

Appendix

A Preliminaries

A.1 Scalable Interpolant Transformer (SiT)

Scalable Interpolant Transformers (SiT) [53] is a diffusion transformer [56] based on *stochastic interpolants*, which defines a class of time-indexed stochastic processes that map a noise sample $\mathbf{x}_0 \sim p_0$ to an intermediate point $\mathbf{x}(t)$ over $t \in [0, 1]$. These interpolants are defined without requiring a target sample \mathbf{x}_1 (as in standard DDPMs), and instead use structured noise perturbation with learned dynamics to generate samples.

The interpolant is defined as:

$$\mathbf{x}(t) = \alpha(t)\mathbf{x}_0 + \sigma(t)\epsilon, \quad \epsilon \sim \mathcal{N}(0, I),$$

where $\alpha(t)$ and $\sigma(t)$ are time-dependent scalar functions, \mathbf{x}_0 is a noise sample (analogous to initial state in forward diffusion). It holds that $\mathbf{x}(0) = \mathbf{x}_0$, and $\mathbf{x}(1)$ ideally follows p_1 , the data distribution.

These interpolants allow generation and learning without explicitly simulating a forward diffusion trajectory from \mathbf{x}_1 to \mathbf{x}_0 as in traditional diffusion models.

Forward SDE. The forward generative process in diffusion models is typically described using a stochastic differential equation of the form:

$$d\mathbf{x} = f(\mathbf{x}, t) dt + g(t) d\mathbf{w}_t,$$

where $f(\mathbf{x}, t)$ is a drift function, $g(t)$ is a diffusion coefficient, \mathbf{w}_t is standard Brownian motion.

Reverse-time SDE. The reverse-time dynamics of this process can be derived from the theory of time-reversal of stochastic processes. The reverse time SDE is:

$$d\mathbf{x} = [f(\mathbf{x}, t) - g(t)^2 \nabla_{\mathbf{x}} \log p_t(\mathbf{x})] dt + g(t) d\bar{\mathbf{w}}_t,$$

where $\nabla_{\mathbf{x}} \log p_t(\mathbf{x})$ is the *score function*, which denotes the gradient of the log-density at time t , $d\bar{\mathbf{w}}_t$ is a reverse-time Brownian motion.

This reverse-time SDE shows that, to sample from the data distribution starting from noise, we need to access the time-dependent score function of intermediate states. This is typically learned using score matching in diffusion models.

Probability flow ODE. An alternative, deterministic formulation of the same marginal distributions is given by the probability flow ODE:

$$\frac{d\mathbf{x}}{dt} = f(\mathbf{x}, t) - \frac{1}{2}g(t)^2 \nabla_{\mathbf{x}} \log p_t(\mathbf{x}).$$

Unlike the reverse SDE, this ODE yields a deterministic mapping from \mathbf{x}_0 to \mathbf{x}_1 . Critically, both the reverse-time SDE and the probability flow ODE share the same marginal distribution $p_t(\mathbf{x})$ for each t .

This connection allows one to model generation either stochastically (via sampling the reverse SDE) or deterministically (via integrating the ODE). In SiT, the idea is to sidestep direct score estimation and instead predict the time-derivative of the interpolant path.

Velocity fields and learning objectives. Rather than explicitly learning the score function $\nabla_{\mathbf{x}} \log p_t(\mathbf{x})$, SiT models the trajectory of the interpolant $\mathbf{x}(t)$ by learning its *velocity field*:

$$\mathbf{v}(\mathbf{x}(t), t) := \frac{d\mathbf{x}(t)}{dt}.$$

Given the analytic form of the interpolant:

$$\mathbf{x}(t) = \alpha(t)\mathbf{x}_0 + \sigma(t)\epsilon,$$

its time derivative is:

$$\frac{d\mathbf{x}(t)}{dt} = \dot{\alpha}(t)\mathbf{x}_0 + \dot{\sigma}(t)\epsilon.$$

Since \mathbf{x}_0 and ϵ are both known (sampled during training), this velocity is analytically computable.

The SiT model learns a velocity estimator $\mathbf{v}_\theta(\mathbf{x}(t), t)$ by minimizing the expected squared error between the predicted and true velocity:

$$\mathcal{L}(\theta) = \mathbb{E}_{\mathbf{x}_0, \epsilon} \left[\int_0^1 \left\| \mathbf{v}_\theta(\mathbf{x}(t), t) - \frac{d\mathbf{x}(t)}{dt} \right\|^2 dt \right].$$

This learning objective removes the need for score estimation or denoising objectives and allows SiT to scale better.

Interpolants design. The choice of interpolant functions $\alpha(t)$, the deterministic scaling of the input noise \mathbf{x}_0 and $\sigma(t)$, the time-dependent standard deviation controlling the magnitude of the stochastic perturbation directly determines the geometry of the stochastic path $\mathbf{x}(t) = \alpha(t)\mathbf{x}_0 + \sigma(t)\epsilon$. This, in turn, influences training dynamics, sample quality, and generalization. Below, we describe two common interpolant designs that capture different trade-offs

The linear interpolant is defined as:

$$\alpha(t) = 1 - t, \quad \sigma(t) = \sqrt{t}.$$

This design induces a *linear* interpolation in the input \mathbf{x}_0 and a *square-root* scaling of the noise. At $t = 0$, we have $\mathbf{x}(0) = \mathbf{x}_0$; at $t = 1$, $\alpha(1) = 0$ and $\sigma(1) = 1$, hence $\mathbf{x}(1) = \epsilon$, a pure noise sample.

