BeLFusion: Latent Diffusion for Behavior-Driven Human Motion Prediction

—— Supplementary Material ——

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

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 [10], with PyTorch 1.9.1 [8] 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.

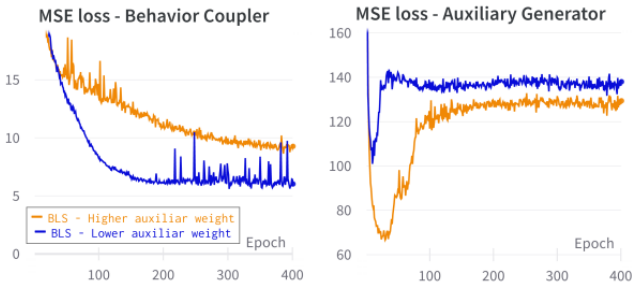

Figure A. **Behavioral disentanglement.** Main (left) and adversarial (right) training losses of the behavioral latent space. As expected, when the auxiliary loss weight is higher (orange), the adversarial interplay intensifies.

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 layers (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 $\mathcal{A}_\omega$ is a clone of $\mathcal{B}_\phi$ with a narrower input shape (128), as only the latent code is fed. Note that the adversarial interplay introduces additional complexity, making convergence more challenging, see Fig. A. For H36M, the behavioral VAE was trained with learning rates of 0.005 and 0.0005 for $\mathcal{L}_{main}$ and $\mathcal{L}_{aux}$, re-

spectively. For AMASS, they were set to 0.001 and 0.005. All learning rates were 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 [3]. To leverage it, the target latent codes were reshaped to a rectangular shape (16x8), as prior work proposed [1]. In particular, our U-Net has 2 attention layers (resolutions of 8 and 4), 16 channels per attention head, a FiLM-like conditioning mechanism [9], 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 [5], which also provided important improvements in our scenario compared to the classic *linear* or *cosine* schedules (see Fig. B). 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. 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 [11].

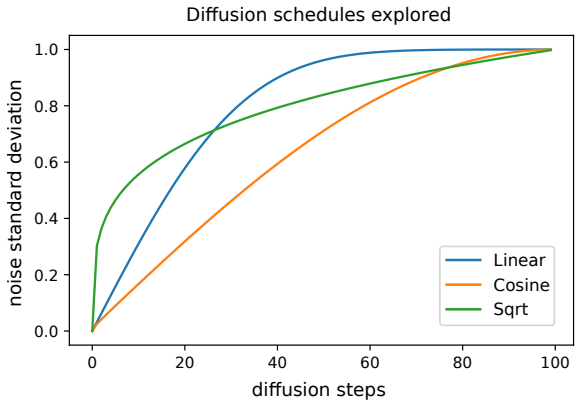

Figure B. **Diffusion schedules.** Schedules explored for diffusing the target latent codes.

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.

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 [6]. However, we also include the new datasets added afterward and with available SMPL+H

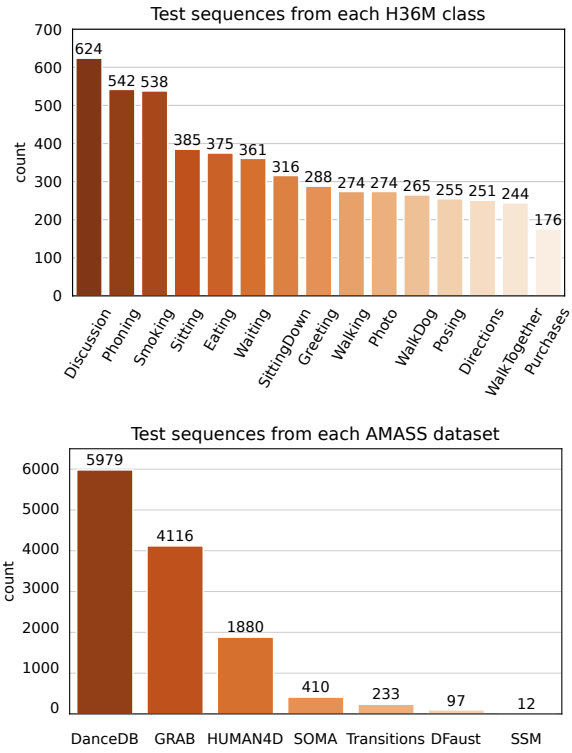

Figure C. **Test set sequences.** We show the number of test sequences evaluated for each class/dataset in H36M/AMASS.

