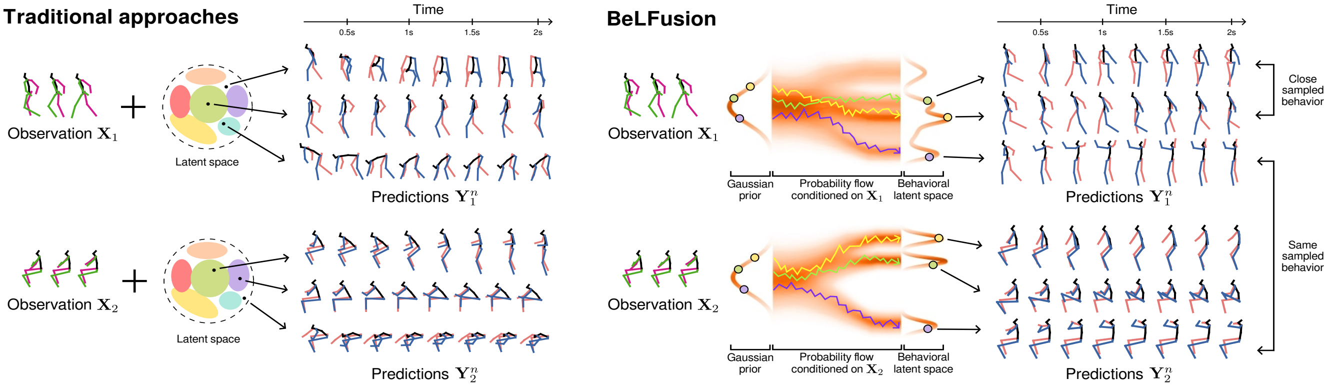
Traditional approaches
Time
0.5s
1s
1.5s
2s
Observation $\mathbf{X}_1$
Latent space
Predictions $\mathbf{Y}_1^n$
Observation $\mathbf{X}_2$
Latent space
Predictions $\mathbf{Y}_2^n$
BeLFusion
Time
0.5s
1s
1.5s
2s
Observation $\mathbf{X}_1$
Gaussian prior
Probability flow conditioned on $\mathbf{X}_1$
Behavioral latent space
Predictions $\mathbf{Y}_1^n$
Close sampled behavior
Same sampled behavior
Observation $\mathbf{X}_2$
Gaussian prior
Probability flow conditioned on $\mathbf{X}_2$
Behavioral latent space
Predictions $\mathbf{Y}_2^n$

Figure 1. Common approaches for stochastic human motion prediction use variational autoencoders to model a latent space. Then, the latent code sampled from it is fed to a decoder conditioned on the observation to generate the prediction. In this scenario, out-of-distribution samples or low KL regularizations lead to unrealistic generated sequences. For example, the first prediction for $\mathbf{X}_1$ shows an abrupt and unrealistic transition from walking to bending down. Instead, BeLFusion leverages latent diffusion models to conditionally sample from a behavioral space. Then, samples codes are decoded into predictions that coherently and smoothly transition into a wide range of behaviors.

Abstract

Stochastic human motion prediction (HMP) has generally been tackled with generative adversarial networks and variational autoencoders. Most prior works aim at predicting highly diverse motion in terms of the skeleton joints' dispersion. This has led to methods predicting fast and divergent movements, which are often unrealistic and incoherent with past motion. Such methods also neglect scenarios where anticipating diverse short-range behaviors with subtle joint displacements is important. To address these issues, we present BeLFusion, a model that, for the first time, leverages latent diffusion models in HMP to sample from a behavioral latent space where behavior is disentangled from pose and motion. Thanks to our behavior coupler, which is able to transfer sampled behavior to ongoing motion, BeLFusion's predictions display a variety of behaviors that are significantly more realistic, and coherent with past motion than the state of the art. To support it, we introduce two metrics, the Area of the Cumulative Motion Distribution, and the Average Pairwise Distance Error, which are correlated to realism according to a qualitative study (126 participants). Finally, we prove BeLFusion's generalization power in a new cross-dataset scenario for stochastic HMP.

1. Introduction

Humans excel at inattentively predicting others' actions and movements. This is key to effectively engaging in social interactions, driving a car, or walking across a crowd. Replicating this ability is imperative in many applications like assistive robots, virtual avatars, or autonomous cars [3, 56]. Many prior works conceive Human Motion Prediction (HMP) from a deterministic point of view, forecasting a single sequence of body poses, or *motion*, given past poses, usually represented with skeleton joints [41]. However, humans are spontaneous and unpredictable creatures by nature, and this deterministic interpretation does not fit contexts where anticipating all possible outcomes is crucial. Accordingly, recent works have attempted to predict the whole distribution of possible future motions (i.e., a *multimodal* distribution) given a short observed motion sequence. We refer to this reformulation as stochastic HMP.

Most prior stochastic works focus on predicting a highly *diverse* distribution of motions. Such diversity has been traditionally defined and evaluated in the coordinate space [70, 18, 45, 58, 42]. This definition biases research toward models that generate fast transitions into very different poses coordinate-wise (see Fig. 1). Although there are scenarios where predicting low-speed diverse motion is important, this is discouraged by prior techniques. For example, in assistive robotics, anticipating *behaviors* (i.e., actions) like whether the interlocutor is about to shake your hand or scratch their head might be crucial for preparing the robot's actuators on time [5, 51]. In a surveillance scenario, a foreseen harmful behavior might not differ much from a well-meaning one when considering only the poses along the motion sequence. We argue that this behavioral perspective is paramount to build next-generation stochastic HMP models. Moreover, results from prior diversity-centric works [45, 18] often suffer from a trade-off that has been persistently overlooked: predicted motion does not look coherent with respect to the latest observed motion. The strong diversity regularization techniques employed often produce abrupt speed changes or direction discontinuities. We argue that consistency with the immediate past is a requirement for prediction plausibility.

To tackle these issues, we present BeLFusion (Fig. 1). By constructing a latent space that disentangles behavior from poses and motion, diversity is no longer limited to the traditional coordinate-based perspective. Instead, diversity is viewed through a behavioral lens, allowing both short- (e.g., hand-waving or smoking) and long-range motions (e.g., standing up or sitting down) to be equally encouraged and represented in the space. Our *behavior coupler* ensures the predicted behavior is decoded into a smooth and plausible continuation of any ongoing motion. Thus, our predicted motions look more realistic and coherent with the near past than alternatives, which we assess through quantitative and qualitative analyses. In addition, BeLFusion is the first approach that exploits conditional latent diffusion models (LDM) [63, 55] for stochastic HMP, achieving state-of-the-art performance. By combining the exceptional capabilities of LDMs to model conditional distributions with the convenient inductive biases of recurrent neural networks (RNNs) for motion modeling [41], BeLFusion represents a powerful method for stochastic HMP.

To summarize, our main contributions are: (1) We propose BeLFusion, a method that generates predictions that are significantly more realistic and coherent with the near past than prior works, while achieving state-of-the-art accuracy on Human 3.6M [32] and AMASS [43] datasets. (2) We improve and extend the usual evaluation pipeline for stochastic HMP. For the first time in this task, a *cross-dataset evaluation* is conducted to assess the robustness against domain shifts, where the superior generalization capabilities of our method are clearly depicted. This setup, built with AMASS [43] dataset, showcases a broad range of actions performed by more than 400 subjects. (3) We propose two new metrics that provide complementary insights on the statistical similarities between a) the predicted and the dataset averaged absolute motion, and b) the predicted and the intrinsic dataset diversity. We show that they are significantly correlated to our definition of realism.