This interpolant is simple and intuitive, but its marginal distribution $p_t(\mathbf{x})$ varies in both mean and variance across time. Specifically:

$$\mathbb{E}[\mathbf{x}(t)] = \alpha(t)\mathbb{E}[\mathbf{x}_0] = 0, \quad \text{Var}[\mathbf{x}(t)] = \alpha(t)^2 + \sigma(t)^2 = (1 - t)^2 + t.$$

Thus, the total variance is time-varying.

To address the variance inconsistency, the GVP interpolant is constructed such that the total variance remains constant over time:

$$\alpha(t)^2 + \sigma(t)^2 = 1.$$

A canonical choice under this constraint is:

$$\alpha(t) = \cos\left(\frac{\pi}{2}t\right), \quad \sigma(t) = \sin\left(\frac{\pi}{2}t\right).$$

This design ensures that:

$$\mathbf{x}(t) \sim \mathcal{N}(0, I), \quad \forall t \in [0, 1].$$

That is, the marginal distribution of $\mathbf{x}(t)$ stays isotropic Gaussian throughout the path. This simplifies score estimation and enhances training stability. Moreover, the smooth transition from \mathbf{x}_0 to noise is nonlinear, leading to smoother gradients and more coherent sample trajectories.

The choice between them depends on downstream task requirements and model capacity.

A.2 Lie Groups for Rigid Body Rotations and Motions

Rigid body rotations and transformations in three-dimensional space are not elements of Euclidean space, but instead belong to structured non-Euclidean manifolds with group structures, specifically Lie groups. This geometric structure is critical for ensuring mathematically consistent operations such as interpolation, averaging, and noise perturbation, which are frequently needed in motion analysis and generation tasks.

The Special Orthogonal Group $\text{SO}(3)$. The space of all 3D rotation matrices forms a Lie group known as the special orthogonal group:

$$\text{SO}(3) = \{R \in \mathbb{R}^{3 \times 3} \mid R^\top R = I, \det(R) = 1\},$$

which is a compact, connected, non-commutative Lie group of dimension 3. Each element of $\text{SO}(3)$ represents a proper rotation in \mathbb{R}^3 , and the group operation is matrix multiplication. The non-Euclidean nature of $\text{SO}(3)$ implies that standard linear operations, such as averaging two rotation matrices or interpolating between them, may lead to results that no longer lie on the manifold.

639 **The Special Euclidean Group SE(3).** For full rigid body transformations, including both rotation
 640 and translation, the appropriate Lie group is SE(3):

$$\text{SE}(3) = \left\{ \begin{bmatrix} R & t \\ 0 & 1 \end{bmatrix} \in \mathbb{R}^{4 \times 4} \mid R \in \text{SO}(3), t \in \mathbb{R}^3 \right\},$$

641 which is a 6-dimensional, non-compact, non-commutative Lie group. The group operation is again
 642 matrix multiplication, and SE(3) encapsulates both orientation and position of a rigid body in space.

643 **Lie Algebras and Local Coordinates.** Associated with each Lie group \mathcal{G} is a Lie algebra \mathfrak{g} , which
 644 serves as the tangent space at the identity element and provides a local, linear coordinate system for
 645 the manifold. For SO(3), the Lie algebra $\mathfrak{so}(3)$ consists of all 3D skew-symmetric matrices:

$$\mathfrak{so}(3) = \{A \in \mathbb{R}^{3 \times 3} \mid A^\top = -A\}.$$

646 A standard isomorphism between \mathbb{R}^3 and $\mathfrak{so}(3)$ is provided by the **hat operator**:

$$[\omega]_\times = \begin{bmatrix} 0 & -\omega_3 & \omega_2 \\ \omega_3 & 0 & -\omega_1 \\ -\omega_2 & \omega_1 & 0 \end{bmatrix}, \quad \omega \in \mathbb{R}^3,$$

647 which maps a vector to its corresponding skew-symmetric matrix. The inverse operation is the **vee**
 648 **operator**, mapping from $\mathfrak{so}(3)$ to \mathbb{R}^3 .

