Text-controlled Motion Mamba: Text-Instructed Temporal Grounding of Human Motion

Xinghan Wang
Peking University
Beijing, China
xinghan_wang@pku.edu.cn

Zixi Kang
Peking University
Beijing, China
forever.kzx0713@stu.pku.edu.cn

Yadong Mu*
Peking University
Beijing, China
myd@pku.edu.cn

ABSTRACT

Human motion understanding is a fundamental task with diverse practical applications, facilitated by the availability of large-scale motion capture datasets. Recent studies focus on text-motion tasks, such as text-based motion generation, editing and question answering. In this study, we introduce the novel task of text-based human motion grounding (THMG), aimed at precisely localizing temporal segments corresponding to given textual descriptions within untrimmed motion sequences. Capturing global temporal information is crucial for the THMG task. However, transformer-based models that rely on global temporal self-attention face challenges when handling long untrimmed sequences due to the quadratic computational cost. We address these challenges by proposing Text-controlled Motion Mamba (TM-Mamba), a unified model that integrates temporal global context, language query control, and spatial graph topology with only linear memory cost. The core of the model is a text-controlled selection mechanism which dynamically incorporates global temporal information based on text query. The model is further enhanced to be topology-aware through the integration of relational embeddings. For evaluation, we introduce BABEL-Grounding, the first text-motion dataset that provides detailed textual descriptions of human actions along with their corresponding temporal segments. Extensive evaluations demonstrate the effectiveness of TM-Mamba on BABEL-Grounding.

CCS CONCEPTS

• **Computing methodologies → Computer vision**.

KEYWORDS

Human Motion Analysis, Temporal Grounding, Mamba

1 INTRODUCTION

Human motion understanding is a crucial task with a wide range of applications. Recent years have witnessed the flourishing of large-scale motion capture databases [31, 51, 52, 65], which greatly facilitate the end-to-end training of various motion-related tasks. Based on these databases, recent studies [21, 26, 41, 43, 58, 68] have augmented motion datasets with textual annotations. These annotated motion-text pairs enable a range of text-motion tasks that require a joint understanding of human motion and language, such as text-based motion generation [56, 69, 83, 85], motion captioning [22, 33], text-based motion editing [15, 37, 86], and motion question answering [12]. However, in real-world scenarios, semantic actions often occur sparsely within lengthy motion sequences.

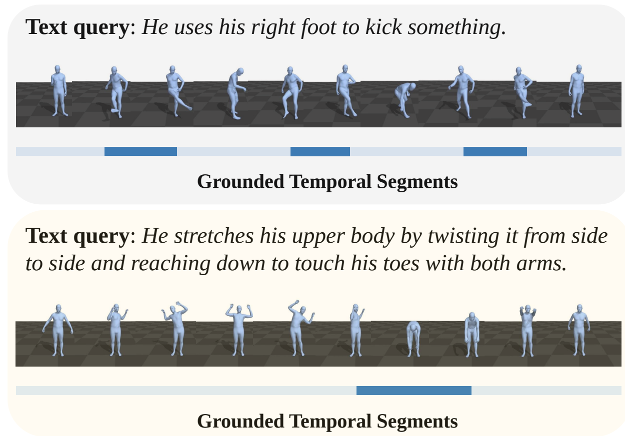

Figure 1: Illustration of the Text-based Human Motion Grounding (TMHG) task and samples of the proposed BABEL-Grounding dataset. Best viewed in color.

Thus, a textual description of particular human actions often corresponds to specific temporal segments of the sequence rather than the entire sequence. Precisely localizing the temporal segments corresponding to the text query presents a substantial challenge.

In this study, we introduce the task of text-based human motion grounding (THMG) for the first time, which aims to identify the start and end timestamps of all segments corresponding to a given textual description from an untrimmed motion sequence. Unlike existing motion temporal action localization tasks [80], the queries in THMG consist of arbitrary natural languages rather than a predefined set of action labels. As shown in Figure 1, the THMG task is highly challenging as it requires simultaneous consideration of several critical factors: (1) Achieving precise grounding of the time interval corresponding to the query within a long sequence demands the model's ability to grasp global temporal context effectively. (2) The model should jointly tackle the information from motion and language, ensuring a thorough fusion and interaction between the two modalities. (3) As human pose representation inherently possesses a graph structure, the model needs to capture the latent spatial topological information.

Effectively capturing global temporal information is crucial for THMG task. Existing frameworks for text-motion analysis mostly adopt recurrent neural networks [23, 72] or temporal convolutions [21, 22, 33, 83, 90, 91] as the main workhorse. However, their ability to capture long-term dependency is quite limited. Recently, there has been a surge of interest in transformer-based models [56, 69, 82] for modeling temporal dependencies. In these approaches,

*Corresponding author.

a global temporal self-attention mechanism is employed across all frames in the sequence. Nonetheless, these methods encounter challenges when dealing with untrimmed sequences that are very long, as computing a global temporal self-attention is exceedingly computationally expensive in such scenarios.

In this work, we aim to attack all the aforementioned challenges through a unified model that seamlessly incorporates temporal global context, language query control, and spatial graph topology. Our primary source of inspiration stems from the recently proposed state space model called Mamba [16], an efficient model for handling long-term dependencies within lengthy sequences, while maintaining linear computational cost. Mamba has demonstrated its power across diverse domains, including language modeling and visual understanding. However, its potential application in human motion tasks remains largely unexplored.

The core of Mamba is an input-dependent selection mechanism, enabling the model to selectively propagate or forget information over time depending on the current input. This innovative design greatly enhances traditional State Space Model (SSM) methods, as Mamba is capable of grasping the global context of long sequences while filtering out irrelevant information. However, in the THMG task, the model must select information based textual queries as well. To address this challenge, a text-controlled selection mechanism is introduced, wherein the key idea is to condition the state transition matrix on both motion and text queries. Unlike existing multimodal Mamba methods [61, 89] that merely concatenate the textual and visual features and feed it into Mamba blocks, our approach is the first work in enabling texts to dictate the selective propagation of input information. This ensures that the model dynamically adjusts its focus based on the interplay between motion and text inputs. Furthermore, as human motion sequences inherently possess a graph topology, the original Mamba model is not suitable as it is designed to operate on univariate time series. To address this, we enhance Mamba by integrating relational information via graph neural networks into its state representation to facilitate topology awareness. The resulting framework, termed Text-Controlled Motion Mamba (TM-Mamba), can selectively extract relevant global context information based on textual queries in the motion sequence.

For evaluation, existing datasets [45, 59] with frame-level temporal annotations are not directly applicable to the THMG task. While BABEL [59] provides temporal boundaries for all actions that occur in the sequence, it lacks complete detailed textual descriptions, offering only simple categorical phrases. Motion-X [45] offers detailed pose descriptions for each frame, but its texts focus on the movement of human body parts at each timestep, making it incapable of establishing a mapping from semantic action descriptions to temporal segments. To address this gap, based on BABEL, a new dataset called BABEL-Grounding is introduced to serve as a benchmark for evaluating THMG task. BABEL-Grounding is the first text-motion dataset that provides detailed textual descriptions of human actions along with their corresponding temporal segments in untrimmed motion sequences. Like real-world scenarios, each query may correspond to multiple temporal segments. Extensive evaluations of TM-Mamba are conducted on the newly introduced dataset, demonstrating its effectiveness in THMG task. Our primary contributions can be summarized as follows:

- We introduce a new text-motion task, text-based human motion grounding (THMG), along with a text-motion dataset called BABEL-Grounding tailored specifically for THMG, which is the first of its kind.
- We proposed TM-Mamba, a unified model with only linear memory cost specially crafted for THMG task, which is the first work that incorporates a text-controlled selection mechanism into the Mamba framework.
- Extensive evaluations on the BABEL-Grounding dataset demonstrates the efficacy of the proposed TM-Mamba.