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

model parameters, 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 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. C 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 [15].

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. D shows a 2-dimensional t-SNE projection of all behavioral encodings of the H36M test sequences [13]. 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[1] 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. Then, we present an extended analysis of the effect of k, which controls the loss *relaxation* level (Sec. D.2). Finally, we compare the inference time of BeLFusion to the state of the art (Sec. D.3).

D.1. Class- and dataset-wise results

Tab. A 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 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. E, 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. B). 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. E as well.

Regarding the CMD, Tab. A and B show how methods that promote highly diverse predictions are biased toward forecasting faster movements than the ones present in the dataset. Fig. F 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

[1] Videos referenced in the supp. material can be found in: https://barquerogerman.github.io/BeLFusion/.

Classes	APD	APDE	ADE	FDE	MMADE	MMFDE	CMD	FID
Directions								
TPK	6.510	2.039	0.447	0.482	0.523	0.544	7.455	1.768
DLow	11.874	3.359	0.415	0.465	0.499	0.514	**2.011**	4.633
GSPS	15.398	6.877	0.407	0.477	0.492	0.522	10.469	4.827
DivSamp	**15.663**	7.142	0.389	0.463	0.502	0.523	10.539	5.489
BeLFusion	7.090	**1.709**	**0.378**	**0.422**	**0.484**	**0.494**	10.110	**1.150**
Discussion								
TPK	6.966	2.572	0.511	0.581	0.570	0.600	7.554	1.090
DLow	11.872	2.659	0.472	0.536	0.533	0.549	**2.695**	1.300
GSPS	14.199	4.992	0.448	0.541	0.526	0.563	8,470	1.870
DivSamp	**15.310**	5.905	0.432	0.526	0.534	0.557	8.975	1.522
BeLFusion	9.172	**1.425**	**0.420**	**0.507**	**0.512**	**0.530**	7.521	**1.055**
Eating								
TPK	6.412	**1.066**	0.388	0.473	0.452	0.472	5.306	4.345
DLow	11.603	4.829	0.358	0.433	0.439	0.452	**3.214**	10.300
GSPS	15.570	8.793	0.334	0.419	0.424	0.448	12.360	11.322
DivSamp	**15.681**	8.904	0.321	0.419	0.428	0.445	13.863	10.270
BeLFusion	5.954	1.297	**0.310**	**0.381**	**0.418**	**0.420**	5.808	**1.439**
Greeting								
TPK	6.779	2.545	0.555	0.615	0.571	0.598	12.313	**2.148**
DLow	11.897	3.112	0.530	0.590	0.542	0.564	**5.994**	3.724
GSPS	14.974	5.950	0.502	0.592	0.532	0.577	10.654	5.488
DivSamp	**15.447**	6.373	0.489	0.579	0.535	0.562	9.044	4.848
BeLFusion	8.482	**1.690**	**0.482**	**0.544**	**0.524**	**0.540**	12.740	2.201
Phoning								
TPK	6.410	**1.400**	0.377	0.475	0.468	0.507	**3.057**	1.882
DLow	11.542	4.605	0.343	0.444	0.451	0.487	4.886	4.847
GSPS	15.050	8.120	0.311	0.413	0.436	0.476	12.292	6.458
DivSamp	**15.751**	8.813	0.296	0.400	0.437	0.471	14.295	5.149
BeLFusion	6.649	1.477	**0.283**	**0.375**	**0.426**	**0.445**	3.388	**0.836**
Photo								
TPK	6.894	1.884	0.541	0.689	0.548	0.633	**3.928**	3.231
DLow	11.931	4.180	0.507	0.655	0.516	0.596	4.013	3.305
