Generating Text Conditioned 3D Human Motion

Shilna Koyileriyan
shilna2000@vjec.ac.in

Department of Computer Science and Engineering, Vimal Jyothi Engineering College

A B S T R A C T

A novel approach to generating 3D human motions from textual descriptions using a two-stage framework that incorporates a temporal variational autoencoder (VAE). The first stage, text-to-length sampling, predicts the length of the motion sequence by approximating a probability distribution conditioned on the input text. This step enables the generation of variable-length motions that align naturally with diverse textual inputs. In the second stage, text-to-motion generation, the model synthesizes human motions consistent with the sampled lengths and input descriptions. At its core, the approach utilizes a temporal VAE, featuring a triplet structure of prior, posterior, and generator networks to learn the mapping between textual semantics and motion dynamics. To further enhance motion fidelity, a motion snippet code representation is proposed as an internal format, encapsulating localized temporal semantics to ensure smooth and realistic motions faithful to the input text. The framework's flexibility accommodates a range of textual complexities, from simple commands to detailed narratives, ensuring both diversity and naturalness in the output. This approach is evaluated on a new large-scale dataset, HumanML3D, which includes 14,616 motion clips and 44,970 text descriptions, as well as on the KIT Motion-Language dataset. Quantitative metrics and user studies demonstrate significant advancements over baseline methods in terms of motion quality, diversity, and alignment with textual input. By addressing the challenges of variable sequence lengths, semantic diversity, and textual complexity, this two-stage VAE-based framework sets a new benchmark for text-driven 3D motion generation, with applications spanning animation, gaming, and human-computer interaction.

Keywords: text-to-motion generation, diffusion models, HumanML3D dataset, 3D human motion, natural language processing, motion synthesis.

1. Introduction

This work explores a two-stage framework for generating 3D human motions from textual descriptions using a temporal variational autoencoder (VAE), addressing the challenges of motion diversity, textual complexity, and variable-length sequence generation. The first stage, text-to-length sampling, predicts the motion sequence length by learning a probability distribution conditioned on the input text. This stage ensures flexibility in generating motions of varying durations, essential for aligning with diverse textual inputs. Specifically, a neural network-based approach, such as pixelCNN, is employed to model the length distribution, allowing the system to handle a wide range of input scenarios, from simple commands to complex narratives. The second stage, text-to-motion generation, synthesizes motion sequences based on the sampled lengths and input text. This stage leverages the capabilities of a temporal VAE with a triplet structure consisting of prior, posterior, and generator networks. The VAE framework facilitates the generation of stochastic, realistic, and semantically faithful motions, ensuring diversity and coherence in the output.

A key innovation in this framework is the introduction of motion snippet codes, a compact and semantic representation of local motion dynamics. By encoding motion sequences into snippets, the model captures

meaningful temporal dependencies, ensuring the smoothness and naturalness of the generated motions. The text encoder, realized using bidirectional GRUs, processes textual descriptions to extract both word-level and sentence-level features, further refined with part-of- speech tags and custom word categorizations. These features are integrated into the VAE to guide motion generation effectively. Additionally, the framework in- corporates local word attention mechanisms to align motion dynamics with critical textual elements, enhancing the semantic accuracy of generated motions.

The approach is evaluated on the new large-scale HumanML3D dataset, which contains over 14,000 motion clips paired with nearly 45,000 textual descriptions, as well as the KIT Motion-Language dataset. HumanML3D represents a diverse collection of human actions, including locomotion, sports, and artistic movements, providing a robust benchmark for motion generation tasks. Empirical evaluations demonstrate that the proposed framework outperforms state-of-the-art baselines across multiple metrics, including motion fidelity, diversity, and alignment with text descriptions. The use of the text-to-length module enables the generation of motions with natural temporal variations, while the temporal VAE ensures rich and diverse motion outputs**.**

User studies validate the visual realism and semantic coherence of the generated motions, with a significant portion of outputs rated highly by evaluators. Ablation studies further highlight the importance of key components, such as motion snippet codes and local word attention, in achieving optimal performance. This two-stage approach effectively addresses the limitations of existing deterministic sequence- to-sequence models, offering a more robust solution for text-driven 3D motion synthesis.

In summary, this framework provides a powerful tool for generating diverse, natural, and text-aligned 3D human motions, with significant implications for applications in virtual reality, gaming, robotics, and human-computer interaction. By integrating innovative techniques such as text-to-length sampling, motion snippet codes, and attention mechanisms, it sets a new standard for the field, enabling ma- chines to produce human-like motion dynamics that are both visually compelling and semantically accurate.

2. Literature Review

Automated generation of 3D human motions from text is a challenging problem. The generated motions are expected to be sufficiently diverse to explore the text- grounded motion space, and more importantly, accurately depicting the content in prescribed text descriptions. Here we tackle this problem with a two-stage approach: text2length sampling and text2motion generation. Text2length involves sampling from the learned distribution function of motion lengths conditioned on the input text. This is followed by our text2motion module using temporal variational autoencoder to synthesize a diverse set of human motions of the sampled lengths. Instead of directly engaging with pose sequences, we propose motion snippet code as our internal motion representation, which captures local semantic motion contexts and is empirically shown to facilitate the generation of plausible motions faithful to the input text. More- over, a large-scale dataset of scripted 3D Human motions, HumanML3D, is constructed, consisting of 14,616 motion clips and 44,970 text descriptions.

3D Human Motion Generation. There are several prior efforts in synthesizing 2D or 3D human motions based on action category, or from modalities including audio and text. To be based on action category, an one-hot condition vector is often engaged in synthesizing pose sequences. In this space, [5, 43] both apply a two-stage generative adversarial network (GAN) framework to progressively extend the partial motion sequence with newly generated poses; the work of [45] instead models the spatiotemporal structures of human dynamics with a GCN- based GAN; mean- while, VAE modelling and transformer architecture are promoted by [11, 12, 29] to incorporate temporal dependencies. In terms of audio signal input, as audio is temporally aligned with its motion output, a common strategy is to employ a temporal sliding-window in translating the acoustic feature representation (e.g. MFCC) to individual human poses using recurrent neural networks (RNNs). In [35], a Bi-Directional LSTM network is adopted to generate up- per body gestures from speech input. Similar LSTM-type models are also examined by [34] to predict upper body dynamics from piano and violin recital audios, and in [36] to capture the music-to-dance mapping. Recent works start to address the stochastic nature of human

dynamics grounded on audio signals. [17] employs a hybrid model of VAE and GAN to produce non-deterministic human dancing movements from music. The work of [14] further supports long- term music-to-dance generation with curriculum training.

