MaTMotion: Text-Driven Human Motion Generation Based on the Mamba-Transformer

Feng Zhou, Yanrui Sun
North China University of Technology

Ju Dai*
Peng Cheng Laboratory

Shihao Zou
Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences

Sen-Zhe Xu
University of Science and Technology Beijing

Wei Zhou, Yu-Kun Lai, Paul L. Rosin
Cardiff University

Abstract

Text-based 3D human motion generation typically employs quantization methods to discretize continuous motion data into a sequence of discrete tokens, enabling the use of powerful sequence-to-sequence models. However, existing approaches, especially those using residual quantization, are generally limited by their single processing architecture. This architecture applies a homogeneous processing strategy to all residual levels, ignoring the differences in information density and task focus across different levels. To address this limitation, we propose MaTMotion, an innovative framework that integrates a hybrid network with layer-aware feature enhancement. First, we design a MambaTrans hybrid backbone network that leverages the strengths of Mamba in modeling temporal continuity and the Transformer's ability to capture global relationships, thereby enhancing residual modeling of motion frame sequences. Second, we introduce a hierarchical adaptive feature enhancement mechanism. The mechanism consists of two modules: an adaptive frame weighting module, serving as a preprocessing step, that dynamically assigns weights to motion frames based on the current prediction residual level, and a multi-scale feature-fusion module, acting as a post-processing step, that employs multiple parallel convolutional blocks to extract motion information across different time scales. Extensive quantitative and qualitative experiments on the HumanML3D and KIT-ML datasets validate the effectiveness of this approach.

Keywords: Text to Motion, Motion Synthesis, Residual Vector Quantization, Residual Transformer, Attention

1. Introduction

With the rapid evolution of multimodal foundation models, the task of generating 3D human motion from textual descriptions has attracted growing attention. Due to its significant role in a wide range of applications, including computer graphics [3, 27], virtual reality [25], and digital entertainment [19], it has become a pivotal yet highly challenging research problem.

Existing approaches can be broadly grouped into two paradigms. The first relies on diffusion models [4, 16, 30, 35], which learn to iteratively denoise latent representations, thereby producing intricate and photorealistic motion sequences. The second adopts Vector Quantization (VQ) [12, 36, 41], which converts continuous motion data into discrete token sequences for modeling, effectively reducing the complexity of the overall generation task.

MotionDiffuse [42] pioneered the application of diffusion model-based text-driven motion generation. It introduces a cross-modality linear transformer that can synthesize motions of an arbitrary length depending on the motion duration. However, the method incurs substantial computational cost. To mitigate this, MLD [4] adopts a VAE (Variational Autoencoder) to compress motion sequences into compact latent codes, thereby reducing the computational burden. However, even after this compression, a large number of denoising steps are still required to obtain the final motion. To alleviate this deficiency, MoMask [10] employs Residual Vector Quantization (RVQ) techniques [40] and

*Corresponding author: daij@pcl.ac.cn

generative masked transformers for text-to-motion synthesis. The key idea is to decompose the complex motion generation task into a hierarchical, coarse-to-fine refinement process, where each residual layer progressively corrects and enriches the motion produced by the preceding layer. This hierarchical design significantly improves the quality and realism of the generated motion.

Nevertheless, despite the remarkable success achieved by hierarchical refinement strategies, we argue that their underlying architectures still leave room for improvement in two key aspects. First, existing methods [10, 41] typically employ a single residual processing module to predict residuals at all levels, often resulting in suboptimal modeling efficiency. Second, they do not adequately adapt to the inherent differences among residuals at different levels. Specifically, when predicting low-level residuals, the model needs to capture and integrate long-term global historical motion sequences to establish the overall motion layout and basic trajectory. To predict high-level residuals and add details, the model must capture fine dynamic changes across a few key frames, which requires stronger local perception capabilities than are provided in the current methods. However, existing methods are somewhat inadequate in handling this aspect, which restricts the model's ability to express features at multiple scales.

To address the aforementioned issues, this paper proposes a hierarchical-aware hybrid residual refinement architecture, which combines Mamba [9] and Transformer [37]. Unlike a single residual, this architecture employs an ensemble of specialized sequence models that are optimized together. This approach both strengthens the sequential modeling capacity for residual prediction and also naturally accommodates the heterogeneous statistics of each residual level. Additionally, to satisfy the distinct requirements of different residual levels, we present a tailored preprocessing and post-processing method that consists of adaptive frame weighting and multi-scale feature fusion modules. These modules enhance the multi-scale expression capability of the hierarchical data. This architecture and its modules are the main contributions of this paper. In particular:

- We introduce MaTMotion, a novel hybrid framework that integrates Mamba's capability in modeling temporal continuity with Transformer's strength in capturing global dependencies.

- Within this framework, we further propose adaptive frame weighting and multi-scale feature fusion modules. They can dynamically pre-process the input information and post-process the output features when predicting residuals at different levels.

- Extensive experiments are conducted on two widely adopted benchmarks, HumanML3D and KIT-ML, and the results show that MaTMotion outperforms other methods in terms of FID evaluation metric, while also exhibiting good time efficiency.

2. Related Work

2.1. Continuous Regression Motion Generation

Early human motion generation approaches predominantly follow the continuous regression paradigm, in which an end-to-end network directly predicts the continuous values of a motion sequence. Within these approaches, several strategies have been explored. For instance, Petrovich et al. [31] and Guo et al. [11] leverage a Variational Auto-Encoder (VAE) [22] to jointly encode text and motion into a shared latent space, enabling unified probabilistic modeling. Lin et al. [23] and Plappert et al. [33] investigate Recurrent Neural Networks [7], while Malek-Podjaski and Deligianni [29] adopt Generative Adversarial Networks [8]. In contrast, Tevet et al. [34] and Lin et al. [24] use Transformer architectures [37] to improve generation quality. Recently, many works [35, 43] adopt a diffusion model [16], which has substantially advanced the continuous generation paradigm and achieved current state-of-the-art results. For instance, ReMoDiffuse [43] introduces a hybrid retrieval mechanism that fuses semantic and kinematic similarities to provide high-quality references for the denoising process. PhysDiff [38] incorporates physical constraints to enhance the plausibility of the generated motions, and Zhang et al. [42] add fine-grained text comprehension to align the generated motions more closely with the text input.

Although continuous regression can, in theory, reconstruct the text-to-motion mapping without information loss, in practice, it faces a fundamental challenge: regressing high-dimensional, continuous motion sequences that encode complex skeletal articulations requires large-scale, domain-specific motion pre-training datasets. Learning the intricate mapping from sparse data often yields suboptimal results.