GSPS	14.310	6.482	0.485	0.663	0.502	0.606	10.855	3.851
DivSamp	**15.330**	7.428	0.474	0.665	0.506	0.607	11.427	4.571
BeLFusion	8.446	**1.726**	**0.434**	**0.601**	**0.462**	**0.546**	4.491	**2.526**
Posing								
TPK	6.520	2.310	0.466	0.538	0.542	0.565	4.740	**1.279**
DLow	11.875	3.116	0.442	0.521	0.510	**0.525**	**3.421**	2.521
GSPS	15.149	6.399	0.415	0.527	**0.498**	0.543	10.720	4.967
DivSamp	**15.429**	6.676	**0.395**	**0.499**	0.510	0.541	11.201	4.143
BeLFusion	8.438	**1.241**	0.406	0.510	**0.498**	0.531	4.729	1.463
Purchases								
TPK	7.450	2.161	0.505	0.522	0.535	0.538	10.298	7.194
DLow	11.947	2.629	0.430	0.422	**0.493**	0.477	**5.090**	6.871
GSPS	13.969	4.552	0.414	0.429	0.497	0.497	7.380	6.521
DivSamp	**14.967**	5.517	**0.388**	**0.404**	0.502	0.478	6.950	**3.758**
BeLFusion	10.272	**1.738**	0.410	0.409	0.494	**0.472**	8.800	5.696
Sitting								
TPK	6.417	**1.167**	0.400	0.547	0.461	0.548	**1.542**	**1.619**
DLow	11.425	4.972	0.364	0.513	0.440	0.523	7.490	3.290
GSPS	14.966	8.494	0.323	0.454	0.411	0.484	14.377	5.717
DivSamp	**15.614**	9.146	0.317	0.465	0.417	0.490	16.828	3.485
BeLFusion	6.495	1.233	**0.306**	**0.446**	**0.400**	**0.461**	1.957	1.836
SittingDown								
TPK	7.393	**1.864**	0.496	0.678	0.531	0.666	**2.889**	**1.987**
DLow	12.044	4.576	0.451	0.605	0.495	0.606	5.651	2.759
GSPS	13.725	6.520	**0.406**	**0.561**	**0.461**	**0.565**	9.301	3.694
DivSamp	**14.899**	7.240	0.413	0.579	0.478	0.586	11.929	3.471
BeLFusion	9.026	2.236	0.413	0.585	0.468	0.587	2.997	2.642
Smoking								
TPK	6.522	1.807	0.422	0.529	0.509	0.560	**3.148**	1.652
DLow	11.549	4.058	0.400	0.515	0.490	0.537	5.123	3.535
GSPS	14.822	7.332	0.366	0.485	0.472	0.530	11.478	4.622
DivSamp	**15.688**	8.153	0.353	0.486	0.475	0.523	14.041	4.258
BeLFusion	6.780	**1.372**	**0.341**	**0.467**	**0.467**	**0.512**	3.849	**0.847**
Waiting								
TPK	6.631	2.080	0.480	0.584	0.526	0.568	4.143	1.022
DLow	11.680	3.398	0.441	0.541	0.497	0.534	**3.866**	1.758
GSPS	15.000	6.702	0.400	0.514	0.475	0.529	10.686	3.277
DivSamp	**15.455**	7.156	**0.387**	0.517	0.486	0.535	11.611	3.108
BeLFusion	7.747	**1.542**	0.390	**0.507**	**0.471**	**0.511**	4.186	**0.981**
WalkDog								
TPK	7.384	2.481	0.560	0.694	0.592	0.665	13.157	3.395
DLow	11.882	2.732	0.490	0.566	0.539	0.570	8.495	3.019
GSPS	13.746	4.569	0.459	0.564	0.530	0.587	8.869	2.647
DivSamp	**15.616**	6.212	0.439	0.555	0.532	0.577	**8.177**	**1.979**
BeLFusion	9.335	**1.893**	**0.432**	**0.530**	**0.527**	**0.569**	11.908	3.193
WalkTogether								
TPK	6.718	**1.791**	0.443	0.548	0.535	0.573	10.814	14.715
DLow	11.951	3.922	0.395	0.495	0.503	0.530	**5.234**	20.315
GSPS	15.030	6.994	0.316	0.440	**0.473**	0.516	10.265	22.212
DivSamp	**16.095**	8.060	0.321	0.458	0.486	0.525	10.584	19.643
BeLFusion	6.378	2.092	**0.296**	**0.393**	0.484	**0.495**	5.613	**4.348**
Walking								
TPK	6.708	**1.875**	0.455	0.533	0.538	0.558	14.279	16.210
DLow	11.904	3.507	0.428	0.518	0.516	0.539	**8.400**	20.670
GSPS	14.797	6.399	**0.351**	**0.469**	**0.490**	**0.528**	10.352	19.394
DivSamp	**15.964**	7.566	0.373	0.535	0.508	0.547	10.024	17.166
BeLFusion	5.116	3.345	0.367	0.471	0.530	0.546	8.588	**3.784**