Generative models play a crucial role in motion synthesis by generating high- quality human motion. Generative Adversarial Nets (GAN) [6] use two sub- models: a generator model that produces new samples, and a discriminator model that attempts to classify samples as either real or fake. These two models compete against each other during training. However, the interpretability of GAN is poor because the learned data distribution lacks an explicit expression, resembling a black box mapping function.

Auto-Encoding Variational Bayes (VAE) [13] is a widely used generative model in motion synthesis. Its primary objective is to generate new samples from the learned distribution of objects by learning latent attributes from the probability distribution of the latent variable space, thereby con- structing new examples. De- spite its usefulness, the quality of samples generated by VAE can be improved. Recently proposed diffusion models [10, 18, 26] have shown immense potential in modeling and present an exciting opportunity to expand into text-driven motion generation. These models utilize the stochastic diffusion process modeled in thermodynamics, which gradually adds noise to the samples of the data distribution. The deep learning model then learns the reverse process of denoising the samples gradually. Diffusion models have the advantage over previous models as they do not make any assumptions about the target distribution, leading to a more diverse generation and better suitability for our task. Therefore, we propose a novel fine-grained human motion generation method that employs the Denoising Diffusion Probability Model [10].

Obtaining human motion sequences through traditional soft- ware is a labor- intensive and tedious process, while motion capture is complex and expensive. Recently, with the advancements in deep learning and computer vision, learning- based human motion generation has emerged as a solution to this problem, leading to the development of associated generation methods based on multimodal data. The input multimodal data include music [12, 15, 25, 27], motion categories [4, 9, 20], text [2, 3, 5, 7, 8, 16, 21, 23, 28, 33], others. Text-driven human motion generation has been a popular research topic, because of its convenience and human-friendliness. In particular, natural language comprises nouns, verbs, adverbs, etc. The mutual connections among different words in a sentence establish its semantics. Verbs define the action's category, while adverbs control the fineness of the action. The interaction between words in syntax plays a vital role in determining the structure and meaning of a sentence. Failure to fully incorporate these text features may result in inadequate text modeling, causing the generated motion sequence to deviate from the intended meaning of the original text. Existing methods can be divided into two branches, including 1) cross-modal alignment of motion and text [2, 3, 5, 7, 8, 16, 21] conditional diffusion models [28, 33]. In the first methods, text sequences and motion sequences are mapped onto separate feature spaces and forcibly aligned, leading to a loss of original information from both domains. In the second methods, the diffusion model incorporates text information as a conditioning factor to learn the probability mapping of human motions.

While there is also work on facial motion generation [8, 11, 23, 43], here we focus on articulated human bodies. Human motion synthesis. While there is a large body of work focusing on future human motion prediction [3, 6, 16, 37, 4, 27] and completion [10, 17], here, we give an overview of methods that generate motions from scratch (i.e., no past or future observations). Generative models of human motion have been designed using GANs [1, 30], VAEs [15, 39], or normalizing flows [18, 15]. In this work, we employ VAEs in the context of Transformer neural network architectures. Recent work suggest that VAEs are effective for human motion generation compared with GANs [15, 39], while being easier to train.

Motion synthesis methods can be broadly divided into two categories: (i) un- constrained generation, which models the entire space of possible motions [31, 26, 18] and (ii) conditioned synthesis, which aims for controllability such as using music [26, 27, 28], speech [17, 14], action [15, 39], and text [1, 3, 13, 29, 30, 45] conditioning. Generative models that synthesize unconstrained motions aim, by design, to sample from a distribution, allowing generation of diverse motions. However, they lack the ability to control the generation process. On the other hand, the conditioned synthesis can be further divided into two categories: deterministic

[2, 13, 29] or probabilistic [1, 15, 26, 27, 30, 39]. In this work, we focus on the latter, motivated by the fact that there are often multiple possible motions for a given condition.

Text-conditioned motion generation. Recent work explores the advances in natural language modeling [9, 36] to design sequence-to-sequence approaches to cast the text-to-motion task as a machine translation problem [1, 29, 41]. Others build joint cross-modal embedding to map the text and motion to the same space [2, 13, 50], which has been a success in other research area [5, 42, 26, 33].

Several methods use an impoverished body motion representation. For ex- ample, some do not model the global trajectory [41, 20], making the motions unrealistic and ignoring the global movement description in the input text. Text2Action [1] uses a sequence-to-sequence model but only models the up- per body motion. This is because Text2Action uses a semi-automatic approach to create training data from the MSR-VTT captioned video dataset [9], which contains frequently occluded lower bodies. They apply 2D pose estimation, lift the joints to 3D, and employ manual cleaning of the input text to make it generic.

Most other work uses 3D motion capture data [2, 13, 29, 30]. DVGANs [30] adapt the CMU MoCap database [47] and Human3.6M [21, 22] for the task of motion generation and completion, and they use the action labels as text- conditioning instead of categorical supervision. More recent works [2, 13, 29] employ the KIT Motion-Language dataset [40], which is also the focus of our work.

Translating text description to human motion is an emerging topic. Prior efforts such as [10, 21, 32, 44] re- sort to classical encoder-decoder RNN models, while it is proposed in [1] to learn a joint embedding space between natural language and 3D human dynamics. [10] considers the hierarchical pose structure as well as utilizing a pose discriminator. These methods however bear undesirable is- sues, as being deterministic one-to-one processes with xed motion length. In contrary, aiming at these issues, our learned model is shown capable of generating stochastic, one-to-many sequence mappings of variable lengths.

