AttT2M: Text-Driven Human Motion Generation with Multi-Perspective Attention Mechanism

Chongyang Zhong[1,2]*, Lei Hu[1,2]*, Zihao Zhang[1]*, Shihong Xia[1,2]†
[1]Institute of Computing Technology, Chinese Academy of Sciences
[2]University of Chinese Academy of Sciences
{zhongchongyang, hulei19z, zhangzihao, xsh}@ict.ac.cn

Abstract

*Generating 3D human motion based on textual descriptions has been a research focus in recent years. It requires the generated motion to be diverse, natural, and conform to the textual description. Due to the complex spatio-temporal nature of human motion and the difficulty in learning the cross-modal relationship between text and motion, text-driven motion generation is still a challenging problem. To address these issues, we propose **AttT2M**, a two-stage method with multi-perspective attention mechanism: **body-part attention** and **global-local motion-text attention**. The former focuses on the motion embedding perspective, which means introducing a body-part spatio-temporal encoder into VQ-VAE to learn a more expressive discrete latent space. The latter is from the cross-modal perspective, which is used to learn the sentence-level and word-level motion-text cross-modal relationship. The text-driven motion is finally generated with a generative transformer. Extensive experiments conducted on HumanML3D and KIT-ML demonstrate that our method outperforms the current state-of-the-art works in terms of qualitative and quantitative evaluation, and achieve fine-grained synthesis and action2motion. Our code is in https://github.com/ZcyMonkey/AttT2M.*

1. Introduction

With the expansion of application scenarios and demands, cross-modal human motion synthesis has become a hot research topic in computer vision and graphics. Among them, text-driven human motion synthesis aims to synthesize natural human motion that matches a given text description, which can be applied to intelligent animation production, virtual reality, game and film industry, human-robotics interaction, etc.
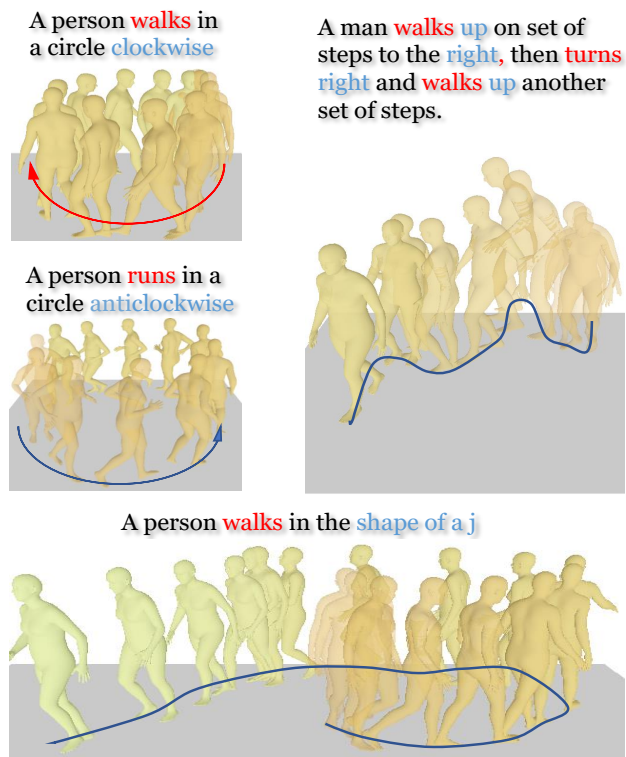
*Equal contribution
†Corresponding author

Figure 1. **Text-driven generation visualization.** Our method achieves high-quality motion that is precisely consistent with semantically rich text descriptions.

However, introducing natural language descriptions as constraints for motion synthesis, which means using cross-modal high-level semantics to control the motion synthesis process, is a difficult interdisciplinary problem involving natural language processing, motion modeling, and cross-modal relationship learning. With the availability of pre-trained language models such as CLIP [35], BERT [10], and GPT [36] for natural language processing, there are still two challenges that remained: (1) human motion is high-dimensional data with irregular spatial structure and highly dynamic temporal characteristics. It is challenging

to perform text-driven synthesis in pose space directly; (2) there are complex local (specific words correspond to specific motion sub-sequences) and global (the overall text description corresponds to the order and connection of these sub-sequences) correspondence between text and motion. How to better learn this cross-modal relationship is still a problem to be solved.

Compared to some works that directly generated text-driven motion in the pose space[4, 3, 8, 53, 41], others attempted to learn the low-dimensional representation of motion first using auto-encoder [13], VAE [49] and VQ-VAE [52] with temporal encoding. However, since human motion has both spatial and temporal characteristics, it is necessary to consider both when dealing with latent representation learning [56, 50]. Therefore, we propose Vector Quantised-Variational AutoEncoder(VQ-VAE) with **Body-Part attention-based Spatio-Temporal(BPST)** encoder to learn an expressive latent space. As for the motion-text cross-modal relationship, some works used cross-modal translation to learn the shared representation of text and motion and directly generated motion from text feature [4, 3, 33, 6]. However, the vast gaps between motion and text data make learning an adequate shared representation difficult. Other researchers attempted to treat text information as a condition during the motion generation [13, 52, 49, 53, 41, 14]. We observe two levels of correspondence in the cross-modal relationship between motion and text: 1). the local correspondence between words of text and sub-segments of motion sequences; 2). the global correspondence between the overall semantics of text and the whole motion sequence. Previous research typically focused on either local [53] or global [52, 41] text information or combined them by simply concatenating [13]. We propose **Global and Local Attention(GLA)** to consider both levels more reasonably better to learn the cross-modal relationship between text and motion.

Specifically, we propose AttT2M, a two-stage text-driven motion generation model with multi-perspective attention(Figure 2(a)). The first attention focuses on motion embedding perspective. We use a spatial transformer [44] based on **body-part attention** and TCN to extract the spatio-temporal features, and then a VQ-VAE [43] is used to quantize the features into a discrete codebook. In the text-diven generation stage, we propose **global and local attention** from the perspective of cross-modal relationship learning. After the global (sentence-level) and local (word-level) text features are extracted using CLIP, we calculate the motion-word cross-attention to learn the local cross-modal relationship and learn the global cross-modal relationship by motion-sentence conditional self-attention. Finally, a generative transformer is used to generate motion sequences.

We conduct extensive qualitative and quantitative exper-

iments on two widely used datasets, KIT-ML [34] and HumanML3D [13]. The experimental results show that our work achieves better text-driven motion generation results than previous work in qualitative and quantitative comparisons. The generated motions highly match the given natural language descriptions and maintain reliable diversity while being highly realistic and natural(seeing Figure 1). Our work can also achieve fine-grained generation and action to motion.

Our contributions can be summarized as follows:

- 1. We propose a Body-Part attention-based Spatio-Temporal VQ-VAE to map motion sequences into a better discrete code book, resulting in a more expressive low-dimensional motion representation.

- 2. We introduce Global and Local Attention to learn the global-local cross-modal relationship between text and motion, achieving precise correspondences between motion and text.

