

# Seamless Human Motion Composition with Blended Positional Encodings

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

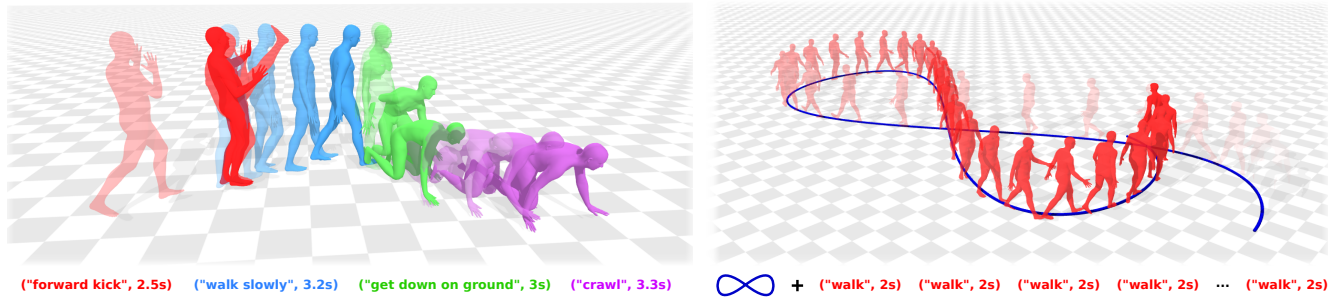


Figure 1. We present FlowMDM, a diffusion-based approach capable of generating seamlessly continuous sequences of human motion from textual descriptions (left). The whole sequence is generated simultaneously and it does not require any postprocessing. FlowMDM also makes strides in the challenging problem of extrapolating and controlling periodic motion such as walking, jumping, or waving (right).

## Abstract

Conditional human motion generation is an important topic with many applications in virtual reality, gaming, and robotics. While prior works have focused on generating motion guided by text, music, or scenes, these typically result in isolated motions confined to short durations. Instead, we address the generation of long, continuous sequences guided by a series of varying textual descriptions. In this context, we introduce FlowMDM, the first diffusion-based model that generates seamless Human Motion Compositions (HMC) without any postprocessing or redundant denoising steps. For this, we introduce the Blended Positional Encodings, a technique that leverages both absolute and relative positional encodings in the denoising chain. More specifically, global motion coherence is recovered at the absolute stage, whereas smooth and realistic transitions are built at the relative stage. As a result, we achieve state-of-the-art results in terms of accuracy, realism, and smoothness on the Babel and HumanML3D datasets. FlowMDM excels when trained with only a single description per motion sequence thanks to its Pose-Centric Cross-Attention, which makes it robust against varying text descriptions at inference time. Finally, to address the limitations of existing HMC metrics, we propose two new metrics: the Peak Jerk and the Area Under the Jerk, to detect abrupt transitions.

## 1. Introduction

In the field of computer vision, recent progress has been made in developing photorealistic avatars [54] for applications like virtual reality, gaming, and robotics [62, 79]. Aside from looking visually realistic, these avatars must also move in a convincing manner. This is challenging due to the intricate nature of human motion, strongly influenced by various factors such as the environment, interactions, and physical contact [14]. Furthermore, complexity increases when attempting to control these motions. Recent advances include the generation of motion sequences from control signals like textual descriptions or actions [109]; however, such methods only produce isolated, standalone motion. Therefore, these approaches fail to handle scenarios where a long motion is driven by distinct control signals on different time slices. Such capability is needed to provide full control over the sequence of desired actions and their duration. In these scenarios, the generated motion needs to feature seamless and realistic transitions between actions. In this work, we tackle this problem, which we refer to as generative Human Motion Composition (HMC). In particular, we focus on generating single-human motion from text, illustrated in Fig. 1.

One of the primary obstacles in HMC is the lack of datasets that offer long motion sequences with diverse textual annotations. Existing datasets typically feature se-

quences of limited duration, often lasting only up to 10 seconds, and with just a single control signal governing the entire sequence [26, 64]. This limitation calls for innovative solutions to address the inherent complexities of the task. Prior works have tackled this problem mostly with autoregressive approaches [4, 45, 48, 66, 104]. These methods iteratively create compositions by using the current motion as a basis to generate subsequent motions. However, they require datasets with multiple consecutive annotated motions, and tend to degenerate in very long HMC scenarios due to error accumulation [107]. Other recent works have leveraged the infilling capabilities of motion diffusion models to generate motion compositions [73, 103]. However, for these, a substantial portion of each motion sequence is generated independently from adjacent motions, and generating transitions requires computing redundant denoising steps. In this work, we propose a novel architecture designed to address these specific challenges. Our main contributions are:

- We propose FlowMDM, the first diffusion-based model that generates seamless human motion compositions without any postprocessing or extra denoising steps. To accomplish it, we introduce Blended Positional Encodings (BPE), a new technique for diffusion Transformers that combines the benefits of both absolute and relative positional encodings during sampling. In particular, the denoising first exploits absolute information to recover the global motion coherence, and then leverages relative positions to build smooth and realistic transitions between actions. As a result, FlowMDM achieves state-of-the-art results in terms of accuracy, realism, and smoothness in the HumanML3D [26] and Babel [65] datasets.
- We introduce a new attention technique tailored for HMC: the Pose-Centric Cross-Attention (PCCAT). This layer ensures each pose is denoised based on its own condition and its neighboring poses. Consequently, FlowMDM can be trained on a dataset with only a single condition available per motion sequence and still generate realistic transitions when using multiple conditions at inference time.
- We reveal the lack of sensitivity of current HMC metrics to identify discontinuous or sharp transitions, and introduce two new metrics that help to detect them: the Peak Jerk (PJ) and the Area Under the Jerk (AUJ).

## 2. Related work

**Conditional human motion generation.** Recent studies in motion generation have shown notable progress in synthesizing movements conditioned on diverse modalities such as text [21, 26, 27, 35, 40, 63, 81, 82, 100–102], music [2, 17, 47, 77, 84, 96, 110], scenes [15, 30, 87–89, 97], interactive objects [1, 18, 42, 92], and even other humans’ behavior [9, 10, 28, 80, 93]. Traditionally, these approaches have been designed to generate motion sequences matching a sin-

gle condition. The progress of this domain has been boosted by the release of big datasets including diverse modalities or manual annotations [12, 26, 28, 47, 51, 60, 64, 65]. Research has also focused on problems like human motion prediction [3, 53, 57, 72, 78, 83, 86, 99] and motion infilling [29, 36, 39, 49, 50, 59, 67, 69, 75, 108], which do not rely on extensive manual annotations but rather on motion itself. Both tasks share a common challenge with HMC: the synthesized motion must not only be plausible but also integrate seamlessly with the neighboring behavior, ensuring fluidity and continuity. In this context, the utilization of human motion priors has been proven to be a successful technique to ensure any generated motion includes natural transitions [8, 46, 91]. In line with these approaches, our method learns a motion prior specifically tailored for HMC.

**Autoregressive human motion composition.** As in many other sequence modeling tasks, HMC was also first tackled with autoregressive methods. The gold standard has been pairing variational autoencoders with autoregressive decoders such as recurrent neural networks [104] or Transformers [4, 45, 48, 66]. Alternative approaches have introduced specialized reinforcement learning frameworks [52, 95, 105]. Autoregressive models rely on the availability of annotated motion transitions, a requirement that constrains the robustness of the models due to the scarcity of such data. To mitigate this issue, some methods include additional postprocessing steps like linear interpolations [4], or affine transformations [45]. However, these can distort the human motion dynamics and require a predetermined estimation of the transitions duration. Furthermore, autoregressive approaches generate motion solely based on the preceding motion. We argue that an accurate model should mimic the humans innate capacity to anticipate their next action and adapt their current behavior accordingly [24, 43].

**Diffusion-based human motion composition.** Diffusion models have excelled at conditional generation [20, 32, 74]. They also possess great zero-shot capabilities for image inpainting [70], and its equivalence in motion: motion infilling. DiffCollage [103], MultiDiffusion [7], and DoubleTake [73] proposed to modify the diffusion sampling process to simultaneously generate temporally superimposed motion sequences, and combine the estimated noise in the overlapped regions so that an infilled transition emerges. DoubleTake complemented such overlapped sampling with a refinement step in which the emerged transition undergoes further unconditional denoising steps. All these methods share two main limitations. First, they are constrained to modeling dependencies among neighboring motion sequences. This becomes a limitation when three or more consecutive actions share semantics and collectively represent a more comprehensive action. In this case, the motion dependencies may extend beyond contiguous actions. Second, they need to set the number of frames

that each transition takes between consecutive actions, for which extra computations are incorporated. Our work seeks to address these constraints by offering a solution able to model longer inter-sequence dynamics without imposing extra computational burdens or predefined transition durations.

### 3. Methodology

**Problem definition.** Our goal consists in generating a motion sequence of  $N$  frames, with the capability of conditioning the generated motion inside non-overlapping intervals  $[0, \tau_1), [\tau_1, \tau_2), \dots, [\tau_j, N)$ , with  $0 < \tau_1 < \dots < \tau_j < N$ . We will refer to the motion inside these intervals as *motion subsequences*, or  $\mathcal{S}_i = \{x_{\tau_i}, \dots, x_{\tau_{i+1}-1}\}$ , each driven by its corresponding condition  $c_i$ , and with a maximum length of  $L$ . It is essential that consecutive subsequences, influenced by different control signals, transition seamlessly and realistically. In particular, we aim at the even more challenging case where motion sequences containing several pairs of  $(\mathcal{S}_i, c_i)$  are not necessarily available in our dataset.

In this section, we present FlowMDM, an architecture with strong inductive biases that promote the emergence of a **robust translation-invariant motion prior**. Such *motion prior* is learned with a diffusion model equipped with a bidirectional (i.e., encoder-only) Transformer, similar to prior works [73, 82]. With it, we overcome the main limitations of autoregressive methods (Sec. 3.1). However, previous works are constrained in terms of motion duration. We could arguably provide extrapolation capabilities to the diffusion model by replacing the absolute positional encoding with a relative alternative, thus making the denoising of each pose *translation invariant*. However, this technique would fail to build complex compositional semantics that require knowledge about the start and end of each subsequence. For example, when generating the motion composition  $\mathcal{S}_i \rightarrow \mathcal{S}_{i+1}$  with  $c_i = \text{'walking'}$  and  $c_{i+1} = \text{'walk and sit down'}$ ,  $\mathcal{S}_{i+1}$  might only feature the action *'sit down'* because, with only relative positional information, the Transformer cannot know if the partially denoised *'walking'* motion preceding the beginning of  $\mathcal{S}_{i+1}$  belongs to  $\mathcal{S}_i$  or  $\mathcal{S}_{i+1}$ . To combine the benefits of both relative and absolute positional encodings, we introduce BPE (Sec. 3.2). This novel technique exploits the iterative nature of diffusion models to promote intra-subsequence global coherence in earlier denoising stages, while making later denoising stages translation invariant, ensuring that realistic and plausible transitions naturally emerge between subsequences. Still, during training, the condition remains unchanged throughout all ground truth motion sequences. In order to make our denoising model *robust* to having multiple conditions per sequence at inference, we introduce a new attention paradigm called PCCAT (Sec. 3.3). As a result, FlowMDM is able to simultaneously generate very

long compositions of human motion subsequences, all in harmony and fostering plausible transitions between them, without explicit supervision on transitions generation.

#### 3.1. Bidirectional diffusion

The cumulative nature of errors in autoregressive models often results in a decline in performance when generating long sequences [107]. This is exacerbated in HMC, where transitions are scarce or even missing in the training corpus, and the model needs to deal with domain shifts at inference. Another limitation of autoregressive methods is that the generated  $\mathcal{S}_i$  only depends on  $\{\mathcal{S}_j\}_{j < i}$ . We discussed in Sec. 2 why this is a suboptimal solution for HMC. Thus, an appropriate model for HMC should also be able to anticipate the following motion,  $\mathcal{S}_{i+1}$ , and possibly adapt  $\mathcal{S}_i$  so that the transition is feasible. We argue that the iterative paradigm of diffusion models provides very appropriate inductive biases for naturally mimicking such ability: the partially denoised  $\mathcal{S}_i$  and  $\mathcal{S}_{i+1}$  are refined later in successive denoising steps. By choosing a bidirectional Transformer as our denoising function [38], we enable the modeling of both past and future dependencies. Therefore, we design our framework as a bidirectional motion diffusion model, similar to MDM [82]. We refer the reader to [94] for more details on the theoretical aspects of diffusion models.

