SFM: Taming State Space Models for Text-to-Motion via Spatial-Frequency Modeling

Shang Gao[1,*], Haicheng Liao[1,*], Wenshuo Chen[2,*], Yumu Xie[1], Jiaxun Zhang[1], Bin Rao[1], Chengyue Wang[1], Yanchen Guan[1], Zhiyong Cui[3], Shiqi Ou[4], Yutao Yue[2], and Zhenning Li[1,†]

[1] State Key Laboratory of Internet of Things for Smart City, University of Macau, China
[2] The Hong Kong University of Science and Technology (Guangzhou), China
[3] Beihang University, China
[4] South China University of Technology, China

Abstract. Text-to-motion generation aims to synthesize realistic and semantically aligned human motions from natural language descriptions. Recent diffusion-based approaches predominantly adopt Transformer or CNN-based denoisers; However, the former suffers from quadratic complexity with respect to sequence length, while the latter is fundamentally constrained by limited receptive fields. To bridge this gap, we propose SFM, a state-space-enhanced diffusion framework that introduces a structured and efficient long-range modeling scheme tailored for motion generation. At its core, SFM integrates two novel components into a UNet backbone: a Frequency State-Space Module (F-SSM), which transforms motion features into the frequency domain with text-gated multi-scale modulation, and a Spatial State-Space Module (S-SSM), which captures bidirectional spatial dependencies through parallel state transitions. This dual-axis decomposition enables SFM to model both the global structure and fine-grained variations of human motion while preserving alignment with linguistic prompts. Extensive experiments on HumanML3D and KIT-ML demonstrate that SFM outperforms strong Transformer- and CNN-based baselines in both fidelity and text-motion alignment, achieving superior R-Precision and FID with optimized speed.

Keywords: Human Motion Generation · Latent Diffusion Models · State Space Model

1 Introduction

Text-to-motion aims to synthesize realistic human motions from natural-language instructions, with a wide range of applications including film, VR/AR, robotics, and autonomous driving [23, 25, 35, 36, 38]. A practical system must satisfy three requirements simultaneously: (i) semantic alignment with the prompt, (ii)

∗ Equal contribution.
† Corresponding author: zhenningli@um.edu.mo

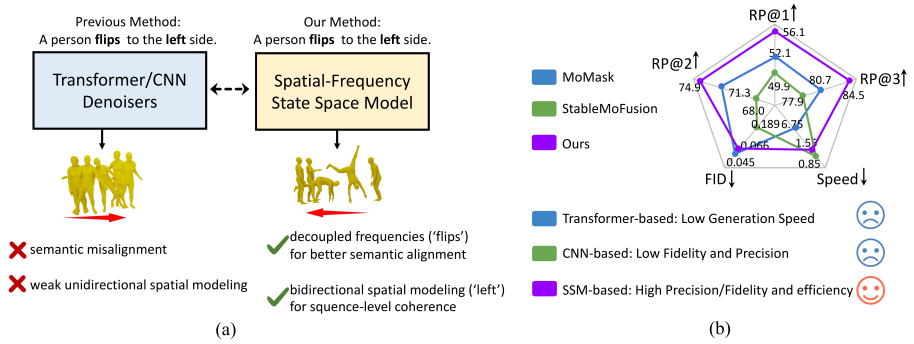

Fig. 1: (a) The limitations of existing methods compared to ours. (b) A comprehensive comparison on HumanML3D across different architectures. RP@1 denotes R-Precision Top1. By introducing dual-axes Spatial-Frequency state-space modeling, our model achieves the best performance–efficiency trade-offs.

long-horizon temporal coherence, and (iii) efficient generation. Existing diffusion pipelines typically choose between Transformer [12, 49–51] or CNN denoisers [6, 18, 21, 48], which makes it difficult to meet all three objectives at once in a unified framework. As shown in Fig. 1, Transformers capture long-range structure but suffer quadratic complexity with respect to sequence length; CNN-based UNets are efficient and adaptable but are constrained by local receptive fields, which weakens long-horizon consistency and text-conditioned detail. This tension underlies much of the current trade-off between fidelity, alignment, and efficiency.

What makes text-to-motion especially hard? Consider two typical prompts: "A person flips to the left side" and "A man crouched down and then rolled forward." The first requires understanding of **directional phrases** ("left side") and a **high-energy action** ("flip"), while the second requires **multi-step sequencing** ("crouched" → "rolled") with smooth transitions. Models must carry global structure over tens to hundreds of frames while preserving fine-grained limb articulations. When Transformer-based methods handle these cases, they often do so at prohibitive cost; when CNN-based methods do, they may miss either the long-range dependency or the fine detail.

State-space models (SSMs) provide an appealing and principled alternative to conventional modeling architectures. Recent formulations such as Mamba [9] recast sequence modeling into selective state transitions with a hardware-friendly parallel scan mechanism, yielding efficient linear-time modeling of ultra-long dependencies. This has effectively closed much of the gap with Transformers in large-scale language tasks. However, simply swapping a vanilla SSM into motion diffusion underperforms, because language-oriented SSMs lack task-specific mechanisms for (1) multi-scale motion dynamics and fine-grained detail and (2) explicit multimodal aggregation required by text-to-motion generation.

To address these challenges, we propose **SFM**, a UNet-based diffusion framework with task-aware state-space modules along two complementary axes:

- **Frequency axis.** Frequency State-Space Module (F-SSM). Operates in the frequency domain to separate global structure and local detail, and aligns multi-scale motion components with the text condition.
- **Spatial axis.** Spatial State-Space Module (S-SSM). Performs bidirectional state-space modeling in the spatial domain to strengthen sequence-level spatial awareness and coherence over long horizons.

These modules produce text-aware, multi-scale representations of motion while maintaining the linear-time scaling of the state-space components.

In summary, our main contributions are as follows:

- We propose SFM, integrating state-space modeling along frequency and spatial axes within a UNet architecture, allowing the model to capture long-range dependencies with linear complexity.
- We design the Frequency State-Space Module for multi-scale, text-aligned modeling in the frequency domain, and the Spatial State-Space Module for enhancing bidirectional spatial awareness and sequence-level coherence.
- SFM attains state-of-the-art or highly competitive R-Precision and FID compared with strong baselines, while simultaneously achieving faster generation speed than Transformer-based methods.

2 Related Works

2.1 Text-to-Motion

Recently, text-to-motion generation has achieved remarkable progress, with two dominant paradigms for motion sequence modeling: Transformer-based approaches and CNN-based methods. In the first paradigm, VAE-based methods [13, 33] leverage conditional VAE, which conditions on previously generated frames and text features, predicting motion according to text length. Autoregressive methods [2, 12, 37, 49, 52] treat motion as token sequences, employing a transformer decoder to generate subsequent tokens. DiT style methods [5,7,50,51] achieve realistic motion sequences synthesis through transformer denoisers. On the other hand, CNN-based methods based on Unet emphasize efficiency and diversity. GMD [21] introduces dense guidance via feature projection to convert sparse signals, often ignored during UNet denoising, into effective guiding signals, thereby enhancing spatial and local pose consistency. PhysDiff [48] integrates a physics-guided denoiser into the UNet structure, establishing a strong baseline for physical realism. Furthermore, MoFusion [6] employs a scheduled weighting strategy for kinematic losses within the UNet to effectively balance motion diversity and generation quality. StableMoFusion [18] employs a foot-ground contact correction mechanism and a direct prediction strategy to achieve robust and high-quality motion synthesis. However, Transformer-based methods suffer

from quadratic complexity with respect to sequence length, limiting their efficiency and practical deployment. CNN-based architectures offer adaptability and greater diversity, but still face challenges in generating faithful motion due to the constraints of local receptive fields. Recent work [53], attempts to integrate naive Mamba into human motion generation for real-time inference, but its performance still lags far behind advanced models. Its design lacks task-specific mechanisms for capturing multi-scale motion details and multimodal latent space aggregation. To push State Space Models to their limits, we equip task-aware state space modules along two complementary axes: the F-SSM performs adaptive frequency decoupling, aligning with semantic conditions in the frequency domain, while the S-SSM performs bidirectional modeling in the spatial domain, enabling the learning of sequence-level spatial awareness and coherence.