- 3. We conduct extensive qualitative and quantitative experiments on KIT-ML and HumanML3D to demonstrate that our method outperforms the previous state-of-the-art methods, which can also handle fine-grained and action-to-motion generation.

2. Related Works

Human motion synthesis is a long-standing research topic [48]. Here we only introduce works based on deep learning and exclude earlier methods. Motion synthesis started with the unconstrained generation, which generated realistic and natural motion sequences in completely random ways [11, 27, 54]. With the development of research and increasing demands, methods based on different conditions have been proposed. One primary condition was motion content. Some works focused on generating motion that satisfied the constraints of low-level control signals (such as speed, direction, trajectory, etc.). They either used a two-stage approach[19, 46, 29] by training a generative model as a prior and then generated motions through optimization, or used an end-to-end manner by parameterizing the control signals as input to predict the control signals and pose for the next frame [25, 18, 51, 37, 38, 39]. Other researchers attempted to use motion graph to synthesize high-quality human motion through searching and blending [9, 17]. In addition, another content constraint was the past motion sequences(referred to as diverse motion prediction). They usually used generative models such as VAE [45, 5] and GAN [7, 23] to encode past motion sequences and then sampled in the latent space to generate multiple prediction results. Recently, motion generation based on the cross-modal condition has been a hot topic. Some works hoped to generate dances from music [24, 26, 42], while other researchers were trying to
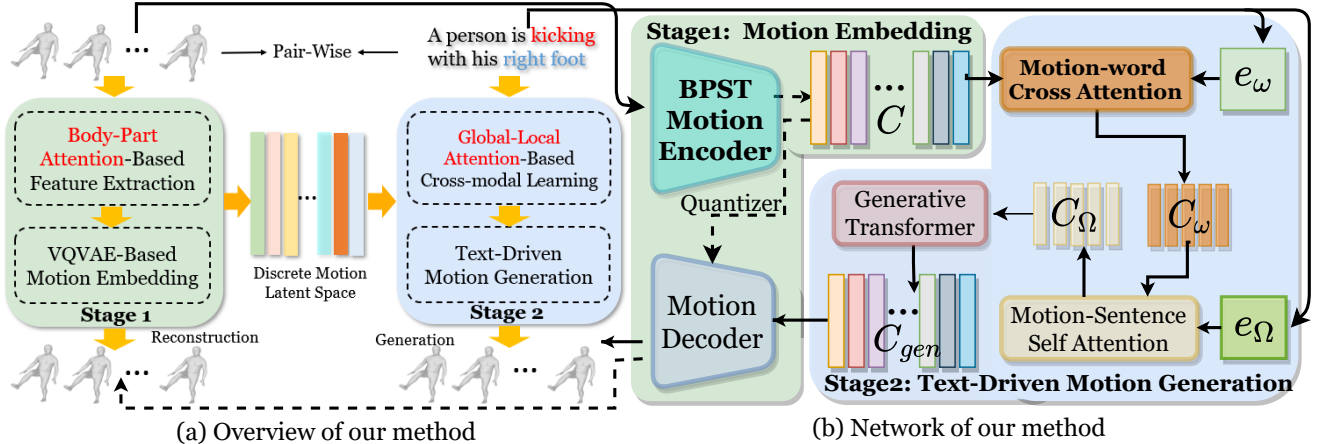
Figure 2. **Overview and network**. (a)We propose a text-driven motion generation method based on a multi-perspective attention mechanism, including body-part attention for expressive spatio-temporal feature extraction in the motion embedding stage and global-local attention for a better cross-modal relationship learning in the text-driven generation stage. (b)In stage 1, we use body-part attention-based spatio-temporal VQ-VAE to learn a discrete latent motion representation. In stage 2, we calculate local motion-word cross-attention and global motion-sentence conditional self-attention to ensure that the motion generated by the generative transformer is highly consistent with text descriptions.

generate motion from the given text, which included action labels [32, 15, 55] as well as natural language descriptions [12, 4, 8, 3, 28, 33, 6, 13, 52, 49, 53, 41, 14, 40, 20].

Motion latent representation learning in motion synthesis is about mapping motions into a low-dimensional latent space due to the high-dimensional and complex nature of the original pose space. Some works utilized autoencoder [13] structures to achieve dimension reduction. Other works attempted to train a generative model based on VAE [49] or VQ-VAE [52] to learn a latent representation of motion with more diversity. Most of these works only focused on temporal characteristics, using RNNs, TCNs, or transformers to learn temporal features. However, in motion prediction and action recognition, researchers had found that the spatial features were also crucial in motion latent representation learning [56, 50]. So we propose Body-Part attention-based SpatiO-Temporal VQ-VAE to learn a more expressive latent representation.

Text-driven motion generation aims to generate 3D human motion based on textual descriptions. We categorize related works into three types. The first type desired to learn shared latent representations of motion and text, and directly converted text to motion. After Text2Action [3] and Language2pose [4] made the first attempts, TEMOS [33] introduced a transformer VAE architecture to make the latent distributions of text and motion as similar as possible, and TEACH [6] achieved coherent motion generation based on multiple action labels. The second type rendered 3D human bodies as images and learned the motion-image-text relationship using contrastive learning with the help of CLIP. MotionCLIP [40] focused on motion generation, while AvatarCLIP [20] generated human appearance and

simple motion simultaneously. The third type treated text as a condition for motion generation. Some works were based on an encoder-decoder framework, using transformers [8] or RNNs [28] to encode motion and concatenating it with text feature as input to the decoder. To increase the diversity of generated motion, generative models were introduced. RNN+CVAE [15], transformer+CVAE [32], and Uncoupled-modulation CVAE [55] are used to generate motion from action labels. With the proposal of a large-scale motion-language dataset called HumanML3D [13], many studies achieved better results. T2M [13] adopted a two-stage approach, learning motion clip codes and conducting text-driven motion generation. They also achieved better results by jointly training text-to-motion and motion-to-text tasks [14]. T2M-GPT [52] proposed to learn the discrete representation with VQ-VAE and used a generative transformer for generation. Other works attempted to use the diffusion model [16] in pose space directly [53, 41] or latent space [49] for text-driven motion generation. This paper proposes a Global-Local Attention mechanism to learn multi-level cross-modal relationships between motion and text, achieving better text-driven motion generation.

3. Our Method

The problem we aim to solve is text-driven motion generation, which refers to generating a T frames sequence of human motions $X = \{x_1, x_2, ..., x_T\}$ based on a given text description $\Omega = \{\Omega_1, \Omega_1, ..., \Omega_N\}$ consisting of N words.