2 RELATED WORK

2.1 Datasets for Text-Motion Learning

This section presents an overview of existing human motion datasets annotated with texts. These text-motion datasets are primarily developed for text-driven motion generation task, hence they typically include textual descriptions at the sequence level for each motion sequence. For example, the KIT Motion Language dataset [58] is the first that provides human motion alongside corresponding sequence-level textual descriptions. Following this direction, several subsequent works have endeavored to construct larger-scale datasets of similar kind, such as HumanML3D [21], InterHuman [43], HumanLong3D [26], FLAG3D [68], STDM [41]. Some datasets go beyond mere textual annotations and incorporate additional contextual information. For instance, HUMANISE [72] integrates 3D scene information to facilitate motion generation within 3D environments, while HOI-Diff [55] incorporates object geometry information to support human-object interaction during generation. Furthermore, there are datasets tailored for diverse text-motion tasks beyond generation alone. Examples include PoseScript [10], which focuses on static pose generation and pose captioning, and PoseFix [11], which targets motion editing tasks.

However, all of the aforementioned datasets only contain sequence-level annotations, rendering them unsuitable for temporal tasks. To address this limitation, various efforts have been made to develop motion datasets with specific temporal information. For instance, BABEL [59] introduces a motion dataset with frame-wise annotations, providing temporal spans for each action label, thereby enabling tasks such as action localization. Constructed upon BABEL, BABEL-QA [12] extends the dataset by incorporating question-answer pairs, aiming to facilitate motion-based question answering. Another example is HuMMan-MoGen [86], which is built upon the HuMMan [6] dataset, where each sequence is divided into predefined action phases, along with phase-level detailed annotations describing the movement of each body part. Recently, Motion-X [45] offers part-level textual annotations for human pose at each frame. However, the annotations of Motion-X are based on individual poses and lack a mapping from textual descriptions of semantic actions to temporal boundaries.

2.2 Text-Motion Multi-modal Learning

Recently, there has been a growing interest in text-motion multi-modal learning. Current research mainly focuses on text-to-motion task (also known as text-driven motion generation), where human motion sequences are generated based on natural language [1–3, 7, 14, 21, 26, 27, 30, 34, 36, 38, 41, 44, 49, 56, 60, 64, 69, 71, 72, 74–76, 79, 81, 82, 84, 86, 87, 90, 92]. The key challenge is to learn a joint embedding space for motion and text. For instance, MotionCLIP [69] aligns human motion embeddings with CLIP space through cosine similarity loss to inherit the semantic structure of CLIP latent space. A traditional transformer-based auto-encoder is employed to generate motion sequences. TEMOS [56] addresses stochastic motion generation using a VAE structure, aligning the embedding space of motion and text via KL-divergence between their latent distributions. Other works focus on employing various modern generation methods for improved conditional generation. For instance, T2M-GPT [83] utilizes Vector Quantised Variational AutoEncoder (VQ-VAE) to encode motion sequences into discrete tokens, enabling GPT-like autoregressive generation and training through next-token prediction, with text embedding serving as prior. MotionDiffuse [85] incorporates Denoising Diffusion Probabilistic Models (DDPM) into the task of text-motion generation, greatly enhancing the diversity and fidelity of generated sequences.

In recent years, some works [9, 22, 33, 37, 91] aim for more general text-motion models capable of seamlessly handling various text-motion tasks simultaneously. For example, TM2T [22] establishes a bi-modal mutual mapping between texts and tokenized human motion using autoregressive neural machine translators (NMT), effectively handling both text-to-motion and motion-to-text tasks. MotionGPT [33] utilizes VQ-VAE to create a motion tokenizer and vocabulary, and then perform pretraining on both motion and text in a unified manner, by treating human motion as a foreign language. Additionally, some works explore various text-motion tasks beyond generation, including motion question answering [12], text-based motion editing [15, 37, 86], text-motion retrieval [54, 57], etc.

2.3 State Space Model

State space models (SSMs) are a series of sequential models renowned for their computation and memory efficiency and the ability to model long-term dependencies. The pioneering work, S4 [19], first proposed applying HiPPO [17] initialization to enable SSMs to maintain long-range memory. Subsequent studies [13, 18, 20, 25, 28, 53, 67] have followed this direction, further improving the space structure and network architecture of S4.

Recently, Mamba [16] introduces an input-dependent selection mechanism into SSMs, demonstrating linear time efficiency in long-sequence modeling and achieving outstanding performance across various sequential tasks. The model has been adapted to diverse tasks, including image restoration [24], image segmentation [46, 50, 63, 73, 77], point cloud [42], video understanding [40, 78], pan-sharpening [29], graph analysis [4, 70], multimodal learning [61, 89]. There're also some efforts [47, 93] aiming to establish a universal visual backbone based on Mamba by sequentializing images using patch-based methods akin to ViT. In a more recent study, [88] introduces a motion generation model that incorporates the Mamba block within a denoising U-Net architecture to effectively enhance motion consistency across frames. The textual information is integrated during the diffusion process via a conditional denoiser, while the Mamba block is solely utilized for spatial-temporal feature extraction, without a multimodal design.

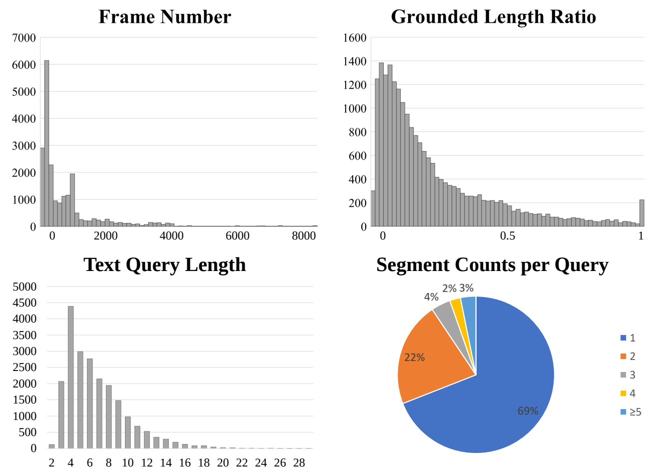

Figure 2: Dataset statistics of BABEL-Grounding. 'Frame Number' refers to the length of motion sequences. 'Text Query Length' denotes the length of textual annotations in the data. 'Grounded Length Ratio' indicates the ratio of the length of temporal segments corresponding to each text query to the total length of the sequence. 'Segment Counts per Query' refers to the number of temporal segments corresponding to each text query.

3 BABEL-GROUNDING DATASET

The task of THMG involves determining the start and end timestamps of all the segments in the motion sequence that align with the given textual description. Our objective is to construct a dataset where each text query depicting specific human actions can be mapped to one or more temporal segments within the motion sequence. To this end, we construct the BABEL-Grounding dataset based on BABEL [59]. BABEL-Grounding provides detailed textual descriptions of human actions alongside their corresponding temporal segments in untrimmed motion sequences, with a total of 5,339 sequences with 21,307 text-segments annotations. Each sequence averages 743 frames, with the ground-truth temporal segments having an average frame count of 112, indicating their sparse distribution within the lengthy motion sequences. Figure 2 illustrates the dataset statistics of BABEL-Grounding. As depicted, the BABEL-Grounding dataset contains comprehensive textual descriptions for motion sequences of diverse lengths. Each text query may correspond to multiple temporal segments, and these segments are sparsely distributed throughout the entire motion sequence, inherently posing a challenge for motion grounding task.