2.2. Quantization-based Motion Generation

To alleviate the inherent difficulties of continuous regression, a more mainstream alternative discretizes the continuous motion signal into a sequence of discrete tokens, thereby transforming the high-dimensional regression task into a more tractable classification problem. This paradigm typically employs a Vector-Quantized Variational Auto-Encoder (VQ-VAE) [36] as the foundational encoder, mapping an input motion sequence into a set of discrete motion tokens. Drawing on the success of natural language processing, researchers have also devised multiple strategies for predicting these token indices [12, 41]. Inspired by BERT [5], Zhu et al. [47] propose a strategy to randomly

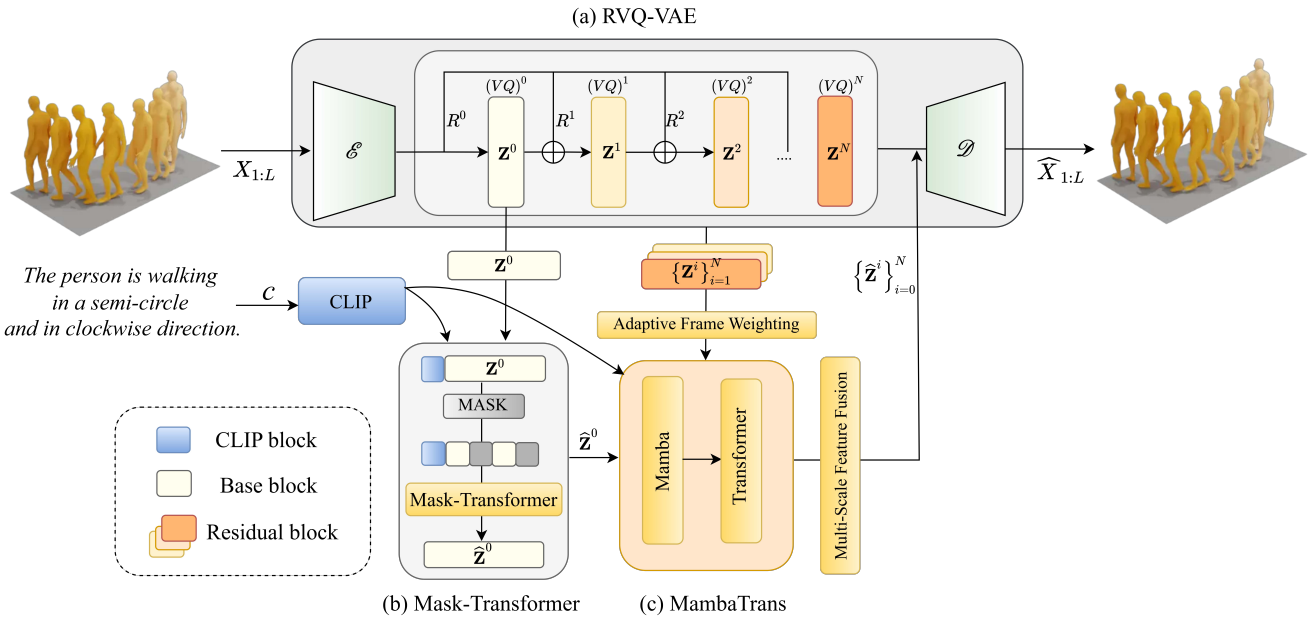

Figure 1. **The Pipeline of MaTMotion.** It mainly consists of three parts: (a) RVQ-VAE encodes the input motion $\mathbf{X}_{1:L}$ into a base token $\mathbf{Z}^0$ for the global structure and a set of residual tokens $\{\mathbf{Z}^i\}_{i=1}^N$ for fine-grained details. (b) Mask-Transformer then predicts the base token $\hat{\mathbf{Z}}^0$, conditioned on a text prompt $\mathbf{C}$. (c) MambaTrans, featuring a novel hybrid Mamba-Transformer architecture, progressively generates the residual tokens $\{\hat{\mathbf{Z}}^i\}_{i=1}^N$. Finally, all predicted tokens are passed to the RVQ-VAE decoder $\mathcal{D}$ to reconstruct the motion $\hat{\mathbf{X}}_{1:L}$.

mask and then predict tokens in parallel to improve generation efficiency. Austin et al. [1] leverage the denoising philosophy of diffusion models in the discrete token space. With the development of Large Language Models (LLMs), some works unify multi-modal control signals into discrete prompting instructions and leverage the emergent capabilities of LLMs to directly generate motion responses [21]. Notably, generative masked-modeling frameworks such as MaskGIT [2] and MoMask [10] break the unidirectional limitation of autoregressive models. These works demonstrate that combining residual vector quantization (RVQ-VAE) [40] with time-step varying mask rate scheduling can naturally decompose complex actions into multi-layer representations with intrinsic hierarchical structures, enabling the model to generate high-quality samples in parallel and iteratively in a non-autoregressive mode.

Although quantization-based methods, which allow models to circumvent the difficulty of directly regressing high-dimensional continuous actions, have greatly improved motion generation, we have observed that existing quantization methods have a limitation that is often overlooked: they usually quantize a whole frame of complete body posture into a single token. This "whole frame quantization" loses the joint space association information contained in the original data, making it difficult for subsequent sequence models to effectively capture and utilize the spatial structure information distributed across frames that is crucial for motion understanding and generation.

2.3. Mamba-based methods

Mamba [9] has a linear computational complexity, achieved through its selective scanning mechanism and hardware-friendly algorithm. This allows the algorithm to perform well in processing long sequences. This breakthrough quickly inspired a series of works that extended Mamba's capabilities to vision and motion generation. However, in general, it is difficult to effectively adapt a model designed for one-dimensional, unidirectional language tasks to visual or motion data with complex spatiotemporal dependencies. To address this, extensive studies have proposed different solutions. Zhu et al. [46] introduce a bidirectional SSM formulation to capture richer global context, at the expense of the additional computational overhead in both the training and the inference due to this bidirectional strategy, while Liu et al. [26] approach this problem by presenting a Cross-Scan Module (CSM) that traverses tokens along four diagonal directions. This method is tailored for vision, and the receptive field is still bounded by the predefined paths, potentially leaving certain long-range interactions unmodeled. Hatamizadeh and Kautz [15] address this issue by combining Mamba with a Transformer instead of using a complex scanning mode, demonstrating that it is remarkably effective at capturing long-range spatial dependencies using a hybrid architecture rather than a single architecture. Our approach also leverages the advantages of Mamba and Transformer to solve the core challenges in hierarchical residual prediction tasks for

motion generation.

3. Methodology

3.1. Overview

Our goal is to generate a 3D human motion sequence $\mathbf{X}_{1:L} \in \mathbb{R}^{L \times D}$ guided by the CLIP-encoded text condition $\mathbf{c}$, where L denotes the sequence length and D denotes the feature dimension. The overall framework of our MaTMotion is as illustrated in Figure 1. Our framework consists of three principal components: RVQ-VAE for residual vectorized representation and motion reconstruction (section 3.2), a masked transformer for base motion-token prediction (section 3.3), and the proposed MambaTrans module together with its tightly-coupled pre- and post-processing modules for progressive residual-token prediction(section3.4).

3.2. RVQ-VAE

The RVQ-VAE uses an encoder-decoder architecture designed to map a continuous input motion sequence, $\mathbf{X}_{1:L}$, into a hierarchical set of discrete tokens and subsequently reconstruct it. The core of this architecture consists of three components: an initial encoder, a symmetric decoder, and a cascade of $N+1$ residual-connected VQ quantizers.

The encoding process begins with an encoder $\mathcal{E}$, which maps $\mathbf{X}_{1:L}$ into a low-dimensional, information-rich, continuous latent representation $\mathbf{R}^0$.

$$\mathbf{R}^0 = \mathcal{E}(\mathbf{X}_{1:L}). \tag{1}$$

Subsequently, $\mathbf{R}^0$ is passed through $N+1$ residual quantizers to be decomposed into a hierarchical set of discrete tokens. This is achieved iteratively: at each layer i, the quantizer $(VQ)^i$ maps the current residual input $\mathbf{R}^i$ to a discrete token $\mathbf{Z}^i$. The following layer then receives a new, finer-grained residual, computed as the difference between the current residual and its discrete token. The final output is a set of discrete tokens, $\{\mathbf{Z}^i\}_{i=0}^N$. The entire quantization process is defined as:

$$\begin{aligned} \mathbf{Z}^i &= (VQ)^i(\mathbf{R}^i), \quad \text{for } i = 0, \ldots, N, \\ \mathbf{R}^{i+1} &= \mathbf{R}^i - \mathbf{Z}^i, \quad \text{for } i = 0, \ldots, N-1. \end{aligned} \tag{2}$$

The objective of this iterative process is to find $\{\mathbf{Z}^i\}_{i=0}^N$, whose element-wise sum serves as a close approximation of the original latent representation $\mathbf{R}^0$.

