Seamless Human Motion Composition with Blended Positional Encodings

— 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/FlowMDM/

In this supplementary material, we first describe the implementation details needed to reproduce our work (Sec. A), extending those included in Sec. 4.1. Then, we explain the details of our evaluation protocol (Sec. B), which is specifically built to enable a more fine-grained analysis (Sec. C.1). Sec. 4.2 from the main paper is complemented with additional experiments in Sec. C. In particular, we give more insights on the effects of the attention horizon (Sec. C.2), the diffusion noise schedule (Sec. C.3), and the classifier-free guidance strength (Sec. C.4). Finally, we show and discuss more qualitative examples (Sec. D).

A. Further implementation details

All values are reported as X/Y for Babel/HumanML3D, or as Z if values are equal for both. Note that motion sequences are downsampled to 30/20 fps.

State-of-the-art models. TEACH is used off-the-shelf [1] with the originally proposed alignment and spherical linear interpolation, and without them (TEACH_B). DoubleTake is used off-the-shelf [2] from their original repository, with the parameters *handshake size* and *blending length* set to 10/20f (frames), and 10/5f, respectively. To fulfill the constraints of their method, the handshake size needs to be shorter than half the shortest sequence we want to generate, which is 30f (1s) for Babel. Since DoubleTake uses the original Motion Diffusion Model [8], whose training discarded very short sequences, it underperforms in our more comprehensive evaluation protocol (see Sec. B). For a fairer comparison, we also evaluate it using our diffusion model with absolute positional encodings (APE), and call it DoubleTake*. DoubleTake* uses the same handshake size and blending length as DoubleTake. DiffCollage and MultiDiffusion were implemented manually, and utilize our model as well for the same reasons mentioned earlier. We set their sampling parameter *transition length* to 10/20f. For DoubleTake, DiffCollage, and MultiDiffusion, we use classifier-free guidance with weights 1.5/2.5 during sampling.

FlowMDM. Our diffusion model uses 1k steps and a cosine noise schedule [4]. FlowMDM is trained with the x_0 parameterization [9], and an L2 reconstruction loss. Denoising timesteps are encoded as a sinusoidal positional encoding that goes through two dense layers into a 512D vector. Textual descriptions are tokenized and embedded with CLIP [5] into 512D vectors. Poses of 135/263D are encoded by a dense layer into a sequence of 512D vectors. If the APE is active, a sinusoidal encoding is added to the embedded poses at this stage. Then, the embedded poses are taken as the *keys* and *values* of a Transformer. Embedded poses are concatenated to the sum of the timesteps and text embeddings, and fed to a dense layer. The resulting 512D vectors are the *queries*. If the relative positional encoding (RPE) is active, rotary embeddings [7] are injected to the queries and keys at this stage. The output of the Transformer is added to the embedded poses with a residual connection. 8 Transformers are stacked together. A final dense layer converts the pose embeddings back to a vector of 135/263D, which are the denoised poses. A dropout of 0.1 is applied to the APE, and to the inputs of the Transformers. The attention span of the Transformers is capped within each subsequence during the APE stage, and within the attention horizon H=100/150f during the RPE stage. We train with blended positional encodings (BPE), i.e., RPE and APE are alternated randomly at a frequency of 0.5. We use Adam [3] with learning rate of 0.0001 as our optimizer, and train for 1.3M/500k steps in a single RTX 3090 (about 4/2 days). During BPE sampling, the binary step schedule transitions from absolute to relative mode after 125/60 denoising steps (out of 1k steps). Classifier-free guidance with weights 1.5/2.5 is used during sampling.

[1] https://github.com/athn-nik/teach/commit/f4285aff0fd556a5b46518a751fc90825d91e68b

[2] https://github.com/priorMDM/priorMDM/commit/8bc565b3120c08182f067e161e83403b0efe7cc9

B. Evaluation details

Generative models are difficult to evaluate and compare due to the limitations of the metrics (discussed in Sec. 4.1) and the stochasticity present during sampling. To alleviate the latter, we run all our evaluation 10 times and provide the 95% confidence intervals. However, we still face another issue in our task: the randomness in the combinations of textual descriptions. The generation difficulty for the combination 'sit down'→'stand up'→'run' is not the same as for 'sit down'→'run'→'stand up'. The evaluation protocol from [6] includes 32 evaluation sequences of 32 randomly sampled textual descriptions from the test set. The generated motion needs to perform sequentially the 32 actions from each evaluation sequence. However, these descriptions are sampled differently in each evaluation run, which hinders reproducibility. In order to ensure proper replication and a fair comparison in future works, we propose a more thorough and fully reproducible evaluation protocol that enables a more fine-grained analysis based on *scenarios* (analysis provided in Sec. C.1):

Babel. We built two scenarios with in-distribution (50%) and out-of-distribution (50%) combinations. For the in-distribution scenario, we first selected test motion sequences showcasing at least three consecutive actions (i.e., textual descriptions) with a total duration of 1.5s. Then, we randomly sampled from them to build 32 sets of 32 combinations of textual descriptions. For the out-of-distribution scenario, 32 sets were built by autoregressively sampling 32 textual descriptions so that consecutive actions did not appear together neither in the training nor in the test set.

HumanML3D. Since annotations in HumanML3D do not include consecutive actions, we cannot build in- and out-of-distribution scenarios. However, this dataset contains a great variability of sequence lengths (3-10s). Therefore, we decided to build four scenarios by varying the length of the subsequences included. More specifically, we created three sets of 6, 8, and 18 combinations (9.4, 12.5, 28.1%) by sampling 32 short (3-5s), medium (5-8s), and long (8-10s) test motions, respectively. Ratios were set so that all together preserved the proportion of short, medium, and long subsequences in the original test set. This is important to keep the validity of statistical measures like FID. Additionally, we included another scenario with 32 sets (50%) of 32 random motion sequences from the test set.

We share the list of evaluation combinations for both the human motion composition and extrapolation tasks in our public code repository[3]. Note that a combination consists of a list of textual descriptions and their associated durations. The 32 textual descriptions used for the extrapolation experiments from Sec. 4 are enumerated in Tab. A.