Table A. 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.

predictions sampled at each training iteration, k), we can increase the diversity of BeLFusion's predictions. We argue that by backpropagating each loss only on its best prediction out of k (Eq. 6): 1) wrong predictions are not penalized, and 2) correct predictions of less frequent behaviors are rewarded. In the disentangled BLS, distinct behaviors are spread out (Fig. D). Thus, high k's implicitly encourage models denoising $\mathcal{N}(0, 1)$ into a distribution with multiple behavioral modes (i.e., into behaviorally diverse futures).

We already showed that by increasing k, the diversity (APD), accuracy (ADE, FDE), and realism (FID) of BeLFusion 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

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 B. 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.

predictions. In Fig. G, 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.

D.3. Inference times

We computed the time it takes BeLFusion and the state-of-the-art models to generate 50 samples for a single prediction on a single GTX 1080Ti GPU. We averaged the values across 50 runs of 100 sequences. BeLFusion (320/323ms for H36M/AMASS) is slower than BeGAN (17/20ms), TPK (30/38ms), DLow(34/43ms), GSPS(5/7ms), and DivSamp (6/9ms). Despite BeLFusion's slower inference (discussed as limitation in Sec. 5,), its z_0 parametrization allows it to be early-stopped and run, if needed, in real-time (BeLFusion_D, 48/52ms) with similar accuracy in return for less diversity and worse APDE, FID, and CMD.

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 [14], DLow [15], GSPS [7], and DivSamp [2]. Videos are identified as '[dataset]_[sample_id]_[class/subdataset]'. For example, 'A_6674_GRAB' is sample 6674, which is part of the GRAB [12] 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 the ongoing motion's 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