3.1. Method Overview

To generate high-quality and text-corresponding motion, we explore motion latent space learning and text-motion cross-modal relationship learning. The proposed two-stage motion generation framework is shown in Figure 2(b). Stage 1 is motion embedding. We rearrange the original motion representation to adapt the body-part segmentation and use a Body-Part attention-based Spatio-Temporal(BPST) motion encoder to learn an expressive spatio-temporal feature. Then a VQ-VAE is used to learn a discrete motion latent space C through quantizer. In Stage 2, we learn the cross-modal relationship between text and motion on the global and local levels to generate motions according to given text Ω. After the word-level embedding e_ω and sentence-level embedding e_Ω are extracted by CLIP, we propose motion-word cross attention to learn the local cross-modal relationship. Moreover, we compute motion-sentence self-attention to learn the global cross-modal relationship. To generate diverse results, we use a generative transformer to generate motion code C_{gen}, and the final generated sequence is obtained by the motion decoder.

3.2. The Motion Embedding Module

The pose of motion exhibits infinite variations, leading to a high-dimensional pose space that makes it difficult to model human motions. Therefore, many researchers attempt to learn a latent space to reduce the dimension [13, 49, 52]. However, 3D human motion is a structured time series with spatial and temporal characteristics. Previous text-driven motion generation works paid more attention to temporal modeling using RNN [15], TCN [13, 52], Transformer [33], the understanding of spatial structure was often overlooked. To better learn the spatio-temporal characteristics and map them to a low-dimensional, expressive latent space, we propose a motion embedding module named Body-Part Attention-based Spatio-Temporal VQ-VAE.

Body-part Attention-based Spatio-Temporal feature extraction Although the human body has many joints, the body part should be the basic unit that best represents the semantic information of motion. The interaction between joints within the same body part ultimately combines into the motion of the entire human. Some previous works have already realized this point [22, 2, 21]. Based on this observation, we propose a spatio-temporal encoder based on body-part attention to extract spatio-temporal features.

As shown in Figure 3(a), we divide the human body with n joints into five body parts: {Torso, Left Arm, Right Arm, Left Leg, Right Leg}, each containing its own set of joints. To apply transformer in the spatial dimension, we rearrange the representation of T2M [13] to gather information of each joint, leading to $n + 1$ motion tokens of each frame: $\bar{x}_i = \{j_{root}, j_1, j_2, ..., j_{n-1}, c^f\}$. Next, we propose a spatial transformer $Trans_{enc}$ based on body-part atten-

tion. Before computing self-attention, we map all tokens to the same dimension through different linear mapping:

$$Q^{\mathcal{I}} = W_q^{\mathcal{I}} \mathcal{J}, K^{\mathcal{I}} = W_k^{\mathcal{I}} \mathcal{J}, V^{\mathcal{I}} = W_v^{\mathcal{I}} \mathcal{J} \qquad (1)$$

where the superscript $\mathcal{I} = \{root, others, contact\}$ of Q, K, V and W_q, W_k, W_v is a denotes set representing the root joint, other joints, and foot contact. And $\mathcal{J} = \{j_{root}, j_i, c_f\}$ are the motion tokens. Finally, we concatenate the mapping results of different tokens to obtain the final Q, K,V. When calculating body-part attention, we define an adjacency mask $\mathcal{M} = \{m_{i,j}\} \in \mathbb{R}^{(n+1)\times(n+1)}$ according to body part division. The dimension of $\mathcal{M}$ becomes $n+1$ because foot contact c_f is treated as an extra joint to avoid foot slide. If joint i and j are in the same body part, $m_{i,j} = 0$, otherwise $-\infty$ we add the adjacency relation mask $\mathcal{M}$. The body-part attention is calculated by:

$$BPAtt = softmax(\frac{QK^T \oplus \mathcal{M}}{\sqrt{D}})V \qquad (2)$$

Finally, we repeat the above steps to calculate multi-head self-attention and obtain the final spatial feature f_s through a feed-forward network(FFN). $\mathcal{M}$ limits the self-attention calculation within the body part. After extracting the spatial features, we use a temporal encoder TCN_{enc} consisting of multiple temporal convolutional layers to obtain the motion spatio-temporal feature f_{st}.

Latent space learning based on VQ-VAE Inspired by T2M-GPT [52], we use VQ-VAE to learn the latent motion space. First, we construct a learnable codebook $C = \{c_i\} \in \mathbb{R}^{K*D_C}$ as a set of discrete latent motion representations. Then, we map the motion sequence X to latent features F_{st} through our spatio-temporal feature extraction module and then quantize it to a discrete latent representation. We use exponential moving average (EMA) and codebook reset [47] to train VQ-VAE and optimize it with the following loss function:

$$\mathcal{L} = \mathcal{L}recon + \alpha \mathcal{L}vel + \mathcal{L}emb + \beta \mathcal{L}com \qquad (3)$$

where:

$$\mathcal{L}rec = \mathcal{L}1(X, \hat{X}), \quad \mathcal{L}vel = \mathcal{L}1(V, \hat{V})$$
$$\mathcal{L}emb = \|sg[F_{st}] - C\|_2, \mathcal{L}com = \| F_{st} - sg[C]\|_2 \qquad (4)$$

V is velocity and ˆ indicates reconstructed. $\mathcal{L}rec$, $\mathcal{L}vel$, $\mathcal{L}emb$ and $\mathcal{L}com$ are the reconstruction term, smooth velocity term, embedding term and commit term, respectively.

3.3. The Text-driven Motion Generation Module

We observe that there are two-level correspondences in the cross-modal relationship between motion and text: (1) the correspondence between words in text and sub-segments in motion sequences. For example, in the sentence

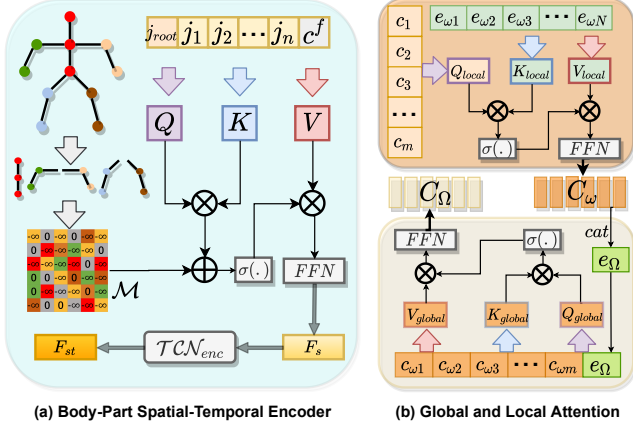**(a) Body-Part Spatial-Temporal Encoder** **(b) Global and Local Attention**

Figure 3. **BPST and GLA module**. (a)We divide the skeleton into five body parts and use the spatial transformer to learn the body-part attention to extract the spatial feature F_s. The final F_{st} is obtained by TCN_{enc}. (b)To better learn the cross-modal relationship between text and motion, we extract the word-level and sentence-level text features(C_ω and C_Ω) and use them to calculate the local motion-word cross-attention and the global motion-sentence conditional self-attention, respectively.

