BeLFusion: Latent Diffusion for Behavior-Driven Human Motion Prediction

German Barquero[1,2] germanbarquero@ub.edu
Sergio Escalera[1,2] sergio@maia.ub.es
Cristina Palmero[1,2] crpalmec7@alumnes.ub.edu
[1]Universitat de Barcelona [2]Computer Vision Center

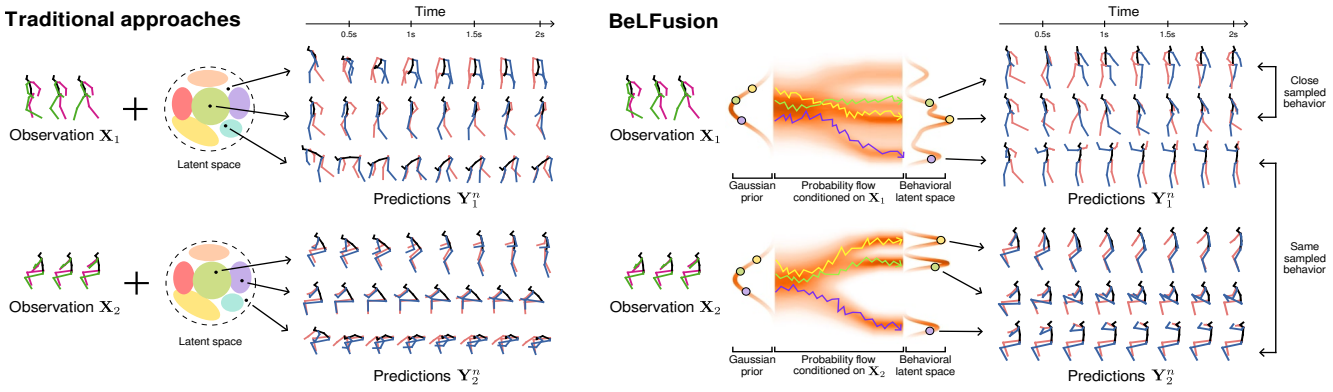

Figure 1. Common approaches for stochastic human motion prediction use variational autoencoders to model a latent space. Then, the latent code sampled from it is fed to a decoder conditioned on the observation to generate the prediction. In this scenario, out-of-distribution samples or low KL regularizations 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, our proposed method, BeLFusion, leverages latent diffusion models to sample from a disentangled behavioral space. As a result, it is able to predict a wide range of future behaviors performed with high realism.

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 movements in terms of the skeleton joints' dispersion. This has led to methods predicting fast and motion-divergent movements, which are often unrealistic and incoherent with past motion. Such methods also neglect contexts that need to anticipate diverse low-range behaviors, or actions, with subtle joint displacements. To address these issues, we present BeLFusion, a model that, for the first time, leverages latent diffusion models in HMP to sample from a latent space where behavior is disentangled from pose and motion. As a result, diversity is encouraged from a behavioral perspective. Thanks to our behavior coupler's ability to transfer sampled behavior to ongoing motion, BeLFusion's predictions display a variety of behaviors that are significantly more realistic 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 our definition of realism according to a qualitative study with 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 interactions with other people, driving a car, or walking across a crowd. Replicating this ability is imperative in many applications like assistive robots, virtual avatars, or autonomous cars [2, 52]. 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 [35]. 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 [15, 36,39,53,65]. This definition biases research toward models that generate fast and motion-divergent motions (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 [4, 44]. In a surveillance scenario, a foreseen noxious 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 often suffer from a trade-off that has been persistently overlooked: predicted motion looks unnatural when observed following the motion of the immediate past. 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[1] (Fig. 1). By building a latent space where behavior is disentangled from poses and motion, diversity is detached from the traditional coordinate-based perspective and promoted from a behavioral viewpoint. The *behavior coupler* ensures the predicted behavior is decoded into a smooth and realistic continuation of the ongoing motion. Thus, our predicted motions look more realistic than alternatives, which we assess through quantitative and qualitative analyses. BeLFusion is the first approach that exploits conditional latent diffusion models (LDM) [51, 58] for stochastic HMP, achieving state-of-the-art performance. Specifically, BeLFusion combines the unique capabilities of LDMs to model conditional distributions with the convenient inductive biases recurrent neural networks (RNNs) have for HMP.

To summarize, our main contributions are: (1) We propose BeLFusion, a method that generates predictions that are significantly more realistic and coherent with respect to the near past than prior works, while achieving state-of-the-art accuracy on Human 3.6M [26] and AMASS [37] datasets. (2) BeLFusion promotes diversity in a behavioral latent space. As a result, both low- (e.g., hand-waving, smoking) and long-range motions (e.g., standing up, sitting down) are equally encouraged. We show that this boosts the capacity to adapt the predictions' diversity to the determinacy of the motion context. (3) 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 [37] dataset, showcases a broad range of actions performed by more than 400 subjects. (4) 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 fairly correlated to our definition of realism.

[1]Code and pretrained models are made publicly available in https://github.com/BarqueroGerman/BeLFusion.

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 usually modeled with autoregressive RNNs [18, 21, 27, 41, 46] or Transformers [1, 11, 42]. Graph Convolutional Networks are typically included as intermediate layers to model the dependencies among joints [14, 30, 31, 40]. Some methods leverage Temporal Convolutional Networks [29, 43] or a simple Multi-Layer Perceptron [23] to predict fixed-size sequences, achieving high performance. Recently, some works claimed the benefits of modeling sequences in the frequency space [11, 38, 40]. 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) [6, 28] and variational autoencoders (VAEs) [12, 39, 60, 62]. These methods often include diversity-promoting losses in order to predict a high variety of motions [39], or incorporate explicit techniques for diverse sampling [15, 65]. 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 within the context of the observed motion. Moreover, prior works neglect situations where a diversity of behaviors, which can sometimes be subtle, is important. We address this by implicitly enforcing such diversity in a behavioral latent space.

Semantic human motion prediction. Few works have attempted to leverage semantically meaningful latent spaces for stochastic HMP [19, 33, 62]. For example, [19] exploits disentangled motion representations for each part of the body to control the HMP. [62] proposes to add 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

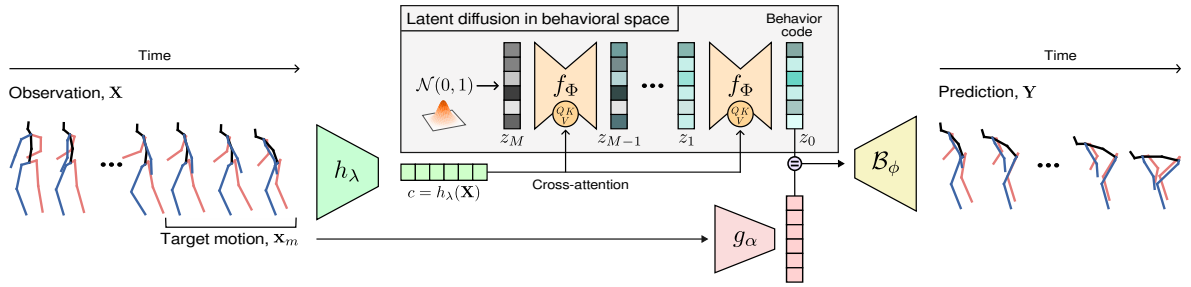

Figure 2. BeLFusion's architecture. A latent diffusion model conditioned on an encoding c of the observation, $\mathbf{X}$, progressively denoises a sample from a zero-mean unit variance multivariate normal distribution into a behavior code. Then, the behavior coupler $\mathcal{B}_\phi$ decodes the prediction 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.

paper, we propose a generalization of such framework that transfers behavior to ongoing movements. Our method exploits this disentanglement to promote behavioral diversity in 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 [25,54]. For conditional generation, the most common strategy consists in applying cross-attention to the conditioning signal at each denoising timestep [16]. Diffusion models have achieved impressive results in fields like video generation, inpainting, or anomaly detection [63]. In a more similar context, [49,57] use diffusion models for time series forecasting and imputation. [20] recently presented a diffusion model for trajectory prediction that controls the uncertainty of the prediction by shortening the denoising chain.

All diffusion models come with an expensive trade-off: immensely slow inference due to the large number of denoising steps required. Latent diffusion models (LDM) accelerate the sampling by applying diffusion to a much lower-resolution latent space learned by a VAE [51, 58]. Thanks to the Kullback–Leibler (KL) regularization, the learned latent space is built close to a normal distribution. As a result, the length of the Markov chain that diffuses 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 describe the methodology of BeLFusion (see Fig. 2). First, we characterize the HMP problem (Sec. 3.1). Then, we adapt the definitions of LDMs to our scenario (Sec. 3.2). Finally, we describe the construction of our behavioral latent space and the derivation of the 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 different 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 $\mathbf{Y}^i = \{p_t^i, p_{t+1}^i, ..., p_{t+T-1}^i\}$, respectively, where $i \in \{1, \dots, N\}$[2], and $p_t^i \in \mathbb{R}^d$ are the Cartesian coordinates of the human joints at time step t.

3.2. Motion latent diffusion

Now, 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 $z = \mathcal{E}(\mathbf{Y}) \in V$[3]. 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 $\epsilon_t = f_\Phi(z_t, t, \mathbf{X})$ of the diffused latent code z_t at each time step t, where $\mathbf{X}$ is the 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 $z_0 = f_\Phi(z_t, t, \mathbf{X})$ [34, 61]. 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

[2] A sampled prediction $\mathbf{Y}^i$ is hereafter referred as $\mathbf{Y}$ for intelligibility.