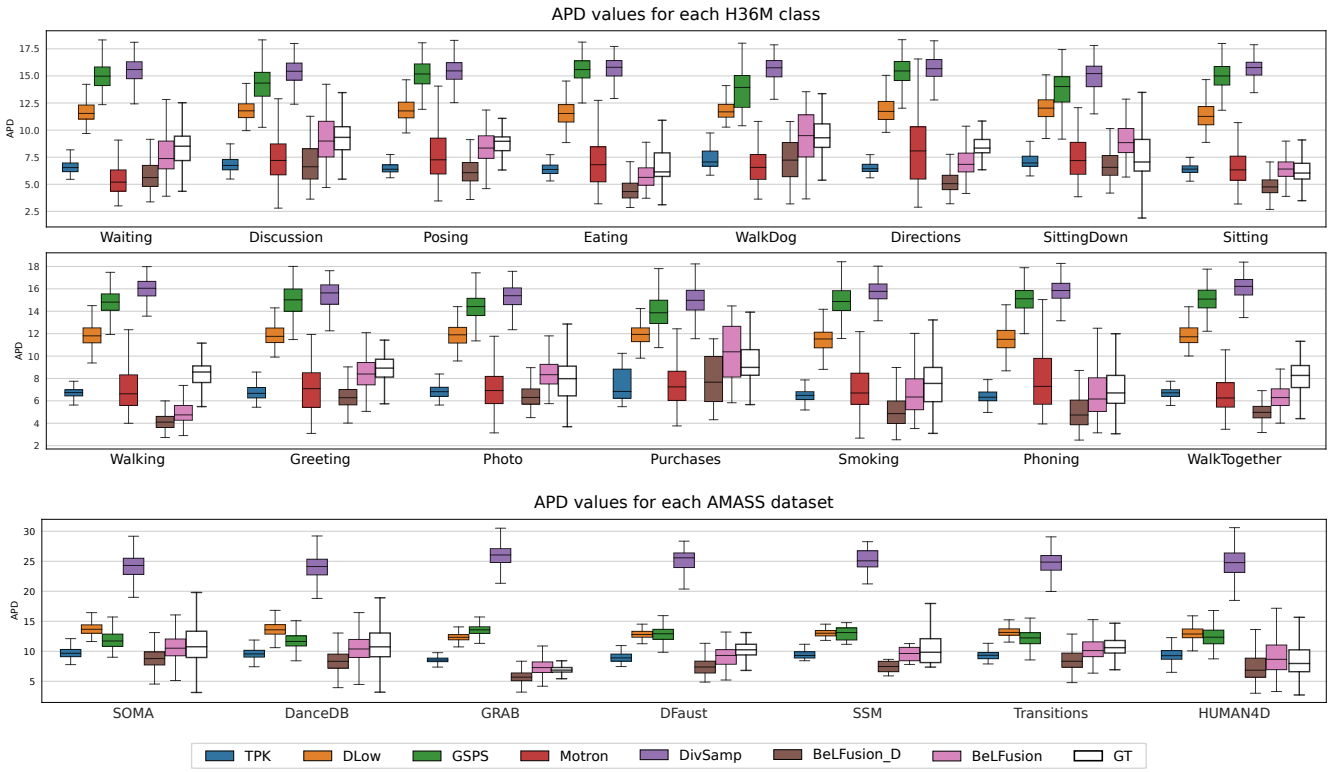

Figure E. **Class- and dataset-wise APD.** GT corresponds to the APD of the multimodal ground truth. BeLFusion is the only method that adjusts the diversity of its predictions to model the intrinsic diversity of each class and dataset. As a result, the APD distributions between BeLFusion and GT are very similar. The proposed APDE metric quantifies such similarity (the lower, the more similar).

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 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[2] 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