Video Generation, and Text-based Video Generation. On generating videos, deep generative models such as GANs and VAEs have been the most popular choice. For example, a recurrent structured GAN is presented in MoCo- GAN [38] to separately model stationary pixels and dynamic motions. This is followed by [37] to incorporate contrastive learning. [8] leverages a VAE with RNN architecture to stochastically predict future frames based on the historical video sequence, which is further extended in [42] to synthesize videos with prescribed start and end frames.

Text-to-video generation is relatively new. To address such task, GAN frameworks have been recruited in several efforts including [20] and [25]. This is followed by [6], where an attention mechanism is additionally engaged to align local video regions with words in text. Moreover, both short-time and long-term cross-domain attentive vectors are utilized in [24] as the inputs to a VAE framework.

Language and 3D Human Motion Data. KIT Motion- Language Dataset [31] is to date the only available dataset comprising both 3D human motions and their textual de- scriptions, which consists of 3,911 motion sequences 6,278 sentences, and is focused on locomotion movements. Moreover, there are a number of existing datasets of 3D motion captured human motions, such as CMU Mocap [7], Hu- man3.6M [15], MoVi [9] and BABEL [33], in the form of everyday actions and sports movements. However, none of them possesses language descriptions of the motions.

3. Methodology

Text-to-3D human motion generation aims to synthesize realistic and diverse human motions driven by natural language descriptions. A novel two-stage framework utilizing a temporal variational autoencoder (VAE) addresses this challenge by ensuring flexibility and semantic fidelity. The first stage, text-to-length sampling, predicts motion sequence lengths based on the input text, accommodating variable temporal durations. The second stage, text-to-motion generation, synthesizes motions consistent with the sampled lengths and textual descriptions using the VAE's stochastic capabilities. This approach introduces motion snippet codes to capture temporal semantics, enabling natural and smooth motions, with applications in virtual reality, robotics, animation,

and human-computer interaction.

As a preprocessing step, a dedicated motion autoencoder is trained on our training motion data to encode a motion sequence into a stream of motion snippet codes, which then could be decoded back into motions. Our training pipeline. Through text encoder, the attentive word features (watt) are used by VAE networks as illustrated in Fig. 3. The triplet structure of temporal VAE involving the prior, posterior, and generator networks is employed to process the motion snippet codes (cs) and the reconstructed ones (ˆcs). This leads to the loss terms evaluating the reconstructed pose sequence (Lmot rec) and the reconstructed code sequence (Lcode rec), respectively. Due to lack of space, some key ingredients are deferred to be presented in Fig.3.1. Our inference pipeline. From the input text, text2length module is activated to sample an intended motion length. Text features extracted through the text encoder are then fed to the prior network, yielding a prior distribution. Generator samples latent vectors from the prior distribution and produces a series of motion snippet codes (cs). The pose sequence is finally obtained by decoding the snippet codes from the motion decoder pre-trained.

From a text description of M words, X = (x1, ..., xM), our goal is to generate a 3D pose sequence, P =(p1, ..., pT), with its length T determined at test time. As shown in Fig. 2, we start by a preprocessing step to train a motion autoencoder. As shown in Fig. 2, we start by a preprocessing step to train a motion autoencoder.

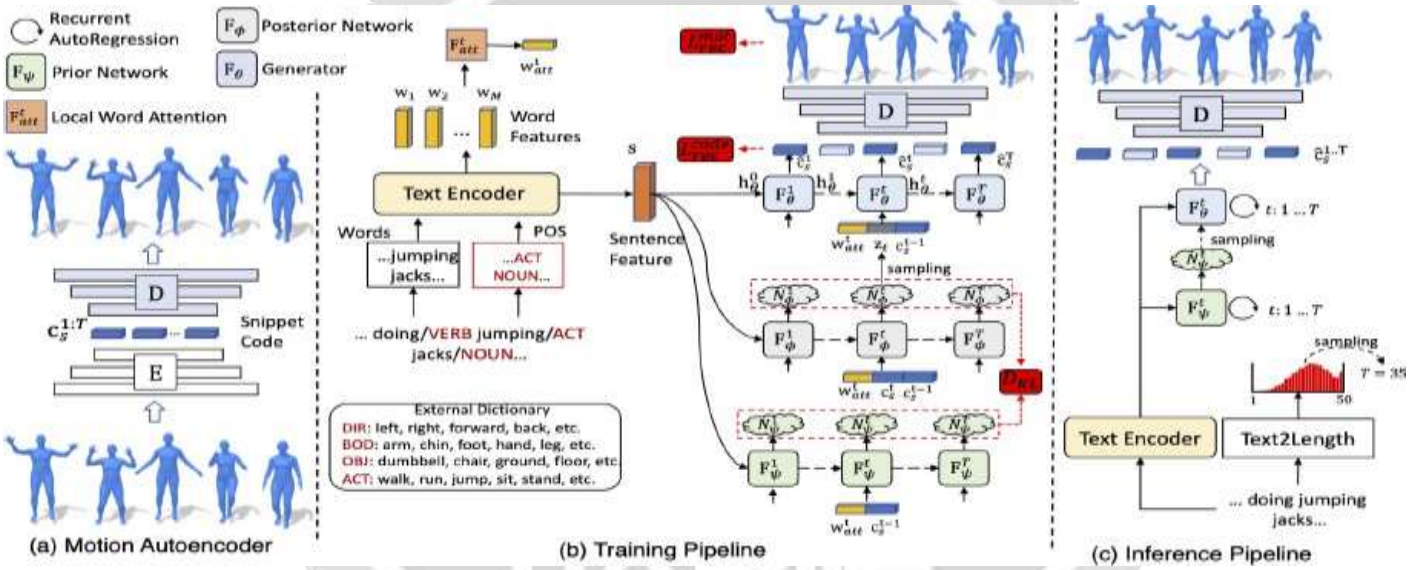


Figure 3.1: Two-Stage Text-to-3D Human Motion Generation Framework

3.1 Motion Autoencoder

As the preprocessing step described in Fig 3.1, an encoder E transforms the pose sequence P = (p1, ..., pT)to a motion snippet code sequence, Cs = c1s, ..., cTs, achieved by applying 1-D convolutions over temporal line;ˆ P is then reconstructed with a deconvolutional decoder, D.Mathematically, this process is formulated as Cs = E(P), ˆ P = D(Cs).