Below, we provide a concise overview of the construction process for BABEL-Grounding annotations. The original BABEL dataset provides dense temporal annotations by labeling each temporal segment with corresponding actions. However, its textual annotations consist of simple categorical phrases rather than detailed and comprehensive sentences, and its data structure is not directly applicable to the THMG task. Therefore, a set of manually-crafted

Algorithm 1 Selection Mechanism (Mamba)

Input: Input sequence $\mathbf{X} \in \mathbb{R}^{\mathsf{V} \times \mathsf{L} \times \mathsf{D}}$
Output: Output sequence $\mathbf{Y} \in \mathbb{R}^{\mathsf{V} \times \mathsf{L} \times \mathsf{D}}$

$\mathbf{B} : (\mathsf{V}, \mathsf{L}, \mathsf{N}) \leftarrow \text{Linear}_{\mathbf{B}}(\mathbf{X})$
$\mathbf{C} : (\mathsf{V}, \mathsf{L}, \mathsf{N}) \leftarrow \text{Linear}_{\mathbf{C}}(\mathbf{X})$
$\boldsymbol{\Delta} : (\mathsf{V}, \mathsf{L}, \mathsf{D}) \leftarrow \log(1 + \exp(\text{Linear}_{\boldsymbol{\Delta}}(\mathbf{X}) + \text{Parameter}_{\boldsymbol{\Delta}}))$
$\overline{\mathbf{A}}, \overline{\mathbf{B}} : (\mathsf{V}, \mathsf{L}, \mathsf{D}, \mathsf{N}) \leftarrow \text{discretize}(\boldsymbol{\Delta}, \text{Parameter}_{\mathbf{A}}, \mathbf{B})$
$\mathbf{Y} \leftarrow \text{SSM}(\overline{\mathbf{A}}, \overline{\mathbf{B}}, \mathbf{C})(\mathbf{X})$
return $\mathbf{Y}$

Algorithm 2 Text-Controlled Selection Mechanism

Input: Input sequence $\mathbf{X} \in \mathbb{R}^{\mathsf{V} \times \mathsf{L} \times \mathsf{D}}$; Text query embedding $q \in \mathbb{R}^{D}$
Output: Output sequence $\mathbf{Y} \in \mathbb{R}^{\mathsf{V} \times \mathsf{L} \times \mathsf{D}}$

$\mathbf{B} : (\mathsf{V}, \mathsf{L}, \mathsf{N}) \leftarrow \text{Linear}_{\mathbf{B}}(\mathbf{X}, q)$
$\mathbf{C} : (\mathsf{V}, \mathsf{L}, \mathsf{N}) \leftarrow \text{Linear}_{\mathbf{C}}(\mathbf{X}, q)$
$\boldsymbol{\Delta} : (\mathsf{V}, \mathsf{L}, \mathsf{D}) \leftarrow \log(1 + \exp(\text{Linear}_{\boldsymbol{\Delta}}(\mathbf{X}, q) + \text{Parameter}_{\boldsymbol{\Delta}}))$
$\overline{\mathbf{A}}, \overline{\mathbf{B}} : (\mathsf{V}, \mathsf{L}, \mathsf{D}, \mathsf{N}) \leftarrow \text{discretize}(\boldsymbol{\Delta}, \text{Parameter}_{\mathbf{A}}, \mathbf{B})$
$\mathbf{Y} \leftarrow \text{SSM}(\overline{\mathbf{A}}, \overline{\mathbf{B}}, \mathbf{C})(\mathbf{X})$
return $\mathbf{Y}$

rules has been employed to augment the dataset, which will be elaborated below.

3.1 Textual augmentation

The quality of textual annotations in BABEL is greatly limited by items that consist of only simple words or phrases, such as 'place', 'turn', and 'step'. These ambiguous and meaningless items make up a considerable proportion of the data and fail to provide detailed descriptions of human body movement involved in the motion. To address this issue, two approaches have been applied:

Utilizing external annotations. BABEL is built upon the AMASS motion capture database [51], while HumanML3D [21] offers detailed textual annotations at the sequence level for the AMASS database. For each entry in the BABEL dataset with overly simplistic annotations, its corresponding entry in HumanML3D is located by the sequence ID. When the sequence-level HumanML3D annotations contain the phrase from the BABEL annotations, human annotators will manually verify their correspondence and supplement the BABEL annotations with detailed textual descriptions from HumanML3D.

Template-based augmentation. As HumanML3D's sequence-level annotations only partially cover the items in BABEL, many low-quality annotations remain unaddressed, particularly those with only one-word labels. To address this gap, we've manually crafted templates to enhance them. For example, 'take' is expanded to 'take something with his hands', 'stir' becomes 'stir something with his hands in a circular motion', and 'place' becomes 'place the object at a specific location'. This substitution process enriches the overly simplistic annotations, rendering them more comprehensible for the model. The final textual annotations are further refined by ChatGPT(gpt-3.5-turbo) to make them more fluent, complete and diverse.

3.2 Temporal Augmentation

Time windows merging. In the original BABEL dataset, temporal segments of multiple annotated items may overlap with each other. To further improve the quality of the textual annotations, we design a time windows merging rule, which merges annotated segments that have a significant overlap. To be specific, if the overlapped portion of two segments counts for more than a certain ratio (which is empirically set to 0.8) of one of them, the textual annotations of them are then merged to describe the motion within the overlapped part. This leads to more detailed texts which comprehensively depict the human motion.

one-to-many mapping. BABEL dataset provides annotations for each motion that occur in a sequence. However, in the THMG task, one query may correspond to multiple temporal segments in the sequence. To implement this feature, a one-to-many mapping from text to segments is established by merging the items with the same textual annotations.

4 METHOD

4.1 Preliminaries on Mamba

In this section, we present a brief review of State Space Models (SSM) and Mamba [16]. SSM is a series of sequential models that maps the input sequence $x(t) \in \mathbb{R}$ to the output sequence $y(t) \in \mathbb{R}$ through a hidden state $h(t) \in \mathbb{R}^{N}$, which can be depicted as a linear ODE:

$$\begin{aligned} h'(t) &= \mathbf{A}h(t) + \mathbf{B}x(t), \\ y(t) &= \mathbf{C}h(t). \end{aligned} \tag{1}$$

where $\mathbf{A} \in \mathbb{R}^{N \times N}, \mathbf{B} \in \mathbb{R}^{N \times 1}, \mathbf{C} \in \mathbb{R}^{1 \times N}$ are the evolution and projection parameters. This continuous ODE can be discretized using a timescale parameter $\boldsymbol{\Delta}$ following the zero-order hold (ZOH) rule:

$$\begin{aligned} \overline{\mathbf{A}} &= \exp(\boldsymbol{\Delta}\mathbf{A}), \\ \overline{\mathbf{B}} &= (\boldsymbol{\Delta}\mathbf{A})^{-1}(\exp(\boldsymbol{\Delta}\mathbf{A}) - \mathbf{I}) \cdot \boldsymbol{\Delta}\mathbf{B}. \end{aligned} \tag{2}$$

The discretized form of the aforementioned formulation can be calculated using linear recurrence:

$$\begin{aligned} h_t &= \overline{\mathbf{A}}h_{t-1} + \overline{\mathbf{B}}x_t, \\ y_t &= \mathbf{C}h_t. \end{aligned} \tag{3}$$

However, linear recurrence requires unfolding in time and is unable to be parallelized. S4 [19] ensures Linear Time Invariance (LTI) by assuming that $\mathbf{A}, \mathbf{B}, \mathbf{C}, \boldsymbol{\Delta}$ remain static, allowing for the implementation using global convolution as $\mathbf{y} = \mathbf{x} * \overline{\mathbf{K}}$, where

$$\overline{\mathbf{K}} = (\mathbf{C}\overline{\mathbf{B}}, \mathbf{C}\overline{\mathbf{A}\mathbf{B}}, \dots, \mathbf{C}\overline{\mathbf{A}}^{L-1}\overline{\mathbf{B}}), \tag{4}$$

and L denotes the length of the input sequence, and $\overline{\mathbf{K}} \in \mathbb{R}^{L}$ represents a structured convolutional kernel. On the other hand, Mamba [16] introduces an input-dependent selection mechanism by making $\overline{\mathbf{A}}, \overline{\mathbf{B}}, \mathbf{C}, \boldsymbol{\Delta}$ become functions of x_t. Such formulation can be efficiently computed via the proposed parallel scan algorithm.

