AMD: Autoregressive Motion Diffusion

Bo Han*
borishan815@zju.edu.cn
Zhejiang University
Hangzhou, China

Hao Peng*
caspian.peng@unity.cn
Unity China
Shanghai, China

Minjing Dong
mdon0736@uni.sydney.edu.au
The University of Sydney
Syndney, Austrilia

Chang Xu
c.xu@sydney.edu.au
The University of Sydney
Syndney, Austrilia

Yi Ren
rayeren613@gmail.com
Zhejiang University
Hangzhou, China

Yixuan Shen
yshe0148@gmail.com
National University of Singapore
Singapore, Singapore

Yuheng Li
tuessica@gmail.com
Zhejiang University
Hangzhou, China

arXiv:2305.09381v2 [cs.MM] 17 May 2023

ABSTRACT

Human motion generation aims to produce plausible human motion sequences according to various conditional inputs, such as text or audio. Despite the feasibility of existing methods in generating motion based on short prompts and simple motion patterns, they encounter difficulties when dealing with long prompts or complex motions. The challenges are two-fold: 1) the scarcity of human motion-captured data for long prompts and complex motions. 2) the high diversity of human motions in the temporal domain and the substantial divergence of distributions from conditional modalities, leading to a many-to-many mapping problem when generating motion with complex and long texts. In this work, we address these gaps by 1) elaborating the first dataset pairing long textual descriptions and 3D complex motions (HumanLong3D), and 2) proposing an autoregressive motion diffusion model (AMD). Specifically, AMD integrates the text prompt at the current timestep with the text prompt and action sequences at the previous timestep as conditional information to predict the current action sequences in an iterative manner. Furthermore, we present its generalization for X-to-Motion with "No Modality Left Behind", enabling for the first time the generation of high-definition and high-fidelity human motions based on user-defined modality input.

CCS CONCEPTS

• **Computing methodologies → Activity recognition and understanding**.

*Both authors contributed equally to this research.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
MM '23, October 29–November 3,2023, Ottawa, Canada
© 2023 Association for Computing Machinery.
ACM ISBN 978-1-4503-XXXX-X/23/10…$15.00
https://doi.org/XXXXXXX.XXXXXXX

KEYWORDS

3D motion generation; text-to-motion; multi-modality

ACM Reference Format:
Bo Han, Hao Peng, Minjing Dong, Chang Xu, Yi Ren, Yixuan Shen, and Yuheng Li. 2023. AMD: Autoregressive Motion Diffusion. In *Proceedings of the 31th ACM International Conference on Multimedia (MM '23), October 29–November 3,2023,Ottawa, Canada.* ACM, New York, NY, USA, 10 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Human motion generation is a crucial task in computer animation and has applications in various fields including gaming, robots, and film. Traditionally, new motion is accessed through motion capture in the gaming industry, which can be costly. As a result, automatically generating motion from textual descriptions or audio signals can be more time-efficient and cost-effective. Related research work is currently flourishing, exploring human motion generation from different modalities [25, 49, 50, 55].

Current text-based conditional human motion synthesis approaches have demonstrated plausible mapping from text to motion [11, 34, 49, 54, 55]. They are mainly divided into three categories: Latent space strategy [1, 34, 48]: This is typically done by separately learning a motion Variational AutoEncoder (VAE) [21] and a text encoder, and then constraining them to a compatible latent space using the Kullback-Leibler (KL) divergence loss. However, since the distributions of natural language and human motion are vastly different, forcibly aligning these two simple Gaussian distributions can result in misalignments and diminished generative diversity. Diffusion-based approach [49, 52, 55]: diffusion models [15, 45] have recently attracted significant attention and have shown remarkable breakthroughs in various areas such as video [27], image [38], and 3D point cloud generation [13], etc. Current motion generation methods based on diffusion models [49, 52, 55] have achieved exceptional results using different denoising strategies.

Typically, MDM [49] proposes a motion diffusion model on raw motion data to learn the relationship between motion and text conditions. However, these models tend to only generate single motions or contain several motion sequences and are often inefficient for complex long texts. Autoregressive method [4, 10, 33]:

they can process varying motion lengths, tackling the issue of fixed motion duration. However, their single-step generation methods often rely on traditional VAE models [21], which are less effective than diffusion models. Despite the progress made by existing methods, text-based conditional human motion generation remains a challenging task for several reasons:

- Lack of enough motion-captured data: At present, there are few widely used text-to-motion datasets [11, 35, 36], which mostly contain simple motions and are deficient in long prompts, i,e., "he is flying kick with his left leg".
- Weak correlation: Due to the differing distributions of natural language and human motion, resulting in a multiple mapping problem [49]. This issue is further exacerbated when generating long text-based human motions.

To address the aforementioned limitations and challenges, we propose Autoregressive Motion Diffusion model (AMD) that can generate motion sequences with complex long content, variable duration, and multiple modalities. It leverages the generative capabilities of the diffusion model and the temporal modeling strengths of the autoregressive model. Considering the high dimensionality of complex long motion sequences, in order to better capture the dependencies between texts and motions in long sequences, AMD combines the text description at the current timestep with the text description and motion information at the previous timestep as conditional information to predict the motion sequence at the current timestep. AMD continuously employs the diffusion method to synthesize the corresponding motion sequence from the previous timestep and finally can generate the motion sequences of all texts. To address the scarcity of human motion-captured data for long prompts and complex motions, we have developed HumanLong3D - the first dataset to pair long textual descriptions with complex 3D human motions, i.e., "A person is doing martial art action raising knees and stretching feet, and then the person performs step forward with his right foot". The dataset comprises 158,179 textual descriptions and 43,696 3D human motions. It encompasses a broad spectrum of complex motion types. Importantly, it features annotations for motion coherence. In addition, we have also developed the HumanMusic dataset to evaluate the generation effect across different modalities. This dataset pairs 137,136 motions with corresponding audio data and follows the format of the HumanML3D dataset [11]. The codes for AMD and demos can be found in the supplementary materials.

In summary, our contributions include:

- We propose a novel continuous autoregressive diffusion model for generating complex and variable motions on long texts.
- We construct two large-scale cross-modal 3D human motion datasets HumanLong3D and HumanMusic, which could serve as the benchmarks for future cross-modal motion generation.
- Our proposed AMD achieves impressive performances on the HumanML3D, HumanLong3D, AIST, and HumanMusic datasets, which highlights its ability to generate high-fidelity motion given inputs with different modalities.

2 RELATED WORK

Human motion generation has been an active area of research for many years [5]. Early work in this field focused on unconditional motion generation [18, 30, 40], with some studies attempting to predict future motion based on an initial pose or starting motion sequence [8, 32]. Statistical models such as Principal Component Analysis (PCA) [31] and Motion Graphs [29] were commonly used for these generative tasks. The development of deep learning has led to the emergence of an increasing number of sophisticated generative architectures [9, 14, 20, 21, 51]. These advanced generative models have encouraged researchers to explore conditional motion generation. Conditional human motion generation can be modulated by a variety of signals that describe the motion, with high-level guidance provided through various means such as action classes [34], audio [3], and natural language [1, 34].