2. Related work

2.1. Human motion prediction

Deterministic scenario. Prior works on HMP define the problem as regressing a single future sequence of skeleton joints matching the immediate past, or *observed* motion. This regression is often modeled with RNNs [22, 33, 47, 26, 52, 39] or Transformers [2, 11, 48]. Graph Convolutional Networks might be included as intermediate layers to model the dependencies among joints [37, 46, 17, 36]. Some methods leverage Temporal Convolutional Networks [35, 49] or a simple Multi-Layer Perceptron [28] to predict fixed-size sequences, achieving high performance. Recently, some works claimed the benefits of modeling sequences in the frequency space [11, 46, 44]. However, none of these solutions can model multimodal distributions of future motions.

Stochastic scenario. To fill this gap, other methods that predict multiple futures for each observed sequence were proposed. Most of them use a generative approach to model the distribution of possible futures. Most popular generative models for HMP are generative adversarial networks (GANs) [7, 34] and variational autoencoders (VAEs) [64, 68, 12, 45]. These methods often include diversity-promoting losses in order to predict a high variety of motions [45], or incorporate explicit techniques for diverse sampling [70, 18, 67]. This diversity is computed with the raw coordinates of the predicted poses. We argue that, as a result, the race for diversity has promoted motions deriving to extremely varied poses very early in the prediction. Most of these predictions are neither realistic nor plausible in the context of the observed motion. Also, prior works neglect situations where a diversity of behaviors, which can sometimes be subtle, is important. We address this by *implicitly* encouraging such diversity in a behavioral latent space.

Semantic human motion prediction. Few works have attempted to leverage semantically meaningful latent spaces for stochastic HMP [68, 38, 24]. For example, [24] exploits disentangled motion representations for each part of the body to control the HMP. [68] adds a sampled latent code to the observed encoding to transform it into a prediction encoding. This inductive bias helps the network disentangle a motion code from the observed poses. However, the strong assumption that a simple arithmetic operation can map both sequences limits the expressiveness of the model. Although not specifically focused on HMP, [10] proposes an adversarial framework to disentangle a behavioral encoding from a sequence of poses. The extracted behavior can then be transferred to any initial pose. In this paper, we propose a generalization of such framework to transfer behavior to ongoing movements. BeLFusion exploits this disentangle-

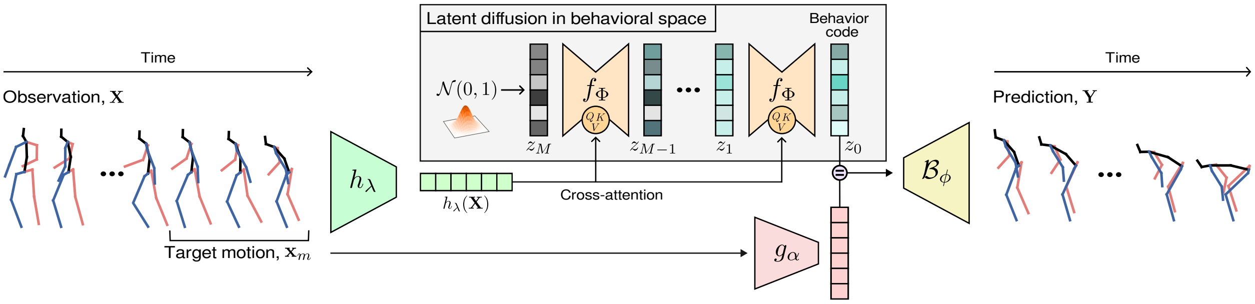

Figure 2. BeLFusion's architecture. A latent diffusion model conditioned on an encoding of the observation, $h_\lambda(\mathbf{X})$, progressively denoises a sample from a zero-mean unit variance multivariate normal distribution into a behavior code. Then, the behavior coupler $\mathcal{B}_\phi$ decodes the prediction by transferring the sampled behavior to the target motion, $\mathbf{x}_m$. In our implementation, f_Φ is a conditional U-Net with cross-attention, g_α is a dense layer, and h_λ, and $\mathcal{B}_\phi$ are one-layer recurrent neural networks.

ment to improve the behavioral coverage of HMP.

2.2. Diffusion models

Denoising diffusion probabilistic models aim at learning to reverse a Markov chain of M diffusion steps (usually $M > 100$) that slowly adds random noise to the target data samples [59, 31]. For conditional generation, a common strategy consists in applying cross-attention to the conditioning signal at each denoising step [19]. Diffusion models have achieved impressive results in fields like video generation, inpainting, or anomaly detection [69]. In a more similar context, [54, 62] use diffusion models for time series forecasting. [25] recently presented a diffusion model for trajectory prediction that controls the prediction uncertainty by shortening the denoising chain. A few concurrent works have explored them for HMP [57, 65, 14, 1, 16].

However, diffusion models have an expensive trade-off: extremely slow inference due to the large number of denoising steps required. Latent diffusion models (LDM) accelerate the sampling by applying diffusion to a low-resolution latent space learned by a VAE [63, 55]. Thanks to the KL regularization, the learned latent space is built close to a normal distribution. As a result, the length of the chain that destroys the latent codes can be greatly reduced, and reversed much faster. In this work, we present the first approach that leverages LDM for stochastic HMP, achieving state-of-the-art performance in terms of accuracy and realism.

3. Methodology

In this section, we first characterize the HMP problem (Sec. 3.1). Then, we present a straightforward adaptation of conditional LDMs to HMP (Sec. 3.2). Finally, we describe BeLFusion's keystones (Fig. 2): our behavioral latent space, the behavioral LDM, and its training losses (Sec. 3.3).

3.1. Problem definition

The goal in HMP consists in, given an observed sequence of B poses (*observation window*), predicting the following T poses (*prediction window*). In stochastic HMP, N prediction windows are predicted for each observation window. Accordingly, we define the set of poses in the observation and prediction windows as $\mathbf{X}=\{p_{t-B}, ..., p_{t-2}, p_{t-1}\}$ and $\mathbf{Y}^i=\{p^i_t, p^i_{t+1}, ..., p^i_{t+T-1}\}$, where $i\in\{1, ..., N\}$[1], and $p^i_t\in\mathbb{R}^d$ are the coordinates of the human joints at timestep t.

3.2. Motion latent diffusion

Here, we define a direct adaptation of LDM to HMP. First, a VAE is trained so that an encoder $\mathcal{E}$ transforms fixed-length target sequences of T poses, $\mathbf{Y}$, into a low-dimensional latent space $V \subset \mathbb{R}^v$. Samples $z \in V$ can be drawn and mapped back to the coordinate space with a decoder $\mathcal{D}$. Then, an LDM conditioned on $\mathbf{X}$ is trained to predict the corresponding latent vector $z = \mathcal{E}(\mathbf{Y}) \in V$[2]. The generative HMP problem is formulated as follows:

$$P(\mathbf{Y}|\mathbf{X}) = P(\mathbf{Y}, z|\mathbf{X}) = P(\mathbf{Y}|z, \mathbf{X})P(z|\mathbf{X}). \quad (1)$$

The first equality holds because $\mathbf{Y}$ is a deterministic mapping from the latent code z. Then, sampling from the true conditional distribution $P(\mathbf{Y}|\mathbf{X})$ is equivalent to sampling z from $P(z|\mathbf{X})$ and decoding $\mathbf{Y}$ with $\mathcal{D}$.