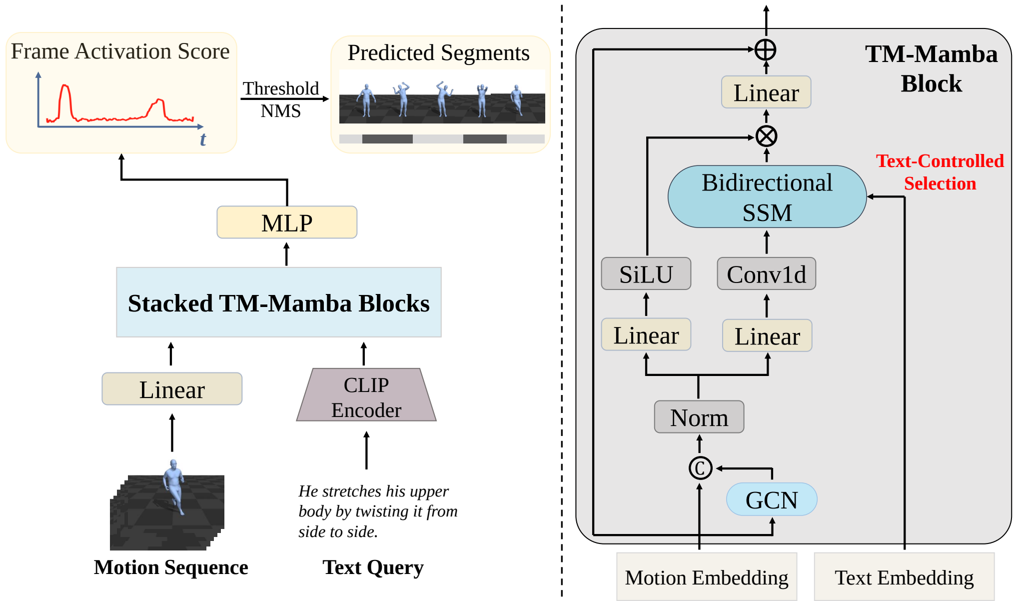

Figure 3: Left: overall architecture of our proposed model. Right: Illustration of TM-Mamba block. 'Bidirectional SSM' refers to the text-controlled selection mechanism demonstrated in Algorithm 2 with bidirectional modeling.

4.2 Text-Controlled Selection Mechanism

Mamba emphasizes the crucial role of selectivity in constructing sequence models. When dealing with very long sequences, it becomes impractical to memorize all the information within the compressed state vector. Therefore, it's essential to design a selection mechanism that controls how information propagates or interacts along the sequence dimension. Mamba addresses this challenge by employing a context-aware parameterization of state transition matrices. This enables the model to focus on or filter out information based on the current input data.

However, in the context of the THMG task, the model needs to dynamically select relevant global information from the sequence based on the text query to achieve better grounding performance. Existing multimodal Mamba methods [61, 89] use simple concatenation to merge textual and visual features as input for Mamba, but the key selection mechanism governing the information flow remains the same. To overcome this limitation, we propose a text-controlled selection mechanism that allows the selection process to depend on both the motion input and the text query. As illustrated in Algorithm 1 and 2, Mamba parameterizes $\overline{\mathbf{A}}, \overline{\mathbf{B}}, \mathbf{C}, \boldsymbol{\Delta}$ as functions of the input, whereas in text-controlled selection, these parameters become functions of input sequence as well as text query.

The algorithm essentially resembles a text-based gating mechanism, dynamically controlling the flow of information based on textual queries. Theorem 1 of [16] implies that Algorithm 2 exhibits similarities to gated RNN under certain conditions:

Lemma 4.1. *When $N = 1, \mathbf{A} = -1, \mathbf{B} = 1$, the text-controlled selection mechanism takes the form of $g_t = \sigma(Linear_{\Delta}(X, q))$ and $h_t = (1 - g_t)h_{t-1} + g_t h_t$, where X denotes input sequence and q denotes query embedding.*

Lemma 4.1 indicates that the text-controlled selective SSM bears resemblance to a gated RNN, wherein the gate g_k relies on both motion input and text query, enabling the text query to control the information flow during propagation. Given that the recurrence process in line 5 of Algorithm 2 remains unchanged, the resultant text-controlled SSM can still be efficiently computed using the parallel scan algorithm outlined in [16]. The new algorithm can be implemented by modifying the forward and gradient backward functions of the original Mamba, which enables end-to-end joint training of SSM and the language backbone.

4.3 Text-Controlled Motion Mamba

Text-Controlled Selective SSMs enjoy linear computational complexity and memory consumption, making them suitable for extracting the global context of very long sequences. This makes it a natural choice for temporal modeling in motion grounding task. However, the human skeleton inherently possesses a latent graph structure, constituting a multivariate time series. Mamba operates on univariate sequences, thus overlooking the interaction of human joints.

In this study, we enhance Mamba through the incorporation of topology awareness, achieved by integrating relational embeddings to convey information regarding neighboring nodes. Suppose an input sequence takes the shape of $\mathbf{X} \in \mathbb{R}^{V \times L \times D}$, where V denotes the number of joints in the human skeleton, L denotes the sequence length, and D denotes the feature dimension. The relational embedding $\mathbf{R}$ is computed as $\mathbf{R} = f(\mathbf{X}) \in \mathbb{R}^{V \times L \times D}$, which encapsulates the graph message at each time step. Here, f denotes a Graph Neural Network (GNN), implemented as the AGCN proposed in [66]. Subsequently, the relational embedding $\mathbf{R}$ is concatenated with the motion features and fed into the text-controlled selective SSM in

Table 1: Ablation studies on BABEL-Grounding dataset. The best results are in bold.

Methods	Text-Controlled	Relational	mAP@IoU (%)							
			0.1	0.2	0.3	0.4	0.5	0.6	0.7	Average
Unidirectional			29.8	27.2	24.9	22.8	20.5	17.8	14.4	22.5
	✓		41.5	37.9	34.4	30.7	27.7	23.7	18.7	30.7
		✓	40.5	37.1	34.0	30.6	27.7	23.9	19.6	30.5
	✓	✓	42.5	38.5	34.9	31.5	28.3	23.9	19.2	31.3
Bidirectional			38.3	35.0	32.0	29.2	26.8	23.4	19.0	29.1
	✓		51.2	48.0	44.1	40.1	36.0	30.5	24.7	39.2
		✓	46.8	43.6	40.0	35.5	32.0	27.1	21.2	35.2
	✓	✓	**53.9**	**50.5**	**46.7**	**42.8**	**38.4**	**32.6**	**26.0**	**41.6**

Algorithm 2, to jointly capture the global temporal information of each joint alongside its topological context.

The overall architecture of the Text-Controlled Motion Mamba is depicted in Figure 3. Unlike the vanilla Mamba block, which adopts unidirectional causal modeling, the task of THMG demands the global context of the entire sequence. To address this issue, a bidirectional non-causal structure, as proposed in Vision Mamba [93], is employed. The input goes through a stack of TM-Mamba blocks, producing the output $\mathbf{Y} \in \mathbb{R}^{V \times L \times D}$. After mean pooling along the V dimension, $\mathbf{Y}$ is subsequently forwarded to an MLP layer. This yields the frame activation score $s_t (t = 1, 2, \ldots, T)$ for each frame, indicating the likelihood of its inclusion in the retrieved temporal segments. The entire framework can be supervised using a simple cross-entropy loss.

$$\mathcal{L}_{ce} = -\frac{1}{T} \sum_{t}^{T} (y_t \log s_t + (1 - y_t) \log (1 - s_t)). \tag{5}$$

where y_t denotes the ground-truth label indicating whether frame t lies within the segments retrieved by the text query. The proposed method achieves effective global context extraction, query-based information selection, and topology modeling within a unified yet simple framework. Compared to transformer-based methods, our approach eliminates the necessity of computing self-attention over the entire sequence along the temporal dimension. This eliminates the need for quadratic memory, rendering it feasible for processing very long sequences.

5 EXPERIMENTS

5.1 Implementation and Evaluation Details

The motion data pre-processing procedure adheres to the methodology outlined in BABEL [59]. The maximum length of the motion sequence is constrained to 2000. The same data split for training and evaluation as BABEL is adopted. We employ CLIP [62] (clip-vit-base-patch32) as the text encoder, with its parameters being jointly tuned together with the entire model during training. The feature dimension D in Algorithm 2 for both the input motion sequence and text query embedding is set to 256. The batch size and base learning rate are configured as 4 and 5×10^{-4} respectively, with the learning rate for CLIP set to 5×10^{-5}. The optimizer is an AdamW [48] with a weight decay of 1×10^{-4}. The number of stacked Motion-Mamba blocks is empirically specified as 3. Model training is conducted on a single Nvidia A40 GPU.

During the inference stage, a series of thresholds are used to obtain the predicted temporal segments, following [5]. Subsequently, non-maximum suppression is performed to remove overlapping segments. The evaluation of grounding performance is conducted using mean Average Precisions (mAPs) under different Intersection of Union (IoU) thresholds, namely $[0.1 : 0.7 : 0.1]$.