2.2 State Space Models

Recently, State Space Models (SSMs) [10, 11, 42], originating from classical control theory [20], have emerged as a promising architecture for sequence modeling due to their linear scaling with sequence length in long-range dependency modeling. The S4 Structured State-Space Sequence Model [10] pioneered long-range sequence modeling, followed by S5 [42], which introduced efficient parallel scanning and MIMO SSM mechanisms. More recently, Gu. [9] proposed a data-dependent SSM layer and developed Mamba, a generic language model backbone featuring selective mechanisms and hardware-aware design. Mamba achieves linear computational scaling and surpasses Transformer [44] architectures across multiple large-scale natural language datasets. Subsequently, Mamba's strong long-range modeling capability and linear scalability have been extended to vision tasks, including image classification [17, 26, 54], long video understanding [22, 45], image super-resolution [15, 16], medical image segmentation [28, 46], and beyond [24, 31, 34, 41, 47]. However, directly transferring vanilla Mamba to text-to-motion generation remains challenging due to its lack of mechanisms for capturing fine-grained motion details aligned with text conditions and for aggregating semantic features. In this work, we further explore the potential of Mamba for text-to-motion generation, introducing restoration-specific designs to establish an effective baseline for future research.

3 Method

3.1 Preliminaries

Text-to-motion diffusion basics. Given a text description, we obtain a text embedding $\mathbf{c}$ from a language encoder. Let $\mathbf{x}_0 \in \mathbb{R}^{N \times D}$ denote the target motion sequence with N frames and per-frame SMPL parameters of dimension D (e.g., joint positions or 6-DoF pose coefficients).

At a timestep $t \in \{1, \ldots, T\}$ sampled uniformly, the Gaussian noise is progressively added to the initial motion sequence $\mathbf{X}_0$ over T timesteps in the for-

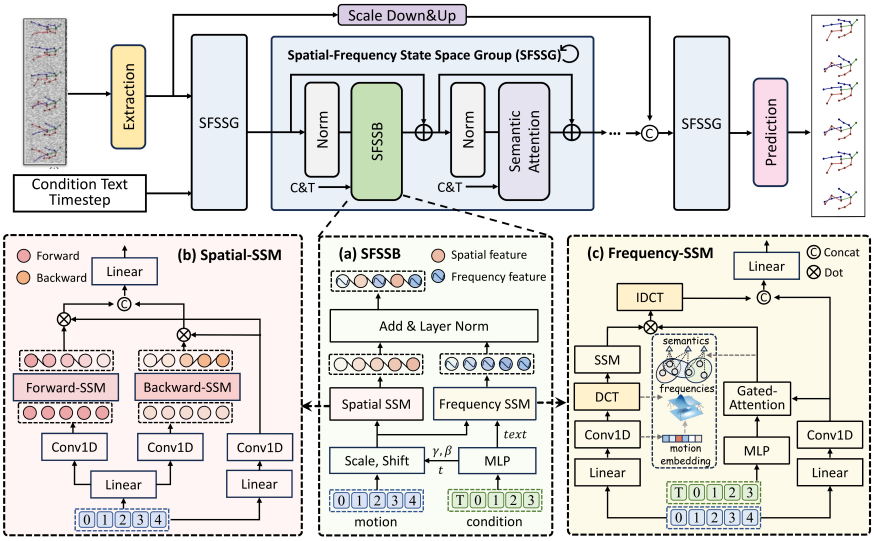

Fig. 2: Overall architecture of our SFM, as well as the (a) Spatial-Frequency State-Space Block (SFSSB), the (b) Spatial State-Space Module (S-SSM), and the (c) Frequency State-Space Module (F-SSM).

ward diffusion process:

$$\mathbf{x}_t = \sqrt{\alpha_t}\,\mathbf{x}_0 + \sqrt{1-\alpha_t}\,\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I}) \tag{1}$$

where $\boldsymbol{\epsilon}$ denotes the Gaussian noise, α_t is the noise schedule, $\alpha_t = \prod_{i=1}^{t}(1-\beta_i)$, and β_t is the variance parameter.

In the reverse process, a U-Net style denoising network G_θ is conditioned on $(t, \mathbf{c})$ and is trained to predict either the added noise or the original motion $\mathbf{x}_0$ from the noisy motion $\mathbf{x}_t$:

$$\min_{\theta} \; \mathbb{E}_{t,\mathbf{x}_0,\boldsymbol{\epsilon},\mathbf{c}} \left\| G_\theta(\mathbf{x}_t, t, \mathbf{c}) - \mathbf{x}_0 \right\|_2^2 \tag{2}$$

At inference, sampling starts from $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and proceeds backward with a standard sampler (e.g., DPM-Solver++ [27]). We apply classifier-free guidance (CFG) with scale ω:

$$G_{\text{CFG}}(\mathbf{x}_t, t, \mathbf{c}) = G_\theta(\mathbf{x}_t, t, \varnothing) + \omega\big[G_\theta(\mathbf{x}_t, t, \mathbf{c}) - G_\theta(\mathbf{x}_t, t, \varnothing)\big] \tag{3}$$

where $\varnothing$ denotes the unconditional input.

Selective State Space Models. Classical continuous-time LTI systems are written as

$$\dot{\mathbf{h}}(t) = \mathbf{A}\,\mathbf{h}(t) + \mathbf{B}\,\mathbf{x}(t), \qquad \mathbf{y}(t) = \mathbf{C}\,\mathbf{h}(t) + \mathbf{D}\,\mathbf{x}(t) \tag{4}$$

Discretizing with step size $\Delta > 0$ gives

$$\mathbf{h}_k = \bar{\mathbf{A}}\,\mathbf{h}_{k-1} + \bar{\mathbf{B}}\,\mathbf{x}_k, \qquad \mathbf{y}_k = \mathbf{C}\,\mathbf{h}_k + \mathbf{D}\,\mathbf{x}_k \tag{5}$$

where $\bar{\mathbf{A}} = e^{\mathbf{A}\Delta}$ and $\bar{\mathbf{B}} = \int_0^{\Delta} e^{\mathbf{A}\tau}\mathbf{B}\, d\tau$. This recurrence is equivalent to a 1D causal convolution. For an input sequence $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_M)$,

$$\mathbf{y} = \mathbf{x} * \mathbf{K}, \qquad \mathbf{K} = \left(\mathbf{C}\bar{\mathbf{B}},\, \mathbf{C}\bar{\mathbf{A}}\bar{\mathbf{B}},\, \ldots,\, \mathbf{C}\bar{\mathbf{A}}^{M-1}\bar{\mathbf{B}}\right) \tag{6}$$

so that $\mathbf{y}_k = \sum_{i=0}^{k-1} \mathbf{K}_i\, \mathbf{x}_{k-i}$

Selective SSMs (e.g., Mamba) make the filter *input-dependent* by parameterizing the timescale and projections as functions of the current input:

$$\Delta_k = \Delta(\mathbf{x}_k), \quad \mathbf{B}_k = \mathbf{B}(\mathbf{x}_k), \quad \mathbf{C}_k = \mathbf{C}(\mathbf{x}_k) \tag{7}$$

together with lightweight gating. A hardware-friendly selective scan computes these data-dependent convolutions in linear time and memory, enabling ultra-long sequence modeling. While such layers are highly effective for language, using them for text-to-motion requires task-specific adaptations (detailed next) to handle multi-scale motion detail and explicit text–motion aggregation.