#### 3.2. Blended positional encodings

Diffusion models can learn strong motion priors that ensure any motion generated is realistic and plausible [73]. In fact, they can also generate smooth transitions between subsequences [7, 73, 103]. However, these capabilities stem from inference-time motion infilling techniques, which we argue do not exploit the full potential of human motion priors. In fact, building a prior that extrapolates well to sequences longer than those observed during training is very challenging. The field of natural language processing has made progress in sequence extrapolation techniques, notably by substituting absolute positional encoding (APE) with a relative (RPE) counterpart [37]. By only providing information regarding how far tokens are between them, they achieve sequence-wise translation invariance and, therefore, can extrapolate their modeling capabilities to longer sequences. Yet, the absolute positions of poses within a motion, including their distances to the start and end of the action, are necessary to build the global semantics of the motion, as exemplified at the beginning of this section.

Here, we propose BPE, a novel positional encoding scheme designed for diffusion models that enables motion extrapolation while preserving the global motion semantics. Our BPE is inspired by the observation that in motion, high frequencies encompass local fine details, whereas low frequencies capture global structures. Similar insights have been drawn for images [61]. Diffusion models ex-

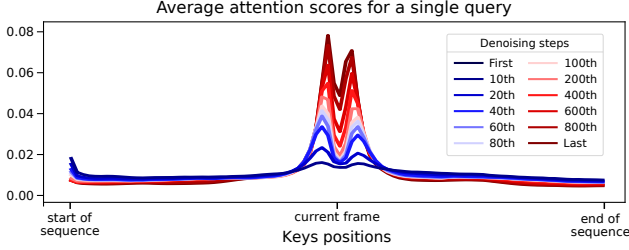


Figure 2. Attention scores of a single query pose (current frame) as a function of the pose attended to (x-axis) in a diffusion-based motion generation model with a sinusoidal absolute positional encoding. Curves show the scores at each denoising step. We observe that, whereas early steps show strong global dependencies (blue), later denoising stages exhibit a clearly local behavior (red).

cel at decomposing the generation process into recovering lower frequencies, and gradually transitioning to higher frequencies. Fig. 2 shows how at early denoising phases, motion diffusion models prioritize global inter-frame dependencies, shifting towards local relative dependencies as the process unfolds. The proposed BPE harmonizes these dynamics *during inference*: at early denoising stages, our denoising model is fed with an APE and, towards the conclusion, with an RPE. A scheduler guides this transition. As a result, intra-subsequence global dependencies are recovered at the beginning of the denoising, and intra- and inter-subsequences motion smoothness and realism are promoted later. To make the model understand APE and RPE at inference, we expose it to both encodings by randomly alternating them during training. As a result, the BPE schedule can be tuned at inference time to balance the intra-subsequence coherence and the inter-subsequence realism trade-off.

**Rotary Position Encoding (RoPE).** Our choice for RPE is rotary embeddings [76]. RoPE integrates a position embedding into the queries and keys, ensuring that after dot-product multiplication, the attention scores’ positional information reflects only the relative pairwise distance between queries and keys. Specifically, let  $W_q$  and  $W_k$  be the projection matrices into the  $d$ -dimensional spaces of queries and keys. Then, RoPE encodes the absolute positions  $m$  and  $n$  of a pair of query ( $q_m = W_q x_m$ ) and key ( $k_n = W_k x_n$ ), respectively, as  $d$ -dimensional rotations  $R_m^d, R_n^d$  over the projected poses  $x_m, x_n$ . The rotation angles are parameterized by  $m$  and  $n$  so that the attention formulation becomes:

$$q_m^T k_n = (R_m^d W_q x_m)^T (R_n^d W_k x_n) = x_m^T W_q R_{n-m}^d W_k x_n. \quad (1)$$

Note that the resulting rotation  $R_{n-m}^d$  only depends on the distance between  $n$  and  $m$ , and any absolute information about  $n$  or  $m$  is removed. RoPE is a natural choice for our RPE due to its simplicity and convenient injection before the attention takes place. As a result, RoPE is compatible with faster attention techniques like FlashAttention [22, 23].

**Sinusoidal Position Encoding.** Our APE is the classic sinusoidal position encoding [85], which leverages sine and

cosine functions to inject positional information. It is added to the queries, keys, and values of the attention layers.

Note that for APE, attention is limited to each subsequence, while for RPE, attention spans all frames up to the attention horizon  $H < L < N$ . Since  $L$  defines the maximum range of motion dynamics learned during RPE training, there is no advantage in setting  $H \geq L$  (Tabs. D/E in supp. material). Leveraging both APE and RPE constraints ensures quadratic complexity over the maximum subsequence length  $L$  in both memory and computation [11]. As a result, FlowMDM’s complexity is equivalent to that of other Transformer-based motion diffusion models [73, 103].

### 3.3. Pose-centric cross-attention

In order to make motion generation with diffusion models efficient, we would like to *simultaneously* generate very long sequences. In motion Transformers, the generation is conditioned at a sequence level by injecting the condition as a token [82], or as a sequence-wise transformation in intermediate layers [102]. Therefore, they cannot be conditioned on multiple signals in different subsequences. For this reason, diffusion-based methods for HMC opted for individually generating sequences and then merging them [73, 103]. To enable such simultaneous heterogeneous conditioning without any extra postprocessing, we propose to inject the condition at every frame. However, we still need to deal with a challenge: the condition never varies at training time. Therefore, at inference time, attention scores are computed with the embeddings  $E_{x_m, c_m}$  and  $E_{x_n, c_n}$  of the pose-condition pairs  $(x_m, c_m)$  and  $(x_n, c_n)$  as:

$$q_m^T k_n = (W_q E_{x_m, c_m})^T (W_k E_{x_n, c_n}) = E_{x_m, c_m}^T W_q^T W_k E_{x_n, c_n}. \quad (2)$$

When  $c_m \neq c_n$ ,  $q_m^T k_n$  was never encountered during training. If instead of injecting the condition at every frame, we used cross-attention layers, distinct conditions would also be temporally mixed, and we would face the same problem. To reduce the presence and impact of such training-inference misalignment, we introduce PCCAT, see Fig. 3, which aims at minimizing the entanglement between conditions and noisy poses. Specifically, PCCAT combines every frame’s noisy pose and condition into queries, while using only noisy poses as keys and values. Thus, Eq. 2 becomes:

$$q_m^T k_n = (W_q E_{x_m, c_m})^T (W_k E_{x_n}) = E_{x_m, c_m}^T W_q^T W_k E_{x_n}. \quad (3)$$

With PCCAT, the attention output for pose  $m$  becomes a weighted average of the value projections of its neighboring noisy poses. A residual connection adds the PCCAT output to the noisy poses. With comprehensive coverage of the motion spectrum in the training dataset, the network observes various poses preceding and following each pose, particularly within its local neighborhood. Therefore, local relationships do not suffer from unseen intermediate representations. Still, there is an obstacle to address: long-range de-



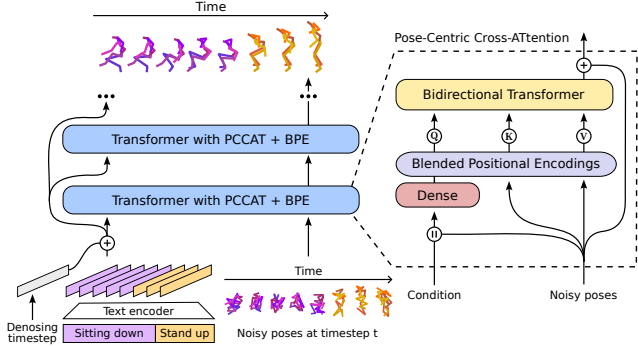


Figure 3. **Pose-centric cross-attention.** Our attention minimizes the entanglement between the control signal (e.g., text, objects) and the noisy motion by feeding the former only to the query. Consequently, our model denoises each frame’s noisy pose only leveraging its own condition, and the neighboring noisy poses.

dependencies. However, as discussed in Sec. 3.2, their importance is mostly confined to the initial stages of denoising. There, the network is exposed to very noisy motion data, thus becoming robust to such unseen combinations of poses. In the latest denoising stages, when the network deals with almost clean input sequences, global dependencies have already been developed and attention is short-ranged (Fig. 2).

## 4. Experiments

### 4.1. Experimental setup

**Datasets.** Our experiments are conducted on the Babel [65] and HumanML3D [26] datasets, with their train and test splits. HumanML3D features multiple textual descriptions of each motion sequence, but lacks explicit transition annotations, making supervised learning infeasible for transition generation. Babel, on the other hand, provides finely-grained textual descriptions at an atomic level, including transitions, which facilitates more precise and dynamic motion control but also presents a greater challenge due to fast and short transitions. To demonstrate the flexibility of FlowMDM, we employ the standard motion representations provided with each dataset. HumanML3D utilizes a 263D pose vector that includes joint coordinates, angles, velocities, and feet contact. By contrast, Babel uses the global position and orientation and a 6D rotation representation [106] of the SMPL model joints [13], as in [63].

**Evaluation.** Our evaluation uses the metrics established by [26], and later refined for this task in [48, 73, 95]. More specifically, motion sequences are synthesized as compositions of 32 pairs of textual descriptions and their durations. The 32 subsequences and the 31 transitions between  $S_{i-1}$  and  $S_i$  pairs are evaluated independently. In particular, each transition is defined as the set of consecutive poses  $\{x_{\tau_i-L_{tr}/2}, \dots, x_{\tau_i+L_{tr}/2-1}\}$ , sharing  $\frac{L_{tr}}{2}$  frames with  $S_{i-1}$  and  $S_i$ . The transition duration  $L_{tr}$  is set to 30

and 60 frames for Babel and HumanML3D (1 and 3 seconds), respectively. The top-3 R-precision (R-prec), and the multimodal distance (MM-Dist) are used to evaluate how well the subsequences’ motion matches their textual description [26]. The FID score and the average pairwise distance among all motion embeddings (diversity) assess the quality and variety of both subsequences and transitions, respectively [26, 31]. All metrics are averaged over 10 runs with 95% confidence intervals reported.

**Closing the gap: the Jerk.** Generative models are hard to evaluate [19, 71, 94]. The FID score [31] has proven to be a very reliable metric in quantifying the similarity between distributions of generated and real motion data while being sensitive to motion artifacts or noise [55]. Nevertheless, exclusively relying on perceptual metrics like FID for assessing transition quality can be misleading due to their insensitivity to motion anomalies such as abrupt accelerations [8], or foot skating [56]. To complement the FID, our work introduces two novel metrics built upon the concept of *jerk* (i.e., the time derivative of acceleration), which is indicative of motion smoothness and known to be sensitive to kinetic irregularities [5, 6, 16, 25, 34, 44, 98]. Given that natural human motion typically exhibits constrained jerk due to relatively consistent acceleration patterns [25, 44], our metrics are tailored to highlight *persistent deviations* from this norm in generated transitions. Firstly, we compute the Peak Jerk (PJ), taking the maximum value found throughout the transition motion over all joints. While this measure captures extreme fluctuations, it may favor models that unnaturally smooth transitions across several wider peaks of jerk. To measure this undesirable effect, we introduce the Area Under the Jerk (AUJ), calculated as the sum of L1-norm differences between a method’s instantaneous jerk and the dataset’s average jerk value. This measure serves as an aggregate indicator of motion smoothness, quantifying the cumulative deviation from natural human movement across the entire transition. The PJ and AUJ of a transition are formally defined as follows:

$$PJ = \max_{\substack{1 \leq i \leq K \\ 1 \leq \tau \leq L_{tr}}} |j_i(\tau)|_1, \quad AUJ = \sum_{\tau=1}^{L_{tr}} \max_{1 \leq i \leq K} |j_i(\tau) - j_{avg}|_1, \quad (4)$$

where  $j_i(\tau)$  is the jerk at time  $\tau$  for joint  $i$ ,  $K$  is the number of joints, and  $j_{avg}$  is the average joints-wise maximum jerk across the dataset.