The decoder $\mathcal{D}$ is responsible for converting the hierarchical discrete tokens $\{\mathbf{Z}^i\}_{i=0}^N$ back into a continuous motion sequence $\hat{\mathbf{X}}_{1:L}$. The decoding process first performs an element-wise summation of all tokens. This aggregated representation, which approximates $\mathbf{R}^0$, is then passed through the decoder to produce the reconstructed motion sequence $\hat{\mathbf{X}}_{1:L}$. The decoding process is summarized as:

$$\hat{\mathbf{X}}_{1:L} = \mathcal{D}\left(\sum_{i=0}^{N} \mathbf{Z}^i\right). \tag{3}$$

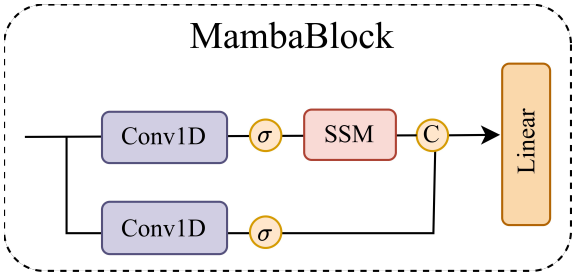

Figure 2. **Illustration of the MambaBlock.** It employs a dual-path architecture to capture both long-range and local information.

3.3. Mask-Transformer

The Mask-Transformer is responsible for masked-token modeling on the base-layer motion tokens $\mathbf{Z}^0$, producing a structurally complete and semantically coherent motion outline. We adopt a standard bidirectional Transformer architecture identical to BERT [5].

During training, we first obtain the ground-truth base-layer tokens $\mathbf{Z}^0$ from the RVQ-VAE encoder. Similar to works like MaskGIT [2] and MoMask [10], we then randomly replace a proportion of these tokens with a special mask token. Conditioned on the text embedding $\mathbf{c}$, the model's objective is to predict the original tokens at the masked positions. The model's output, denoted as $\hat{\mathbf{Z}}^0$, represents the predicted probability distribution for these masked tokens.

3.4. MambaTrans

The MambaTrans module aims to progressively model the residual motion tokens $\{\hat{\mathbf{Z}}^i\}_{i=1}^N$ in a layer-wise fashion, thereby gradually superimposing fine motion details onto the base outline provided by $\hat{\mathbf{Z}}^0$. During training for a target residual layer t, the model receives three inputs:

- the sum of all previously predicted tokens, i.e., $\sum_{i=0}^{t-1} \hat{\mathbf{Z}}^i, t > 0$;
- the semantic condition $\mathbf{c}$ from CLIP;
- the current layer index t.

The model then predicts the motion token $\hat{\mathbf{Z}}^t$ for the target layer. After such N steps, the complete set of residual tokens $\{\hat{\mathbf{Z}}^i\}_{i=1}^N$ is obtained.

Our MambaTrans is a heterogeneous, stage-wise processor designed to leverage the complementary strengths of Mamba and Transformer architectures. It consists of 12 layers in total: the first 8 layers are Mamba blocks for efficient

sequential processing, while the final 4 layers are standard Transformer blocks, which excel at modeling global relationships.

The architecture of the Mamba block is illustrated in Figure 2. We enhance the standard Mamba by adding a symmetric, purely convolutional branch that operates in parallel to the main State Space Model (SSM) branch. The outputs from these two paths are concatenated and projected via a linear layer. This dual-path design ensures the representation incorporates both long-range sequential dependencies and local patterns. The process is defined as follows. First, the input to the first block, $\mathbf{F}_{\text{in}}$, is constructed:

$$\mathbf{F}_{\text{in}} = \text{Concat}\left(\sum_{i=0}^{t-1} \hat{\mathbf{Z}}^i, \mathbf{c}, \text{Emb}(t)\right), \quad t > 0. \tag{4}$$

Then, Mamba block processes $\mathbf{F}_{\text{in}}$ to produce output $\mathbf{F}_{\text{out}}$:

$$\begin{aligned} \mathbf{F}_1 &= \text{Scan}(\sigma(\text{Conv}(\mathbf{F}_{\text{in}}))), \\ \mathbf{F}_2 &= \sigma(\text{Conv}(\mathbf{F}_{\text{in}})), \\ \mathbf{F}_{\text{out}} &= \text{Linear}(\text{Concat}(\mathbf{F}_1, \mathbf{F}_2)), \end{aligned} \tag{5}$$

where Conv and Concat represent 1D convolution and concatenation operations, Scan is the selective scan operation as in [9], and σ is the activation function for which SiLU is used. Linear is a linear projection. Separately, $\text{Emb}(t)$ refers to an Embedding layer that converts the scalar layer index t into a learnable feature vector.

The output from these 8 Mamba blocks is then passed to the 4 Transformer blocks. Each block employs a standard Transformer architecture [37], with its core component being the multi-head self-attention. This mechanism allows the model to refine the features it receives from the Mamba stage by modeling the global relationships across the entire sequence, which is crucial for generating coherent and physically plausible motion.

Adaptive Frame Weighting. Before Mamba and Transformer blocks process the data, Adaptive Frame Weighting (AFW) is leveraged to perform layer-aware feature recalibration for the input feature, as illustrated in Figure 3. It processes $\mathbf{F}_{\text{in}}$ through a lightweight MLP [18], followed by a Sigmoid function, to generate a dynamic weight matrix $\mathbf{W}_t$. This matrix represents the frame-wise salience for the current prediction task. The original input features are then modulated by these weights, effectively amplifying relevant historical information and suppressing noise. This refined feature tensor, $\mathbf{F}'_{\text{in}}$, is then passed to the first Mamba block. The process is defined as:

$$\begin{aligned} \mathbf{W}_t &= \text{Sigmoid}(\text{MLP}(\mathbf{F}_{\text{in}})), \\ \mathbf{F}'_{\text{in}} &= \mathbf{F}_{\text{in}} \odot \mathbf{W}_t. \end{aligned} \tag{6}$$

This pre-processing step enables MambaTrans to adaptively focus on the most critical temporal frames for each residual layer, significantly enhancing predictive accuracy.

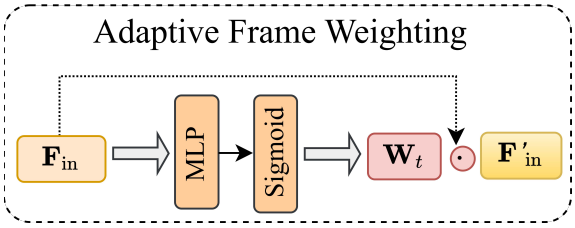

Figure 3. **Illustration of AFW module.** It processes $\mathbf{F}_{\text{in}}$ using an MLP and a Sigmoid function to compute dynamic weights $\mathbf{W}_t$. Through element-wise multiplication, these weights recalibrate $\mathbf{F}_{\text{in}}$, selectively amplifying the most salient frames.

Multi-scale Feature Fusion. After MambaTrans predicts the complete set of motion tokens $\{\hat{\mathbf{Z}}^i\}_{i=0}^N$, the Multi-scale Feature Fusion (MSFF) module is applied as a post-processing step, as illustrated in Figure 4. It refines each token $\hat{\mathbf{Z}}^i$ by adaptively re-weighting its own representation based on multi-scale temporal features.

For each token, its multi-scale features are extracted through a set of parallel 1D convolutions with unique kernel sizes ($k \in \{1, 3, 5, 7\}$), aiming to capture a range of temporal patterns:

- **Small kernels** ($k = 1, 3$)**:** Capture instantaneous pose details and short-term motion dynamics.
- **Large kernels** ($k = 5, 7$)**:** Model broader, mid- to long-term motion rhythms.