Babel	HumanML3D
walk forward	a person walks in a curved path to the left.
swim movement	a person stands still and does not move.
stretch arms	a person walks straight forward.
walk	a person does jumping jacks.
stand	a person start to dance with legs.
step backwards	person walking in an s shape.
t-pose	a person walks to his right.
throw the ball	a person slowly walked forward.
run	the person is standing still doing body stretches.
circle right arm backwards	the person is dancing the waltz.
wave right	the person is clapping.
ginga dance	walking side to side.
forward kick	a person stayed on the place.
look around	person is jogging in place.
steps to the right	a person walks backward for 3 steps.
side steps	person is running in a circle.
hop forward	the person is waving hi.
dance with arms	a person walks in a circular path.
jog	swinging arms up and down.
walk slowly	a man walks counterclockwise in a circle.
jump jacks series	the person is walking towards the left.
run in half a circle	the person is walking on the treadmill.
walk a few steps ahead	the man is moving his left arm.
move head up and down	the person is doing basketball signals.
rotate right ankle	a person remained sitting down.
play guitar	a person hits his drums.
jump forward	person is doing a dance.
move both hands around chest	a person takes some steps forward.
swing back and forth	a person slowly walks forward five steps.
wave	a person jumps in place.
shake it	this person appears to be painting.
walk in circle	a person wiping a surface with something.

Table A. Extrapolated motions for Babel and HumanML3D.

C. More experimental results

C.1. Fine-grained comparison

Tab. B shows the comparison of FlowMDM with the state of the art in both in-distribution and out-of-distribution scenarios. We observe that, while all methods maintain similar performance in both scenarios for the subsequence generation, they generate less realistic and more abrupt transitions in the out-of-distribution case. FlowMDM performs the best at most metrics in both scenarios, with an important gap with respect to the previous state of the art regarding transition smoothness. Tab. C shows the scenario-wise results for HumanML3D, where FlowMDM also performs the best in most metrics and scenarios. Interestingly, MultiDiffusion is, after ours, the most stable method in terms of transition smoothness across scenarios (PJ and AUJ), whereas DiffCollage and DoubleTake show severe transition degeneration in combinations of long sequences. Such degeneration is mostly due to their methodological need to pad the motion sequence during sampling. When dealing with long

[3] https://barquerogerman.github.io/FlowMDM/

	Subsequence				Transition			
	R-prec ↑	FID ↓	Div →	MM-Dist ↓	FID ↓	Div →	PJ →	AUJ ↓
GT	$0.715^{\pm0.003}$	$0.00^{\pm0.00}$	$8.42^{\pm0.15}$	$3.36^{\pm0.00}$	$0.00^{\pm0.00}$	$6.20^{\pm0.06}$	$0.02^{\pm0.00}$	$0.00^{\pm0.00}$
In-distribution								
TEACH_B	$\mathbf{0.727}^{\pm0.004}$	$2.26^{\pm0.03}$	$8.20^{\pm0.12}$	$\mathbf{3.35}^{\pm0.01}$	$\underline{2.77}^{\pm0.05}$	$6.32^{\pm0.07}$	$1.03^{\pm0.00}$	$2.20^{\pm0.01}$
TEACH	$0.665^{\pm0.003}$	$2.09^{\pm0.03}$	$8.06^{\pm0.09}$	$3.73^{\pm0.02}$	$2.78^{\pm0.06}$	$6.31^{\pm0.07}$	$\underline{0.07}^{\pm0.00}$	$\underline{0.42}^{\pm0.01}$
DoubleTake*	$0.620^{\pm0.006}$	$3.04^{\pm0.06}$	$7.49^{\pm0.07}$	$4.19^{\pm0.02}$	$3.04^{\pm0.12}$	$\underline{6.21}^{\pm0.06}$	$0.28^{\pm0.00}$	$1.01^{\pm0.01}$
DoubleTake	$0.682^{\pm0.008}$	$\underline{1.52}^{\pm0.03}$	$7.90^{\pm0.07}$	$3.67^{\pm0.04}$	$3.47^{\pm0.08}$	$6.16^{\pm0.07}$	$0.17^{\pm0.00}$	$0.62^{\pm0.01}$
MultiDiffusion	$0.724^{\pm0.008}$	$2.00^{\pm0.05}$	$\underline{8.36}^{\pm0.10}$	$\underline{3.38}^{\pm0.02}$	$6.33^{\pm0.13}$	$5.91^{\pm0.06}$	$0.17^{\pm0.00}$	$0.65^{\pm0.01}$
DiffCollage	$0.690^{\pm0.006}$	$1.92^{\pm0.07}$	$7.92^{\pm0.09}$	$3.67^{\pm0.02}$	$4.25^{\pm0.15}$	$\mathbf{6.19}^{\pm0.07}$	$0.19^{\pm0.01}$	$0.82^{\pm0.02}$
FlowMDM (Ours)	$\underline{0.726}^{\pm0.006}$	$\mathbf{1.36}^{\pm0.05}$	$\mathbf{8.47}^{\pm0.10}$	$3.40^{\pm0.03}$	$\mathbf{2.26}^{\pm0.08}$	$6.60^{\pm0.08}$	$\mathbf{0.05}^{\pm0.00}$	$\mathbf{0.11}^{\pm0.00}$
Out-of-distribution								
TEACH_B	$\underline{0.680}^{\pm0.006}$	$1.75^{\pm0.04}$	$8.15^{\pm0.11}$	$3.51^{\pm0.01}$	$3.53^{\pm0.06}$	$\underline{6.04}^{\pm0.10}$	$1.14^{\pm0.01}$	$2.49^{\pm0.01}$
TEACH	$0.644^{\pm0.004}$	$2.06^{\pm0.03}$	$7.94^{\pm0.12}$	$3.70^{\pm0.01}$	$4.08^{\pm0.08}$	$6.00^{\pm0.09}$	$\mathbf{0.07}^{\pm0.00}$	$\underline{0.46}^{\pm0.00}$
DoubleTake*	$0.572^{\pm0.007}$	$3.78^{\pm0.07}$	$7.53^{\pm0.12}$	$4.15^{\pm0.02}$	$3.83^{\pm0.09}$	$\mathbf{6.12}^{\pm0.07}$	$0.28^{\pm0.00}$	$1.07^{\pm0.02}$
DoubleTake	$0.654^{\pm0.009}$	$1.65^{\pm0.07}$	$8.06^{\pm0.08}$	$3.66^{\pm0.02}$	$\mathbf{2.98}^{\pm0.06}$	$6.03^{\pm0.07}$	$0.17^{\pm0.00}$	$0.66^{\pm0.01}$
MultiDiffusion	$\mathbf{0.681}^{\pm0.009}$	$2.11^{\pm0.06}$	$\mathbf{8.35}^{\pm0.08}$	$\mathbf{3.47}^{\pm0.03}$	$6.97^{\pm0.12}$	$5.67^{\pm0.05}$	$0.19^{\pm0.00}$	$0.71^{\pm0.01}$
DiffCollage	$0.652^{\pm0.004}$	$\underline{1.60}^{\pm0.07}$	$7.91^{\pm0.09}$	$3.74^{\pm0.01}$	$4.65^{\pm0.19}$	$6.00^{\pm0.09}$	$0.20^{\pm0.00}$	$0.86^{\pm0.01}$
FlowMDM (Ours)	$0.679^{\pm0.004}$	$\mathbf{1.26}^{\pm0.06}$	$\underline{8.16}^{\pm0.08}$	$\underline{3.50}^{\pm0.03}$	$\underline{3.17}^{\pm0.12}$	$6.44^{\pm0.09}$	$\mathbf{0.07}^{\pm0.00}$	$\mathbf{0.17}^{\pm0.00}$