**Baselines.** We compare our method to publicly released related works that can generate sequential motions from text: the autoregressive TEACH [4], and the diffusion sampling techniques DoubleTake [73], DiffCollage [103], and MultiDiffusion [7]. Sampling techniques are evaluated with PCCAT and APE for a fairer comparison. Additionally, we evaluate TEACH with its spherical linear interpolation over transitions turned off (TEACH\_B), and DoubleTake with

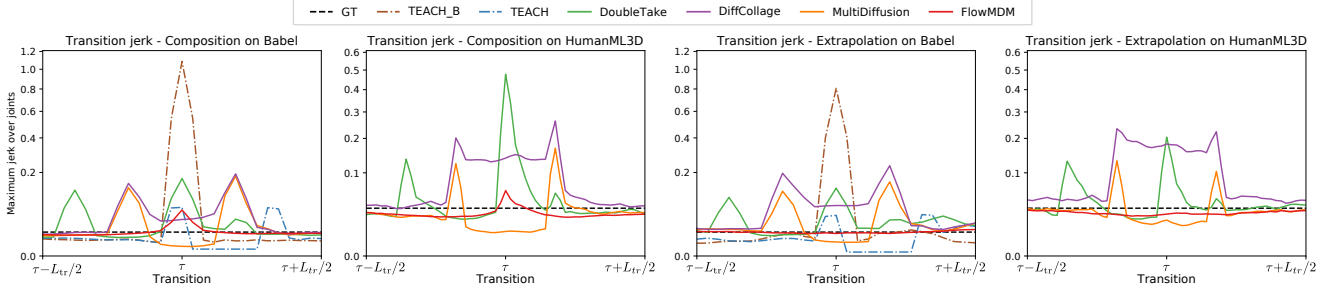


Figure 4. **Transitions smoothness.** Average maximum jerk over joints at each frame of the transitions for both motion composition (left) and extrapolation (right) tasks. While other methods show severe smoothness artifacts in the beginning and end of their transition refinement processes, FlowMDM’s jerk curve has the shortest peak for composition, and an absence of peaks for extrapolation.

	R-prec $\uparrow$	Subsequence			FID $\downarrow$	Transition		
		FID $\downarrow$	Div $\rightarrow$	MM-Dist $\downarrow$		Div $\rightarrow$	PJ $\rightarrow$	AUJ $\downarrow$
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
TEACH_B	<b>0.703</b> $\pm$ 0.002	1.71 $\pm$ 0.03	8.18 $\pm$ 0.14	<b>3.43</b> $\pm$ 0.01	3.01 $\pm$ 0.04	<b>6.23</b> $\pm$ 0.05	1.09 $\pm$ 0.00	2.35 $\pm$ 0.01
TEACH	0.655 $\pm$ 0.002	1.82 $\pm$ 0.02	7.96 $\pm$ 0.11	3.72 $\pm$ 0.01	3.27 $\pm$ 0.04	6.14 $\pm$ 0.06	0.07 $\pm$ 0.00	0.44 $\pm$ 0.00
DoubleTake*	0.596 $\pm$ 0.005	3.16 $\pm$ 0.06	7.53 $\pm$ 0.11	4.17 $\pm$ 0.02	3.33 $\pm$ 0.06	<u>6.16</u> $\pm$ 0.05	0.28 $\pm$ 0.00	1.04 $\pm$ 0.01
DoubleTake	0.668 $\pm$ 0.005	<u>1.33</u> $\pm$ 0.04	7.98 $\pm$ 0.12	3.67 $\pm$ 0.03	3.15 $\pm$ 0.05	6.14 $\pm$ 0.07	0.17 $\pm$ 0.00	0.64 $\pm$ 0.01
MultiDiffusion	0.702 $\pm$ 0.005	1.74 $\pm$ 0.04	<b>8.37</b> $\pm$ 0.13	<b>3.43</b> $\pm$ 0.02	6.56 $\pm$ 0.12	5.72 $\pm$ 0.07	0.18 $\pm$ 0.00	0.68 $\pm$ 0.00
DiffCollage	0.671 $\pm$ 0.003	1.45 $\pm$ 0.05	7.93 $\pm$ 0.09	3.71 $\pm$ 0.01	4.36 $\pm$ 0.09	6.09 $\pm$ 0.08	0.19 $\pm$ 0.00	0.84 $\pm$ 0.01
FlowMDM	<u>0.702</u> $\pm$ 0.004	<b>0.99</b> $\pm$ 0.04	<u>8.36</u> $\pm$ 0.13	3.45 $\pm$ 0.02	<b>2.61</b> $\pm$ 0.06	6.47 $\pm$ 0.05	<b>0.06</b> $\pm$ 0.00	<b>0.13</b> $\pm$ 0.00

Table 1. Comparison of FlowMDM with the state of the art in Babel. Symbols  $\uparrow$ ,  $\downarrow$ , and  $\rightarrow$  indicate that higher, lower, or values closer to the ground truth (GT) are better, respectively. Evaluation is run 10 times and  $\pm$  specifies the 95% confidence intervals.

	R-prec $\uparrow$	Subsequence			FID $\downarrow$	Transition		
		FID $\downarrow$	Div $\rightarrow$	MM-Dist $\downarrow$		Div $\rightarrow$	PJ $\rightarrow$	AUJ $\downarrow$
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
DoubleTake*	0.643 $\pm$ 0.005	0.80 $\pm$ 0.02	9.20 $\pm$ 0.11	3.92 $\pm$ 0.01	1.71 $\pm$ 0.05	<b>8.82</b> $\pm$ 0.13	0.52 $\pm$ 0.01	2.10 $\pm$ 0.03
DoubleTake	0.628 $\pm$ 0.005	1.25 $\pm$ 0.04	9.09 $\pm$ 0.12	4.01 $\pm$ 0.01	4.19 $\pm$ 0.09	8.45 $\pm$ 0.09	0.48 $\pm$ 0.00	1.83 $\pm$ 0.02
MultiDiffusion	0.629 $\pm$ 0.002	1.19 $\pm$ 0.03	<b>9.38</b> $\pm$ 0.08	4.02 $\pm$ 0.01	4.31 $\pm$ 0.06	8.37 $\pm$ 0.10	0.17 $\pm$ 0.00	1.06 $\pm$ 0.01
DiffCollage	0.615 $\pm$ 0.005	1.56 $\pm$ 0.04	8.79 $\pm$ 0.08	4.13 $\pm$ 0.02	4.59 $\pm$ 0.10	8.22 $\pm$ 0.11	0.26 $\pm$ 0.00	2.85 $\pm$ 0.09
FlowMDM	<b>0.685</b> $\pm$ 0.004	<b>0.29</b> $\pm$ 0.01	9.58 $\pm$ 0.12	<b>3.61</b> $\pm$ 0.01	<b>1.38</b> $\pm$ 0.05	8.79 $\pm$ 0.09	<b>0.06</b> $\pm$ 0.00	<b>0.51</b> $\pm$ 0.01

Table 2. Comparison of FlowMDM with the state of the art in HumanML3D.

MDM, as originally proposed (DoubleTake\*). TEACH and TEACH\_B cannot be trained for HumanML3D due to the lack of pairs of consecutive actions and textual descriptions.

**Implementation details.** We tune the hyperparameters of all models with grid search. The attention horizon for RPE,  $H$ , is set to 100/150 for Babel/HumanML3D. The number of diffusion steps is 1K for all experiments. Our model is trained with the  $x_0$  parameterization [90], and minimizes the L2 reconstruction loss. During training, RPE and APE are alternated randomly at a frequency of 0.5. We use classifier-free guidance with weights 1.5/2.5 [33]. We use a binary step function to guide the BPE sampling, yielding 125/60 initial APE steps. The minimum/maximum lengths for training subsequences are set to 30/200 and 70/200 frames (i.e., 1/6.7s and 3.5/10s). For Babel, training subsequences include consecutive ground truth motions with distinct textual descriptions in order to increase the motions variability, and make the network explicitly robust to multiple conditions. The ablation study includes two conditioning baselines: 1) concatenating each frame’s condition

and noisy pose, and replacing the PCCAT with vanilla self-attention (SAT), and 2) injecting the condition with cross-attention layers (CAT). See more details in supp. material Sec. A.

## 4.2. Quantitative analysis

**Comparison with the state of the art on HMC.** Tables 1 and 2 show the comparison of FlowMDM with current state-of-the-art models in Babel and HumanML3D datasets, respectively. In HumanML3D, our model outperforms by a fair margin the other methods in terms of subsequence accuracy-wise metrics (R-prec and MM-Dist), and FID. In Babel, it matches the state of the art in accuracy and excels in FID score. FlowMDM produces transitions of higher quality and smoothness on both datasets, as indicated by FID, PJ, and AUJ metrics. The lack of correlation between the FID score and the AUJ underscores the importance of the latter as a complementary metric for assessing smoothness. Fig. 4-left shows the average jerk values across the generated transitions. We observe that state-of-

Cond.	Train. PE	Inf. PE	Subsequence				Transition			
			R-prec $\uparrow$	FID $\downarrow$	Div $\rightarrow$	MM-Dist $\downarrow$	FID $\downarrow$	Div $\rightarrow$	PJ $\rightarrow$	AUJ $\downarrow$
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
PCCAT	A	A	0.699 $\pm$ 0.004	1.34 $\pm$ 0.04	<b>8.36</b> $\pm$ 0.12	3.40 $\pm$ 0.02	4.26 $\pm$ 0.07	5.98 $\pm$ 0.06	1.81 $\pm$ 0.01	3.73 $\pm$ 0.01
PCCAT	R	R	0.635 $\pm$ 0.006	1.28 $\pm$ 0.03	8.05 $\pm$ 0.11	4.02 $\pm$ 0.02	2.18 $\pm$ 0.07	<b>6.14</b> $\pm$ 0.08	<u>0.03</u> $\pm$ 0.00	0.20 $\pm$ 0.00
PCCAT	B	A	0.716 $\pm$ 0.006	1.20 $\pm$ 0.04	8.31 $\pm$ 0.14	3.32 $\pm$ 0.02	3.01 $\pm$ 0.06	6.35 $\pm$ 0.07	1.78 $\pm$ 0.01	3.66 $\pm$ 0.02
PCCAT	B	R	0.635 $\pm$ 0.004	<b>0.85</b> $\pm$ 0.02	8.25 $\pm$ 0.12	3.98 $\pm$ 0.02	<u>2.14</u> $\pm$ 0.04	6.44 $\pm$ 0.09	0.04 $\pm$ 0.00	0.15 $\pm$ 0.00
SAT	B	B	0.681 $\pm$ 0.004	1.52 $\pm$ 0.04	8.22 $\pm$ 0.11	3.61 $\pm$ 0.02	<b>1.91</b> $\pm$ 0.03	6.41 $\pm$ 0.07	0.06 $\pm$ 0.00	<u>0.12</u> $\pm$ 0.00
CAT	B	B	<b>0.719</b> $\pm$ 0.004	1.29 $\pm$ 0.02	8.16 $\pm$ 0.13	<b>3.27</b> $\pm$ 0.02	2.57 $\pm$ 0.08	<u>6.06</u> $\pm$ 0.07	<b>0.02</b> $\pm$ 0.00	<b>0.07</b> $\pm$ 0.00
PCCAT	B	B	0.702 $\pm$ 0.004	<u>0.99</u> $\pm$ 0.04	<b>8.36</b> $\pm$ 0.13	3.45 $\pm$ 0.02	2.61 $\pm$ 0.06	6.47 $\pm$ 0.05	0.06 $\pm$ 0.00	0.13 $\pm$ 0.00

Table 3. Ablation study in Babel. Cond. indicates the conditioning scheme, Train./Inf. PE specify the positional encodings (PE) used at training/inference time, and A, R, and B refer to absolute, relative, and blended PE, respectively.  $\uparrow$ ,  $\downarrow$ , and  $\rightarrow$  indicate that higher, lower, or values closer to the ground truth (GT) are better, respectively. Evaluation is run 10 times and  $\pm$  specifies the 95% confidence intervals.