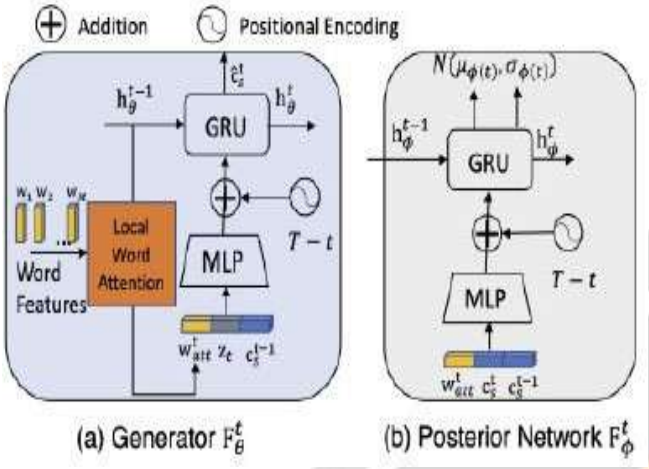
To avoid foot sliding, our decoder D additionally predict foot contacts at each frame which are not given to the encoder E. It is also necessary to constrain the snippet code values and the differences of consecutive codes to encourage sparsity and temporal smoothness. The nal objective function becomes

$$L_{E,D} = \sum_{t'} \|\hat{p}_{t'} - p_{t'}\|_1 + \lambda_{\text{spr}} \sum_{t} \|c_t^s\|_1 + \lambda_{\text{smt}} \sum_{t} \|c_t^s - c^{t-1}\|$$

The autoencoder consists of two-layer convolutions with lter size of 4 and stride 2, whose structure is detailed in supplementary le. As a result, a motion snippet code a ct s has a 8-frame receptive eld, amounting to around 0.5 second for 20 frame-per-second (fps) pose streaming; it also leads to a more compact internal code sequence with ′ — ′ Compared to individual

poses, snippet code captures temporal semantic information that is crucial in smooth and

faithful motion generation.



3.2 Text2Length Sampling

In order to obtain a discrete time length T at the inference stage by sampling from this learned distribution function, p(T—x1,..., xM), given an input text, our text2length sampling module aims to approximate the probability distribution of discrete motion length T conditioned on text. Thus, this module makes it possible for our method to produce movements of different lengths.

There are numerous workable solutions for this common density estimation problem, and we

choose to use the pixelCNN neural network approach. Our goal specifically consists of determining the length of snippet codes because a motion sequence is internally represented in our work as a collection of snippet codes. Sentence-level characteristics are extracted from the input text by a text encoder and fed into an MLP layer with softmax activation for inference. This results in a multinomial distribution across discrete length indices 1, 2,..., Tmax. In this case, one increment is equivalent to four pose frames, and setting Tmax = 50 is equivalent to 200 frames, or 10 seconds for a video at 20 frames per second. The cross entropy loss defines its training goal.

3.3Text2Motion Generation

Our text2motion generator contains a text encoder, and a temporal VAE model consisting of a triplet network of generator F_ϑ, posterior F_ϕ, and prior F_ψ. The text encoder extracts both the word-level $w_{1:M}$ and sentence-level s features from input text; our VAE generates motion snippet codes one by one with a recurrent architecture: at time t, our posterior network F_ϕ approximates the posterior distribution conditioned on partial code sequence as well as word and sentence features $c = (w_{1:M}, s, \dots)$. Instead of relating the posterior distribution to a prior normal distribution $N(0, I)$ as used in the literature, here it is related to a learned prior distribution which is obtained by our prior network based on the previous state and conditions. Overall, our VAE is trained

$$\log p(C_i) \geq \sum_{t=1} E_{q(z_{1:t}|c^t,c)} \log p_\theta(c^t|c^t_{1:t-1}, z_{1:t}) - \lambda_{KLD} D_{KL}(q_\phi(z_t|c^t_{1:t}, c) \| p_\psi(z_t|c^t_{1:t-1}, c))$$

The first term is to reduce reconstruction error Lrec, while the second term penal- izes the KL-divergence LKL

between the posterior and the prior distributions.

3.4 Training Scheme

To address the variable length sequence to- sequence generation task, our training process utilizes both curriculum learning [3] and scheduled sampling [2] strategies, as follows. Starting from aiming to generate rst Tcur snippet codes in sequence, we optimize our model on training data that owns snippet code lengths equal or longer than Tcur. As long as the reconstruction loss on the validation starts raising, then we move on to the next stage by appending one more snippet code in the target sequence ; the complexity of the task is progressively increased at every stage till the maximum time step Tmax of prediction is reached (i.e, Tcur = Tmax). In addition, to bridge the gap of training and inference for sequence prediction, teacher forcing is applied for the entire target snippet code sequence c1:T s with probability of ptf , which means the ground-truth snippet code is taken as input for the generation at next step. Accordingly, the generated snippet code will instead serve as the input with probability 1 ptf . As a boundary condition, c0s is a constant vector that encodes mean poses using motion encoder E.

3.5 HumanML3D Dataset

Motion sequences from two sizable, publicly available datasets of 3D human motion captures, HumanAct12 and AMASSS, are combined to create the Hu- manML3D dataset. They include motions from a range of human activities, including sports (like swimming and karate), daily activities (like walking and leaping), acrobatics (like cartwheeling), and artistic endeavors (like dancing). Sadly, there are no written explanations of the motions in these datasets. Data normalization involves the following processing processes. After scaling to 20 frames per second, motions longer than 10 seconds are arbitrarily reduced to 10- second ones, retargeted to a default human skeletal template, and initially correctly rotated to face Z+. After that, fluent English speakers who have an average work approval rating of more than 92% are hired and requested to characterize a motion in at least five words as part of a textual annotation process conducted through Amazon Mechanical Turk (AMT). Each motion clip has three written descriptions that we gather from different employees. After that, a human postprocessing step is performed to remove any unusual textual descriptions. With 14,616 motions and 44,970 descriptions made up of 5,371 unique words, our HumanML3D dataset thus becomes the largest and most varied collection of programmable human motions that we are aware of. The average duration of the motions is 7.1 seconds, with a total duration of 28.59 hours. The duration is two seconds at the lowest and ten seconds at the maximum. Their median and average textual description lengths are 10 and 12, respectively. In Table 1, our HumanML3D is tabulated against the only motion-text dataset currently available, KIT Motion-Language.