649 **Exponential and Logarithmic Maps.** The **exponential map** $\exp : \mathfrak{g} \rightarrow \mathcal{G}$ and its inverse, the
 650 **logarithmic map** $\log : \mathcal{G} \rightarrow \mathfrak{g}$, provide the tools to move between the manifold and its tangent space.
 651 For SO(3), the exponential map is given in closed form by Rodrigues' formula:

$$\exp([\omega]_\times) = I + \frac{\sin \theta}{\theta} [\omega]_\times + \frac{1 - \cos \theta}{\theta^2} [\omega]_\times^2, \quad \theta = \|\omega\|.$$

652 This constructs a rotation matrix corresponding to a rotation of angle θ around axis $\omega/\|\omega\|$. The
 653 logarithmic map inverts this operation, computing the minimal-axis rotation vector ω corresponding
 654 to a given rotation matrix R :

$$\log(R) = \frac{\theta}{2 \sin \theta} (R - R^\top), \quad \theta = \cos^{-1} \left(\frac{\text{Tr}(R) - 1}{2} \right).$$

655 **Interpolation and Perturbation on Lie Groups.** A major benefit of the Lie group structure is
 656 that interpolation and noise perturbation can be carried out in the tangent space, ensuring that the
 657 results lie back on the manifold after mapping. Given two rotations $R_1, R_2 \in \text{SO}(3)$, a geodesic
 658 interpolation can be defined via:

$$R(t) = R_1 \cdot \exp(t \cdot \log(R_1^\top R_2)), \quad t \in [0, 1],$$

659 which traces the shortest path on the manifold between R_1 and R_2 . More generally, any operation of
 660 the form:

$$R = \exp(\omega), \quad \omega \sim \mathcal{N}(0, \Sigma),$$

661 defines a distribution on SO(3) by sampling from a Gaussian in the tangent space \mathbb{R}^3 and mapping
 662 to the manifold via the exponential map. Such constructions are widely used in manifold-aware
 663 generative models and motion synthesis.

664 **Extensions to SE(3).** For rigid body motion in SE(3), the corresponding Lie algebra $\mathfrak{se}(3)$ is a
 665 6-dimensional space encoding translational and rotational velocities. Elements of $\mathfrak{se}(3)$ are typically
 666 expressed using twist coordinates $(v, \omega) \in \mathbb{R}^6$, and the exponential or logarithmic maps general-
 667 ize accordingly. The Baker-Campbell-Hausdorff formula governs the non-linear composition of
 668 transformations, and matrix representations of exp and log are available via the theory of screw
 669 motions.

670 This Lie group machinery forms the mathematical foundation for handling rotation and transformation
 671 data in a consistent, geometry-aware manner, and is indispensable in domains such as robotics,
 672 graphics, and motion modeling.

B User Study

This study conducts a comprehensive user evaluation of our method compared with MotionLCM [17] and MoGenTS [83]. We assess the real-world applicability of motion sequences generated by Motion Anything and baseline models using a Google Forms survey completed by 50 participants. As shown in Figure 7, the user interface presents 3–4 motion clips (Videos 1–3/4) generated by the same model, followed by a comparative set (Videos A–C) from different models. Participants evaluate each animation based on motion accuracy and overall user experience, using a 3-point scale (1 = low, 3 = high). In the comparison section, users select the model they perceive as most realistic and engaging. This evaluation is designed to measure both the fidelity of the generated motion to real-world human movement and the overall effectiveness of each model in delivering visually compelling results.

Results:

- Our method achieved a motion quality rating of **2.92**, with **94%** of participants agreeing that it produces high-quality motion with minimal jitter, sliding, or unrealistic artifacts.
- For motion diversity, we received a rating of **2.86**, with **90%** of participants indicating that our method generates complex and varied motion sequences.
- In terms of text-motion alignment, our model scored **2.80**, and **82%** of users reported that the generated motions were well-aligned with the given text descriptions.
- Notably, **92%** of participants preferred our method over competing approaches.

C Qualitative Results

To qualitatively evaluate our performance in text-to-motion generation, we compare the visualizations generated by our method with those produced by both state-of-the-art diffusion and VQ-VAE based methods specializing in text-to-motion generation, including MotionLCM [17] and MoGenTS [83]. The text prompts are customized based on the HumanML3D [31] test set. As shown in Figure 8 and video demos, our method generates motions with superior quality, greater diversity, and better alignment between text and motion compared to the previous state-of-the-art methods.

D Full Comparison Tables

To comprehensively evaluate our method on text-to-motion generation, we report full comparisons with prior approaches in Table 5 and 6. Our method consistently achieves state-of-the-art performance on both HumanML3D [31] and KIT-ML [59], outperforming existing baselines across all metrics.

E Broader Impacts

Our work advances the field of 3D human motion generation by addressing key limitations in efficiency and scalability that hinder real-world deployment. By introducing frequency-aware sparsification and a scalable transformer architecture with principled geometric modeling, FlashMo offers a more practical solution for generating realistic human motion at scale. This has broad implications for downstream applications such as human-robot interaction, AR/VR environments, animation, and digital avatars. By reducing computational demands without sacrificing quality, our approach lowers the barrier to adoption in resource-constrained settings and paves the way for real-time, interactive, and embodied AI systems.

Table 5: Comparison of text-to-motion generation on HumanML3D [31] dataset. \rightarrow indicates the closer to real data, the better. **Bold** and underline indicate best and second best results. Efficient motion diffusion models are highlighted in blue.