LDMs are typically trained to predict the perturbation $\epsilon_t = f_\Phi(z_t, t, h_\lambda(\mathbf{X}))$ of the diffused latent code z_t at each timestep t, where $h_\lambda(\mathbf{X})$ is the encoded conditioning observation. Once trained, the network f_Φ can reverse the diffusion Markov chain of length M and infer z from a random sample $z_M \sim \mathcal{N}(0, 1)$. Instead, we choose to use a more convenient parameterization so that $z_0 = f_\Phi(z_t, t, h_\lambda(\mathbf{X}))$ [66, 40]. With this, an approximation of z is predicted in every denoising step z_0, and used to sample the input of the next denoising step z_{t-1}, by diffusing it $t-1$ times. We use $q(z_{t-1}|z_0)$ to refer to this diffusion process. With this parameterization, the LDM loss (or *latent* loss) becomes:

$$\mathcal{L}_{lat}(\mathbf{X}, \mathbf{Y}) = \sum_{t=1}^{T} \mathop{\mathbb{E}}_{q(z_t|z_0)} \|f_\Phi(z_t, t, h_\lambda(\mathbf{X})) - \mathcal{E}(\mathbf{Y})\|_1. \quad (2)$$

[1] A sampled prediction $\mathbf{Y}^i$ is hereafter referred as $\mathbf{Y}$ for intelligibility.
[2] For simplicity, we use $\mathcal{E}(\mathbf{Y})$ to refer to the expected value of $\mathcal{E}(z|\mathbf{Y})$.

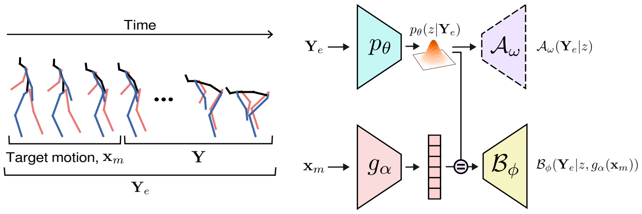
Time
Target motion, $\mathbf{x}_m$
$\mathbf{Y}$
$\mathbf{Y}_e$
$\mathbf{Y}_e$
p_θ
$p_\theta(z|\mathbf{Y}_e)$
$\mathcal{A}_\omega$
$\mathcal{A}_\omega(\mathbf{Y}_e|z)$
$\mathbf{x}_m$
g_α
$\mathcal{B}_\phi$
$\mathcal{B}_\phi(\mathbf{Y}_e|z, g_\alpha(\mathbf{x}_m))$

Figure 3. Framework for behavioral disentanglement. By adversarially training the auxiliary generator, $\mathcal{A}_\omega$, against the behavior coupler, $\mathcal{B}_\phi$, the behavior encoder, p_θ, learns to generate a disentangled latent space of behaviors, $p_\theta(z|\mathbf{Y}_e)$. At inference, $\mathcal{B}_\phi$ decodes a sequence of poses that smoothly transitions from *any* target motion $\mathbf{x}_m$ to performing the behavior extracted from $\mathbf{Y}$.

Having an approximate prediction at any denoising step allows us to 1) apply regularization in the coordinates space (Sec. 3.3), and 2) stop the inference at any step and still have a meaningful prediction (Sec. 4.3).

3.3. Behavioral latent diffusion

In HMP, small discontinuities between the last observed pose and the first predicted pose can look unrealistic. Thus, the LDM (Sec. 3.2) must be highly accurate in matching the coordinates of the first predicted pose to the last observed pose. An alternative consists in autoencoding the offsets between poses in consecutive frames. Although this strategy minimizes the risk of discontinuities in the first frame, motion speed or direction discontinuities are still bothersome.

Our proposed architecture, **Be**havioral **L**atent dif**Fusion**, or BeLFusion, solves both problems. It reduces the latent space complexity by relegating the adaption of the motion speed and direction to the decoder. It does so by learning a representation of posture-independent human dynamics: a *behavioral representation*. In this framework, the decoder learns to transfer any behavior to an ongoing motion by building a coherent and smooth transition. Here, we first describe how the behavioral latent space is learned, and then detail the BeLFusion pipeline for behavior-driven HMP.

Behavioral Latent Space (BLS). The behavioral representation learning is inspired by [10], which presents a framework to disentangle behavior from motion. Once disentangled, such behavior can be transferred to any static initial pose. We propose an extension of their work to a general and challenging scenario: behavioral transference to ongoing motions. The proposed architecture is shown in Fig. 3.

First, we define the last C observed poses as the *target motion*, $\mathbf{x}_m = \{p_{t-C}, ..., p_{t-2}, p_{t-1}\} \subset \mathbf{X}$, and $\mathbf{Y}_e = \mathbf{x}_m \cup \mathbf{Y}$. $\mathbf{x}_m$ informs us about the motion speed and direction of the last poses of $\mathbf{X}$, which should be coherent with $\mathbf{Y}$[3]. The goal is to disentangle the behavior from the motion and poses in $\mathbf{Y}_e$. To do so, we adversarially train two generators, the behavior coupler $\mathcal{B}_\phi$, and the auxiliary generator $\mathcal{A}_\omega$, such that a behavior encoder p_θ learns to generate a disentangled latent space of behaviors $p_\theta(z|\mathbf{Y}_e)$. Both $\mathcal{B}_\phi$ and $\mathcal{A}_\omega$ have access to such latent space, but $\mathcal{B}_\phi$ is additionally fed with an encoding of the target motion, $g_\alpha(\mathbf{x}_m)$. During adversarial training, $\mathcal{A}_\omega$ aims at preventing p_θ from encoding pose and motion information by trying to reconstruct poses of $\mathbf{Y}_e$ directly from $p_\theta(z|\mathbf{Y}_e)$. This training allows $\mathcal{B}_\phi$ to decode a sequence of poses that smoothly transitions from $\mathbf{x}_m$ to perform the behavior extracted from $\mathbf{Y}_e$. At inference time, $\mathcal{A}_\omega$ is discarded.

More concretely, the disentanglement is learned by alternating two objectives at each training iteration. The first objective, which optimizes the parameters ω of the auxiliary generator, forces it to predict $\mathbf{Y}_e$ given the latent code z:

$$\max_\omega \mathcal{L}_{\text{aux}} = \max_\omega \mathbb{E}_{p_\theta(z|\mathbf{Y}_e)}(\log \mathcal{A}_\omega(\mathbf{Y}_e|z)). \tag{3}$$

The second objective acts on the parameters of the target motion encoder, α, the behavior encoder, θ, and the behavior coupler, ϕ. It forces $\mathcal{B}_\phi$ to learn an accurate $\mathbf{Y}_e$ reconstruction through the construction of a normally distributed intermediate latent space:

$$\begin{aligned}\max_{\alpha,\theta,\phi} \mathcal{L}_{\text{main}} = \max_{\alpha,\theta,\phi} \mathbb{E}_{p_\theta(z|\mathbf{Y}_e)}[\log \mathcal{B}_\phi(\mathbf{Y}_e|z, g_\alpha(\mathbf{x}_m))] \\ -D_{\text{KL}}(p_\theta(z|\mathbf{Y}_e)||p(z))) - \mathcal{L}_{\text{aux}}.\end{aligned} \tag{4}$$

Note that the parameters ω are not optimized when training with Eq. 4, and α, θ, ϕ with Eq. 3. The prior $p(z)$ is a multivariate $\mathcal{N}(0, I)$. The inclusion of $-\mathcal{L}_{\text{aux}}$ in Eq. 4 penalizes any accurate reconstruction of $\mathbf{Y}_e$ through $\mathcal{A}_\omega$, forcing p_θ to filter any postural information out. Since $\mathcal{B}_\phi$ has access to the target posture and motion provided by $\mathbf{x}_m$, it only needs $p_\theta(z|\mathbf{Y}_e)$ to encode the behavioral dynamics. One could argue that a valid alternative strategy for p_θ would consist in disentangling motion from postures. However, motion dynamics can still be used to extract a good pose approximation. See supp. material Sec. C for more details and visual examples of behavioral transference to several motions $\mathbf{x}_m$.