Dataset	#Motions	#texts	Duration	Vocab.
HumanML3D	14,616	44,970	28.59h	5,371
KIT-ML	3,911	6,278	10.33h	1,623

Table 3.1: Comparisons of 3D human motion-language datasets

4.Results and Discussion

Evaluation Metrics include Frechet Inception Distance (FID), diversity and multi- modality. For quantitative evaluation, a motion feature extractor and text feature extractor is trained under contrastive loss to produce geometrically close feature vectors for matched text-motion pairs, and vice versa. Further explanations of aforementioned metrics as well as the specific textual and motion feature extractor are relegated to the supplementary file due to space limit. In addition, the R-precision and MultiModal distance are proposed in this work as complementary metrics, as follows. Consider Rprecision: for each generated motion, its ground- truth text description and 31 randomly selected mismatched descriptions from the test set form a description pool. This is followed by calculating and ranking the Euclidean distances between the motion feature and the text feature of each de- scription

in the pool. We then count the average accuracy at top-1, top-2 and top-3 places. The ground truth entry falling into the top-k candidates is treated as successful retrieval, otherwise it fails. Meanwhile, MultiModal distance is computed as the average Euclidean distance between the motion feature of each generated motion and the text feature of its corresponding description in test set.

We compare our work to three state of the-art methods: Seq2Seq, Language2Pose and Text2Gesture. As with all existing methods, they are deterministic methods. Considering the stochastic nature of our task, we adapt two non-deterministic methods from related fields for more fair and thorough evaluations: MoCo- GAN and Dance2Music. The former is widely used for conditioned video synthesis, and the latter produces 2D dancing motion sequences from audio signals. Proper changes are made to allow these methods generating 3D motions from text.

4.1 Quantitative Evaluation

The quantitative findings for the KIT-ML and HumanML3D datasets are shown in Tables 2 and 3, respectively. Each experiment is conducted 20 times for fair comparison, and a 95% confidence level statistical interval is presented. We also examine a variation of our strategy by eliminating the text2length sample module for fair comparison, as all baseline approaches generate a new motion directly using the ground-truth motion length.

The dependability of the suggested R-precision metric, which establishes a maxi- mum performance limit for all approaches, is demonstrated by the high R precision of genuine motions. Overall, the findings from Tables 2 and 3 are as follows. First, our strategy performs significantly better than all comparison approaches on both datasets and across all measures. With their neural machine translation architecture of encoder-decoder and transformer, Seq2Seq and Text2Gesture immediately translate textual input to human dynamics; yet, they struggle to maintain the illusion of realistic motions throughout their operations. As a result, FID values are high and motion-based text retrieval precision is low. By include a co- embedding space, Language2Pose improves generation quality; yet, the outcomes are far from realistic. Due to their disloyalty to the input text, the motions produced by MoCoGAN and Dance2Music's non-deterministic techniques are regrettably of extremely poor quality, as evidenced by their low diversity and multimodality scores. Conversely, the version of our method that uses real motion length directly (Ours w/ real length) performs best across nearly all measures. Our de- fault method, which employs text2length sampling (Ours), performs similarly in R-precision and FID scores, but it is better at synthesizing a variety of motions, as evidenced particularly in the diversity multimodality scores. Leads to large FID values and poor motion-based text retrieval precision. By include a co-embedding space, Language2Pose improves generation quality; yet, the outcomes are far from realistic.

Methods	R Precision↑			FID ↓	MultiModal Dist↓	Diversity→	MultiModality ↑
(lr)2-4	Top 1	Top 2	Top 3				
Real motions	$0.511_{\pm 0.003}$	$0.703_{\pm 0.003}$	$0.797_{\pm 0.002}$	$0.002_{\pm 0.000}$	$2.974_{\pm 0.008}$	$9.503_{\pm 0.065}$	-
Seq2Seq [?]	$0.180_{\pm 0.002}$	$0.300_{\pm 0.002}$	$0.396_{\pm 0.002}$	$11.75_{\pm 0.035}$	$5.529_{\pm 0.007}$	$6.223_{\pm 0.061}$	-
Language2Pose [?]	$0.246_{\pm 0.002}$	$0.387_{\pm 0.002}$	$0.486_{\pm 0.002}$	$11.02_{\pm 0.046}$	$5.296_{\pm 0.008}$	$7.676_{\pm 0.058}$	-
Text2Gesture [?]	$0.165_{\pm 0.001}$	$0.267_{\pm 0.003}$	$0.345_{\pm 0.002}$	$7.66_{\pm 0.030}$	$6.030_{\pm 0.008}$	$6.409_{\pm 0.071}$	-
MoCoGAN [?]	$0.037_{\pm 0.002}$	$0.072_{\pm 0.001}$	$0.106_{\pm 0.001}$	$94.41_{\pm 1.021}$	$9.643_{\pm 0.006}$	$0.462_{\pm 0.008}$	$0.019_{\pm 0.000}$
Dance2Music [?]	$0.033_{\pm 0.002}$	$0.065_{\pm 0.001}$	$0.097_{\pm 0.001}$	$66.98_{\pm 0.016}$	$8.116_{\pm 0.006}$	$0.725_{\pm 0.011}$	$0.043_{\pm 0.001}$
Ours w/ real length	$\mathbf{0.457_{\pm 0.003}}$	$\mathbf{0.639_{\pm 0.003}}$	$\mathbf{0.740_{\pm 0.003}}$	$\mathbf{1.067_{\pm 0.002}}$	$\mathbf{3.340_{\pm 0.008}}$	$\mathbf{9.188_{\pm 0.082}}$	$\mathbf{2.090_{\pm 0.083}}$
Ours	$\mathbf{0.455_{\pm 0.003}}$	$\mathbf{0.636_{\pm 0.003}}$	$\mathbf{0.736_{\pm 0.002}}$	$\mathbf{1.087_{\pm 0.021}}$	$\mathbf{3.347_{\pm 0.008}}$	$\mathbf{9.175_{\pm 0.083}}$	$\mathbf{2.219_{\pm 0.074}}$

