T2M-GPT: Generating Human Motion from Textual Descriptions with Discrete Representations

Jianrong Zhang[1,3*], Yangsong Zhang[2,3*], Xiaodong Cun[3], Shaoli Huang[3], Yong Zhang[3]
Hongwei Zhao[1], Hongtao Lu[2], Xi Shen[3,†]

[*]Equal contribution [†]Corresponding author

[1]Jilin University [2]Shanghai Jiao Tong University [3]Tencent AI Lab

Abstract

In this work, we investigate a simple and must-known conditional generative framework based on Vector Quantised-Variational AutoEncoder (VQ-VAE) and Generative Pre-trained Transformer (GPT) for human motion generation from textural descriptions. We show that a simple CNN-based VQ-VAE with commonly used training recipes (EMA and Code Reset) allows us to obtain high-quality discrete representations. For GPT, we incorporate a simple corruption strategy during the training to alleviate training-testing discrepancy. Despite its simplicity, our T2M-GPT shows better performance than competitive approaches, including recent diffusion-based approaches. For example, on HumanML3D, which is currently the largest dataset, we achieve comparable performance on the consistency between text and generated motion (R-Precision), but with FID 0.116 largely outperforming MotionDiffuse of 0.630. Additionally, we conduct analyses on HumanML3D and observe that the dataset size is a limitation of our approach. Our work suggests that VQ-VAE still remains a competitive approach for human motion generation. Our implementation is available on the project page: https://mael-zys.github.io/T2M-GPT/.

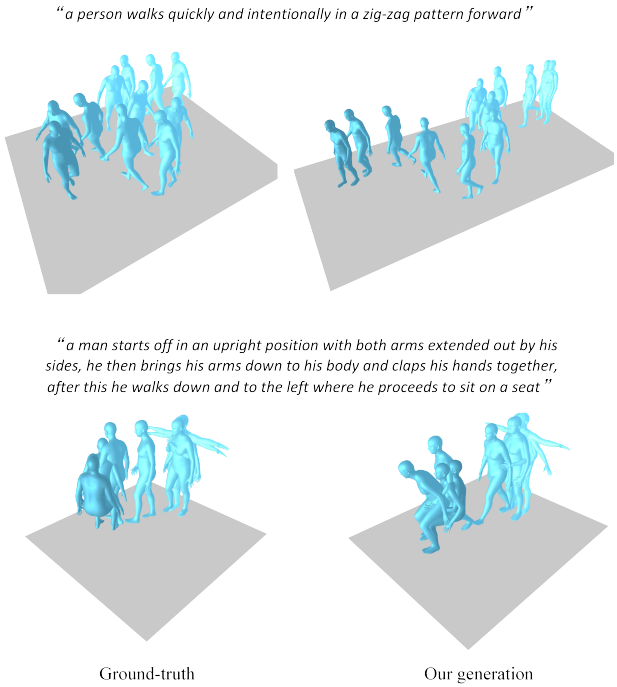


Figure 1. **Visual results on HumanML3D [21].** Our approach is able to generate precise and high-quality human motion consistent with challenging text descriptions. More visual results are on the project page.

1. Introduction

Generating motion from textual descriptions can be used in numerous applications in the game industry, film-making, and animating robots. For example, a typical way to access new motion in the game industry is to perform motion capture, which is expensive. Therefore automatically generating motion from textual descriptions, which allows producing meaningful motion data, could save time and be more economical.

Motion generation conditioned on natural language is challenging, as motion and text are from different modalities. The model is expected to learn precise mapping from the language space to the motion space. To this end, many works propose to learn a joint embedding for language and motion using auto-encoders [3, 20, 63] and VAEs [50, 51]. MotionClip [63] aligns the motion space to CLIP [53] space. ACTOR [50] and TEMOES [51] propose transformer-based VAEs for action-to-motion and text-to-motion respectively. These works show promising performances with simple descriptions and are limited to producing high-quality motion when textual descriptions become long and complicated. Guo *et al.* [21] and TM2T [22] aim to generate motion

sequences with more challenging textual descriptions. However, both approaches are not straightforward, involve three stages for text-to-motion generation, and sometimes fail to generate high-quality motion consistent with the text (See Figure 4 and more visual results on the project page). Recently, diffusion-based models [31] have shown impressive results on image generation [58], which are then introduced to motion generation by MDM [64] and MotionDiffuse [71] and dominates text-to-motion generation task. However, we find that compared to classic approaches, such as VQ-VAE [66], the performance gain of the diffusion-based approaches [64, 71] might not be that significant. In this work, we are inspired by recent advances from learning the discrete representation for generation [5, 15, 16, 18, 56, 66, 68] and investigate a simple and classic framework based on Vector Quantized Variational Autoencoders (VQ-VAE) [66] and Generative Pre-trained Transformer (GPT) [54, 67] for text-to-motion generation.

Precisely, we propose a two-stage method for motion generation from textual descriptions. In stage 1, we use a standard 1D convolutional network to map motion sequences to discrete code indices. In stage 2, a standard GPT-like model [54, 67] is learned to generate sequences of code indices from pre-trained text embedding. We find that the naive training of VQ-VAE [66] suffers from code collapse. One effective solution is to leverage two standard recipes during the training: *EMA* and *Code Reset*. We provide a full analysis of different quantization strategies. For GPT, the next token prediction brings inconsistency between the training and inference. We observe that simply corrupting sequences during the training alleviates this discrepancy. Moreover, throughout the evolution of image generation, the size of the dataset has played an important role. We further explore the impact of dataset size on the performance of our model. The empirical analysis suggests that the performance of our model can potentially be improved with larger datasets.

Despite its simplicity, our approach can generate high-quality motion sequences that are consistent with challenging text descriptions (Figure 1 and more on the project page). Empirically, we achieve comparable or even better performances than concurrent diffusion-based approaches MDM [64] and HumanDiffuse [71] on two widely used datasets: HumanML3D [21] and KIT-ML [52]. For example, on HumanML3D, which is currently the largest dataset, we achieve comparable performance on the consistency between text and generated motion (R-Precision), but with FID 0.116 largely outperforming MotionDiffuse of 0.630. We conduct comprehensive experiments to explore this area, and hope that these experiments and conclusions will contribute to future developments.

In summary, our contributions include:

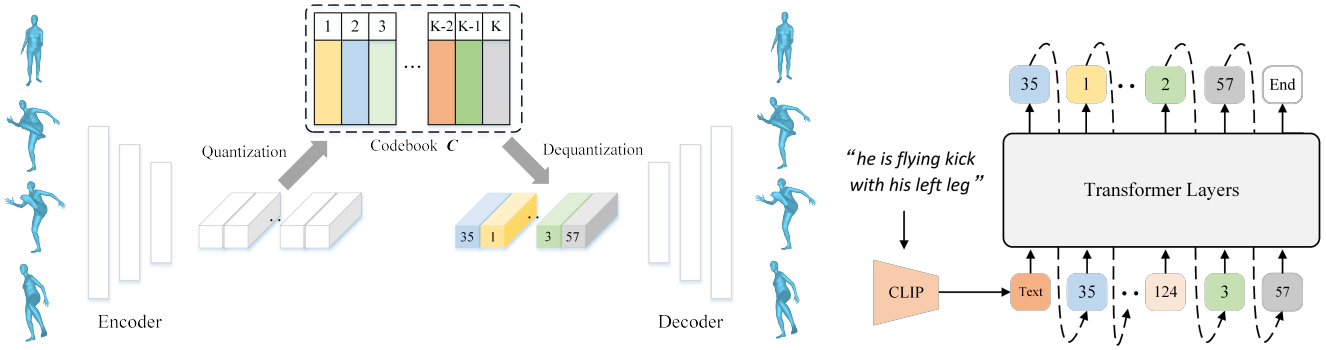
- We present a simple yet effective approach for motion generation from textual descriptions. Our approach achieves state-of-the-art performance on HumanML3D [21] and KIT-ML [52] datasets.
- We show that GPT-like models incorporating discrete representations still remain a very competitive approach for motion generation.
- We provide a detailed analysis of the impact of quantization strategies and dataset size. We show that a larger dataset might still offer a promising prospect to the community.

Our implementation is available on the project page.