These extracted features are then processed via an attention network [37] and a Softmax function to compute a set of normalized attention scores, $\{\mathbf{A}_k\}$. These scores dynamically assess the importance of each temporal scale for the given token. The final refined token $\hat{\mathbf{Z}}^{i'}$ is produced by an attention-weighted sum, where these scores are applied back to the original input token:

$$\hat{\mathbf{Z}}^{i'} = \sum_{k \in \{1,3,5,7\}} (\mathbf{A}_k \odot \hat{\mathbf{Z}}^i). \tag{7}$$

This self-adaptive refinement provides the decoder with a richer, context-aware representation for each motion component, ultimately enhancing the coherence and realism of the final generated motion.

4. Experiments

4.1. Datasets

To verify MaTMotion, we conduct empirical evaluations on two widely used motion-language benchmarks, namely HumanML3D and KIT-ML.

HumanML3D is a large-scale, highly diverse dataset that aggregates 14,616 motion sequences from AMASS [28] and HumanAct12 [13]. It is currently one of

Table 1. Performance comparison of various methods across multiple metrics on HumanML3D dataset. The best results are in **bold**, and the second best results are underlined. ↓ means the lower is better while ↑ means the higher is better.

Methods	R-Precision ↑			FID↓	MultiModal Dist↓	MultiModality↑	AITS ↓
	Top1↑	Top2↑	Top3↑				
Diffusion-based Models							
MotionDiffuse[42]	$0.491^{\pm.001}$	$0.681^{\pm.001}$	$0.782^{\pm.001}$	$0.630^{\pm.011}$	$3.113^{\pm.001}$	$1.553^{\pm.072}$	10.89
MDM[14]	$0.320^{\pm.005}$	$0.498^{\pm.004}$	$0.611^{\pm.007}$	$0.544^{\pm.044}$	$5.556^{\pm.027}$	$\mathbf{2.799}^{\pm.072}$	18.20
MLD[4]	$0.481^{\pm.003}$	$0.673^{\pm.003}$	$0.772^{\pm.002}$	$0.473^{\pm.013}$	$3.196^{\pm.010}$	$2.413^{\pm.079}$	0.22
ReMoDiffuse[42]	$0.510^{\pm.005}$	$0.698^{\pm.006}$	$0.795^{\pm.004}$	$0.103^{\pm.004}$	$2.974^{\pm.016}$	$1.795^{\pm.043}$	-
DiverseMotion[39]	$0.515^{\pm.003}$	$0.706^{\pm.002}$	$0.802^{\pm.002}$	$0.072^{\pm.004}$	$2.941^{\pm.007}$	$1.869^{\pm.089}$	-
StableMoFusion[20]	$\mathbf{0.553}^{\pm.003}$	$\mathbf{0.748}^{\pm.002}$	$\mathbf{0.841}^{\pm.002}$	$0.098^{\pm.003}$	$\mathbf{2.770}^{\pm.006}$	$1.774^{\pm.051}$	-
B2A-HDM[45]	$0.511^{\pm.002}$	$0.699^{\pm.002}$	$0.791^{\pm.002}$	$0.084^{\pm.004}$	$3.020^{\pm.010}$	$1.914^{\pm.078}$	-
MotionMamba[44]	$0.502^{\pm.003}$	$0.693^{\pm.002}$	$0.792^{\pm.002}$	$0.281^{\pm.010}$	$3.036^{\pm.005}$	$2.294^{\pm.058}$	-
VQ-VAE-based Models							
TM2T[12]	$0.457^{\pm.002}$	$0.639^{\pm.003}$	$0.740^{\pm.003}$	$1.067^{\pm.020}$	$3.340^{\pm.011}$	$2.090^{\pm.083}$	0.76
T2M-GPT[41]	$0.492^{\pm.003}$	$0.679^{\pm.002}$	$0.775^{\pm.002}$	$0.141^{\pm.005}$	$3.121^{\pm.009}$	$1.831^{\pm.048}$	0.38
MoMask[10]	$0.521^{\pm.002}$	$0.713^{\pm.002}$	$0.807^{\pm.002}$	$\underline{0.045}^{\pm.002}$	$2.958^{\pm.008}$	$1.241^{\pm.040}$	0.12
MMM[6]	$0.502^{\pm.002}$	$0.692^{\pm.004}$	$0.788^{\pm.006}$	$0.053^{\pm.007}$	$3.037^{\pm.003}$	$0.810^{\pm.023}$	-
BAD[17]	$0.504^{\pm.002}$	$0.696^{\pm.003}$	$0.794^{\pm.003}$	$0.080^{\pm.003}$	$2.998^{\pm.008}$	$1.164^{\pm.044}$	-
BAMM[32]	$\underline{0.525}^{\pm.002}$	$\underline{0.720}^{\pm.003}$	$\underline{0.814}^{\pm.003}$	$0.055^{\pm.002}$	$\underline{2.919}^{\pm.008}$	$1.687^{\pm.051}$	-
Ours	$0.518^{\pm.003}$	$0.712^{\pm.002}$	$0.804^{\pm.002}$	$\mathbf{0.039}^{\pm.001}$	$2.958^{\pm.006}$	$1.273^{\pm.048}$	0.13

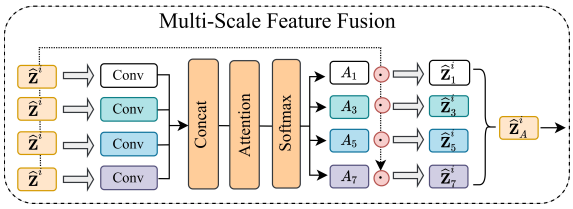

Figure 4. **Illustration of MSFF module.** The input is processed by parallel Conv1D branches with varying kernels to extract multi-scale temporal features. An attention mechanism then computes fusion weights $\mathbf{A}_k$. The output $\mathbf{Z}^{i'}$ is an attention-weighted sum, achieving adaptive information fusion across different time scales.

the largest and most diverse datasets that jointly contains 3-D human motion and natural-language descriptions. It consists of 14,616 actions and 44,970 descriptions consisting of 5,371 different words, with an average motion length of 7.1 seconds and an average description length of 12 words. Each motion clip has 3-4 descriptions, the sequence is downsampled to 20 frames per second, and each clip lasts 2-10 seconds.

KIT-ML is a large-scale and open dataset that associates human actions with natural language annotations, designed to support the development and evaluation of systems for generating semantic representations of human and robot activities. It contains 3,911 actions, a total duration of 11.23 hours, and 6,278 natural language annotations (about 52,903 words). It is relatively small in size compared to the HumanML3D dataset and can be used as a benchmark for evaluating the generalization ability of models under limited data conditions.

4.2. Evaluations

Following T2M [11], we leverage the Frèchet Inception Distance (FID), R-Precision (Top-1, 2, 3), Multimodal Distance (MM-Dist), Multimodality, and AITS (Average Inference Time per Sentence) to evaluate model performance. All metrics are computed in the shared embedding space produced by a pre-trained text–motion encoder.

- **FID:** The core indicator of generation quality, FID measures both the "realism" and "diversity" of generated actions. It is computed as the distance between the feature distributions of generated actions and real actions, with lower values indicating higher quality.
- **R-Precision:** This metric evaluates semantic alignment between generated actions and input text. For a given text query, it calculates the proportion of correct matches among the top-R retrieved results, where R equals the number of ground-truth actions associated with the query. Higher R-Precision values indicate stronger text and motion matching.

- **MM-Dist:** Defined as the Euclidean distance between paired text and motion embeddings, MM-Dist directly measures semantic closeness. Smaller distances correspond to stronger alignment.

- **Multimodality:** This metric captures the diversity of generated actions conditioned on the same input text. It assesses how many distinct yet valid motions can be produced for a single prompt, thereby reflecting the model's ability to generate varied outputs.