Table 4.1: Quantitative evaluation on the HumanML3D test set

Methods	R Precision↑			FID ↓	MultiModal Dist↓	Diversity→	MultiModality ↑
(lr)2-4	Top 1	Top 2	Top 3				
Real motions	$0.424_{\pm 0.005}$	$0.649_{\pm 0.006}$	$0.779_{\pm 0.006}$	$0.031_{\pm 0.004}$	$2.788_{\pm 0.012}$	$11.08_{\pm 0.097}$	-
Seq2Seq [?]	$0.103_{\pm 0.003}$	$0.178_{\pm 0.003}$	$0.241_{\pm 0.002}$	$24.86_{\pm 0.348}$	$7.960_{\pm 0.031}$	$6.744_{\pm 0.106}$	-
Language2Pose [?]	$0.221_{\pm 0.005}$	$0.374_{\pm 0.004}$	$0.483_{\pm 0.005}$	$15.54_{\pm 0.072}$	$5.147_{\pm 0.030}$	$9.073_{\pm 0.100}$	-
Text2Gesture [?]	$0.156_{\pm 0.004}$	$0.255_{\pm 0.003}$	$0.342_{\pm 0.002}$	$22.12_{\pm 0.183}$	$6.964_{\pm 0.029}$	$9.334_{\pm 0.079}$	-
MoCoGAN [?]	$0.022_{\pm 0.002}$	$0.042_{\pm 0.003}$	$0.063_{\pm 0.003}$	$62.69_{\pm 1.242}$	$10.47_{\pm 0.012}$	$3.091_{\pm 0.043}$	$0.250_{\pm 0.009}$
Dance2Music [?]	$0.031_{\pm 0.002}$	$0.058_{\pm 0.002}$	$0.084_{\pm 0.003}$	$15.41_{\pm 0.240}$	$10.40_{\pm 0.016}$	$0.241_{\pm 0.004}$	$0.062_{\pm 0.002}$
Ours w/ real length	$\mathbf{0.370_{\pm 0.005}}$	$\mathbf{0.569_{\pm 0.007}}$	$\mathbf{0.693_{\pm 0.007}}$	$\mathbf{0.770_{\pm 0.109}}$	$\mathbf{3.401_{\pm 0.088}}$	$\mathbf{10.91_{\pm 0.119}}$	$\mathbf{1.482_{\pm 0.065}}$
Ours	$\mathbf{0.361_{\pm 0.006}}$	$\mathbf{0.559_{\pm 0.007}}$	$\mathbf{0.681_{\pm 0.007}}$	$\mathbf{0.022_{\pm 0.107}}$	$\mathbf{3.488_{\pm 0.028}}$	$\mathbf{10.72_{\pm 0.145}}$	$\mathbf{2.052_{\pm 0.107}}$

Table 4.2: Quantitative evaluation on the KIT-ML test set

Methods	R Precision↑			FID ↓
(lr)2-4	Top 1	Top 2	Top 3	
Ours	$\mathbf{0.455_{\pm 0.003}}$	$\mathbf{0.636_{\pm 0.003}}$	$\mathbf{0.736_{\pm 0.002}}$	$\mathbf{1.087_{\pm 0.021}}$
w/o SnC	$0.370_{\pm 0.002}$	$0.538_{\pm 0.003}$	$0.642_{\pm 0.003}$	$1.200_{\pm 0.027}$
w/o Att	$0.396_{\pm 0.002}$	$0.570_{\pm 0.002}$	$0.674_{\pm 0.003}$	$1.833_{\pm 0.032}$
w/o PoS	$0.443_{\pm 0.003}$	$0.622_{\pm 0.003}$	$0.723_{\pm 0.003}$	$1.157_{\pm 0.016}$
w/o PoE	$0.444_{\pm 0.005}$	$0.627_{\pm 0.003}$	$0.729_{\pm 0.002}$	$1.229_{\pm 0.020}$

Table 4.3: Performance Comparison

4.2 Qualitative Evaluation

Qualitative comparisons between our method and the top-performing baseline, Language2Pose, are shown in Fig. 4. The SMPL human shapes cannot render motions from other comparison methods because they are too deformed. Partial concepts, such ”sit down,” are occasionally captured by Language2Pose in the input text. Nevertheless, it is unable to comprehend the global textual information. Furthermore, after a short time, the motions that are produced usually become fixed. Our method, on the other hand, can produce aesthetically pleasing motions that faithfully capture the specifics of written descriptions, including gesture, actions, body parts, and time.

Additionally, our generated motions are sufficiently varied from the same input text. Additional findings are shown in the appendix.