Table B. Scenario-wise comparison in Babel. Symbols ↑, ↓, and → indicate that higher, lower, or values closer to the ground truth (GT) are better, respectively. Evaluation is run 10 times and ± specifies the 95% confidence intervals.

	Subsequence				Transition			
	R-prec ↑	FID ↓	Div →	MM-Dist ↓	FID ↓	Div →	PJ →	AUJ ↓
GT	$0.796^{\pm0.004}$	$0.00^{\pm0.00}$	$9.34^{\pm0.08}$	$2.97^{\pm0.01}$	$0.00^{\pm0.00}$	$9.54^{\pm0.15}$	$0.04^{\pm0.00}$	$0.07^{\pm0.00}$
Short								
DoubleTake*	$0.649^{\pm0.012}$	$\mathbf{3.03}^{\pm0.18}$	$\mathbf{9.52}^{\pm0.11}$	$3.72^{\pm0.05}$	$\underline{3.56}^{\pm0.14}$	$\underline{8.92}^{\pm0.14}$	$0.13^{\pm0.01}$	$0.79^{\pm0.05}$
DoubleTake	$0.704^{\pm0.022}$	$4.85^{\pm0.20}$	$10.01^{\pm0.15}$	$\underline{3.25}^{\pm0.09}$	$4.40^{\pm0.24}$	$8.88^{\pm0.17}$	$\underline{0.09}^{\pm0.00}$	$\underline{0.73}^{\pm0.02}$
MultiDiffusion	$\mathbf{0.717}^{\pm0.011}$	$5.49^{\pm0.15}$	$10.14^{\pm0.17}$	$\mathbf{3.23}^{\pm0.07}$	$4.66^{\pm0.27}$	$8.68^{\pm0.08}$	$0.10^{\pm0.00}$	$0.92^{\pm0.02}$
DiffCollage	$0.705^{\pm0.012}$	$\underline{4.69}^{\pm0.18}$	$\underline{9.73}^{\pm0.14}$	$3.30^{\pm0.04}$	$4.81^{\pm0.32}$	$8.49^{\pm0.12}$	$0.15^{\pm0.01}$	$1.13^{\pm0.10}$
FlowMDM (Ours)	$\underline{0.714}^{\pm0.015}$	$4.75^{\pm0.26}$	$9.90^{\pm0.20}$	$3.31^{\pm0.06}$	$\mathbf{3.17}^{\pm0.17}$	$\mathbf{9.03}^{\pm0.14}$	$\mathbf{0.04}^{\pm0.00}$	$\mathbf{0.59}^{\pm0.04}$
Medium								
DoubleTake*	$0.644^{\pm0.009}$	$\underline{2.18}^{\pm0.08}$	$\underline{9.18}^{\pm0.12}$	$3.72^{\pm0.04}$	$\mathbf{3.34}^{\pm0.30}$	$\mathbf{8.73}^{\pm0.12}$	$0.14^{\pm0.00}$	$\underline{0.70}^{\pm0.03}$
DoubleTake	$0.642^{\pm0.014}$	$2.34^{\pm0.05}$	$9.59^{\pm0.09}$	$3.79^{\pm0.05}$	$5.42^{\pm0.30}$	$\underline{8.61}^{\pm0.11}$	$0.12^{\pm0.00}$	$0.83^{\pm0.02}$
MultiDiffusion	$\mathbf{0.673}^{\pm0.007}$	$3.22^{\pm0.10}$	$9.91^{\pm0.07}$	$\mathbf{3.54}^{\pm0.04}$	$6.24^{\pm0.34}$	$8.11^{\pm0.12}$	$\underline{0.10}^{\pm0.00}$	$1.14^{\pm0.01}$
DiffCollage	$0.661^{\pm0.010}$	$\mathbf{2.03}^{\pm0.07}$	$\mathbf{9.38}^{\pm0.10}$	$3.60^{\pm0.04}$	$4.95^{\pm0.27}$	$8.13^{\pm0.09}$	$0.14^{\pm0.00}$	$\mathbf{0.66}^{\pm0.05}$
FlowMDM (Ours)	$\underline{0.669}^{\pm0.012}$	$3.18^{\pm0.15}$	$9.68^{\pm0.08}$	$\underline{3.55}^{\pm0.04}$	$\underline{4.18}^{\pm0.43}$	$8.52^{\pm0.07}$	$\mathbf{0.04}^{\pm0.00}$	$0.86^{\pm0.03}$
Long								
DoubleTake*	$\underline{0.616}^{\pm0.006}$	$\underline{2.51}^{\pm0.09}$	$\underline{8.77}^{\pm0.08}$	$\underline{4.09}^{\pm0.03}$	$\underline{3.38}^{\pm0.18}$	$\underline{8.50}^{\pm0.11}$	$0.89^{\pm0.02}$	$3.52^{\pm0.07}$
DoubleTake	$0.605^{\pm0.006}$	$4.07^{\pm0.13}$	$8.19^{\pm0.11}$	$4.18^{\pm0.01}$	$8.45^{\pm0.33}$	$7.79^{\pm0.12}$	$0.81^{\pm0.02}$	$3.04^{\pm0.07}$
MultiDiffusion	$0.569^{\pm0.012}$	$5.02^{\pm0.15}$	$8.07^{\pm0.07}$	$4.49^{\pm0.05}$	$8.56^{\pm0.32}$	$7.91^{\pm0.10}$	$\underline{0.23}^{\pm0.01}$	$\underline{1.16}^{\pm0.01}$
DiffCollage	$0.557^{\pm0.008}$	$5.79^{\pm0.13}$	$7.75^{\pm0.09}$	$4.61^{\pm0.02}$	$9.00^{\pm0.36}$	$7.75^{\pm0.09}$	$0.38^{\pm0.01}$	$5.04^{\pm0.14}$
FlowMDM (Ours)	$\mathbf{0.666}^{\pm0.012}$	$\mathbf{1.93}^{\pm0.08}$	$\mathbf{8.81}^{\pm0.09}$	$\mathbf{3.81}^{\pm0.04}$	$\mathbf{2.85}^{\pm0.22}$	$\mathbf{8.54}^{\pm0.11}$	$\mathbf{0.08}^{\pm0.00}$	$\mathbf{0.45}^{\pm0.03}$
All								
DoubleTake*	$\underline{0.655}^{\pm0.007}$	$\underline{0.84}^{\pm0.04}$	$\mathbf{9.29}^{\pm0.10}$	$\underline{3.92}^{\pm0.03}$	$\underline{1.91}^{\pm0.12}$	$\mathbf{8.79}^{\pm0.11}$	$0.51^{\pm0.01}$	$2.11^{\pm0.05}$
DoubleTake	$0.621^{\pm0.006}$	$1.49^{\pm0.07}$	$8.91^{\pm0.04}$	$4.13^{\pm0.02}$	$4.75^{\pm0.13}$	$8.39^{\pm0.06}$	$0.47^{\pm0.01}$	$1.84^{\pm0.03}$
MultiDiffusion	$0.632^{\pm0.003}$	$1.17^{\pm0.04}$	$\mathbf{9.29}^{\pm0.09}$	$4.05^{\pm0.02}$	$4.42^{\pm0.16}$	$8.37^{\pm0.08}$	$\underline{0.17}^{\pm0.00}$	$\underline{1.06}^{\pm0.01}$
DiffCollage	$0.615^{\pm0.007}$	$1.73^{\pm0.07}$	$8.73^{\pm0.05}$	$4.18^{\pm0.04}$	$4.98^{\pm0.24}$	$8.09^{\pm0.06}$	$0.26^{\pm0.00}$	$2.71^{\pm0.12}$
FlowMDM	$\mathbf{0.695}^{\pm0.008}$	$\mathbf{0.30}^{\pm0.02}$	$9.55^{\pm0.08}$	$\mathbf{3.58}^{\pm0.02}$	$\mathbf{1.49}^{\pm0.06}$	$\underline{8.78}^{\pm0.11}$	$\mathbf{0.06}^{\pm0.00}$	$\mathbf{0.50}^{\pm0.01}$