Behavior-driven HMP. BeLFusion's goal is to sample the appropriate behavior code given the observation $\mathbf{X}$, see Fig. 2. To that end, a conditional LDM is trained to optimize $\mathcal{L}_{lat}(\mathbf{X}, \mathbf{Y}_e)$ (Eq. 2), with $\mathcal{E} = p_\theta$, so that it learns to predict the behavioral encoding of $\mathbf{Y}_e$: the expected value of $p_\theta(z|\mathbf{Y}_e)$. Then, the behavior coupler, $\mathcal{B}_\phi$, transfers the predicted behavior to the target motion, $\mathbf{x}_m$, to reconstruct the poses of the prediction. However, the reconstruction of $\mathcal{B}_\phi$ is also conditioned on $\mathbf{x}_m$. Such dependency cannot be modeled by the $\mathcal{L}_{lat}$ objective alone. Thanks to our parameterization (Sec. 3.2), we can also use the traditional MSE loss in the reconstruction space:

$$\begin{aligned}\mathcal{L}_{rec}(\mathbf{X}, \mathbf{Y}_e) = \sum_{t=1}^{T} \mathop{\mathbb{E}}_{q(z_t|z_0)} \|\mathcal{B}_\phi(f_\Phi(z_t, t, h_\lambda(\mathbf{X})), g_\alpha(\mathbf{x}_m)) \\ -\mathcal{B}_\phi(p_\theta(\mathbf{Y}_e), g_\alpha(\mathbf{x}_m))\|_2.\end{aligned} \tag{5}$$

The second term of Eq. 5 is the reconstructed $\mathbf{Y}_e$. Optimizing the objective within the solutions space upper

[3] In practice, decoding $\mathbf{x}_m$ also helped stabilize the BLS training.

bounded by $\mathcal{B}_\phi$'s reconstruction capabilities helps stabilize the training. Note that only the future poses $\mathbf{Y} \subset \mathbf{Y}_e$ form the prediction. The observation encoder, h_λ, is pretrained in an autoencoding framework that reconstructs $\mathbf{X}$. We found experimentally that h_λ does not benefit from further training, so its parameters λ are frozen when training the LDM. The target motion encoder, g_α, and the behavior coupler, $\mathcal{B}_\phi$, are also pretrained as described before and kept frozen. f_Φ is conditioned on $h_\lambda(\mathbf{X})$ with cross-attention.

Implicit diversity loss. Although training BeLFusion with Eqs. 2 and 5 leads to accurate predictions, their diversity is poor. We argue that this is due to the strong regularization of both losses. Similarly to [21, 29], we propose to relax them by sampling k predictions at each training iteration and only backpropagating the gradients through the two predictions that separately minimize the latent or the reconstructed loss (further discussion in supp. material Sec. D.2):

$$\min_k \mathcal{L}_{lat}(\mathbf{X}, \mathbf{Y}_e^k) + \lambda \min_k \mathcal{L}_{rec}(\mathbf{X}, \mathbf{Y}_e^k), \qquad (6)$$

where λ controls the trade-off between the latent and the reconstruction errors. Regularization relaxation usually leads to out-of-distribution predictions [45]. This is often solved by employing additional complex techniques like pose priors, or bone-length losses that regularize the other predictions [45, 9]. BeLFusion can dispense with it due to mainly two reasons: 1) Denoising diffusion models are capable of faithfully capturing a greater breadth of the training distribution than GANs or VAEs [19]; 2) The variational training of the behavior coupler makes it more robust to errors in the predicted behavior code.

4. Experimental evaluation

Our experimental evaluation is tailored toward two objectives. First, we aim at proving BeLFusion's generalization capabilities for both seen and unseen scenarios. For the latter, we propose a challenging cross-dataset evaluation setup. Second, we want to demonstrate the superiority of our model with regard to the realism of its predictions compared to state-of-the-art approaches. In this sense, we propose two metrics and perform a qualitative study.

4.1. Evaluation setup

Datasets. We evaluate our proposed methodology on Human3.6M [32] (H36M), and AMASS [43]. H36M consists of clips where 11 subjects perform 15 actions, totaling 3.6M frames recorded at 50 Hz, with action class labels available. We use the splits proposed by [70] and adopted by most subsequent works [45, 58, 42, 18] (16 joints). Accordingly, 0.5s (25 frames) are used to predict the following 2s (100 frames). AMASS is a large-scale dataset that, as of today, unifies 24 extremely varied datasets with a common joints configuration, with a total of 9M frames when downsampled to 60Hz. Whereas latest deterministic HMP approaches already include a within-dataset AMASS configuration in their evaluation protocol [44, 2, 49], the dataset remains unexplored in the stochastic context yet. To determine whether state-of-the-art methods can generalize their learned motion predictive capabilities to other contexts (i.e., other datasets), we propose a new cross-dataset evaluation protocol with AMASS. The training, validation, and test sets include 11, 4, and 7 datasets, and 406, 33, and 54 subjects (21 joints), respectively. We set the observation and prediction windows to 0.5s and 2s (30 and 120 frames after downsampling), respectively. AMASS does not provide action labels. See supp. material Sec. B for more details.

Baselines. We include the zero-velocity baseline, which has been proven very competitive in HMP [47, 6], and a version of our model that replaces the LDM with a GAN, BeGAN [23]. We train three versions with $k = 1, 5, 50$. We also compare against state-of-the-art methods for stochastic HMP (referenced in Tab. 1). For H36M, we took all evaluation values from their respective works. For AMASS, we retrained state-of-the-art methods with publicly available code that showed competitive performance for H36M.

Implementation details. We trained BeLFusion with N=50, M=10, k=50, a U-Net with cross-attention [19] as f_Φ, one-layer RNNs as h_λ, and $\mathcal{B}_\phi$, and a dense layer as g_α. For H36M, λ=5, and for AMASS, λ=1. At inference, we use an exponential moving average of the trained model with a decay of 0.999. Sampling was conducted with a DDIM sampler [60]. As explained in Sec. 3.2, our implementation of LDM can be early-stopped at any step of the chain of length M and still have access to an approximation of the behavioral latent code. Thus, we also include BeLFusion's results when inference is early-stopped right after the first denoising step (i.e., x10 faster): BeLFusion_D. Further details are included in the supp. material Sec. A.

4.2. Evaluation metrics

To compare BeLFusion with prior works, we follow the well-established evaluation pipeline proposed in [70]. The Average and the Final Displacement Error metrics (ADE, and FDE, respectively) quantify the error on the most similar prediction compared to the ground truth. While the ADE averages the error along all timesteps, the FDE only does it for the last predicted frame. Their multimodal versions for stochastic HMP, MMADE and MMFDE, compare all predicted futures with the multimodal ground truth of the observation. To obtain the latter, each observation window $\mathbf{X}$ is grouped with other observations $\mathbf{X}_i$ with a similar last observed pose in terms of L2 distance. The corresponding prediction windows $\mathbf{Y}_i$ form the *multimodal ground truth* of $\mathbf{X}$. The Average Pairwise Distance (APD) quantifies the diversity by computing the L2 distance among all pairs of predicted poses at each timestep. Following [27, 53, 18, 9],