2. Related work

VQ-VAE. Vector Quantized Variational Autoencoders (VQ-VAE), which is a variant of VAE [35], is initially proposed in [66]. VQ-VAE is composed of an AutoEncoder architecture, which aims at learning reconstruction with discrete representations. Recently, VQ-VAE achieves promising performance on generative tasks across different modalities, which includes: image synthesis [18, 68], text-to-image generation [56], speech gesture generation [5], music generation [15, 16] etc. The success of VQ-VAE for generation might be attributed to its decoupling of learning the discrete representation and the prior. A naive training of VQ-VAE suffers from the codebook collapse, *i.e.*, only a number of codes are activated, which importantly limited the performances of the reconstruction as well as generation. To alleviate the problem, a number of techniques can be used during training, including stop-gradient along with some losses [66] to optimize the codebook, exponential moving average (EMA) for codebook update [68], reset inactivated codes during the training (Code Reset [68]), etc.

Human motion synthesis. Research on human motion synthesis has a long history [8]. One of the most active research fields is human motion prediction, which aims at predicting the future motion sequence based on past observed motion. Approaches mainly focus on efficiently and effectively fusing spatial and temporal information to generate deterministic future motion through different models: RNN [12, 19, 48, 49], GAN [9, 29], GCN [47], Attention [46] or even simply MLP [11, 24]. Some approaches aim at generating diverse motion through VAE [4, 25, 70]. In addition to synthesizing motion conditioning on past motion, another related topic is generating motion "in-betweening" that takes both past and future poses and fills motion between them [17, 26, 27, 34, 62]. [49] considers the generation of locomotion sequences from a given trajectory for simple actions such as: walking and running. Motion can also be generated with music to produce 3D dance motion [6, 13, 36, 37, 39, 40]. For unconstrained generations, [69] generates a long sequence altogether by transforming from a sequence of latent



(a) Motion VQ-VAE (b) T2M-GPT

Figure 2. **Overview of our framework for text-driven motion generation.** It includes two modules: Motion VQ-VAE (Figure 2a) and T2M-GPT (Figure 2b). In T2M-GPT, an additional learnable *End* token is inserted to indicate the stop of the generation. During the inference, we first generate code indexes in an auto-regressive fashion and then obtain the motion using the decoder in Motion VQ-VAE.

vectors sampled from a Gaussian process. In graphics literature, many works focus on animator control. Holden *et al.* [33] learn a convolutional autoencoder to reconstruct motion, the learned latent representation can be used to synthesize and edit motion. [32] proposes phase functioned neural network to perform the control task. [61] uses a deep auto-regressive framework to scene interaction behaviors. Starke *et al.* [60] proposes to reconstruct motion through periodic features, the learned periodic embedding improves motion synthesis. Recently, inspired by SinGAN [59] for image synthesis, Li *et al.* [38] propose a generative model approach for motion synthesis from a single sequence.

Text-driven human motion generation. Text-driven human motion generation aims at generating 3D human motion from textual descriptions. Text2Action [2] trains an RNN-based model to generate motion conditioned on a short text. Language2Pose [3] employs a curriculum learning approach to learn a joint embedding space for both text and pose. The decoder can thus take text embedding to generate motion sequences. Ghost *et al.* [20] learn two manifold representations for the upper body and the lower body movements, which shows improved performance compared to Language2Pose [3]. Similarly, MotionCLIP [63] also tends to align text and motion embedding but proposes to utilize CLIP [53] as the text encoder and employ rendered images as extra supervision. It shows the ability to generate out-of-distribution motion and enable latent code editing. However, the generated motion sequences are not in high-quality and are without global translation. ACTOR [50] proposes a transformer-based VAE to generate motion in a non-autoregressive fashion from a pre-defined action class. TEMOS [51] extends the architecture of ACTOR [50] by introducing an additional text encoder and producing diverse motion sequences given text descriptions. TEMOS demonstrates its effect on KIT Motion-Language [52] with mainly short sentences and suffers from out-of-distribution descriptions [51]. TEACH [7] further extends TEMOS to generate temporal motion compositions from a series of natural language descriptions. Recently, a large-scale dataset HumanML3D is proposed in [21]. Guo *et al.* [21] also propose to incorporate motion length prediction from text to produce motion with reasonable length. TM2T [22] considers text-to-motion and motion-to-text tasks. It also shows additional improvement can be obtained through jointly training both tasks. As concurrent works, diffusion-based models are introduced for text-to-motion generation by MDM [64] and MotionDiffuse [71]. In this work, we show that without any sophisticated designs, the classic VQ-VAE framework could achieve competitive or even better performance with a classical framework and some standard training recipes.

3. Method

Our goal is to generate high-quality motion that is consistent with text descriptions. The overall framework consists of two modules: Motion VQ-VAE and T2M-GPT, which is illustrated in Figure 2. The former learns a mapping between motion data and discrete code sequences, the latter generates code indices conditioned on the text description. With the decoder in Motion VQ-VAE, we are able to recover the motion from the code indices. In Section 3.1, we present the VQ-VAE module. The T2M-GPT is introduced in Section 3.2.

3.1. Motion VQ-VAE

VQ-VAE, proposed in [66], enables the model to learn discrete representations for generative models. Given a motion sequence $X = [x_1, x_2, \ldots, x_T]$ with $x_t \in \mathbb{R}^d$, where T is the number of frames and d is the dimension of the motion, we aim to recover the motion sequence through an autoencoder and a learnable codebook containing K codes $C = \{c_k\}_{k=1}^{K}$ with $c_k \in \mathbb{R}^{d_c}$, where d_c is the dimension of

codes. The overview of VQ-VAE is presented in Figure 2a. With encoder and decoder of the autoencoder denoted by E and D, the latent feature Z can be computed as $Z = E(X)$ with $Z = [z_1, z_2, ..., z_{T/l}]$ and $z_i \in \mathbb{R}^{d_c}$, where l represents the temporal downsampling rate of the encoder E. For i-th latent feature z_i, the quantization through C is to find the most similar element in C, which can be properly written as:

$$\hat{z}_i = \underset{c_k \in C}{\arg\min} \|z_i - c_k\|_2 \tag{1}$$

Optimization goal. To optimize VQ-VAE, the standard optimization goal [66] $\mathcal{L}_{\text{vq}}$ contains three components: a reconstruction loss $\mathcal{L}_{\text{re}}$, the embedding loss $\mathcal{L}_{embed}$ and the commitment loss $\mathcal{L}_{commit}$.

$$\mathcal{L}_{vq} = \mathcal{L}_{re} + \underbrace{||Z - sg[\hat{Z}]||_2}_{\mathcal{L}_{embed}} + \beta \underbrace{||sg[Z] - \hat{Z}||_2}_{\mathcal{L}_{commit}} \tag{2}$$

where β is a hyper-parameter for the commitment loss and sg is the stop-gradient operator. For the reconstruction, we find that L1 smooth loss $\mathcal{L}_1^{smooth}$ performs best and an additional regularization on the velocity enhances the generation quality. Let X_{re} be the reconstructed motion of X, i.e., $X_{\text{re}} = D(\hat{Z})$, V(X) be the velocity of X where $V = [v_1, v_2, \ldots, v_{T-1}]$ with $v_i = x_{i+1} - x_i$. Therefore, the objective of the reconstruction is as follows:

$$\mathcal{L}_{re} = \mathcal{L}_1^{smooth}(X, X_{re}) + \alpha \mathcal{L}_1^{smooth}(V(X), V(X_{re})) \tag{3}$$

where α is a hyper-parameter to balance the two losses. We provide an ablation study on α as well as different reconstruction losses ($\mathcal{L}_1$, $\mathcal{L}_1^{smooth}$ and $\mathcal{L}_2$) in Section B of Appendix.

Quantization strategy. A naive training of VQ-VAE suffers from codebook collapse [57, 66]. Two training recipes [57] are commonly used to improve the codebook utilization: exponential moving average *(EMA)* and codebook reset *(Code Reset)*. *EMA* makes the codebook C evolve smoothly: $C^t \leftarrow \lambda C^{t-1} + (1-\lambda)C^{t-1}$, where C^t is the codebook at iteration t and λ is the exponential moving constant. *Code Reset* finds inactivate codes during the training and reassigns them according to input data. We provide an ablation study on the quantization strategy in Section 4.3.

Architecture. We use a simple convolutional architecture composed of 1D convolution, residual block [28], and ReLU. Our VQ-VAE architecture is illustrated in Figure 3. The architecture is inspired by [18, 40]. We use convolution with stride 2 and nearest interpolation for temporal downsampling and upsampling respectively. The downsampling rate is thus $l = 2^L$, where L denotes the number of residual blocks. We provide an ablation study on the architecture in Section 4.3. The detail of the architecture is provided in Section F of the Appendix.