3.2 SFM Architecture

As illustrated in Fig. 2, our proposed **SFM** consists of three main stages: *Extraction*, *Spatial-Frequency State Space Group (SFSSG)*, and *Prediction* head. Let $\mathbf{M}_t \in \mathbb{R}^{N\times D}$ denote the noisy motion sequence at diffusion timestep t, $\mathbf{M}_0 \in \mathbb{R}^{N\times D}$ represent the ground-truth motion, and $\mathbf{c}_{\text{text}} \in \mathbb{R}^{L\times d}$ be the textual embedding, where N and L denote the number of motion frames and the text length, respectively, and D and d indicate the feature dimensions of motion and text embeddings. The *Extraction* module is composed of two 1D convolutional layers with GroupNorm and Mish activations, which are employed to extract shallow motion features $\mathbf{F} \in \mathbb{R}^{N\times D'}$, where $D' > D$. This stage can be formulated as:

$$\mathbf{F} = \{\text{Mish \& Norm \& Conv1D}\}_{\times 2}(\mathbf{M}_t) \tag{8}$$

Subsequently, the shallow feature $\mathbf{F}$ undergoes a feature denoising stage to obtain the deep feature $\mathbf{F}^l$ at the l-th layer, where $l \in \{1, 2, \ldots, L\}$. This stage adopts a UNet-style architecture stacked by multiple Spatial-Frequency State Space Groups (SFSSGs), with each containing several Residual Spatial-Frequency State Space Blocks(SFSSBs). This stage can be formulated as:

$$\mathbf{F}^l = \text{Stacked}_{SFSSGs}(\mathbf{F}^{l-1}, \mathbf{c}_{\text{text}}, t), l \in \{1, 2, \ldots, L\} \tag{9}$$

Finally, the *Prediction* head, composed of a single 1D convolutional layer, is used to directly predict the real motion sequence $\mathbf{M}_0^{'} \in \mathbb{R}^{N\times D}$. This stage can be formulated as:

$$\mathbf{M}_0^{'} = \text{Conv1D}(\mathbf{F}^L) \tag{10}$$

3.3 Spatial-Frequency State Space Group

The Spatial-Frequency State Space Group adopts a UNet style architecture built upon State Space Models (SSMs), with Scale-down and Scale-up modules.

As illustrated in Fig. 2, for the l-th layer motion feature $\mathbf{F}^l$, we first apply a Normalization (LN) Layer, followed by a Spatial-Frequency State Space Block (SFSSB) to capture the spatial-frequency motion features aligned with the text condition, denoted as $\mathbf{F}^l_{sf}$. Subsequently, $\mathbf{F}^l_{sf}$ undergoes residual connection and normalization. Furthermore, a Semantic Attention layer is introduced at the end of the group to further enhance semantic fusion. Specifically, it uses the text condition $\mathbf{c}_{\text{text}}$ as the key and value, and the spatial-frequency motion features $\mathbf{F}^l_{sf}$ as the query. The added dropout is employed to enhance the model's generalization. The Scale-down and Scale-up structures are implemented using 1D convolution and transposed convolution, respectively, to downsample and upsample the motion features. The details are provided in the Appendix.

3.4 Spatial-Frequency State Space Block

In this section, we introduce the Spatial-Frequency State Space Block (SFSSB), a core component of our model that performs parallel long-range modeling of motion features in both the spatial and frequency domains.

As shown in Fig. 2, given a pose embedding $p^{(i)}$ of the i-th frame after Conv1D blocks dimensionality adjustment [18], a time step t, and text condition c_{text}, we employ a conditional linear modulation strategy. The time step t and condition embeddings c_{text} are processed by multilayer perceptrons (MLPs) to produce *scale* and *shift* parameters, which are then used for feature-wise linear modulation as follows:

$$f^{(i)} = p^{(i)} \times (1 + \text{scale}) + \text{shift} \tag{11}$$

The modulated features $\mathbf{F}^l = f^{(i)}, i \in \{1, \ldots, N\}$ at the $l-$th layer are then processed in parallel by two complementary modules: the **Frequency State-Space Module (F-SSM)** for capturing multi-scale details aligning with semantic conditions in the frequency domain, and the **Spatial State-Space Module (S-SSM)** for learning latent spatial-aware representations. Finally, the outputs from the spatial and frequency branches are combined through element-wise summation and normalized by LayerNorm, yielding the integrated spatio-frequency motion representation $\mathbf{F}^l_{sf}$.

3.5 Frequency State-Space Module

To enhance the vanilla SSM's ability to model multi-scale motion features aligning with the semantic condition, we introduce the Frequency State-Space Module (F-SSM). This module transforms motion features from the spatial domain into the frequency domain, where the different scale details can be naturally disentangled. The low-frequency components primarily correspond to the *structural patterns* of motion, while the high-frequency components correspond to the *fine-grained details* of motion.

As illustrated in Fig. 2 (c), given the input motion features $\mathbf{F}^{l-1}$ at layer $l-1$, we linearly project them into two embeddings, $\mathbf{x}_m$ and $\mathbf{z}_m$, each with dimensionality D and length N. A 1D convolution is then applied to both embeddings to

produce $\mathbf{x}'_m$ and $\mathbf{z}'_m$. Next, a discrete cosine transform (DCT) is applied to $\mathbf{x}'_m$ to convert it into the frequency domain and further modeled by the SSM layer to get the frequency feature $\mathbf{y}_f$. The DCT of $\mathbf{x}'_m$ at index k is defined:

$$c^d_{\mathbf{x}'_m}[k] = \alpha(k) \sum_{n=0}^{N-1} x^d_m[n] \cos\left(\frac{\pi(2n+1)k}{2N}\right) \quad k = 0, 1, \ldots, N-1 \tag{12}$$

where the normalization factor $\alpha(k)$ is defined as:

$$\alpha(k) = \begin{cases} \sqrt{\frac{1}{N}}, & \text{if } k = 0, \\ \sqrt{\frac{2}{N}}, & \text{if } k > 0. \end{cases} \tag{13}$$

Intuitively, we could predefine the first K coefficients to rigidly get the low- and high-frequency components. However, such a hard thresholding strategy fails to adaptively capture multi-scale motion features that align with the textual descriptions. Moreover, since the entire motion sequence is perturbed by uniform timestep-dependent noise, different temporal segments correspond to distinct semantic parts of the text. To this end, we propose a text-gated adaptive mechanism that models frequency-specific motion–text correlations in a decoupled frequency space, thereby improving the model's ability to follow textual semantics. The text-gated adaptive mechanism takes the text embedding $\mathbf{c}_{\text{text}}$ as the key, value, and auxiliary branch motion feature $\mathbf{z}'_m$ as the query. After gating, the fused text-aware features are processed by an MLP followed by a Sigmoid to generate a modulation mask $\mathbf{M} \in [0, 1]^{L \times 1}$, formulated as:

$$\mathbf{M} = \text{Sigmoid}\left(\text{MLP}\left(\text{CrossAttn}\left(\mathbf{z}'_m, \mathbf{c}_{\text{text}}\right)\right)\right) \tag{14}$$

The resulting mask $\mathbf{M}$ is subsequently used to modulate the frequency-domain motion feature $\mathbf{y}_f$, emphasizing the frequency components that are semantically relevant to the given text condition. This modulation can be formulated as:

$$\mathbf{y}'_f = \mathbf{M} \odot \mathbf{y}_f \tag{15}$$

where $\odot$ denotes element-wise multiplication. Finally, the modulated frequency-domain feature $\mathbf{y}'_f$ is transformed back to the spatial domain via inverse DCT, and concatenated with the auxiliary motion feature $\mathbf{z}'_m$ along the feature dimension. The combined representation is linearly projected to produce the final output $\mathbf{F}^l_f$, formulated as:

$$\mathbf{F}^l_f = \text{Linear}\left(\text{Concat}\left[\text{IDCT}(\mathbf{y}'_f), \mathbf{z}'_m\right]\right) \tag{16}$$

3.6 Spatial State-Space Module

To enable the vanilla SSM to learn spatially aware motion representations and further enhance sequence-level coherence, we introduce the Spatial State-Space

Module (S-SSM). This module models the motion sequence from both the forward and backward directions.