	Human3.6M [32]								AMASS [43]						
	APD	APDE	ADE	FDE	MMADE	MMFDE	CMD	FID*	APD	APDE	ADE	FDE	MMADE	MMFDE	CMD
Zero-Velocity	0.000	8.079	0.597	0.884	0.683	0.909	22.812	0.606	0.000	9.292	0.755	0.992	0.814	1.015	39.262
BeGAN k=1	0.675	7.411	0.494	0.729	0.605	0.769	12.082	0.542	0.717	8.595	0.643	0.834	0.688	0.843	24.483
BeGAN k=5	2.759	5.335	0.495	0.697	0.584	0.718	13.973	0.578	5.643	4.043	0.631	0.788	0.667	0.787	24.034
BeGAN k=50	6.230	2.200	0.470	0.637	0.561	0.661	8.406	0.569	7.234	2.548	0.613	0.717	0.650	0.720	22.625
HP-GAN [7]	7.214	-	0.858	0.867	0.847	0.858	-	-	-	-	-	-	-	-	-
DSF [71]	9.330	-	0.493	0.592	0.550	0.599	-	-	-	-	-	-	-	-	-
DeLiGAN [30]	6.509	-	0.483	0.534	0.520	0.545	-	-	-	-	-	-	-	-	-
GMVAE [20]	6.769	-	0.461	0.555	0.524	0.566	-	-	-	-	-	-	-	-	-
TPK [64]	6.723	1.906	0.461	0.560	0.522	0.569	6.326	0.538	9.283	2.265	0.656	0.675	0.658	0.674	17.127
MT-VAE [68]	0.403	-	0.457	0.595	0.716	0.883	-	-	-	-	-	-	-	-	-
BoM [8]	6.265	-	0.448	0.533	0.514	0.544	-	-	-	-	-	-	-	-	-
DLow [70]	11.741	3.781	0.425	0.518	0.495	0.531	**4.927**	1.255	13.170	4.243	0.590	0.612	0.618	0.617	**15.185**
MultiObj [42]	14.240	-	0.414	0.516	-	-	-	-	-	-	-	-	-	-	-
GSPS [45]	14.757	6.749	0.389	0.496	0.476	0.525	10.758	2.103	12.465	4.678	0.563	0.613	0.609	0.633	18.404
Motron [58]	7.168	2.583	0.375	0.488	0.509	0.539	40.796	13.743	-	-	-	-	-	-	-
DivSamp [18]	**15.310**	7.479	0.370	0.485	0.475	0.516	11.692	2.083	**24.724**	15.837	0.564	0.647	0.623	0.667	50.239
BeLFusion_D	5.777	2.571	**0.367**	**0.472**	**0.469**	**0.506**	8.508	0.255	7.458	2.663	**0.508**	0.567	**0.564**	0.591	19.497
BeLFusion	7.602	**1.662**	0.372	0.474	0.473	0.507	5.988	**0.209**	9.376	**1.977**	0.513	**0.560**	0.569	**0.585**	16.995

Table 1. Comparison of BeLFusion_D (single denoising step) and BeLFusion (all denoising steps) with state-of-the-art methods for stochastic human motion prediction on Human3.6M and AMASS datasets. Bold and underlined results correspond to the best and second-best results, respectively. Lower is better for all metrics except APD. *Only showed for Human3.6M due to lack of class labels for AMASS.

we also include the Fréchet Inception Distance (FID), which leverages the output of the last layer of a pretrained action classifier to quantify the similarity between the distributions of predicted and ground truth motions.

Area of the Cumulative Motion Distribution (CMD). The plausibility and realism of human motion are difficult to assess quantitatively. However, some metrics can provide an intuition of when a set of predicted motions are not plausible. For example, consistently predicting high-speed movements given a context where the person was standing still might be plausible but does not represent a statistically coherent distribution of possible futures. We argue that prior works have persistently ignored this. We propose a simple complementary metric: the area under the cumulative motion distribution. First, we compute the average of the L2 distance between the joint coordinates in two consecutive frames (displacement) across the whole test set, $\bar{M}$. Then, for each frame t of all predicted motions, we compute the average displacement M_t. Then:

$$\text{CMD} = \sum_{i=1}^{T-1} \sum_{t=1}^{i} \|M_t - \bar{M}\|_1 = \sum_{t=1}^{T-1} (T-t) \|M_t - \bar{M}\|_1. \quad (7)$$

Our choice to accumulate the distribution is motivated by the fact that early motion irregularities in the predictions impact the quality of the remaining sequence. Intuitively, this metric gives an idea of how the predicted average displacement per frame deviates from the expected one. However, the expected average displacement could arguably differ among actions and datasets. To account for this, we compute the total CMD as the weighted average of the CMD for each H36M action, or each AMASS test dataset, weighted by the action or dataset relative frequency.

Average Pairwise Distance Error (APDE). There are many elements that condition the distribution of future movements and, therefore, the appropriate motion diversity levels. To analyze to which extent the diversity is properly modeled, we introduce the average pairwise distance error. We define it as the absolute error between the APD of the multimodal ground truth and the APD of the predicted samples. Samples without any multimodal ground truth are dismissed. See supp. material Fig. E for a visual illustration.

4.3. Results

Comparison with the state of the art. As shown in Tab. 1, BeLFusion achieves state-of-the-art performance in all accuracy metrics for both datasets. The improvements are especially important in the cross-dataset AMASS configuration, proving its superior robustness against domain shifts. We hypothesize that such good generalization capabilities are due to 1) the exhaustive coverage of behaviors modeled in the disentangled latent space, and 2) the potential of LDMs to model the conditional distribution of future behaviors. In fact, after a single denoising step, our model already achieves state-of-the-art uni- and multimodal ADE and FDE (BeLFusion_D) in return for less diversity and realism. When going through all denoising steps (BeLFusion), our method also excels at realism-related metrics like CMD and FID (see the *Implicit diversity* section below for a detailed discussion on the topic). By contrast, Fig. 5 shows that predictions from GSPS and DivSamp consistently accelerate at the beginning, presumably toward divergent poses that promote high diversity values. As a result, they yield high CMD values, especially for H36M. The predictions from methods that leverage transformations in the frequency space freeze at the very long-term horizon. Motron's high CMD depicts an important jitter in its predictions, missed by all other metrics. BeLFusion's low APDE highlights its good ability to adapt to the observed context.

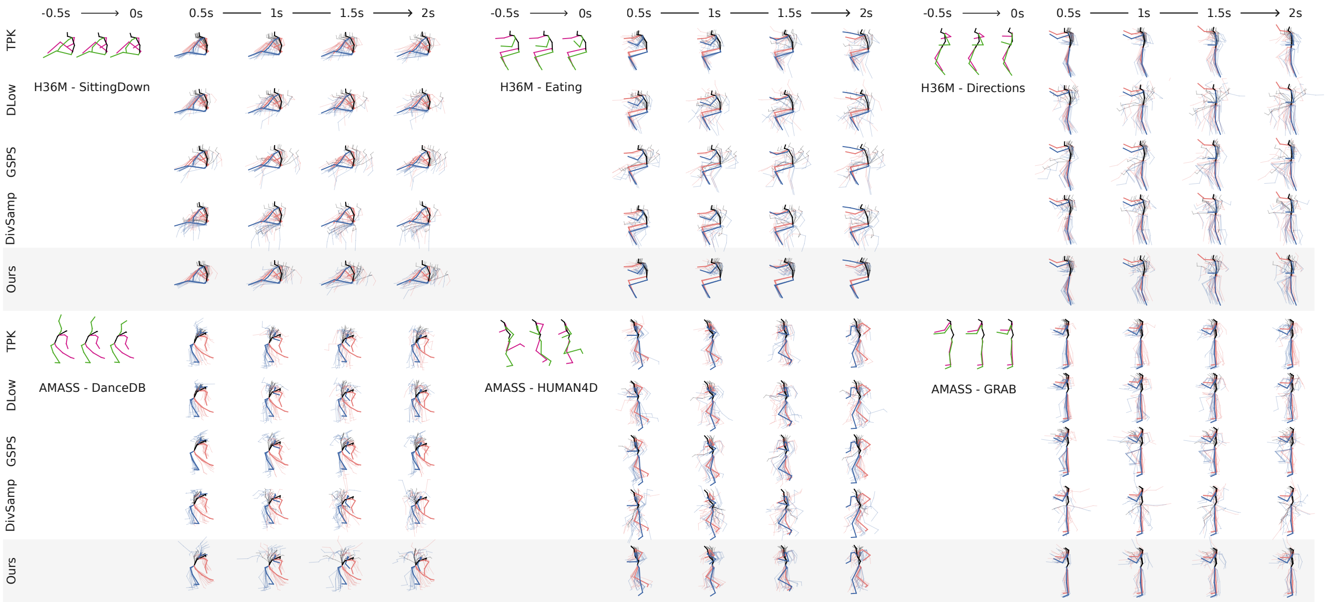

Figure 4. Qualitative results show the adaption of BeLFusion's diversity to the observation context in both within- (H36M, top) and cross-dataset (AMASS, bottom). At each future timestep, 10 predicted samples are superimposed below the thicker ground truth.