Method	Venue	AIT(s) \downarrow	R-Precision \uparrow			FID \downarrow	MM Dist \downarrow	Diversity \rightarrow	MModality \uparrow
			Top-1	Top-2	Top-3				
Real	-	-	0.511 \pm .003	0.703 \pm .003	0.797 \pm .002	0.002 \pm .000	2.974 \pm .008	9.503 \pm .065	-
TM2T [32]	ECCV 2022	0.760	0.424 \pm .003	0.618 \pm .003	0.729 \pm .002	1.501 \pm .017	3.467 \pm .011	8.589 \pm .076	2.424 \pm .093
T2M-GPT [90]	CVPR 2023	0.380	0.492 \pm .003	0.679 \pm .002	0.775 \pm .002	0.141 \pm .005	3.121 \pm .009	9.722 \pm .082	1.831 \pm .048
CoMo [38]	ECCV 2024	0.620	0.502 \pm .002	0.692 \pm .007	0.790 \pm .002	0.262 \pm .004	3.032 \pm .015	9.936 \pm .066	1.013 \pm .046
MMM [58]	CVPR 2024	0.081	0.504 \pm .003	0.696 \pm .003	0.794 \pm .002	0.080 \pm .003	2.998 \pm .007	9.411 \pm .058	1.164 \pm .041
MoMask [30]	CVPR 2024	0.120	0.521 \pm .002	0.713 \pm .002	0.807 \pm .002	0.045 \pm .002	2.958 \pm .008	-	1.241 \pm .040
BAMM [57]	ECCV 2024	0.411	0.525 \pm .002	0.720 \pm .003	0.814 \pm .003	0.055 \pm .002	2.919 \pm .008	9.717 \pm .089	1.687 \pm .051
MoGenTS [83]	NeurIPS 2024	0.181	0.529 \pm .003	0.719 \pm .002	0.812 \pm .002	0.033 \pm .001	2.867 \pm .006	9.570 \pm .077	-
MDM [68]	ICLR 2023	24.74	0.320 \pm .005	0.498 \pm .004	0.611 \pm .007	0.544 \pm .044	5.566 \pm .027	9.559 \pm .086	2.799 \pm .072
MotionDiffuse [91]	TPAMI 2024	14.74	0.491 \pm .001	0.681 \pm .001	0.782 \pm .001	0.630 \pm .001	3.113 \pm .001	9.410 \pm .049	1.553 \pm .042
MLD [7]	CVPR 2023	0.217	0.481 \pm .003	0.673 \pm .003	0.772 \pm .002	0.473 \pm .013	3.196 \pm .010	9.724 \pm .082	2.413 \pm .079
ReMoDiffuse [92]	ICCV 2023	0.624	0.510 \pm .005	0.698 \pm .006	0.795 \pm .004	0.103 \pm .004	2.974 \pm .016	9.018 \pm .075	1.795 \pm .043
M2DM [43]	ICCV 2023	-	0.497 \pm .003	0.682 \pm .002	0.763 \pm .003	0.352 \pm .005	3.134 \pm .010	9.926 \pm .073	3.587 \pm .072
Fg-T2M [71]	ICCV 2023	-	0.492 \pm .002	0.683 \pm .003	0.783 \pm .002	0.243 \pm .019	3.109 \pm .007	9.278 \pm .072	1.614 \pm .049
FineMoGen [93]	NeurIPS 2023	-	0.504 \pm .002	0.690 \pm .002	0.784 \pm .002	0.151 \pm .008	2.998 \pm .008	9.263 \pm .094	2.696 \pm .079
GraphMotion [41] (50-step)	NeurIPS 2023	0.776	0.496 \pm .003	0.686 \pm .003	0.778 \pm .002	0.118 \pm .008	3.143 \pm .009	9.796 \pm .069	2.603 \pm .095
GraphMotion [41] (150-step)	NeurIPS 2023	2.552	0.504 \pm .003	0.699 \pm .002	0.785 \pm .002	0.116 \pm .007	3.070 \pm .008	9.692 \pm .067	2.766 \pm .096
B2A-HDM [78]	AAAI 2024	-	0.511 \pm .002	0.699 \pm .002	0.791 \pm .002	0.084 \pm .004	3.020 \pm .010	9.526 \pm .080	1.914 \pm .078
M2D2M [12]	ECCV 2024	-	-	-	0.799 \pm .002	0.087 \pm .004	3.018 \pm .008	9.672 \pm .086	2.115 \pm .079
MotionLCM [17] (1-step)	ECCV 2024	0.030	0.502 \pm .003	0.701 \pm .002	0.803 \pm .002	0.467 \pm .012	3.022 \pm .009	9.631 \pm .066	2.172 \pm .082
MotionLCM [17] (2-step)	ECCV 2024	0.035	0.505 \pm .003	0.705 \pm .002	0.805 \pm .002	0.368 \pm .011	2.986 \pm .008	9.640 \pm .052	2.187 \pm .094
MotionLCM [17] (4-step)	ECCV 2024	0.043	0.502 \pm .003	0.698 \pm .002	0.798 \pm .002	0.304 \pm .012	3.012 \pm .007	9.607 \pm .066	2.259 \pm .092