[3] For simplicity, we make an abuse of notation by using $\mathcal{E}(\mathbf{Y})$ to refer to the mean of the distribution $\mathcal{E}(z|\mathbf{Y})$.

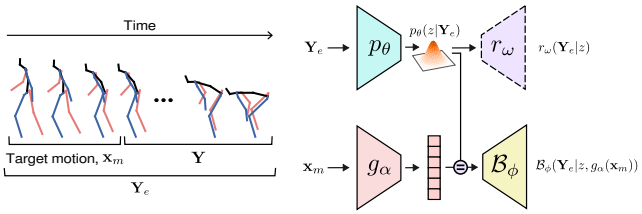

Figure 3. Framework for behavioral disentanglement. By adversarially training the auxiliary generator, r_ω, against the behavior coupler, $\mathcal{B}_\phi$, the behavior encoder, p_θ, learns to generate a disentangled latent space of behaviors, $p_\theta(z|\mathbf{Y}_e)$. At inference, $\mathcal{B}_\phi$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

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 objective loss (or *latent* loss) becomes:

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

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 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 architecture proposed 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 $\mathbf{Y}$. The goal is to disentangle the behavior from the motion and poses in $\mathbf{Y}$. To do so, we adversarially train two generators, the behavior coupler $\mathcal{B}_\phi$, and the auxiliary decoder r_ω, 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 r_ω 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, r_ω 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, r_ω 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:

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

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

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

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 multi-variate $\mathcal{N}(0, I)$. The inclusion of $-\mathcal{L}_{\text{aux}}$ in Eq. 4 penalizes any accurate reconstruction of $\mathbf{Y}_e$ by the auxiliary generator. Since $\mathcal{B}_\phi$ has access to the target posture and motion provided by $\mathbf{x}_m$, the main decoder $\mathcal{B}_\phi$ only needs p_θ to encode the behavioral dynamics. The erasing of any postural information from z is encouraged by the $-\mathcal{L}_{\text{aux}}$ term in $\mathcal{L}_{\text{main}}$. One could argue that a valid and simple solution for p_θ would consist in disentangling motion from postures. However, motion dynamics can still be used to easily extract a good posture approximation. Further details and visual examples of behavioral transference to several motions $\mathbf{x}_m$ are included in the supp. material.

Behavior-driven HMP. BeLFusion's goal is to sample the appropriate behavior code given the observation $\mathbf{X}$, see Fig. 2. To that end, an LDM conditioned on $c = h_\lambda(\mathbf{X})$ 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:

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

The second term of Eq. 5 leverages the autoencoded $\mathbf{Y}_e$. We optimize the objective within the solutions space bounded at the top by the autoencoder capabilities to help stabilize the training. Note that only the future poses $\mathbf{Y} \subset \mathbf{Y}_e$ form the prediction. The encoder h_λ is pretrained in an autoencoder framework that reconstructs $\mathbf{X}$. We found experimentally that h_λ does not benefit from further training, so its parameters λ are frozen during BeLFusion's training. 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 c 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 caused by the strong regularization of both losses. We propose to relax them by sampling k predictions at each training iteration and only backpropagating the gradients through the two predictions that each minimize the latent or the reconstructed loss:

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

where λ controls the trade-off between the latent and the reconstruction errors. Regularization relaxation usually leads to out-of-distribution predictions. This is often solved by employing additional complex techniques like pose priors, or bone-length losses that regularize the other predictions [9, 39]. 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 [16]; 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 [26] (H36M), and AMASS [37]. 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 [65] and adopted by most subsequent works [15, 36, 39, 53] (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 [1, 38, 43], 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 class labels. See the supp. material for more details.

Baselines. We include the zero-velocity baseline, which has been proven very competitive in HMP [5, 41], and a version of our model that replaces the LDM with a GAN, BeGAN. 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 $M = 10$, $k = 50$, a U-Net with cross-attention [16] as f_Φ, and one-layer RNNs as h_λ, g_α, and $\mathcal{B}_\phi$. For H36M, $\lambda = 5$, and for AMASS, $\lambda = 1$. The model used for inference was an exponential moving average of the trained model with a decay of 0.999. Sampling was conducted with a DDIM sampler [55]. 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 diffusion step (i.e., x10 faster): BeLFusion_D. Further implementation details are included in the supp. material.

4.2. Evaluation metrics

To compare BeLFusion with prior works, we follow the well-established evaluation pipeline proposed in [65]. 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 $\mathbf{X}$

	Human3.6M [26]								AMASS [37]						
	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 [6]	7.214	-	0.858	0.867	0.847	0.858	-	-	-	-	-	-	-	-	-
DSF [64]	9.330	-	0.493	0.592	0.550	0.599	-	-	-	-	-	-	-	-	-
DeLiGAN [24]	6.509	-	0.483	0.534	0.520	0.545	-	-	-	-	-	-	-	-	-
GMVAE [17]	6.769	-	0.461	0.555	0.524	0.566	-	-	-	-	-	-	-	-	-
TPK [60]	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 [62]	0.403	-	0.457	0.595	0.716	0.883	-	-	-	-	-	-	-	-	-
BoM [8]	6.265	-	0.448	0.533	0.514	0.544	-	-	-	-	-	-	-	-	-
DLow [65]	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 [36]	14.240	-	0.414	0.516	-	-	-	-	-	-	-	-	-	-	-
GSPS [39]	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 [53]	7.168	2.583	0.375	0.488	0.509	0.539	40.796	13.743	-	-	-	-	-	-	-
DivSamp [15]	**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.

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 [9, 15, 22, 48], 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:

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

The distribution accumulation 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 for additional details.

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 the best accuracy results (BeLFusion_D). Our method also excels at realism-related metrics like CMD and FID, which benefit from going through all denoising steps. 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, espe-

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.

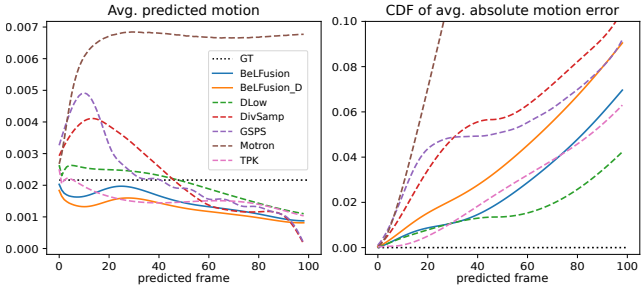

Figure 5. Left. Average predicted motion of state-of-the-art methods in H36M. Right. Cumulative distribution function (CDF) of the weighted absolute errors in the left with respect to the ground truth. CMD is the area under this curve.

cially for H36M. The predictions from methods that leverage transformations in the frequency space freeze at the very long-term horizon. The high CMD value of Motron depicts an important jitter in its predictions. BeLFusion shows low APDE, highlighting its good ability to adjust to the observed context. 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.

Fig. 4 shows the evolution of 10 superimposed predictions along time in three actions from H36M (sitting down, eating, and giving directions), and three datasets from AMASS (DanceDB[4], HUMAN4D [13], and GRAB [56]). It confirms visually what the CMD and APDE metrics already suggested. 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.

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) tun-

[4]Dance Motion Capture DB, http://dancedb.cs.ucy.ac.cy.

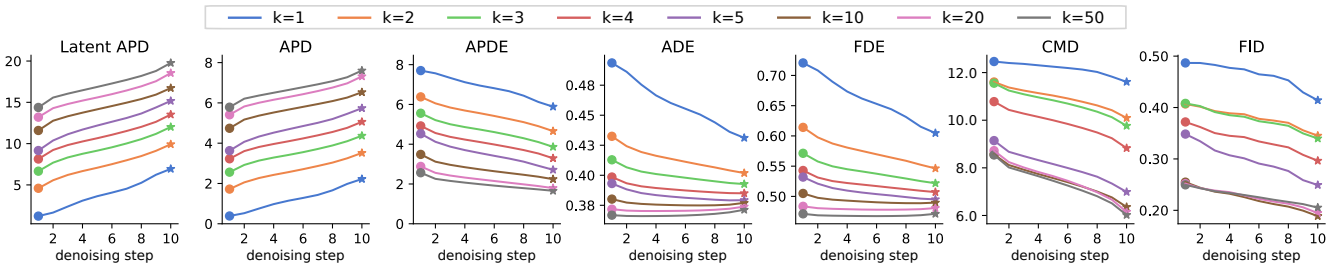

Figure 6. Evolution of evaluation metrics (y-axis) along denoising steps (x-axis) at inference time, for different values of k. Early stopping can be applied at any time, between the first (•) and the last step ($\star$). Accuracy saturates at $k = 50$, with gains for all metrics when increasing k, especially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

			Human3.6M [26]						AMASS [37]				
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 [26]		AMASS [37]	
	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	$\mathbf{1.415 \pm 0.217}$	**68.7%**	$\mathbf{1.565 \pm 0.332}$	**55.5%**

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

ing 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 (APD, CMD, FID), and makes the fast single-step inference extremely 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 stochastically 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 the equivalence of 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). Concurrent research has observed a similar effect on image generation [3].

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 (more details in supp. material). 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. This might 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.

