InfiniDreamer: Arbitrarily Long Human Motion Generation via Segment Score Distillation

Supplementary Material

8. Theoretical Analysis: Segment Score Distillation and Global Consistency

We provide a theoretical analysis to demonstrate that if the Segment Score Distillation loss $\mathcal{L}_{\text{align}} = \mathbb{E}_{t,\epsilon}\big[w(t)\|\hat{x}_0^i - x_0^i\|_2^2\big]$ converges, the resulting long motion sequence M is guaranteed to be globally coherent and smooth.

8.1. Problem Setup

Let $M = \{m_1, t_1, m_2, t_2, ..., m_N\}$ denote a long motion sequence M, where m_i represents the motion segment corresponding to the i-th text prompt, and t_i represents the transition segment between m_i and m_{i+1}. The initial sequence M is constructed by concatenating motion segments $\{m_i\}$ with randomly initialized transition $\{t_i\}$. Our goal is to iteratively optimize M so that:
(i) Each motion segment m_i and transition t_i conforms to a learned motion prior $p(x)$, ensuring realism.
(ii) M achieves global coherence and smoothness.

To optimize M, we introduce Segment Score Distillation (SSD), which operates as follows:
(i) Using a sliding window, we sample overlapping short sequences $\{x_0^i\}_{i=1}^K$ from M, where each $x_0^i = M[i : i + L]$ spans motion segments $(m_i, m_{i+1}$ and transitions (t_i).
(ii) Add noise to each sampled sequence to obtain $x_t^i \sim q(x_t^i|x_0^i)$.
(iii) Denoise x_t^i using the Motion Diffusion Model ϕ to predict $\hat{x}_0^i$, and compute the alignment loss:

$$\mathcal{L}_{align} = \mathbf{E}_{t,\epsilon}[w(t)||\hat{x}_0^i - x_0^i||_2^2], \tag{9}$$

(iv) The loss $\mathcal{L}_{align}$ is back-propagated to optimize M, ensuring that both m_i and t_i align with $p(x)$

8.2. Theoretical Analysis

We now prove that minimizing $\mathcal{L}_{align}$ ensures global coherence and smoothness in M.

8.2.1. Local Consistency Through Loss Convergence

The Motion Diffusion Model ϕ is trained to model the conditional distribution of motion x_0^i at time step t:

$$\hat{x}_0^i = \phi(x_t^i, t), \;\; x_t^i \sim q(x_t^i|x_0^i), \tag{10}$$

where $q(x_t^i|x_0^i)$ represents the forward diffusion process.

When $\mathcal{L}_{align} \to 0$, the predicted $\hat{x}_0^i$ aligns with the x_0^i for all sampled segments. This implies:
(i) Each segment x_0^i conforms to the motion prior $p(x)$, which encodes realistic and smooth dynamics.
(ii) The transitions within x_0^i are temporally consistent.

8.2.2. Global Coherence via Overlapping Optimization

The sliding window mechanism ensures that adjacent sampled sequences overlap. Let:

$$x_0^i = M[i : i + L], \;\; x_0^{i+1} = M[i + \delta : i + L + \delta], \tag{11}$$

where L is the segment length, and δ is the overlap step size. The overlapping region $O^i = x_0^i \bigcap x_0^{i+1}$ satisfies:

$$\hat{x}_0^i[O^i] \to x_0^i[O^i], \;\; \hat{x}_0^{i+1}[O^i] \to x_0^{i+1}[O^i], \tag{12}$$

by transitivity:

$$||\hat{x}_0^i[O^i] - \hat{x}_0^{i+1}[O^i]||_2^2 \to 0, \tag{13}$$

as $\mathcal{L}_{align} \to 0$. By enforcing consistency within overlapping regions, the optimization propagates coherence across the entire sequence M.

8.2.3. Smoothness Guarantee

Smoothness in M is ensured through two mechanisms:

(1) **Local Smoothness**: The loss $\mathcal{L}_{align}$ minimizes the discrepancy between $\hat{x}_0^i$ and x_0^i, ensuring smooth dynamics within each short segment.

(2) **Global Smoothness**: Overlapping regions O^i propagate smoothness across segment boundaries.

By the end of optimization, all segments and transitions are aligned with $p(x)$, resulting in a globally smooth sequence M.

9. Further implementation details

To facilitate better reproducibility of our work, we provide additional details about our implementation in this section. For HumanML3D, we set the fps as 20, and encode timesteps as a sinusoidal positional encoding. We utilize a dense layer to encode poses of 263D into a sequence of 512D vectors. For BABEL, we set fps as 30. We encode poses of 135D into a sequence of 512D vectors. In the first stage, we utilize guidance scale 2.5 to generate each single motion segments, and in the process of Segment Score Distillation, we utilize 7.5 to optimize the entire motion sequence. We set the weighting function $w(t)$ as $1 - \alpha(t)$ for all experiments.