EMDM [102]	ECCV 2024	0.050	0.498 \pm .007	0.684 \pm .006	0.786 \pm .006	0.112 \pm .019	3.110 \pm .027	<u>9.551</u> \pm .078	1.641 \pm .078
Motion Mamba [97]	ECCV 2024	0.058	0.502 \pm .003	0.693 \pm .002	0.792 \pm .002	0.281 \pm .009	3.060 \pm .058	9.871 \pm .084	2.294 \pm .058
StableMoFusion [37]	MM 2024	0.499	0.553 \pm .003	0.748 \pm .002	0.841 \pm .002	0.098 \pm .003	-	9.748 \pm .092	1.774 \pm .051
MotionLCM-V2 [16] (1-step)	Preprint 2024	0.031	0.546 \pm .003	0.743 \pm .002	0.837 \pm .002	0.072 \pm .003	2.767 \pm .007	9.577 \pm .070	1.858 \pm .056
MotionLCM-V2 [16] (2-step)	Preprint 2024	0.038	0.551 \pm .003	0.745 \pm .002	0.836 \pm .002	0.049 \pm .003	2.765 \pm .008	9.584 \pm .066	1.833 \pm .052
MotionLCM-V2 [16] (4-step)	Preprint 2024	0.050	0.553 \pm .003	0.746 \pm .002	0.837 \pm .002	0.056 \pm .003	2.773 \pm .009	9.598 \pm .067	1.758 \pm .056
MMDM-t [8]	Preprint 2024	-	0.464 \pm .006	0.654 \pm .007	0.754 \pm .005	0.319 \pm .026	3.288 \pm .023	9.299 \pm .064	2.741 \pm .112
MMDM-b [8]	Preprint 2024	-	0.435 \pm .006	0.627 \pm .006	0.733 \pm .007	0.285 \pm .032	3.363 \pm .029	9.398 \pm .088	2.701 \pm .083
FTMoMamba [46]	Preprint 2024	-	0.489 \pm .003	0.680 \pm .002	0.777 \pm .002	0.181 \pm .009	3.151 \pm .009	9.789 \pm .085	2.277 \pm .099
Light-T2M [87]	AAAI 2025	0.151	0.511 \pm .003	0.699 \pm .002	0.795 \pm .002	0.040 \pm .002	3.002 \pm .008	-	1.670 \pm .061
Free-MDM [6]	Preprint 2025	0.045	0.466 \pm .008	0.657 \pm .007	0.757 \pm .005	0.256 \pm .045	-	9.666 \pm .080	-
Free-StableMoFusion [6]	Preprint 2025	0.036	0.520 \pm .013	0.707 \pm .003	0.803 \pm .006	0.051 \pm .002	-	9.480 \pm .005	-
MotionPCM [40] (1-step)	Preprint 2025	0.031	0.560 \pm .002	0.752 \pm .003	0.844 \pm .002	0.044 \pm .003	<u>2.711</u> \pm .008	9.559 \pm .081	1.772 \pm .067
MotionPCM [40] (2-step)	Preprint 2025	0.036	0.555 \pm .002	0.749 \pm .002	0.839 \pm .002	0.033 \pm .002	<u>2.739</u> \pm .007	9.618 \pm .088	1.760 \pm .068
MotionPCM [40] (4-step)	Preprint 2025	0.045	0.559 \pm .003	0.752 \pm .003	0.842 \pm .002	0.030 \pm .002	<u>2.716</u> \pm .008	9.575 \pm .082	1.714 \pm .062
Fg-T2M++ [72]	ICCV 2025	-	0.513 \pm .002	0.702 \pm .002	0.801 \pm .003	0.089 \pm .004	2.925 \pm .007	9.223 \pm .114	2.625 \pm .084
BioMoDiffuse [42]	Preprint 2025	-	0.547 \pm .003	0.743 \pm .002	0.835 \pm .002	0.071 \pm .003	2.784 \pm .008	9.567 \pm .086	1.919 \pm .063
HiSTF Mamba [88] (10-step)	Preprint 2025	0.690	0.488 \pm .005	0.685 \pm .004	0.784 \pm .005	0.189 \pm .018	3.101 \pm .022	9.712 \pm .090	2.529 \pm .044
HiSTF Mamba [88] (15-step)	Preprint 2025	-	0.504 \pm .005	0.699 \pm .005	0.798 \pm .005	0.249 \pm .023	3.053 \pm .022	9.383 \pm .091	2.276 \pm .036
ACMo [74]	Preprint 2025	-	0.493 \pm .002	0.698 \pm .003	0.795 \pm .002	0.102 \pm .003	2.973 \pm .006	9.749 \pm .082	2.614 \pm .100
FlashMo (Ours)	-	0.027	<u>0.562</u> \pm .004	<u>0.754</u> \pm .005	<u>0.847</u> \pm .005	0.041 \pm .002	<u>2.711</u> \pm .006	9.614 \pm .056	2.812 \pm .046
FlashMo w/ pretrain (Ours)	-	0.027	0.568 \pm .005	0.761 \pm .002	0.851 \pm .003	0.029 \pm .002	2.703 \pm .005	9.601 \pm .073	<u>2.851</u> \pm .069