2.1 Text-To-Motion

Due to the language descriptors are the most user-friendly and convenient. Text-to-motion has been driving and dominating research frontiers. In recent years, the leading approach for the Text-to-Motion task is to learn a shared latent space for language and motion. JL2P [1] learns from the KIT-ML dataset [35] with an auto-encoder, limiting one-to-one mapping from text to motion. TEMOS [1] and T2M [11] propose using a VAE [21] to map a text prompt into a normal distribution in latent space. Recently, MotionCLIP [48] has leveraged the shared text-image latent space learned by CLIP to expand text-to-motion beyond data limitations and enable latent space editing. However, due to the inconsistency of the two data distributions of natural language and human motion, it is very difficult to align them in the shared latent space. Diffusion Generative Models [44] achieve significant success in the image synthesis domain, such as Imagen [42], DALL2 [38] and Stable Diffusion [39]. Inspired by their works, most recent methods [49, 52, 55] leverage diffusion models for human motion synthesis. MotionDiffuse [55] is the first work to generate human motion that corresponds to text utilizing a diffusion model. Recently, MDM [49] has been proposed, which operates on raw motion data to learn the relationship between motion and input conditions. Inspired by Stable Diffusion [39], MLD [52] implements the human motion diffusion process in the latent space. Despite their ability to produce exceptional results, these models are typically limited to short text descriptions and simple motions. Additionally, several works [4, 10, 33] have been developed based on the concept of autoregression, which can generate human actions of any length. Consequently, for long text prompts, we combine the advantages of the diffusion model in generating motion for short text descriptions with the concept of autoregression to achieve superior human motion results for continuous long text.

2.2 Motion Datasets

Common forms of description for human motion data are 2D keypoints, 3D keypoints, and statistical model parameters [6, 53]. For the text-conditioned motion generation task, KIT [35] is the first 3D human motion dataset with matching text annotations for each motion sequence. HumanML3D [11] provides more textual annotation for some motions of AMASS [28]. They are also our focus in the text-to-motion task. Babel [36] also collects motions from AMASS [28] and provides action and behavior annotations, it annotates each frame of the action sequence, thereby dividing compound

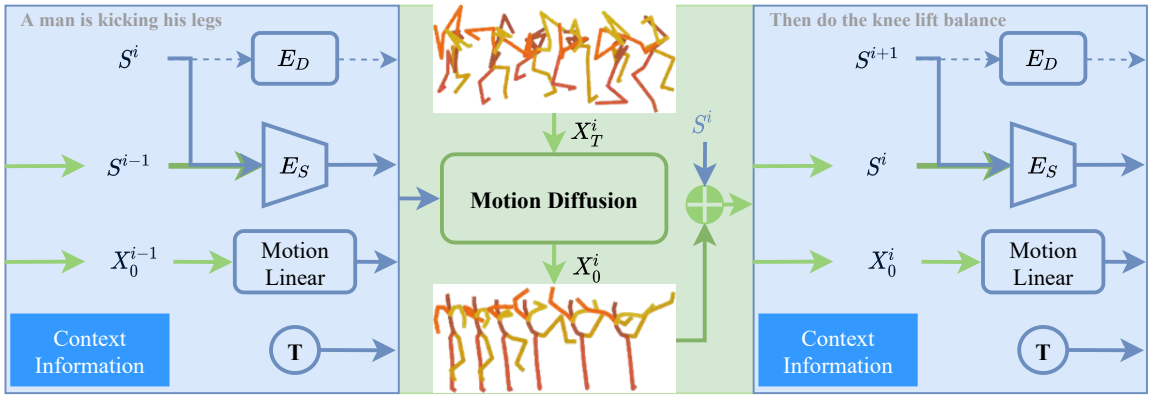


Figure 1: Overview of the Autoregressive Motion Diffusion model (AMD). Given the current timestep text prompt S^i (A man is kicking his legs), the last timestep text prompt S^{i-1}, and motion X_0^{i-1} (green arrow), we first encode the context information (blue block). Then, we feed the input conditions and corrupted motion X_T^i to Motion Diffusion Mudule (Figure. 2) to generate the original cleaned motion X_0^i. Afterward, we send the current timestep text prompt S^i and motion X_0^i to the next time step. Iteratively, we can obtain motion sequences for long text prompts.

actions into simple action groups. In this paper, we use the HumanML3D dataset to evaluate the proposed methods for simple motions and short prompts. In addition, we collected and labeled pairs of complex motion data and text prompts (HumanLong3D). More importantly, we provided temporal motion-coherence information to support long text-to-motion generation tasks.

2.3 Audio-To-Motion

Generating natural and realistic human motion from audio is also a challenging problem. Many early approaches follow a motion retrieval paradigm [7, 23]. A traditional approach to motion synthesis involves constructing motion maps. New motions are synthesized by combining different motion segments and optimizing transition costs along graph paths [41]. More recent approaches employ RNN [2, 17, 47], GANs [22, 46], Transformer [24, 25, 43], and CNN [16] models to map the given music to a joint sequence of the continuous human pose space directly. Such methods would regress to nonstandard poses that are beyond the dancing subspace during inference. In contrast, our proposed method does not produce the phenomenon of limb drift.

3 OUR APPROACH

In this section, we introduce the problem formulation for semantic-driven human motion generation. To enable adaptive motion generation for different prompts, we propose the inclusion of a motion duration prediction network to approximate the duration. To generate human motions that correspond to continuous long text descriptions, we establish a connection between an autoregressive encoder and the diffusion model, incorporating information from both the previous motion sequence and the text prompt.

3.1 Problem Description

To generate complex motion sequences with long-term text prompts, we propose to feed multiple text prompts in order. Given N text prompts $S^{1:N} = \{S^1, S^2, \ldots, S^N\}$, the model is required to generate N motion segments $X^{1:N} = \{X^1, X^2, \ldots, X^N\}$ consistent with the text descriptions, where N denotes the number of motion segments involved in the entire motion sequence. Each motion segment is defined as $X^i = \left\{x^1, x^2, \ldots, x^{F^i}\right\}$, where F^i is the total number of frames of the motion segment X^i and x^j denotes the 3D human body pose representation of the j-th frame. It is imperative that each generated motion segment and the corresponding number of motion frames adhere to the specifications outlined in the text prompt. Additionally, a seamless transition from X^{i-1} to X^i is crucial for the generation of high-fidelity motion.