5.2 Ablation Studies

We conduct extensive ablative studies on BABEL-Grounding dataset to demonstrate the effect of each component in our proposed TM-Mamba model.

Text-Controlled Selection Mechanism. The text-controlled selection mechanism lies at the core of the TM-Mamba model. Removing this component directly from Algorithm 2 for ablation causes the model incapable of perceiving text query information. To enable comparison, the ablative models follow the practice of [61, 89], where textual embeddings are concatenated with the sequence input. An MLP is then employed for feature fusion. As demonstrated in Table 1, the text-controlled selection mechanism accounts for a substantial performance gain. This underscores the significance of dynamically regulating the propagation of information based on textual queries.

Unidirectional v.s. Bidirectional. The Vanilla Mamba model adopts a unidirectional approach, whereby the model can only access the sequence history while processing the current step. This unidirectional approach proves advantageous for tasks such as language modeling. However, in the context of the THMG task, the model requires a comprehensive understanding of the global context of the entire sequence, demanding a non-causal bidirectional structure. The results presented in Table 1 substantiate this claim, demonstrating that the bidirectional model significantly outperforms its unidirectional counterpart.

Relational Embedding. The importance of topology modeling in THMG task is also evaluated by removing the relational embedding from TM-Mamba. As shown, the inclusion of relational embeddings $\mathbf{R}$ improves the performance of the model by effectively capturing the underlying graph structure within the skeletal data.

Table 2: Performance comparisons to baseline methods on BABEL-Grounding dataset. The best results are in bold.

Methods	mAP@IoU (%)							
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	Avg
S4D-LegS [18]	29.1	26.2	23.0	19.6	15.9	12.0	8.9	19.2
S4D-Lin [18]	30.1	26.9	23.8	20.5	16.6	12.8	9.5	20.0
2s-AGCN [66]	31.2	24.7	20.5	17.1	14.0	10.5	7.8	18.0
InfoGCN [8]	49.5	42.5	36.1	30.5	26.0	20.8	15.3	31.5
MomentDETR [39]	51.1	46.0	39.1	32.9	26.8	20.5	13.6	32.9
EaTR [32]	53.4	48.5	43.6	37.3	31.0	23.5	16.2	36.2
STCAT [35]	47.1	44.2	40.7	37.1	33.3	28.9	23.6	36.4
TM-Mamba	**53.9**	**50.5**	**46.7**	**42.8**	**38.4**	**32.6**	**26.0**	**41.6**

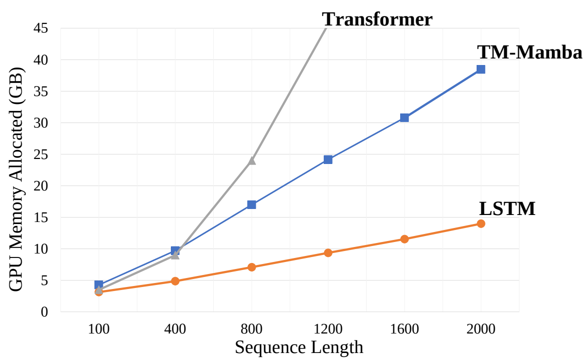

Figure 4: Comparison of memory consumption for TM-Mamba and its transformer and LSTM counterparts under varying motion sequence lengths. The transformer model runs out of GPU memory when the sequence length reaches 1200.

5.3 Comparison with Baselines

Given that the THMG task is novel and lacks existing works for comparison, several baseline methods are implemented for analysis. To begin, several prominent models are chosen from the video moment retrieval task: MomentDETR [39], EaTR [32] and STCAT [35]. However, these models rely on global temporal transformers, which require quadratic computational memory and are impractical for handling the lengthy sequences in THMG task. As a remedy, their global temporal transformers are substituted with efficient GRUs which require only linear memory. We also implement two baselines based on powerful human motion backbones using spatial-temporal graph convolutions, namely 2s-AGCN [66] and InfoGCN [8]. Text embedding is incorporated through concatenation and MLP fusion to integrate query information for text-based grounding. Finally, we implemented several baselines based on recent SSM-based models [18] to demonstrate the advantages of TM-Mamba over other SSM methods. As illustrated in Table 2, TM-Mamba outperforms the baseline models in terms of mAP at various IoU thresholds, thereby demonstrating its effectiveness on the BABEL-Grounding.

Table 3: Performance comparison of TM-Mamba and its temporal transformer counterpart under different maximum sequence length.

Length	Model	mAP@IoU (%)							
		0.1	0.2	0.3	0.4	0.5	0.6	0.7	Avg
300	TM-Mamba	77.0	73.7	68.8	63.0	58.2	50.2	39.5	61.5
	Transformer	72.9	67.3	60.9	54.6	48.4	39.4	30.2	53.4
500	TM-Mamba	74.0	70.2	65.4	59.7	53.7	45.7	37.1	58.0
	Transformer	67.3	62.3	56.7	50.5	44.7	37.0	28.2	49.5
1000	TM-Mamba	58.0	54.5	50.5	46.5	41.4	35.9	28.4	45.0
	Transformer	53.8	48.4	42.9	37.7	32.6	26.8	20.6	37.5
1500	TM-Mamba	55.6	51.9	48.1	44.2	39.5	33.8	26.6	42.8
	Transformer	*– Out of Memory –*							
2000	TM-Mamba	53.9	50.5	46.7	42.8	38.4	32.6	26.0	41.6
	Transformer	*– Out of Memory –*							

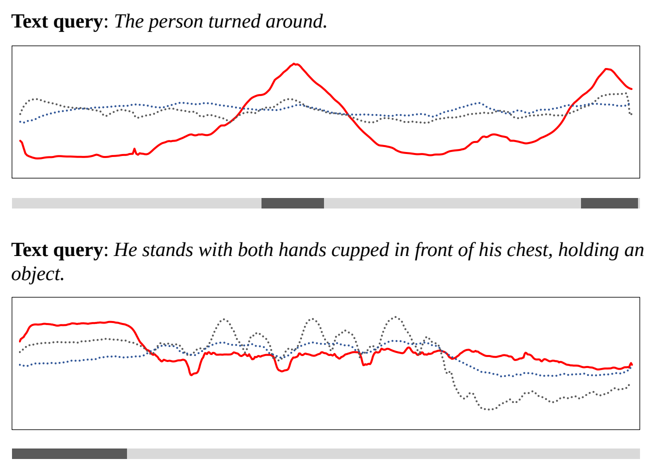

Figure 5: Visualizations of predicted frame activation score. The solid red line denotes the predicted score of our full model, while the dashed blue and gray lines denote the model without text control and relational embeddings, respectively. The gray bars below illustrate the ground-truth temporal segments. Best viewed in color.

5.4 More Analysis

Memory Consumption. The memory usage of the Mamba-based models increases linearly with the length of the sequence, which allows them to effectively handle longer sequences. Figure 4 validates this by comparing the GPU memory consumption of our TM-Mamba model with its transformer and LSTM counterparts when processing sequences of varying lengths. As depicted in the figure, the memory cost of the transformer-based model increases quadratically with sequence length, causing it to quickly run out of memory. On the other hand, both the LSTM and TM-Mamba models exhibit linear memory consumption, making them applicable for processing long motion sequences when extracting global context information. The slope of TM-Mamba is larger than its LSTM counterpart due to larger hidden states.