Cond.	Train. PE	Inf. PE	Subsequence				Transition			
			R-prec $\uparrow$	FID $\downarrow$	Div $\rightarrow$	MM-Dist $\downarrow$	FID $\downarrow$	Div $\rightarrow$	PJ $\rightarrow$	AUJ $\downarrow$
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
PCCAT	A	A	0.689 $\pm$ 0.005	0.66 $\pm$ 0.02	9.73 $\pm$ 0.12	3.63 $\pm$ 0.02	3.90 $\pm$ 0.12	8.29 $\pm$ 0.08	1.50 $\pm$ 0.01	3.40 $\pm$ 0.02
PCCAT	R	R	0.531 $\pm$ 0.005	1.75 $\pm$ 0.07	8.71 $\pm$ 0.10	4.80 $\pm$ 0.03	2.53 $\pm$ 0.12	8.62 $\pm$ 0.08	<b>0.03</b> $\pm$ 0.00	0.58 $\pm$ 0.01
PCCAT	B	A	<b>0.699</b> $\pm$ 0.005	0.61 $\pm$ 0.02	9.76 $\pm$ 0.10	3.54 $\pm$ 0.02	2.42 $\pm$ 0.09	8.39 $\pm$ 0.09	1.40 $\pm$ 0.01	3.29 $\pm$ 0.02
PCCAT	B	R	0.554 $\pm$ 0.007	1.06 $\pm$ 0.06	9.02 $\pm$ 0.11	4.54 $\pm$ 0.02	<b>1.12</b> $\pm$ 0.04	<b>9.00</b> $\pm$ 0.10	0.05 $\pm$ 0.00	0.53 $\pm$ 0.01
SAT	B	B	0.692 $\pm$ 0.004	0.49 $\pm$ 0.02	9.08 $\pm$ 0.09	<b>3.51</b> $\pm$ 0.01	3.19 $\pm$ 0.08	8.09 $\pm$ 0.11	<u>0.04</u> $\pm$ 0.00	<b>0.36</b> $\pm$ 0.02
CAT	B	B	0.622 $\pm$ 0.005	1.27 $\pm$ 0.04	8.86 $\pm$ 0.15	4.10 $\pm$ 0.01	3.93 $\pm$ 0.14	8.23 $\pm$ 0.10	<u>0.04</u> $\pm$ 0.00	<u>0.49</u> $\pm$ 0.02
PCCAT	B	B	0.685 $\pm$ 0.004	<b>0.29</b> $\pm$ 0.01	<b>9.58</b> $\pm$ 0.12	3.61 $\pm$ 0.01	<u>1.38</u> $\pm$ 0.05	<u>8.79</u> $\pm$ 0.09	0.06 $\pm$ 0.00	0.51 $\pm$ 0.01

Table 4. Ablation study in HumanML3D.

the-art methods exhibit severe smoothness artifacts. During TEACH’s spherical linear interpolation, the jerk quickly reaches values near zero. By contrast, DiffCollage leans toward higher-than-average jerk values, while MultiDiffusion exhibits the opposite trend. DoubleTake shows three peaks, caused by their two-stage noise estimation process. In comparison, FlowMDM successfully minimizes peak jerk values, producing the smoothest transitions between subsequences. See supp. material Sec. C for in-depth analyses.

**Human motion extrapolation.** In single text-to-motion, the duration of the generated motion is limited to the longest subsequence length  $L$  available in the training set. Extrapolating periodic actions into sequences longer than those in the ground truth presents a notable challenge. Achieving this through HMC requires the harmonization of periodicity across adjacent subsequences. However, common strategies that combine independently generated subsequences often disrupt the periodicity of the motion. To assess our model’s capabilities in addressing this issue, we construct an evaluation set comprising 32 consecutive repetitions of 32 different extrapolatable actions such as ‘walk forward’, ‘jumping’, or ‘playing the guitar’, extracted from the Babel and HumanML3D test sets (more details in supp. material Sec. B). Fig. 4-right displays the motion jerk across transitions for all models on this task. We observe that, while other models exhibit smoothness anomalies similar to those shown in the HMC evaluation, FlowMDM closely mirrors the ground truth jerk. This observation indicates that the jerk peak noted in FlowMDM for the composition task is

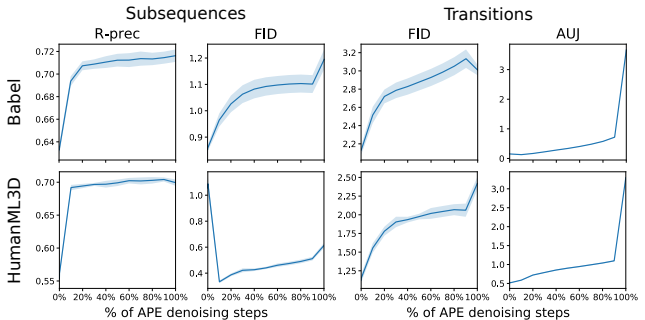


Figure 5. **BPE trade-offs.** Increasing the number of APE steps undergone during BPE sampling improves the correspondence between motion and textual description (R-prec), but reduces the transition realism and smoothness (FID and AUJ). The best balance is reached around 10% of APE denoising steps.

likely attributed to smoothness irregularities in more complex transitions.

**Ablation study.** The effectiveness of BPE and PCCAT is presented in Tables 3 and 4. Reasonably, the baseline model trained solely with APE fails to generate smooth transitions. Conversely, a model trained only with RPE, despite producing the smoothest transitions, struggles to model global motion dependencies and accurately reflect the corresponding textual descriptions. Interestingly, training with BPE improves the performance of both APE- and RPE-only samplings. Sampling with BPE combines the best of both worlds by preserving the excellent AUJ values of the RPE models and reaching the state-of-the-art accuracy and FID

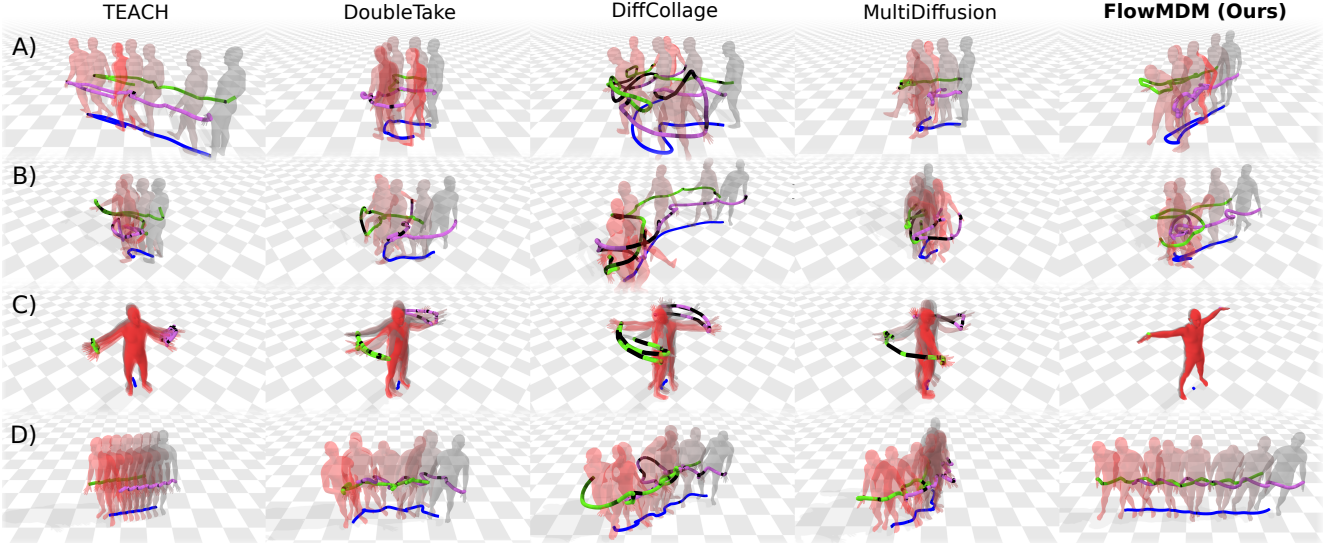


Figure 6. **Qualitative analysis (Babel)**. A) and B) show compositions of 3 motions (‘walk straight’→‘side steps’→‘walk backward’, and ‘walk’→‘turn around’→‘sit on the bench’, respectively), and C) and D) illustrate extrapolations that repeat 6 times a static (‘t-pose’) and a dynamic (‘step to the right’) action, respectively. 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. FlowMDM exhibits the most fluid motion and preserves the staticity or periodicity of extrapolated actions, in contrast to other methods that show spontaneous high jerk values and fail to keep the motion coherence in extrapolations.

scores of the APE models. Fig. 5 illustrates this balance. Specifically, increasing the number of APE steps enhances the motion’s congruence with the textual description, at the cost of reducing the smoothness and realism of the transitions. In HumanML3D, the SAT and CAT conditioning schemes lead to worse transitions in terms of FID and diversity. This is caused by the coexistence of different conditions in the local neighborhood of the transition at inference, which never happens during training. Our PCCAT conditioning technique effectively solves this problem. In Babel, such effect is not present because the training motion sequences include several subsequences, thus increasing the model’s robustness to transitions with varying conditions.

**On the efficiency of FlowMDM.** Diffusion-based state-of-the-art methods such as MultiDiffusion and DiffCollage denoise poses from the transition more than once in order to harmonize it with the adjacent motions. DoubleTake’s transitions undergo an additional denoising process, which adds computational burden and can not be parallelized. Oppositely, FlowMDM does not apply redundant denoising steps to any pose. In particular, our model goes through 47.1%, 28.4%, and 16.5% less pose-wise denoising steps than DoubleTake, DiffCollage, and MultiDiffusion, respectively.

### 4.3. Qualitative results

Fig. 6 illustrates how our quantitative findings translate into visual outcomes on the human motion composition and extrapolation tasks. First, as anticipated by Fig. 4, we confirm that state-of-the-art methods produce short intervals of

jerk peaks around transitions. These do not typically match long-range motion scenarios, where such jerks might be contextually appropriate. Contrarily, FlowMDM produces motion that is realistic, accurate, and smooth. Particularly, we notice that DiffCollage’s bias toward producing constantly high jerk values around transitions is perceived as an overall chaotic motion. Due to the independent generation of their subsequences, DoubleTake, DiffCollage, and MultiDiffusion are unable to maintain the static or periodic nature of actions when extrapolating them. Only TEACH and FlowMDM are able to successfully extrapolate a static ‘t-pose’, and ours is the only one capable of extrapolating a ‘step to the right’ sequence realistically. Finally, FlowMDM also inherits the trajectory control capabilities of motion diffusion models as shown in Fig. 1-right.

## 5. Conclusion

We presented FlowMDM, the first approach that generates human motion compositions simultaneously, without undergoing postprocessing or redundant denoising diffusion steps. We also introduced the blended positional encodings to combine the benefits of absolute and relative positional encodings during the denoising chain. Finally, we presented the pose-centric cross-attention, a technique that improves the generation of transitions when training with only a single condition per motion sequence.

**Limitations and future work.** The absolute stage of BPE does not model relationships between subsequences.



Consequently, their low-frequency spectrum is generated independently. This limitation could be addressed in future work by incorporating an intention planning module. Finally, our method learns a strong motion prior that generates transitions between combinations of actions never seen at training time. Such capability could theoretically be used with different models leveraging different control signals, assuming they all are trained under the same framework. Future work will experimentally validate this hypothesis.