- **AITS:** This metric quantifies a model's generation efficiency. It is computed as the average time in seconds required to synthesize a complete motion sequence from a single text prompt. This measurement explicitly excludes model and data loading overhead to isolate the core inference process. Lower AITS values signify superior efficiency.

4.3. Implementation Details

All our experiments are conducted on a single NVIDIA 3090 GPU in which the models were implemented in PyTorch. The residual vector quantization autoencoder (RVQ-VAE) consisted of 6 quantization layers, each codebook contained 512 codewords, and the dimension of each codeword was 512. Here, MambaTrans consists of 8 hybrid blocks, including 6 Mamba modules and 2 Transformer modules. For the Mamba module, its internal expansion ratio was set to 4, the SSM state expansion factor was set to 32, and the internal convolution kernel size was set to 5.

We tailor specific training hyperparameters for different datasets. On the HumanML3D dataset, the optimizer parameters are set to $\beta_1 = 0.9$, $\beta_2 = 0.95$, and weight decay to 0.01. The initial learning rate is set to 2×10^{-4} with a batch size of 64. On the KIT-ML dataset, we adopt a more aggressive regularization strategy to mitigate overfitting risks. Specifically, we increase the dropout and drop_path_rate in the hybrid modules to 0.5, and raise the weight decay to 0.1. Additionally, we use a more conservative initial learning rate of 1×10^{-4} and adjust the batch size to 32. All models are trained using the AdamW [4] optimizer.

4.4. Comparison with SOTA Methods

We compare MaTMotion against state-of-the-art methods in text to motion, including TM2T[12], T2M-GPT[41], MotionDiffuse [42], MDM [14], as well as additional baselines detailed in Table 1.

Quantitative Comparisons. Following MoMask [10], we systematically and quantitatively evaluate our model, and the reported values represent the mean with a 95% confidence interval. The quantitative assessment primarily focuses on the model's Top3 R-Precision, FID, MultiModal Distance, and MultiModality. Comparison results presented in Table 1 and Table 2 demonstrate that our model obtains the lowest scores in FID.

The experimental results indicate that our method has achieved significant breakthroughs in the core quality metric FID, outperforming all previous methods. Specifically, the core metric FID dropped to 0.039 (approximately 13.3% improvement over MoMask) on HumanML3D and decreased to 0.174 (approximately 14.7% improvement over MoMask) on KIT-ML. We attribute this leap in performance to the innovative architecture of our hybrid backbone. The frontend Mamba module, engineered with a dual-path parallel mechanism, effectively mitigates the long-range information decay that hampers traditional Transformers in long-sequence tasks. This allows for the preservation of fine-grained temporal details. Subsequently, the backend Transformer stage leverages these high-fidelity features to refine global motion coherence. It is this powerful synergy that significantly enhances the physical plausibility of the generated actions, directly contributing to the marked reduction in FID. Additionally, we observed a slight decrease in the semantic alignment metric R-Precision compared to MoMask on both HumanML3D and KIT-ML datasets. We argue that this reflects a strategic prioritization of kinematic realism and structural integrity over simple token-to-text matching. While high R-Precision indicates better discrete token alignment, it does not always correlate with the physical smoothness and temporal consistency required for natural human motion. Our significant improvement in FID demonstrates that MaTMotion generates more realistic and physically plausible motions, which is a critical requirement for high-quality synthesis.

Qualitative Comparisons. To more intuitively demonstrate the superiority of our model in terms of generation quality and text understanding, we provide qualitative comparison results of our method with current state-of-the-art models, namely MLD [4] and MoMask [10], on three different text descriptions, as shown in Figure 5. For instance, in the text description "A person is pushed from the side and stumbles, but regains their balance.", the motions generated by MoMask and MLD appear more like an unnatural, rigid lateral shift, which fails to capture the prompt; "stumbles, but regains their balance". In the example of "A man walks forward with both hands above head.", MoMask completes the basic motion, while its ability to maintain stability during leg movement is relatively insufficient. For the example of "The person is walking in a semi-circle and in a clockwise direction, " MLD incorrectly exhibits counterclockwise motion, indicating a significant deviation from the intended textual semantics.

In contrast, our method demonstrates a good understanding of text details and physical transitions. For: "The person is walking in a semi-circle and in a clockwise direc-

Table 2. Performance comparison of various methods across multiple metrics on the KIT-ML dataset. The best results are in **bold**, and the second best results are underlined. ↓ means the lower is better while ↑ means the higher is better.

Methods	R-Precision ↑			FID↓	MultiModal Dist↓	MultiModality↑	AITS↓
	Top1↑	Top2↑	Top3↑				
Diffusion-based Models							
MotionDiffuse[42]	$0.417^{\pm.004}$	$0.621^{\pm.004}$	$0.739^{\pm.004}$	$1.954^{\pm.062}$	$2.958^{\pm.005}$	$0.730^{\pm.013}$	10.89
MDM[14]	$0.164^{\pm.004}$	$0.291^{\pm.004}$	$0.396^{\pm.004}$	$0.497^{\pm.021}$	$9.191^{\pm.022}$	$1.907^{\pm.214}$	18.20
MLD[4]	$0.390^{\pm.008}$	$0.609^{\pm.008}$	$0.734^{\pm.007}$	$0.404^{\pm.027}$	$3.204^{\pm.027}$	$2.192^{\pm.071}$	0.22
ReMoDiffuse[43]	$0.427^{\pm.014}$	$0.641^{\pm.004}$	$0.765^{\pm.055}$	$\mathbf{0.155}^{\pm.006}$	$2.814^{\pm.012}$	$1.239^{\pm.028}$	-
DiverseMotion[39]	$0.416^{\pm.005}$	$0.637^{\pm.008}$	$0.760^{\pm.011}$	$0.468^{\pm.098}$	$2.892^{\pm.041}$	$2.062^{\pm.079}$	-
StableMoFusion[20]	$\mathbf{0.445}^{\pm.006}$	$0.660^{\pm.005}$	$0.782^{\pm.004}$	$0.258^{\pm.029}$	–	$1.362^{\pm.062}$	-
B2A-HDM[45]	$0.436^{\pm.006}$	$0.653^{\pm.006}$	$0.773^{\pm.005}$	$0.367^{\pm.020}$	$2.946^{\pm.024}$	$1.291^{\pm.047}$	-
MotionMamba[44]	$0.419^{\pm.006}$	$0.645^{\pm.005}$	$0.765^{\pm.006}$	$0.307^{\pm.041}$	$3.021^{\pm.025}$	$1.678^{\pm.064}$	-
VQ-VAE-based Models							
TM2T[12]	$0.280^{\pm.005}$	$0.463^{\pm.006}$	$0.587^{\pm.005}$	$3.599^{\pm.153}$	$4.591^{\pm.026}$	$\mathbf{3.292}^{\pm.081}$	0.76
T2M-GPT[41]	$0.416^{\pm.006}$	$0.627^{\pm.006}$	$0.745^{\pm.006}$	$0.514^{\pm.029}$	$3.007^{\pm.023}$	$1.570^{\pm.039}$	0.38
MoMask[10]	$0.433^{\pm.007}$	$0.656^{\pm.005}$	$0.781^{\pm.005}$	$0.204^{\pm.011}$	$2.779^{\pm.022}$	$1.131^{\pm.043}$	0.12
MMM[6]	$0.404^{\pm.005}$	$0.621^{\pm.005}$	$0.744^{\pm.004}$	$0.316^{\pm.028}$	$2.977^{\pm.019}$	$1.232^{\pm.039}$	-
BAD[17]	$0.417^{\pm.006}$	$0.631^{\pm.006}$	$0.750^{\pm.006}$	$0.221^{\pm.012}$	$2.941^{\pm.025}$	$1.170^{\pm.047}$	-
BAMM[32]	$0.438^{\pm.009}$	$0.661^{\pm.009}$	$\mathbf{0.788}^{\pm.005}$	$0.183^{\pm.013}$	$\mathbf{2.723}^{\pm.026}$	$1.609^{\pm.065}$	-
Ours	$0.416^{\pm.006}$	$\mathbf{0.663}^{\pm.007}$	$0.756^{\pm.007}$	$0.174^{\pm.009}$	$2.819^{\pm.018}$	$1.252^{\pm.047}$	0.13