References

[1] 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

[2] 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

[3] 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

[4] 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

[5] 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

[6] 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

[7] Miguel Angel Bautista, Pengsheng Guo, Samira Abnar, Walter Talbott, Alexander Toshev, Zhuoyuan Chen, Laurent Dinh, Shuangfei Zhai, Hanlin Goh, Daniel Ulbricht, et al. Gaudi: A neural architect for immersive 3d scene generation. *arXiv preprint arXiv:2207.13751*, 2022. 12

[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. Hit-dvae: Human motion generation via hierarchical transformer dynamical vae. *arXiv preprint arXiv:2204.01565*, 2022. 5, 6

[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] 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

[15] Lingwei Dang, Yongwei Nie, Chengjiang Long, Qing Zhang, and Guiqing Li. Diverse human motion prediction via gumbel-softmax sampling from an auxiliary space; diverse human motion prediction via gumbel-softmax sampling from an auxiliary space. *ACM Multimedia*, 2022. 1, 2, 5, 6, 16

[16] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34, 2021. 3, 5, 12

[17] 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

[18] 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

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

[20] Tianpei Gu, Guangyi Chen, Junlong Li, Chunze Lin, Yongming Rao, Jie Zhou, and Jiwen Lu. Stochastic trajectory prediction via motion indeterminacy diffusion. *arXiv preprint arXiv:2203.13777*, 2022. 3

[21] 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

[22] 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. 6

[23] Wen Guo, Yuming Du, Xi Shen, Vincent Lepetit, Xavier Alameda-Pineda, and Francesc Moreno-Noguer. Back to mlp: A simple baseline for human motion prediction. *arXiv preprint arXiv:2207.01567*, 2022. 2

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

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

[26] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*, 36(7):1325–1339, 2013. 2, 5, 6, 8, 21

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

[28] 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

[29] 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

[30] 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

[31] 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

[32] Xiang Lisa Li, John Thickstun, Ishaan Gulrajani, Percy Liang, and Tatsunori B Hashimoto. Diffusion-lm improves controllable text generation. *arXiv preprint arXiv:2205.14217*, 2022. 12

[33] 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*, volume 35, pages 2225–2232, 2021. 2

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

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

[36] 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

[37] 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, 13, 21

[38] 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

[39] 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, 16

[40] 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

[41] 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

[42] 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

[43] 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

[44] 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

[45] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 12

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

[47] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. 12

[48] 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. 6

[49] Kashif Rasul, Calvin Seward, Ingmar Schuster, and Roland Vollgraf. Autoregressive denoising diffusion models for mul-

tivariate probabilistic time series forecasting. In *International Conference on Machine Learning*, pages 8857–8868. PMLR, 2021. 3

[50] Sashank J Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of adam and beyond. *arXiv preprint arXiv:1904.09237*, 2019. 12

[51] 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

[52] 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

[53] 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

[54] 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

[55] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *International Conference on Learning Representations*, 2021. 5, 13

[56] 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, 17

[57] 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

[58] 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

[59] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 14

[60] 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, 16

[61] Zhisheng Xiao, Karsten Kreis, and Arash Vahdat. Tackling the generative learning trilemma with denoising diffusion gans. *arXiv preprint arXiv:2112.07804*, 2021. 3

[62] 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

[63] 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

[64] Ye Yuan and Kris Kitani. Diverse trajectory forecasting with determinantal point processes. *arXiv preprint arXiv:1907.04967*, 2019. 6

[65] 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, 14, 16

Supplementary Material

In this supplementary material, we first include additional implementation details to those provided in Sec. 4.1 needed to reproduce our work (Sec. A). Then, we complement Sec. 4.1 by providing all the information needed to follow the proposed cross-dataset AMASS evaluation protocol (Sec. B). Sec. 3.3 is also extended with a 2D visualization of the disentangled behavioral latent space, and several video examples of behavioral transference (Sec. C). Class- and dataset-wise results from Sec. 4.3 are included and discussed (Sec. D), as well as a detailed discussion on several video examples comparing BeLFusion against the state of the art (Sec. E). Finally, we provide a thorough description and extended results of the qualitative assessment presented at the end of Sec. 4.3 (Sec. F).

A. Implementation details

To ensure reproducibility, we include in this section all the details regarding BeLFusion's architecture and training procedure (Sec. A.1). We also cover the details on the implementation of the state-of-the-art models retrained with AMASS (Sec. A.2). We follow the terminology used in Fig. 2 and 3 from the main paper.

Note that we only report the hyperparameter values of the best models. For their selection, we conducted grid searches that included learning rate, losses weights, and most relevant network parameters. Data augmentation for all models consisted in randomly rotating from 0 to 360 degrees around the Z axis and mirroring the body skeleton with respect to the XZ- and YZ-planes. The axis and mirroring planes were selected to preserve the floor position and orientation. All models were trained with the ADAM optimizer with AMSGrad [50], with PyTorch 1.9.1 [45] and CUDA 11.1 on a single NVIDIA GeForce RTX 3090. The whole BeLFusion training pipleine was trained in 12h for H36M, and 24h for AMASS.

A.1. BeLFusion

Behavioral latent space. The behavioral VAE consists of four modules. The behavior encoder p_θ, which receives the flattened coordinates of all the joints, is composed of a single Gated Recurrent Unit (GRU) cell (hidden state of size 128) followed by a set of a 2D convolutional layer (kernel size of 1, stride of 1, padding of 0) with L2 weight normalization and learned scaling parameters that maps the GRU state to the mean of the latent distribution, and another set to its variance. The behavior coupler $\mathcal{B}_\phi$ consists of a GRU (input shape of 256, hidden state of size 128) followed by a linear layer that maps, at each timestep, its hidden state to the offsets of each joint coordinates with respect to their last observed position. The context encoder g_α is an MLP (hidden state of 128) that is fed with the flattened joints coordinates of the target motion $\mathbf{x}_m$, that includes C=3 frames. Finally, the auxiliary decoder r_ω is a clone of $\mathcal{B}_\phi$ with a narrower input shape (128), as only the latent code is fed. For H36M, the behavioral VAE was trained with learning rates of 0.005, and 0.0005 for $\mathcal{L}_{main}$ and $\mathcal{L}_{aux}$, respectively. For AMASS, they were set to 0.001 and 0.005. They were all decayed with a ratio of 0.9 every 50 epochs. The batch size was set to 64. Each epoch consisted of 5000 and 10000 iterations for H36M and AMASS, respectively. The weight of the $-\mathcal{L}_{aux}$ term in $\mathcal{L}_{main}$ was set to 1.05 for H36M and to 1.00 for AMASS. The KL term was assigned a weight of 0.0001 in both datasets. Once trained, the behavioral VAE was further fine-tuned for 500 epochs with the behavior encoder p_θ frozen, to enhance the reconstruction capabilities without modifying the disentangled behavioral latent space. Note that for the ablation study, the non-behavioral latent space was built likewise by disabling the adversarial training framework, and optimizing the model only with the log-likelihood and KL terms of $\mathcal{L}_{main}$ (main paper, Eq. 4), as in a traditional VAE framework.

Observation encoding. The observation encoder h_λ was pretrained as an autoencoder with an L2 reconstruction loss. It consists of a single-cell GRU layer (hidden state of 64) fed with the flattened joints coordinates. The hidden state of the GRU layer is fed to three MLP layers (output sizes of 300, 200, and 64), and then set as the hidden state of the GRU decoder unit (hidden state of size 64). The sequence is reconstructed by predicting the offsets with respect to the last observed joint coordinates.

Latent diffusion model. BeLFusion's LDM borrowed its U-Net from [16]. To leverage it, the target latent codes were reshaped to a rectangular shape (16x8), as prior work proposed [7]. In particular, our U-Net has 2 attention layers (resolutions of 8 and 4), 16 channels per attention head, a FiLM-like conditioning mechanism [47], residual blocks for up and downsampling, and a single residual block. Both the observation and target behavioral encodings were normalized between -1 and 1. The LDM was trained with the *sqrt* noise schedule ($s = 0.0001$) proposed in [32], which also provided important improvements in our scenario compared to the classic *linear* or *cosine* schedules (see Fig. 7). With this schedule, the diffusion process is started with a higher noise level, which increases rapidly in the middle of the chain. The length of the Markov diffusion chain was set to 10, the batch size to 64, the learning rate to 0.0005, and the learning rate decay to a rate of 0.9 every 100 epochs. Each epoch included 10000 samples in both H36M and AMASS training scenarios. Early stopping with a patience of 100 epochs was applied to both, and the epoch where it was triggered was used for the final training with both validation and training sets together. Thus, BeLFusion was trained for 217 epochs in H36M and 1262 for AMASS.

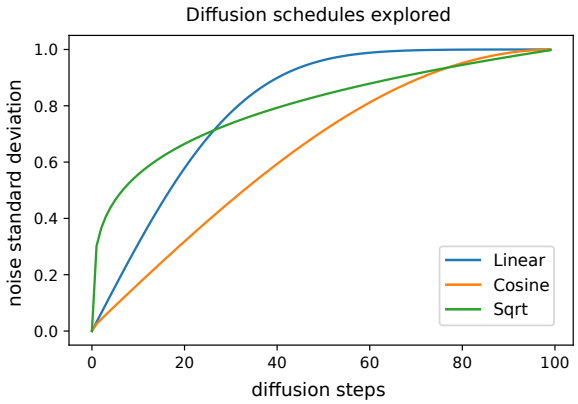

Figure 7. **Diffusion schedules.** Schedules explored for diffusing the target latent codes.

For both datasets, the LDM was trained with an exponential moving average (EMA) with a decay of 0.999, triggered every 10 batch iterations, and starting after 1000 initial iterations. The EMA helped reduce the overfitting in the last denoising steps. Predictions were inferred with DDIM sampling [55].

A.2. State-of-the-art models

The publicly available codes from TPK, DLow, GSPS, and DivSamp were adapted to be trained and evaluated under the AMASS cross-dataset protocol. The best values for their most important hyperparameters were found with grid search. The number of iterations per epoch for all of them was set to 10000.