10. More experimental results

In this section, we present more experimental results to validate the effectiveness of our framework.

	Motion				Transition			
	R-prec ↑	FID ↓	Div →	MM-Dist ↓	FID ↓	Div →	PJ →	AUJ ↓
Ground Truth	0.796 ± 0.004	0.00 ± 0.00	9.34 ± 0.08	2.97 ± 0.01	0.00 ± 0.00	9.54 ± 0.15	0.02 ± 0.00	0.00 ± 0.00
DoubleTake*	$0.643^{\pm 0.005}$	$0.80^{\pm 0.02}$	$9.20^{\pm 0.11}$	$3.92^{\pm 0.01}$	$1.71^{\pm 0.05}$	$\underline{8.82}^{\pm 0.13}$	$0.52^{\pm 0.01}$	$2.10^{\pm 0.03}$
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	$0.629^{\pm 0.002}$	$1.19^{\pm 0.03}$	$\mathbf{9.38}^{\pm 0.08}$	$4.02^{\pm 0.01}$	$4.31^{\pm 0.06}$	$8.37^{\pm 0.10}$	$\underline{0.17}^{\pm 0.00}$	$\underline{1.06}^{\pm 0.01}$
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	$\underline{0.685}^{\pm 0.004}$	$\mathbf{0.29}^{\pm 0.01}$	$9.58^{\pm 0.12}$	$3.61^{\pm 0.01}$	$\mathbf{1.38}^{\pm 0.05}$	$8.79^{\pm 0.09}$	$\mathbf{0.06}^{\pm 0.00}$	$\mathbf{0.51}^{\pm 0.01}$
InfiniDreamer (ours)	$0.679^{\pm 0.007}$	$\underline{0.47}^{\pm 0.12}$	$9.58^{\pm 0.15}$	$\mathbf{3.15}^{\pm 0.01}$	$2.04^{\pm 0.05}$	$8.69^{\pm 0.11}$	-	-
+ FlowMDM [3]	$0.674^{\pm 0.004}$	$0.68^{\pm 0.02}$	$9.27^{\pm 0.11}$	$3.78^{\pm 0.01}$	$1.64^{\pm 0.05}$	$8.77^{\pm 0.11}$	-	-
+ MLD [6]	$\mathbf{0.713}^{\pm 0.003}$	$0.52^{\pm 0.15}$	$\underline{9.46}^{\pm 0.11}$	$\underline{3.17}^{\pm 0.01}$	$\underline{1.47}^{\pm 0.05}$	$\mathbf{8.87}^{\pm 0.11}$	-	-

Table 4. Comparison of InfiniDreamer with the state of the art in HumanML3D. Symbols ↑, ↓, and → mean that higher, lower, or closer to the ground truth (GT) value are better, respectively. We run each evaluation 10 times to obtain the final results. We use **Bold** to indicate the best result, and use underline to indicate the second-best result.