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

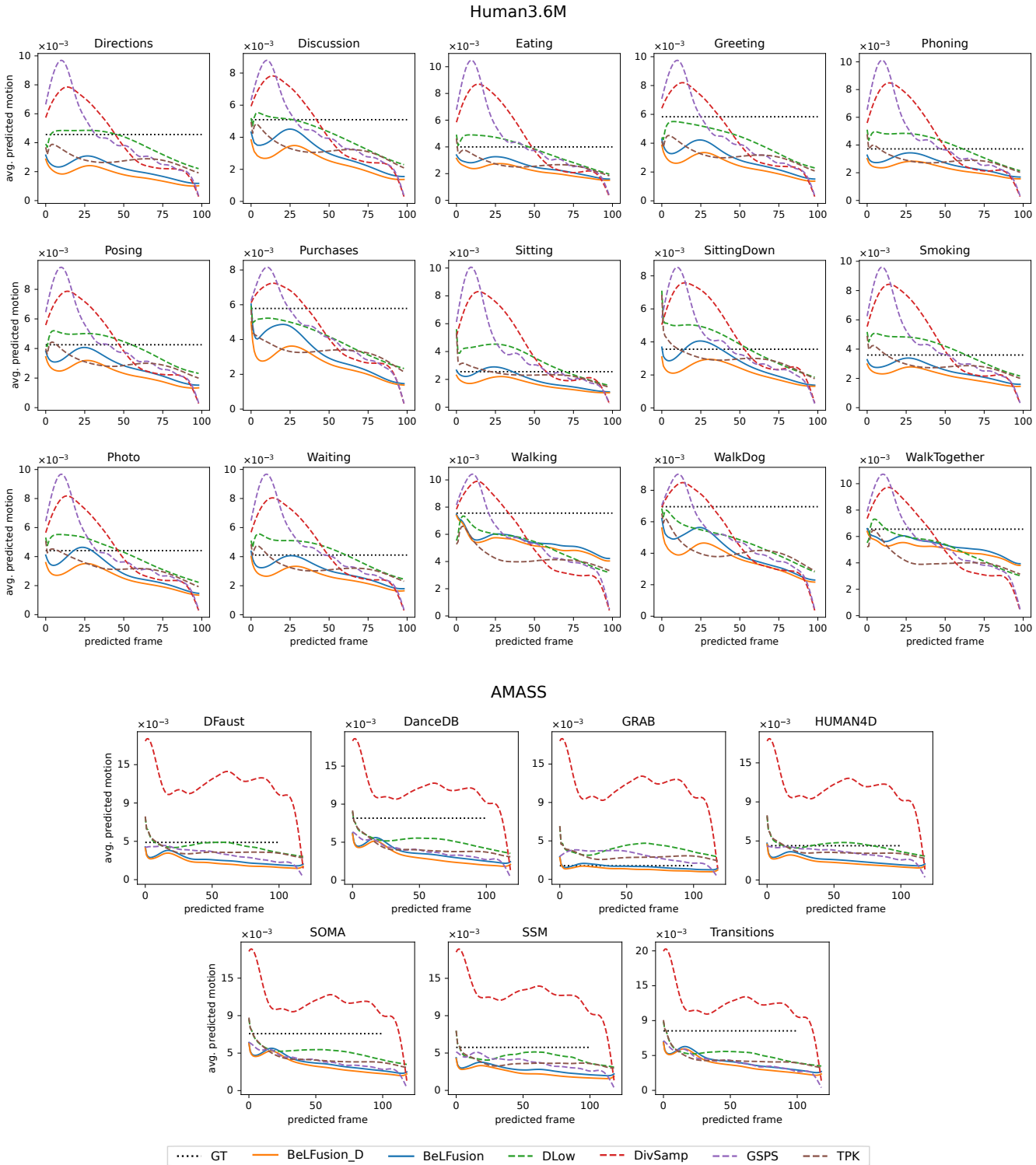

Figure F. **Predicted motion analysis.** For each timestep in the future (predicted frame), the plots above show the displacement predicted averaged across all test sequences. For H36M, GSPS and DivSamp predictions accelerate in the beginning, leading to unrealistic transitions. For AMASS, DivSamp shows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