TPK's loss weights were set to 1000 and 0.1 for the transition and KL losses, respectively. The learning rate was set to 0.001. DLow was trained on top of the TPK model with a learning rate of 0.0001. Its reconstruction and diversity losses weights were set to 2 and 25. For GSPS, the upper- and lower-body joint indices were adapted to the AMASS skeleton configuration. The multimodal ground truth was generated with an upper L2 distance of 0.1, and a lower APD threshold of 0.3. The body angle limits were recomputed with the AMASS statistics. The GSPS learning rate was set to 0.0005, and the weights of the upper- and lower-body diversity losses were set to 5 and 10, respectively. For DivSamp, we used the multimodal ground truth from GSPS, as for H36M they originally borrowed such information from GSPS. For the first training stage (VAE), the learning rate was set to 0.001, and the KL weight to 1. For the second training stage (sampling model), the learning rate was set to 0.0001, the reconstruction loss weight was set to 40, and the diversity loss weight to 20. For all of them, unspecified parameters were set to the values reported in their original H36M implementations.

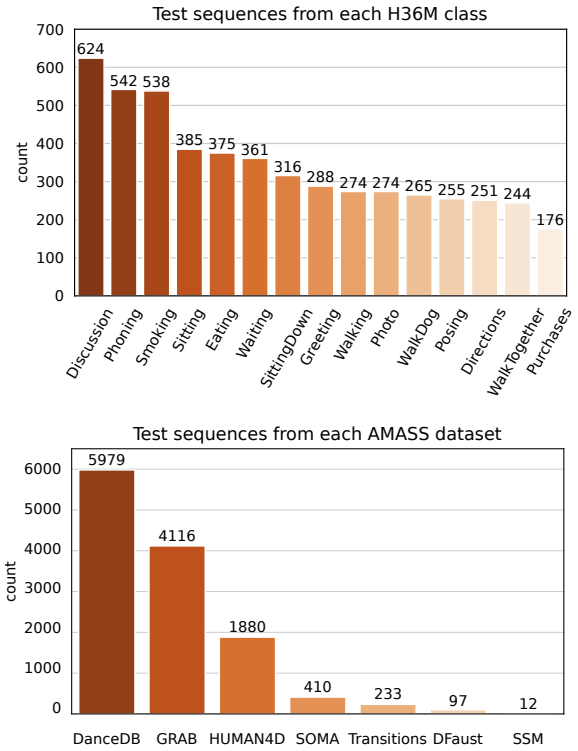

Figure 8. **Test set sequences.** We show the number of test sequences evaluated for each class/dataset in H36M/AMASS.

B. AMASS cross-dataset protocol

In this section, we give more details to ensure the reproducibility of the cross-dataset AMASS evaluation protocol.

Training splits. The training, validation, and test splits are based on the official AMASS splits from the original publication [37]. However, we also include the new datasets added afterward, up to date. Accordingly, the training set contains the ACCAD, BMLhandball, BMLmovi, BMLrub, CMU, EKUT, EyesJapanDataset, KIT, PosePrior, TCD-Hands, and TotalCapture datasets, and the validation set contains the HumanEva, HDM05, SFU, and MoSh datasets. The remaining datasets are all part of the test set: DFaust, DanceDB, GRAB, HUMAN4D, SOMA, SSM, and Transitions. AMASS datasets showcase a wide range of behaviors at both intra- and inter-dataset levels. For example, DanceDB, GRAB, and BMLhandball contain sequences of dancing, grabbing objects, and sport actions, respectively. Other datasets like HUMAN4D offer a wide intra-dataset variability of behaviors by themselves. As a result, this evaluation protocol represents a very complete and challenge benchmark for HMP.

Test sequences. For each dataset clip (previously downsampled to 60Hz), we selected all sequences starting from frame 180 (3s), with a stride of 120 (2s). This was done to ensure that for any segment to predict (prediction window), up to 3s of preceding motion was available. As a

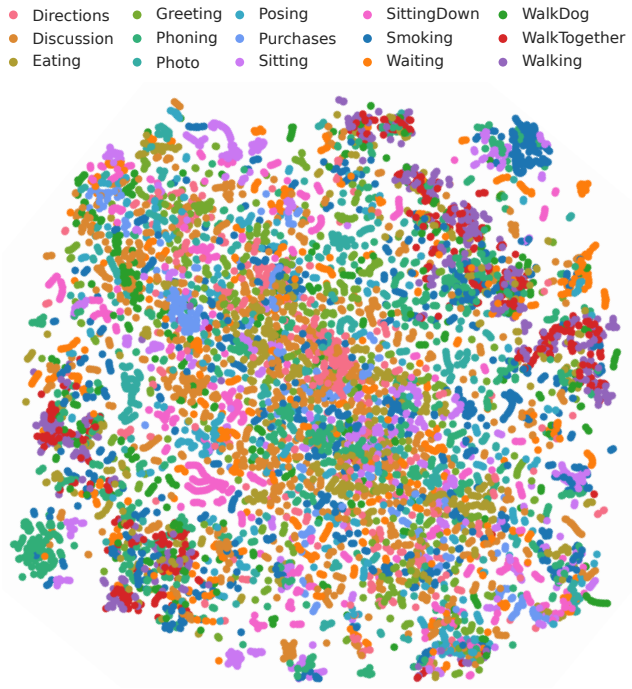

Figure 9. **Behavioral latent space.** 2D projection of the behavioral encodings of all H36M test sequences generated with t-SNE.

result, future work will be able to explore models exploiting longer observation windows while still using the same prediction windows and, therefore, be compared to our results. A total of 12728 segments were selected, around 2.5 times the amount of H36M test sequences. Note that those clips with no framerate available in AMASS metadata were ignored. Fig. 8 shows the number of segments extracted from each test dataset. 94.1% of all test samples belong to either DanceDB, GRAB, or HUMAN4D. Most SSM clips had to be discarded due to lengths shorter than 300 frames (5s). The list of sequence indices is made available along the project code for easing reproducibility.

Multimodal ground truth. The L2 distance threshold used for the generation of the multimodal ground truth was set to 0.4 so that the average number of resulting multimodal ground truths for each sequence was similar to that of H36M with a threshold of 0.5 [65].

C. Behavioral latent space

In this section, we present 1) a t-SNE plot for visualizing the behavioral latent space of the H36M test segments, and 2) visual examples of transferring behavior to ongoing motions.

2D projection. Fig. 9 shows a 2-dimensional t-SNE projection of all behavioral encodings of the H36M test sequences [59]. Note that, despite its class label, a sequence may show actions of another class. For example, *Waiting* sequences include sub-sequences where the person walks or sits down. Interestingly, we can observe that most walking-related sequences (*WalkDog*, *WalkTogether*, *Walking*) are clustered together in the top-right and bottom-left corners. Such entanglement within those clusters suggests that the task of choosing the way to keep walking might be relegated to the behavior coupler, which has information on how the action is being performed. Farther in those corners, we can also find very isolated clusters of *Phoning* and *Smoking*, whose proximity to the walking behaviors suggests that such sequences may involve a subject making a call or smoking while walking. However, without fine-grained annotations at the sequence level, we cannot come to any strong conclusion.

Transference of behaviors. We include several videos[5] showing the capabilities of the behavior coupler to transfer a behavioral latent code to any ongoing motion. The motion tagged as *behavior* shows the target behavior to be encoded and transferred. All the other columns show the ongoing motions where the behavior will be transferred to. They are shown with blue and orange skeletons. Once the behavior is transferred, the color of the skeletons switches to green and pink. In 'H1' (H36M), the walking action or behavior is transferred to the target ongoing motions. For ongoing motions where the person is standing, they start walking towards the direction they are facing (#1, #2, #4, #5). Such transition is smooth and coherent with the observation. For example, the person making a phone call in #7 keeps the arm next to the ear while starting to walk. When sitting or bending down, the movement of the legs is either very little (#3 and #6), or very limited (#8). 'H2' and 'H3' show the transference of subtle and long-range behaviors, respectively. For AMASS, such behavioral encoding faces a huge domain drift. However, we still observe good results at this task. For example, 'A1' shows how a *stretching* movement is successfully transferred to very distinct ongoing motions by generating smooth and realistic transitions. Similarly, 'A2' and 'A3' are examples of transferring subtle and aggressive behaviors, respectively. Even though the dancing behavior in 'A3' was not seen at training time, it is transferred and adapted to the ongoing motion fairly realistically.

D. Further experimental results

In this section, we present a class- and dataset-wise comparison to the state of the art for H36M and AMASS, respectively (Sec. D.1). We also include the distributions of predicted displacement for each class/dataset, which are used for the CMD calculation. Finally, we present an extended analysis of the effect of k, which controls the loss *relaxation* level (Sec. D.2).

[5] Videos referenced in the supp. material are available in https://barquerogerman.github.io/BeLFusion/.