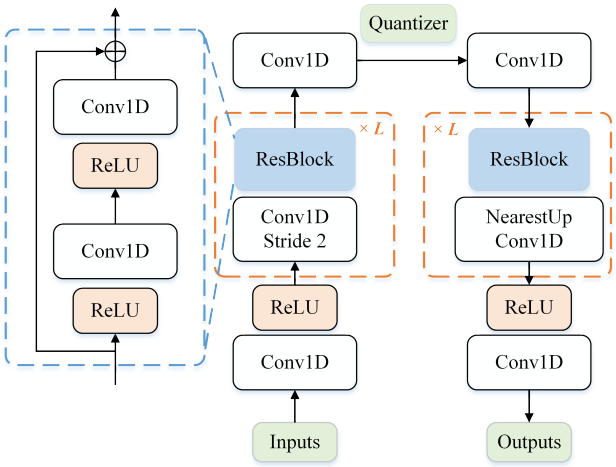


Figure 3. **Architecture of the motion VQ-VAE.** We use a standard CNN-based architecture with 1D convolution (*Conv1D*), residual block (*ResBlock*) and ReLU activation. 'L' denotes the number of residual blocks. We use convolution with stride 2 and nearest interpolation for temporal downsampling and upsampling.

3.2. T2M-GPT

With a learned motion VQ-VAE, a motion sequence $X = [x_1, x_2, \ldots, x_T]$ can be mapped to a sequence of indices $S = [s_1, s_2, \ldots, s_{T/l}, End]$ with $s_i \in [1, 2, \ldots, s_{T/l}]$, which are indices from the learned codebook. Note that a special End token is added to indicate the stop of the motion, which is different from [21] that leverages an extra module to predict motion length. By projecting S back to their corresponding codebook entries, we obtain $\hat{Z} = [\hat{z}_1, \hat{z}_2, \ldots, \hat{z}_{T/l}]$ with $\hat{z}_i = c_{s_i}$, which can be decoded to a motion X_{re} through the decoder D. Therefore, text-to-motion generation can be formulated as an autoregressive next-index prediction: given previous $i-1$ indices, i.e., $S_{<i}$, and text condition c, we aim to predict the distribution of possible next indices $p(S_i|c, S_{<i})$, which can be addressed with transformer [67]. The overview of our transformer is shown in Figure 2b.

Optimization goal. Denoting the likelihood of the full sequence as $p(S|c) = \prod_{i=1}^{|S|} p(S_i|c, S_{<i})$, we directly maximize the log-likelihood of the data distribution:

$$\mathcal{L}_{trans} = \mathbb{E}_{S \sim p(S)}[-\log p(S|c)] \tag{4}$$

We leverage CLIP [53] to extract text embedding c, which has shown its effectiveness in relevant tasks [14, 55, 63].

Causal Self-attention. We apply the causal self-attention [54] in T2M-GPT. Precisely, the output of the causal self-attention is calculated as follows:

$$\text{Attention} = \text{Softmax}\left(\frac{QK^T \times mask}{\sqrt{d_k}}\right) \tag{5}$$

Methods	R-Precision ↑			FID ↓	MM-Dist ↓	Diversity ↑	MModality ↑
	Top-1	Top-2	Top-3				
Real motion	$0.511^{\pm.003}$	$0.703^{\pm.003}$	$0.797^{\pm.002}$	$0.002^{\pm.000}$	$2.974^{\pm.008}$	$9.503^{\pm.065}$	-
Our VQ-VAE (Recons.)	$0.501^{\pm.002}$	$0.692^{\pm.002}$	$0.785^{\pm.002}$	$0.070^{\pm.001}$	$3.072^{\pm.009}$	$9.593^{\pm.079}$	-
Seq2Seq [41]	$0.180^{\pm.002}$	$0.300^{\pm.002}$	$0.396^{\pm.002}$	$11.75^{\pm.035}$	$5.529^{\pm.007}$	$6.223^{\pm.061}$	-
Language2Pose [3]	$0.246^{\pm.002}$	$0.387^{\pm.002}$	$0.486^{\pm.002}$	$11.02^{\pm.046}$	$5.296^{\pm.008}$	$7.676^{\pm.058}$	-
Text2Gesture [10]	$0.165^{\pm.001}$	$0.267^{\pm.002}$	$0.345^{\pm.002}$	$5.012^{\pm.030}$	$6.030^{\pm.008}$	$6.409^{\pm.071}$	-
Hier [20]	$0.301^{\pm.002}$	$0.425^{\pm.002}$	$0.552^{\pm.004}$	$6.532^{\pm.024}$	$5.012^{\pm.018}$	$8.332^{\pm.042}$	-
MoCoGAN [65]	$0.037^{\pm.000}$	$0.072^{\pm.001}$	$0.106^{\pm.001}$	$94.41^{\pm.021}$	$9.643^{\pm.006}$	$0.462^{\pm.008}$	$0.019^{\pm.000}$
Dance2Music [36]	$0.033^{\pm.000}$	$0.065^{\pm.001}$	$0.097^{\pm.001}$	$66.98^{\pm.016}$	$8.116^{\pm.006}$	$0.725^{\pm.011}$	$0.043^{\pm.001}$
TM2T [22]	$0.424^{\pm.003}$	$0.618^{\pm.003}$	$0.729^{\pm.002}$	$1.501^{\pm.017}$	$3.467^{\pm.011}$	$8.589^{\pm.076}$	$2.424^{\pm.093}$
Guo *et al.* [21]	$0.455^{\pm.003}$	$0.636^{\pm.003}$	$0.736^{\pm.002}$	$1.087^{\pm.021}$	$3.347^{\pm.008}$	$9.175^{\pm.083}$	$2.219^{\pm.074}$
MDM [64]§	-	-	$0.611^{\pm.007}$	$0.544^{\pm.044}$	$5.566^{\pm.027}$	$9.559^{\pm.086}$	$2.799^{\pm.072}$
MotionDiffuse [71]§	$0.491^{\pm.001}$	$0.681^{\pm.001}$	$0.782^{\pm.001}$	$0.630^{\pm.001}$	$3.113^{\pm.001}$	$9.410^{\pm.049}$	$1.553^{\pm.042}$
Our GPT ($\tau = 0$)	$0.417^{\pm.003}$	$0.589^{\pm.002}$	$0.685^{\pm.003}$	$0.140^{\pm.006}$	$3.730^{\pm.009}$	$9.844^{\pm.095}$	$3.285^{\pm.070}$
Our GPT ($\tau = 0.5$)	$0.491^{\pm.003}$	$0.680^{\pm.003}$	$0.775^{\pm.002}$	$0.116^{\pm.004}$	$3.118^{\pm.011}$	$9.761^{\pm.081}$	$1.856^{\pm.011}$
Our GPT ($\tau \in \mathcal{U}[0, 1]$)	$0.492^{\pm.003}$	$0.679^{\pm.002}$	$0.775^{\pm.002}$	$0.141^{\pm.005}$	$3.121^{\pm.009}$	$9.722^{\pm.082}$	$1.831^{\pm.048}$

§ reports results using ground-truth motion length.

Table 1. **Comparison with the state-of-the-art methods on HumanML3D [21] test set.** We compute standard metrics following Guo *et al.* [21]. For each metric, we repeat the evaluation 20 times and report the average with 95% confidence interval. Red and Blue indicate the best and the second best result.

where $Q \in \mathbb{R}^{T \times d_k}$ and $K \in \mathbb{R}^{T \times d_k}$ are query and key respectively, while $mask$ is the causal mask with $mask_{i,j} = -\infty \times \mathbf{1}(i > j) + \mathbf{1}(i \leq j)$, where $\mathbf{1}(\cdot)$ is the indicator function. This causal mask ensures that future information is not allowed to attend the calculation of current tokens. For inference, we start from the text embedding and generate indices in an autoregressive fashion, the generation process will be stopped if the model predicts the End token. Note that we are able to generate diverse motions by sampling from the predicted distributions given by the transformer.

Corrupted sequences for the training-testing discrepancy. There is a discrepancy between training and testing. For training, $i - 1$ correct indices are used to predict the next index. While for inference, there is no guarantee that indices serving as conditions are correct. To address this problem, we adopt a simple data augmentation strategy: we replace $\tau \times 100\%$ ground-truth code indices with random ones during training. τ can be a hyper-parameter or randomly sampled from $\tau \in \mathcal{U}[0, 1]$. We provide an ablation study on this strategy in Section C of the Appendix.