Table C. Scenario-wise comparison in HumanML3D.

sequences, sequences might be extended beyond the maximum sequence length at training time. Therefore, given that the APE does not extrapolate well, the generation in the padded motion, or transition, tends to degenerate. Our method naturally avoids this limitation.

		Subsequence				Transition			
H (frames)	Inf. PE	R-prec ↑	FID ↓	Div →	MM-Dist ↓	FID ↓	Div →	PJ →	AUJ ↓
GT	-	$0.715^{\pm 0.003}$	$0.00^{\pm 0.00}$	$8.42^{\pm 0.15}$	$3.36^{\pm 0.00}$	$0.00^{\pm 0.00}$	$6.20^{\pm 0.06}$	$0.02^{\pm 0.00}$	$0.00^{\pm 0.00}$
50	R	$0.641^{\pm 0.004}$	$1.03^{\pm 0.04}$	$7.99^{\pm 0.11}$	$3.92^{\pm 0.03}$	$\mathbf{2.04}^{\pm 0.06}$	$6.30^{\pm 0.05}$	$\mathbf{0.04}^{\pm 0.00}$	$0.15^{\pm 0.00}$
100	R	$0.635^{\pm 0.004}$	$\mathbf{0.85}^{\pm 0.02}$	$8.25^{\pm 0.12}$	$3.98^{\pm 0.02}$	$\underline{2.14}^{\pm 0.04}$	$6.44^{\pm 0.09}$	$\mathbf{0.04}^{\pm 0.00}$	$0.15^{\pm 0.00}$
150	R	$0.641^{\pm 0.005}$	$\underline{0.99}^{\pm 0.04}$	$8.24^{\pm 0.15}$	$3.88^{\pm 0.03}$	$2.43^{\pm 0.06}$	$6.43^{\pm 0.06}$	$\mathbf{0.04}^{\pm 0.00}$	$0.15^{\pm 0.00}$
200	R	$0.601^{\pm 0.005}$	$1.48^{\pm 0.04}$	$7.85^{\pm 0.14}$	$4.17^{\pm 0.02}$	$3.18^{\pm 0.09}$	$\mathbf{6.16}^{\pm 0.05}$	$\mathbf{0.04}^{\pm 0.00}$	$0.19^{\pm 0.00}$
50	B	$0.698^{\pm 0.006}$	$1.07^{\pm 0.03}$	$8.19^{\pm 0.11}$	$3.44^{\pm 0.02}$	$2.34^{\pm 0.06}$	$\underline{6.24}^{\pm 0.07}$	$0.06^{\pm 0.00}$	$\mathbf{0.13}^{\pm 0.00}$
100	B	$\underline{0.702}^{\pm 0.004}$	$\underline{0.99}^{\pm 0.04}$	$\mathbf{8.36}^{\pm 0.13}$	$3.45^{\pm 0.02}$	$2.61^{\pm 0.06}$	$6.47^{\pm 0.05}$	$0.06^{\pm 0.00}$	$\mathbf{0.13}^{\pm 0.00}$
150	B	$\mathbf{0.704}^{\pm 0.004}$	$1.24^{\pm 0.03}$	$\underline{8.34}^{\pm 0.12}$	$\underline{3.43}^{\pm 0.02}$	$2.54^{\pm 0.08}$	$6.40^{\pm 0.08}$	$0.06^{\pm 0.00}$	$\mathbf{0.13}^{\pm 0.00}$
200	B	$0.694^{\pm 0.006}$	$1.13^{\pm 0.02}$	$8.25^{\pm 0.13}$	$\mathbf{3.42}^{\pm 0.02}$	$3.31^{\pm 0.08}$	$6.38^{\pm 0.09}$	$0.06^{\pm 0.00}$	$0.14^{\pm 0.01}$