Classes	APD	APDE	ADE	FDE	MMADE	MMFDE	CMD
Directions							
TPK	6.510	2.039	0.447	0.482	0.523	0.544	4.687
DLow	11.874	3.359	0.415	0.465	0.499	0.514	**3.471**
GSPS	15.398	6.877	0.407	0.477	0.492	0.522	11.312
DivSamp	**15.663**	7.142	0.389	0.463	0.502	0.523	11.557
BeLFusion	7.090	**1.709**	**0.378**	**0.422**	**0.484**	**0.494**	7.364
Discussion							
TPK	6.966	2.572	0.511	0.581	0.570	0.600	6.795
DLow	11.872	2.659	0.472	0.536	0.533	0.549	**2.708**
GSPS	14.199	4.992	0.448	0.541	0.526	0.563	8.641
DivSamp	**15.310**	5.905	0.432	0.526	0.534	0.557	9.410
BeLFusion	9.172	**1.425**	**0.420**	**0.507**	**0.512**	**0.530**	6.789
Eating							
TPK	6.412	**1.066**	0.388	0.473	0.452	0.472	5.451
DLow	11.603	4.829	0.358	0.433	0.439	0.452	**3.278**
GSPS	15.570	8.793	0.334	0.419	0.424	0.448	12.099
DivSamp	**15.681**	8.904	0.321	0.419	0.428	0.445	13.621
BeLFusion	5.954	1.297	**0.310**	**0.381**	**0.418**	**0.420**	5.904
Greeting							
TPK	6.779	2.545	0.555	0.615	0.571	0.598	10.781
DLow	11.897	3.112	0.530	0.590	0.542	0.564	**4.262**
GSPS	14.974	5.950	0.502	0.592	0.532	0.577	10.539
DivSamp	**15.447**	6.373	0.489	0.579	0.535	0.562	9.280
BeLFusion	8.482	**1.690**	**0.482**	**0.544**	**0.524**	**0.540**	11.293
Phoning							
TPK	6.410	**1.400**	0.377	0.475	0.468	0.507	**3.458**
DLow	11.542	4.605	0.343	0.444	0.451	0.487	4.390
GSPS	15.050	8.120	0.311	0.413	0.436	0.476	12.037
DivSamp	**15.751**	8.813	0.296	0.400	0.437	0.471	13.780
BeLFusion	6.649	1.477	**0.283**	**0.375**	**0.426**	**0.445**	4.070
Photo							
TPK	6.894	1.884	0.541	0.689	0.548	0.633	**2.536**
DLow	11.931	4.180	0.507	0.655	0.516	0.596	6.320
GSPS	14.310	6.482	0.485	0.663	0.502	0.606	11.734
DivSamp	**15.330**	7.428	0.474	0.665	0.506	0.607	12.886
BeLFusion	8.446	**1.726**	**0.434**	**0.601**	**0.462**	**0.546**	3.317
Posing							
TPK	6.520	2.310	0.466	0.538	0.542	0.565	3.518
DLow	11.875	3.116	0.442	0.521	0.510	**0.525**	4.500
GSPS	15.149	6.399	0.415	0.527	**0.498**	0.543	11.039
DivSamp	**15.429**	6.676	**0.395**	**0.499**	0.510	0.541	11.861
BeLFusion	8.438	**1.241**	0.406	0.510	**0.498**	0.531	**3.331**
Purchases							
TPK	7.450	2.161	0.505	0.522	0.535	0.538	11.063
DLow	11.947	2.629	0.430	0.422	**0.493**	0.477	**5.880**
GSPS	13.969	4.552	0.414	0.429	0.497	0.497	7.651
DivSamp	**14.967**	5.517	**0.388**	**0.404**	0.502	0.478	6.922
BeLFusion	10.272	**1.738**	0.410	0.409	0.494	**0.472**	9.564

Classes	APD	APDE	ADE	FDE	MMADE	MMFDE	CMD
Sitting							
TPK	6.417	**1.167**	0.400	0.547	0.461	0.548	2.004
DLow	11.425	4.972	0.364	0.513	0.440	0.523	8.565
GSPS	14.966	8.494	0.323	0.454	0.411	0.484	15.597
DivSamp	**15.614**	9.146	0.317	0.465	0.417	0.490	17.710
BeLFusion	6.495	1.233	**0.306**	**0.446**	**0.400**	**0.461**	**1.956**
SittingDown							
TPK	7.393	**1.864**	0.496	0.678	0.531	0.666	**2.475**
DLow	12.044	4.576	0.451	0.605	0.495	0.606	6.147
GSPS	13.725	6.520	**0.406**	**0.561**	**0.461**	**0.565**	9.930
DivSamp	**14.899**	7.240	0.413	0.579	0.478	0.586	12.477
BeLFusion	9.026	2.236	0.413	0.585	0.468	0.587	2.853
Smoking							
TPK	6.522	1.807	0.422	0.529	0.509	0.560	**3.146**
DLow	11.549	4.058	0.400	0.515	0.490	0.537	4.989
GSPS	14.822	7.332	0.366	0.485	0.472	0.530	11.445
DivSamp	**15.688**	8.153	0.353	0.486	0.475	0.523	14.019
BeLFusion	6.780	**1.372**	**0.341**	**0.467**	**0.467**	**0.512**	3.787
Waiting							
TPK	6.631	2.080	0.480	0.584	0.526	0.568	3.794
DLow	11.680	3.398	0.441	0.541	0.497	0.534	4.336
GSPS	15.000	6.702	0.400	0.514	0.475	0.529	10.923
DivSamp	**15.455**	7.156	**0.387**	0.517	0.486	0.535	11.597
BeLFusion	7.747	**1.542**	0.390	**0.507**	**0.471**	**0.511**	**3.609**
WalkDog							
TPK	7.384	2.481	0.560	0.694	0.592	0.665	12.615
DLow	11.882	2.732	0.490	0.566	0.539	0.570	**7.967**
GSPS	13.746	4.569	0.459	0.564	0.530	0.587	8.864
DivSamp	**15.616**	6.212	0.439	0.555	0.532	0.577	8.104
BeLFusion	9.335	**1.893**	**0.432**	**0.530**	**0.527**	**0.569**	11.334
WalkTogether							
TPK	6.718	**1.791**	0.443	0.548	0.535	0.573	9.378
DLow	11.951	3.922	0.395	0.495	0.503	0.530	4.309
GSPS	15.030	6.994	0.316	0.440	**0.473**	0.516	10.010
DivSamp	**16.095**	8.060	0.321	0.458	0.486	0.525	10.683
BeLFusion	6.378	2.092	**0.296**	**0.393**	0.484	**0.495**	**4.138**
Walking							
TPK	6.708	**1.875**	0.455	0.533	0.538	0.558	9.540
DLow	11.904	3.507	0.428	0.518	0.516	0.539	4.429
GSPS	14.797	6.399	**0.351**	**0.469**	**0.490**	**0.528**	9.395
DivSamp	**15.964**	7.566	0.373	0.535	0.508	0.547	11.190
BeLFusion	5.116	3.345	0.367	0.471	0.530	0.546	**4.386**

Table 4. Comparison of BeLFusion with state-of-the-art methods on H36M. Bold and underlined results correspond to the best and second-best results, respectively. Lower is better for all metrics except APD.

D.1. Class- and dataset-wise results

Tab. 4 shows that BeLFusion achieves state-of-the-art results in most metrics in all H36M classes. We stress that our model is especially good at predicting the future in contexts where the observation strongly determines the following action. For example, when the person is *Smoking*, or *Phoning*, a model should predict a coherent future that also involves holding a cigar, or a phone. BeLFusion succeeds at it, showing improvements of 9.1%, 6.3%, and 3.7% for FDE with respect to other methods for *Eating*, *Phoning*, and *Smoking*, respectively. Our model also excels in classes where the determinacy of each part of the body needs to be assessed. For example, for *Directions*, and *Photo*, which often involve a static lower-body, and diverse upper-body

Datasets	APD	APDE	ADE	FDE	MMADE	MMFDE	CMD
DFaust							
TPK	8.998	**2.435**	0.591	0.555	0.637	0.601	8.263
DLow	12.805	2.755	0.521	0.505	0.565	0.539	**3.640**
GSPS	12.870	3.218	0.504	0.508	0.564	0.556	8.150
DivSamp	**25.016**	14.691	0.479	0.495	0.569	0.569	57.256
BeLFusion	9.285	2.456	**0.441**	**0.424**	**0.514**	**0.498**	14.174
DanceDB							
TPK	9.665	2.812	0.810	0.798	0.815	0.796	25.232
DLow	13.703	3.307	0.763	0.760	0.769	0.756	**18.800**
GSPS	11.792	3.121	0.747	0.764	0.758	0.765	27.113
DivSamp	**23.984**	13.008	0.757	0.815	0.777	0.818	31.244
BeLFusion	10.619	**2.780**	**0.690**	**0.713**	**0.709**	**0.717**	28.874
GRAB							
TPK	8.590	1.555	0.415	0.457	0.463	0.469	9.646
DLow	12.376	5.180	0.338	0.383	0.407	0.411	15.502
GSPS	13.515	6.331	0.300	0.381	0.404	0.435	11.642
DivSamp	**25.882**	18.686	0.287	0.394	0.407	0.447	76.817
BeLFusion	7.421	**1.111**	**0.260**	**0.323**	**0.375**	**0.388**	**1.321**
HUMAN4D							
TPK	9.451	2.618	0.657	0.732	0.662	0.705	6.305
DLow	13.083	4.571	0.562	0.629	0.583	0.612	**2.888**
GSPS	12.449	4.764	0.514	0.609	0.563	0.617	4.099
DivSamp	**24.665**	16.149	0.519	0.632	0.581	0.641	57.120
BeLFusion	9.262	**2.020**	**0.471**	**0.568**	**0.526**	**0.576**	10.909
SOMA							
TPK	9.823	3.166	0.806	0.835	0.798	0.817	20.689
DLow	13.761	3.402	0.726	0.746	0.722	0.737	**15.123**
GSPS	11.867	3.665	0.715	0.779	0.710	0.765	22.222
DivSamp	**24.131**	13.296	0.724	0.802	0.728	0.795	35.350
BeLFusion	10.765	**3.106**	**0.647**	**0.691**	**0.655**	**0.685**	23.727
SSM							
TPK	9.459	2.741	0.595	0.486	0.662	0.615	13.479
DLow	13.029	3.290	0.498	0.379	0.559	**0.466**	**8.491**
GSPS	12.973	3.467	0.490	0.412	0.556	0.504	12.369
DivSamp	**24.993**	14.164	0.474	0.416	0.580	0.568	56.610
BeLFusion	9.576	**1.916**	**0.433**	**0.356**	**0.502**	0.470	19.351
Transitions							
TPK	9.525	2.217	0.696	0.672	0.706	0.658	26.234
DLow	13.308	2.461	0.599	**0.538**	0.615	**0.550**	**21.308**
GSPS	12.169	2.470	0.636	0.642	0.655	0.648	27.634
DivSamp	**24.612**	14.092	0.648	0.724	0.687	0.725	33.953
BeLFusion	10.499	**2.085**	**0.577**	0.578	**0.611**	0.596	27.361