4. Experiment

In this section, we present our experimental results. In Section 4.1, we introduce standard datasets as well as evaluation metrics. We compare our results to competitive approaches in Section 4.2. Finally, we provide analysis and discussion in Section 4.3.

4.1. Datasets and evaluation metric

We conduct experiments on two standard datasets for text-driven motion generations: HumanML3D [21] and KIT Motion-Language (KIT-ML) [52]. Both datasets are commonly used in the community. We follow the evaluation protocol proposed in [21].

KIT Motion-Language (KIT-ML). KIT-ML [52] contains 3,911 human motion sequences and 6,278 textual annotations. The total vocabulary size, that is the number of unique words disregarding capitalization and punctuation, is 1,623. Motion sequences are selected from KIT [45] and CMU [1] datasets but downsampled into 12.5 frame-per-second (FPS). Each motion sequence is described by from 1 to 4 sentences. The average length of descriptions is approximately 8. Following [21, 22], the dataset is split into training, validation, and test sets with proportions of 80%, 5%, and 15%, respectively. We select the model that achieves the best FID on the validation set and reports its performance on the test set.

HumanML3D. HumanML3D [21] is currently the largest 3D human motion dataset with textual descriptions. The dataset contains 14,616 human motion and 44,970 text descriptions. The entire textual descriptions are composed of 5,371 distinct words. The motion sequences are originally from AMASS [44] and HumanAct12 [23] but with specific pre-processing: motion is scaled to 20 FPS; those that are longer than 10 seconds are randomly cropped to 10-second

Methods	R-Precision ↑			FID ↓	MM-Dist ↓	Diversity ↑	MModality ↑
	Top-1	Top-2	Top-3				
Real motion	$0.424^{\pm.005}$	$0.649^{\pm.006}$	$0.779^{\pm.006}$	$0.031^{\pm.004}$	$2.788^{\pm.012}$	$11.08^{\pm.097}$	-
Our VQ-VAE (Recons.)	$0.399^{\pm.005}$	$0.614^{\pm.005}$	$0.740^{\pm.006}$	$0.472^{\pm.011}$	$2.986^{\pm.027}$	$10.994^{\pm.120}$	-
Seq2Seq [41]	$0.103^{\pm.003}$	$0.178^{\pm.005}$	$0.241^{\pm.006}$	$24.86^{\pm.348}$	$7.960^{\pm.031}$	$6.744^{\pm.106}$	-
Language2Pose [3]	$0.221^{\pm.005}$	$0.373^{\pm.004}$	$0.483^{\pm.005}$	$6.545^{\pm.072}$	$5.147^{\pm.030}$	$9.073^{\pm.100}$	-
Text2Gesture [10]	$0.156^{\pm.004}$	$0.255^{\pm.004}$	$0.338^{\pm.005}$	$12.12^{\pm.183}$	$6.964^{\pm.029}$	$9.334^{\pm.079}$	-
Hier [20]	$0.255^{\pm.006}$	$0.432^{\pm.007}$	$0.531^{\pm.007}$	$5.203^{\pm.107}$	$4.986^{\pm.027}$	$9.563^{\pm.072}$	-
MoCoGAN [65]	$0.022^{\pm.002}$	$0.042^{\pm.003}$	$0.063^{\pm.003}$	$82.69^{\pm.242}$	$10.47^{\pm.012}$	$3.091^{\pm.043}$	$0.250^{\pm.009}$
Dance2Music [36]	$0.031^{\pm.002}$	$0.058^{\pm.002}$	$0.086^{\pm.003}$	$115.4^{\pm.240}$	$10.40^{\pm.016}$	$0.241^{\pm.004}$	$0.062^{\pm.002}$
TM2T [22]	$0.280^{\pm.005}$	$0.463^{\pm.006}$	$0.587^{\pm.005}$	$3.599^{\pm.153}$	$4.591^{\pm.026}$	$9.473^{\pm.117}$	$3.292^{\pm.081}$
Guo *et al.* [21]	$0.361^{\pm.006}$	$0.559^{\pm.007}$	$0.681^{\pm.007}$	$3.022^{\pm.107}$	$3.488^{\pm.028}$	$10.72^{\pm.145}$	$2.052^{\pm.107}$
MDM [64]§	-	-	$0.396^{\pm.004}$	$0.497^{\pm.021}$	$9.191^{\pm.022}$	$10.847^{\pm.109}$	$1.907^{\pm.214}$
MotionDiffuse [71]§	$0.417^{\pm.004}$	$0.621^{\pm.004}$	$0.739^{\pm.004}$	$1.954^{\pm.062}$	$2.958^{\pm.005}$	$11.10^{\pm.143}$	$0.730^{\pm.013}$
Our GPT ($\tau = 0$)	$0.392^{\pm.007}$	$0.600^{\pm.007}$	$0.716^{\pm.006}$	$0.737^{\pm.049}$	$3.237^{\pm.027}$	$11.198^{\pm.086}$	$2.309^{\pm.055}$
Our GPT ($\tau = 0.5$)	$0.402^{\pm.006}$	$0.619^{\pm.005}$	$0.737^{\pm.006}$	$0.717^{\pm.041}$	$3.053^{\pm.026}$	$10.862^{\pm.094}$	$1.912^{\pm.036}$
Our GPT ($\tau \in \mathcal{U}[0, 1]$)	$0.416^{\pm.006}$	$0.627^{\pm.006}$	$0.745^{\pm.006}$	$0.514^{\pm.029}$	$3.007^{\pm.023}$	$10.921^{\pm.108}$	$1.570^{\pm.039}$

§ reports results using ground-truth motion length.

Table 2. **Comparison with the state-of-the-art methods on KIT-ML [52] test set.** We compute standard metrics following Guo *et al.* [21]. For each metric, we repeat the evaluation 20 times and report the average with 95% confidence interval. Red and Blue indicate the best and the second best result.

ones; they are then re-targeted to a default human skeletal template and properly rotated to face Z+ direction initially. Each motion is paired with at least 3 precise textual descriptions. The average length of descriptions is approximately 12. According to [21], the dataset is split into training, validation, and test sets with proportions of 80%, 5%, and 15%, respectively. We select the model that achieves the best FID on the validation set and reports its performance on the test set.

Implementation details. For Motion VQ-VAE, the codebook size is set to 512×512. The downsampling rate l is 4. We provide an ablation on the number of codes in Section D of Appendix. For both HumanML3D [21] and KIT-ML [52] datasets, the motion sequences are cropped to $T = 64$ for training. We use AdamW [43] optimizer with $[\beta_1, \beta_2] = [0.9, 0.99]$, batch size of 256, and exponential moving constant $\lambda = 0.99$. We train the first 200K iterations with a learning rate of 2e-4, and 100K with a learning rate of 1e-5. β and α in $\mathcal{L}_{\text{vq}}$ and $\mathcal{L}_{\text{re}}$ are set to 1 and 0.5, respectively. Following [21], the dataset KIT-ML and HumanML3D are extracted into motion features with dimensions 251 and 263 respectively, which correspond to local joints position, velocity, and rotations in root space as well as global translation and rotations. These features are computed from 21 and 22 joints of SMPL [42]. More details about the motion representations are provided in Section E of Appendix.

For the T2M-GPT, we employ 18 transformer [67] layers with a dimension of 1,024 and 16 heads. The ablation for different scales of transformer is provided in Section A of Appendix. Following Guo *et al*. [21], the maximum length of Motion is 196 on both datasets, and the minimum lengths are 40 and 24 for HumanML3D [21] and KIT-ML [52] respectively. The maximum length of the code index sequence is $T' = 50$. We train an extra *End* token as a signal to stop index generation. The transformer is optimized using AdamW [43] with $[\beta_1, \beta_2] = [0.5, 0.99]$ and batch size 128. The initialized learning rate is set to 1e-4 for 150K iterations and decayed to 5e-6 for another 150K iterations.

Training Motion VQ-VAE and T2M-GPT take about 14 hours and 78 hours respectively on a single Tesla V100-32G GPU.

Evaluation metric. Following [21], global representations of motion and text descriptions are first extracted with the pre-trained network in [21], and then measured by the following five metrics:

- *R-Precision.* Given one motion sequence and 32 text descriptions (1 ground-truth and 31 randomly selected mismatched descriptions), we rank the Euclidean distances between the motion and text embeddings. Top-1, Top-2, and Top-3 accuracy of motion-to-text retrieval are reported.