As shown in Fig. 2 (b), similar to the F-SSM, the input motion sequence $\mathbf{F}^{l-1}$ is first linearly projected into two representations, $\mathbf{x_m}$ and $\mathbf{z_m}$. In each direction, a 1D convolution is applied to $\mathbf{x_m}$ to obtain $\mathbf{x}'_o$, which is subsequently linearly transformed to yield $\mathbf{B}_o$, $\mathbf{C}_o$, and Δ_o in Eq.5. The parameter Δ_o is utilized to modulate $\mathbf{A}_o$ and $\mathbf{B}_o$, respectively. Next, the bidirectional state-space modeling is performed via the SSM layer, producing $\mathbf{y}_{\text{forward}}$ and $\mathbf{y}_{\text{backward}}$. Finally, the two directional outputs are gated by $\mathbf{z_m}$ and concatenated to project the output motion feature $\mathbf{F}^l_s$.

$$\begin{aligned} \mathbf{z}'_m &= \{\text{SiLU\&Conv1D}\}(\mathbf{z_m}) \\ \mathbf{F}^l_s &= \text{Linear}\Big(\text{Concat}\big[\mathbf{y}_{\text{forward}} \odot \mathbf{z}'_m,\, \mathbf{y}_{\text{backward}} \odot \mathbf{z}'_m\big]\Big) \end{aligned} \tag{17}$$

4 Experiments

4.1 Dataset

We conduct experiments on two widely used text-to-motion benchmarks: HumanML3D [13] and KIT-ML [39]. HumanML3D contains 14,616 motion sequences from AMASS [29] and HumanAct12 [14], each paired with three textual descriptions, for a total of 44,970 annotations that cover a broad set of actions (walking, exercising, dancing, etc.). KIT-ML is smaller, with 3,911 motions and 6,278 descriptions. We follow the official train/validation/test splits and the standard preprocessing used in prior work [18].

4.2 Implementation Details

We follow the architecture and setup of previous works [4, 18]. The model is trained for 200,000 steps with AdamW (batch size 64, learning rate 1×10^{-4}). The diffusion process uses $T = 1000$ with a linear β schedule. During training, we apply classifier-free guidance via unconditional dropout $p = 0.1$; at inference, we use a guidance scale of 3. Sampling uses the DPM-Solver++ sampler with 10 inference steps. All experiments run on NVIDIA A800 GPUs (80 GB).

4.3 Comparison to State-of-the-art Approaches

Quantitative results. We compare with recent Transformer- and CNN-based baselines, following the protocol in [12]: each metric is averaged over 20 runs and reported with 95% confidence intervals. Prior work [6, 30] notes that very low FID may not align with human preference in text-to-motion; we therefore treat R-Precision as the primary indicator and FID as a reference.

As shown in Table 1, our method attains the best R-Precision on HumanML3D (Top-1/2/3: 0.561/0.749/0.845) with the second-best FID (0.066). Against a CNN-based method (StableMoFusion), Top-3 improves from 0.779 to 0.845 and

Table 1: Quantitative results on HumanML3D and KIT-ML. ± denotes 95% confidence intervals. ↓: lower is better; ↑: higher is better; →: closer to Ground Truth (GT) is better. Red / Blue mark best / second-best per metric.

Method	R-Precision ↑			FID ↓	Diversity →	MM-Dist↓	Multimodality↑
	Top1	Top2	Top3				
HumanML3D							
Ground Truth	$0.511^{\pm0.003}$	$0.703^{\pm0.003}$	$0.797^{\pm0.002}$	$0.002^{\pm0.000}$	$9.503^{\pm0.065}$	-	-
Seq2Seq [40]	$0.180^{\pm.002}$	$0.300^{\pm.002}$	$0.396^{\pm.002}$	$11.75^{\pm.035}$	$6.223^{\pm.061}$	$5.529^{\pm.007}$	-
LJ2P [1]	$0.246^{\pm.001}$	$0.387^{\pm.002}$	$0.486^{\pm.002}$	$11.02^{\pm.046}$	$7.676^{\pm.058}$	$5.296^{\pm.008}$	-
T2G [3]	$0.165^{\pm.001}$	$0.267^{\pm.002}$	$0.345^{\pm.002}$	$7.664^{\pm.030}$	$6.409^{\pm.071}$	$6.030^{\pm.008}$	-
Hier [8]	$0.301^{\pm.002}$	$0.425^{\pm.002}$	$0.552^{\pm.004}$	$6.532^{\pm.024}$	$8.332^{\pm.042}$	$5.012^{\pm.018}$	-
TEMOS [33]	$0.424^{\pm.002}$	$0.612^{\pm.002}$	$0.722^{\pm.002}$	$3.734^{\pm.028}$	$8.973^{\pm.071}$	$3.703^{\pm.008}$	$0.368^{\pm.018}$
MLD [5]	$0.481^{\pm0.003}$	$0.673^{\pm0.003}$	$0.772^{\pm0.002}$	$0.473^{\pm0.013}$	$9.724^{\pm0.082}$	$3.196^{\pm0.010}$	$2.413^{\pm0.079}$
ReMoDiffuse [51]	$0.510^{\pm0.005}$	$0.698^{\pm0.006}$	$0.795^{\pm0.004}$	$0.103^{\pm0.004}$	$9.018^{\pm0.075}$	$3.025^{\pm0.008}$	$1.795^{\pm0.043}$
MotionDiffuse [50]	$0.491^{\pm0.001}$	$0.681^{\pm0.001}$	$0.782^{\pm0.001}$	$0.630^{\pm0.001}$	$9.410^{\pm0.049}$	$3.113^{\pm0.001}$	$1.553^{\pm0.042}$
MotionLCM [7]	$0.502^{\pm0.003}$	$0.701^{\pm0.002}$	$0.803^{\pm0.002}$	$0.467^{\pm0.012}$	$9.361^{\pm0.660}$	$3.012^{\pm0.007}$	$2.172^{\pm0.082}$
T2M-GPT [49]	$0.492^{\pm0.003}$	$0.679^{\pm0.002}$	$0.775^{\pm0.002}$	$0.141^{\pm0.005}$	$9.722^{\pm0.082}$	$3.121^{\pm0.009}$	$1.831^{\pm0.048}$
MMM [38]	$0.515^{\pm0.002}$	$0.708^{\pm0.002}$	$0.804^{\pm0.002}$	$0.089^{\pm0.002}$	$9.577^{\pm0.050}$	$2.926^{\pm0.007}$	$1.226^{\pm0.035}$
MoMask [12]	$0.521^{\pm0.002}$	$0.713^{\pm0.002}$	$0.807^{\pm0.002}$	$0.045^{\pm0.002}$	-	$2.958^{\pm0.008}$	$1.241^{\pm0.040}$
MDM [43]	$0.320^{\pm0.005}$	$0.498^{\pm0.004}$	$0.611^{\pm0.007}$	$0.544^{\pm0.044}$	$9.559^{\pm0.086}$	$5.556^{\pm0.027}$	$2.799^{\pm0.072}$
MoFusion [6]	$0.492^{\pm0.000}$	-	-	-	$8.82^{\pm0.000}$	-	$2.521^{\pm0.000}$
PhysDiff [48]	-	-	$0.631^{\pm0.000}$	$0.433^{\pm0.000}$	-	-	-
GMD [21]	-	-	$0.652^{\pm0.000}$	$0.235^{\pm0.000}$	$9.726^{\pm0.000}$	-	-
StableMoFusion [18]	$0.499^{\pm0.004}$	$0.680^{\pm0.006}$	$0.779^{\pm0.007}$	$0.189^{\pm0.003}$	$9.466^{\pm0.002}$	-	-
Motion Mamba [53]	$0.502^{\pm0.003}$	$0.693^{\pm0.002}$	$0.792^{\pm0.002}$	$0.281^{\pm0.009}$	$9.871^{\pm0.084}$	$3.060^{\pm0.058}$	$2.294^{\pm0.058}$
SFM	$0.561^{\pm0.004}$	$0.749^{\pm0.003}$	$0.845^{\pm0.003}$	$0.066^{\pm0.003}$	$9.610^{\pm0.076}$	$2.758^{\pm0.086}$	$1.923^{\pm0.064}$
KIT-ML							
Ground Truth	$0.424^{\pm0.005}$	$0.649^{\pm0.006}$	$0.779^{\pm0.006}$	$0.031^{\pm0.004}$	$11.080^{\pm0.097}$	-	-
Seq2Seq [40]	$0.103^{\pm.003}$	$0.178^{\pm.005}$	$0.241^{\pm.006}$	$24.86^{\pm.348}$	$6.744^{\pm.106}$	$7.960^{\pm.031}$	-
T2G [3]	$0.156^{\pm.004}$	$0.255^{\pm.004}$	$0.338^{\pm.005}$	$12.12^{\pm.183}$	$9.334^{\pm.079}$	$6.964^{\pm.029}$	-
LJ2P [1]	$0.221^{\pm.005}$	$0.373^{\pm.004}$	$0.483^{\pm.005}$	$6.545^{\pm.072}$	$9.073^{\pm.100}$	$5.147^{\pm.030}$	-
Hier [8]	$0.255^{\pm.006}$	$0.432^{\pm.007}$	$0.531^{\pm.007}$	$5.203^{\pm.107}$	$9.563^{\pm.072}$	$4.986^{\pm.027}$	$2.090^{\pm.083}$
TEMOS [33]	$0.353^{\pm.006}$	$0.561^{\pm.007}$	$0.687^{\pm.005}$	$3.717^{\pm.051}$	$10.84^{\pm.100}$	$3.417^{\pm.019}$	$0.532^{\pm.034}$
MLD [5]	$0.390^{\pm0.008}$	$0.609^{\pm0.008}$	$0.734^{\pm0.007}$	$0.404^{\pm0.027}$	$10.800^{\pm0.117}$	$3.204^{\pm0.027}$	$2.192^{\pm0.071}$
MDM [43]	$0.164^{\pm0.004}$	$0.291^{\pm0.004}$	$0.396^{\pm0.004}$	$0.497^{\pm0.021}$	$10.847^{\pm0.119}$	$9.191^{\pm0.022}$	$1.907^{\pm0.214}$
ReMoDiffuse [51]	$0.427^{\pm0.014}$	$0.641^{\pm0.004}$	$0.765^{\pm0.055}$	$0.155^{\pm0.006}$	$10.800^{\pm0.105}$	$1.239^{\pm0.028}$	$1.239^{\pm0.028}$
MotionDiffuse [50]	$0.417^{\pm0.004}$	$0.621^{\pm0.004}$	$0.739^{\pm0.004}$	$1.954^{\pm0.062}$	$11.100^{\pm0.143}$	$2.958^{\pm0.005}$	$0.730^{\pm0.013}$
T2M-GPT [49]	$0.416^{\pm0.006}$	$0.627^{\pm0.006}$	$0.745^{\pm0.006}$	$0.514^{\pm0.029}$	$10.921^{\pm0.108}$	$3.007^{\pm0.023}$	$1.570^{\pm0.039}$
MotionGPT [19]	$0.366^{\pm0.005}$	$0.558^{\pm0.004}$	$0.680^{\pm0.005}$	$0.510^{\pm0.016}$	$10.350^{\pm0.084}$	$3.527^{\pm0.021}$	$2.328^{\pm0.117}$
MMM [38]	$0.404^{\pm0.005}$	$0.621^{\pm0.005}$	$0.744^{\pm0.004}$	$0.316^{\pm0.028}$	$10.910^{\pm0.101}$	$2.977^{\pm0.019}$	$1.232^{\pm0.039}$
MoMask [12]	$0.433^{\pm0.007}$	$0.656^{\pm0.005}$	$0.781^{\pm0.005}$	$0.204^{\pm0.011}$	-	$2.779^{\pm0.022}$	$1.131^{\pm0.043}$
StableMoFusion [18]	$0.445^{\pm0.006}$	$0.660^{\pm0.005}$	$0.782^{\pm0.004}$	$0.258^{\pm0.029}$	$10.936^{\pm0.077}$	$2.800^{\pm0.018}$	$1.362^{\pm0.062}$
Motion Mamba [53]	$0.419^{\pm0.006}$	$0.645^{\pm0.005}$	$0.765^{\pm0.006}$	$0.307^{\pm0.041}$	$11.02^{\pm0.098}$	$3.021^{\pm0.025}$	$1.678^{\pm0.064}$
SFM	$0.453^{\pm0.005}$	$0.680^{\pm0.005}$	$0.798^{\pm0.004}$	$0.228^{\pm0.024}$	$11.140^{\pm0.122}$	$2.548^{\pm0.034}$	$1.852^{\pm0.050}$