Table 5. Comparison of BeLFusion with state-of-the-art methods on AMASS. Bold and underlined results correspond to the best and second- best results, respectively. Lower is better for all metrics except APD.

movements, BeLFusion improves FDE by an 8.9%, and an 8.0%, respectively. We also highlight the adaptive APD that our model shows, in contrast to the constant variety of motions predicted by the state-of-the-art methods. Such effect is better observed in Fig. 10, where BeLFusion is the method that best replicates the intrinsic multimodal diversity of each class (i.e., APD of the multimodal ground truth, see Sec. 4.2). The variety of motions present in each AMASS dataset impedes such a detailed analysis. However, we also observe that the improvements with respect to the other methods are consistent across datasets (Tab. 5). The only dataset where BeLFusion is beaten in an accuracy metric (FDE) is Transitions, where the sequences consist of transitions among different actions, without any behavioral cue that allows the model to anticipate it. We also observe that our model yields a higher variability of APD across datasets that adapts to the sequence context, clearly depicted in Fig. 10 as well.

Regarding the CMD, Tab. 4 and 5 show how methods that promote highly diverse predictions are biased toward forecasting faster movements than the ones present in the dataset. Fig. 11 shows a clearer picture of this bias by plotting the average predicted displacement at all predicted frames. We observe how in all H36M classes, GSPS and DivSamp accelerate very early and eventually stop by the end of the prediction. We argue that such early divergent motion favors high diversity values, at expense of realistic transitions from the ongoing to the predicted motion. By contrast, BeLFusion produces movements that resemble those present in the dataset. While DivSamp follows a similar trend in AMASS than in H36M, GSPS does not. Although DLow is far from state-of-the-art accuracy, it achieves the best performance with regard to this metric in both datasets. Interestingly, BeLFusion slightly decelerates at the first frames and then achieves the motion closest to that of the dataset shortly after. We hypothesize that this effect is an artifact of the behavioral coupling step, where the ongoing motion smoothly transitions to the predicted behavior.

D.2. Ablation study: implicit diversity

As described in Sec. 3.3 and 4.3 of the main paper, by relaxing the loss regularization (i.e., increasing the number of predictions sampled at each training iteration, k), we can increase the diversity of BeLFusion's predictions. We already showed that by increasing k, the diversity (APD), accuracy (ADE, FDE), and realism (FID) improves. In fact, for large k (> 5), a single denoising step becomes enough to achieve state-of-the-art accuracy. Still, going through the whole reverse Markov diffusion chain helps the predicted behavior code move closer to the latent space manifold, thus generating more realistic predictions. In Fig. 12, we include the same analysis for all the models in the ablation study of the main paper. The results prove that the implicit diversity effect is not exclusive of either BeLFusion's loss or behavioral latent space.

E. Examples *in motion*

For each dataset, we include several videos where 10 predictions of BeLFusion are compared to those of methods showing competitive performance for H36M: TPK [60], DLow [65], GSPS [39], and DivSamp [15]. Videos are identified as '[dataset]_[sample_id]_[class/subdataset]'. For example, 'A_6674_GRAB' is sample 6674, which is part of

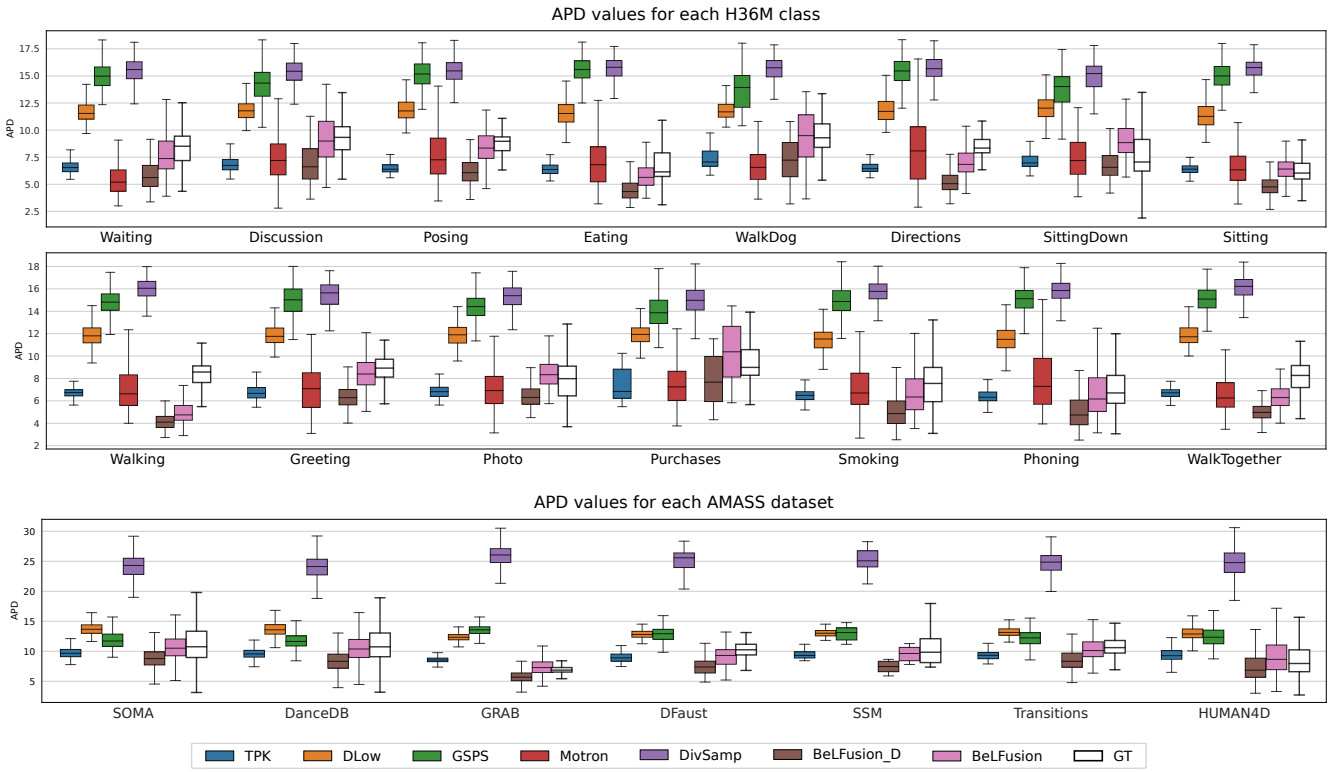

Figure 10. **Class- and dataset-wise APD.** GT corresponds to the APD of the multimodal ground truth. BeLFusion is the only method that adjusts the diversity of its predictions to model the intrinsic diversity of each class and dataset. As a result, the APD distributions between BeLFusion and GT are very similar.

the GRAB [56] dataset within AMASS (prefix 'A_'), and 'H_1246_Sitting' is the sample 1246, which is part of a 'Sitting' sequence of H36M (prefix 'H_'). The *Context* column shows the observed sequence and freezes at the last observed pose. The *GT* column shows the ground truth motion.

In this section, we discuss the visual results by highlighting the main advantages provided by BeLFusion and showing some failure examples.

Realistic transitioning. By means of the behavior coupler, BeLFusion is able to transfer predicted behaviors to any ongoing motion with high realism. This is supported quantitatively by the FID and CMD metrics, and perceptually by our qualitative assessment (Sec. 4.3). Now, we assess it by visually inspecting several examples. For example, when the observation shows an ongoing fast motion ('H_608_Walking', 'H_1928_Eating' or 'H_2103_Photo'), BeLFusion is the only model that consistently generates a coherent transition between the observation and the predicted behavior. Other methods mostly predict a sudden stop of the previous action. This is also appreciated in the cross-dataset evaluation. For example, although the observation window of the 'A_103_Transitions' clearly showcases a fast rotational dancing step, none of the state-of-the-art methods are able to generate a plausible continuation of the observed motion, and all of their predictions abruptly stop rotating. BeLFusion is the only method that generates predictions that slowly decrease its rotational momentum to start performing a different action. A similar effect is observed in 'A_2545_DanceDB', and 'A_10929_HUMAN4D'.