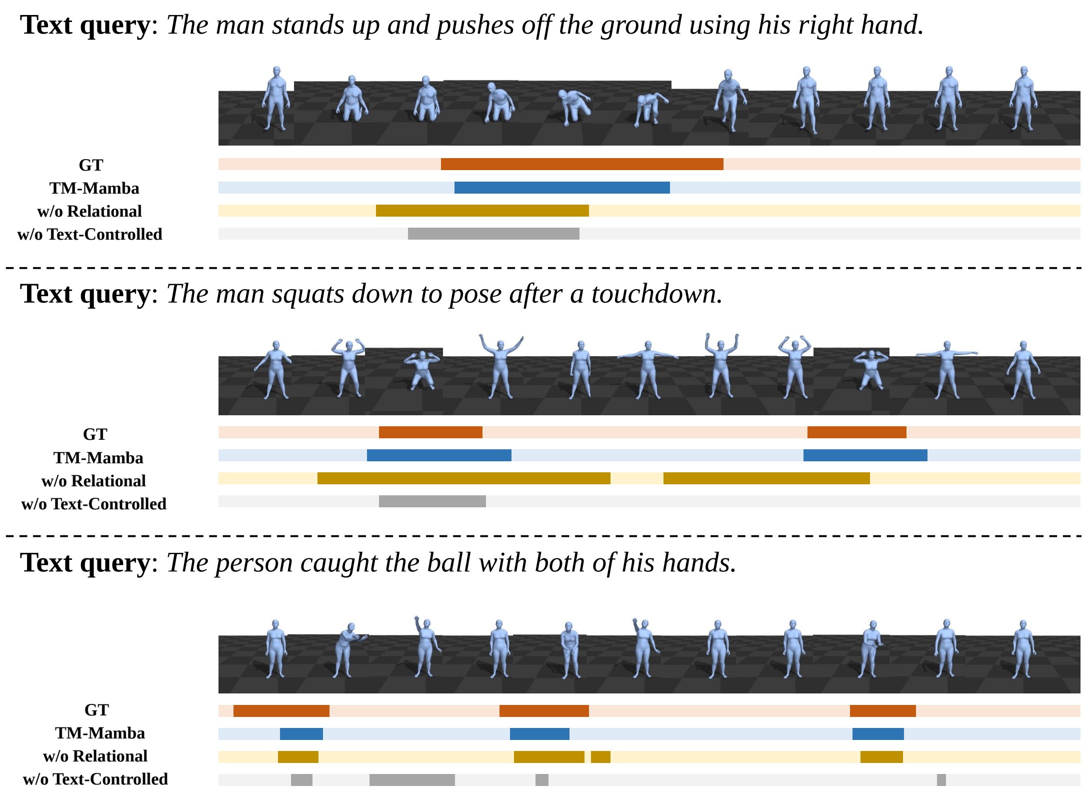

Figure 6: Visualizations of the grounding results of different models. 'GT' denotes ground-truth temporal segments corresponding to the text query. Our TM-Mamba demonstrates superior performance in motion grounding, in terms of the number of retrieved segments and the temporal boundaries of each segment. Best viewed in color.

Comparison with Transformer. Table 3 presents a comparison between TM-Mamba and its temporal transformer counterpart across various maximum sequence lengths on the BABEL-Grounding dataset. As shown, TM-Mamba exhibits superior performance to temporal transformer on shorter motion sequence. Meanwhile, the transformer model runs out of GPU memory at sequence lengths exceeding 1000 due to its quadratic memory consumption. In contrast, TM-Mamba, benefiting from its linear memory cost, manages longer sequences effectively. It can be observed that increasing sequence length complicates the extraction of temporal global context, leading to a decline in motion grounding performance.

Visualizations. In order to demonstrate the effectiveness of our proposed model, we provide visualizations on the BABEL-Grounding dataset to showcase the comparisons between the TM-Mamba model and two ablative models (w/o text control and w/o relational embedding). Figure 5 presents the frame activation scores and the ground-truth temporal segments corresponding to the text query. As illustrated, our full model manifests higher activation scores within the ground-truth temporal segments in contrast to other models, thereby resulting in enhanced accuracy in the grounding performance. Figure 6 further visualizes the grounding results predicted by the model in comparison with the ground-truth. As depicted, the ablative models struggle to accurately predict the time spans corresponding to the text query, while TM-Mamba achieves significantly improved grounding precision.

6 CONCLUSION

This work introduces a novel task called text-based human motion grounding (THMG), which aims to determine the start and end timestamps of all segments from an untrimmed motion sequence given a textual description. The key challenge lies in extracting global temporal information from lengthy untrimmed sequences based on text query, while transformer-based methods suffer from quadratic memory cost. To this end, we draw inspiration from recent advances in state space models, and propose a unified framework called TM-Mamba with linear memory cost. TM-Mamba incorporates a novel text-controlled selection mechanism into the Mamba algorithm, enabling the model to dynamically propagate input information based on text queries and extract relevant global context. A relational embedding is incorporated to model the underlying graph topology of the human skeleton. For evaluation, a text-motion dataset called BABEL-Grounding is constructed, which is the first one that provides detailed textual descriptions with their corresponding temporal segments annotation. Rigorous experiments demonstrate the effectiveness of the proposed model.

REFERENCES

[1] Chaitanya Ahuja and Louis-Philippe Morency. 2019. Language2pose: Natural language grounded pose forecasting. In *2019 International Conference on 3D Vision (3DV)*. IEEE, 719–728.

[2] Nikos Athanasiou, Mathis Petrovich, Michael J Black, and Gül Varol. 2022. Teach: Temporal action composition for 3d humans. In *2022 International Conference on 3D Vision (3DV)*. IEEE, 414–423.

[3] Samaneh Azadi, Akbar Shah, Thomas Hayes, Devi Parikh, and Sonal Gupta. 2023. Make-An-Animation: Large-Scale Text-conditional 3D Human Motion Generation. *arXiv preprint arXiv:2305.09662* (2023).

[4] Ali Behrouz and Farnoosh Hashemi. 2024. Graph Mamba: Towards Learning on Graphs with State Space Models. *arXiv preprint arXiv:2402.08678* (2024).

[5] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. 2015. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*. 961–970.

[6] Zhongang Cai, Daxuan Ren, Ailing Zeng, Zhengyu Lin, Tao Yu, Wenjia Wang, Xiangyu Fan, Yang Gao, Yifan Yu, Liang Pan, et al. 2022. Humman: Multi-modal 4d human dataset for versatile sensing and modeling. In *European Conference on Computer Vision*. Springer, 557–577.

[7] Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, and Gang Yu. 2023. Executing your Commands via Motion Diffusion in Latent Space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18000–18010.

[8] Hyung-gun Chi, Myoung Hoon Ha, Seunggeun Chi, Sang Wan Lee, Qixing Huang, and Karthik Ramani. 2022. Infogcn: Representation learning for human skeleton-based action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 20186–20196.

[9] Rishabh Dabral, Muhammad Hamza Mughal, Vladislav Golyanik, and Christian Theobalt. 2023. Mofusion: A framework for denoising-diffusion-based motion synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9760–9770.

[10] Ginger Delmas, Philippe Weinzaepfel, Thomas Lucas, Francesc Moreno-Noguer, and Grégory Rogez. 2022. PoseScript: 3D human poses from natural language. In *European Conference on Computer Vision*. Springer, 346–362.

[11] Ginger Delmas, Philippe Weinzaepfel, Francesc Moreno-Noguer, and Grégory Rogez. 2023. Posefix: Correcting 3d human poses with natural language. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 15018–15028.

[12] Mark Endo, Joy Hsu, Jiaman Li, and Jiajun Wu. 2023. Motion Question Answering via Modular Motion Programs. *arXiv preprint arXiv:2305.08953* (2023).

[13] Daniel Y Fu, Tri Dao, Khaled K Saab, Armin W Thomas, Atri Rudra, and Christopher Ré. 2022. Hungry hungry hippos: Towards language modeling with state space models. *arXiv preprint arXiv:2212.14052* (2022).

[14] Anindita Ghosh, Noshaba Cheema, Cennet Oguz, Christian Theobalt, and Philipp Slusallek. 2021. Synthesis of compositional animations from textual descriptions. In *Proceedings of the IEEE/CVF international conference on computer vision*. 1396–1406.

[15] Purvi Goel, Kuan-Chieh Wang, C Karen Liu, and Kayvon Fatahalian. 2023. Iterative Motion Editing with Natural Language. *arXiv preprint arXiv:2312.11538* (2023).

[16] Albert Gu and Tri Dao. 2023. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752* (2023).

[17] Albert Gu, Tri Dao, Stefano Ermon, Atri Rudra, and Christopher Ré. 2020. Hippo: Recurrent memory with optimal polynomial projections. *Advances in neural information processing systems* 33 (2020), 1474–1487.

[18] Albert Gu, Karan Goel, Ankit Gupta, and Christopher Ré. 2022. On the parameterization and initialization of diagonal state space models. *Advances in Neural Information Processing Systems* 35 (2022), 35971–35983.