FID from 0.189 to 0.066. Compared with a Transformer baseline (MoMask), our model achieves higher accuracy (Top-3 0.845 vs. 0.807) and lower MM-Dist (2.758 vs. 2.958) while remaining competitive in FID (0.066 vs. 0.045). Relative to a naïve Mamba variant, we observe clear gains (Top-3 0.845 vs. 0.792, FID 0.066 vs. 0.281), indicating that our design better captures fine-grained dynamics. On KIT-ML, our method reaches the best R-Precision (Top-3 0.798) and competitive FID (0.228), with strong MM-Dist. Overall, these results support that the proposed spatial–frequency state-space design improves text–motion alignment and sequence fidelity.

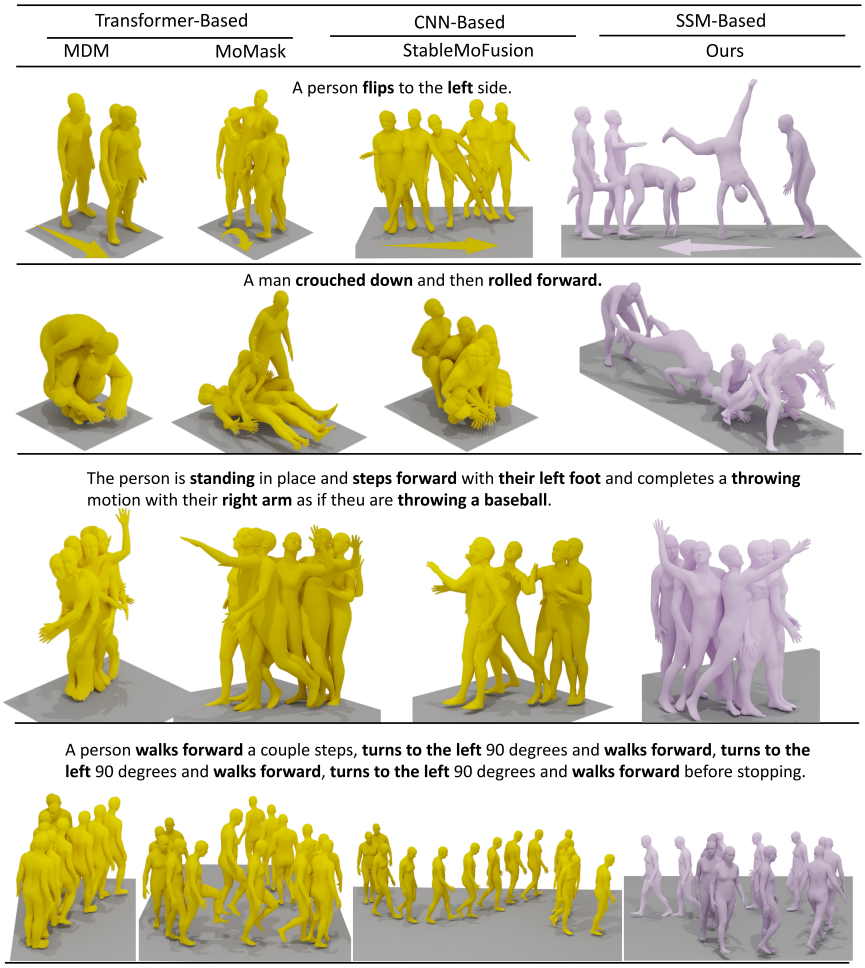

Fig. 3: Qualitative comparison on different prompts. Columns show representative Transformer (MDM, MoMask), CNN (StableMoFusion), and our SSM-based model under the same text. Our results follow directional phrases and execute multi-step actions with smooth transitions and contact-preserving details.

Qualitative results. Fig. 3 compares our model with state-of-the-art baselines from different paradigms under identical text conditions. For "A person flips to the left side," Transformer-based MDM and MoMask do not realize a clear flip, and the motion remains largely static or unstable; the CNN baseline StableMoFusion produces locomotion but disregards the directional phrase "left side." Our model shows stronger spatial awareness, completing a leftward flip with consistent take-off and landing orientation and clear limb articulation.