"With both arms outstretched at the side, bent at elbow, the person raises arms straight over head", words such as "outstretched", "bent", "raises" and "arm" are essential. If we do not consider this correspondence, important words may be ignored, resulting in generated motion inconsistent with the given text; (2) the correspondence between the overall semantics of the text and the whole motion sequence. The temporal properties of the motion are continuous, but the textual ones may not be. In the above sentence, "outstretched" and "bent" has a sequential order, while the actions they represent occur simultaneously. Thus the overall correspondence between the sentence and motion also plays an indispensable role in text-driven motion generation. Previous work usually only considered one aspect of the correspondences [53, 41, 33, 52, 49] or simply concatenate two of them [13]. To achieve better text-driven motion generation, we propose a Global and Local Attention Generative Transformer to simultaneously consider these two correspondences, using different levels of attention to better learn the cross-modal relationship between text and motion (as shown in Figure 3(b)).

Local cross-modal relationship For a text description Ω with N words and the corresponding motion discrete representations $C = \{c_i\}_{i=1}^m$, we first extract word-level features $e_\omega = \{e_{\omega_i}\}_{i=1}^N$ using the pre-trained model CLIP [35]. To learn the importance of each word in the cross-modal correspondence, we propose a local cross-modal transformer $Trans_{local}$ to calculate the cross-attention between e_ω and C. Specifically, we use e_ω as the key and value and C as the query:

$$Q_{local} = W_q C, K_{local} = W_k e_\omega, V_{local} = W_v e_\omega \quad (5)$$

Then, we compute their cross-attention:

$$CrossAtt = softmax(\frac{Q_{local}K_{local}^T}{\sqrt{D}})V_{local}$$

In this way, we can estimate the importance of each word in the sentence for motion generation(seeing Fig. 8) and learn the cross-modal relationship at the local level. Finally, we input $CrossAtt$ into a FFN to obtain the joint motion-words embedding $C_\omega = \{c_{\omega i}\}_{i=1}^m$.

Global cross-modal relationship Similar to extracting word-level features, we also use a CLIP to extract sentence-level features $e_\Omega \in \mathbb{R}^{1 \times D}$ from the text. Since the size of e_Ω is $1 \times D$, performing a softmax weighted sum on it is meaningless. To better learn global cross-modal relationship, we adopt the idea of T2M-GPT [52] and use a transformer based on conditional self-attention, called $Trans_{global}$. Specifically, we concatenate e_Ω with C_ω on temporal dimension to calculate self-attention between e_Ω and C_ω. During this process, e_Ω is fully involved in the attention calculation with C_ω, which is beneficial for learning the global cross-modal relationship.

To obtain more diverse generation results, we input C_Ω into a generative transformer TCN_{gen} to predict the probability $p(c_{m+1}|C_\Omega)$ of the next motion code in the codebook, and optimize the maximum likelihood:

$$\mathcal{L} = -\mathbb{E}_{c_{m+1} \sim p(C)}[logp(c_{m+1}|C_\Omega)] \quad (7)$$

During testing, we directly input the sentence-level feature e_Ω into TCN_{global} to predict the first motion code c_1, then continuously generate a motion code sequence C_{gen}. Finally, we input it into the motion decoder TCN_{enc} to obtain the human motion sequence $\hat{X} = \{\hat{x}_i\}_{i=1}^T$.

4. Experiments

In this section, we first introduce the **dataset and evaluation metrics** in Sec.4.1, **the implementation details** are demonstrated in supplementary materials. **Quantitative and qualitative evaluations** will be presented in Sec.4.2 and 4.3, respectively. Then, we analyze the main components of the method through **ablation study** in Sec.4.4. Finally, we further discuss our **attention mechanism** in Sec.4.5. Due to space limitations, we demonstrate our **failure cases** and **limitations** in the supplementary material.

4.1. Dataset and Metrics

KIT Motion-Language [34] is the first 3D human motion dataset with text labels, consisting of a subset of KIT [31] and CMU [1] datasets. It contains 3911 motion sequences and 6353 sequence-level text annotations, with an average of 9.5 words per annotation.

HumanML3D [13]: HumanML3D added detailed text labels to the motion in AMASS dataset, creating a motion-language dataset. It contains 14616 motion segments with

Methods	R Precision↑			FID.↓	MM-D.↓	Div.→	MM.↑
	Top-1	Top-2	Top-3				
Real	$0.511 \pm .003$	$0.703 \pm .003$	$0.797 \pm .002$	$0.002 \pm .000$	$2.974 \pm .008$	$9.503 \pm .065$	-
Hier [12]	$0.301 \pm .002$	$0.425 \pm .002$	$0.552 \pm .004$	$6.523 \pm .024$	$5.012 \pm .018$	$8.332 \pm .042$	-
TM2T [14]	$0.424 \pm .003$	$0.618 \pm .003$	$0.729 \pm .002$	$1.501 \pm .017$	$3.467 \pm .011$	$8.589 \pm .076$	$2.424 \pm .093$
T2M [13]	$0.455 \pm .003$	$0.636 \pm .003$	$0.736 \pm .002$	$1.087 \pm .021$	$3.347 \pm .008$	$9.175 \pm .083$	$2.219 \pm .074$
MDM [41]	$0.320 \pm .005$	$0.498 \pm .004$	$0.611 \pm .007$	$0.544 \pm .044$	$5.566 \pm .027$	$\mathbf{9.559} \pm .086$	$\mathbf{2.799} \pm .072$
MD [53]	$\underline{0.491} \pm .001$	$\underline{0.681} \pm .001$	$\underline{0.782} \pm .001$	$0.630 \pm .001$	$\underline{3.113} \pm .001$	$\underline{9.410} \pm .049$	$1.553 \pm .042$
MLD [49]	$0.481 \pm .003$	$0.673 \pm .003$	$0.772 \pm .002$	$0.473 \pm .013$	$3.196 \pm .010$	$9.724 \pm .082$	$2.413 \pm .079$
T2M-GPT [52]	$\underline{0.491} \pm .003$	$0.680 \pm .003$	$0.775 \pm .002$	$\underline{0.116} \pm .004$	$3.118 \pm .011$	$9.761 \pm .081$	$1.856 \pm .011$
Ours	$\mathbf{0.499} \pm .003$	$\mathbf{0.690} \pm .002$	$\mathbf{0.786} \pm .002$	$\mathbf{0.112} \pm .006$	$\mathbf{3.038} \pm .007$	$9.700 \pm .090$	$\underline{2.452} \pm .051$

Table 1. **Quantitative evaluation on the testset of HumanML3D.** We report the metrics following T2M [13] and repeat 20 times to get the average results with 95% confidence interval. The $\downarrow, \uparrow, and \rightarrow$ denote the lower, higher, and closer to Real are better, respectively. The best results are marked in bold and the second best is underlined. Our method achieves significant improvement on almost all metrics.