Table D. Attention horizon effect in Babel. All models correspond to FlowMDM, trained with BPE. Inf. PE indicates the type of positional encoding used during sampling: B for BPE, and R for only RPE. Symbols ↑, ↓, and → indicate that higher, lower, or values closer to the ground truth (GT) are better, respectively. Evaluation is run 10 times and ± specifies the 95% confidence intervals.

		Subsequence				Transition			
H (frames)	Inf. PE	R-prec ↑	FID ↓	Div →	MM-Dist ↓	FID ↓	Div →	PJ →	AUJ ↓
GT	-	$0.796^{\pm 0.004}$	$0.00^{\pm 0.00}$	$9.34^{\pm 0.08}$	$2.97^{\pm 0.01}$	$0.00^{\pm 0.00}$	$9.54^{\pm 0.15}$	$0.04^{\pm 0.00}$	$0.07^{\pm 0.00}$
50	R	$0.583^{\pm 0.005}$	$1.08^{\pm 0.07}$	$9.03^{\pm 0.15}$	$4.30^{\pm 0.02}$	$1.88^{\pm 0.06}$	$8.85^{\pm 0.10}$	$\underline{0.04}^{\pm 0.00}$	$0.70^{\pm 0.01}$
100	R	$0.591^{\pm 0.005}$	$1.07^{\pm 0.03}$	$9.02^{\pm 0.13}$	$4.29^{\pm 0.02}$	$1.51^{\pm 0.08}$	$8.90^{\pm 0.08}$	$\underline{0.04}^{\pm 0.00}$	$0.56^{\pm 0.01}$
150	R	$0.554^{\pm 0.007}$	$1.06^{\pm 0.06}$	$9.02^{\pm 0.11}$	$4.54^{\pm 0.02}$	$\mathbf{1.12}^{\pm 0.04}$	$\mathbf{9.00}^{\pm 0.10}$	$0.05^{\pm 0.00}$	$0.53^{\pm 0.01}$
200	R	$0.528^{\pm 0.004}$	$1.37^{\pm 0.04}$	$8.87^{\pm 0.07}$	$4.68^{\pm 0.01}$	$1.72^{\pm 0.05}$	$\underline{8.97}^{\pm 0.09}$	$\mathbf{0.03}^{\pm 0.00}$	$0.97^{\pm 0.01}$
50	B	$0.671^{\pm 0.004}$	$\mathbf{0.25}^{\pm 0.01}$	$\mathbf{9.37}^{\pm 0.14}$	$3.66^{\pm 0.02}$	$\underline{1.27}^{\pm 0.04}$	$8.79^{\pm 0.08}$	$0.06^{\pm 0.00}$	$\underline{0.52}^{\pm 0.01}$
100	B	$\underline{0.684}^{\pm 0.003}$	$0.36^{\pm 0.02}$	$9.55^{\pm 0.09}$	$\mathbf{3.61}^{\pm 0.02}$	$2.04^{\pm 0.11}$	$8.59^{\pm 0.06}$	$0.06^{\pm 0.00}$	$0.56^{\pm 0.01}$
150	B	$\mathbf{0.685}^{\pm 0.004}$	$\underline{0.29}^{\pm 0.01}$	$9.58^{\pm 0.12}$	$\mathbf{3.61}^{\pm 0.01}$	$1.38^{\pm 0.05}$	$8.79^{\pm 0.09}$	$0.06^{\pm 0.00}$	$\mathbf{0.51}^{\pm 0.01}$
200	B	$0.658^{\pm 0.006}$	$0.47^{\pm 0.03}$	$\mathbf{9.37}^{\pm 0.13}$	$3.77^{\pm 0.02}$	$2.27^{\pm 0.07}$	$8.69^{\pm 0.08}$	$0.06^{\pm 0.00}$	$0.68^{\pm 0.01}$

Table E. Attention horizon effect in HumanML3D. All models correspond to FlowMDM, trained with BPE. Inf. PE indicates the type of positional encoding used during sampling: B for BPE, and R for only RPE.

C.2. On the attention horizon

In Tabs. D and E, we show the effect of the attention horizon when using RPE for either a purely relative inference schedule, or our proposed BPE inference schedule. We observe how increasing it too much (H=200) makes the network perform worse at transition generation in both datasets (FID and AUJ), and also in subsequence generation for HumanML3D (R-prec and MM-Dist). Conversely, when decreasing it too much (H=50), the capacity to model long-range dynamics becomes limited, thus reducing the accuracy of the generated subsequences (R-prec and MM-Dist). As the performance with H of 100 and 150 is similar in both datasets, we chose values that are closest to the average sequence length in each dataset, i.e., 100/150f for Babel/HumanML3D.