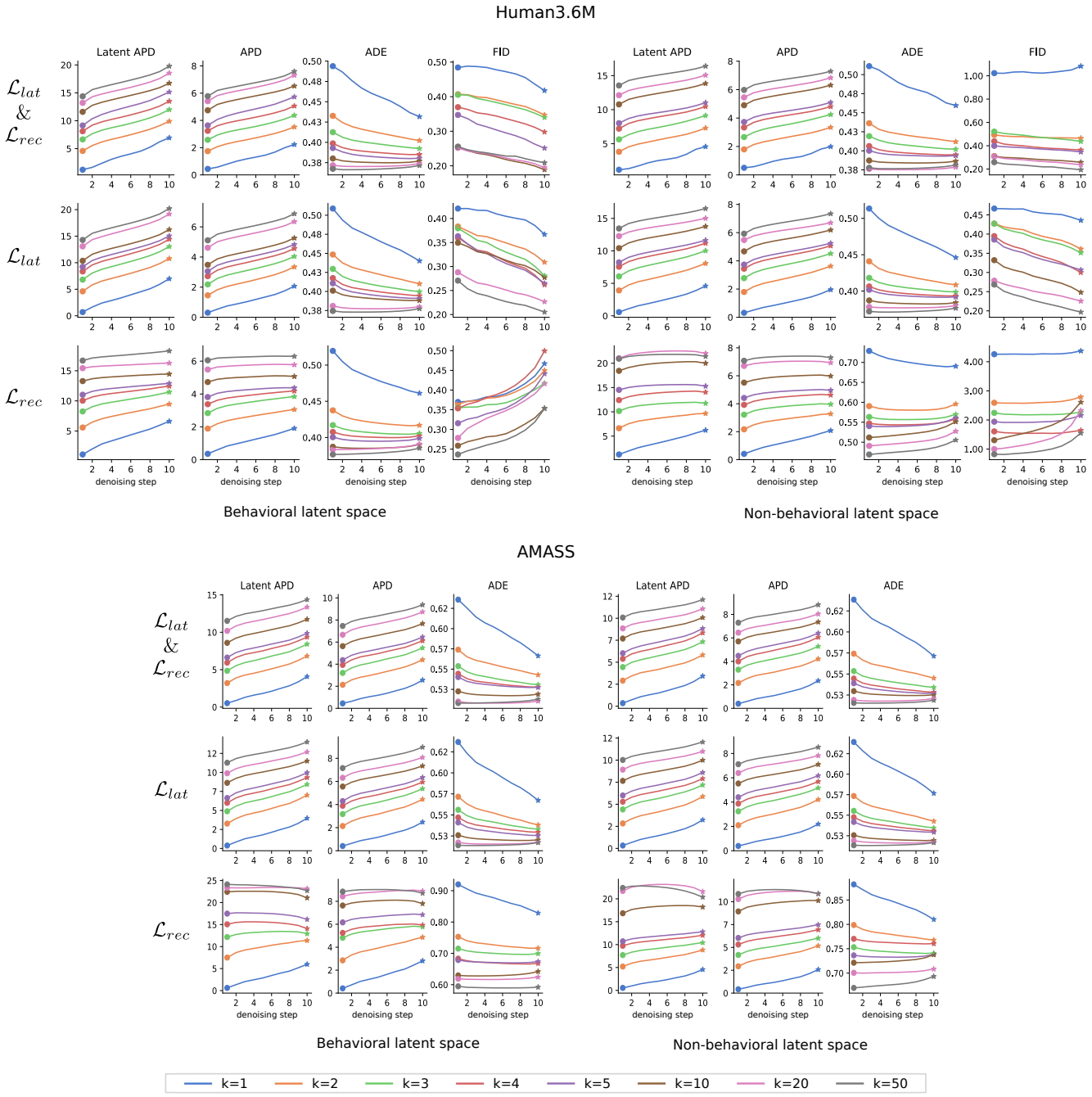

Figure G. **Implicit diversity.** By increasing the value of k, the diversity is implicitly promoted in both the latent and reconstructed spaces (Latent APD, and APD). We observe that this effect is not particular to the loss choice ($\mathcal{L}_{lat}$, $\mathcal{L}_{rec}$, or both) or the latent space 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.

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*[3] platform. Users accessed it through a link generated with *NimbleLinks*[4], which randomly redirected them to one of the tests. Fig. H 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. C. 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.

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

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

References

[1] Miguel Ángel Bautista, Pengsheng Guo, Samira Abnar, Walter Talbott, Alexander T Toshev, Zhuoyuan Chen, Laurent Dinh, Shuangfei Zhai, Hanlin Goh, Daniel Ulbricht, et al. Gaudi: A neural architect for immersive 3d scene generation. In *Advances in Neural Information Processing Systems*, 2022. 2

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

[3] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34, 2021. 2

[4] 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. 11

[5] Xiang Lisa Li, John Thickstun, Ishaan Gulrajani, Percy Liang, and Tatsunori Hashimoto. Diffusion-lm improves controllable text generation. In *Advances in Neural Information Processing Systems*, 2022. 2

[6] 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, 11

[7] 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. 6

[8] 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. 1

[9] 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*, 2018. 2

[10] Sashank J Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of adam and beyond. *arXiv preprint arXiv:1904.09237*, 2019. 1

[11] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *International Conference on Learning Representations*, 2021. 2