Methods	R Precision↑			FID.↓	MM-D.↓	Div.→	MM.↑
	Top-1	Top-2	Top-3				
Real	$0.424 \pm .005$	$0.649 \pm .006$	$0.779 \pm .006$	$0.031 \pm .004$	$2.788 \pm .012$	$11.08 \pm .097$	-
Hier [12]	$0.255 \pm .006$	$0.432 \pm .007$	$0.531 \pm .007$	$5.203 \pm .107$	$4.986 \pm .027$	$9.563 \pm .072$	-
TM2T [14]	$0.280 \pm .005$	$0.463 \pm .006$	$0.587 \pm .005$	$3.599 \pm .153$	$4.591 \pm .026$	$9.473 \pm .117$	$\mathbf{3.292} \pm .081$
T2M [13]	$0.361 \pm .006$	$0.559 \pm .007$	$0.681 \pm .007$	$3.022 \pm .107$	$3.488 \pm .028$	$10.72 \pm .145$	$2.052 \pm .107$
MDM [41]	$0.164 \pm .004$	$0.291 \pm .004$	$0.396 \pm .004$	$\underline{0.497} \pm .021$	$9.191 \pm .022$	$10.85 \pm .109$	$1.907 \pm .214$
MD [53]	$\mathbf{0.417} \pm .004$	$\underline{0.621} \pm .004$	$\underline{0.739} \pm .004$	$1.954 \pm .064$	$\mathbf{2.958} \pm .005$	$\mathbf{11.10} \pm .143$	$0.730 \pm .013$
MLD [49]	$0.390 \pm .008$	$0.609 \pm .008$	$0.734 \pm .007$	$\mathbf{0.404} \pm .027$	$3.204 \pm .027$	$10.80 \pm .117$	$2.192 \pm .071$
T2M-GPT [52]	$0.402 \pm .006$	$0.619 \pm .005$	$0.737 \pm .006$	$0.717 \pm .041$	$3.053 \pm .026$	$10.86 \pm .094$	$1.912 \pm .036$
Ours	$\underline{0.413} \pm .006$	$\mathbf{0.632} \pm .006$	$\mathbf{0.751} \pm .006$	$0.870 \pm .039$	$\underline{3.039} \pm .021$	$\underline{10.96} \pm .123$	$\underline{2.281} \pm .047$

Table 2. **Quantitative evaluation on the testset of KIT-ML.** The experimental settings are the same as Table 1. We report the metrics following T2M [13] and repeat 20 times to get the average results with 95% confidence interval. The best results are marked in bold and the second best is underlined.

a total duration of 28.59 hours. The average length of per segment is 7.1 seconds, and the longest and shortest motions are 10 and 2 seconds, respectively. The dataset contains 44970 text descriptions with an average length of 12 words, covering 5317 unique words. We follow the split of T2M [13] for our training, validation, and testing on both KIT-ML and HumanML3D.

Evaluation Metrics We adopt the evaluation metrics following T2M [13], including Frechet Inception Distance(FID.), R-precision, Multi-Modal distance(MM-D.), Diversity(Div.) and Multi-Modality(MM.), which are evaluating the distribution distance between generated motion and ground-truth, the consistency of text and generated motion, the euclidean distance of motion feature and text feature, the diversity of whole generated motion and the diversity of generated motion from the same text, respectively.

4.2. Quantitative Evaluation

We use the metrics mentioned above to evaluate the quantitative effects of our method and compare them with recent state-of-the-art works, including Hier [12], TM2T [14], T2M [13], MDM [41], MotionDiffuse(MD) [53], MLD [49], and T2M-GPT [52]. The results of these methods were obtained from their pre-trained models or related papers.

HumanML3D: As shown in Table 1, our method have significant improvements in almost all evaluation metrics compared to previous works. The performance of text-driven motion generation has two aspects: the quality of the generated motion itself and its consistency with the given text description. In terms of text consistency, our method has a significant improvement in the accuracy of top 1, 2, and 3 compared to previous works, which are also very close to the results of real data. This indicates that the consistency of our generated motion with the text is the highest, which is also evidenced by our method's lowest MM-D. As for the quality of the generated motion itself, our method has the lowest FID of all methods, indicating that the quality of our generated motion is natural, realistic, and close to real data. It is worth noting that Div can only measure the diversity of motion to a certain extent, as Div may also be large when the generated motion has various artifacts.
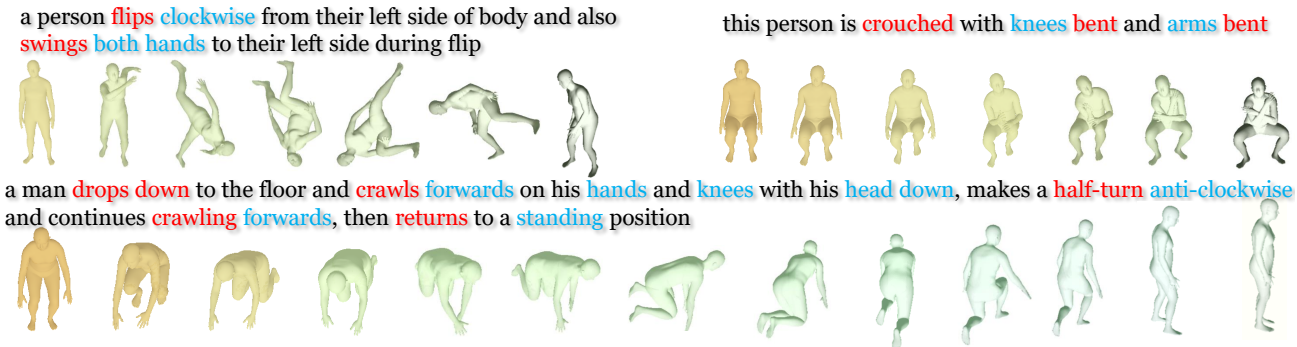
a person flips clockwise from their left side of body and also swings both hands to their left side during flip

this person is crouched with knees bent and arms bent

a man drops down to the floor and crawls forwards on his hands and knees with his head down, makes a half-turn anti-clockwise and continues crawling forwards, then returns to a standing position

Figure 4. **More visualization results on HumanML3D**. Our method achieves high-quality motion, including complex actions like "flip", "crouch", and "crawl", which are realistic and highly consistent with the given text. The change of color indicates the passage of time.

The more compelling diversity metric is MM, of which our method is the second highest.

KIT-ML: We evaluate the same metrics on KIT-ML. The results are shown in Table 2. Since MD [53], MDM [41], and MLD [49] used ground-truth lengths in their evaluation, it is understandable that they achieved good results whereas our method has better or comparable performance in all metrics(except FID) without the need for ground-truth lengths. And there is a significant improvement over other methods that do not require ground-truth length, like T2M [13] and T2M-GPT [52].

4.3. Qualitative Evaluation

Text-driven Motion Generation: We present the visualization results of our approach for long text generation with SMPL [30] model(Figure 1 and 4). The figure shows multiple motion results generated by our method, which are high quality and match the textual description very well. The generated motion types include walking on the ground and upstairs, running, crouching, crawling, flipping, and other actions that involve both movements in situ and global displacement, which fully demonstrate the superiority of our method. Moreover, our method can generate diverse sequences for the same input text while ensuring motion consistency. The diversity results and more visualization results are shown in the supplementary materials and demo.