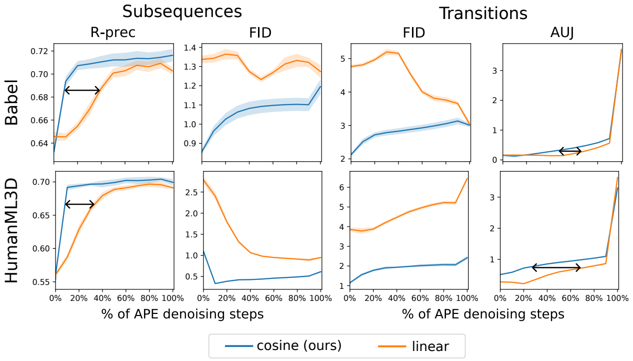

Figure A. **Diffusion noise schedules.** The cosine noise schedule destroys the motion signal slower and in a more evenly distributed way than the linear schedule. As a result, FlowMDM is able to exploit better the low-to-high frequencies decomposition along the denoising chain and generate better subsequences and transitions. The faster motion destruction in the linear schedule translates to needing more APE steps to reconstruct global dependencies inside subsequences (black arrows ↔).

C.3. On the diffusion schedule

The discussion and the BPE design in Sec. 3.2 are motivated by the low-to-high frequencies decomposition during the denoising stage of diffusion models. However, the denoising process depends on how the noise is injected, or the *noise schedule*. The linear and the cosine (our choice) noise schedules are the most common schedules. The linear schedule destroys the motion very fast, reaching a non-recognizable state after going through the 75% of the diffusion steps [4]. Instead, the cosine schedule destroys the motion signal slower and in a more evenly distributed way. Fig. A shows the performance of FlowMDM during

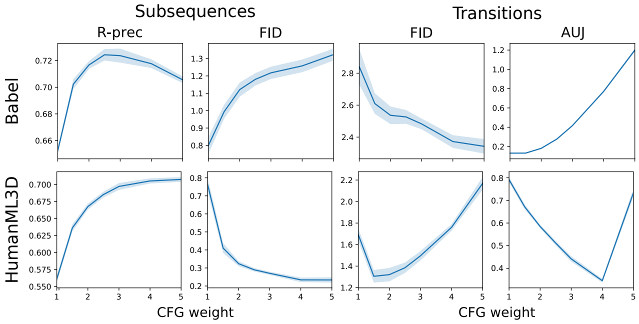

Figure B. **Classifier-free guidance.** In line with prior works, we also observe an accuracy improvement (R-prec) when increasing the strength (i.e., *weight*) of the classifier-free guidance (CFG). However, above certain values, the performance degrades, especially in terms of smoothness (AUJ). This is caused by the misalignment of CFG directions on each side of the transition.

BPE sampling with both schedules. First, we observe that FlowMDM benefits from the steadier noise injection of the cosine schedule, achieving better performance in all realism and accuracy metrics (R-prec and FID). Second, we identify a displacement in the accuracy (R-prec) and smoothness (AUJ) curves (see black arrows). Given that with the linear schedule global dependencies start being recovered later, more APE steps are needed to achieve the accuracy and smoothness reached with the cosine schedule.

C.4. On the classifier-free guidance

The classifier-free guidance is an important add-on for diffusion sampling that intensifies the conditioning signal, thus improving the quality and accuracy of the generated samples [2]. It is implemented by first computing the conditionally denoised motion x_c, and the unconditionally denoised motion x. Then, the denoised sample is computed as $x + w(x_c - x)$. If w=1, the classifier-free guidance is deactivated. When generating motion from single textual descriptions with classifier-free guidance, we keep steering the denoising toward motions matching better the textual description. However, when building human motion compositions with our method, two different conditions coexist in the neighborhoods of the transitions. There, the classifier-free guidance pushes the denoising towards dispar directions. As a result, if w is too high, the transition will become sharper, and if w is too low, subsequences might not be accurate enough. Fig. B shows these effects for FlowMDM. We notice a sweet point around w=$1.5/2.5$ for Babel/HumanML3D, where FlowMDM reaches the maximum accuracy and quality for subsequences and a good trade-off for quality and smoothness of transitions.

D. Qualitative results

Figs. C and D show six human motion compositions (A to F), and two extrapolations (G and H) for Babel and HumanML3D, respectively. The compositions are subsets of the evaluation combinations composed of 32 actions, so the beginning and end of these can contain partial transitions toward other actions. Note that we can represent the motions from Babel with SMPL body meshes thanks to its motion representation including the SMPL parameters [1]. For HumanML3D, we use skeletons, as its motion representation only includes the 3D coordinates of the joints.

Discussion. The hands trajectories and the jerk color indicators in Figs. C and D highlight that FlowMDM generates the smoothest transitions between subsequences. Notably, state-of-the-art methods exhibit frequent smoothness artifacts (black segments) in the boundaries of their transitions. We notice that the compositions produced by TEACH lack realism due to the use of a naive spherical linear interpolation, disrupting the motion dynamics. This becomes more apparent in extrapolations G and H of both datasets, where the periodicity of the movement is clearly compromised. On the other side, DoubleTake, DiffCollage, and MultiDiffusion share two significant limitations. Firstly, they adhere to a predetermined transition length, which may not fit all situations. For example, in Babel-A, the ‘picking’ actions occur very rapidly due to the insufficient length for generating a natural transition. By contrast, our approach is able to leverage more transitioning time from either transition side if needed, without artificial constraints. Secondly, the denoising process in these methods only considers a small portion of the neighboring subsequences, leading to poor performance in dynamic motion extrapolations. For example, in HumanML3D-G, they all generate erratic jumping jacks. While our method also independently generates the low-frequency motion spectrum, it effectively rectifies inconsistencies in later stages, yielding realistic and periodic motion. In the case of Babel-H, where successfully extrapolating the ‘hop forward’ action requires synchronizing each subsequence with the whole neighboring motion, our model is the only one able to generate a smooth, coherent, and realistic extrapolation.