3.2 Motion Duration Prediction Network

Given a semantic prompt S^i describing a motion, the duration of each X^i may vary. For instance, in the HumanML3D dataset [11], the prompt "a man kicks something or someone with his left leg" corresponds to 116 motion frames, while the prompt "a person squats down then jumps" corresponds to 35 motion frames. Consequently, we propose predicting the motion duration in order to generate motions with adaptive length. Following T2M [11] we use probability density estimation to determine the number of frames required for the motion synthesis based on text prompts. Due to the diversity of the duration of motion clips, it is more reasonable to model the mapping problem of text-to-duration as a density estimation problem than directly regressing the specific value. By utilizing the semantic prompt S^i as input for the motion duration prediction network, a probability density estimation is conducted on the discrete group encompassing all possible motion durations $L = \{L_{min}, L_{min}+1, \ldots, L_{max}\}$. The discrete duration probability density can be formulated as:

$$p(L|S^i) = \left\{p(L_{min}|S^i), p(L_{min}+1|S^i), \ldots, p(L_{max}|S^i)\right\}. \quad (1)$$

Therefore, the loss function of the network is designed as the cross-entropy loss of multi-classification, as depicted in Equation 2:

$$\mathcal{L}_{CE} = -\sum_{d=L_{min}}^{L_{max}} l_d \log(p(d|S^i)), \tag{2}$$

where l is the one-hot encoding of the ground truth duration, if and only when the duration is d, l_d is equal to 1, otherwise 0.

3.3 Autoregressive Iteration

It is important to note that daily human motions encompass not only simple, single motions but also complex, prolonged motions that more accurately reflect real-life scenarios. Specifically, given a series of semantic prompts $S^{1:N}$, a series of randomly sampled temporal motion sequences $X_T^{1:N} \sim \mathcal{N}(0, I)$ obeying the standard normal distribution, and a maximum noise scale $T \in \mathbb{N}$ where each semantic prompt S^i describes a single and distinct motion. Our goal is to generate noise-free temporal motion sequences $X_0^{1:N}$, which are guided by the semantic prompts, with smooth transitions between adjacent motions X_0^{i-1} and X_0^i. The overall process is illustrated in Figure 1, and each pair of blue and green blocks represents each step of the AMD model. $S^{1:N}$ employs the model iteratively to synthesize motion $X_0^{1:N}$. The blue block represents the context encoder and the green block is the motion diffusion module.

3.3.1 Context Encoder. It includes the motion duration prediction Network E_D, the semantic conditional encoder E_S, and the motion linear layer. The CLIP model [37] is utilized as the semantic conditional encoder. Given that our primary focus is on long text-to-motion generation, it is necessary to consider timing-related information associated with long texts. To this end, we encode the previous motion X_0^{i-1} by the motion linear layer to obtain z_m^{i-1} and encode semantic information S^{i-1} by the semantic conditional encoder E_S to obtain z_c^{i-1}. These are then concatenated to form the final prior condition feature z_{past}^{i-1}. Simultaneously, the current semantic information is input into the motion duration prediction network and semantic conditional encoder to obtain F^i and z_c^i, respectively. In order to avoid overfitting, we perform a random mask on the semantic conditional information z_c^i. For the corrupted motion X_t^i, the same motion linear layer is utilized to obtain the encoded information z_m^i. We feed the diffusion time scale t to a Multi-layer Perceptron (MLP) to obtain the time embedding z_t. The final condition information z is defined as follows:

$$z = C(C(C(z_m^{i-1}, z_c^{i-1}) + RM(z_c^i), z_t), z_m^i, PE(F^i)) \tag{3}$$

In Equation 3 above, C represents the concatenation operation, RM denotes Random Mask, and PE refers to potion embedding. It is important to note that during training, we utilize the actual motion duration present in the dataset, whereas during the inference phase, the predicted duration information is used.

3.3.2 Motion Diffusion. The network architecture of the motion diffusion module is depicted in Figure 2. The denoising process (red) and the diffusion process (yellow) span a total of T timesteps, where T represents the pre-defined maximum time scale. The objective of the denoising process is to predict the original, cleaned motion X_0^i, while the diffusion process operates in the opposite direction. During the denoising process, we commence from the current denoising timescale DE_t and directly predict the coarse raw motion $\hat{X}_0^i$, as shown in Equation 4.

$$p_\theta(\hat{X}_0^i) := p(X_t^i) \prod_{t=1}^{DE_t} p_\theta(X_{t-1}^i | X_t^i, z). \tag{4}$$

The single step of the denoising process is essentially the transfer process from X_t^i to X_{t-1}^i, the transfer strategy requires a network with parameters θ to learn the sampling distribution as:

$$\begin{aligned} p_\theta(X_{t-1}^i | X_t^i, z) &:= q(X_{t-1}^i | \hat{X}_0^\theta(X_t^i, t, z)) \\ &= \mathcal{N}(X_t^i; \sqrt{\bar{\alpha}_t}\hat{X}_0^\theta(X_t^i, t, z), (1-\bar{\alpha}_t)I), \end{aligned} \tag{5}$$

where $\hat{X}_0^\theta(X_t^i, t, z)$ represents the neural network with parameter θ, which takes in X_t^i, t, and conditional information z as input. In the diffusion process, we start from the coarse raw motion $\hat{X}_0^i$ predicted by the last denoising process and the current diffusion timescale DI_t. $\hat{X}_0^i$ passes through a series of Markov random noises and finally transfers to a noise motion that approximately obeys the standard Gaussian normal distribution. The single step of the diffusion process is essentially the transform from X_{t-1}^i to X_t^i, which is defined in Equation 6, where the β_t is pre-defined. The transition probability formula of the diffusion process from X_0^i to X_t^i can be derived from Equation 6, which is defined as shown in Equation 7, where $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^t \alpha_s$.

$$q(X_t^i | X_{t-1}^i) := \mathcal{N}(X_t^i; \sqrt{1-\beta_t}X_{t-1}^i, \beta_t I), \tag{6}$$

$$q(X_t^i | \hat{X}_0^i) := \prod_{t=1}^{DI_t} q(X_t^i | X_{t-1}^i) = \mathcal{N}(X_t^i; \sqrt{\bar{\alpha}_t}\hat{X}_0^i, (1-\bar{\alpha}_t)I). \tag{7}$$

The relationship of the timestep t, denoising timescale DE_t, and diffusion timescale DI_t is $t + DE_t = T$ and $t + DI_t = T - 1$, where $t = [0, 1, ..., T-1]$. For example, in the timestep 0, the input is the denoising timescale T, the noise motion X_T^i, and the current semantic conditional feature z, the output is the predicted rough original motion $\hat{X}_0^i = (\hat{X}_0^{i1}, \hat{X}_0^{i2}, ...\hat{X}_0^{iF^i})$, where F^i is the frame number. At the same time, the input of the coupled diffusion process is diffusion timescale $T-1$ and $\hat{X}_0^i$, the output is $\hat{X}_{T-1}^i$. The time iteration process is shown in Figure 3. The lengths of the red and yellow blocks represent the size of the denoising timescale and the diffusion timescale, respectively. It is worth mentioning that the last iteration only has the denoising process.