Comparison with Previous Works: To demonstrate the effectiveness of our method, we conduct qualitative comparison on HumanML3D with MD [53] and T2M-GPT [52], which achieve the most similar results to us in the quantitative evaluation. The results are shown in Figure 5. The left sequences correspond to the text "a person quickly waves with their right hand", which emphasizes the words "quickly" and "wave right hand". We can see that the motion generated by MD [53] waves the left hand, while the motion generated by T2M-GPT [52] ignores the word "quickly" and pauses for a moment before waving. Our result matches the text description best. The sequences on
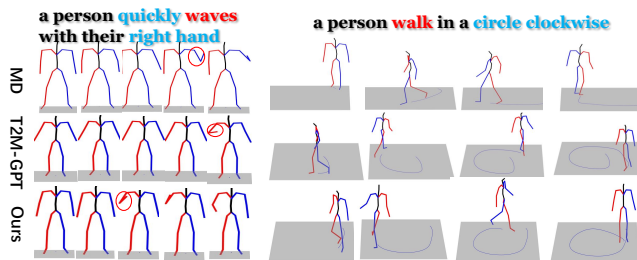

Figure 5. **Qualitative comparison.** We compare our method qualitatively with MD and T2M-GPT, who achieved the most similar results to us in the quantitative evaluation. The results generated by ours more precisely match the textual descriptions. For a more intuitive comparison, we demonstrate skeleton sequences here.
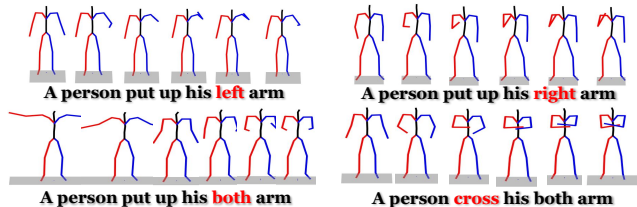

Figure 6. **Fine-grained generation**. Our method can generate high-quality motion when a single word in a given text is replaced, whether it's a verb or an adjective. For a more intuitive visualization, we demonstrate skeleton sequences here.

the right correspond to the text "a person walks in a circle clockwise". The motion generated by MD [53] stops after only half a circle, while the trajectory generated by T2M-GPT [52] is not a true circle. Our result's trajectory is a more seamless circle. The better performance on text-driven motion generation relies on the learning of the cross-modal relationship between text and motion, which allows our model to focus more attention on essential words.

Fine-grained Motion Generation: When only one word is changed in a sentence, the resulting change in the text encoding may be very small, but the corresponding motion sequence can be completely different. Thanks to the body-part attention-based spatial features extraction and
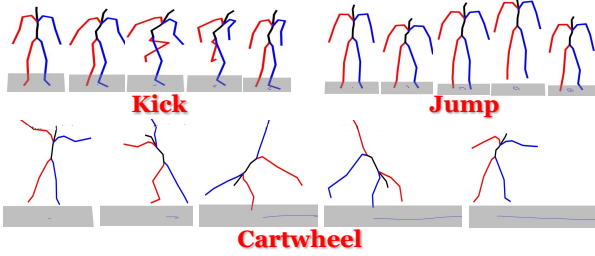
Figure 7. **Action to motion**. Our method can generate high-quality motion with a given action label expressed by a word or a phrase.

BPST	GA	LA	R Precision↑			FID.↓
			Top-1	Top-2	Top-3	
			$0.472 \pm .002$	$0.665 \pm .002$	$0.759 \pm .002$	$0.359 \pm .007$
✓	✓		$0.484 \pm .002$	$0.670 \pm .003$	$0.769 \pm .002$	$0.183 \pm .005$
✓		✓	$0.493 \pm .002$	$0.681 \pm .002$	$0.779 \pm .002$	$0.156 \pm .004$
	✓	✓	$\underline{0.495} \pm .002$	$\underline{0.682} \pm .002$	$\underline{0.781} \pm .002$	$0.213 \pm .004$
✓	✓	✓	$\mathbf{0.499} \pm .003$	$\mathbf{0.690} \pm .002$	$\mathbf{0.786} \pm .002$	$\mathbf{0.112} \pm .006$

Table 3. **Ablation study**. Ablation studies are conducted on HumanML3D to verify the effectiveness of body-part attention(BPST), local attention(LA), and global attention(GA). The results are labeled from top to bottom as AS1-AS5. The best results are marked in bold and the second best is underlined.

	Recon. Error↓	Commit Error↓
T2M-GPT(w/o BPST)	0.6512	2.1943
Ours(w/ BPST)	**0.6427**	**2.0701**

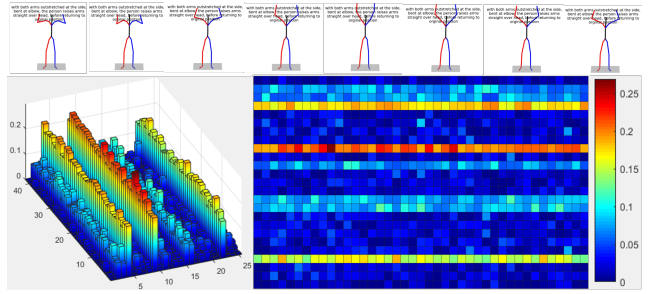Table 4. Comparison on reconstruction and commitment error of our VQ-VAE and T2M-GPT's.

Figure 8. **Motion-word cross-attention visualization**. We visualize the motion-word cross-attention in 2D and 3D heatmaps to demonstrate that our method does learn the different importance of words in the sentence in the process of text-driven generation.

the local cross-modal relationship learning, our method can achieve fine-grained motion synthesis by changing a single word in the sentence(Figure 6). As we modify the word "left" to "right" and "both", the generated motions change from raising the left hand to raising the right and both arms. In addition to replacing adjectives, when we change the verb by replacing "put up" with "cross", the model can also generate the motion of crossing both arms. This demonstrates that our method is very sensitive to small changes in the input text, which is more helpful in keeping the consistency between the generated motion and the text.

Action to Motion Generation: In addition to long text-driven motion generation, our method can also achieve high quality action to motion generation without any extra training, as shown in Figure 7. With only a single word or phrase representing the action as input, our method can still generate motion with high quality.

4.4. Ablation Study

We conducted an ablation study to investigate the effect of each module in our method on two aspects: the quality of the generated motion itself and its consistency with the given text description.

Effect of BPST: The high-quality low-dimensional motion representation can simplify motion modeling, making the generated motion more realistic and helpful for cross-modal learning between text and motion. To demonstrate the effect of our proposed BPST module, we remove the BPST and directly input the motion representation into the temporal convolutional layer to learn the VQ-VAE. The AS4 in Table 3 shows the results of this ablation study. We can see that the three R-Precision metrics decrease to some extent compared to the complete version(AS6), while FID increases significantly. R-Precision measures the consistency between the generated motion and text description, while FID reflects both the quality of the generated motion itself and the consistency between text and motion. The experimental results indicate that our BPST module significantly impacts the quality of generated motion while also improving the consistency between text and motion to some extent. In addition, we further demonstrated the effective-

ness of BPST by comparing the reconstruction error and commitment error of VQ-VAE with T2M-GPT(Table. 4).