	Motion				Transition			
	R-prec ↑	FID ↓	Div →	MM-Dist ↓	FID ↓	Div →	PJ →	AUJ ↓
Ground Truth	$0.715^{\pm 0.003}$	$0.00^{\pm 0.00}$	$8.42^{\pm 0.15}$	$3.36^{\pm 0.00}$	$0.00^{\pm 0.00}$	$6.20^{\pm 0.06}$	$0.02^{\pm 0.00}$	$0.00^{\pm 0.00}$
TEACH_B	$\mathbf{0.703}^{\pm 0.002}$	$1.71^{\pm 0.03}$	$8.18^{\pm 0.14}$	$\mathbf{3.43}^{\pm 0.01}$	$3.01^{\pm 0.04}$	$\mathbf{6.23}^{\pm 0.05}$	$1.09^{\pm 0.00}$	$2.35^{\pm 0.01}$
TEACH	$0.655^{\pm 0.002}$	$1.82^{\pm 0.02}$	$7.96^{\pm 0.11}$	$3.72^{\pm 0.01}$	$3.27^{\pm 0.04}$	$6.14^{\pm 0.06}$	$\underline{0.07}^{\pm 0.00}$	$\underline{0.44}^{\pm 0.00}$
DoubleTake*	$0.596^{\pm 0.005}$	$3.16^{\pm 0.06}$	$7.53^{\pm 0.11}$	$4.17^{\pm 0.02}$	$3.33^{\pm 0.06}$	$\underline{6.16}^{\pm 0.05}$	$0.28^{\pm 0.00}$	$1.04^{\pm 0.01}$
DoubleTake	$0.668^{\pm 0.005}$	$1.33^{\pm 0.04}$	$7.98^{\pm 0.12}$	$3.67^{\pm 0.03}$	$3.15^{\pm 0.05}$	$6.14^{\pm 0.07}$	$0.17^{\pm 0.00}$	$0.64^{\pm 0.01}$
MultiDiffusion	$\underline{0.702}^{\pm 0.005}$	$1.74^{\pm 0.04}$	$\mathbf{8.37}^{\pm 0.13}$	$\mathbf{3.43}^{\pm 0.02}$	$6.56^{\pm 0.12}$	$5.72^{\pm 0.07}$	$0.18^{\pm 0.00}$	$0.68^{\pm 0.00}$
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	$\underline{0.702}^{\pm 0.004}$	$\underline{0.99}^{\pm 0.04}$	$\underline{8.36}^{\pm 0.13}$	$3.45^{\pm 0.02}$	$\underline{2.61}^{\pm 0.06}$	$6.47^{\pm 0.05}$	$\mathbf{0.06}^{\pm 0.00}$	$\mathbf{0.13}^{\pm 0.00}$
InfiniDreamer (ours)	$0.543^{\pm 0.009}$	$\mathbf{0.97}^{\pm 0.09}$	$8.31^{\pm 0.06}$	$5.80^{\pm 0.01}$	$\mathbf{2.07}^{\pm 0.30}$	$7.95^{\pm 0.07}$	-	-
+ FlowMDM [3]	$0.667^{\pm 0.005}$	$1.49^{\pm 0.03}$	$7.94^{\pm 0.14}$	$3.53^{\pm 0.01}$	$2.97^{\pm 0.08}$	$6.31^{\pm 0.07}$	-	-

Table 5. Comparison of InfiniDreamer with the state of the art in BABEL. Symbols ↑, ↓, and → mean that higher, lower, or closer to the ground truth (GT) value are better, respectively. We run each evaluation 10 times to obtain the final results. We use **Bold** to indicate the best result, and use underline to indicate the second-best result.

10.1. More quantitative results

We also compare our framework with the latest work, FlowMDM [3], which introduces Blended Positional Encodings, a technique that combines absolute and relative positional encodings in the denoising process. We follow the evaluation protocol of FlowMDM [3]. However, we only use FlowMDM [3] to generate individual motion segments, and then use our method to generate the entire long motion sequence. As shown in Tab. 4 and Tab. 5, we find that InfiniDreamer performs slightly worse than FlowMDM [3] but outperforms previous training-free methods. We speculate that this is because FlowMDM [3] is fine-tuned on long human motion sequences using both absolute and relative positional encodings, which introduces some interference in individual short motion segments. Our method, which adds further interference, therefore achieves slightly lower performance compared to FlowMDM [3]. Nonetheless, the experimental results demonstrate the advantages of our approach over DoubleTake [50].

10.2. More qualitative results

We also conducted qualitative experiments to compare the results of our framework with those of FlowMDM [3]. In this section, we use the open-source model of FlowMDM [3] to generate its results, while for our method, we use MDM [62] as our motion prior. As shown in Fig. 5, we present two examples of long motion sequence generation. The first example, at the top of the figure, is generated using the text prompts 'jogging forward slowly' + 'a person is walking down the stairs' + 'jogging forward slowly'. The results show that both FlowMDM and our method can infer