For the multi-action prompt "A man crouched down and then rolled forward," models must generate a coherent two-stage sequence. MoMask and StableMoFu-

Table 2: The average inference time per sentence (AITS) on HumanML3D dataset (A100).

Method	Arch.	AITS↓	Top-1↑	Top-2↑	Top-3↑	FID↓
MDM	Transformer	15.33	$0.320^{\pm0.005}$	$0.498^{\pm0.004}$	$0.611^{\pm0.007}$	$0.544^{\pm0.044}$
MoMask	Transformer	0.25	$0.521^{\pm0.002}$	$0.713^{\pm0.002}$	$0.807^{\pm0.002}$	$0.045^{\pm0.002}$
StableMoFusion	CNN	0.13	$0.499^{\pm0.004}$	$0.680^{\pm0.006}$	$0.779^{\pm0.007}$	$0.189^{\pm0.003}$
SFM	SSM	0.19	$0.561^{\pm0.004}$	$0.749^{\pm0.003}$	$0.845^{\pm0.003}$	$0.066^{\pm0.003}$

Table 3: Results on HumanML3D-LS long-sequence benchmark (>190 frames).

Method	Arch.	Top-1↑	Top-2↑	Top-3↑	FID↓
MLD	Transformer	$0.403^{\pm0.005}$	$0.584^{\pm0.005}$	$0.690^{\pm0.005}$	$0.952^{\pm0.020}$
MoMask	Transformer	$0.446^{\pm0.005}$	$0.623^{\pm0.005}$	$0.727^{\pm0.005}$	$0.125^{\pm0.006}$
StableMoFusion	CNN	$0.390^{\pm0.005}$	$0.565^{\pm0.005}$	$0.665^{\pm0.005}$	$0.442^{\pm0.02}$
Motion Mamba	SSM	$0.417^{\pm0.003}$	$0.606^{\pm0.003}$	$0.713^{\pm0.004}$	$0.668^{\pm0.019}$
SFM	SSM	$0.510^{\pm0.003}$	$0.704^{\pm0.004}$	$0.801^{\pm0.004}$	$0.190^{\pm0.009}$

sion fail to express the crouch phase reliably, and MDM loses the roll or breaks temporal continuity. Our method executes the *crouch* → *roll* sequence in the correct order with smooth phase transitions and contact cues (knees/shoulders to floor), yielding text-aligned and temporally coherent motion.

For more semantically complex multi-action prompts, existing methods either misinterpret directional cues or fail to synthesize step-by-step motions that follow the textual description. In contrast, SFM performs best on semantically rich long sequences, maintaining accurate semantic alignment and stable long-horizon spatial awareness and coherence. These visual results corroborate our quantitative improvements in text–motion alignment and spatial consistency. Additional visualizations are provided in the appendix.

Efficiency. To evaluate efficiency in practical applications, we report the average inference time per sentence (AITS) across SOTA methods of different architectures. As shown in Table 2, our model is substantially faster than a Transformer denoiser (0.19s vs.15.33s) while improving fidelity and accuracy(Top-3 0.845 vs. 0.611, FID 0.066 vs. 0.544). Compared with a CNN UNet (StableMoFusion), we obtain large gains in R-Precision (Top-3 0.845 vs. 0.779) with a modest increase in runtime (0.19s vs. 0.13s). Relative to a naïve Mamba variant, our design better captures fine-grained dynamics (Top-3 0.845 vs. 0.792, FID 0.066 vs. 0.281) while retaining linear-time scaling. Overall, the results indicate strong accuracy–efficiency trade-offs suitable for human motion generation.

HumanML3D-LS evaluation. We conduct experiments on HumanML3D-LS [53], a long-sequence benchmark consisting of motion sequences longer than 190 frames from the original HumanML3D test set. As shown in Table 3, SFM achieves the best R-Precision on the subset (Top-1 0.510, Top-2 0.704, Top-3 0.801) with a competitive FID (0.190), outperforming Motion Mamba (Top-3 0.713, FID 0.668) and StableMoFusion (Top-3 0.665, FID 0.442). These results indicate that our model can effectively capture long-range motion dependencies while preserving semantic consistency and temporal coherence. Additional SL-HumanML3D subset results in the appendix also show consistent trends.

Human evaluation. We conducted a blinded study on HumanML3D with 20 raters and 50 randomly sampled prompts. For each prompt, motions from MoMask, MMM, StableMoFusion, and our model were rendered (Motion Mamba excluded due to no public code); raters first judged semantic accuracy (yes/no) and then performed pairwise preference between our result and each baseline, with randomized order. As shown in Fig. 4, our method achieved the highest accuracy (75.2%) and the highest win rate (80.3%), indicating stronger text–motion alignment and temporal coherence, consistent with human preferences.

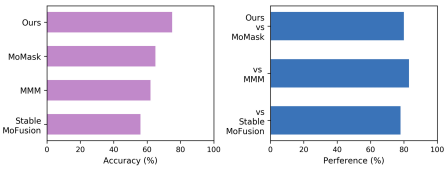

Fig. 4: Human evaluation results on HumanML3D. Left: semantic accuracy, measured as the percentage of "yes" responses, indicating whether the generated motion is semantically consistent with the input text for each method.Right: pairwise preference scores (win rates, %) comparing our method against each baseline (MoMask, MMM, and StableMoFusion). Bars show means over 20 raters.

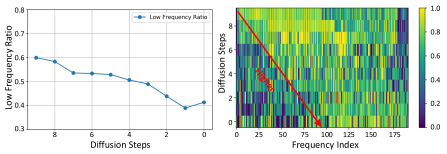

Fig. 5: Frequency modulation across diffusion steps on a sample prompt. Left: ratio of low-frequency components versus timestep t (9 = early, high noise; 0 = late, low noise). Right: heatmap of text-gated frequency masks by timestep (y) and frequency index (x, low $\rightarrow$ high). Early steps favor low frequencies; later steps progressively admit higher frequencies to refine details.

4.4 Ablation Study

Table 4: Ablation of core components on HumanML3D (± 95% CI).

Components	Top-1↑	Top-2↑	Top-3↑	FID↓
w/o F-SSM	$0.462^{\pm 0.006}$	$0.641^{\pm 0.005}$	$0.733^{\pm 0.005}$	$0.157^{\pm 0.003}$
w/o S-SSM	$0.512^{\pm 0.002}$	$0.708^{\pm 0.002}$	$0.779^{\pm 0.003}$	$0.104^{\pm 0.004}$
w/o Semantic Attention	$0.490^{\pm 0.005}$	$0.670^{\pm 0.005}$	$0.758^{\pm 0.005}$	$0.116^{\pm 0.003}$
w/o Text-Gated	$0.528^{\pm 0.007}$	$0.727^{\pm 0.006}$	$0.804^{\pm 0.006}$	$0.093^{\pm 0.004}$
Full Model	$0.561^{\pm 0.004}$	$0.749^{\pm 0.003}$	$0.845^{\pm 0.003}$	$0.066^{\pm 0.003}$

Analysis of core components. Table 4 reports ablations of SFM's key modules. Removing the Frequency State-Space Module (F-SSM) causes a large drop (Top-3 0.845→0.733; FID 0.066→0.157), showing its role in modeling semantically consistent multi-scale frequency cues. Removing the Spatial State-Space Module (S-SSM) also degrades performance (Top-3 0.845→0.779; FID 0.066→0.104), highlighting the importance of bidirectional spatial awareness and consistency. Removing either the F-SSM text-gated frequency modulation or the semantic attention applied after dual-axis fusion leads to a substantial drop (Top-3 0.845→0.727/0.670; FID 0.066→0.093/0.116), demonstrating that both components play a critical role in improving the model's interaction-following ability.