## References

- [1] Vida Adeli, Mahsa Ehsanpour, Ian Reid, Juan Carlos Niebles, Silvio Savarese, Ehsan Adeli, and Hamid Rezaatoughi. Tripod: Human trajectory and pose dynamics forecasting in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13390–13400, 2021. 2
- [2] Simon Alexanderson, Rajmund Nagy, Jonas Beskow, and Gustav Eje Henter. Listen, denoise, action! audio-driven motion synthesis with diffusion models. *ACM Transactions on Graphics (TOG)*, 42(4):1–20, 2023. 2
- [3] Sadegh Aliakbarian, Microsoft Fatemeh Saleh ACRV, Stephen Gould ACRV, and Anu Mathieu Salzmman CVLab. Contextually plausible and diverse 3d human motion prediction. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 2
- [4] Nikos Athanasiou, Mathis Petrovich, Michael J Black, and Gül Varol. Teach: Temporal action composition for 3d humans. In *2022 International Conference on 3D Vision (3DV)*, pages 414–423. IEEE, 2022. 2, 5
- [5] Sivakumar Balasubramanian, Alejandro Melendez-Calderon, and Etienne Burdet. A robust and sensitive metric for quantifying movement smoothness. *IEEE transactions on biomedical engineering*, 59(8):2126–2136, 2011. 5
- [6] Sivakumar Balasubramanian, Alejandro Melendez-Calderon, Agnes Roby-Brami, and Etienne Burdet. On the analysis of movement smoothness. *Journal of neuroengineering and rehabilitation*, 12(1):1–11, 2015. 5
- [7] Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. Multidiffusion: Fusing diffusion paths for controlled image generation. *arXiv preprint arXiv:2302.08113*, 2023. 2, 3, 5
- [8] German Barquero, Sergio Escalera, and Cristina Palmero. Belfusion: Latent diffusion for behavior-driven human motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2317–2327, 2023. 2, 5
- [9] 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*, pages 139–178. PMLR, 2022. 2
- [10] 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*, pages 107–138. PMLR, 2022. 2
- [11] Iz Beltagy, Matthew E Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020. 4
- [12] Bharat Lal Bhatnagar, Xianghui Xie, Ilya A Petrov, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Behave: Dataset and method for tracking human object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15935–15946, 2022. 2
- [13] 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 European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part V 14*, pages 561–578. Springer, 2016. 5, 20
- [14] Paulo Vinicius Koerich Borges, Nicola Conci, and Andrea Cavallaro. Video-based human behavior understanding: A survey. *IEEE transactions on circuits and systems for video technology*, 23(11):1993–2008, 2013. 1
- [15] Zhe Cao, Hang Gao, Karttikeya Mangalam, Qi-Zhi Cai, Minh Vo, and Jitendra Malik. Long-term human motion prediction with scene context. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 387–404. Springer, 2020. 2
- [16] Angela Castillo, Maria Escobar, Guillaume Jeanneret, Albert Pumarola, Pablo Arbeláez, Ali Thabet, and Arsiom Sanakoyeu. Bodiffusion: Diffusing sparse observations for full-body human motion synthesis. *arXiv preprint arXiv:2304.11118*, 2023. 5
- [17] Kang Chen, Zhipeng Tan, Jin Lei, Song-Hai Zhang, Yuan-Chen Guo, Weidong Zhang, and Shi-Min Hu. Choreomaster: choreography-oriented music-driven dance synthesis. *ACM Transactions on Graphics (TOG)*, 40(4):1–13, 2021. 2
- [18] Enric Corona, Albert Pumarola, Guillem Alenya, and Francesc Moreno-Noguer. Context-aware human motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6992–7001, 2020. 2
- [19] Antonia Creswell, Tom White, Vincent Dumoulin, Kai Arulkumaran, Biswa Sengupta, and Anil A Bharath. Generative adversarial networks: An overview. *IEEE signal processing magazine*, 35(1):53–65, 2018. 5
- [20] Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 2
- [21] Rishabh Dabral, Muhammad Hamza Mughal, Vladislav Golyanik, and Christian Theobalt. Mofusion: A framework for denoising-diffusion-based motion synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9760–9770, 2023. 2
- [22] Tri Dao. Flashattention-2: Faster attention with better parallelism and work partitioning. *arXiv preprint arXiv:2307.08691*, 2023. 4

- [23] Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in Neural Information Processing Systems*, 35:16344–16359, 2022. 4
- [24] David A Engström, JA Scott Kelso, and Tom Holroyd. Reaction-anticipation transitions in human perception-action patterns. *Human movement science*, 15(6):809–832, 1996. 2
- [25] Philipp Gulde and Joachim Hermsdörfer. Smoothness metrics in complex movement tasks. *Frontiers in neurology*, 9:615, 2018. 5
- [26] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5152–5161, 2022. 2, 5
- [27] Chuan Guo, Xinxin Zuo, Sen Wang, and Li Cheng. Tm2t: Stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts. In *European Conference on Computer Vision*, pages 580–597. Springer, 2022. 2
- [28] Wen Guo, Xiaoyu Bie, Xavier Alameda-Pineda, and Francesc Moreno-Noguer. Multi-person extreme motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13053–13064, 2022. 2
- [29] Félix G Harvey, Mike Yurick, Derek Nowrouzezahrai, and Christopher Pal. Robust motion in-betweening. *ACM Transactions on Graphics (TOG)*, 39(4):60–1, 2020. 2
- [30] Mohamed Hassan, Duygu Ceylan, Ruben Villegas, Jun Saito, Jimei Yang, Yi Zhou, and Michael J Black. Stochastic scene-aware motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11374–11384, 2021. 2
- [31] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 5
- [32] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 2
- [33] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 6, 16
- [34] Neville Hogan and Dagmar Sternad. Sensitivity of smoothness measures to movement duration, amplitude, and arrests. *Journal of motor behavior*, 41(6):529–534, 2009. 5
- [35] Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. Motiongpt: Human motion as a foreign language. *arXiv preprint arXiv:2306.14795*, 2023. 2
- [36] Manuel Kaufmann, Emre Aksan, Jie Song, Fabrizio Pece, Remo Ziegler, and Otmar Hilliges. Convolutional autoencoders for human motion infilling. In *2020 International Conference on 3D Vision (3DV)*, pages 918–927. IEEE, 2020. 2
- [37] Amirhossein Kazemnejad, Inkit Padhi, Karthikeyan Natesan Ramamurthy, Payel Das, and Siva Reddy. The impact of positional encoding on length generalization in transformers. *arXiv preprint arXiv:2305.19466*, 2023. 3
- [38] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186, 2019. 3
- [39] Jihoon Kim, Taehyun Byun, Seungyou Shin, Jungdam Won, and Sungjoon Choi. Conditional motion in-betweening. *Pattern Recognition*, 132:108894, 2022. 2
- [40] Jihoon Kim, Jiseob Kim, and Sungjoon Choi. Flame: Free-form language-based motion synthesis & editing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 8255–8263, 2023. 2
- [41] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 14
- [42] Nilesh Kulkarni, Davis Rempe, Kyle Genova, Abhijit Kundu, Justin Johnson, David Fouhey, and Leonidas Guibas. Nifty: Neural object interaction fields for guided human motion synthesis. *arXiv preprint arXiv:2307.07511*, 2023. 2
- [43] Wilfried Kunde, Katrin Elsner, and Andrea Kiesel. No anticipation–no action: the role of anticipation in action and perception. *Cognitive Processing*, 8:71–78, 2007. 2
- [44] Caroline Larboulette and Sylvie Gibet. A review of computable expressive descriptors of human motion. In *Proceedings of the 2nd International Workshop on Movement and Computing*, pages 21–28, 2015. 5
- [45] Taeryung Lee, Gyeongsik Moon, and Kyoung Mu Lee. Multiact: Long-term 3d human motion generation from multiple action labels. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 1231–1239, 2023. 2
- [46] Jiaman Li, Ruben Villegas, Duygu Ceylan, Jimei Yang, Zhengfei Kuang, Hao Li, and Yajie Zhao. Task-generic hierarchical human motion prior using vaes. In *2021 International Conference on 3D Vision (3DV)*, pages 771–781. IEEE, 2021. 2
- [47] Ruilong Li, Shan Yang, David A Ross, and Angjoo Kanazawa. Ai choreographer: Music conditioned 3d dance generation with aist++. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13401–13412, 2021. 2
- [48] Shuai Li, Sisi Zhuang, Wenfeng Song, Xinyu Zhang, Hejia Chen, and Aimin Hao. Sequential texts driven cohesive motions synthesis with natural transitions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9498–9508, 2023. 2, 5
- [49] Weiyu Li, Xuelin Chen, Peizhuo Li, Olga Sorkine-Hornung, and Baoquan Chen. Example-based motion synthesis via generative motion matching. *arXiv preprint arXiv:2306.00378*, 2023. 2
- [50] Yunhao Li, Zhenbo Yu, Yucheng Zhu, Bingbing Ni, Guangtao Zhai, and Wei Shen. Skeleton2humanoid: Animating simulated characters for physically-plausible motion in-betweening. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 1493–1502, 2022. 2
- [51] Jing Lin, Ailing Zeng, Shunlin Lu, Yuanhao Cai, Ruimao Zhang, Haoqian Wang, and Lei Zhang. Motion-x: A large-scale 3d expressive whole-body human motion dataset. In *Thirty-seventh Conference on Neural Information Process-*

- ing Systems Datasets and Benchmarks Track*, 2023. [2](#)
- [52] Zhengyi Luo, Jinkun Cao, Kris Kitani, Weipeng Xu, et al. Perpetual humanoid control for real-time simulated avatars. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10895–10904, 2023. [2](#)
- [53] 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. [2](#)
- [54] Shugao Ma, Tomas Simon, Jason Saragih, Dawei Wang, Yuecheng Li, Fernando De La Torre, and Yaser Sheikh. Pixel codec avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 64–73, 2021. [1](#)
- [55] Antoine Maiorca, Youngwoo Yoon, and Thierry Dutoit. Evaluating the quality of a synthesized motion with the fréchet motion distance. In *ACM SIGGRAPH 2022 Posters*, pages 1–2, 2022. [5](#)
- [56] Antoine Maiorca, Youngwoo Yoon, and Thierry Dutoit. Validating objective evaluation metric: Is fréchet motion distance able to capture foot skating artifacts? In *Proceedings of the 2023 ACM International Conference on Interactive Media Experiences*, pages 242–247, 2023. [5](#)
- [57] 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. [2](#)
- [58] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021. [14](#), [15](#)
- [59] Boris N Oreshkin, Antonios Valkanas, Félix G Harvey, Louis-Simon Ménard, Florent Bocquet, and Mark J Coates. Motion in-betweening via deep delta-interpolator. *IEEE Transactions on Visualization and Computer Graphics*, 2023. [2](#)
- [60] Cristina Palmero, German Barquero, Julio CS Jacques Junior, Albert Clapés, Johnny Núñez, 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](#)
- [61] Zizheng Pan, Jianfei Cai, and Bohan Zhuang. Fast vision transformers with hilo attention. *Advances in Neural Information Processing Systems*, 35:14541–14554, 2022. [3](#)
- [62] Sang-Min Park and Young-Gab Kim. A metaverse: Taxonomy, components, applications, and open challenges. *IEEE access*, 10:4209–4251, 2022. [1](#)
- [63] Mathis Petrovich, Michael J Black, and Gül Varol. Temos: Generating diverse human motions from textual descriptions. In *European Conference on Computer Vision*, pages 480–497. Springer, 2022. [2](#), [5](#)
- [64] Matthias Plappert, Christian Mandery, and Tamim Asfour. The kit motion-language dataset. *Big data*, 4(4):236–252, 2016. [2](#)
- [65] Abhinanda R Punnakal, Arjun Chandrasekaran, Nikos Athanasiou, Alejandra Quiros-Ramirez, and Michael J Black. Babel: Bodies, action and behavior with english labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 722–731, 2021. [2](#), [5](#)
- [66] Yijun Qian, Jack Urbanek, Alexander G Hauptmann, and Jungdam Won. Breaking the limits of text-conditioned 3d motion synthesis with elaborative descriptions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2306–2316, 2023. [2](#)
- [67] Jia Qin, Youyi Zheng, and Kun Zhou. Motion in-betweening via two-stage transformers. *ACM Transactions on Graphics (TOG)*, 41(6):1–16, 2022. [2](#)
- [68] 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. [14](#)
- [69] Tianxiang Ren, Jubo Yu, Shihui Guo, Ying Ma, Yutao Ouyang, Zijiao Zeng, Yazhan Zhang, and Yipeng Qin. Diverse motion in-betweening with dual posture stitching. *arXiv preprint arXiv:2303.14457*, 2023. [2](#)
- [70] 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](#)
- [71] Mehdi SM Sajjadi, Olivier Bachem, Mario Lucic, Olivier Bousquet, and Sylvain Gelly. Assessing generative models via precision and recall. *Advances in neural information processing systems*, 31, 2018. [5](#)
- [72] 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. [2](#)
- [73] Yonatan Shafir, Guy Tevet, Roy Kapon, and Amit H Bermano. Human motion diffusion as a generative prior. *arXiv preprint arXiv:2303.01418*, 2023. [2](#), [3](#), [4](#), [5](#), [14](#)
- [74] 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. [2](#)
- [75] Paul Starke, Sebastian Starke, Taku Komura, and Frank Steinicke. Motion in-betweening with phase manifolds. *Proceedings of the ACM on Computer Graphics and Interactive Techniques*, 6(3):1–17, 2023. [2](#)
- [76] 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. [4](#), [14](#)
- [77] Guofei Sun, Yongkang Wong, Zhiyong Cheng, Mohan S Kankanhalli, Weidong Geng, and Xiangdong Li. Deepdance: music-to-dance motion choreography with adversarial learning. *IEEE Transactions on Multimedia*, 23:497–509, 2020. [2](#)
- [78] Jiarui Sun and Girish Chowdhary. Towards globally consistent stochastic human motion prediction via motion dif-