3.3.3 Loss Function. Following each timestep, we directly predict the original motion sequence $\hat{X}_0^i$ and optimize the diffusion model parameters by measuring the MSE loss between $\hat{X}_0^i$ and ground truth X_0^i. However, the motion sequence parameters include the rotation, position, speed information of the human body posture, and the static judgment information of the foot joints. In order to more accurately measure the difference between the generated motion and the real motion, we designed the loss function, including $\mathcal{L}_{height}$, $\mathcal{L}_{pos}$, $\mathcal{L}_{rot}$, $\mathcal{L}_{vel}$, and $\mathcal{L}_{foot}$. Refer to the sector 4.2, we define the motion of the jth frame as

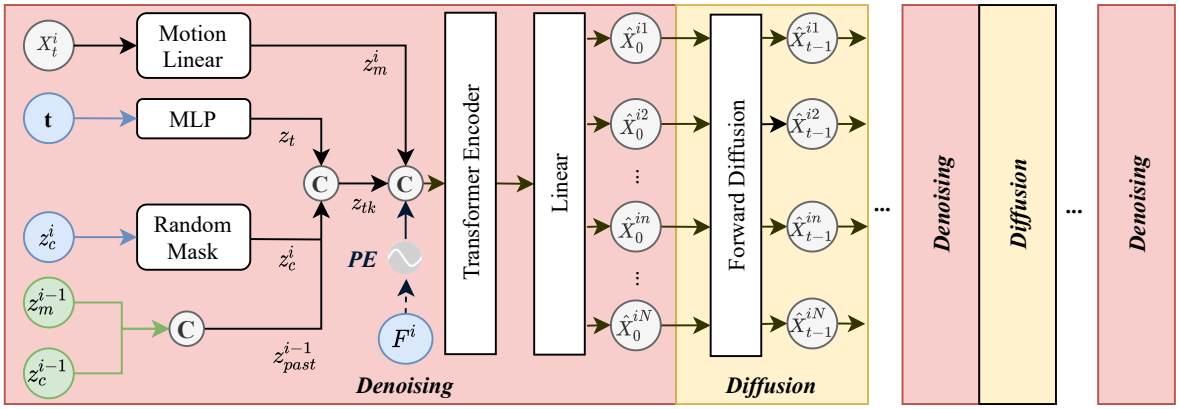


Figure 2: Motion Diffusion Module. The red blocks denote the denoising process, while the yellow blocks represent the diffusion process. Within the motion diffusion module, they appear in pairs T times (with the exception of the last one).

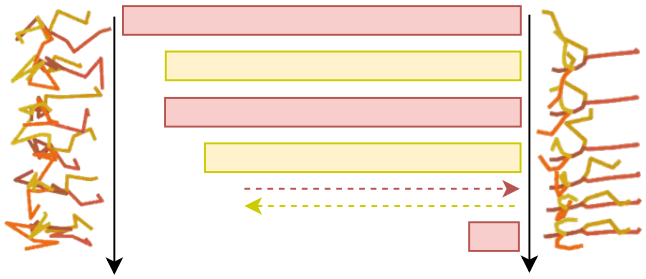


Figure 3: Time iteration. The red block and yellow block represent the denoising process and diffusion process respectively. The left side shows the result of the diffusion process, and the right side is the result of the denoising process.

$x_j = \left\{\dot{r}_j^{root}, \dot{g}_j^{root}, h_j^{root}, g_j^{loc}, r_j^{loc}, \dot{g}_j^{loc}, f_j\right\}$. The loss function of each part of human motion is defined as:

$$\mathcal{L}_{height} = \frac{1}{F}\sum_{j=1}^{F} \| \hat{h}_j^{root} - h_j^{root} \|_2^2, \tag{8}$$

$$\mathcal{L}_{pos} = \frac{1}{F}\sum_{j=1}^{F} \| \hat{g}_j^{loc} - g_j^{loc} \|_2^2, \tag{9}$$

$$\mathcal{L}_{rot} = \frac{1}{F}\sum_{j=1}^{F} \| \hat{r}_j^{loc} - r_j^{loc} \|_2^2, \tag{10}$$

$$\mathcal{L}_{vel} = \frac{1}{F}\sum_{j=1}^{F} (\| \hat{\dot{r}}_j^{root} - \dot{r}_j^{root} \|_2^2 + \| \hat{\dot{g}}_j^{root} - \dot{g}_j^{root} \|_2^2 + \| \hat{\dot{g}}_j^{loc} - \dot{g}_j^{loc} \|_2^2), \tag{11}$$

$$\mathcal{L}_{foot} = \frac{1}{F}\sum_{j=1}^{F} \| \hat{f}_j - f_j \|_2^2 . \tag{12}$$

Among them, $\mathcal{L}_{height}$ represents the height loss, which is used to measure the mean square error of the y-axis height between the generated motion and the real motion. The $\mathcal{L}_{pos}$ represents the joint position loss, which is used to measure the mean square error of the three-dimensional joint point position coordinates in the local coordinate system between the generated motion and the real motion. The $\mathcal{L}_{rot}$ represents the joint rotation loss, which is used to measure the mean square error of the six-dimensional joint rotation in the local coordinate system between the generated motion and the real motion. The $\mathcal{L}_{vel}$ represnts the joint velocity loss, which is used to measure the mean square error of the linear velocity and angular velocity of each joint between the generated motion and the real motion. The $\mathcal{L}_{foot}$ represents the sliding foot loss, which is used to measure the mean square error of the static labels of the foot joints between the generated motion and the real motion. Finally, the loss function is defined as

$$\mathcal{L}_{train} = \lambda_{height}\mathcal{L}_{height} + \lambda_{pos}\mathcal{L}_{pos} + \lambda_{rot}\mathcal{L}_{rot} + \lambda_{vel}\mathcal{L}_{vel} + \lambda_{foot}\mathcal{L}_{foot}, \tag{13}$$

where λ denotes the coefficients to balance the loss terms.

With the proposed ADM, we are able to generate motion sequences according to ordered semantic prompts in an iterative manner. Specifically, we commence from the first prompt S^1 and utilize the autoregressive diffusion model to synthesize the corresponding clean motion sequence X_0^1. The remaining high-fidelity motion sequences $X_0^{2:N}$ can be synthesized using prior condition information as well as $S^{2:N}$. Ultimately, a coherent motion sequence of any length can be synthesized.