Analysis of frequency-path design. Table 5 compares *hard* DCT coefficient selection with *soft*, text-gated adaptation in the frequency pathway. While increasing k (randomly kept frequency coefficients) improves hard selection, the performance shows no significant gap as it gradually approaches the full setting (208). However, it remains markedly inferior to the soft text-gated design. This observation indicates that different frequency components contribute unequally, and adaptively modulating and selectively propagating the semantics-consistent frequencies is therefore crucial for enhancing motion accuracy and fidelity.

Table 5: Ablation of frequency-path design in F-SSM on HumanML3D (± 95% CI). Hard selection keeps the k DCT coefficients; soft adaptation applies text-gated filtering.

DCT Design		Top1↑	Top2↑	Top3↑	FID↓
Hard selection	k=32	$0.201^{\pm0.005}$	$0.290^{\pm0.005}$	$0.397^{\pm0.005}$	$12.520^{\pm0.004}$
	k=64	$0.359^{\pm0.004}$	$0.492^{\pm0.003}$	$0.587^{\pm0.003}$	$2.986^{\pm0.004}$
	k=128	$0.430^{\pm0.005}$	$0.646^{\pm0.004}$	$0.724^{\pm0.005}$	$1.642^{\pm0.004}$
	k=Full	$0.528^{\pm0.007}$	$0.727^{\pm0.006}$	$0.804^{\pm0.006}$	$0.093^{\pm0.004}$
Soft Adaptation	Text-Gated	$0.561^{\pm0.004}$	$0.749^{\pm0.003}$	$0.845^{\pm0.003}$	$0.066^{\pm0.003}$

Analysis of frequency transforms. DCT offers an ordered global frequency basis and exhibits strong energy compaction while remaining real-valued, which makes it stable and efficient for training. In contrast, FFT, DWT, or STFT introduce complex-valued representations, time–frequency localization effects, and additional sensitivity to window-length hyperparameters, which can complicate optimization. As shown in Table 6, DCT performs best, while all frequency transforms outperform the no-frequency baseline. Based on these observations, we adopt DCT as the frequency-domain transformation in our model.

Table 6: Comparison of frequency transforms on HumanML3D (± 95% CI).

Frequency Transforms	Top-1↑	Top-2↑	Top-3↑	FID↓
STFT	$0.517^{\pm0.005}$	$0.704^{\pm0.004}$	$0.796^{\pm0.004}$	$0.112^{\pm0.004}$
DWT	$0.552^{\pm0.004}$	$0.736^{\pm0.004}$	$0.834^{\pm0.003}$	$0.085^{\pm0.003}$
FFT	$0.541^{\pm0.005}$	$0.718^{\pm0.004}$	$0.815^{\pm0.004}$	$0.103^{\pm0.003}$
DCT	$0.561^{\pm0.004}$	$0.749^{\pm0.003}$	$0.845^{\pm0.003}$	$0.066^{\pm0.003}$

Analysis of frequency preference across diffusion. As shown in Fig. 5, we visualize text-gated frequency masks and the low-frequency energy ratio in the last block across the timesteps to examine how F-SSM allocates spectrum. We observe that during the denoising phase (step $9 \rightarrow 0$), the model initially focuses primarily on low-frequency features, with their relative contribution gradually decreasing (from 60% to 40%), while high-frequency features are predominantly optimized in the later stages. This phenomenon is consistent with similar observations in the text-to-image generation task [32]. For the prompt "the man started walking and turned around and started running and then stop quick.", early (high-noise) steps emphasize low frequencies that establish the global pose and phase (e.g., *walking*, *turning*); as denoising proceeds, once the overall motion framework is well established, focus gradually shifts to higher frequencies to refine contacts and limb articulations, producing smoother, more coherent transitions. The theoretical analysis and derivations are provided in the appendix.

5 Conclusion

We presented SFM, a text-to-motion framework that augments a UNet denoiser with task-aware state-space modeling along two complementary axes. The Frequency State-Space Module aligns global structure and local detail in the frequency domain, while the Spatial State-Space Module enhances spatial awareness and coherence over long horizons. This dual-axis design preserves linear-time scaling and yields text-aware, multi-scale features. Experiments on HumanML3D and KIT-ML show state-of-the-art and best performance–efficiency trade-offs.

Acknowledgements

This work was supported by the Science and Technology Development Fund of Macau [0007/2025/RIC, 0122/2024/RIB2, 0215/2024/AGJ, 0074/2025/AMJ, 001/2024/SKL, 0002/2025/EQP], the Research Services and Knowledge Transfer Office, University of Macau [SRG2023-00037-IOTSC, MYRG-GRG2024-00284-IOTSC], the Shenzhen-Hong Kong-Macau Science and Technology Program Category C [SGDX20230821095159012], the Science and Technology Planning Project of Guangdong [2025A0505010016], National Natural Science Foundation of China [52572354], the State Key Lab of Intelligent Transportation System [2024-B001], and the Jiangsu Provincial Science and Technology Program [BZ2024055].

References

1. Ahuja, C., Morency, L.P.: Language2pose: Natural language grounded pose forecasting. In: 2019 International conference on 3D vision (3DV). pp. 719–728. IEEE (2019)
2. Athanasiou, N., Petrovich, M., Black, M.J., Varol, G.: Teach: Temporal action composition for 3d humans. In: 2022 International Conference on 3D Vision (3DV). pp. 414–423. IEEE (2022)
3. Bhattacharya, U., Rewkowski, N., Banerjee, A., Guhan, P., Bera, A., Manocha, D.: Text2gestures: A transformer-based network for generating emotive body gestures for virtual agents. In: 2021 IEEE virtual reality and 3D user interfaces (VR). pp. 1–10. IEEE (2021)
4. Chen, W., Yu, K., Haozhe, J., Yuan, K., Huang, Z., Tian, B., Lai, S., Xiao, H., Zhang, E., Wang, L., et al.: Ant: Adaptive neural temporal-aware text-to-motion model. In: Proceedings of the 33rd ACM International Conference on Multimedia. pp. 9852–9861 (2025)
5. Chen, X., Jiang, B., Liu, W., Huang, Z., Fu, B., Chen, T., Yu, G.: Executing your commands via motion diffusion in latent space. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 18000–18010 (2023)
6. Dabral, R., Mughal, M.H., Golyanik, V., Theobalt, C.: Mofusion: A framework for denoising-diffusion-based motion synthesis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9760–9770 (2023)
7. Dai, W., Chen, L.H., Wang, J., Liu, J., Dai, B., Tang, Y.: Motionlcm: Real-time controllable motion generation via latent consistency model. In: European Conference on Computer Vision. pp. 390–408. Springer (2024)