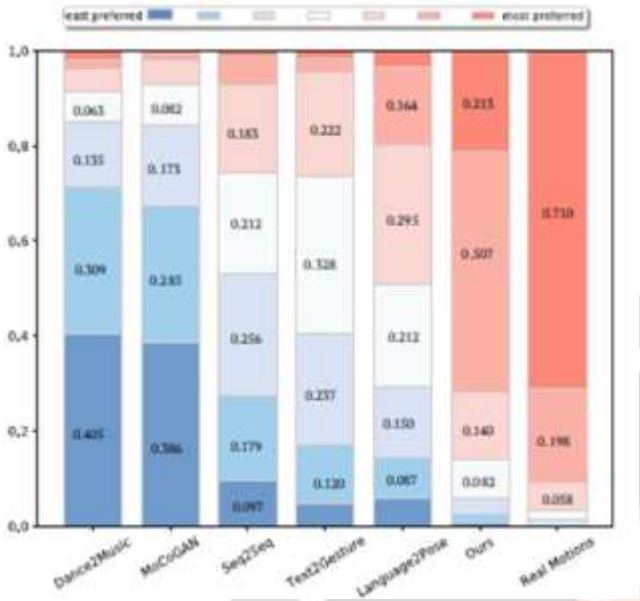


Figure4.1 Quantitative evaluation of user preference among the generated motions

5. Conclusion

the proposed two-stage framework for text-to-3D human motion generation, combining text-to-length sampling and text-to-motion generation with a temporal variational autoencoder (VAE), effectively addresses the challenges of generating realistic, diverse, and text-aligned motions. By introducing the innovative text-to-length module, the framework accommodates variable-length motion sequences, ensuring temporal alignment with textual inputs of varying complexity. The text-to-motion generation stage, driven by the VAE's stochastic capabilities, enables the synthesis of natural and semantically faithful human motions, capturing the inherent variability of human dynamics. The use of motion snippet codes further enhances the coherence and smoothness of the generated motions, ensuring they align accurately with described actions.

This approach outperforms traditional deterministic methods by offering greater diversity, semantic fidelity, and adaptability, as demonstrated through extensive evaluations on the HumanML3D and KIT-ML datasets. The framework's ability to handle complex textual descriptions and generate lifelike motions has significant implications for applications in virtual reality, animation, robotics, and human- computer interaction. With its flexible and scalable design, this method sets a new benchmark in text-driven 3D motion synthesis, paving the way for future advancements in creating expressive and context-aware human motions across a wide range of industries.

References

[1] haitanya Ahuja and Louis-Philippe Morency. Lan- guage2pose: Natural lan- guage grounded pose forecasting. In International Conference on 3D Vision (3DV), pages 719– 728. IEEE, 2019. 1, 2, 6, 7, 8

[2] Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. page 1171–1179, 2015. 5

[3] oshua Bengio, J´erˆome Louradour, Ronan Collobert, and Ja- son Weston. Cur- riculum learning. In Proceedings of the 26th annual International Conference on Machine Learning, pages 41–48, 2009. 5

[4] 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 IEEE Virtual Re- ality and 3D User Interfaces (VR), pages 1–10. IEEE, 2021. 1, 6, 7

[5] Haoye Cai, Chunyan Bai, Yu-Wing Tai, and Chi-Keung Tang. Deep video generation, prediction and completion of human action sequences. In Proceedings of the European Conference on Computer Vision (ECCV), pages 366–382, 2018. 2

[6] Qi Chen, Qi Wu, Jian Chen, Qingyao Wu, Anton van den Hengel, and Mingkui Tan. Scripted video generation with a bottom-up generative adversarial network.IEEE Transac- tions on Image Processing, 29:7454–7467, 2020. 2

[7] MU. Cmu graphics lab motion capture database. 2003. 3Emily Denton and Rob Fergus. Stochastic video generation with a learned prior. In International Conference on Machine Learning, pages 1174–1183. PMLR, 2018. 2

[8] Saeed Ghorbani, Kimia Mahdaviani, Anne Thaler, Konrad Kording, Dougla James Cook, Gunnar Blohm, and Niko- laus F Troje. Movi: A large multipurpose motion and video dataset. arXiv preprint arXiv:2003.01888, 2020. 3

[9] Anindita Ghosh, Noshaba Cheema, Cennet Oguz, Chris- tian Theobalt, and Philipp Slusallek. Synthesis of compo- sitional animations from textual descrip- tions. arXiv preprint arXiv:2103.14675, 2021. 2

[10] Chuan Guo, Xinxin Zuo, Sen Wang, Xinshuang Liu, Shihao Zou, Minglun Gong, and Li Cheng. Action2video: Gener- ating videos of human 3d actions.International Journal of Computer Vision, pages 1–31, 2022. 2

[11] Chuan Guo, Xinxin Zuo, Sen Wang, Shihao Zou, Qingyao Sun, Annan Deng, Minglun Gong, and Li Cheng. Ac- tion2motion: Conditioned generation of 3d human motions. In Proceedings of the 28th ACM International Conference on Multimedia, pages 2021–2029, 2020. 2, 5, 6

[12] Daniel Holden, Taku Komura, and Jun Saito. Phase- functioned neural net- works for character control. ACM Transactions on Graphics (TOG), 36(4):1–13, 2017. 6

[13] Ruozi Huang, Huang Hu, Wei Wu, Kei Sawada, Mi Zhang, and Daxin Jiang. Dance revolution: Long-term dance gen- eration with music via curriculum learn- ing. In International Conference on Learning Representations (ICLR), 2021. 2

[14] atalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Hu- man3. 6m: Large scale datasets and pre- dictive methods for 3d human sensing in natural environ- ments. IEEE transactions on pattern analysis and machine intelligence, 36(7):1325–1339, 2013. 3

[15] Atsuhiro Kojima, Takeshi Tamura, and Kunio Fukunaga. Natural language description of human activities from video images based on concept hierarchy of actions. International Journal of Computer Vision, 50(2):171–184, 2002. 2

[16] Hsin-Ying Lee, Xiaodong Yang, Ming-Yu Liu, Ting-Chun Wang, Yu-Ding Lu, Ming-Hsuan Yang, and Jan Kautz. Dancing to music. In Proceedings of the 33rd Interna- tional Conference on Neural Information Processing Sys- tems, 2019. 2,

[17] 6Jie Lei, Liwei Wang, Yelong Shen, Dong Yu, Tamara L Berg, and Mohit Bansal. Mart: Memory-augmented recurrent transformer for coherent video para- graph captioning. arXiv preprint arXiv:2005.05402, 2020. 2

[18] Lijun Li and Boqing Gong. End-to-end video captioning with multitask reinforcement learning. In 2019 IEEE Win- ter Conference on Applications of Computer Vision (WACV), pages 339–348. IEEE, 2019.