3.3.4 Classifier Free. The unconditionally guided synthetic network and the conditionally guided synthetic network are trained simultaneously, and the training sample size of the unconditionally guided synthetic network accounts for 10% of the training set. When sampling in the inference stage, the result $\hat{X}_0^{\theta'}(X_t, t, c)$ of the denoising network will be linearly interpolated by the results of $\hat{X}_0^{\theta}(X_t, t)$ and $\hat{X}_0^{\theta}(X_t, t, c)$, and the interpolation calculation method is shown in equation 14, where the interpolation weight $\omega = 2.5$

$$\hat{X}_0^{\theta'}(X_t, t, c) = (1-\omega)\hat{X}_0^{\theta}(X_t, t) + \omega\hat{X}_0^{\theta}(X_t, t, c) \tag{14}$$

Method	R-Precision(top3)↑	FID↓	MultiModal Dist↓	Diversity→	MultiModality↑
T2M [11]	$0.740^{\pm.003}$	$1.067^{\pm.002}$	$3.340^{\pm.008}$	$9.188^{\pm.002}$	$2.090^{\pm.083}$
MotionDiffuse [55]	$0.782^{\pm.001}$	$0.630^{\pm.001}$	$3.113^{\pm.001}$	$9.410^{\pm.049}$	$1.553^{\pm.042}$
MDM [49]	$0.611^{\pm.007}$	$0.544^{\pm.044}$	$5.566^{\pm.027}$	$9.559^{\pm.086}$	$2.799^{\pm.072}$
T2M-GPT [54]	$0.775^{\pm.002}$	$0.141^{\pm.005}$	$3.121^{\pm.009}$	$9.722^{\pm.082}$	$1.831^{\pm.048}$
MLD [52]	$0.772^{\pm.002}$	$0.473^{\pm.013}$	$3.196^{\pm.010}$	$9.724^{\pm.082}$	$2.413^{\pm.079}$
Ours	$0.617^{\pm.014}$	$0.586^{\pm.107}$	$5.469^{\pm.063}$	$9.769^{\pm.096}$	$2.512^{\pm.232}$
GT	$0.797^{\pm.002}$	$0.002^{\pm.000}$	$2.974^{\pm.008}$	$9.503^{\pm.065}$	-

Table 1: Single motion synthesis evaluation on HumanML3D Dataset. All methods use the real motion length from the ground truth except ours and T2M-GPT. → means results are better if the metric is closer to the real distribution. We run all the evaluation 20 times (except MultiModality runs 5 times) and ± indicates the 95% confidence interval.

4 EXPERIMENTS

4.1 Datasets and Evaluation Metrics

HumanML3D[11] The dataset involves the textual re-annotation of motion capture data from the AMASS [28] and HumanAct12 [12], comprising 14,616 motions annotated with 44,970 textual descriptions.

HumanLong3D We collected motion data using motion capture equipment and online sources, and annotated each motion sequence with various semantic labels to create the HumanLong3D dataset. The data format of the HumanLong3D dataset is consistent with that of HumanML3D, and it additionally includes coherence information for motion sequences to support temporal motion generation tasks. Further details about the HumanLong3D dataset can be found in the Supplementary Material.

HumanMuisic We collected dance videos from online sources and extracted the pose parameters of the dancers in the videos, converting the motion data into the HumanML3D format. For the music data, we used the public audio toolbox Librosa [19] to extract music features, including mel frequency cepstral coefficients (MFCC), MFCC delta, constant-Q chromagram, tempogram, and onset strength, resulting in a total of 438 dimensions. In total, we obtained 137,136 paired dance and music data samples, with each dance sample consisting of 200 frames. Further details about the HumanMusic dataset can be found in the Supplementary Material.

AIST++ [25] This dataset comprises 992 high-quality 3D pose sequences in SMPL format [26], captured at 60 FPS, with 952 sequences designated for training and 40 for evaluation. We followed the approach of Bailando [43] to partition the AIST++ dataset.

Dataset	Motion	Textual descriptions	Duration
KIT-ML	3911	6248	10.33h
HumanML3D	14616	44970	28.59h
HumanLong3D	43696	158179	85.87h

Table 2: Dataset description

Evaluation Metrics For text-to-motion evaluation, we employ metrics consistent with existing methods [49, 55]. Specifically, (a) Frechet Inception Distance (FID) is used as the primary metric to evaluate the feature distributions between generated and real motions in feature space [11], and (b) R-Precision (top 3) calculates the top 3 matching accuracy between text and motion in feature space.(c) *MultiModal Dist* calculates the distance between motions and texts. (d) *Diversity* measures variance through features. (e) *MultiModality* assesses the diversity of generated motions for the same text. For music-to-dance evaluation, we employ metrics consistent with existing methods [17, 43]. Refer to Supplementary, Chapter 4 for more evaluation metric details.

4.2 Experimental Settings

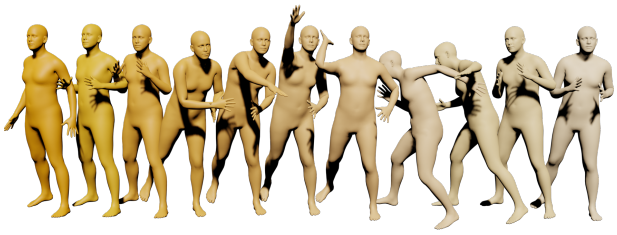
Motion Representation Our motion representation adopts the same format as HumanML3D [11], i.e., $X \in \mathbb{R}^{263\times F}$. Each frame of motion is 263-dimensional data, including the position, linear velocity, angular velocity, joint space rotation of three-dimensional human joints, and label information for judging whether the foot joints are still. The motion of a single frame is represented as $x = \left\{\dot{r}^{root}, \dot{g}^{root}, h^{root}, g^{loc}, r^{loc}, \dot{g}^{loc}, f\right\}$. Since images are often represented as $I \in \mathbb{R}^{W\times H\times C}$, in order to naturally transfer motions to image-based diffusion models, we upscale $X \in \mathbb{R}^{263\times F}$ to $X \in \mathbb{R}^{263\times F\times 1}$. Refer to Supplementary, Chapter 3 for more details.

Motion Duration Prediction Network L_{min} is set to 10 and L_{max} is 50, each unit increment corresponds to 4 motion frames, i.e., 0.2s motion duration, so the duration prediction range covers the lower bound of 2s and the upper bound of 9.8s of the data samples. The motion duration prediction network and the diffusion model are trained independently, with the motion duration prediction network being used only during inference.

Motion Diffusion We set the maximum noise scale T to be 1000, the coefficient $\beta_{1:T}$ is set to a linear increment from 10^{-4} to 0.02, latent vector dimensions are 512, the number of layers of the motion encoder is 6, and the number of heads of the multi-head attention mechanism is set to 6, the learning rate is fixed at 10^{-4}, the number of training steps is 200000, and we use AdamW optimizer.

Other Settings The output dimension of the motion linear layer and the latent vector dimension of the motion diffusion module are both 512. The semantic conditional encoder adopts the CLIP-ViT-B/32 model. During inference, the semantic prompt S^i is input into the motion duration prediction network E_D to obtain the estimated value F^i of the motion sequence duration, which is used to determine the timing dimension for motion sequence sampling.