Limitations. However, FlowMDM is not without its imperfections. We noticed that our method struggles with very complex descriptions, such as the first one in HumanML3D-B. Instead of executing the intricate description that includes ‘walk backwards, sit, stand, and walk forward again’, it only walks backwards. Given that the partial execution of actions is also observed in other methods, we consider it a challenge associated with the broader text-to-motion task. Indeed, our model could theoretically also benefit from improved conditioning schemes such as using better text embeddings. Another acknowledged limitation of our model, discussed in Sec. 5, is the independent gen-

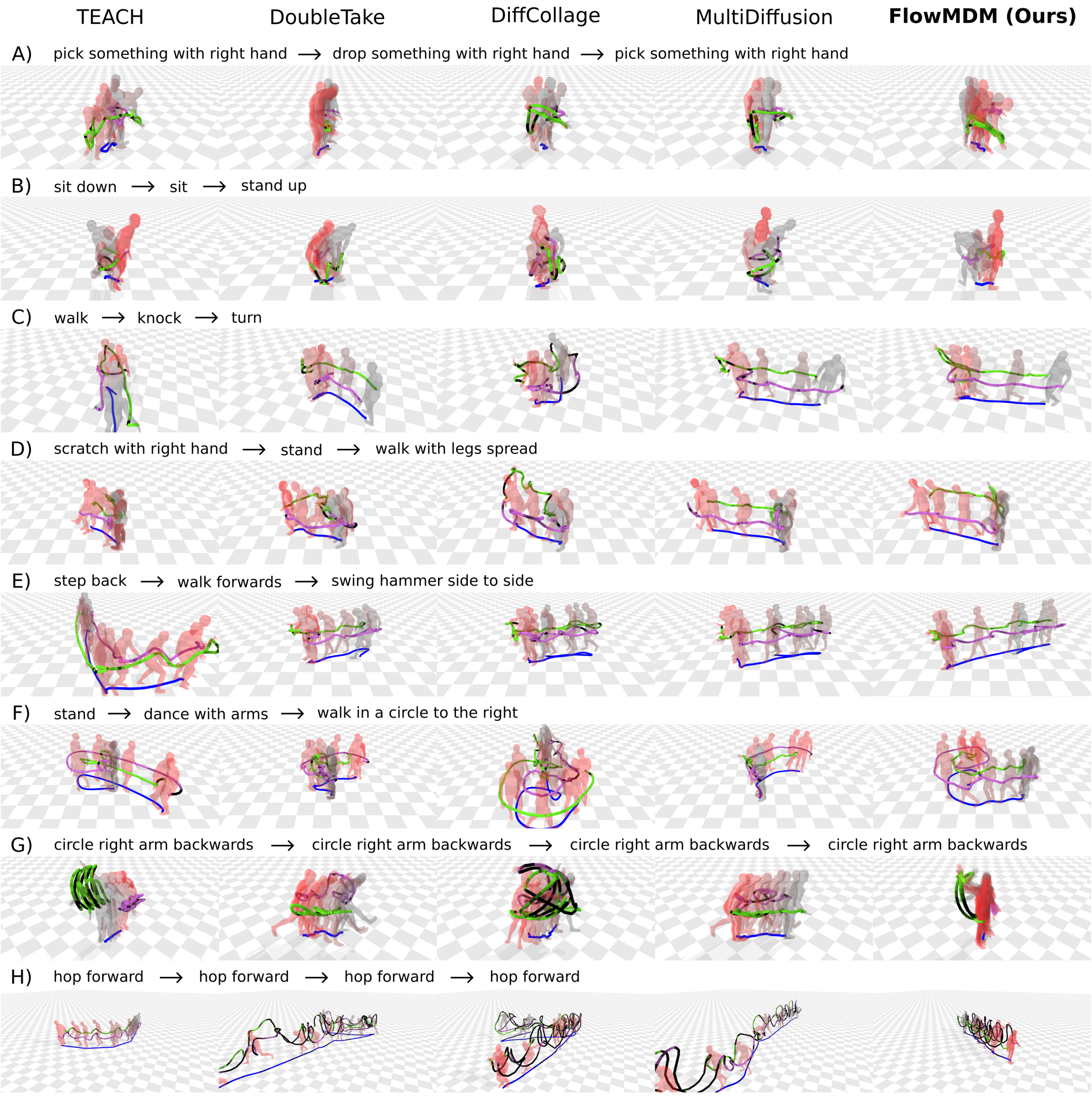
TEACH
DoubleTake
DiffCollage
MultiDiffusion
FlowMDM (Ours)
A) pick something with right hand ⟶ drop something with right hand ⟶ pick something with right hand
B) sit down ⟶ sit ⟶ stand up
C) walk ⟶ knock ⟶ turn
D) scratch with right hand ⟶ stand ⟶ walk with legs spread
E) step back ⟶ walk forwards ⟶ swing hammer side to side
F) stand ⟶ dance with arms ⟶ walk in a circle to the right
G) circle right arm backwards ⟶ circle right arm backwards ⟶ circle right arm backwards ⟶ circle right arm backwards
H) hop forward ⟶ hop forward ⟶ hop forward ⟶ hop forward

Figure C. **Qualitative examples (Babel).** A-F feature six human motion compositions, and G-H two human motion extrapolations. According to the scenarios defined in Sec. B, A, B, C belong to in-distribution combinations, and D, E, F to out-of-distribution combinations. Solid curves match the trajectories of the global position (blue) and left/right hands (purple/green). Darker colors indicate instantaneous jerk deviations from the median value, saturating at twice the jerk's standard deviation in the dataset (black segments). Abrupt transitions manifest as black segments amidst lighter ones.

eration of low-frequency components. In Babel-B, for example, a slight mismatch between the sitting and standing positions is observed. Nonetheless, in contrast to DiffCollage, MultiDiffusion, and DoubleTake which also exhibit this effect, FlowMDM produces a smoother result.