2

[19] Yitong Li, Martin Min, Dinghan Shen, David Carlson, and Lawrence Carin. Video generation from text. In Proceedings of the AAAI Conference on Articial Intelligence, volume 32, 2018. 2

[20] Angela S Lin, Lemeng Wu, Rodolfo Corona, Kevin Tai, Qix- ing Huang, and Raymond J Mooney. Generating animated videos of human activities from natural language descrip- tions. Learning, 2018:1, 2018. 1, 2, 6, 7

[21] 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. 6, 8

[22] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Ger- ard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In Pro- ceedings of the IEEE/CVF International Conference on Computer Vision, pages 5442–5451, 2019. 5

[23] Tanya Marwah, Gaurav Mittal, and Vineeth N Balasubrama- nian. Attentive semantic video generation using captions. In Proceedings of the IEEE Interna- tional Conference on Com- puter Vision, pages 1426–1434, 2017. 2

[24] Yingwei Pan, Zhaofan Qiu, Ting Yao, Houqiang Li, and Tao Mei. To create what you tell: Generating videos from cap- tions. In Proceedings of the 25th ACM international confer- ence on Multimedia, pages 1789–1798, 2017. 2

[25] Jae Sung Park, Marcus Rohrbach, Trevor Darrell, and Anna Rohrbach. Ad- versarial inference for multi-sentence video description. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 6598– 6608, 2019. 2

[26] Ramakanth Pasunuru and Mohit Bansal. Reinforced video captioning with entailment rewards. arXiv preprint arXiv:1708.02300, 2017. 2

[27] Wenjie Pei, Jiyuan Zhang, Xiangrong Wang, Lei Ke, Xiaoy- ong Shen, and Yu-Wing Tai. Memory-attended recurrent net- work for video captioning. In Pro- ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogni- tion, pages 8347–8356, 2019. 2

[28] Mathis Petrovich, Michael J Black, and G¨ul Varol. Action- conditioned 3d human motion synthesis with transformer vae. 2021. 2

[29] Yongri Piao, Wei Ji, Jingjing Li, Miao Zhang, and Huchuan Lu. Depth- induced multi-scale recurrent attention network for saliency detection. In ICCV, pages 7254–7263, 2019. 2 [31] Matthias Plappert, Christian Mandery, and TamimAsfour. The kit motion-language dataset. Big data, 4(4):236–252, 2016. 1, 3, 5

[30] Matthias Plappert, Christian Mandery, and Tamim Asfour. Learning a bidi- rectional mapping between human whole- body motion and natural language using deep recurrent neu- ral networks. Robotics and Autonomous Systems, 109:13–26, 2018. 1, 2

[31] Abhinanda R Punnakkal, Arjun Chandrasekaran, Nikos Athanasiou, Alejan- dra Quiros-Ramirez, and Michael J Black. Babel: Bodies, action and behavior with english la- bels. In Proceedings of the IEEE/CVF Conference on Com- puter Vision and Pattern Recognition, pages 722–731, 2021. 3

[32] Eli Shlizerman, Lucio Dery, Hayden Schoen, and Ira Kemelmacher-Shlizerman. Audio to body dynamics. In Pro- ceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 7574–7583, 2018. 2

[33] Kenta Takeuchi, Dai Hasegawa, Shinichi Shirakawa, Naoshi Kaneko, Hiroshi Sakuta, and Kazuhiko Sumi. Speech-to- gesture generation: A challenge in deep learning approach with bi-directional lstm. In Proceedings of the 5th Interna- tional Conference on Human Agent Interaction, pages 365– 369, 2017. 2

[34] Taoran Tang, Jia Jia, and Hanyang Mao. Dance with melody: An lstm- autoencoder approach to music-oriented dance syn- thesis. In Proceedings of the 26th ACM international confer- ence on Multimedia, pages 1598–1606, 2018. 2, 7

[35] Yu Tian, Jian Ren, Menglei Chai, Kyle Olszewski, Xi Peng, Dimitris N Metaxas, and Sergey Tulyakov. A good image generator is what you need for high-resolution video synthe- sis. arXiv preprint

arXiv:2104.15069, 2021. 2

[36] Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. Mocogan: Decomposing motion and content for video generation. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1526–1535, 2018. 2, 6, 7

[37] Aaron Van Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recur- rent neural networks. In In- ternational Conference on Machine Learning, pages 1747–1756. PMLR, 2016. 4

[38] Subhashini Venugopalan, Marcus Rohrbach, Jeffrey Don- ahue, Raymond Mooney, Trevor Darrell, and Kate Saenko. Sequence to sequence-video to text. In Proceedings of the IEEE international conference on computer vision, pages 4534–4542, 2015. 2

[39] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment net- works: Towards good practices for deep action recognition. In European conference on computer vision, pages 20–36. Springer, 2016. 2

[40] Tsun-Hsuan Wang, Yen-Chi Cheng, Chieh Hubert Lin, Hwann-Tzong Chen, and Min Sun. Point-to-point video gen- eration. In Proceedings of the IEEE/CVF International Con- ference on Computer Vision, pages 10491–10500, 2019. 2

[41] Zhenyi Wang, Ping Yu, Yang Zhao, Ruiyi Zhang, Yufan Zhou, Junsong Yuan, and Changyou Chen. Learning di- verse stochastic human-action generators by learning smooth latent transitions. In Proceedings of the AAAI Conference on Articial Intelligence, volume 34, pages 12281–12288, 2020. 2

[42] Tatsuro Yamada, Hiroyuki Matsunaga, and Tetsuya Ogata. Paired recurrent autoencoders for bidirectional translation between robot actions and linguistic descriptions. IEEE Robotics and Automation Letters, 3(4):3441–3448, 2018. 1, 2

[43] Ping Yu, Yang Zhao, Chunyuan Li, Junsong Yuan, and Changyou Chen. Structure-aware human-action generation. In European Conference on Computer Vision, pages 18–34. Springer, 2020. 2

[44] Miao Zhang, Jingjing Li, Wei Ji, Yongri Piao, and Huchuan Lu. Memory- oriented decoder for light eld salient object detection. In NeurIPS, pages 896–906, 2019. 2

[45] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neu- ral networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 5745– 5753, 2019. 6