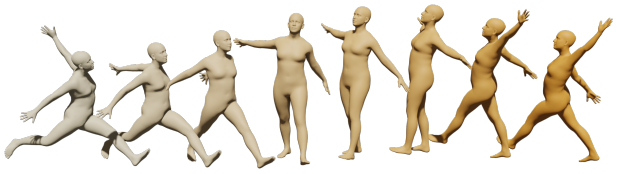
(a) the person picks an object up off the floor with their left hand



(b) a person throws an object with his right hand.

Figure 4: Visualization on HumanML3D Dataset. The darker colors indicate the later in time.

(a) the motion he is doing is called Mulan boxing fifth road.



(b) it is an motion named support the ground with bending knees.

Figure 5: Visualization on HumanLong3D Dataset

Comparisons on Single Motion We compared single motion generation with existing state-of-the-art methods. For single motion generation, our conditional information includes the estimated motion duration value and semantic information but excludes prior motion and semantic information. The visualization results are shown in Figure 4 and Figure 5. It can be seen that AMD is capable of generating corresponding motion in response to text prompts containing a single motion while achieving smooth transitions. As can be seen from Table 1 and Table 3, AMD achieves SOTA performance on single motion generation and can infer the motion duration, which is beneficial for actual choreographing motions.

Since the motion duration prediction network is trained independently and used only during inference, we evaluate its performance separately. We perform four-fold cross-validation on the KIT-ML, HumanML3D, and HumanLong3D datasets, using brier-score and cross-entropy loss as density estimation evaluation metrics. Refer to Supplementary, Chapter 2 for details.

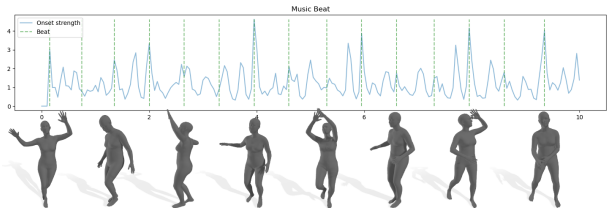


Figure 6: Visualization Results. The top figure represents the musical beat. The bottom figure illustrates the motion generated by our model.

Comparisons on Music Dance We conducted a comparison between our method and the state-of-the-art (SOTA) approach using the AIST dataset and HumanMusic dataset. Our dataset division methodology was identical to that of the SOTA methods [43], and we converted the data in AIST++ into HumanML3D format. As illustrated in Figure 6, the movements generated by AMD were in sync with the beat of the music. The results of this comparison are presented in Table 4. Our method achieved performance on par with the SOTA approach, with particularly notable improvements in terms of diversity. Further details about the evaluation of the music-dance task can be found in the Supplementary Material.

4.3 Comparisons on Compound Motion

Method	Motion Quality		Motion Diversity		
	$FID_k \downarrow$	$FID_g^\dagger \downarrow$	$Div_k \uparrow$	$Div_g^\dagger \uparrow$	BAS ↑
DanceNet [56]	69.18	25.59	2.86	2.85	0.1430
DanceRevolution [17]	73.42	25.92	3.52	4.87	0.1950
FACT [25]	35.35	22.11	5.94	6.18	0.2209
Bailando [43]	28.16	9.62	7.83	6.34	0.2332
Ours	32.21	18.72	21.24	16.53	0.2158

Table 4: Music-Dance evaluation on AIST++ Dataset

We compare compound motion generation with SOTA methods [49, 52, 54, 55]. Since the HumanML3D dataset does not contain motion coherence information, we conducted this experiment only on the HumanLong3D dataset, and we divided the dataset into training, test, and validation sets using a ratio of 0.85:0.10:0.05. Additionally, we designed three benchmarks based on TEACH [4]: 1) Joint prediction (ours-J): The long semantic prompt $S^{i-1:i}$ formed by the combination of two coherent prompts are used as the input of the diffusion model, and a coherent time-series motion sequence $X_0^{i-1:i}$ is obtained by direct joint prediction. 2) Linear interpolation (ours-I): This method interpolates the results of two independent motion synthesis. 3) Motion filling (ours-F): Similar to linear interpolation, two independent motion synthesis are required to obtain X_0^{i-1} and X_0^i, and the time window is set to 10% of the motion sequence duration. All frame data except for the time window are fixed, and the frame data within the time window are filled with random normal distribution noise. The coherent motion sequence is then restored through the denoising process.

Method	R-Precision(top3)↑	FID↓	MultiModal Dist↓	Diversity→	MultiModality↑
T2M [11]	$0.197^{\pm.007}$	$1.352^{\pm.033}$	$5.777^{\pm.021}$	$5.584^{\pm.073}$	$2.742^{\pm.315}$
MotionDiffuse [55]	$0.191^{\pm.007}$	$1.171^{\pm.049}$	$5.801^{\pm.016}$	$5.221^{\pm.091}$	$3.123^{\pm.098}$
MDM [49]	$0.152^{\pm.004}$	$0.721^{\pm.024}$	$8.058^{\pm.021}$	$5.035^{\pm.084}$	$2.727^{\pm.027}$
T2M-GPT [54]	$0.189^{\pm.003}$	$0.350^{\pm.018}$	$5.613^{\pm.016}$	$5.046^{\pm.061}$	$2.735^{\pm.057}$
MLD [52]	$0.173^{\pm.003}$	$0.857^{\pm.023}$	$5.815^{\pm.012}$	$4.815^{\pm.052}$	$3.052^{\pm.080}$
Ours	$0.150^{\pm.004}$	$0.745^{\pm.027}$	$8.062^{\pm.019}$	$5.047^{\pm.075}$	$2.835^{\pm.257}$
GT	$0.208^{\pm.003}$	$0.006^{\pm.000}$	$5.294^{\pm.007}$	$4.977^{\pm.048}$	-

Table 3: Single motion synthesis evaluation on HumanLong3D Dataset.