- *Frechet Inception Distance (FID).* We calculate the distribution distance between the generated and real motion using FID [30] on the extracted motion features.

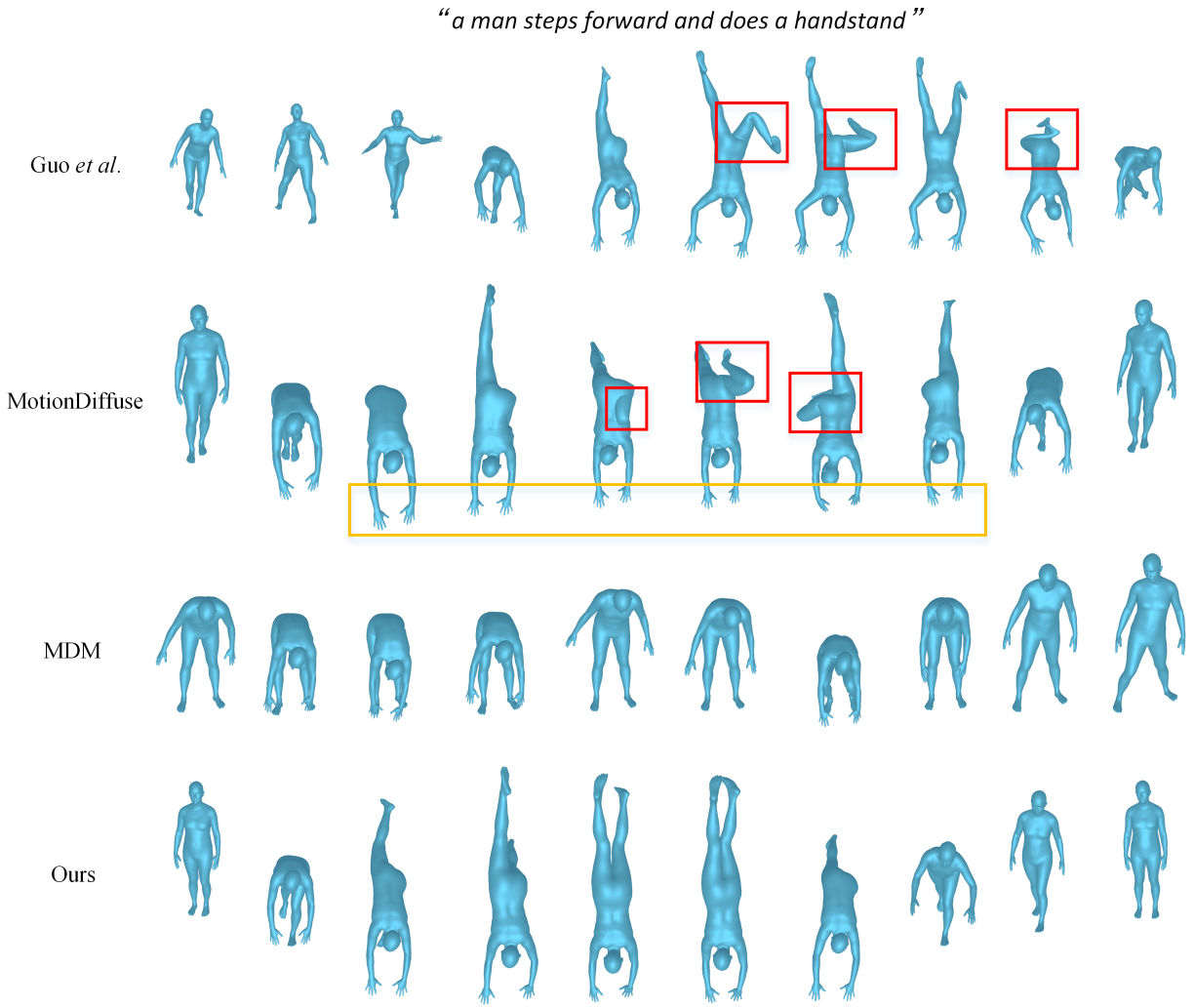


Figure 4. **Visual results on HumanML3D [21] dataset.** We compare our generation with Guo *et al.* [21], MotionDiffuse [71], and MDM [64]. Distorted motions (red) and sliding (yellow) are highlighted. More visual results can be found on the project page project page.

- *Multimodal Distance (MM-Dist).* The average Euclidean distances between each text feature and the generated motion feature from this text.
- *Diversity.* From a set of motion, we randomly sample 300 pairs of motion. We extract motion features and compute the average Euclidean distances of the pairs to measure motion diversity in the set.
- *Multimodality (MModality).* For one text description, we generate 20 motion sequences forming 10 pairs of motion. We extract motion features and compute the average Euclidean distances of the pairs. We finally report the average over all the text descriptions.

Note that more details about the evaluation metrics are provided in Section E of Appendix.

4.2. Comparison to state-of-the-art approaches

We compare to existing state-of-the-art methods [3, 10, 20–22, 36, 41, 64, 65, 71] on the test set of HumanML3D [21] and KIT-ML [52]. Note that more visual results are provided on our project page.

Quantitative results. We show the comparison results in Table 1 and Table 2 on HumanML3D [21] test set and KIT-ML [52] test set. On both datasets, our reconstruction with VQ-VAE reaches close performances to real motion, which suggests high-quality discrete representations learned by our VQ-VAE. For the generation, our approach achieves comparable performance on text-motion consistency (R-Precision and MM-Dist) compared to state-of-the-art method MotionDiffuse [71], while significantly outperforms MotionDif-

Quantizer		Reconstruction		Generation	
Code Reset	EMA	FID ↓	Top-1 ↑	FID ↓	Top-1 ↑
		$0.492^{\pm.004}$	$0.436^{\pm.003}$	$42.797^{\pm.156}$	$0.048^{\pm.001}$
	✓	$0.097^{\pm.001}$	$0.499^{\pm.002}$	$0.176^{\pm.008}$	$0.490^{\pm.002}$
✓		$0.102^{\pm.001}$	$0.494^{\pm.003}$	$0.248^{\pm.009}$	$0.461^{\pm.002}$
✓	✓	$\mathbf{0.070}^{\pm.001}$	$\mathbf{0.501}^{\pm.002}$	$\mathbf{0.116}^{\pm.004}$	$\mathbf{0.491}^{\pm.003}$

Table 3. **Analysis of VQ-VAE quantizers on HumanML3D [21] test set.** For all the quantizers, we set $\tau = 0.5$ and use the same architectures (VQ-VAE and GPT) described in Section 4.1. We report FID and Top-1 for both reconstruction and generation. For each metric, we repeat the evaluation 20 times and report the average with 95% confidence interval.

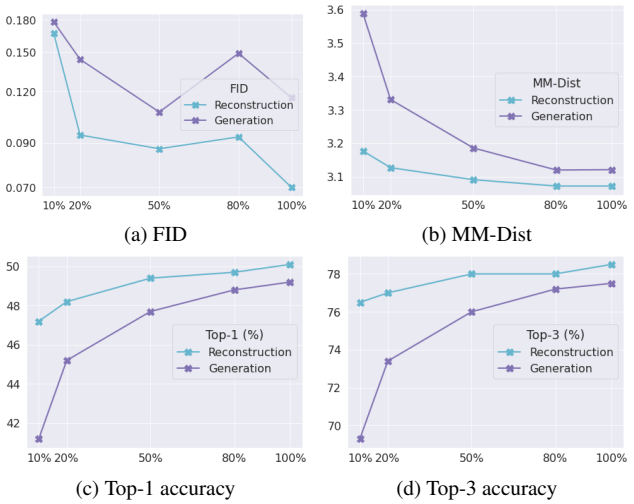
fuse with FID metric. KIT-ML [52] and HumanML3D [21] are in different scales, which demonstrates the robustness of the proposed approach. Manually corrupting sequences during the training of GPT brings consistent improvement ($\tau = 0.5$ v.s. $\tau = 0$). A more detailed analysis is provided in Section C. Unlike Guo *et al.* [21] involving an extra module to predict motion length, we implicitly learn the motion length through an additional *End* token, which is more straightforward and shown to be more effective. Note that MDM [64] and MotionDiffuse [71] evaluate their models with the ground-truth motion length, which is not practical for real applications.