- fusion. *arXiv preprint arXiv:2305.12554*, 2023. 2
- [79] Ryo Suzuki, Adnan Karim, Tian Xia, Hooman Hedayati, and Nicolai Marquardt. Augmented reality and robotics: A survey and taxonomy for ar-enhanced human-robot interaction and robotic interfaces. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–33, 2022. 1
- [80] Julian Tanke, Linguang Zhang, Amy Zhao, Chengcheng Tang, Yujun Cai, Lezi Wang, Po-Chen Wu, Juergen Gall, and Cem Keskin. Social diffusion: Long-term multiple human motion anticipation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9601–9611, 2023. 2
- [81] Guy Tevet, Brian Gordon, Amir Hertz, Amit H Bermano, and Daniel Cohen-Or. Motionclip: Exposing human motion generation to clip space. In *European Conference on Computer Vision*, pages 358–374. Springer, 2022. 2
- [82] 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. 2, 3, 4, 14
- [83] Sibotian, Minghui Zheng, and Xiao Liang. Transfusion: A practical and effective transformer-based diffusion model for 3d human motion prediction. *arXiv preprint arXiv:2307.16106*, 2023. 2
- [84] Jonathan Tseng, Rodrigo Castellon, and Karen Liu. Edge: Editable dance generation from music. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 448–458, 2023. 2
- [85] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 4
- [86] 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
- [87] Jingbo Wang, Yu Rong, Jingyuan Liu, Sijie Yan, Dahua Lin, and Bo Dai. Towards diverse and natural scene-aware 3d human motion synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20460–20469, 2022. 2
- [88] Jiashun Wang, Huazhe Xu, Jingwei Xu, Sifei Liu, and Xiaolong Wang. Synthesizing long-term 3d human motion and interaction in 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9401–9411, 2021.
- [89] Jingbo Wang, Sijie Yan, Bo Dai, and Dahua Lin. Scene-aware generative network for human motion synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12206–12215, 2021. 2
- [90] Zhisheng Xiao, Karsten Kreis, and Arash Vahdat. Tackling the generative learning trilemma with denoising diffusion GANs. In *International Conference on Learning Representations (ICLR)*, 2022. 6, 14
- [91] Jiachen Xu, Min Wang, Jingyu Gong, Wentao Liu, Chen Qian, Yuan Xie, and Lizhuang Ma. Exploring versatile prior for human motion via motion frequency guidance. In *2021 International Conference on 3D Vision (3DV)*, pages 606–616. IEEE, 2021. 2
- [92] Sirui Xu, Zhengyuan Li, Yu-Xiong Wang, and Liang-Yan Gui. Interdiff: Generating 3d human-object interactions with physics-informed diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14928–14940, 2023. 2
- [93] Sirui Xu, Yu-Xiong Wang, and Liangyan Gui. Stochastic multi-person 3d motion forecasting. In *The Eleventh International Conference on Learning Representations*, 2022. 2
- [94] Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications. *ACM Computing Surveys*, 2022. 3, 5
- [95] Zhao Yang, Bing Su, and Ji-Rong Wen. Synthesizing long-term human motions with diffusion models via coherent sampling. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 3954–3964, 2023. 2, 5
- [96] Zijie Ye, Haozhe Wu, Jia Jia, Yaohua Bu, Wei Chen, Fanbo Meng, and Yanfeng Wang. Choreonet: Towards music to dance synthesis with choreographic action unit. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 744–752, 2020. 2
- [97] Hongwei Yi, Chun-Hao P Huang, Shashank Tripathi, Lea Hering, Justus Thies, and Michael J Black. Mime: Human-aware 3d scene generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12965–12976, 2023. 2
- [98] Xinyu Yi, Yuxiao Zhou, and Feng Xu. Transpose: Real-time 3d human translation and pose estimation with six inertial sensors. *ACM Transactions on Graphics (TOG)*, 40(4):1–13, 2021. 5
- [99] Ye Yuan and Kris Kitani. Dlow: Diversifying latent flows for diverse human motion prediction. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, pages 346–364. Springer, 2020. 2
- [100] Ye Yuan, Jiaming Song, Umar Iqbal, Arash Vahdat, and Jan Kautz. Physdiff: Physics-guided human motion diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16010–16021, 2023. 2
- [101] Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Yong Zhang, Hongwei Zhao, Hongtao Lu, Xi Shen, and Ying Shan. Generating human motion from textual descriptions with discrete representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14730–14740, 2023.
- [102] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiandiffuse: Text-driven human motion generation with diffusion model. *arXiv preprint arXiv:2208.15001*, 2022. 2, 4
- [103] Qinsheng Zhang, Jiaming Song, Xun Huang, Yongxin Chen, and Ming-Yu Liu. Diffcollage: Parallel generation of large content with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10188–10198, 2023. 2, 3, 4, 5
- [104] Yan Zhang, Michael J Black, and Siyu Tang. Perpetual motion: Generating unbounded human motion. *arXiv preprint*



- arXiv:2007.13886*, 2020. [2](#)
- [105] Yan Zhang and Siyu Tang. The wanderings of odysseus in 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20481–20491, 2022. [2](#)
  - [106] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5745–5753, 2019. [5](#)
  - [107] Yi Zhou, Zimo Li, Shuangjiu Xiao, Chong He, Zeng Huang, and Hao Li. Auto-conditioned recurrent networks for extended complex human motion synthesis. In *International Conference on Learning Representations*, 2018. [2](#), [3](#)
  - [108] Yi Zhou, Jingwan Lu, Connelly Barnes, Jimei Yang, Sitao Xiang, et al. Generative tweening: Long-term inbetweening of 3d human motions. *arXiv preprint arXiv:2005.08891*, 2020. [2](#)
  - [109] Wentao Zhu, Xiaoxuan Ma, Dongwoo Ro, Hai Ci, Jinlu Zhang, Jiaxin Shi, Feng Gao, Qi Tian, and Yizhou Wang. Human motion generation: A survey. *arXiv preprint arXiv:2307.10894*, 2023. [1](#)
  - [110] Wenlin Zhuang, Congyi Wang, Jinxiang Chai, Yangang Wang, Ming Shao, and Siyu Xia. Music2dance: Dancenet for music-driven dance generation. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 18(2):1–21, 2022. [2](#)

# Supplementary Material

## 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<sup>1</sup> with the originally proposed alignment and spherical linear interpolation, and without them (TEACH.B). DoubleTake is used off-the-shelf<sup>2</sup> 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 [82], 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 [58]. FlowMDM is trained with the  $x_0$  parameterization [90], 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 [68] 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 [76] 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 Trans-

<sup>1</sup><https://github.com/athn-nik/teach/commit/f4285aff0fd556a5b46518a751fc90825d91e68b>

<sup>2</sup><https://github.com/priorMDM/priorMDM/commit/8bc565b3120c08182f067e161e83403b0efe7cc9>

formers. 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 [41] 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.

## 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 [73] 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 impor-

tant 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<sup>3</sup>. 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

<sup>3</sup><https://barqueroerman.github.io/FlowMDM/>

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 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.

### 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.

### 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 [58]. Instead, the cosine schedule destroys the motion signal slower and in a more evenly distributed way. Fig. A shows the performance of FlowMDM during 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

	Subsequence				Transition			
	R-prec $\uparrow$	FID $\downarrow$	Div $\rightarrow$	MM-Dist $\downarrow$	FID $\downarrow$	Div $\rightarrow$	PJ $\rightarrow$	AUJ $\downarrow$
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
In-distribution								
TEACH_B	<b>0.727</b> $\pm$ 0.004	2.26 $\pm$ 0.03	8.20 $\pm$ 0.12	<b>3.35</b> $\pm$ 0.01	2.77 $\pm$ 0.05	6.32 $\pm$ 0.07	1.03 $\pm$ 0.00	2.20 $\pm$ 0.01
TEACH	0.665 $\pm$ 0.003	2.09 $\pm$ 0.03	8.06 $\pm$ 0.09	3.73 $\pm$ 0.02	2.78 $\pm$ 0.06	6.31 $\pm$ 0.07	<b>0.07</b> $\pm$ 0.00	<b>0.42</b> $\pm$ 0.01
DoubleTake*	0.620 $\pm$ 0.006	3.04 $\pm$ 0.06	7.49 $\pm$ 0.07	4.19 $\pm$ 0.02	3.04 $\pm$ 0.12	6.21 $\pm$ 0.06	0.28 $\pm$ 0.00	1.01 $\pm$ 0.01
DoubleTake	0.682 $\pm$ 0.008	<u>1.52</u> $\pm$ 0.03	7.90 $\pm$ 0.07	3.67 $\pm$ 0.04	3.47 $\pm$ 0.08	6.16 $\pm$ 0.07	0.17 $\pm$ 0.00	0.62 $\pm$ 0.01
MultiDiffusion	0.724 $\pm$ 0.008	2.00 $\pm$ 0.05	<u>8.36</u> $\pm$ 0.10	<u>3.38</u> $\pm$ 0.02	6.33 $\pm$ 0.13	5.91 $\pm$ 0.06	0.17 $\pm$ 0.00	0.65 $\pm$ 0.01
DiffCollage	0.690 $\pm$ 0.006	1.92 $\pm$ 0.07	7.92 $\pm$ 0.09	3.67 $\pm$ 0.02	4.25 $\pm$ 0.15	<b>6.19</b> $\pm$ 0.07	0.19 $\pm$ 0.01	0.82 $\pm$ 0.02
FlowMDM (Ours)	<u>0.726</u> $\pm$ 0.006	<b>1.36</b> $\pm$ 0.05	<b>8.47</b> $\pm$ 0.10	3.40 $\pm$ 0.03	<b>2.26</b> $\pm$ 0.08	6.60 $\pm$ 0.08	<b>0.05</b> $\pm$ 0.00	<b>0.11</b> $\pm$ 0.00
Out-of-distribution								
TEACH_B	<u>0.680</u> $\pm$ 0.006	1.75 $\pm$ 0.04	8.15 $\pm$ 0.11	3.51 $\pm$ 0.01	3.53 $\pm$ 0.06	6.04 $\pm$ 0.10	1.14 $\pm$ 0.01	2.49 $\pm$ 0.01
TEACH	0.644 $\pm$ 0.004	2.06 $\pm$ 0.03	7.94 $\pm$ 0.12	3.70 $\pm$ 0.01	4.08 $\pm$ 0.08	6.00 $\pm$ 0.09	<b>0.07</b> $\pm$ 0.00	<u>0.46</u> $\pm$ 0.00
DoubleTake*	0.572 $\pm$ 0.007	3.78 $\pm$ 0.07	7.53 $\pm$ 0.12	4.15 $\pm$ 0.02	3.83 $\pm$ 0.09	<b>6.12</b> $\pm$ 0.07	0.28 $\pm$ 0.00	1.07 $\pm$ 0.02
DoubleTake	0.654 $\pm$ 0.009	1.65 $\pm$ 0.07	8.06 $\pm$ 0.08	3.66 $\pm$ 0.02	<b>2.98</b> $\pm$ 0.06	6.03 $\pm$ 0.07	0.17 $\pm$ 0.00	0.66 $\pm$ 0.01
MultiDiffusion	<b>0.681</b> $\pm$ 0.009	2.11 $\pm$ 0.06	<b>8.35</b> $\pm$ 0.08	<b>3.47</b> $\pm$ 0.03	6.97 $\pm$ 0.12	5.67 $\pm$ 0.05	0.19 $\pm$ 0.00	0.71 $\pm$ 0.01
DiffCollage	0.652 $\pm$ 0.004	<u>1.60</u> $\pm$ 0.07	7.91 $\pm$ 0.09	3.74 $\pm$ 0.01	4.65 $\pm$ 0.19	6.00 $\pm$ 0.09	0.20 $\pm$ 0.00	0.86 $\pm$ 0.01
FlowMDM (Ours)	0.679 $\pm$ 0.004	<b>1.26</b> $\pm$ 0.06	<u>8.16</u> $\pm$ 0.08	3.50 $\pm$ 0.03	3.17 $\pm$ 0.12	6.44 $\pm$ 0.09	<b>0.07</b> $\pm$ 0.00	<b>0.17</b> $\pm$ 0.00