Effect of GLA: The role of the GLA module is to learn the cross-modal relationship between text and motion, ensuring consistency between the generated motion and the given text. To demonstrate the effectiveness of our GLA, we keep the settings of stage 1 unchanged and remove $Trans_{local}$ and $Trans_{global}$ in turn. The AS2 and AS3 in Table 3 shows the results of this ablation study. We can see that compared to AS5, the R Precision of AS2 and AS3 decreases significantly, and FID decreases slightly. This indicates that GLA can significantly improve the realism of motion and consistency with the text description. In addition, we found that the performance of AS2 is worse than that of AS3, which indicates that LA plays a more important role in text-driven motion generation than GA.

Finally, we remove both BPST and GLA, and use vanilla VQ-VAE and concatenation of motion and text in vanilla transformer. The result is shown in AS1 of Table 3. All metrics became much worse, indicating that our method is reasonable and effective.
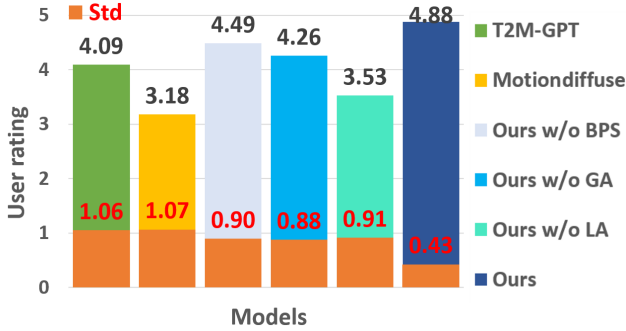
Figure 9. User perceptual study. The motion generated by our method is the one that best matches the textual description from the user's perspective.

4.5. Attention Visualization

To further explain the effectiveness and rationality of our method, we visualize the cross-attention between words and motion. We selecte a generated motion sequence of length 41 described by the text "With both arms outstretched at the side, bent at the elbow, the person raises arms straight overhead, before returning to the original position." The visualization of attention is shown in Figure 8 using 2D and 3D heatmaps. Rows 4, 15, and 22 of the 2D heatmap correspond to the words "outstretched," "raise," and "returning," which represent the main actions of this motion sequence and have the highest attention weights. Rows 2, 3, 9, 11, and 16 correspond to words such as "both", "arms", "bent", "elbow", and "arms", which are related to the body parts and states of them and relevant to the motion sequence. Other words and punctuation in the sentence have weak relationships with the motion and therefore have correspondingly lower weights. This visualization demonstrates that our method can indeed learn local cross-modal attention for each word, which is helpful for learning cross-modal relationships between text and motion. The body-part attention is shown in the supplementary materials.

4.6. User Perceptual Study

To illustrate the visual superiority of our approach, we did a user perceptual study, and the results are shown in Fig. 9. Specifically, we randomly select 30 texts in testset of HumanML3D and set up 10 manual texts to generate motion using the following 6 models: T2M-GPT, Motiondiffuse, Ours w/o BPS, Ours w/o GA, Ours w/o LA, Ours. We have 30 participants and ask them to rate the generated motion on a scale of 1-6(6 is the best). The evaluation criteria is a combination of motion realism and text-motion consistency. The highest score and lowest std indicate that our method works better and is more robust.

5. Limitations and Failure Cases

Despite achieving the state-of-the-art performance, our method still has some limitations and corresponding failure cases as follows(See the supplemental video for the visualization results).

(1) Insufficient diversity in long and detailed text-driven generation: Our method can generate motions with diversity. However, for very long texts with very detailed descriptions in the test set, the diversity of the motions generated by our method is insufficient. We observe that there are usually few motion sequences corresponding to long and detailed text in the HumanML3D dataset. When we want to generate multiple motions with them, the generated motions will be very similar.

(2) Fine-grained generation without ground truth or any similar sample in the dataset: As mentioned in the main paper, our approach enables fine-grained motion generation by changing a word or phrase in a sentence. However, when the motion represented by the changed word does not have ground truth in the dataset, or does not have any similar motion, our method does not yield results that match the text description well.

(3) Out-of-distribution Generation: When the given text does not lie within the distribution of the dataset, our method is unable to generate motions consistent to the text. But our method still tries very hard to find the motion similar to this text description.

6. Conclusion

We propose AttT2M, a two-stage method with multi-perspective attention mechanism for text-driven motion generation. Specifically, we use the spatial transformer based on body-part attention and temporal convolution to extract spatio-temporal motion features, and map them into a discrete latent space using VQ-VAE. Next, we learn the cross-modal relationship between the latent motion representation and the text utilizing global and local attention, which means calculating sentence-level conditional self-attention and word-level cross-attention. Finally, a generative transformer is trained to perform motion generation. Extensive experiments demonstrate that our method outperforms state-of-the-art methods in both qualitative and quantitative evaluations, and can accomplish fine-grained motion synthesis and action to motion.

References

[1] Carnegie mellon university: Cmu mocap dataset, http://mocap.cs.cmu.edu.

[2] Kfir Aberman, Peizhuo Li, Dani Lischinski, Olga Sorkine-Hornung, Daniel Cohen-Or, and Baoquan Chen. Skeleton-aware networks for deep motion retargeting. *ACM Transactions on Graphics (TOG)*, 39(4):62–1, 2020.

[3] Hyemin Ahn, Timothy Ha, Yunho Choi, Hwiyeon Yoo, and Songhwai Oh. Text2action: Generative adversarial synthesis from language to action. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5915–5920. IEEE, 2018.

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

[5] Sadegh Aliakbarian, Fatemeh Sadat Saleh, Mathieu Salzmann, Lars Petersson, and Stephen Gould. A stochastic conditioning scheme for diverse human motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5223–5232, 2020.

[6] Nikos Athanasiou, Mathis Petrovich, Michael J Black, and Gül Varol. Teach: Temporal action composition for 3d humans. In *International Conference on 3D Vision 2022*, 2022.

[7] Emad Barsoum, John Kender, and Zicheng Liu. Hp-gan: Probabilistic 3d human motion prediction via gan. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 1418–1427, 2018.

[8] Uttaran Bhattacharya, Nicholas Rewkowski, Abhishek Banerjee, Pooja Guhan, Aniket Bera, and Dinesh Manocha. Text2gestures: A transformer-based network for generating emotive body gestures for virtual agents. In *2021 IEEE virtual reality and 3D user interfaces (VR)*, pages 1–10. IEEE, 2021.

[9] Simon Clavet. Motion matching and the road to next-gen animation. In *Proc. of GDC*, 2016.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[11] Katerina Fragkiadaki, Sergey Levine, Panna Felsen, and Jitendra Malik. Recurrent network models for human dynamics. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4346–4354, 2015.

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

[13] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li 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.