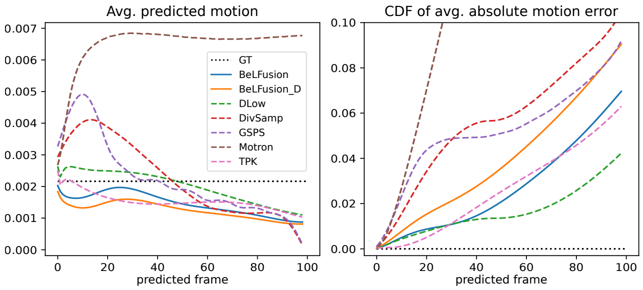

Figure 5. Left. Average predicted motion of state-of-the-art methods in H36M. Right. Cumulative distribution function (CDF) of the weighted absolute errors in the left with respect to the ground truth. CMD is the area under this curve.

This is achieved thanks to 1) the pretrained encoding of the whole observation window, and 2) the behavior coupling to the *target motion*. In contrast, higher APDE values of GSPS and DivSamp are caused by their tendency toward predicting movements more diverse than those present in the dataset. Action- (H36M) and dataset-wise (AMASS) results are included in supp. material Sec. D.1.

Fig. 4 displays 10 overlaid predictions over time for three actions from H36M (sitting down, eating, and giving directions), and three datasets from AMASS (DanceDB [50], HUMAN4D [13], and GRAB [61]). The purpose of this visualization is to confirm the observations made by the CMD and APDE metrics. First, the acceleration of GSPS and DivSamp at the beginning of the prediction leads to extreme poses very fast, abruptly transitioning from the observed motion. Second, it shows the capacity of BeLFusion to adapt the diversity predicted to the context. For example, the diversity of motion predicted while eating focuses on the arms, and does not include holistic extreme poses. Interestingly, when just sitting, the predictions include a wider range of full-body movements like laying down, or bending over. A similar context fitting is observed in the AMASS cross-dataset scenario. For instance, BeLFusion correctly identifies that the diversity must target the upper body in the GRAB dataset, or the arms while doing a dance step. Examples *in motion* can be found in supp. material Sec. E.

Ablation study. Here, we analyze the effect of each of our contributions in the final model quantitatively. This includes the contributions of $\mathcal{L}_{lat}$ and $\mathcal{L}_{rec}$, and the benefits of disentangling behavior from motion in the latent space construction. Results are summarized in Tab. 2. Although training is stable and losses decrease similarly in all cases, solely considering the loss at the coordinate space ($\mathcal{L}_{rec}$) leads to poor generalization capabilities. This is especially noticeable in the cross-dataset scenario, where models with both latent space constructions are the least accurate among all loss configurations. We observe that the latent loss ($\mathcal{L}_{lat}$) boosts the metrics in both datasets, and can be further enhanced when considered along with the reconstruction loss. Overall, the BLS construction benefits all loss configurations in terms of accuracy on both datasets, proving it a very promising strategy to be further explored in HMP.

Implicit diversity. As explained in Sec. 3.3, the parameter k regulates the *relaxation* of the training loss (Eq. 6) on BeLFusion. Fig. 6 shows how metrics behave when 1) tuning k, and 2) moving forward in the reverse diffusion chain (i.e., progressively applying denoising steps). In general, increasing k enhances the samples' diversity, accuracy, and realism. For $k \leq 5$, going through the whole chain of denoising steps boosts accuracy. However, for $k > 5$, further denoising only boosts diversity- and realism-wise metrics

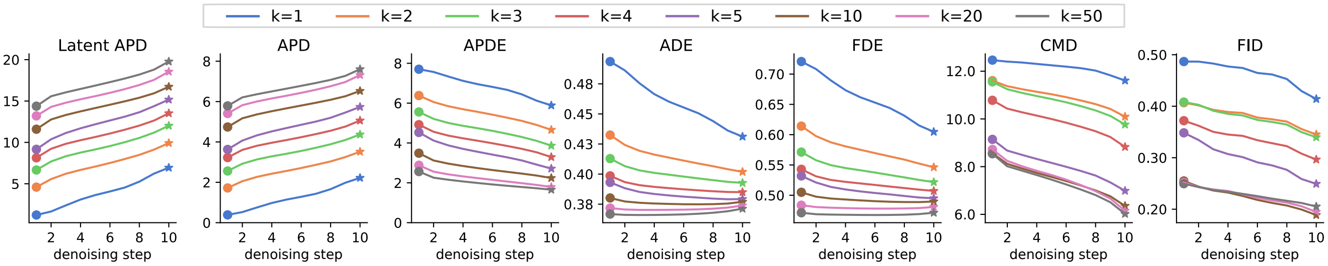

Figure 6. Evolution of evaluation metrics (y-axis) along denoising steps (x-axis) at inference time, for different values of k. Early stopping can be applied at any time, between the first (●) and the last step ($\star$). Accuracy saturates at $k = 50$, with gains for all metrics when increasing k, especially for diversity (APD). Qualitative metrics (CMD, FID) decrease after each denoising step across all k values.

			Human3.6M [32]						AMASS [43]				
BLS	$\mathcal{L}_{lat}$	$\mathcal{L}_{rec}$	APD	APDE	ADE	FDE	CMD	FID	APD	APDE	ADE	FDE	CMD
		✓	**7.622**	**1.276**	0.510	0.795	5.110	2.530	**10.788**	3.032	0.697	0.881	**16.628**
✓		✓	6.169	2.240	0.386	0.505	8.432	0.475	9.555	2.216	0.593	0.685	17.036
	✓		7.475	1.773	0.388	0.490	**4.643**	**0.177**	8.688	2.079	0.528	0.572	18.429
✓	✓		6.760	1.974	0.377	0.485	6.615	0.233	8.885	2.009	0.516	0.565	17.576
	✓	✓	7.301	2.012	0.380	0.484	4.870	0.195	8.832	2.034	0.519	0.568	17.618
✓	✓	✓	7.602	1.662	**0.372**	**0.474**	5.988	0.209	9.376	**1.977**	**0.513**	**0.560**	16.995

Table 2. Results from the ablation analysis of BeLFusion. We assess the contribution of the latent ($\mathcal{L}_{lat}$) and reconstruction ($\mathcal{L}_{rec}$) losses, as well as the benefits of applying latent diffusion to a disentangled behavioral latent space (BLS).

	Human3.6M[32]		AMASS[43]	
	Avg. rank	Ranked 1st	Avg. rank	Ranked 1st
GSPS	2.246 ± 0.358	17.9%	2.003 ± 0.505	30.5%
DivSamp	2.339 ± 0.393	13.4%	2.432 ± 0.408	14.0%
BeLFusion	**1.415 ± 0.217**	**68.7%**	**1.565 ± 0.332**	**55.5%**

Table 3. Qualitative study. 126 participants ranked sets of samples from GSPS, DivSamp, and BeLFusion by their realism. Lower average rank (± std. dev.) is better.