Qualitative comparison. Figure 4 shows visual results on HumanML3D [21]. We compare our generations with the current state-of-the-art models: Guo *et al.* [21], MDM [64] and MotionDiffuse [71]. From the example in Figure 4, one can figure out that our model generates human motion with better quality than the others, and we highlight in red for unrealistic motion generated by Guo *et al.* [21] and MotionDiffuse [71]. Moreover, the generated motion of MDM [64] is not related to the semantics of the description. Note that more visualization results are provided on the project page.

4.3. Discussion

Quantization strategies. We first investigate the impact of different quantization strategies presented in Section 3.1. The results are illustrated in Table 3 for both reconstruction and generation. We notice that naive VQ-VAE training is not able to reconstruct nor generate high-quality motion. However, training with *EMA* or *Code Reset* can importantly boost the performances for both reconstruction and generation.

Impact of dataset size. We further analyze the impact of dataset size. To understand whether the largest dataset HumanML3D [21] contains enough data for motion generation, we train our motion VQ-VAE and T2M-GPT on different subsets of the training data, which consists of 10%, 20%, 50%, 80% and 100% of the training data respectively. The trained models are evaluated on the entire test set. The results



(a) FID (b) MM-Dist

(c) Top-1 accuracy (d) Top-3 accuracy

Figure 5. **Impact of dataset size on HumanML3D [21].** We train our motion VQ-VAE (*Reconstruction*) and T2M-GPT (*Generation*) on the subsets of HumanML3D [21] composed of 10%, 20%, 50%, 80%, and 100% training set respectively. All the models are evaluated on the entire test set. We report FID, MM-Dist, Top-1, and Top-3 accuracy for all the models. Results suggest that our model might benefit from more training data.

are illustrated in Figure 5. We evaluate reconstruction for our motion VQ-VAE and generation for our T2M-GPT using four metrics: FID, MM-Dist, Top-1, and Top-3 accuracies. Several insights can be figured out: *i)* metric for motion quality (FID) and metric for motion-text consistency (MM-Dist, Top-1, and Top-3) should be considered at the same time. With only 10% data, the motion might be of good quality, however, the model is not able to generate a correct motion that corresponds to the text description; *ii)* the performances become better with more training data. This trend suggests that additional training data could bring non-negligible improvement to both reconstruction and generation.

5. Conclusion

In this work, we investigated a classic framework based on VQ-VAE and GPT to synthesize human motion from textual descriptions. Our method achieved comparable or even better performances than concurrent diffusion-based approaches, suggesting that this classic framework remains a very competitive approach for motion generation. We explored in detail the effect of various quantization strategies on human motion reconstruction and generation. Moreover, we provided an analysis of the dataset size. Our finding suggests that a larger dataset could still bring additional improvement to our approach.

Acknowledgement We thank Mathis Petrovich, Yuming Du, Yingyi Chen, Dexiong Chen, and Xuelin Chen for inspiring discussions and valuable feedback.

References

[1] Cmu graphics lab motion capture database. http://mocap.cs.cmu.edu/. Accessed: 2022-11-11. 5

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

[3] Chaitanya Ahuja and Louis-Philippe Morency. Language2pose: Natural language grounded pose forecasting. In *International Conference on 3D Vision (3DV)*, 2019. 1, 3, 5, 6, 7

[4] Sadegh Aliakbarian, Fatemeh Sadat Saleh, Mathieu Salzmann, Lars Petersson, and Stephen Gould. A stochastic conditioning scheme for diverse human motion prediction. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[5] Tenglong Ao, Qingzhe Gao, Yuke Lou, Baoquan Chen, and Libin Liu. Rhythmic gesticulator: Rhythm-aware co-speech gesture synthesis with hierarchical neural embeddings. In *SIGGRAPH Asia*, 2022. 2

[6] Andreas Aristidou, Anastasios Yiannakidis, Kfir Aberman, Daniel Cohen-Or, Ariel Shamir, and Yiorgos Chrysanthou. Rhythm is a dancer: Music-driven motion synthesis with global structure. *arXiv*, 2021. 2

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

[8] Norman I Badler, Cary B Phillips, and Bonnie Lynn Webber. *Simulating humans: computer graphics animation and control*. Oxford University Press, 1993. 2

[9] Emad Barsoum, John Kender, and Zicheng Liu. Hp-gan: Probabilistic 3d human motion prediction via gan. In *Proceedings of the Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2018. 2

[10] 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 *Virtual Reality and 3D User Interfaces (VR)*, 2021. 5, 6, 7

[11] Arij Bouazizi, Adrian Holzbock, Ulrich Kressel, Klaus Dietmayer, and Vasileios Belagiannis. Motionmixer: Mlp-based 3d human body pose forecasting. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2022. 2

[12] Judith Butepage, Michael J Black, Danica Kragic, and Hedvig Kjellstrom. Deep representation learning for human motion prediction and classification. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[13] Kang Chen, Zhipeng Tan, Jin Lei, Song-Hai Zhang, Yuan-Chen Guo, Weidong Zhang, and Shi-Min Hu. Choreomaster: choreography-oriented music-driven dance synthesis. *ACM Transactions on Graphics (TOG)*, 2021. 2

[14] Katherine Crowson, Stella Biderman, Daniel Kornis, Dashiell Stander, Eric Hallahan, Louis Castricato, and Edward Raff. Vqgan-clip: Open domain image generation and editing with natural language guidance. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 4

[15] Prafulla Dhariwal, Heewoo Jun, Christine Payne, Jong Wook Kim, Alec Radford, and Ilya Sutskever. Jukebox: A generative model for music. *arXiv*, 2020. 2

[16] Sander Dieleman, Aaron van den Oord, and Karen Simonyan. The challenge of realistic music generation: modelling raw audio at scale. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. 2

[17] Yinglin Duan, Tianyang Shi, Zhengxia Zou, Yenan Lin, Zhehui Qian, Bohan Zhang, and Yi Yuan. Single-shot motion completion with transformer. *arXiv*, 2021. 2

[18] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2, 4

[19] Katerina Fragkiadaki, Sergey Levine, Panna Felsen, and Jitendra Malik. Recurrent network models for human dynamics. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2015. 2

[20] Anindita Ghosh, Noshaba Cheema, Cennet Oguz, Christian Theobalt, and Philipp Slusallek. Synthesis of compositional animations from textual descriptions. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021. 1, 3, 5, 6, 7

[21] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 2, 3, 4, 5, 6, 7, 8, 12, 13

[22] Chuan Guo, Xinxin Zuo, Sen Wang, and Li Cheng. Tm2t: Stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 1, 3, 5, 6, 7

[23] Chuan Guo, Xinxin Zuo, Sen Wang, Shihao Zou, Qingyao Sun, Annan Deng, Minglun Gong, and Li Cheng. Action2motion: Conditioned generation of 3d human motions. In *Proceedings of the ACM International Conference on Multimedia (ACMMM)*, 2020. 5

[24] Wen Guo, Yuming Du, Xi Shen, Vincent Lepetit, Xavier Alameda-Pineda, and Francesc Moreno-Noguer. Back to mlp: A simple baseline for human motion prediction. In *Proceedings of the Winter Conference on Applications of Computer Vision (WACV)*, 2022. 2

[25] Ikhsanul Habibie, Daniel Holden, Jonathan Schwarz, Joe Yearsley, and Taku Komura. A recurrent variational autoencoder for human motion synthesis. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2017. 2

[26] Félix G Harvey and Christopher Pal. Recurrent transition networks for character locomotion. In *SIGGRAPH Asia 2018 Technical Briefs*, 2018. 2

[27] Félix G Harvey, Mike Yurick, Derek Nowrouzezahrai, and Christopher Pal. Robust motion in-betweening. *ACM Transactions on Graphics (TOG)*, 2020. 2

[28] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings*

of the Conference on Computer Vision and Pattern Recognition (CVPR), 2016. 4

[29] Alejandro Hernandez, Jurgen Gall, and Francesc Moreno-Noguer. Human motion prediction via spatio-temporal inpainting. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[30] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 6

[31] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 2

[32] Daniel Holden, Taku Komura, and Jun Saito. Phase-functioned neural networks for character control. *ACM Transactions on Graphics (TOG)*, 2017. 3

[33] Daniel Holden, Jun Saito, and Taku Komura. A deep learning framework for character motion synthesis and editing. *ACM Transactions on Graphics (TOG)*, 2016. 3

[34] Manuel Kaufmann, Emre Aksan, Jie Song, Fabrizio Pece, Remo Ziegler, and Otmar Hilliges. Convolutional autoencoders for human motion infilling. In *International Conference on 3D Vision (3DV)*, 2020. 2