References

[1] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *Computer Vision–ECCV 2016: 14th Euro-*

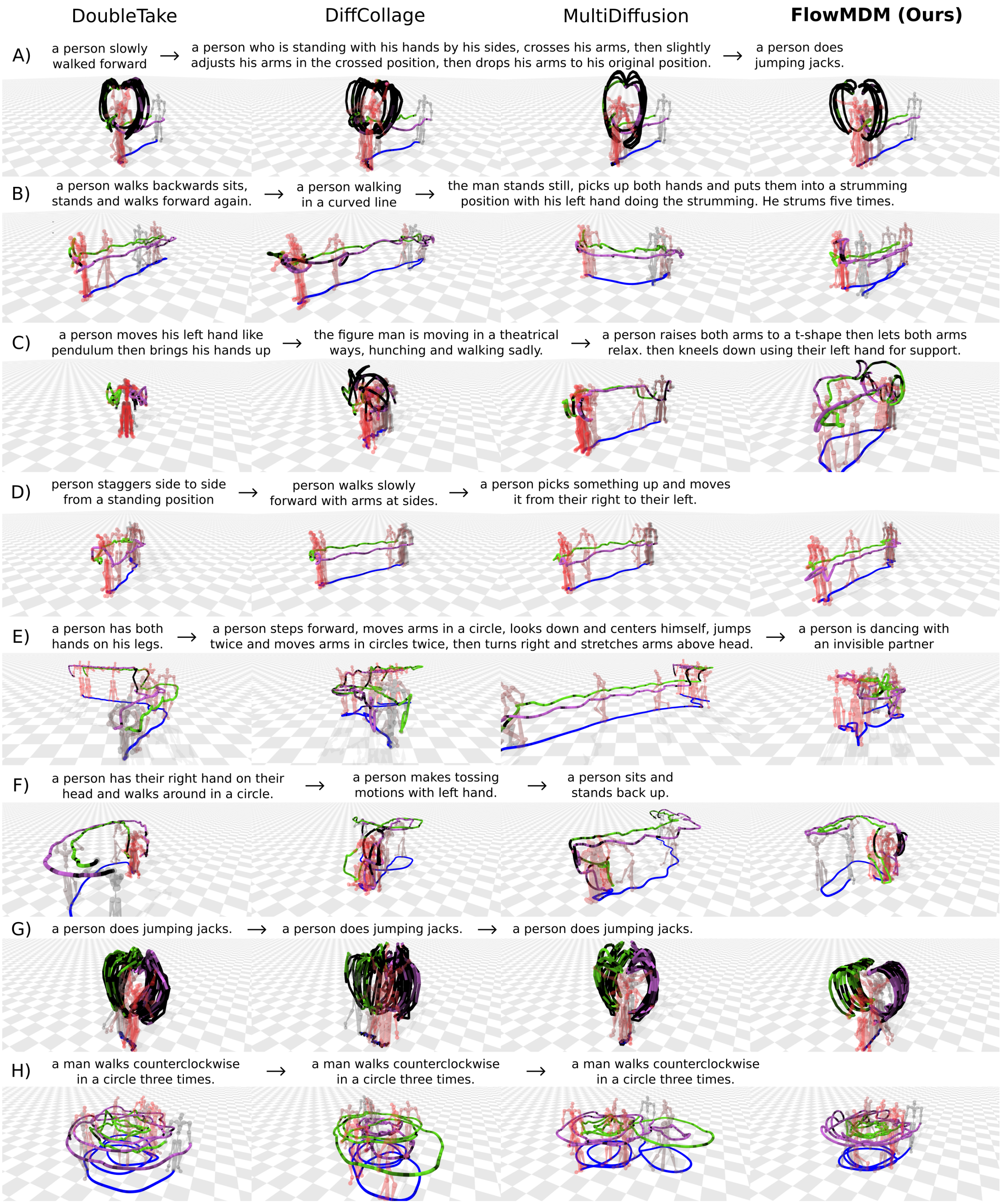
DoubleTake
DiffCollage
MultiDiffusion
FlowMDM (Ours)
A)
a person slowly walked forward →
a person who is standing with his hands by his sides, crosses his arms, then slightly adjusts his arms in the crossed position, then drops his arms to his original position. →
a person does jumping jacks.
B)
a person walks backwards sits, stands and walks forward again. →
a person walking in a curved line →
the man stands still, picks up both hands and puts them into a strumming position with his left hand doing the strumming. He strums five times.
C)
a person moves his left hand like pendulum then brings his hands up →
the figure man is moving in a theatrical ways, hunching and walking sadly. →
a person raises both arms to a t-shape then lets both arms relax. then kneels down using their left hand for support.
D)
person staggers side to side from a standing position →
person walks slowly forward with arms at sides. →
a person picks something up and moves it from their right to their left.
E)
a person has both hands on his legs. →
a person steps forward, moves arms in a circle, looks down and centers himself, jumps twice and moves arms in circles twice, then turns right and stretches arms above head. →
a person is dancing with an invisible partner
F)
a person has their right hand on their head and walks around in a circle. →
a person makes tossing motions with left hand. →
a person sits and stands back up.
G)
a person does jumping jacks. →
a person does jumping jacks. →
a person does jumping jacks.
H)
a man walks counterclockwise in a circle three times. →
a man walks counterclockwise in a circle three times. →
a man walks counterclockwise in a circle three times.

Figure D. **Qualitative examples (HumanML3D).** A-F feature six human motion compositions, and G-H two human motion extrapolations. According to the scenarios defined in Sec. B, A, B, C are samples from the short, medium, and long scenarios, respectively, and D, E, F from the mixed scenario.

pean Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14, pages 561–578. Springer, 2016. 5

[2] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 5

[3] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 1

[4] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021. 1, 4

[5] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1

[6] Yonatan Shafir, Guy Tevet, Roy Kapon, and Amit H Bermano. Human motion diffusion as a generative prior. *arXiv preprint arXiv:2303.01418*, 2023. 2

[7] Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *arXiv preprint arXiv:2104.09864*, 2021. 1

[8] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *The Eleventh International Conference on Learning Representations*, 2022. 1

[9] Zhisheng Xiao, Karsten Kreis, and Arash Vahdat. Tackling the generative learning trilemma with denoising diffusion GANs. In *International Conference on Learning Representations (ICLR)*, 2022. 1