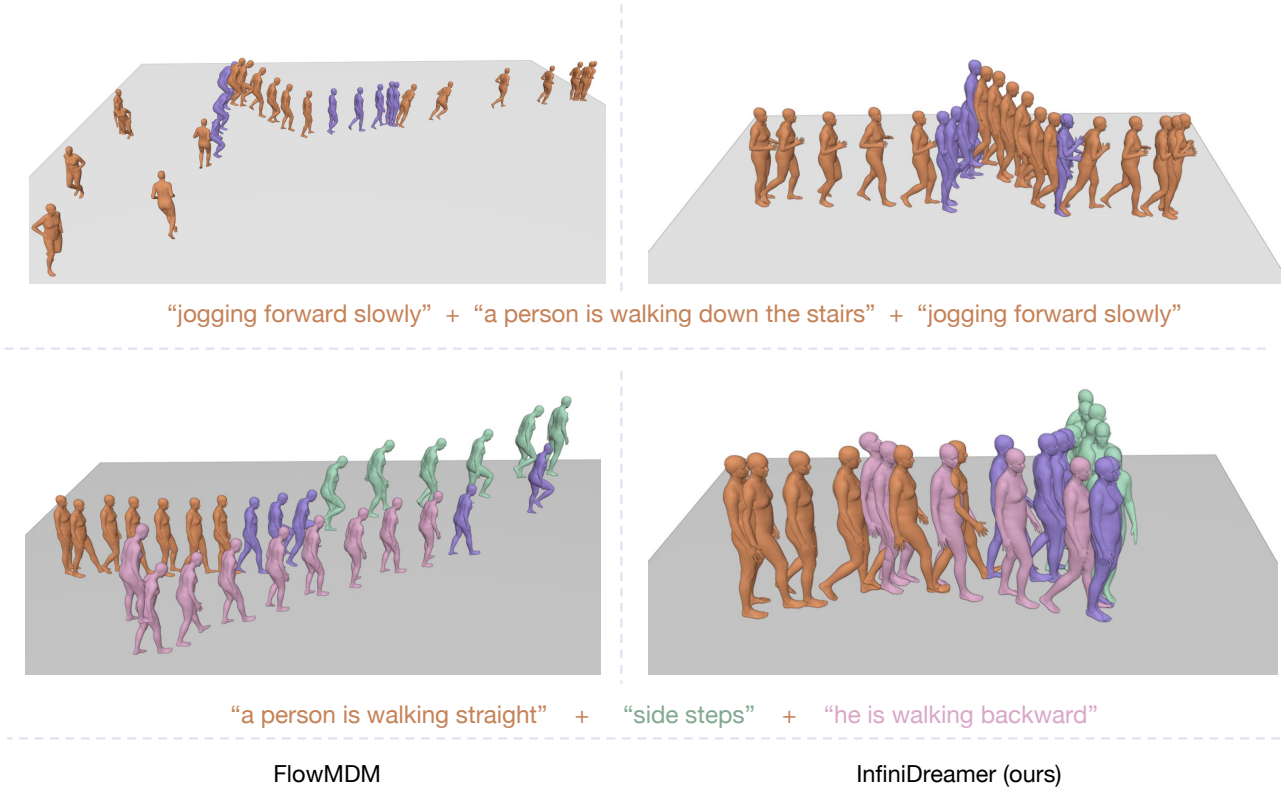
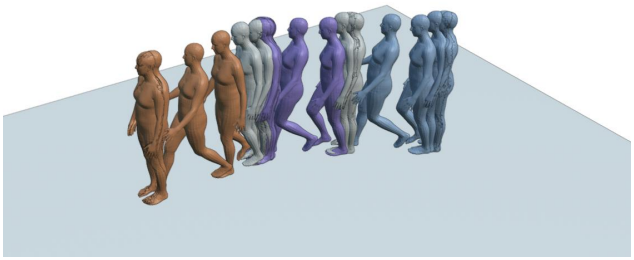


Figure 5. Qualitative Comparisons to FlowMDM for Long Motion Generation. We present two examples: in the top row, our framework demonstrates strong contextual understanding, guiding the transition segment to "go upstairs" in response to the following "downstairs" prompt. In contrast, FlowMDM shows slightly motion drift in this segment. In the bottom row, we use a more fine-grained textual prompt, where the FlowMDM exhibits issues with motion drift and semantic errors, failing to generate the "side steps" segment. Our framework, however, produces a higher-quality sequence with enhanced fine-grained comprehension of the text.

	Motion				Transition (30 frames)	
	R-precision ↑	FID ↓	Diversity →	MultiModal-Dist ↓	FID ↓	Diversity →
Ground Truth	0.797 ± 0.003	$1.6 \cdot 10^{-3} \pm 0.00$	9.59 ± 0.13	2.98 ± 0.01	$1.8 \cdot 10^{-3} \pm 0.00$	9.55 ± 0.09
InfiniDreamer (ours)	$\mathbf{0.679 \pm 0.007}$	$\mathbf{0.47 \pm 0.12}$	$\mathbf{9.58 \pm 0.15}$	$\mathbf{3.15 \pm 0.01}$	$\mathbf{2.04 \pm 0.28}$	$\mathbf{8.69 \pm 0.09}$
w/o text selection strategy:						
+ "transition"	0.657 ± 0.011	0.53 ± 0.13	9.51 ± 0.13	3.32 ± 0.01	2.32 ± 0.32	8.43 ± 0.09
+ "motion"	0.650 ± 0.013	0.56 ± 0.14	9.54 ± 0.13	3.39 ± 0.01	2.36 ± 0.33	8.46 ± 0.10
+ uncondition	$\underline{0.664 \pm 0.010}$	$\underline{0.49 \pm 0.09}$	$\underline{9.55 \pm 0.13}$	$\underline{3.23 \pm 0.01}$	$\underline{2.15 \pm 0.28}$	$\underline{8.57 \pm 0.11}$