[14] Chuan Guo, Xinxin Zuo, Sen Wang, and Li Cheng. Tm2t: Stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXV*, pages 580–597. Springer, 2022.

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

[16] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

[17] Daniel Holden, Oussama Kanoun, Maksym Perepichka, and Tiberiu Popa. Learned motion matching. *ACM Transactions on Graphics (TOG)*, 39(4):53–1, 2020.

[18] Daniel Holden, Taku Komura, and Jun Saito. Phase-functioned neural networks for character control. *ACM Transactions on Graphics (TOG)*, 36(4):1–13, 2017.

[19] Daniel Holden, Jun Saito, and Taku Komura. A deep learning framework for character motion synthesis and editing. *ACM Transactions on Graphics (TOG)*, 35(4):1–11, 2016.

[20] Fangzhou Hong, Mingyuan Zhang, Liang Pan, Zhongang Cai, Lei Yang, and Ziwei Liu. Avatarclip: Zero-shot text-driven generation and animation of 3d avatars. *arXiv preprint arXiv:2205.08535*, 2022.

[21] Lei Hu, Zihao Zhang, Chongyang Zhong, Boyuan Jiang, and Shihong Xia. Pose-aware attention network for flexible motion retargeting by body part. *IEEE Transactions on Visualization and Computer Graphics*, 2023.

[22] Deok-Kyeong Jang, Soomin Park, and Sung-Hee Lee. Motion puzzle: Arbitrary motion style transfer by body part. *ACM Transactions on Graphics (TOG)*, 41(3):1–16, 2022.

[23] Jogendra Nath Kundu, Maharshi Gor, and R Venkatesh Babu. Bihmp-gan: Bidirectional 3d human motion prediction gan. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 8553–8560, 2019.

[24] Hsin-Ying Lee, Xiaodong Yang, Ming-Yu Liu, Ting-Chun Wang, Yu-Ding Lu, Ming-Hsuan Yang, and Jan Kautz. Dancing to music. *Advances in Neural Information Processing Systems*, 32, 2019.

[25] Kyungho Lee, Seyoung Lee, and Jehee Lee. Interactive character animation by learning multi-objective control. *ACM Transactions on Graphics (TOG)*, 37(6):1–10, 2018.

[26] Jiaman Li, Yihang Yin, Hang Chu, Yi Zhou, Tingwu Wang, Sanja Fidler, and Hao Li. Learning to generate diverse dance motions with transformer. *arXiv preprint arXiv:2008.08171*, 2020.

[27] Zimo Li, Yi Zhou, Shuangjiu Xiao, Chong He, Zeng Huang, and Hao Li. Auto-conditioned recurrent networks for extended complex human motion synthesis. *arXiv preprint arXiv:1707.05363*, 2017.

[28] Angela S. Lin, Lemeng Wu, Rodolfo Corona, Kevin W. H. Tai, Qixing Huang, and Raymond J. Mooney. Generating animated videos of human activities from natural language descriptions. 2018.

[29] Hung Yu Ling, Fabio Zinno, George Cheng, and Michiel Van De Panne. Character controllers using motion vaes. *ACM Transactions on Graphics (TOG)*, 39(4):40–1, 2020.

[30] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 34(6):1–16, 2015.

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

[32] Mathis Petrovich, Michael J Black, and Gül Varol. Action-conditioned 3d human motion synthesis with transformer vae. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10985–10995, 2021.

[33] Mathis Petrovich, Michael J Black, and Gül Varol. Temos: Generating diverse human motions from textual descriptions. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXII*, pages 480–497. Springer, 2022.

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

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

[36] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training.

[37] Sebastian Starke, He Zhang, Taku Komura, and Jun Saito. Neural state machine for character-scene interactions. *ACM Trans. Graph.*, 38(6):209–1, 2019.

[38] Sebastian Starke, Yiwei Zhao, Taku Komura, and Kazi Zaman. Local motion phases for learning multi-contact character movements. *ACM Transactions on Graphics (TOG)*, 39(4):54–1, 2020.

[39] Sebastian Starke, Yiwei Zhao, Fabio Zinno, and Taku Komura. Neural animation layering for synthesizing martial arts movements. *ACM Transactions on Graphics (TOG)*, 40(4):1–16, 2021.

[40] Guy Tevet, Brian Gordon, Amir Hertz, Amit H Bermano, and Daniel Cohen-Or. Motionclip: Exposing human motion generation to clip space. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXII*, pages 358–374. Springer, 2022.

[41] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *The Eleventh International Conference on Learning Representations*, 2023.

[42] Guillermo Valle-Pérez, Gustav Eje Henter, Jonas Beskow, Andre Holzapfel, Pierre-Yves Oudeyer, and Simon Alexanderson. Transflower: probabilistic autoregressive dance generation with multimodal attention. *ACM Transactions on Graphics (TOG)*, 40(6):1–14, 2021.

[43] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

[44] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[45] Jacob Walker, Kenneth Marino, Abhinav Gupta, and Martial Hebert. The pose knows: Video forecasting by generating pose futures. In *Proceedings of the IEEE international conference on computer vision*, pages 3332–3341, 2017.

[46] Zhiyong Wang, Jinxiang Chai, and Shihong Xia. Combining recurrent neural networks and adversarial training for human motion synthesis and control. *IEEE transactions on visualization and computer graphics*, 27(1):14–28, 2019.

[47] Will Williams, Sam Ringer, Tom Ash, David MacLeod, Jamie Dougherty, and John Hughes. Hierarchical quantized autoencoders. *Advances in Neural Information Processing Systems*, 33:4524–4535, 2020.

[48] Shihong Xia, Lin Gao, Yu-Kun Lai, Ming-Ze Yuan, and Jinxiang Chai. A survey on human performance capture and animation. *Journal of Computer Science and Technology*, 32:536–554, 2017.

[49] Chen Xin, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, Jingyi Yu, and Gang Yu. Executing your commands via motion diffusion in latent space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023.

[50] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[51] He Zhang, Sebastian Starke, Taku Komura, and Jun Saito. Mode-adaptive neural networks for quadruped motion control. *ACM Transactions on Graphics (TOG)*, 37(4):1–11, 2018.

[52] Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Shaoli Huang, Yong Zhang, Hongwei Zhao, Hongtao Lu, and Xi Shen. T2m-gpt: Generating human motion from textual descriptions with discrete representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023.

[53] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *arXiv preprint arXiv:2208.15001*, 2022.

[54] Yan Zhang, Michael J Black, and Siyu Tang. Perpetual motion: Generating unbounded human motion. *arXiv preprint arXiv:2007.13886*, 2020.

[55] Chongyang Zhong, Lei Hu, Zihao Zhang, and Shihong Xia. Learning uncoupled-modulation cvae for 3d action-conditioned human motion synthesis. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXI*, pages 716–732. Springer, 2022.

[56] Chongyang Zhong, Lei Hu, Zihao Zhang, Yongjing Ye, and Shihong Xia. Spatio-temporal gating-adjacency gcn for human motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6447–6456, 2022.