tion.", the method accurately generates a smooth circular trajectory, matching the description of the spatial path in the text. In the "hands above head" task, the character not only moved forward but also maintained a very stable posture and natural gait, proving the strong capability of our hybrid architecture in coordinating different parts while maintaining physical stability. In the complex instruction "A person is pushed from the side and stumbles, but regains their balance.", our method vividly reproduces the complete physical process from losing balance to regaining a steady state with significantly reduced foot sliding compared to baselines. These results highlight that MaTMotion is particularly beneficial for actions involving complex spatial constraints and intricate coordination, as Mamba's efficient temporal modeling ensures long-range coherence that Transformers often struggle to maintain.

4.5. Ablation Experiments

In this section, we conduct ablation studies to comprehensively validate the impact of each component of our model on the overall performance. The overall results are shown in Table 3.

Effect of AFW and MSFF Modules. We evaluate the Adaptive Frame Weighting (AFW) and Multi-Scale Feature Fusion (MSFF) modules by comparing the model's performance with and without these modules, in order to gauge their individual contribution to hierarchical perception. Results show that adding AFW and MSFF modules reduces the FID from 0.045 to 0.043 (approximately a 4.4% decrease), while the R-Precision metric remains at the same level. This clearly verifies our core hypothesis: the AFW module focuses on "pre-emptive concentration" of hierarchical perception of input information, and the MSFF module further refines the output features through multi-scale processing, effectively enhancing the final prediction accuracy.

Effect of MambaTrans Module. To validate the effectiveness of the MambaTrans hybrid network, we replace the MambaTrans hybrid network in our model with a single Transformer network (w/o MambaTrans). The experimental results show significant improvements: the FID is reduced from 0.043 to 0.039, and the Multimodality is increased from 1.216 to 1.273. Beyond these quantitative gains, we attribute this performance leap to the architectural advantage of the frontend Mamba module. Its dual-path parallel mechanism effectively mitigates the long-range information decay that typically hampers traditional Transformers in long-sequence tasks. These results fully demonstrate that MambaTrans can more deeply understand and model complex motion sequences compared to a pure Transformer, especially in capturing the temporal dynamics and diversity of motion, highlighting its strong potential as

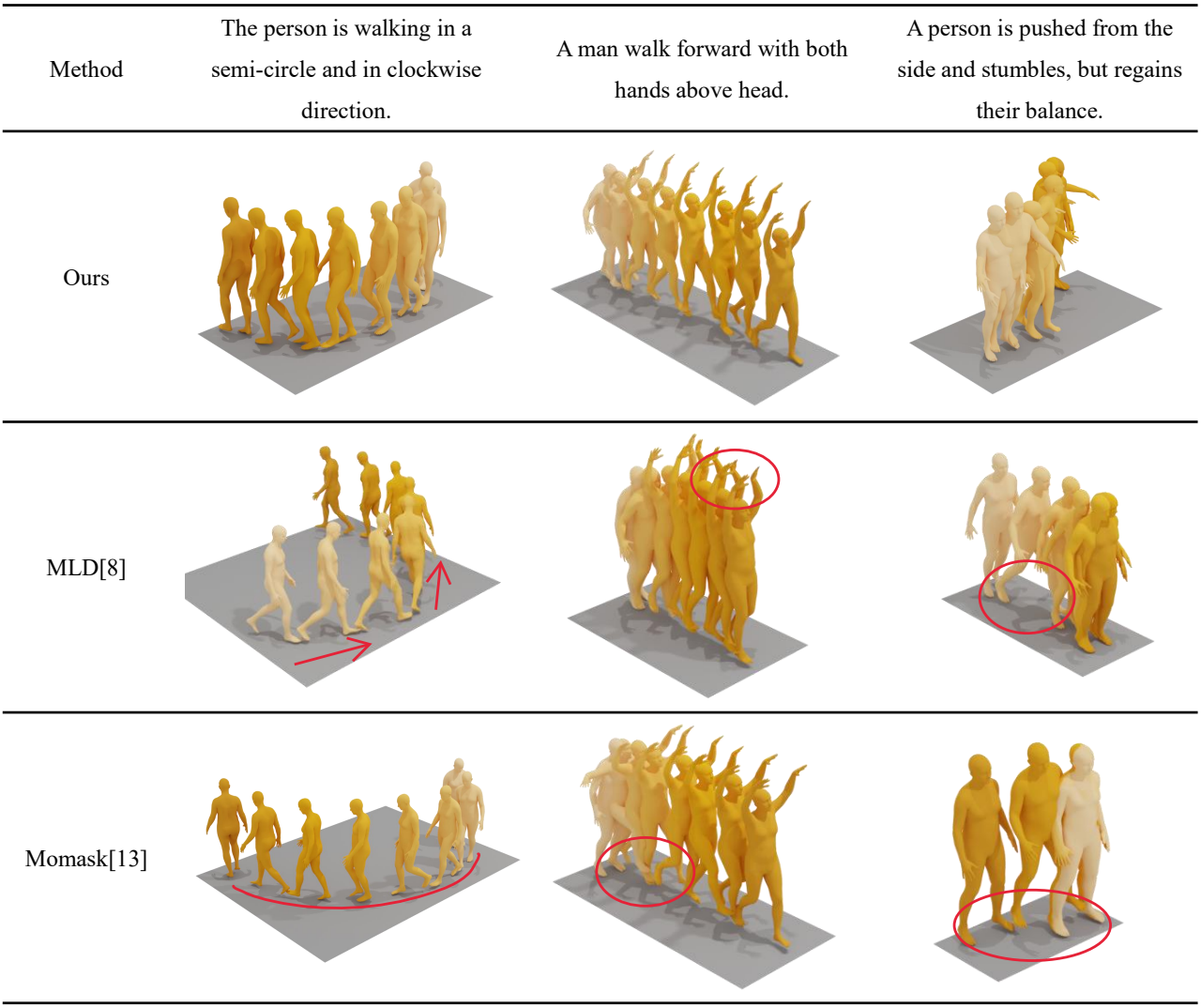

Figure 5. Visual comparisons between the different methods given three distinct text descriptions from the HumanML3D testset.The red circles and arrows in the figure are used to highlight typical failure cases in the baseline model, such as inaccurate motion trajectory, unstable upper limb posture, and non-physical foot contact.

Table 3. Ablation studies with different modules added on the HumanML3D dataset.

Methods	R-Precision ↑			FID↓	MultiModal Dist↓	MultiModality↑
	Top1↑	Top2↑	Top3↑			
w/o MambaTrans	**0.520**$^{\pm.002}$	**0.717**$^{\pm.002}$	**0.812**$^{\pm.002}$	0.043$^{\pm.001}$	3.037$^{\pm.004}$	1.216$^{\pm.050}$
w/o AFW&MSFF	0.516$^{\pm.002}$	0.710$^{\pm.001}$	0.805$^{\pm.002}$	0.041$^{\pm.002}$	2.998$^{\pm.004}$	1.272$^{\pm.036}$
Ours	0.518$^{\pm.003}$	0.712$^{\pm.002}$	0.804$^{\pm.002}$	**0.039**$^{\pm.001}$	**2.958**$^{\pm.006}$	**1.273**$^{\pm.048}$

a motion sequence processing engine.