[12] 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. 6

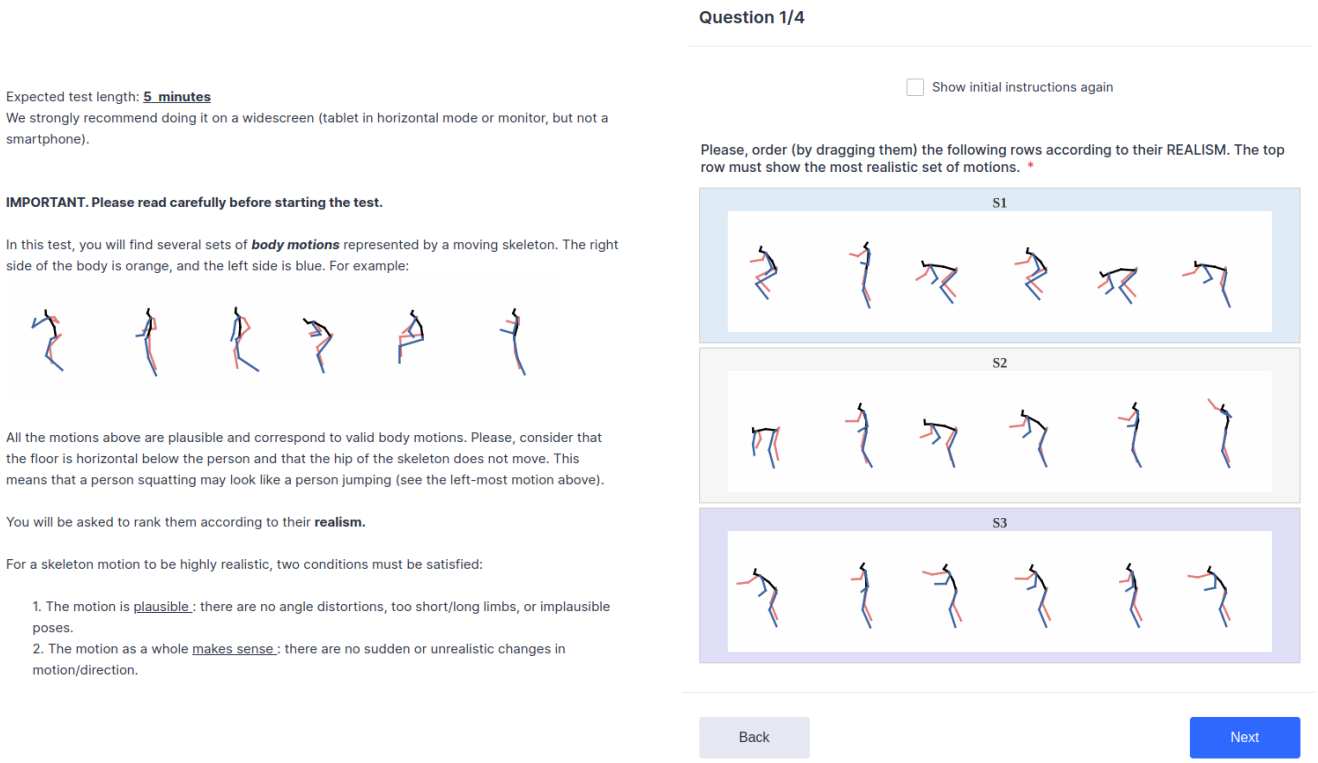

Figure H. **Questionnaire example.** On the left, instructions shown to the participant at the beginning. On the right, the interface for ranking the skeleton motions. All skeletons correspond to *gif* images that repeatedly show the observation and prediction motion sequences.

	Human3.6M[4]				AMASS[6]			
	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 C. **Qualitative assessment.** 126 participants ranked sets of samples from GSPS, DivSamp, and BeLFusion by their realism. Lower average rank (± std. dev.) is better.

[13] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 3

[14] 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. 6

[15] Ye Yuan, Kris Kitani, Y Yuan, and K Kitani. Dlow: Diversifying latent flows for diverse human motion prediction. *European Conference on Computer Vision*, 2020. 3, 6