Context-driven prediction. BeLFusion's state-of-the-art APDE and CMD metrics show its superior ability to adjust both the *motion speed* and *motion determinacy* to the observed context. This results in sets of predictions that are, overall, more coherent with respect to the observed context. For example, whereas for 'H_4_Sitting' BeLFusion's predicted motions showcase a high variety of arms-related actions, its predictions for sequences where the arms are used in an ongoing action ('H_402_Smoking', 'H_446_Smoking', and 'H_541_Phoning') have a more limited variety of arms motion. In contrast, predictions from state-of-the-art methods do not have such behavioral consistency with respect to the observed motion. This is more evident in diversity-promoting methods like DLow, GSPS, and DivSamp, where the motion predicted is usually implausible for a person that is smoking or making a phone call. Similarly, in 'H_962_WalkTogether', our method predicts motions that are compatible with the ongoing action of walking next to someone, whereas other methods ignore such possibility. In AMASS, BeLFusion's capability to adapt to the context

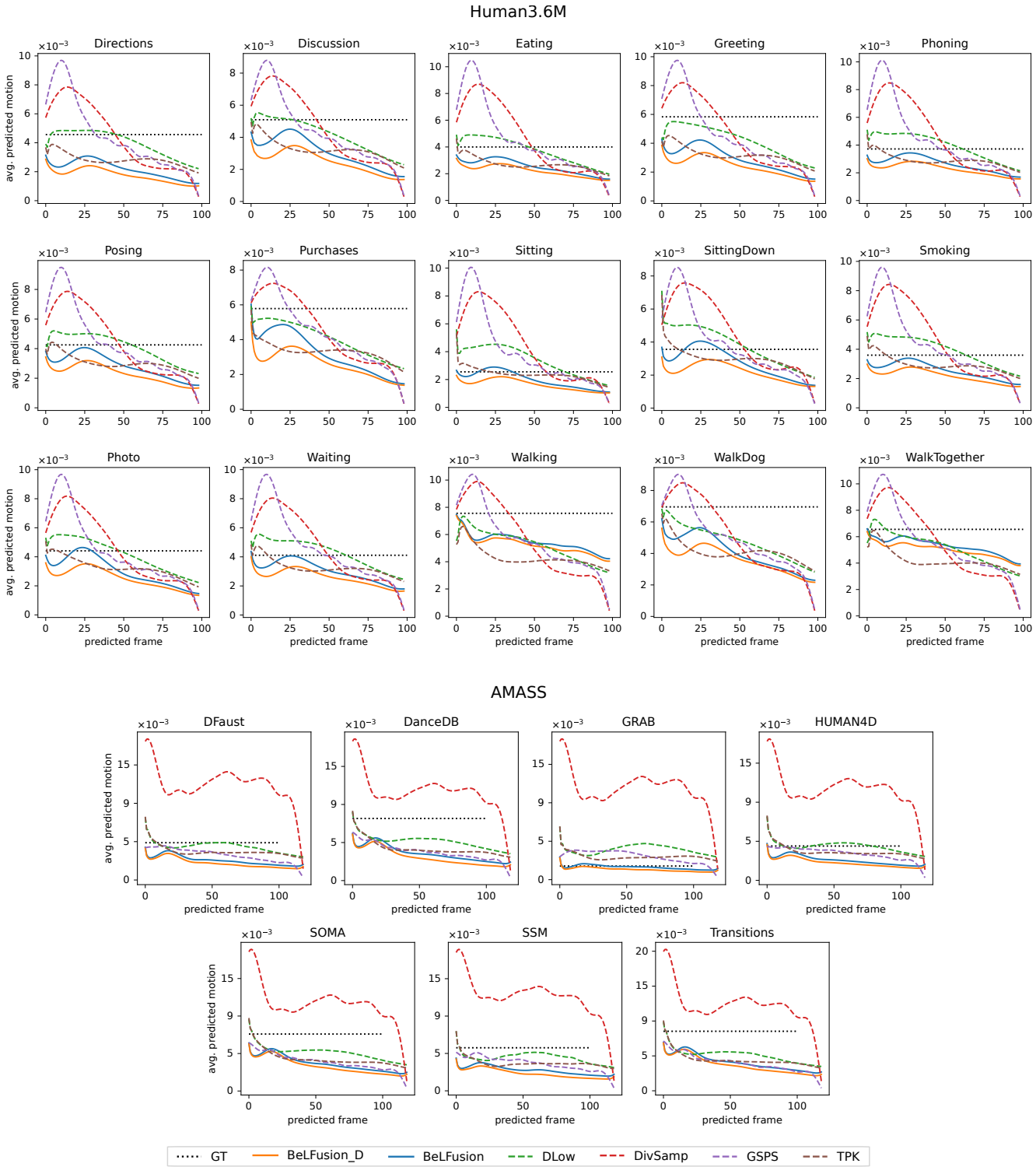

Figure 11. **Predicted motion analysis.** For each timestep in the future (predicted frame), the plots above show the displacement predicted averaged across all test sequences. For H36M, GSPS and DivSamp predictions accelerate in the beginning, leading to unrealistic transitions. For AMASS, DivSamp shows a similar behavior, and DLow beats all methods except in GRAB, where BeLFusion matches very well the average dataset motion.

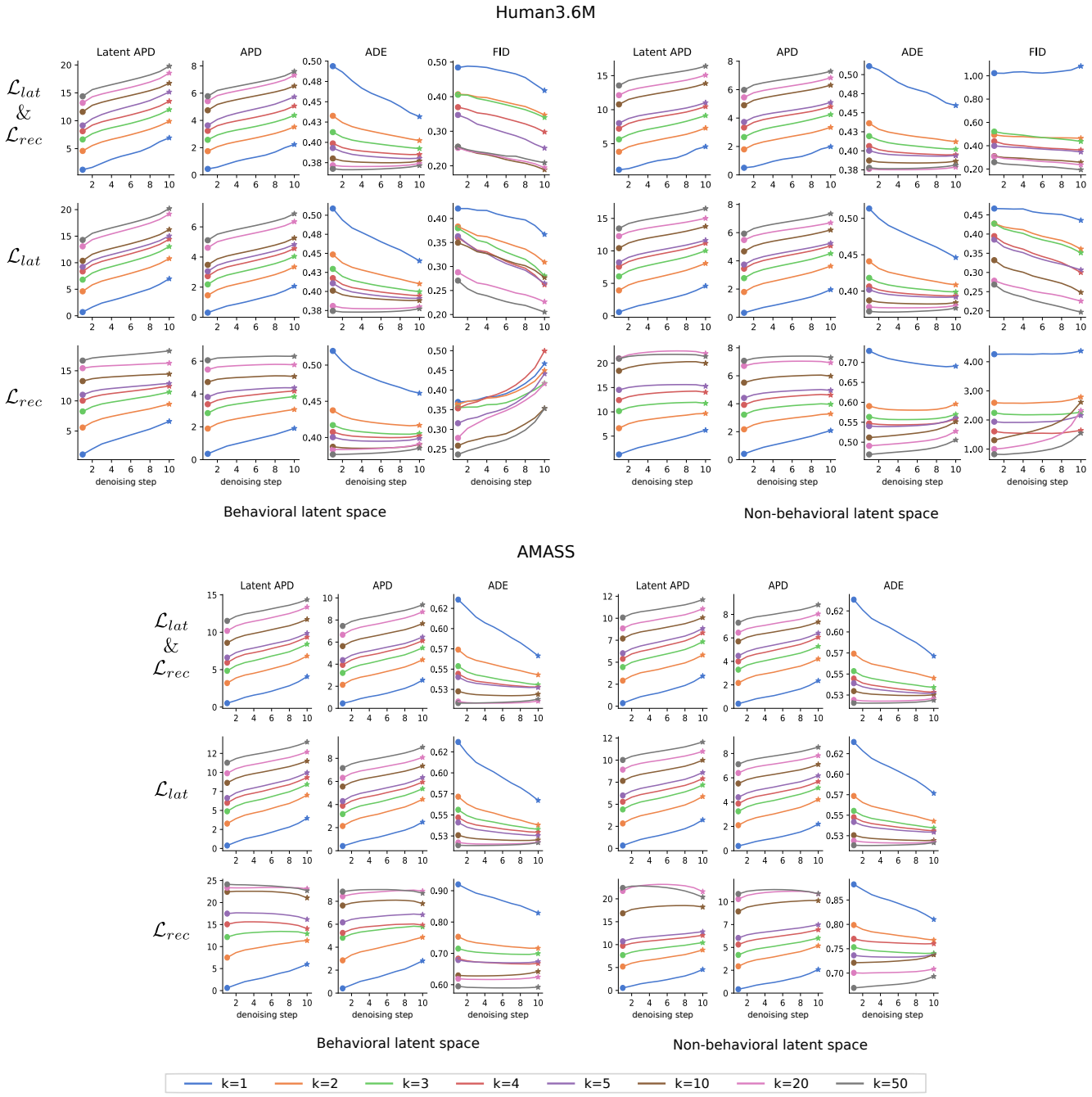

Figure 12. **Implicit diversity.** By increasing the value of k, the diversity is implicitly promoted in both the latent and reconstructed spaces (Latent APD, and APD). We observe that this is effect is not particular to the loss choice ($\mathcal{L}_{lat}$, $\mathcal{L}_{rec}$, or both) or the latent space construction (behavioral or not). Using the LDM to reverse the whole Markov chain of 10 steps (x-axis) helps improve diversity (APD), accuracy (ADE), and realism (FID) in general. Note that for $k > 5$, only the diversity and the realism are further improved, and a single denoising step becomes enough to generate the most accurate predictions.

is clearly depicted in sequences with low-range motion, or where motion is focused on particular parts of the body. For example, BeLFusion adapts the diversity of predictions to the 'grabbing' action present in the GRAB dataset. While other methods predict coordinate-wise diverse inaccurate predictions, our model encourages diversity within the short spectrum of the plausible behaviors that can follow (see 'A_7667_GRAB', 'A_7750_GRAB', or 'A_9274_GRAB'). In fact, in 'A_11074_HUMAN4D' and 'A_12321_SOMA', our model is the only able to anticipate the intention of

laying down by detecting subtle cues inside the observation window (samples #6 and #8). In general, BeLFusion provides good coverage of all plausible futures given the contextual setting. For example, in 'H_910_SittingDown', and 'H_861_SittingDown' our model's predictions contain as many different actions as all other methods, with no realism trade-off as for GSPS or DivSamp.