[35] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations (ICLR)*, 2014. 2

[36] Hsin-Ying Lee, Xiaodong Yang, Ming-Yu Liu, Ting-Chun Wang, Yu-Ding Lu, Ming-Hsuan Yang, and Jan Kautz. Dancing to music. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 2, 5, 6, 7

[37] Jiaman Li, Yihang Yin, Hang Chu, Yi Zhou, Tingwu Wang, Sanja Fidler, and Hao Li. Learning to generate diverse dance motions with transformer. *arXiv*, 2020. 2

[38] Peizhuo Li, Kfir Aberman, Zihan Zhang, Rana Hanocka, and Olga Sorkine-Hornung. Ganimator: Neural motion synthesis from a single sequence. *ACM Transactions on Graphics (TOG)*, 2022. 3

[39] Ruilong Li, Shan Yang, David A Ross, and Angjoo Kanazawa. Ai choreographer: Music conditioned 3d dance generation with aist++. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[40] Siyao Li, Weijiang Yu, Tianpei Gu, Chunze Lin, Quan Wang, Chen Qian, Chen Change Loy, and Ziwei Liu. Bailando: 3d dance generation by actor-critic gpt with choreographic memory. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 4

[41] Angela S Lin, Lemeng Wu, Rodolfo Corona, Kevin Tai, Qixing Huang, and Raymond J Mooney. Generating animated videos of human activities from natural language descriptions. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. 5, 6, 7

[42] 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)*, 2015. 6

[43] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019. 6

[44] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2019. 5

[45] Christian Mandery, Ömer Terlemez, Martin Do, Nikolaus Vahrenkamp, and Tamim Asfour. The kit whole-body human motion database. In *International Conference on Robotics and Automation (ICRA)*, 2015. 5

[46] Wei Mao, Miaomiao Liu, and Mathieu Salzmann. History repeats itself: Human motion prediction via motion attention. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 2

[47] Wei Mao, Miaomiao Liu, Mathieu Salzmann, and Hongdong Li. Learning trajectory dependencies for human motion prediction. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2019. 2

[48] Julieta Martinez, Michael J Black, and Javier Romero. On human motion prediction using recurrent neural networks. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[49] Dario Pavllo, David Grangier, and Michael Auli. Quaternet: A quaternion-based recurrent model for human motion. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2018. 2

[50] Mathis Petrovich, Michael J. Black, and Gül Varol. Action-conditioned 3D human motion synthesis with transformer VAE. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021. 1, 3

[51] Mathis Petrovich, Michael J. Black, and Gül Varol. TEMOS: Generating diverse human motions from textual descriptions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 1, 3

[52] Matthias Plappert, Christian Mandery, and Tamim Asfour. The kit motion-language dataset. *Big data*, 2016. 2, 3, 5, 6, 7, 8, 13

[53] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021. 1, 3, 4

[54] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. 2, 4

[55] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv*, 2022. 4

[56] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning (ICML)*, 2021. 2

[57] Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. In *Advances*

in Neural Information Processing Systems (NeurIPS), 2019. 4

[58] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[59] Tamar Rott Shaham, Tali Dekel, and Tomer Michaeli. Singan: Learning a generative model from a single natural image. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3

[60] Sebastian Starke, Ian Mason, and Taku Komura. Deepphase: periodic autoencoders for learning motion phase manifolds. *ACM Transactions on Graphics (TOG)*, 2022. 3

[61] Sebastian Starke, He Zhang, Taku Komura, and Jun Saito. Neural state machine for character-scene interactions. *ACM Transactions on Graphics (TOG)*, 2019. 3

[62] Xiangjun Tang, He Wang, Bo Hu, Xu Gong, Ruifan Yi, Qilong Kou, and Xiaogang Jin. Real-time controllable motion transition for characters. *ACM Transactions on Graphics (TOG)*, 2022. 2

[63] Guy Tevet, Brian Gordon, Amir Hertz, Amit H Bermano, and Daniel Cohen-Or. Motionclip: Exposing human motion generation to clip space. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 1, 3, 4

[64] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Amit H Bermano, and Daniel Cohen-Or. Human motion diffusion model. *arXiv*, 2022. 2, 3, 5, 6, 7, 8

[65] Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. Mocogan: Decomposing motion and content for video generation. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 5, 6, 7

[66] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 2, 3, 4

[67] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 2, 4, 6

[68] Will Williams, Sam Ringer, Tom Ash, David MacLeod, Jamie Dougherty, and John Hughes. Hierarchical quantized autoencoders. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 2

[69] Sijie Yan, Zhizhong Li, Yuanjun Xiong, Huahan Yan, and Dahua Lin. Convolutional sequence generation for skeleton-based action synthesis. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2019. 2

[70] Xinchen Yan, Akash Rastogi, Ruben Villegas, Kalyan Sunkavalli, Eli Shechtman, Sunil Hadap, Ersin Yumer, and Honglak Lee. Mt-vae: Learning motion transformations to generate multimodal human dynamics. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 2

[71] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *arXiv*, 2022. 2, 3, 5, 6, 7, 8

Num. layers	Num. dim	Num. heads	FID ↓	Top-1 ↑	Training time (hours).
4	512	8	$0.469^{\pm.014}$	$0.469^{\pm.002}$	17
8	512	8	$0.339^{\pm.010}$	$0.481^{\pm.002}$	23
8	768	8	$0.338^{\pm.009}$	$0.490^{\pm.003}$	30
8	768	12	$0.296^{\pm.009}$	$0.484^{\pm.002}$	31
12	768	12	$0.273^{\pm.007}$	$0.487^{\pm.002}$	40
12	1024	16	$0.149^{\pm.007}$	$0.489^{\pm.002}$	55
16	768	12	$0.145^{\pm.006}$	$0.486^{\pm.003}$	47
16	1024	16	$0.143^{\pm.007}$	$0.490^{\pm.004}$	59
18	768	12	$\mathbf{0.130}^{\pm.006}$	$0.483^{\pm.003}$	51
18	1024	16	$0.141^{\pm.005}$	$\mathbf{0.492}^{\pm.003}$	78

Table 4. **Ablation study of T2M-GPT architecture on HumanML3D [21] test set.** For all the architectures, we use the same motion VQ-VAE. The T2M-GPT is trained with $\tau \in \mathcal{U}[0, 1]$. The training time is evaluated on a single Tesla V100-32G GPU.

Appendix

In this appendix, we present:

- Section A: ablation study of T2M-GPT architecture.
- Section B: ablation study of the reconstruction loss ($\mathcal{L}_{re}$ in Equation 1) for motion VQ-VAE.
- Section C: ablation study of τ for the corruption strategy in T2M-GPT training.
- Section D: ablation study of the number of codes in VQ-VAE.
- Section E: more details on the evaluation metrics and the motion representations.
- Section F: the detail of the Motion VQ-VAE architecture.

A. Ablation study of T2M-GPT architecture

In this section, we present results with different transformer architectures for T2M-GPT. The results are provided in Table 4. We notice that better performance can be obtained with a larger architecture. We finally leverage an 18-layer transformer with 16 heads and 1,024 dimensions.

B. Impact of the reconstruction loss in motion VQ-VAE

In this section, we study the effect of the reconstruction loss ($\mathcal{L}_{re}$ in Equation 1) and the hyper-parameter α (Equation 3). The results are presented in Table 5. We find that L1 Smooth achieves the best performance on reconstruction, and the performance of L1 loss is close to L1 Smooth loss. For the hyper-parameter α, we find that $\alpha = 0.5$ leads to the best performance.

$\mathcal{L}_{cons}$	α	Reconstruction	
		FID ↓	Top-1 (%)
L1	0	$0.095^{\pm.001}$	$0.493^{\pm.002}$
L1	0.5	$0.144^{\pm.001}$	$0.495^{\pm.003}$
L1	1	$0.160^{\pm.001}$	$0.496^{\pm.003}$
L1Smooth	0	$0.112^{\pm.001}$	$0.496^{\pm.003}$
L1Smooth	0.5	$\mathbf{0.070}^{\pm.001}$	$\mathbf{0.501}^{\pm.002}$
L1Smooth	1	$0.128^{\pm.001}$	$0.499^{\pm.003}$
L2	0	$0.321^{\pm.002}$	$0.478^{\pm.003}$
L2	0.5	$0.292^{\pm.002}$	$0.483^{\pm.002}$
L2	1	$0.213^{\pm.002}$	$0.490^{\pm.003}$

Table 5. **Ablation of losses for VQ-VAE on HumanML3D [21] test set.** We report FID and Top1 metric for the models trained 300K iterations.