(APD, CMD, FID), and makes the fast single-step inference very accurate. With large enough k values, the LDM learns to cover the conditional space of future behaviors to a great extent and can therefore make a fast and reliable first prediction. The successive denoising steps refine such approximations at expenses of larger inference time. Thus, each denoising step 1) promotes diversity within the latent space, and 2) brings the predicted latent code closer to the true behavioral distribution. Both effects can be observed in the latent APD and FID plots in Fig. 6. The latent APD is equivalent to the APD in the latent space of predictions and is computed likewise. Note that these effects are not favored by neither the loss choice nor the BLS (see supp. material Fig. G). Concurrent works have also highlighted the good performance achievable by single-step denoising [4, 15].

Qualitative assessment. We performed a qualitative study to assess the realism of BeLFusion's predictions compared to those of the most accurate methods: DivSamp and GSPS. For each method, we sampled six predictions for 24 randomly sampled observation segments from each dataset (48 in total). We then generated a *gif* that showed both the observed and predicted sequences of the six predictions at the same time. Each participant was asked to order the three sets according to the average realism of the samples. Four questions from either H36M or AMASS were asked to each participant (see supp. material Sec. F). A total of 126 people participated in the study. The statistical significance of the results was assessed with the Friedman and Nemenyi tests. Results are shown in Tab. 3. BeLFusion's predictions are significantly more realistic than both competitors' in both datasets ($p<0.01$). GSPS could only be proved significantly more realistic than DivSamp for AMASS ($p<0.01$). Interestingly, the participant-wise average realism ranks of each method are highly correlated to each method's CMD (r=0.730, and r=0.601) and APDE (r=0.732, and r=0.612), for both datasets (H36M, and AMASS, respectively), in terms of Pearson's correlation ($p<0.001$).

5. Conclusion

We presented BeLFusion, a latent diffusion model that exploits a behavioral latent space to make more realistic, accurate, and context-adaptive human motion predictions. BeLFusion takes a major step forward in the cross-dataset AMASS configuration. This suggests the necessity of future work to pay attention to domain shifts. These are present in any on-the-wild scenario and therefore on our way toward making highly capable predictive systems.

Limitations and future work. Although sampling with BeLFusion only takes 10 denoising steps, this is still slower than sampling from GANs or VAEs (see supp. material Sec. D.3.). This may limit its applicability to a real-life scenario. Future work includes exploring our method's capabilities for exploiting a longer observation time-span, and for being auto-regressively applied to predict longer-term sequences.

Acknowledgements. This work has been partially supported by the Spanish project PID2022-136436NB-I00 and by ICREA under the ICREA Academia programme.

References

[1] Hyemin Ahn and Dongheui Mascaro, Valls Esteve an Lee. Can we use diffusion probabilistic models for 3d motion prediction? In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, May 2023. 3

[2] Emre Aksan, Manuel Kaufmann, Peng Cao, and Otmar Hilliges. A spatio-temporal transformer for 3d human motion prediction. In *2021 International Conference on 3D Vision (3DV)*, pages 565–574. IEEE, 2021. 2, 5

[3] Sean Andrist, Bilge Mutlu, and Adriana Tapus. Look like me: matching robot personality via gaze to increase motivation. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, pages 3603–3612, 2015. 1

[4] Arpit Bansal, Eitan Borgnia, Hong-Min Chu, Jie S Li, Hamid Kazemi, Furong Huang, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Cold diffusion: Inverting arbitrary image transforms without noise. *arXiv preprint arXiv:2208.09392*, 2022. 8

[5] German Barquero, Johnny Núñez, Sergio Escalera, Zhen Xu, Wei-Wei Tu, Isabelle Guyon, and Cristina Palmero. Didn't see that coming: a survey on non-verbal social human behavior forecasting. In *Understanding Social Behavior in Dyadic and Small Group Interactions*, Proceedings of Machine Learning Research, 2022. 2

[6] German Barquero, Johnny Núñez, Zhen Xu, Sergio Escalera, Wei-Wei Tu, Isabelle Guyon, and Cristina Palmero. Comparison of spatio-temporal models for human motion and pose forecasting in face-to-face interaction scenarios. In *Understanding Social Behavior in Dyadic and Small Group Interactions*, Proceedings of Machine Learning Research, 2022. 5

[7] Emad Barsoum, John Kender, and Zicheng Liu. Hp-gan: Probabilistic 3d human motion prediction via gan. *Proceedings of the IEEE conference on computer vision and pattern recognition workshops.*, 2018. 2, 6

[8] Apratim Bhattacharyya, Bernt Schiele, and Mario Fritz. Accurate and diverse sampling of sequences based on a "best of many" sample objective. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8485–8493, 2018. 6

[9] Xiaoyu Bie, Wen Guo, Simon Leglaive, Lauren Girin, Francesc Moreno-Noguer, and Xavier Alameda-Pineda. Hitdvae: Human motion generation via hierarchical transformer dynamical vae. *arXiv preprint arXiv:2204.01565*, 2022. 5

[10] Andreas Blattmann, Timo Milbich, Michael Dorkenwald, and Björn Ommer. Behavior-driven synthesis of human dynamics. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 2, 4

[11] 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 *European Conference on Computer Vision*, pages 226–242. Springer, 2020. 2

[12] Yujun Cai, Yiwei Wang, Yiheng Zhu, Tat-Jen Cham, Jianfei Cai, Junsong Yuan, Jun Liu, Chuanxia Zheng, Sijie Yan, Henghui Ding, Xiaohui Shen, Ding Liu, and Nadia Magnenat Thalmann. A unified 3d human motion synthesis model via conditional variational auto-encoder. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 2

[13] Anargyros Chatzitofis, Leonidas Saroglou, Prodromos Boutis, Petros Drakoulis, Nikolaos Zioulis, Shishir Subramanyam, Bart Kevelham, Caecilia Charbonnier, Pablo Cesar, Dimitrios Zarpalas, et al. Human4d: A human-centric multimodal dataset for motions and immersive media. *IEEE Access*, 8:176241–176262, 2020. 7

[14] Ling-Hao Chen, Jiawei Zhang, Yewen Li, Yiren Pang, Xiaobo Xia, and Tongliang Liu. Humanmac: Masked motion completion for human motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 3

[15] Shoufa Chen, Peize Sun, Yibing Song, and Ping Luo. Diffusiondet: Diffusion model for object detection. *arXiv preprint arXiv:2211.09788*, 2022. 8

[16] Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, and Gang Yu. Executing your commands via motion diffusion in latent space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18000–18010, 2023. 3

[17] 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*, pages 11467–11476, 2021. 2

[18] Lingwei Dang, Yongwei Nie, Chengjiang Long, Qing Zhang, and Guiqing Li. Diverse human motion prediction via gumbel-softmax sampling from an auxiliary space. *ACM Multimedia*, 2022. 1, 2, 5, 6

[19] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34, 2021. 3, 5

[20] 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. 6

[21] Haoqiang Fan, Hao Su, and Leonidas J Guibas. A point set generation network for 3d object reconstruction from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 605–613, 2017. 5

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

[23] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. In *Advances in Neural Information Processing Systems*, volume 27, 2014. 5

[24] Chunzhi Gu, Jun Yu, and Chao Zhang. Learning disentangled representations for controllable human motion prediction. *arXiv preprint arXiv:2207.01388*, 2022. 2

[25] Tianpei Gu, Guangyi Chen, Junlong Li, Chunze Lin, Yongming Rao, Jie Zhou, and Jiwen Lu. Stochastic trajectory prediction via motion indeterminacy diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17113–17122, 2022. 3

[26] 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. 2

[27] 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 28th ACM International Conference on Multimedia*, pages 2021–2029, 2020. 5