8. Ghosh, A., Cheema, N., Oguz, C., Theobalt, C., Slusallek, P.: Synthesis of compositional animations from textual descriptions. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1396–1406 (2021)
9. Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. In: First conference on language modeling (2024)
10. Gu, A., Goel, K., Ré, C.: Efficiently modeling long sequences with structured state spaces. arXiv preprint arXiv:2111.00396 (2021)
11. Gu, A., Johnson, I., Goel, K., Saab, K., Dao, T., Rudra, A., Ré, C.: Combining recurrent, convolutional, and continuous-time models with linear state space layers. Advances in neural information processing systems **34**, 572–585 (2021)
12. Guo, C., Mu, Y., Javed, M.G., Wang, S., Cheng, L.: Momask: Generative masked modeling of 3d human motions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1900–1910 (2024)
13. Guo, C., Zou, S., Zuo, X., Wang, S., Ji, W., Li, X., Cheng, L.: Generating diverse and natural 3d human motions from text. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5152–5161 (2022)
14. Guo, C., Zuo, X., Wang, S., Zou, S., Sun, Q., Deng, A., Gong, M., Cheng, L.: Action2motion: Conditioned generation of 3d human motions. In: Proceedings of the 28th ACM international conference on multimedia. pp. 2021–2029 (2020)
15. Guo, H., Guo, Y., Zha, Y., Zhang, Y., Li, W., Dai, T., Xia, S.T., Li, Y.: Mambairv2: Attentive state space restoration. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 28124–28133 (2025)
16. Guo, H., Li, J., Dai, T., Ouyang, Z., Ren, X., Xia, S.T.: Mambair: A simple baseline for image restoration with state-space model. In: European conference on computer vision. pp. 222–241. Springer (2024)
17. Hatamizadeh, A., Kautz, J.: Mambavision: A hybrid mamba-transformer vision backbone. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 25261–25270 (2025)
18. Huang, Y., Yang, H., Luo, C., Wang, Y., Xu, S., Zhang, Z., Zhang, M., Peng, J.: Stablemofusion: Towards robust and efficient diffusion-based motion generation framework. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 224–232 (2024)
19. Jiang, B., Chen, X., Liu, W., Yu, J., Yu, G., Chen, T.: Motiongpt: Human motion as a foreign language. Advances in Neural Information Processing Systems **36**, 20067–20079 (2023)
20. Kalman, R.E.: A new approach to linear filtering and prediction problems (1960)
21. Karunratanakul, K., Preechakul, K., Suwajanakorn, S., Tang, S.: Guided motion diffusion for controllable human motion synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2151–2162 (2023)
22. Li, K., Li, X., Wang, Y., He, Y., Wang, Y., Wang, L., Qiao, Y.: Videomamba: State space model for efficient video understanding. In: European conference on computer vision. pp. 237–255. Springer (2024)
23. Li, Z., Cui, Z., Liao, H., Ash, J., Zhang, G., Xu, C., Wang, Y.: Steering the future: Redefining intelligent transportation systems with foundation models. Chain **1**(1), 46–53 (2024)
24. Liang, D., Zhou, X., Xu, W., Zhu, X., Zou, Z., Ye, X., Tan, X., Bai, X.: Pointmamba: A simple state space model for point cloud analysis. Advances in neural information processing systems **37**, 32653–32677 (2024)
25. Liao, H., Li, Z., Wang, C., Wang, B., Kong, H., Guan, Y., Li, G., Cui, Z., Xu, C.: A cognitive-driven trajectory prediction model for autonomous driving in mixed autonomy environment. arXiv preprint arXiv:2404.17520 (2024)

26. Liu, Y., Tian, Y., Zhao, Y., Yu, H., Xie, L., Wang, Y., Ye, Q., Jiao, J., Liu, Y.: Vmamba: Visual state space model. Advances in neural information processing systems **37**, 103031–103063 (2024)
27. Lu, C., Zhou, Y., Bao, F., Chen, J., Li, C., Zhu, J.: Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. Machine Intelligence Research **22**(4), 730–751 (2025)
28. Ma, J., Li, F., Wang, B.: U-mamba: Enhancing long-range dependency for biomedical image segmentation. arXiv preprint arXiv:2401.04722 (2024)
29. Mahmood, N., Ghorbani, N., Troje, N.F., Pons-Moll, G., Black, M.J.: Amass: Archive of motion capture as surface shapes. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 5442–5451 (2019)
30. Meng, Z., Xie, Y., Peng, X., Han, Z., Jiang, H.: Rethinking diffusion for text-driven human motion generation. arXiv preprint arXiv:2411.16575 (2024)
31. Nguyen, E., Goel, K., Gu, A., Downs, G., Shah, P., Dao, T., Baccus, S., Ré, C.: S4nd: Modeling images and videos as multidimensional signals with state spaces. Advances in neural information processing systems **35**, 2846–2861 (2022)
32. Ning, M., Li, M., Su, J., Jia, H., Liu, L., Beneš, M., Chen, W., Salah, A.A., Ertugrul, I.O.: Dctdiff: Intriguing properties of image generative modeling in the dct space. arXiv preprint arXiv:2412.15032 (2024)
33. Petrovich, M., Black, M.J., Varol, G.: Temos: Generating diverse human motions from textual descriptions. In: European Conference on Computer Vision. pp. 480–497. Springer (2022)
34. Phung, H., Dao, Q., Dao, T., Phan, H., Metaxas, D., Tran, A.: Dimsum: Diffusion mamba–a scalable and unified spatial-frequency method for image generation. arXiv preprint arXiv:2411.04168 (2024)
35. Pinyoanuntapong, E., Saleem, M.U., Karunratanakul, K., Wang, P., Xue, H., Chen, C., Guo, C., Cao, J., Ren, J., Tulyakov, S.: Controlmm: Controllable masked motion generation. arXiv preprint arXiv:2410.10780 (2024)
36. Pinyoanuntapong, E., Saleem, M.U., Wang, P., Lee, M., Das, S., Chen, C.: Bamm: Bidirectional autoregressive motion model. In: European Conference on Computer Vision. pp. 172–190. Springer (2024)
37. Pinyoanuntapong, E., Saleem, M.U., Wang, P., Lee, M., Das, S., Chen, C.: Bamm: Bidirectional autoregressive motion model. In: European Conference on Computer Vision. pp. 172–190. Springer (2024)
38. Pinyoanuntapong, E., Wang, P., Lee, M., Chen, C.: Mmm: Generative masked motion model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1546–1555 (2024)
39. Plappert, M., Mandery, C., Asfour, T.: The kit motion-language dataset. Big data **4**(4), 236–252 (2016)
40. Plappert, M., Mandery, C., Asfour, T.: Learning a bidirectional mapping between human whole-body motion and natural language using deep recurrent neural networks. Robotics and Autonomous Systems **109**, 13–26 (2018)
41. Qin, S., Wang, J., Zhou, Y., Chen, B., Luo, T., An, B., Dai, T., Xia, S., Wang, Y.: Mambavc: Learned visual compression with selective state spaces. arXiv preprint arXiv:2405.15413 (2024)
42. Smith, J.T., Warrington, A., Linderman, S.W.: Simplified state space layers for sequence modeling. arXiv preprint arXiv:2208.04933 (2022)
43. Tevet, G., Raab, S., Gordon, B., Shafir, Y., Cohen-Or, D., Bermano, A.H.: Human motion diffusion model. arXiv preprint arXiv:2209.14916 (2022)

44. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
45. Wang, J., Zhu, W., Wang, P., Yu, X., Liu, L., Omar, M., Hamid, R.: Selective structured state-spaces for long-form video understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6387–6397 (2023)
46. Xing, Z., Ye, T., Yang, Y., Liu, G., Zhu, L.: Segmamba: Long-range sequential modeling mamba for 3d medical image segmentation. In: International conference on medical image computing and computer-assisted intervention. pp. 578–588. Springer (2024)
47. Yan, J.N., Gu, J., Rush, A.M.: Diffusion models without attention. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8239–8249 (2024)
48. Yuan, Y., Song, J., Iqbal, U., Vahdat, A., Kautz, J.: Physdiff: Physics-guided human motion diffusion model. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 16010–16021 (2023)
49. Zhang, J., Zhang, Y., Cun, X., Huang, S., Zhang, Y., Zhao, H., Lu, H., Shen, X.: T2m-gpt: Generating human motion from textual descriptions with discrete representations. arXiv preprint arXiv:2301.06052 (2023)
50. Zhang, M., Cai, Z., Pan, L., Hong, F., Guo, X., Yang, L., Liu, Z.: Motiondiffuse: Text-driven human motion generation with diffusion model. IEEE transactions on pattern analysis and machine intelligence **46**(6), 4115–4128 (2024)
51. Zhang, M., Guo, X., Pan, L., Cai, Z., Hong, F., Li, H., Yang, L., Liu, Z.: Remodiffuse: Retrieval-augmented motion diffusion model. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 364–373 (2023)
52. Zhang, Z., Liu, A., Chen, Q., Chen, F., Reid, I., Hartley, R., Zhuang, B., Tang, H.: Infinimotion: Mamba boosts memory in transformer for arbitrary long motion generation. arXiv preprint arXiv:2407.10061 (2024)
53. Zhang, Z., Liu, A., Reid, I., Hartley, R., Zhuang, B., Tang, H.: Motion mamba: Efficient and long sequence motion generation. In: European Conference on Computer Vision. pp. 265–282. Springer (2024)
54. Zhu, L., Liao, B., Zhang, Q., Wang, X., Liu, W., Wang, X.: Vision mamba: Efficient visual representation learning with bidirectional state space model. arXiv preprint arXiv:2401.09417 (2024)