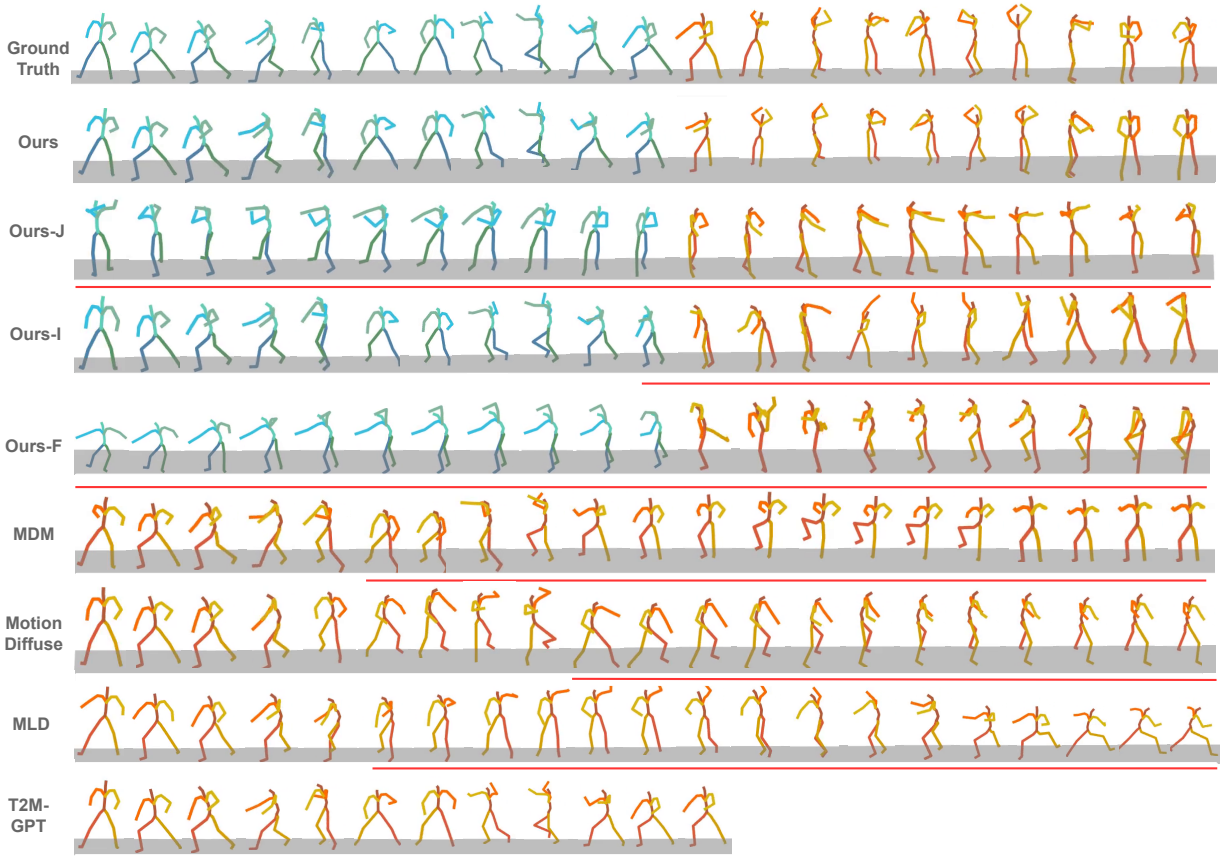


Figure 7: Result for compound motion synthesis (blue: "there is a man doing left smash right cover." yellow: "the motion he is doing is called step forward and turn around"). The part delineated by the red line indicates a discrepancy between the generated motion and the ground truth.

As shown in Table 5, Among the five evaluation metrics, AMD achieved top 3 performance in three of them, with its FID and Diversity scores ranking first. Notably, AMD outperformed other methods by a significant margin in the FID metric, which measures the similarity of generated motions. As illustrated in Figure 7, compared to the ground truth, AMD keeps with the highest degree of similarity, while MDM, MotionDiffuse, and MLD all exhibited varying degrees of limb stiffness. Although T2M-GPT achieves results comparable to the ground truth in the first half of motion generation, its performance deteriorates in longer text-to-motion generation tasks. This is due to its premature prediction of the terminator, resulting in a lack of corresponding motion sequence for the second half of the text.

5 CONCLUSION

In this paper, we present the HumanLong3D - the first dataset that pairs complex motions with long textual descriptions to address the

Method	R-Precision(Top3)↑	FID↓	MultiModal Dist↓	Diversity→	MultiModality↑
Ours-J	$0.122^{\pm.005}$	$12.813^{\pm.085}$	$9.898^{\pm.032}$	$4.610^{\pm.104}$	$3.527^{\pm.264}$
Ours-I	$0.141^{\pm.004}$	$0.739^{\pm.030}$	$7.978^{\pm.043}$	$4.364^{\pm.084}$	$2.281^{\pm.223}$
Ours-F	$0.138^{\pm.008}$	$0.589^{\pm.029}$	$7.872^{\pm.021}$	$4.337^{\pm.076}$	$2.248^{\pm.229}$
MDM [49]	$0.096^{\pm.005}$	$27.348^{\pm.349}$	$7.203^{\pm.039}$	$0.781^{\pm.040}$	$0.547^{\pm.037}$
MotionDiffuse [55]	$0.157^{\pm.004}$	$6.860^{\pm.113}$	$6.783^{\pm.028}$	$4.529^{\pm.076}$	$2.409^{\pm.204}$
T2M-GPT [54]	$0.166^{\pm.003}$	$1.249^{\pm.026}$	$5.903^{\pm.017}$	$4.895^{\pm.047}$	$3.093^{\pm.104}$
MLD [52]	$0.144^{\pm.002}$	$3.843^{\pm.058}$	$6.540^{\pm.011}$	$4.365^{\pm.033}$	$2.831^{\pm.072}$
Ours	$0.154^{\pm.005}$	$0.215^{\pm.017}$	$7.719^{\pm.039}$	$4.515^{\pm.135}$	$1.242^{\pm.118}$
GT	$0.160^{\pm.003}$	$0.001^{\pm.000}$	$7.309^{\pm.017}$	$4.452^{\pm.069}$	-

Table 5: Compound motion generation evaluation on HumanLong3D Dataset. For each metric, we repeat the evaluation 20 times (except MultiModality runs 5 times). Red, Blue, and Green indicate the first, the second, and the third best result.

scarcity of such data. Given the suboptimal performance of current motion generation methods on long text descriptions, we introduce a novel network architecture AMD, which combines autoregressive and diffusion models to effectively capture the information contained in long texts. Furthermore, we extend our approach to incorporate audio conditional input and construct a large-scale music-dance dataset - HumanMusic.

ACKNOWLEDGMENTS

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] Omid Alemi, Jules Françoise, and Philippe Pasquier. 2017. GrooveNet: Real-time music-driven dance movement generation using artificial neural networks. *networks* 8, 17 (2017), 26.

[3] Andreas Aristidou, Anastasios Yiannakidis, Kfir Aberman, Daniel Cohen-Or, Ariel Shamir, and Yiorgos Chrysanthou. 2022. Rhythm is a Dancer: Music-Driven Motion Synthesis with Global Structure. *IEEE Transactions on Visualization & Computer Graphics* 01 (2022), 1–1.

[4] Nikos Athanasiou, Mathis Petrovich, Michael J. Black, and Gül Varol. 2022. TEACH: Temporal Action Compositions for 3D Humans. In *International Conference on 3D Vision (3DV)*.

[5] Norman I Badler, Cary B Phillips, and Bonnie Lynn Webber. 1993. *Simulating humans: computer graphics animation and control*. Oxford University Press.