[28] 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. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4809–4819, 2023. 2

[29] Agrim Gupta, Justin Johnson, Li Fei-Fei, Silvio Savarese, and Alexandre Alahi. Social gan: Socially acceptable trajectories with generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2255–2264, 2018. 5

[30] Swaminathan Gurumurthy, Ravi Kiran Sarvadevabhatla, and R Venkatesh Babu. Deligan: Generative adversarial networks for diverse and limited data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 166–174, 2017. 6

[31] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 3

[32] 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. 2, 5, 6, 8

[33] Ashesh Jain, Amir R Zamir, Silvio Savarese, and Ashutosh Saxena. Structural-rnn: Deep learning on spatio-temporal graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5308–5317, 2016. 2

[34] Jogendra Nath Kundu, Maharshi Gor, and R. Venkatesh Babu. Bihmp-gan: Bidirectional 3d human motion prediction gan. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019. 2

[35] 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. 2

[36] Maosen Li, Siheng Chen, Zihui Liu, Zijing Zhang, Lingxi Xie, Qi Tian, and Ya Zhang. Skeleton graph scattering networks for 3d skeleton-based human motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 854–864, 2021. 2

[37] Maosen Li, Siheng Chen, Yangheng Zhao, Ya Zhang, Yanfeng Wang, and Qi Tian. Dynamic multiscale graph neural networks for 3d skeleton based human motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 214–223, 2020. 2

[38] Zhenguang Liu, Kedi Lyu, Shuang Wu, Haipeng Chen, Yanbin Hao, and Shouling Ji. Aggregated multi-gans for controlled 3d human motion prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021. 2

[39] Zhenguang Liu, Shuang Wu, Shuyuan Jin, Qi Liu, Shijian Lu, Roger Zimmermann, and Li Cheng. Towards natural and accurate future motion prediction of humans and animals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10004–10012, 2019. 2

[40] Calvin Luo. Understanding diffusion models: A unified perspective. *arXiv preprint arXiv:2208.11970*, 2022. 3

[41] Kedi Lyu, Haipeng Chen, Zhenguang Liu, Beiqi Zhang, and Ruili Wang. 3d human motion prediction: A survey. *Neurocomputing*, 489:345–365, 2022. 1, 2

[42] Hengbo Ma, Jiachen Li, Ramtin Hosseini, Masayoshi Tomizuka, and Chiho Choi. Multi-objective diverse human motion prediction with knowledge distillation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 1, 5, 6

[43] 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*, pages 5442–5451, 2019. 2, 5, 6, 8

[44] Wei Mao, Miaomiao Liu, and Mathieu Salzmann. History repeats itself: Human motion prediction via motion attention. In *European Conference on Computer Vision*, pages 474–489. Springer, 2020. 2, 5

[45] Wei Mao, Miaomiao Liu, and Mathieu Salzmann. Generating smooth pose sequences for diverse human motion prediction. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 1, 2, 5, 6

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

[47] Julieta Martinez, Michael J Black, and Javier Romero. On human motion prediction using recurrent neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2891–2900, 2017. 2, 5

[48] Angel Martínez-González, Michael Villamizar, and Jean-Marc Odobez. Pose transformers (potr): Human motion prediction with non-autoregressive transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2276–2284, 2021. 2

[49] Omar Medjaouri and Kevin Desai. Hr-stan: High-resolution spatio-temporal attention network for 3d human motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2540–2549, 2022. 2, 5

[50] Graphics & Extended Reality Lab (University of Cyprus). Dance motion capture database. Available: http://dancedb.eu/ [Accessed: 01-Sep.-2022]. 7

[51] Cristina Palmero, German Barquero, Julio CS Jacques Junior, Albert Clapés, Johnny Núnez, David Curto, Sorina Smeureanu, Javier Selva, Zejian Zhang, David Saeteros, et al. Chalearn lap challenges on self-reported personality recognition and non-verbal behavior forecasting during social dyadic interactions: Dataset, design, and results. In *Understanding Social Behavior in Dyadic and Small Group Interactions*, pages 4–52. PMLR, 2022. 2

[52] Dario Pavllo, David Grangier, and Michael Auli. Quaternet: A quaternion-based recurrent model for human motion. In *British Machine Vision Conference (BMVC)*, 2018. 2

[53] Mathis Petrovich, Michael J Black, and Gül Varol. Action-conditioned 3d human motion synthesis with transformer vae. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10985–10995, 2021. 5

[54] Kashif Rasul, Calvin Seward, Ingmar Schuster, and Roland Vollgraf. Autoregressive denoising diffusion models for multivariate probabilistic time series forecasting. In *International Conference on Machine Learning*, pages 8857–8868. PMLR, 2021. 3

[55] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 2, 3

[56] Andrey Rudenko, Luigi Palmieri, Michael Herman, Kris M Kitani, Dariu M Gavrila, and Kai O Arras. Human motion trajectory prediction: A survey. *The International Journal of Robotics Research*, 39(8):895–935, 2020. 1

[57] Saeed Saadatnejad, Ali Rasekh, Mohammadreza Mofayezi, Yasamin Medghalchi, Sara Rajabzadeh, Taylor Mordan, and Alexandre Alahi. A generic diffusion-based approach for 3d human pose prediction in the wild. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 8246–8253. IEEE, 2023. 3

[58] Tim Salzmann, Marco Pavone, and Markus Ryll. Motron: Multimodal probabilistic human motion forecasting. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 1, 5, 6

[59] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015. 3

[60] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *International Conference on Learning Representations*, 2021. 5

[61] Omid Taheri, Nima Ghorbani, Michael J Black, and Dimitrios Tzionas. Grab: A dataset of whole-body human grasping of objects. In *European conference on computer vision*, pages 581–600. Springer, 2020. 7

[62] Yusuke Tashiro, Jiaming Song, Yang Song, and Stefano Ermon. Csdi: Conditional score-based diffusion models for probabilistic time series imputation. *Advances in Neural Information Processing Systems*, 34, 2021. 3

[63] Arash Vahdat, Karsten Kreis, and Jan Kautz. Score-based generative modeling in latent space. *Advances in Neural Information Processing Systems*, 34:11287–11302, 2021. 2, 3

[64] Jacob Walker, Kenneth Marino, Abhinav Gupta, and Martial Hebert. The pose knows: Video forecasting by generating pose futures. *Proceedings of the IEEE international conference on computer vision*, 2017. 2, 6

[65] Dong Wei, Huaijiang Sun, Bin Li, Jianfeng Lu, Weiqing Li, Xiaoning Sun, and Shengxiang Hu. Human joint kinematics diffusion-refinement for stochastic motion prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 6110–6118, 2023. 3

[66] Zhisheng Xiao, Karsten Kreis, and Arash Vahdat. Tackling the generative learning trilemma with denoising diffusion gans. *International Conference on Learning Representations*, 2021. 3

[67] Sirui Xu, Yu-Xiong Wang, and Liang-Yan Gui. Diverse human motion prediction guided by multi-level spatial-temporal anchors. In *European Conference on Computer Vision*, pages 251–269. Springer, 2022. 2

[68] 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. *Proceedings of the European conference on computer vision (ECCV)*, 2018. 2, 6

[69] Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Yingxia Shao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications. *arXiv preprint arXiv:2209.00796*, 2022. 3

[70] Ye Yuan, Kris Kitani, Y Yuan, and K Kitani. Dlow: Diversifying latent flows for diverse human motion prediction. *European Conference on Computer Vision*, 2020. 1, 2, 5, 6

[71] Ye Yuan and Kris M Kitani. Diverse trajectory forecasting with determinantal point processes. In *International Conference on Learning Representations*, 2019. 6