To visually verify the effectiveness of each module in our framework, we further conduct qualitative analysis. As shown in Figure 6, we select two representative text descriptions and visualize the results of three model configurations in the ablation study: (a) w/o AFW & MSFF, (b) w/o MambaTrans, (c) Ours.

Scenario 1 ("*A person sidesteps continuously to the right for several paces.*"): Although (a) conforms to the semantics, it lacks stability and fluidity in body movement compared to our method (c), demonstrating the direct impact of the AFW and MSFF modules on understanding fine mo-

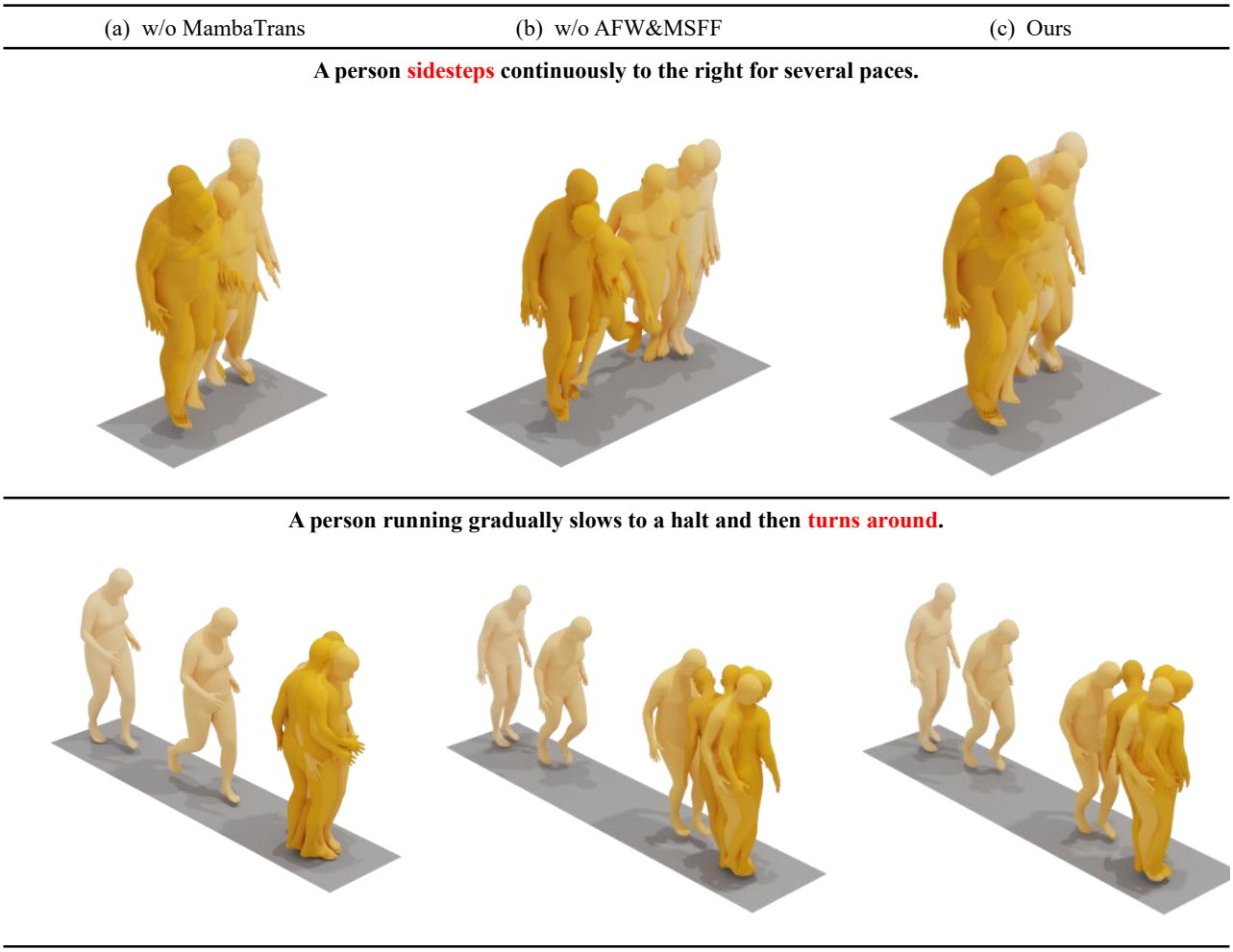

Figure 6. Visual comparisons with and without our proposed three modules given two distinct text descriptions from the HumanML3D testset.

tion details. (b) Fails to accurately capture the key phrase "sidesteps continuously to the right," resulting in stiff and disjointed motion despite its competitive R-Precision score. In contrast, our model (c) accurately performs the sidestepping motion, generating smoother and more stable motion, effectively demonstrating that high R-Precision in baselines does not necessarily guarantee physical realism.

Scenario 2 ("*A person running gradually slows to a halt and then turns around.*"): For this text description, (b) only shows a simple deceleration, appearing unnatural for the critical phase of "turning", our method (c) successfully performs the turning motion, proving its strong capability in capturing and modeling long-range motion sequences and complex motion transitions which are often poorly reflected by static R-Precision metric.

5. Conclusion and Future Work

This paper addresses the issue of singular feature processing mechanisms in existing hierarchical residual methods, which struggle to adapt to the needs of different hierarchical tasks. We propose an innovative solution, MaTMotion, that integrates the MambaTrans and a hierarchically-aware refinement AFW and MSFF modules. By combining Mamba's temporal modeling capabilities with Transformer's global relationship capturing abilities, and leveraging adaptive frame weighting and multi-scale fusion for dynamic feature pre- and post-processing, we significantly enhance the quality and realism of 3D human motion generation. Experimental results demonstrate promising performance on the HumanML3D and KIT-ML datasets, validating the effectiveness of our proposed heterogeneous network design and hierarchical adaptive strategy.

However, despite the fidelity and faithfulness in generating high-quality results, there is still room for further explo-

ration in handling complex interactive tasks and computational efficiency. MaTMotion primarily focuses on single-character motion generation, and its generation capabilities for complex scenes requiring precise physical interactions or multi-person collaboration have not been fully verified. Therefore, future work can focus on extending the framework to interactive scenarios to generate more realistic, complex multi-person motions and on investigating model lightweighting methods to optimize computational efficiency.

Acknowledgments

This work is supported by R&D Program of Beijing Municipal Education Commission (KM202310009002), Humanities and Social Science Fund of Ministry of Education (24YJCZH458) and Yuxiu Innovation Project of NCUT (2024NCUTYXCX202).