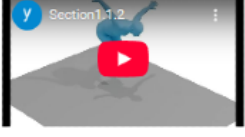
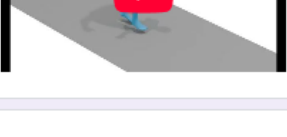

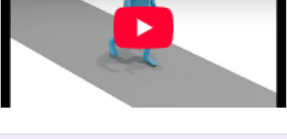
<p>Section 1</p> <p>There are two major questions for Section 1. The first question requests you to overall qualify the motions provided, and for the second question please choose the best motion from a set of motions.</p>	<p>Section 1.2</p>
<p>Section 1.1</p>	
<p>The sim appears to walk forward bend slightly grabbing an object with their left hand.</p>	
	<p>a</p> 
<p>A person is spinning with is arms spread out and then he falls over.</p>	
	<p>b</p> 
<p>He starts to dance a lot.</p>	
	<p>c</p> 
<p>Question 1</p> <p>Please assess the overall quality of the provided motion generations. Specifically, determine whether they exhibit noticeable jitter, sliding, or unrealistic movement, and select the statement that best reflects your evaluation.</p> <ul style="list-style-type: none"> <input type="radio"/> The motion generation is completely unusable. <input type="radio"/> The visualization is acceptable, but there are several aspects that need improvement. <input type="radio"/> The visualization is strong, highlighting the potential impact and promising future for applications. 	<p>Question 4</p> <p>Which video you believe that has the best motion quality?</p> <ul style="list-style-type: none"> <input type="radio"/> a <input type="radio"/> b <input type="radio"/> c
<p>Question2</p> <p>Please evaluate the average diversity of the motion generations provided and select the statement that best reflects your assessment</p> <ul style="list-style-type: none"> <input type="radio"/> The motions lack diversity, showing repetitive and uniform patterns. <input type="radio"/> The motions exhibit moderate diversity, though certain elements are still predictable. <input type="radio"/> The motions are highly diverse, showcasing a wide range of movement patterns and creative variations. 	<p>Question 5</p> <p>Which video you believe that has the highest diversity?</p> <ul style="list-style-type: none"> <input type="radio"/> a <input type="radio"/> b <input type="radio"/> c
<p>Question3</p> <p>Please evaluate the overall degree of match between the texts and their corresponding motions provided, and select the option that best reflects your assessment:</p> <ul style="list-style-type: none"> <input type="radio"/> There is little to no alignment between the texts and motions. <input type="radio"/> The texts and motions align fairly well, though some details inconsistencies remain. <input type="radio"/> The texts and motions are well-aligned, accurately reflecting each other. 	<p>Question 6</p> <p>Which motion do you believe best aligns with or reflects the text: "Person walks forwards straight while stumbling"?</p> <ul style="list-style-type: none"> <input type="radio"/> a <input type="radio"/> b <input type="radio"/> c

Figure 7: User study Google Forms. The User Interface (UI) used in our user study.