Table B. Scenario-wise comparison in Babel. Symbols  $\uparrow$ ,  $\downarrow$ , and  $\rightarrow$  indicate that higher, lower, or values closer to the ground truth (GT) are better, respectively. Evaluation is run 10 times and  $\pm$  specifies the 95% confidence intervals.

	Subsequence				Transition			
	R-prec $\uparrow$	FID $\downarrow$	Div $\rightarrow$	MM-Dist $\downarrow$	FID $\downarrow$	Div $\rightarrow$	PJ $\rightarrow$	AUJ $\downarrow$
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
Short								
DoubleTake*	0.649 $\pm$ 0.012	<b>3.03</b> $\pm$ 0.18	<b>9.52</b> $\pm$ 0.11	3.72 $\pm$ 0.05	3.56 $\pm$ 0.14	8.92 $\pm$ 0.14	0.13 $\pm$ 0.01	0.79 $\pm$ 0.05
DoubleTake	0.704 $\pm$ 0.022	4.85 $\pm$ 0.20	10.01 $\pm$ 0.15	3.25 $\pm$ 0.09	4.40 $\pm$ 0.24	8.88 $\pm$ 0.17	<u>0.09</u> $\pm$ 0.00	<u>0.73</u> $\pm$ 0.02
MultiDiffusion	<b>0.717</b> $\pm$ 0.011	5.49 $\pm$ 0.15	10.14 $\pm$ 0.17	<b>3.23</b> $\pm$ 0.07	4.66 $\pm$ 0.27	8.68 $\pm$ 0.08	0.10 $\pm$ 0.00	0.92 $\pm$ 0.02
DiffCollage	0.705 $\pm$ 0.012	<u>4.69</u> $\pm$ 0.18	<u>9.73</u> $\pm$ 0.14	3.30 $\pm$ 0.04	4.81 $\pm$ 0.32	8.49 $\pm$ 0.12	0.15 $\pm$ 0.01	1.13 $\pm$ 0.10
FlowMDM (Ours)	<u>0.714</u> $\pm$ 0.015	4.75 $\pm$ 0.26	9.90 $\pm$ 0.20	3.31 $\pm$ 0.06	<b>3.17</b> $\pm$ 0.17	<b>9.03</b> $\pm$ 0.14	<b>0.04</b> $\pm$ 0.00	<b>0.59</b> $\pm$ 0.04
Medium								
DoubleTake*	0.644 $\pm$ 0.009	<u>2.18</u> $\pm$ 0.08	<u>9.18</u> $\pm$ 0.12	3.72 $\pm$ 0.04	<b>3.34</b> $\pm$ 0.30	<b>8.73</b> $\pm$ 0.12	0.14 $\pm$ 0.00	0.70 $\pm$ 0.03
DoubleTake	0.642 $\pm$ 0.014	2.34 $\pm$ 0.05	9.59 $\pm$ 0.09	3.79 $\pm$ 0.05	5.42 $\pm$ 0.30	8.61 $\pm$ 0.11	0.12 $\pm$ 0.00	0.83 $\pm$ 0.02
MultiDiffusion	<b>0.673</b> $\pm$ 0.007	3.22 $\pm$ 0.10	9.91 $\pm$ 0.07	<b>3.54</b> $\pm$ 0.04	6.24 $\pm$ 0.34	8.11 $\pm$ 0.12	<u>0.10</u> $\pm$ 0.00	1.14 $\pm$ 0.01
DiffCollage	0.661 $\pm$ 0.010	<b>2.03</b> $\pm$ 0.07	<b>9.38</b> $\pm$ 0.10	3.60 $\pm$ 0.04	4.95 $\pm$ 0.27	8.13 $\pm$ 0.09	0.14 $\pm$ 0.00	<b>0.66</b> $\pm$ 0.05
FlowMDM (Ours)	<u>0.669</u> $\pm$ 0.012	3.18 $\pm$ 0.15	9.68 $\pm$ 0.08	<u>3.55</u> $\pm$ 0.04	<u>4.18</u> $\pm$ 0.43	8.52 $\pm$ 0.07	<b>0.04</b> $\pm$ 0.00	0.86 $\pm$ 0.03
Long								
DoubleTake*	0.616 $\pm$ 0.006	<u>2.51</u> $\pm$ 0.09	<u>8.77</u> $\pm$ 0.08	4.09 $\pm$ 0.03	3.38 $\pm$ 0.18	8.50 $\pm$ 0.11	0.89 $\pm$ 0.02	3.52 $\pm$ 0.07
DoubleTake	0.605 $\pm$ 0.006	4.07 $\pm$ 0.13	8.19 $\pm$ 0.11	4.18 $\pm$ 0.01	8.45 $\pm$ 0.33	7.79 $\pm$ 0.12	0.81 $\pm$ 0.02	3.04 $\pm$ 0.07
MultiDiffusion	0.569 $\pm$ 0.012	5.02 $\pm$ 0.15	8.07 $\pm$ 0.07	4.49 $\pm$ 0.05	8.56 $\pm$ 0.32	7.91 $\pm$ 0.10	<u>0.23</u> $\pm$ 0.01	<u>1.16</u> $\pm$ 0.01
DiffCollage	0.557 $\pm$ 0.008	5.79 $\pm$ 0.13	7.75 $\pm$ 0.09	4.61 $\pm$ 0.02	9.00 $\pm$ 0.36	7.75 $\pm$ 0.09	0.38 $\pm$ 0.01	5.04 $\pm$ 0.14
FlowMDM (Ours)	<b>0.666</b> $\pm$ 0.012	<b>1.93</b> $\pm$ 0.08	<b>8.81</b> $\pm$ 0.09	<b>3.81</b> $\pm$ 0.04	<b>2.85</b> $\pm$ 0.22	<b>8.54</b> $\pm$ 0.11	<b>0.08</b> $\pm$ 0.00	<b>0.45</b> $\pm$ 0.03
All								
DoubleTake*	0.655 $\pm$ 0.007	<u>0.84</u> $\pm$ 0.04	<b>9.29</b> $\pm$ 0.10	3.92 $\pm$ 0.03	1.91 $\pm$ 0.12	<b>8.79</b> $\pm$ 0.11	0.51 $\pm$ 0.01	2.11 $\pm$ 0.05
DoubleTake	0.621 $\pm$ 0.006	1.49 $\pm$ 0.07	8.91 $\pm$ 0.04	4.13 $\pm$ 0.02	4.75 $\pm$ 0.13	8.39 $\pm$ 0.06	0.47 $\pm$ 0.01	1.84 $\pm$ 0.03
MultiDiffusion	0.632 $\pm$ 0.003	1.17 $\pm$ 0.04	<b>9.29</b> $\pm$ 0.09	4.05 $\pm$ 0.02	4.42 $\pm$ 0.16	8.37 $\pm$ 0.08	<u>0.17</u> $\pm$ 0.00	<u>1.06</u> $\pm$ 0.01
DiffCollage	0.615 $\pm$ 0.007	1.73 $\pm$ 0.07	8.73 $\pm$ 0.05	4.18 $\pm$ 0.04	4.98 $\pm$ 0.24	8.09 $\pm$ 0.06	0.26 $\pm$ 0.00	2.71 $\pm$ 0.12
FlowMDM	<b>0.695</b> $\pm$ 0.008	<b>0.30</b> $\pm$ 0.02	9.55 $\pm$ 0.08	<b>3.58</b> $\pm$ 0.02	<b>1.49</b> $\pm$ 0.06	<u>8.78</u> $\pm$ 0.11	<b>0.06</b> $\pm$ 0.00	<b>0.50</b> $\pm$ 0.01

Table C. Scenario-wise comparison in HumanML3D.

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 [33]. It is implemented by first computing the con-

ditionally 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



H (frames)	Inf. PE	R-prec $\uparrow$	Subsequence			Transition			
			FID $\downarrow$	Div $\rightarrow$	MM-Dist $\downarrow$	FID $\downarrow$	Div $\rightarrow$	PJ $\rightarrow$	AUJ $\downarrow$
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	2.04 $\pm$ 0.06	6.30 $\pm$ 0.05	0.04 $\pm$ 0.00	0.15 $\pm$ 0.00
100	R	0.635 $\pm$ 0.004	<b>0.85</b> $\pm$ 0.02	8.25 $\pm$ 0.12	3.98 $\pm$ 0.02	2.14 $\pm$ 0.04	6.44 $\pm$ 0.09	0.04 $\pm$ 0.00	0.15 $\pm$ 0.00
150	R	0.641 $\pm$ 0.005	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	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	<b>6.16</b> $\pm$ 0.05	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	6.24 $\pm$ 0.07	0.06 $\pm$ 0.00	<b>0.13</b> $\pm$ 0.00
100	B	0.702 $\pm$ 0.004	0.99 $\pm$ 0.04	<b>8.36</b> $\pm$ 0.13	3.45 $\pm$ 0.02	2.61 $\pm$ 0.06	6.47 $\pm$ 0.05	0.06 $\pm$ 0.00	<b>0.13</b> $\pm$ 0.00
150	B	<b>0.704</b> $\pm$ 0.004	1.24 $\pm$ 0.03	8.34 $\pm$ 0.12	3.43 $\pm$ 0.02	2.54 $\pm$ 0.08	6.40 $\pm$ 0.08	0.06 $\pm$ 0.00	<b>0.13</b> $\pm$ 0.00
200	B	0.694 $\pm$ 0.006	1.13 $\pm$ 0.02	8.25 $\pm$ 0.13	<b>3.42</b> $\pm$ 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  $\uparrow$ ,  $\downarrow$ , and  $\rightarrow$  indicate that higher, lower, or values closer to the ground truth (GT) are better, respectively. Evaluation is run 10 times and  $\pm$  specifies the 95% confidence intervals.