[19] Albert Gu, Karan Goel, and Christopher Ré. 2021. Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396* (2021).

[20] Albert Gu, Isys Johnson, Aman Timalsina, Atri Rudra, and Christopher Ré. 2022. How to train your hippo: State space models with generalized orthogonal basis projections. *arXiv preprint arXiv:2206.12037* (2022).

[21] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. 2022. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5152–5161.

[22] Chuan Guo, Xinxin Zuo, Sen Wang, and Li Cheng. 2022. Tm2t: Stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts. In *European Conference on Computer Vision*. Springer, 580–597.

[23] Chuan Guo, Xinxin Zuo, Sen Wang, Shihao Zou, Qingyao Sun, Annan Deng, Minglun Gong, and Li Cheng. 2020. Action2motion: Conditioned generation of 3d human motions. In *Proceedings of the 28th ACM International Conference on Multimedia*. 2021–2029.

[24] Hang Guo, Jinmin Li, Tao Dai, Zhihao Ouyang, Xudong Ren, and Shu-Tao Xia. 2024. MambaIR: A Simple Baseline for Image Restoration with State-Space Model. *arXiv preprint arXiv:2402.15648* (2024).

[25] Ankit Gupta, Albert Gu, and Jonathan Berant. 2022. Diagonal state spaces are as effective as structured state spaces. *Advances in Neural Information Processing Systems* 35 (2022), 22982–22994.

[26] Bo Han, Hao Peng, Minjing Dong, Chang Xu, Yi Ren, Yixuan Shen, and Yuheng Li. 2023. AMD Autoregressive Motion Diffusion. *arXiv preprint arXiv:2305.09381* (2023).

[27] Gaoge Han, Shaoli Huang, Mingming Gong, and Jinglei Tang. 2023. HuTuMotion: Human-Tuned Navigation of Latent Motion Diffusion Models with Minimal Feedback. *arXiv preprint arXiv:2312.12227* (2023).

[28] Ramin Hasani, Mathias Lechner, Tsun-Hsuan Wang, Makram Chahine, Alexander Amini, and Daniela Rus. 2022. Liquid structural state-space models. *arXiv preprint arXiv:2209.12951* (2022).

[29] Xuanhua He, Ke Cao, Keyu Yan, Rui Li, Chengjun Xie, Jie Zhang, and Man Zhou. 2024. Pan-Mamba: Effective pan-sharpening with State Space Model. *arXiv preprint arXiv:2402.12192* (2024).

[30] Nhat M Hoang, Kehong Gong, Chuan Guo, and Michael Bi Mi. 2024. MotionMix: Weakly-Supervised Diffusion for Controllable Motion Generation. *arXiv preprint arXiv:2401.11115* (2024).

[31] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. 2013. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence* 36, 7 (2013), 1325–1339.

[32] Jinhyun Jang, Jungin Park, Jin Kim, Hyeongjun Kwon, and Kwanghoon Sohn. 2023. Knowing where to focus: Event-aware transformer for video grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 13846–13856.

[33] Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. 2024. Motiongpt: Human motion as a foreign language. *Advances in Neural Information Processing Systems* 36 (2024).

[34] Peng Jin, Yang Wu, Yanbo Fan, Zhongqian Sun, Yang Wei, and Li Yuan. 2023. Act as you wish: Fine-grained control of motion diffusion model with hierarchical semantic graphs. *arXiv preprint arXiv:2311.01015* (2023).

[35] Yang Jin, Zehuan Yuan, Yadong Mu, et al. 2022. Embracing consistency: A one-stage approach for spatio-temporal video grounding. *Advances in Neural Information Processing Systems* 35 (2022), 29192–29204.

[36] Korrawe Karunratanakul, Konpat Preechakul, Supasorn Suwajanakorn, and Siyu Tang. 2023. Guided motion diffusion for controllable human motion synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2151–2162.

[37] Jihoon Kim, Jiseob Kim, and Sungjoon Choi. 2023. Flame: Free-form language-based motion synthesis & editing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 8255–8263.

[38] Hanyang Kong, Kehong Gong, Dongze Lian, Michael Bi Mi, and Xinchao Wang. 2023. Priority-centric human motion generation in discrete latent space. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 14806–14816.

[39] Jie Lei, Tamara L Berg, and Mohit Bansal. 2021. Detecting moments and highlights in videos via natural language queries. *Advances in Neural Information Processing Systems* 34 (2021), 11846–11858.

[40] Kunchang Li, Xinhao Li, Yi Wang, Yinan He, Yali Wang, Limin Wang, and Yu Qiao. 2024. VideoMamba: State Space Model for Efficient Video Understanding. *arXiv preprint arXiv:2403.06977* (2024).

[41] Shuai Li, Sisi Zhuang, Wenfeng Song, Xinyu Zhang, Hejia Chen, and Aimin Hao. 2023. Sequential texts driven cohesive motions synthesis with natural transitions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 9498–9508.

[42] Dingkang Liang, Xin Zhou, Xinyu Wang, Xingkui Zhu, Wei Xu, Zhikang Zou, Xiaoqing Ye, and Xiang Bai. 2024. PointMamba: A Simple State Space Model for Point Cloud Analysis. *arXiv preprint arXiv:2402.10739* (2024).

[43] Han Liang, Wenqian Zhang, Wenxuan Li, Jingyi Yu, and Lan Xu. 2023. InterGen: Diffusion-based Multi-human Motion Generation under Complex Interactions. *arXiv preprint arXiv:2304.05684* (2023).

[44] Junfan Lin, Jianlong Chang, Lingbo Liu, Guanbin Li, Liang Lin, Qi Tian, and Chang-wen Chen. 2023. 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*. 23222–23231.

[45] Jing Lin, Ailing Zeng, Shunlin Lu, Yuanhao Cai, Ruimao Zhang, Haoqian Wang, and Lei Zhang. 2023. Motion-x: A large-scale 3d expressive whole-body human motion dataset. *arXiv preprint arXiv:2307.00818* (2023).

[46] Jiarun Liu, Hao Yang, Hong-Yu Zhou, Yan Xi, Lequan Yu, Yizhou Yu, Yong Liang, Guangming Shi, Shaoting Zhang, Hairong Zheng, et al. 2024. Swin-umamba: Mamba-based unet with imagenet-based pretraining. *arXiv preprint arXiv:2402.03302* (2024).

[47] Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, and Yunfan Liu. 2024. Vmamba: Visual state space model. *arXiv preprint arXiv:2401.10166* (2024).

[48] Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101* (2017).

[49] Jianxin Ma, Shuai Bai, and Chang Zhou. 2022. Pretrained diffusion models for unified human motion synthesis. *arXiv preprint arXiv:2212.02837* (2022).

[50] Jun Ma, Feifei Li, and Bo Wang. 2024. U-mamba: Enhancing long-range dependency for biomedical image segmentation. *arXiv preprint arXiv:2401.04722* (2024).

[51] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. 2019. AMASS: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF international conference on computer vision*. 5442–5451.

[52] Christian Mandery, Ömer Terlemez, Martin Do, Nikolaus Vahrenkamp, and Tamim Asfour. 2015. The KIT whole-body human motion database. In *2015 International Conference on Advanced Robotics (ICAR)*. IEEE, 329–336.

[53] Harsh Mehta, Ankit Gupta, Ashok Cutkosky, and Behnam Neyshabur. 2022. Long range language modeling via gated state spaces. *arXiv preprint arXiv:2206.13947* (2022).

[54] Nicola Messina, Jan Sedmidubsky, Fabrizio Falchi, and Tomás Rebok. 2023. Text-to-motion retrieval: Towards joint understanding of human motion data and natural language. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2420–2425.

[55] Xiaogang Peng, Yiming Xie, Zizhao Wu, Varun Jampani, Deqing Sun, and Huaizu Jiang. 2023. HOI-Diff: Text-Driven Synthesis of 3D Human-Object Interactions using Diffusion Models. *arXiv preprint arXiv:2312.06553* (2023).

[56] Mathis Petrovich, Michael J Black, and Gül Varol. 2022. TEMOS: Generating diverse human motions from textual descriptions. In *European Conference on Computer Vision*. Springer, 480–497.