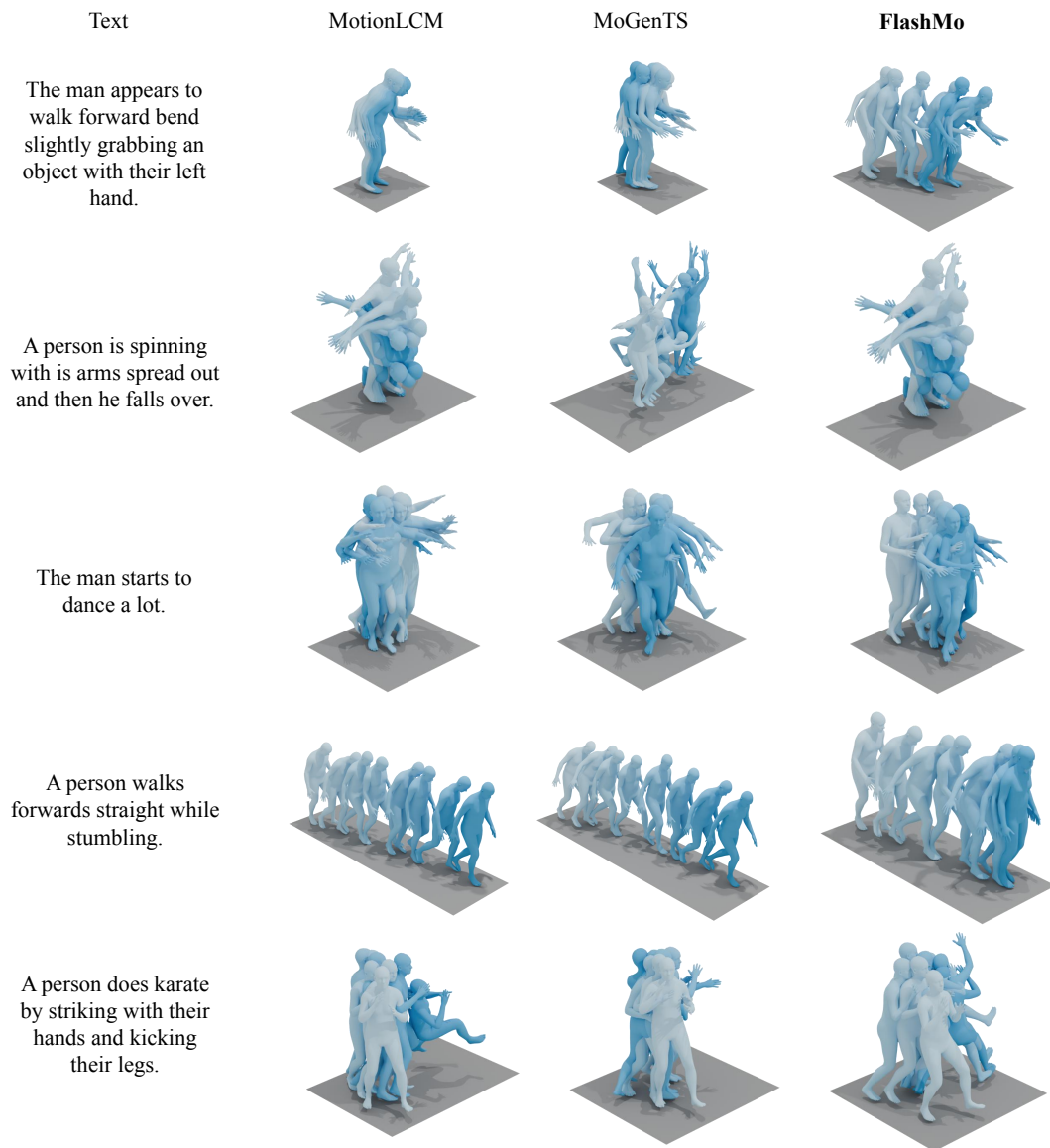


Figure 8: **Qualitative evaluation on HumanML3D [31] test set.** We qualitatively compared the visualizations generated by our method with those produced by MotionLCM [17] and MoGenTS [83].

Table 6: Comparison of text-to-motion generation on KIT-ML [59] dataset. \rightarrow indicates the closer to real data, the better. **Bold** and underline indicate best and second best results. Efficient motion diffusion models are highlighted in **blue**.