H (frames)	Inf. PE	R-prec $\uparrow$	Subsequence			Transition			
			FID $\downarrow$	Div $\rightarrow$	MM-Dist $\downarrow$	FID $\downarrow$	Div $\rightarrow$	PJ $\rightarrow$	AUJ $\downarrow$
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	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	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	<b>1.12</b> $\pm$ 0.04	<b>9.00</b> $\pm$ 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	8.97 $\pm$ 0.09	<b>0.03</b> $\pm$ 0.00	0.97 $\pm$ 0.01
50	B	0.671 $\pm$ 0.004	<b>0.25</b> $\pm$ 0.01	<b>9.37</b> $\pm$ 0.14	3.66 $\pm$ 0.02	1.27 $\pm$ 0.04	8.79 $\pm$ 0.08	0.06 $\pm$ 0.00	0.52 $\pm$ 0.01
100	B	0.684 $\pm$ 0.003	0.36 $\pm$ 0.02	9.55 $\pm$ 0.09	<b>3.61</b> $\pm$ 0.02	2.04 $\pm$ 0.11	8.59 $\pm$ 0.06	0.06 $\pm$ 0.00	0.56 $\pm$ 0.01
150	B	<b>0.685</b> $\pm$ 0.004	0.29 $\pm$ 0.01	9.58 $\pm$ 0.12	<b>3.61</b> $\pm$ 0.01	1.38 $\pm$ 0.05	8.79 $\pm$ 0.09	0.06 $\pm$ 0.00	<b>0.51</b> $\pm$ 0.01
200	B	0.658 $\pm$ 0.006	0.47 $\pm$ 0.03	<b>9.37</b> $\pm$ 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.

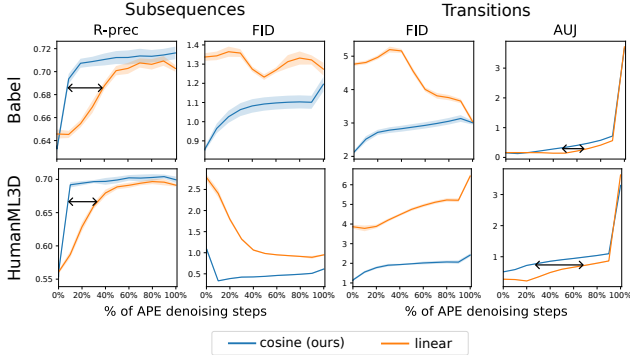


Figure A. **Diffusion noise schedules.** The cosine noise schedule destroys the motion signal slower and in a more evenly distributed way than the linear schedule. As a result, FlowMDM is able to exploit better the low-to-high frequencies decomposition along the denoising chain and generate better subsequences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

compositions with our method, two different conditions co-exist 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

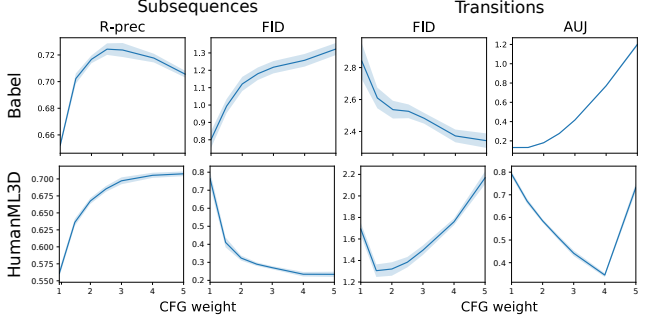


Figure B. **Classifier-free guidance.** In line with prior works, we also observe an accuracy improvement (R-prec) when increasing the strength (i.e., *weight*) of the classifier-free guidance (CFG). However, above certain values, the performance degrades, especially in terms of smoothness (AUJ). This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

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.

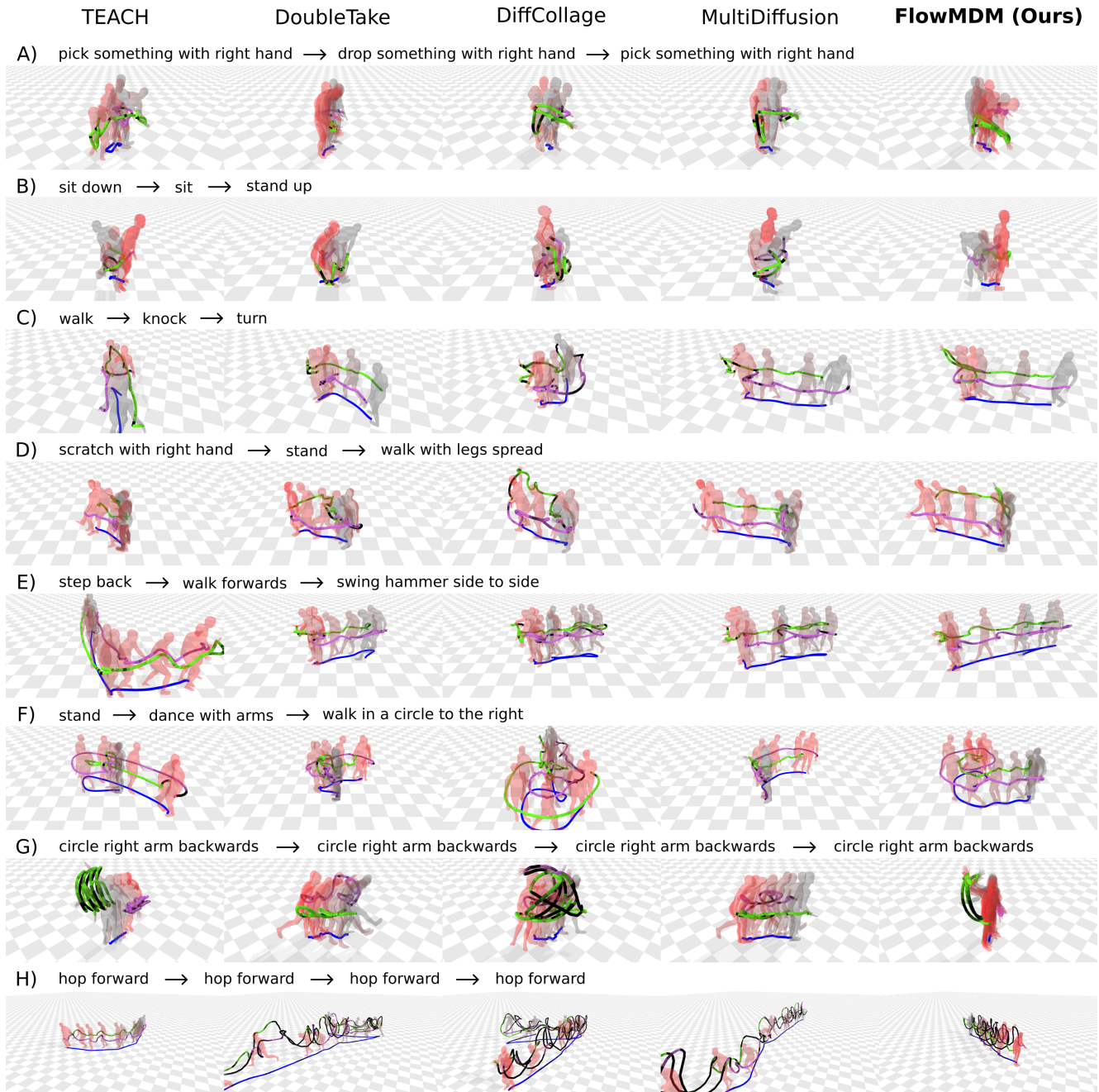


Figure C. **Qualitative examples (Babel)**. A-F feature six human motion compositions, and G-H two human motion extrapolations. According to the scenarios defined in [Sec. B](#), A, B, C belong to in-distribution combinations, and D, E, F to out-of-distribution combinations. Videos of all samples are also included as part of this supplementary material. 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.

## 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. Motion videos are also included as part of the supplementary material. Note that we can represent the motions from Babel with SMPL body meshes thanks to its motion representation including the SMPL param-

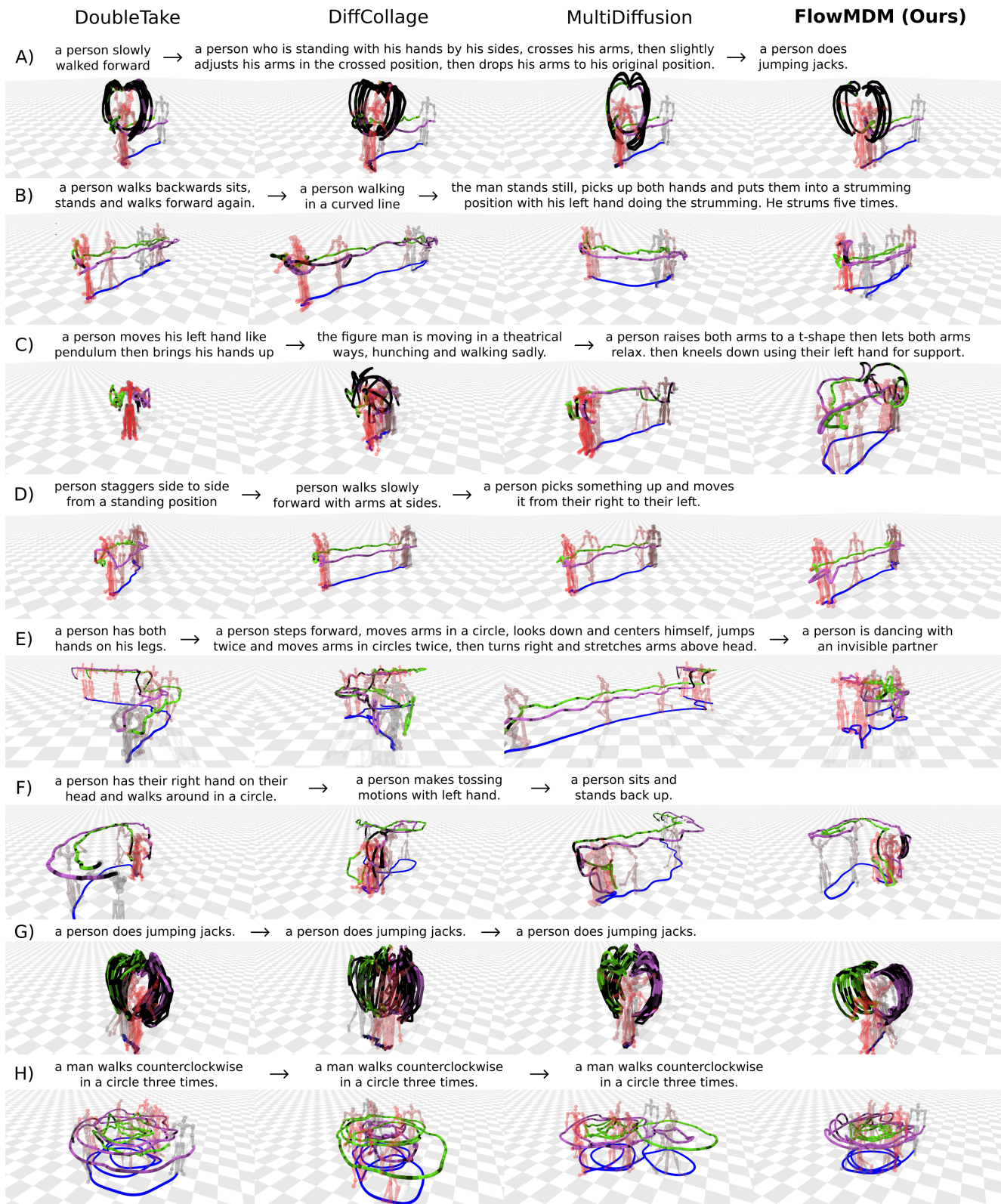


Figure D. **Qualitative examples (HumanML3D)**. A-F feature six human motion compositions, and G-H two human motion extrapolations. According to the scenarios defined in [Sec. B](#), A, B, C are samples from the short, medium, and long scenarios, respectively, and D, E, F from the mixed scenario. Videos of all samples are also included as part of this supplementary material.



ters [13]. 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 and the videos 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 generation 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.