Generalization to unseen contexts. As a result of the two properties above (realistic transitioning and context-driven prediction), BeLFusion shows superior generalization to unseen situations. This is quantitatively supported by the big step forward in the results of the cross-dataset evaluation. Such generalization capabilities are especially perceptible in the DanceDB[6] sequences, which include dance moves unseen at training time. For instance, 'A_2054_DanceDB' shows how BeLFusion can predict, up to some extent, the correct continuation of a dance move, while other methods either almost freeze or simply predict an out-of-context movement. Similarly, 'A_2284_DanceDB' and 'A_1899_DanceDB' show how BeLFusion is able to detect that the dance moves involve keeping the arms arising while moving or rotating. In comparison, DLow, GSPS, and DivSamp simply predict other unrelated movements. TPK is only able to predict a few samples with fairly good continuations to the dance step. Also, in 'A_12391_SOMA', BeLFusion is the only method able to infer how a very challenging repetitive stretching movement will follow.

We also include some examples where our model fails to generate a coherent and plausible set of predictions. This mostly happens under aggressive domain shifts. For example, in 'A_1402_DanceDB', the first-seen handstand behavior in the observation leads to BeLFusion generating several wrong movement continuations. Similarly to the other state-of-the-art methods, BeLFusion also struggles with modeling high-frequencies. For example, in 'A_1087_DanceDB', the fast legs motion during the observation is not reflected in any prediction, although BeLFusion slightly shows it in samples #4 and #7. Even though less clearly, this is also observed in H36M. For example, in 'H_148_WalkDog', none of the models is able to model the high-speed walking movement from the ground truth. Robustness against huge domain drifts and modeling of high-frequencies are interesting and challenging limitations that need to be addressed as future work.

F. Qualitative assessment

Selection criteria. In order to ensure the assessment of a wide range of scenarios, we randomly sampled from three sampling pools per dataset. To generate them, we first ordered all test sequences according to the average joint displacement D_i in the last 100 ms of observation. Then, we selected the pools by taking sequences with D_i within 1) the top 10% (high-speed transition), 2) 40-60% (medium-speed transition), and 3) the bottom 10% (low-speed transition). Then, 8 sequences were randomly sampled for each group. A total of 24 samples for each dataset were selected. These were randomly distributed in groups of 4 and used to generate 6 tests per dataset. Since each dataset has different joint configurations, we did not mix samples from both datasets in the same test to avoid confusion.

Assessment details. The tests were built with the *JotForm*[7] platform. Users accessed it through a link generated with *NimbleLinks*[8], which randomly redirected them to one of the tests. Fig. 13 shows an example of the instructions and definition of realism shown to the user before starting the test (left), and an example of the interface that allowed the user to order the methods according to the realism showcased (right). Note that the instructions showed either AMASS or H36M ground truth samples, as both skeletons have a different number of joints. A total of 126 people answered the test, with 67 participating in the H36M study, and 59 participating in the AMASS one.

Extended results. Extended results for the qualitative study are shown in Tab. 6. We also show the results for each sampling pool, i.e., grouping sequences by the speed of the transition. The average rank was computed as the average of all samples' mean ranks, and the 1st/2nd/3rd position percentages as the number of times a sample was placed at 1st/2nd/3rd position over the total amount of samples available. We observe that the realism superiority of BeLFusion is particularly notable in the sequences with medium-speed transitions (77.0% and 64.9% ranked first in H36M and AMASS, respectively). We argue that this is partly promoted by the good capabilities of the behavior coupler to adapt the prediction to the movement speed and direction observed. This is also seen in the high-speed set (ranked third only in 9.8% and 14.1% of the cases), despite GSPS showing competitive performance on it.

[6] Dance Motion Capture DB, http://dancedb.cs.ucy.ac.cy.

[7] https://www.jotform.com/

[8] https://www.nimblelinks.com/

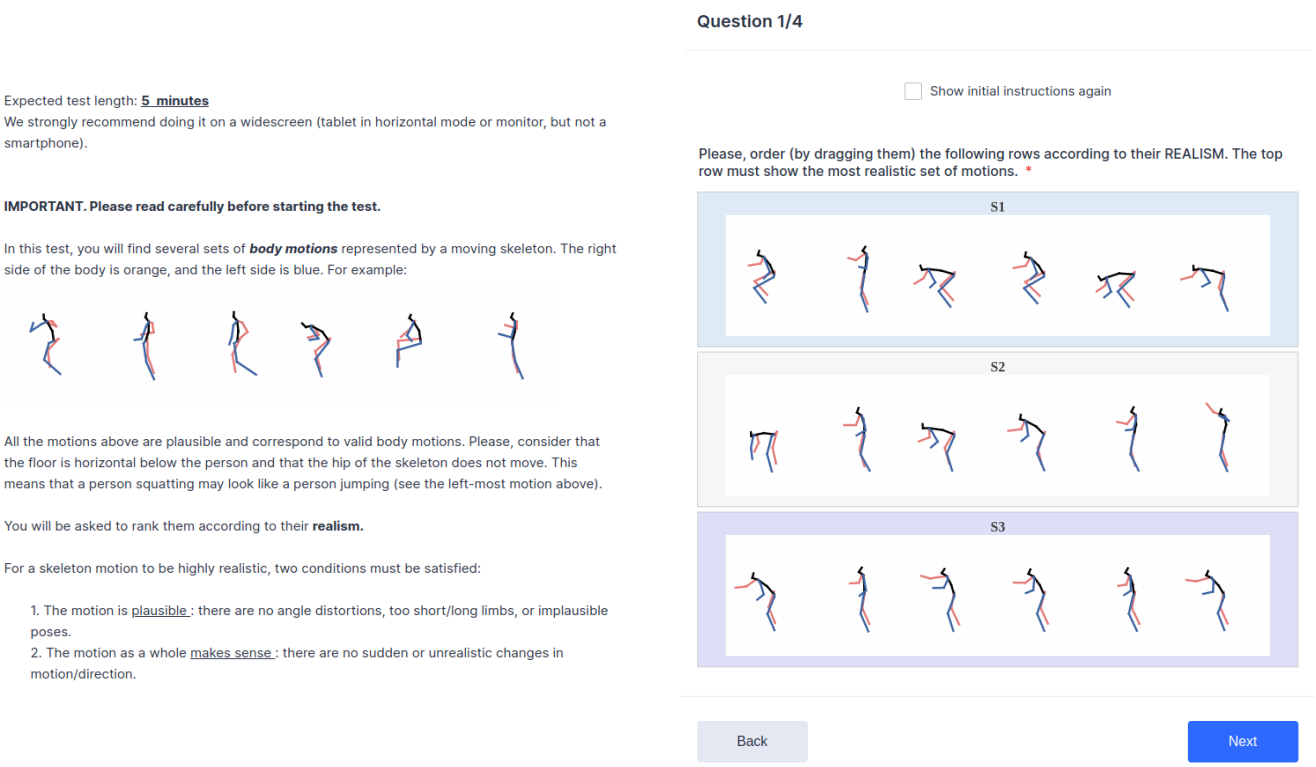

Figure 13. **Questionnaire example.** On the left, instructions shown to the participant at the beginning. On the right, the interface for ranking the skeleton motions. All skeletons correspond to *gif* images that repeatedly show the observation and prediction motion sequences.

	Human3.6M [26]				AMASS [37]			
	Avg. rank	Ranked 1st	Ranked 2nd	Ranked 3rd	Avg. rank	Ranked 1st	Ranked 2nd	Ranked 3rd
Low-speed transition								
GSPS	2.238 ± 0.305	18.0%	**40.4%**	41.6%	2.156 ± 0.595	22.6%	**38.1%**	39.3%
DivSamp	2.276 ± 0.459	15.7%	39.3%	**44.9%**	2.210 ± 0.373	23.8%	31.0%	**45.2%**
BeLFusion	**1.486 ± 0.225**	**66.3%**	20.2%	13.5%	**1.634 ± 0.294**	**53.6%**	31.0%	15.5%
Medium-speed transition								
GSPS	2.305 ± 0.466	13.8%	**48.3%**	37.9%	2.025 ± 0.449	24.3%	**50.0%**	25.7%
DivSamp	2.396 ± 0.451	9.2%	36.8%	**54.0%**	2.497 ± 0.390	10.8%	28.4%	**60.8%**
BeLFusion	**1.299 ± 0.243**	**77.0%**	14.9%	8.0%	**1.478 ± 0.424**	**64.9%**	21.6%	13.5%
High-speed transition								
GSPS	2.194 ± 0.320	21.7%	**40.2%**	38.0%	1.828 ± 0.468	44.9%	35.9%	19.2%
DivSamp	2.345 ± 0.292	15.2%	32.6%	**52.2%**	2.589 ± 0.409	6.4%	26.9%	**66.7%**
BeLFusion	**1.461 ± 0.149**	**63.0%**	27.2%	9.8%	**1.583 ± 0.287**	**48.7%**	**37.2%**	14.1%
All								
GSPS	2.246 ± 0.358	17.9%	**42.9%**	39.2%	2.003 ± 0.505	30.5%	**41.1%**	28.4%
DivSamp	2.339 ± 0.393	13.4%	36.2%	**50.4%**	2.432 ± 0.408	14.0%	28.8%	**57.2%**
BeLFusion	**1.415 ± 0.217**	**68.7%**	20.9%	10.4%	**1.565 ± 0.332**	**55.5%**	30.1%	14.4%

Table 6. **Qualitative assessment.** 126 participants ranked sets of samples from GSPS, DivSamp, and BeLFusion by their realism. Lower average rank (± std. dev.) is better.