Method	Venue	R-Precision \uparrow			FID \downarrow	MM Dist \downarrow	Diversity \rightarrow	MModality \uparrow
		Top-1	Top-2	Top-3				
Real	-	0.424 \pm .005	0.649 \pm .006	0.779 \pm .006	0.031 \pm .004	2.788 \pm .012	11.08 \pm .097	-
TM2T [32]	ECCV 2022	0.280 \pm .005	0.463 \pm .006	0.587 \pm .005	3.599 \pm .153	4.591 \pm .026	9.473 \pm .117	3.292 \pm .081
T2M-GPT [90]	CVPR 2023	0.416 \pm .006	0.627 \pm .006	0.745 \pm .006	0.514 \pm .029	3.007 \pm .023	10.92 \pm .108	1.570 \pm .039
CoMo [38]	ECCV 2024	0.422 \pm .009	0.638 \pm .007	0.765 \pm .011	0.332 \pm .045	2.873 \pm .021	10.95 \pm .196	1.249 \pm .008
MMM [58]	CVPR 2024	0.404 \pm .005	0.621 \pm .005	0.744 \pm .004	0.316 \pm .028	2.977 \pm .019	10.91 \pm .101	1.232 \pm .039
MoMask [30]	CVPR 2024	0.433 \pm .007	0.656 \pm .005	0.781 \pm .005	0.204 \pm .011	2.779 \pm .022	-	1.131 \pm .043
BAMM [57]	ECCV 2024	0.438 \pm .009	0.661 \pm .009	0.788 \pm .005	0.183 \pm .013	2.723 \pm .026	11.01 \pm .094	1.609 \pm .065
MoGenTS [83]	NeurIPS 2024	0.445 \pm .006	<u>0.671</u> \pm .006	0.797 \pm .005	0.143 \pm .004	2.711 \pm .024	10.92 \pm .090	-
MDM [68]	ICLR 2023	0.164 \pm .004	0.291 \pm .004	0.396 \pm .004	0.497 \pm .021	9.191 \pm .022	10.85 \pm .109	1.907 \pm .214
MotionDiffuse [91]	TPAMI 2024	0.417 \pm .004	0.621 \pm .004	0.739 \pm .004	1.954 \pm .062	2.958 \pm .005	11.10 \pm .143	0.730 \pm .013
MLD [7]	CVPR 2023	0.390 \pm .008	0.609 \pm .008	0.734 \pm .007	0.404 \pm .027	3.204 \pm .027	10.80 \pm .117	2.192 \pm .071
ReMoDiffuse [92]	ICCV 2023	0.427 \pm .014	0.641 \pm .004	0.765 \pm .055	0.155 \pm .006	2.814 \pm .012	10.80 \pm .105	1.239 \pm .028
M2DM [43]	ICCV 2023	0.405 \pm .003	0.629 \pm .005	0.739 \pm .004	0.502 \pm .049	3.012 \pm .015	11.38 \pm .079	3.273 \pm .045
Fg-T2M [71]	ICCV 2023	0.418 \pm .005	0.626 \pm .004	0.745 \pm .004	0.571 \pm .047	3.114 \pm .015	10.93 \pm .083	1.019 \pm .029
FineMoGen [93]	NeurIPS 2023	0.432 \pm .006	0.649 \pm .005	0.772 \pm .006	0.178 \pm .007	2.869 \pm .014	10.85 \pm .115	1.877 \pm .093
GraphMotion [41] (50-step)	NeurIPS 2023	0.417 \pm .008	0.635 \pm .006	0.755 \pm .004	0.262 \pm .021	3.085 \pm .031	11.21 \pm .106	3.568 \pm .132
GraphMotion [41] (150-step)	NeurIPS 2023	0.429 \pm .007	0.648 \pm .006	0.769 \pm .006	0.313 \pm .013	3.076 \pm .022	11.12 \pm .135	3.627 \pm .113
B2A-HDM [78]	AAAI 2024	0.436 \pm .006	0.653 \pm .006	0.773 \pm .005	0.367 \pm .020	2.946 \pm .024	10.86 \pm .124	1.291 \pm .047
M2D2M [12]	ECCV 2024	-	-	0.753 \pm .006	0.378 \pm .023	3.012 \pm .021	10.71 \pm .121	2.061 \pm .067
EMDM [102]	ECCV 2024	0.443 \pm .006	0.660 \pm .006	0.780 \pm .005	0.261 \pm .014	2.874 \pm .015	10.96 \pm .093	1.343 \pm .089
Motion Mamba [97]	ECCV 2024	0.419 \pm .006	0.645 \pm .005	0.765 \pm .006	0.307 \pm .041	3.021 \pm .025	11.02 \pm .098	1.678 \pm .064
StableMoFusion [37]	MM 2024	0.445 \pm .006	0.660 \pm .005	0.782 \pm .004	0.258 \pm .029	-	10.94 \pm .077	1.362 \pm .062
MMDM-t [8]	Preprint 2024	0.432 \pm .006	0.643 \pm .007	0.760 \pm .006	0.237 \pm .013	2.938 \pm .025	10.84 \pm .125	1.457 \pm .129
MMDM-b [8]	Preprint 2024	0.386 \pm .007	0.603 \pm .006	0.729 \pm .006	0.408 \pm .022	3.215 \pm .026	10.53 \pm .100	2.261 \pm .144
Light-T2M [87]	AAAI 2025	0.444 \pm .006	0.670 \pm .007	0.794 \pm .005	0.161 \pm .009	2.746 \pm .016	-	1.005 \pm .036
Free-MDM [6]	Preprint 2025	0.382 \pm .006	0.587 \pm .006	0.707 \pm .007	0.401 \pm .033	-	10.73 \pm .102	-
Free-StableMoFusion [6]	Preprint 2025	0.431 \pm .003	<u>0.671</u> \pm .001	0.789 \pm .002	0.155 \pm .079	-	10.90 \pm .045	-
MotionPCM [40] (1-step)	Preprint 2025	0.433 \pm .007	0.654 \pm .007	0.781 \pm .008	0.355 \pm .011	2.820 \pm .022	10.78 \pm .078	1.337 \pm .047
MotionPCM [40] (2-step)	Preprint 2025	0.437 \pm .005	0.664 \pm .005	0.787 \pm .006	0.294 \pm .011	2.844 \pm .018	10.83 \pm .094	1.254 \pm .050
MotionPCM [40] (4-step)	Preprint 2025	0.443 \pm .005	0.664 \pm .004	0.789 \pm .005	0.336 \pm .013	2.881 \pm .023	10.76 \pm .096	1.258 \pm .056
Fg-T2M++ [72]	IJCV 2025	0.442 \pm .006	0.657 \pm .005	0.781 \pm .004	<u>0.135</u> \pm .004	2.696 \pm .011	10.99 \pm .105	1.255 \pm .078
BioMoDiffuse [42]	Preprint 2025	0.448 \pm .008	0.666 \pm .005	0.788 \pm .005	0.211 \pm .011	2.772 \pm .017	<u>11.11</u> \pm .094	1.380 \pm .050
HiSTF Mamba [88] (10-step)	Preprint 2025	0.437 \pm .006	0.651 \pm .006	0.772 \pm .006	0.289 \pm .021	2.846 \pm .018	10.92 \pm .096	1.512 \pm .088
HiSTF Mamba [88] (15-step)	Preprint 2025	0.440 \pm .006	0.657 \pm .006	0.774 \pm .006	0.293 \pm .017	2.819 \pm .015	10.93 \pm .099	1.347 \pm .056
FlashMo (Ours)	-	<u>0.449</u> \pm .002	0.670 \pm .004	<u>0.799</u> \pm .002	0.152 \pm .004	2.709 \pm .005	10.64 \pm .074	3.287 \pm .042
FlashMo w/ pretrain (Ours)	-	0.453 \pm .001	0.679 \pm .004	0.807 \pm .003	0.132 \pm .005	2.701 \pm .005	10.79 \pm .093	<u>3.591</u> \pm .070