References

[1] J. Austin, D. D. Johnson, J. Ho, D. Tarlow, and R. Van Den Berg. Structured denoising diffusion models in discrete state-spaces. *Advances in neural information processing systems*, 34:17981–17993, 2021. 3

[2] H. Chang, H. Zhang, L. Jiang, C. Liu, and W. T. Freeman. MaskGIT: Masked generative image transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11315–11325, 2022. 3, 4

[3] L.-H. Chen, J. Zhang, Y. Li, Y. Pang, X. Xia, and T. Liu. HumanMAC: Masked motion completion for human motion prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9544–9555, 2023. 1

[4] X. Chen, B. Jiang, W. Liu, Z. Huang, B. Fu, T. Chen, and G. Yu. Executing your commands via motion diffusion in latent space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18000–18010, 2023. 1, 6, 7, 8

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 4171–4186, 2019. 2, 4

[6] M. L. Ekkasit Pinyoanuntapong, Pu Wang and C. Chen. MMM: Generative masked motion model. In *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, page 1546–1555, 2024. 6, 8

[7] K. Fragkiadaki, S. Levine, P. Felsen, and J. Malik. Recurrent network models for human dynamics. In *Proceedings of the IEEE international conference on computer vision*, pages 4346–4354, 2015. 2

[8] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 2

[9] A. Gu and T. Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023. 2, 3, 5

[10] C. Guo, Y. Mu, M. G. Javed, S. Wang, and L. Cheng. MoMask: Generative masked modeling of 3d human motions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1900–1910, 2024. 1, 2, 3, 4, 6, 7, 8

[11] C. Guo, S. Zou, X. Zuo, S. Wang, W. Ji, X. Li, and L. Cheng. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5152–5161, 2022. 2, 6

[12] C. Guo, X. Zuo, S. Wang, and L. Cheng. TM2T: Stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts. In *European Conference on Computer Vision*, pages 580–597. Springer, 2022. 1, 2, 6, 7, 8

[13] C. Guo, X. Zuo, S. Wang, S. Zou, Q. Sun, A. Deng, M. Gong, and L. Cheng. Action2motion: Conditioned generation of 3d human motions. In *Proceedings of the ACM international conference on multimedia*, pages 2021–2029, 2020. 5

[14] B. G. Y. S. D. C.-o. Guy Tevet, Sigal Raab and A. H. Bermano. Human motion diffusion model. In *International Conference on Learning Representations*, 2023. 6, 7, 8

[15] A. Hatamizadeh and J. Kautz. MambaVision: A hybrid mamba-transformer vision backbone. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 25261–25270, 2025. 3

[16] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 1, 2

[17] S. R. Hosseyni, A. A. Rahmani, S. J. Seyed-Mohammadi, S. Seyedin, and A. Mohammadi. BAD: bidirectional auto-regressive diffusion for text-to-motion generation. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1–5, 2025. 6, 8

[18] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 5

[19] C. Huang, C.-E. Lin, Z. Yang, Y. Kong, P. Chen, X. Yang, and K.-T. Cheng. Learning to film from professional human motion videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4244–4253, 2019. 1

[20] Y. Huang, Y. Hui, C. Luo, Y. Wang, S. Xu, Z. Zhang, M. Zhang, and J. Peng. StableMoFusion: Towards robust and efficient diffusion-based motion generation framework. In *In Proceedings of the ACM International Conference on Multimedia*, page 224–232, 2024. 6, 8

[21] B. Jiang, X. Chen, W. Liu, J. Yu, G. Yu, and T. Chen. MotionGPT: Human motion as a foreign language. *Advances in Neural Information Processing Systems*, 36:20067–20079, 2023. 3

[22] D. P. Kingma, M. Welling, et al. Auto-encoding variational bayes, 2013. 2

[23] A. S. Lin, L. Wu, R. Corona, K. Tai, Q. Huang, and R. J. Mooney. Generating animated videos of human activities

from natural language descriptions. In *Proceedings of the NeurIPS Workshop on Visually Grounded Interaction and Language*, 2018. 2

[24] J. Lin, J. Chang, L. Liu, G. Li, L. Lin, Q. Tian, and C.-w. Chen. Being comes from not-being: Open-vocabulary text-to-motion generation with wordless training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 23222–23231, 2023. 2

[25] J. Lin, A. Zeng, S. Lu, Y. Cai, R. Zhang, H. Wang, and L. Zhang. Motion-X: A large-scale 3d expressive whole-body human motion dataset. *Advances in Neural Information Processing Systems*, 36:25268–25280, 2023. 1

[26] Y. Liu, Y. Tian, Y. Zhao, H. Yu, L. Xie, Y. Wang, Q. Ye, J. Jiao, and Y. Liu. VMamba: Visual state space model. *Advances in neural information processing systems*, 37:103031–103063, 2024. 3

[27] K. Lyu, H. Chen, Z. Liu, B. Zhang, and R. Wang. 3d human motion prediction: A survey. *Neurocomputing*, 489:345–365, 2022. 1

[28] N. Mahmood, N. Ghorbani, N. F. Troje, G. Pons-Moll, and M. J. Black. Amass: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5442–5451, 2019. 5

[29] M. Malek-Podjaski and F. Deligianni. Adversarial attention for human motion synthesis. In *IEEE Symposium Series on Computational Intelligence*, pages 69–74, 2023. 2

[30] A. Q. Nichol and P. Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pages 8162–8171, 2021. 1

[31] M. Petrovich, M. J. Black, and G. Varol. TEMOS: Generating diverse human motions from textual descriptions. In *European Conference on Computer Vision*, pages 480–497, 2022. 2

[32] E. Pinyoanuntapong, M. U. Saleem, P. Wang, M. Lee, S. Das, and C. Chen. BAMM: bidirectional autoregressive motion model. In *European Conference on Computer Vision*, volume 15073, pages 172–190, 2024. 6, 8

[33] M. Plappert, C. Mandery, and T. Asfour. 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. 2

[34] G. Tevet, B. Gordon, A. Hertz, A. H. Bermano, and D. Cohen-Or. MotionCLIP: Exposing human motion generation to CLIP space. In *European Conference on Computer Vision*, pages 358–374, 2022. 2

[35] G. Tevet, S. Raab, B. Gordon, Y. Shafir, D. Cohen-Or, and A. H. Bermano. Human motion diffusion model. In *International Conference on Learning Representations*, 2023. 1, 2

[36] A. Van Den Oord, O. Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 1, 2

[37] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2, 5

[38] Y. Yuan, J. Song, U. Iqbal, A. Vahdat, and J. Kautz. PhysDiff: Physics-guided human motion difffusion model. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 16010–16021, 2023. 2

[39] Y. W. X. W. Yunhong Lou, Linchao Zhu and Y. Yang. DiverseMotion: Towards diverse human motion generation via discrete diffusion. *arXiv preprint arXiv:2309.01372*, 2023. 6, 8

[40] N. Zeghidour, A. Luebs, A. Omran, J. Skoglund, and M. Tagliasacchi. SoundStream: An end-to-end neural audio codec. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:495–507, 2021. 1, 3

[41] J. Zhang, Y. Zhang, X. Cun, Y. Zhang, H. Zhao, H. Lu, X. Shen, and Y. Shan. Generating human motion from textual descriptions with discrete representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14730–14740, 2023. 1, 2, 6, 7, 8

[42] M. Zhang, Z. Cai, L. Pan, F. Hong, X. Guo, L. Yang, and Z. Liu. MotionDiffuse: Text-driven human motion generation with diffusion model. *IEEE transactions on pattern analysis and machine intelligence*, 46(6):4115–4128, 2024. 1, 2, 6, 7, 8

[43] M. Zhang, X. Guo, L. Pan, Z. Cai, F. Hong, H. Li, L. Yang, and Z. Liu. ReMoDiffuse: Retrieval-augmented motion diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 364–373, 2023. 2, 8

[44] Z. Zhang, A. Liu, I. Reid, R. Hartley, B. Zhuang, and H. Tang. Motion Mamba: Efficient and long sequence motion generation. In *European Conference on Computer Vision*, pages 265–282, 2025. 6, 8

[45] X. G. Z. S. W. Y. Zhenyu Xie, Yang Wu and X. Liang. Towards detailed text-to-motion synthesis via basic-to-advanced hierarchical diffusion model. In *In Proceedings of the AAAI Conference on Artificial Intelligence*, page 6252–6260, 2024. 6, 8

[46] L. Zhu, B. Liao, Q. Zhang, X. Wang, W. Liu, and X. Wang. Vision Mamba: Efficient visual representation learning with bidirectional state space model. 2024. 3

[47] W. Zhu, X. Ma, Z. Liu, L. Liu, W. Wu, and Y. Wang. MotionBERT: A unified perspective on learning human motion representations. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15085–15099, 2023. 2