[57] Mathis Petrovich, Michael J Black, and Gül Varol. 2023. TMR: Text-to-Motion Retrieval Using Contrastive 3D Human Motion Synthesis. *arXiv preprint arXiv:2305.00976* (2023).

[58] Matthias Plappert, Christian Mandery, and Tamim Asfour. 2016. The KIT motion-language dataset. *Big data* 4, 4 (2016), 236–252.

[59] Abhinanda R Punnakkal, Arjun Chandrasekaran, Nikos Athanasiou, Alejandra Quiros-Ramirez, and Michael J Black. 2021. BABEL: Bodies, action and behavior with english labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 722–731.

[60] Yijun Qian, Jack Urbanek, Alexander G Hauptmann, and Jungdam Won. 2023. Breaking The Limits of Text-conditioned 3D Motion Synthesis with Elaborative Descriptions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2306–2316.

[61] Yanyuan Qiao, Zheng Yu, Longteng Guo, Sihan Chen, Zijia Zhao, Mingzhen Sun, Qi Wu, and Jing Liu. 2024. VL-Mamba: Exploring State Space Models for Multimodal Learning. *arXiv preprint arXiv:2403.13600* (2024).

[62] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.

[63] Jiacheng Ruan and Suncheng Xiang. 2024. Vm-unet: Vision mamba unet for medical image segmentation. *arXiv preprint arXiv:2402.02491* (2024).

[64] Yonatan Shafir, Guy Tevet, Roy Kapon, and Amit H Bermano. 2023. Human motion diffusion as a generative prior. *arXiv preprint arXiv:2303.01418* (2023).

[65] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. 2016. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1010–1019.

[66] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. 2019. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 12026–12035.

[67] Jimmy TH Smith, Andrew Warrington, and Scott W Linderman. 2022. Simplified state space layers for sequence modeling. *arXiv preprint arXiv:2208.04933* (2022).

[68] Yansong Tang, Jinpeng Liu, Aoyang Liu, Bin Yang, Wenxun Dai, Yongming Rao, Jiwen Lu, Jie Zhou, and Xiu Li. 2023. Flag3d: A 3d fitness activity dataset with language instruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 22106–22117.

[69] Guy Tevet, Brian Gordon, Amir Hertz, Amit H Bermano, and Daniel Cohen-Or. 2022. Motionclip: Exposing human motion generation to clip space. In *European Conference on Computer Vision*. Springer, 358–374.

[70] Chloe Wang, Oleksii Tsepa, Jun Ma, and Bo Wang. 2024. Graph-mamba: Towards long-range graph sequence modeling with selective state spaces. *arXiv preprint arXiv:2402.00789* (2024).

[71] Yin Wang, Zhiying Leng, Frederick WB Li, Shun-Cheng Wu, and Xiaohui Liang. 2023. Fg-t2m: Fine-grained text-driven human motion generation via diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 22035–22044.

[72] Zan Wang, Yixin Chen, Tengyu Liu, Yixin Zhu, Wei Liang, and Siyuan Huang. 2022. Humanise: Language-conditioned human motion generation in 3d scenes. *Advances in Neural Information Processing Systems* 35 (2022), 14959–14971.

[73] Ziyang Wang and Chao Ma. 2024. Semi-Mamba-UNet: Pixel-Level Contrastive Cross-Supervised Visual Mamba-based UNet for Semi-Supervised Medical Image Segmentation. *arXiv preprint arXiv:2402.07245* (2024).

[74] Dong Wei, Huaijiang Sun, Bin Li, Xiaoning Sun, Shengxiang Hu, Weiqing Li, and Jianfeng Lu. 2023. NeRM: Learning Neural Representations for High-Framerate Human Motion Synthesis. In *The Twelfth International Conference on Learning Representations*.

[75] Yiming Xie, Varun Jampani, Lei Zhong, Deqing Sun, and Huaizu Jiang. 2023. Omnicontrol: Control any joint at any time for human motion generation. *arXiv preprint arXiv:2310.08580* (2023).

[76] Zhenyu Xie, Yang Wu, Xuehao Gao, Zhongqian Sun, Wei Yang, and Xiaodan Liang. 2023. Towards detailed text-to-motion synthesis via basic-to-advanced hierarchical diffusion model. *arXiv preprint arXiv:2312.10960* (2023).

[77] Zhaohu Xing, Tian Ye, Yijun Yang, Guang Liu, and Lei Zhu. 2024. Segmamba: Long-range sequential modeling mamba for 3d medical image segmentation. *arXiv preprint arXiv:2401.13560* (2024).

[78] Yijun Yang, Zhaohu Xing, and Lei Zhu. 2024. Vivim: a video vision mamba for medical video object segmentation. *arXiv preprint arXiv:2401.14168* (2024).

[79] Zhao Yang, Bing Su, and Ji-Rong Wen. 2023. Synthesizing Long-Term Human Motions with Diffusion Models via Coherent Sampling. In *Proceedings of the 31st ACM International Conference on Multimedia*. 3954–3964.

[80] Qing Yu and Kent Fujiwara. 2023. Frame-level label refinement for skeleton-based weakly-supervised action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 3322–3330.

[81] Ye Yuan, Jiaming Song, Umar Iqbal, Arash Vahdat, and Jan Kautz. 2023. Physdiff: Physics-guided human motion diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 16010–16021.

[82] Yuanhao Zhai, Mingzhen Huang, Tianyu Luan, Lu Dong, Ifeoma Nwogu, Siwei Lyu, David Doermann, and Junsong Yuan. 2023. Language-guided Human Motion Synthesis with Atomic Actions. In *Proceedings of the 31st ACM International Conference on Multimedia*. 5262–5271.

[83] Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Shaoli Huang, Yong Zhang, Hongwei Zhao, Hongtao Lu, and Xi Shen. 2023. T2m-gpt: Generating human motion from textual descriptions with discrete representations. *arXiv preprint arXiv:2301.06052* (2023).

[84] Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Yong Zhang, Hongwei Zhao, Hongtao Lu, Xi Shen, and Ying Shan. 2023. Generating Human Motion From Textual Descriptions With Discrete Representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14730–14740.

[85] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. 2024. Motiondiffuse: Text-driven human motion generation with diffusion model. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2024).

[86] Mingyuan Zhang, Huirong Li, Zhongang Cai, Jiawei Ren, Lei Yang, and Ziwei Liu. 2023. FineMoGen: Fine-Grained Spatio-Temporal Motion Generation and Editing. *arXiv preprint arXiv:2312.15004* (2023).

[87] Yaqi Zhang, Di Huang, Bin Liu, Shixiang Tang, Yan Lu, Lu Chen, Lei Bai, Qi Chu, Nenghai Yu, and Wanli Ouyang. 2023. MotionGPT: Finetuned LLMs are General-Purpose Motion Generators. *arXiv preprint arXiv:2306.10900* (2023).

[88] Zeyu Zhang, Akide Liu, Ian Reid, Richard Hartley, Bohan Zhuang, and Hao Tang. 2024. Motion Mamba: Efficient and Long Sequence Motion Generation with Hierarchical and Bidirectional Selective SSM. *arXiv preprint arXiv:2403.07487* (2024).

[89] Han Zhao, Min Zhang, Wei Zhao, Pengxiang Ding, Siteng Huang, and Donglin Wang. 2024. Cobra: Extending Mamba to Multi-Modal Large Language Model for Efficient Inference. *arXiv preprint arXiv:2403.14520* (2024).

[90] Chongyang Zhong, Lei Hu, Zihao Zhang, and Shihong Xia. 2023. Attt2m: Text-driven human motion generation with multi-perspective attention mechanism. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 509–519.

[91] Zixiang Zhou, Yu Wan, and Baoyuan Wang. 2023. AvatarGPT: All-in-One Framework for Motion Understanding, Planning, Generation and Beyond. *arXiv preprint arXiv:2311.16468* (2023).

[92] Zixiang Zhou and Baoyuan Wang. 2023. Ude: A unified driving engine for human motion generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5632–5641.

[93] Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. 2024. Vision mamba: Efficient visual representation learning with bidirectional state space model. *arXiv preprint arXiv:2401.09417* (2024).