[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 *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VII*. Springer, 557–577.

[7] Rukun Fan, Songhua Xu, and Weidong Geng. 2011. Example-based automatic music-driven conventional dance motion synthesis. *IEEE transactions on visualization and computer graphics* 18, 3 (2011), 501–515.

[8] Dariu M Gavrila. 1999. The visual analysis of human movement: A survey. *Computer vision and image understanding* 73, 1 (1999), 82–98.

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2020. Generative adversarial networks. *Commun. ACM* 63, 11 (2020), 139–144.

[10] Anand Gopalakrishnan, Ankur Mali, Dan Kifer, Lee Giles, and Alexander G Ororbia. 2019. A neural temporal model for human motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12116–12125.

[11] 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.

[12] 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.

[13] Bo Han, Yitong Liu, and Yixuan Shen. 2023. Zero3D: Semantic-Driven Multi-Category 3D Shape Generation. *arXiv preprint arXiv:2301.13591* (2023).

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

[15] Jonathan Ho and Tim Salimans. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598* (2022).

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

[17] Ruozi Huang, Huang Hu, Wei Wu, Kei Sawada, Mi Zhang, and Daxin Jiang. 2020. Dance revolution: Long-term dance generation with music via curriculum learning. *arXiv preprint arXiv:2006.06119* (2020).

[18] Leslie Ikemoto, Okan Arikan, and David Forsyth. 2009. Generalizing motion edits with gaussian processes. *ACM Transactions on Graphics (TOG)* 28, 1 (2009), 1–12.

[19] Yanghua Jin, Jiakai Zhang, Minjun Li, Yingtao Tian, Huachun Zhu, and Zhihao Fang. 2017. Towards the automatic anime characters creation with generative adversarial networks. *arXiv preprint arXiv:1708.05509* (2017).

[20] Durk P Kingma and Prafulla Dhariwal. 2018. Glow: Generative flow with invertible 1x1 convolutions. *Advances in neural information processing systems* 31 (2018).

[21] Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).

[22] Hsin-Ying Lee, Xiaodong Yang, Ming-Yu Liu, Ting-Chun Wang, Yu-Ding Lu, Ming-Hsuan Yang, and Jan Kautz. 2019. Dancing to music. *Advances in neural information processing systems* 32 (2019).

[23] Minho Lee, Kyogu Lee, and Jaeheung Park. 2013. Music similarity-based approach to generating dance motion sequence. *Multimedia tools and applications* 62 (2013), 895–912.

[24] Buyu Li, Yongchi Zhao, Shi Zhelun, and Lu Sheng. 2022. Danceformer: Music conditioned 3d dance generation with parametric motion transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 1272–1279.

[25] Ruilong Li, Shan Yang, David A Ross, and Angjoo Kanazawa. 2021. Ai choreographer: Music conditioned 3d dance generation with aist++. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 13401–13412.

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

[27] Zhengxiong Luo, Dayou Chen, Yingya Zhang, Yan Huang, Liang Wang, Yujun Shen, Deli Zhao, Jingren Zhou, and Tieniu Tan. 2023. VideoFusion: Decomposed Diffusion Models for High-Quality Video Generation. *arXiv e-prints* (2023), arXiv–2303.

[28] 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.

[29] Jianyuan Min and Jinxiang Chai. 2012. Motion graphs++ a compact generative model for semantic motion analysis and synthesis. *ACM Transactions on Graphics (TOG)* 31, 6 (2012), 1–12.

[30] Tomohiko Mukai and Shigeru Kuriyama. 2005. Geostatistical motion interpolation. In *ACM SIGGRAPH 2005 Papers*. 1062–1070.

[31] Dirk Ormoneit, Michael J Black, Trevor Hastie, and Hedvig Kjellström. 2005. Representing cyclic human motion using functional analysis. *Image and Vision Computing* 23, 14 (2005), 1264–1276.

[32] Joseph O'rourke and Norman I Badler. 1980. Model-based image analysis of human motion using constraint propagation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 6 (1980), 522–536.

[33] Dario Pavllo, Christoph Feichtenhofer, Michael Auli, and David Grangier. 2020. Modeling human motion with quaternion-based neural networks. *International Journal of Computer Vision* 128 (2020), 855–872.

[34] Mathis Petrovich, Michael J Black, and Gül Varol. 2022. 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*. Springer, 480–497.

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

[36] 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.

[37] 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.

[38] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125* (2022).

[39] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10684–10695.

[40] Charles Rose, Michael F Cohen, and Bobby Bodenheimer. 1998. Verbs and adverbs: Multidimensional motion interpolation. *IEEE Computer Graphics and Applications* 18, 5 (1998), 32–40.

[41] Alla Safonova and Jessica K Hodgins. 2007. Construction and optimal search of interpolated motion graphs. In *ACM SIGGRAPH 2007 papers*. 106–es.

[42] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems* 35 (2022), 36479–36494.

[43] Li Siyao, Weijiang Yu, Tianpei Gu, Chunze Lin, Quan Wang, Chen Qian, Chen Change Loy, and Ziwei Liu. 2022. Bailando: 3d dance generation by actor-critic gpt with choreographic memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11050–11059.

[44] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*. PMLR, 2256–2265.

[45] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. 2020. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456* (2020).

[46] Guofei Sun, Yongkang Wong, Zhiyong Cheng, Mohan S Kankanhalli, Weidong Geng, and Xiangdong Li. 2020. DeepDance: music-to-dance motion choreography with adversarial learning. *IEEE Transactions on Multimedia* 23 (2020), 497–509.

[47] Taoran Tang, Jia Jia, and Hanyang Mao. 2018. Dance with melody: An lstm-autoencoder approach to music-oriented dance synthesis. In *Proceedings of the 26th ACM international conference on Multimedia*. 1598–1606.

[48] Guy Tevet, Brian Gordon, Amir Hertz, Amit H Bermano, and Daniel Cohen-Or. 2022. 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*. Springer, 358–374.

[49] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. 2022. Human motion diffusion model. *arXiv preprint arXiv:2209.14916* (2022).

[50] Jonathan Tseng, Rodrigo Castellon, and C Karen Liu. 2022. EDGE: Editable Dance Generation From Music. *arXiv preprint arXiv:2211.10658* (2022).

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

[52] Chen Xin, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, Jingyi Yu, 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 (CVPR)*.

[53] Zhixuan Yu, Jae Shin Yoon, In Kyu Lee, Prashanth Venkatesh, Jaesik Park, Jihun Yu, and Hyun Soo Park. 2020. Humbi: A large multiview dataset of human body expressions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2990–3000.

[54] 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).

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

[56] Wenlin Zhuang, Congyi Wang, Jinxiang Chai, Yangang Wang, Ming Shao, and Siyu Xia. 2022. Music2dance: Dancenet for music-driven dance generation. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 18, 2 (2022), 1–21.