Table 6. Ablation Study on textual prompt. We remove the original text selection strategy and instead optimize using a single text prompt. We present two types of prompt: "transition" and "motion", as well as an unconditional optimization scenario. We find that mismatched textual conditions lead to a decline in performance, while the unconditional setting produces an sub-optimal result. We believe this is because the text prompts used are not well-suited to capture the semantics of diverse transition segments. This validate the effectiveness of our text condition selection strategy.

the transitional 'walking up the stairs' segment before descending. However, FlowMDM exhibits slight motion drift during this segment. In the second example, with the text prompts 'a person is walking straight' + 'side steps' + 'he is walking backward', we observe that FlowMDM generates the 'side steps' motion incorrectly and also shows motion drift at the end. In contrast, our method avoids these issues, producing more accurate and coherent results. Additionally, we observe that the motions generated by FlowMDM exhibit a larger displacement range, while our method produces smoother and more controlled movements.



"A person takes 3 steps backward" (520 frames)

Figure 6. We demonstrate InfiniDreamer's capability to generate long motion sequence from a single text prompt. Starting with a base model designed to generate short sequences (approximately 70 to 200 frames), our framework extends its generation range to 520 frames while ensuring that the generated motions remain semantically consistent with the input text.

10.3. More ablation study on prompts

In this section, we present additional ablation studies. We explore the use of different text conditions to guide the optimization of Segment Score Distillation (SSD). In this experiments, we remove the original text selection strategy and instead optimize using a single text prompt. We present two types of prompts: "transition" and "motion ", and set the guidance scale as 7.5. We also present an unconditional optimization scenario. The results are shown in Tab. 6, we observe that incorporating suboptimial prompts negatively impacts InfiniDreamer's performance. Using the unconditional optimization results in a slight performance decline, whereas using "transition" or "motion" as prompts leads to a more significant degradation in performance. We believe this is because the chosen text prompts are struggle to capture the semantic diversity of various transition segments.

10.4. Long motion generation with Single Prompt

Our framework has an additional capability: it can generate long motion sequences from a single text prompt. It is a feature that currently beyond the reach of other models. Specifically, given a short-sequence generation model, Motion Diffusion Model (MDM) [62], we set the total frame count of the long sequence to 520 frames and the frame count of each short sequence to 120 frames. We employ conditional Segment Score Distillation (SSD), using the text prompt as the conditioning input. At the beginning, we randomly initialize a long motion sequence. In this experiment, we omit the first stage of InfiniDreamer, meaning that we do not use MDM to generate the initial short motion sequence. In the subsequent stages, we set the guidance scale to 10.0 and the learning rate to 0.005. As shown in Fig. 6, we use "a person takes 3 steps backward" as our textual prompt. InfiniDreamer, through conditional optimization, extends the generation capability of the original Motion Diffusion Model (MDM) from 70-200 frames to 520 frames, while maintaining alignment between the generated motions and the input text.

Algorithm 1 Segment Score Distillation (SSD)

Require: Initial motion sequence M_0, motion diffusion model ϕ, number of iteration N, windows size W, stride S, learning rate η
Ensure: Optimized long motion sequence M

Initialize: $M \leftarrow M_0$
Set: Gradient masks
for $n = 1$ to N **do**
  Sample a start index i from $\{1, 1+S, \ldots, \text{len}(M) - W\}$
  Extract a motion segment $x_0^i \leftarrow M[i : i+W]$
  Choose the textual prompt based on strategy
  Sample a timestep $t \sim \mathcal{U}(20, 980)$
  Add noise to obtain $x_t^i = \sqrt{\alpha_t} x_0^i + \sqrt{\sigma_t}\epsilon$, where $\epsilon \sim \mathcal{N}(0, I)$
  Compute total SSD loss via Eq. 8
  Update M by backpropagating $\mathcal{L}_{ssd}$ and adjusting the values of x_0^i in M
end for

return Optimized long motion sequence M

11. Limitation

InfiniDreamer is capable of generating arbitrarily long motion sequences based on text prompts, even in single-text scenarios. However, our framework still has some limitations. For example, the generation of sub-motions is constrained by the performance of the short-sequence generation model. Additionally, our method is slower compared to other sampling approaches, taking approximately 4 minutes to generate a 520-frame sequence. In the future, we plan to improve its efficiency and enhance InfiniDreamer's performance by advancing the capabilities of the short-sequence generation model.