C. Impact of τ for the corruption strategy in T2M-GPT training

τ	FID ↓	Top-1 ↑	MM-Dist ↓
0.0	$0.140^{\pm.006}$	$0.417^{\pm.003}$	$3.730^{\pm.009}$
0.1	$0.131^{\pm.005}$	$0.453^{\pm.002}$	$3.357^{\pm.007}$
0.3	$0.147^{\pm.006}$	$0.485^{\pm.002}$	$3.157^{\pm.007}$
0.5	$\mathbf{0.116}^{\pm.004}$	$0.491^{\pm.003}$	$\mathbf{3.118}^{\pm.011}$
0.7	$0.155^{\pm.006}$	$0.480^{\pm.004}$	$3.183^{\pm.011}$
$\mathcal{U}[0, 1]$	$0.141^{\pm.005}$	$\mathbf{0.492}^{\pm.003}$	$3.121^{\pm.009}$

Table 6. **Analysis of τ on HumanML3D [21] test set.**

In this section, we study τ, which is used for corrupting sequences during the training of T2M-GPT. The results are provided in Table 6. We can see that the training with corrupted sequences $\tau = 0.5$ significantly improves over Top-1 accuracy and FID compared to $\tau = 0$. Compared to $\tau \in \mathcal{U}[0, 1]$, $\tau = 0.5$ is probably preferable for HumanML3D [21], as it achieves comparable Top-1 accuracy compared to $\tau \in \mathcal{U}[0, 1]$ but with much better FID.

Num. code	Reconstruction	
	FID ↓	Top-1 (%)
256	$0.145^{\pm.001}$	$0.497^{\pm.002}$
512	$\mathbf{0.070}^{\pm.001}$	$\mathbf{0.501}^{\pm.002}$
1024	$0.090^{\pm.001}$	$0.498^{\pm.003}$

Table 7. **Study on the number of code in codebook on HumanML3D [21] test set.**

Dilation rate	Reconstruction	
	FID ↓	Top-1 (%)
1, 1, 1	$0.145^{\pm.001}$	$0.500^{\pm.003}$
4, 2, 1	$0.138^{\pm.001}$	$\mathbf{0.502}^{\pm.002}$
9, 3, 1	$\mathbf{0.070}^{\pm.001}$	$0.501^{\pm.002}$
16, 4, 1	$57.016^{\pm.084}$	$0.032^{\pm.001}$

Table 8. **Ablation study of different dilation rate in VQ-VAE on HumanML3D [21] test set.**

D. Ablation study of the number of codes in VQ-VAE

We investigate the number of codes in the codebook in Table 7. We find that the performance of 512 codes is slightly better than 1,024 codes. The results show that 256 codes are not sufficient for reconstruction.

E. More details on the evaluation metrics and the motion representations.

E.1. Evaluation metrics

We detail the calculation of several evaluation metrics, which are proposed in [21]. We denote ground-truth motion features, generated motion features, and text features as f_{gt}, f_{pred}, and f_{text}. Note that these features are extracted with pretrained networks in [21].

FID. FID is widely used to evaluate the overall quality of the generation. We obtain FID by

$$\text{FID} = \|\mu_{gt} - \mu_{pred}\|^2 - \text{Tr}(\Sigma_{gt} + \Sigma_{pred} - 2(\Sigma_{gt}\Sigma_{pred})^{\frac{1}{2}}) \tag{6}$$

where μ_{gt} and μ_{pred} are mean of f_{gt} and f_{pred}. Σ is the covariance matrix and Tr denotes the trace of a matrix.

MM-Dist. MM-Dist measures the distance between the text embedding and the generated motion feature. Given N randomly generated samples, the MM-Dist measures the feature-level distance between the motion and the text. Precisely, it computes the average Euclidean distances between each text feature and the generated motion feature from this text:

$$\text{MM-Dist} = \frac{1}{N}\sum_{i=1}^{N}\|f_{pred,i} - f_{text,i}\| \tag{7}$$

where $f_{pred,i}$ and $f_{text,i}$ are the features of the i-th text-motion pair.

Diversity. Diversity measures the variance of the whole motion sequences across the dataset. We randomly sample S_{dis} pairs of motion and each pair of motion features is denoted by $f_{pred,i}$ and $f'_{pred,i}$. The diversity can be calculated by

$$\text{Diversity} = \frac{1}{S_{dis}}\sum_{i=1}^{S_{dis}}||f_{pred,i} - f'_{pred,i}|| \tag{8}$$

In our experiments, we set S_{dis} to 300 as [21].

MModality. MModality measures the diversity of human motion generated from the same text description. Precisely, for the i-th text description, we generate motion 30 times and then sample two subsets containing 10 motion. We denote features of the j-th pair of the i-th text description by ($f_{pred,i,j}$, $f'_{pred,i,j}$). The MModality is defined as follows:

$$\text{MModality} = \frac{1}{10N}\sum_{i=1}^{N}\sum_{j=1}^{10}\|f_{pred,i,j} - f'_{pred,i,j}\| \tag{9}$$

E.2. Motion representations

We use the same motion representations as [21]. Each pose is represented by $(\dot{r}^a, \dot{r}^x, \dot{r}^z, r^y, j^p, j^v, j^r, c^f)$, where $\dot{r}^a \in \mathbb{R}$ is the global root angular velocity; $\dot{r}^x \in \mathbb{R}, \dot{r}^z \in \mathbb{R}$ are the global root velocity in the X-Z plan; $j^p \in \mathbb{R}^{3j}, j^v \in \mathbb{R}^{3j}, j^r \in \mathbb{R}^{6j}$ are the local pose positions, velocity and rotation with j the number of joints; $c^f \in \mathbb{R}^4$ is the foot contact features calculated by the heel and toe joint velocity.

F. VQ-VAE Architecture

We illustrate the detailed architecture of VQ-VAE in Table 9. The dimensions of the HumanML3D [21] and KIT-ML [52] datasets feature are 263 and 259 respectively.

Dilation rate. We investigate the impact of different dilation rates of the convolution layers used in VQ-VAE, and the results are presented in Table 8 for reconstruction. We notice that setting the dilation rate as (9, 3, 1) gives the most effective and stable performance.

Components	Architecture
VQ-VAE Encoder	(0): Conv1D(D_{in}, 512, kernel_size=(3,), stride=(1,), padding=(1,)) (1): ReLU() (2): 2 × Sequential((0): Conv1D(512, 512, kernel_size=(4,), stride=(2,), padding=(1,)) (1): Resnet1D((0): ResConv1DBlock((activation1): ReLU() (conv1): Conv1D(512, 512, kernel_size=(3,), stride=(1,), padding=(9,), dilation=(9,)) (activation2): ReLU() (conv2): Conv1D(512, 512, kernel_size=(1,), stride=(1,))) (1): ResConv1DBlock((activation1): ReLU() (conv1): Conv1D(512, 512, kernel_size=(3,), stride=(1,), padding=(3,), dilation=(3,)) (activation2): ReLU() (conv2): Conv1D(512, 512, kernel_size=(1,), stride=(1,))) (2): ResConv1DBlock((activation1): ReLU() (conv1): Conv1D(512, 512, kernel_size=(3,), stride=(1,), padding=(1,)) (activation2): ReLU() (conv2): Conv1D(512, 512, kernel_size=(1,), stride=(1,)))))
Codebook	nn.Parameter((512, 512), requires_grad=False)
VQ-VAE Decoder	(0): 2 × Sequential((0): Conv1D(512, 512, kernel_size=(3,), stride=(1,), padding=(1,)) (1): Resnet1D((0): ResConv1DBlock((activation1): ReLU() (conv1): Conv1D(512, 512, kernel_size=(3,), stride=(1,), padding=(9,), dilation=(9,)) (activation2): ReLU() (conv2): Conv1D(512, 512, kernel_size=(1,), stride=(1,))) (1): ResConv1DBlock((activation1): ReLU() (conv1): Conv1D(512, 512, kernel_size=(3,), stride=(1,), padding=(3,), dilation=(3,)) (activation2): ReLU() (conv2): Conv1D(512, 512, kernel_size=(1,), stride=(1,))) (2): ResConv1DBlock((activation1): ReLU() (conv1): Conv1D(512, 512, kernel_size=(3,), stride=(1,), padding=(1,)) (activation2): ReLU() (conv2): Conv1D(512, 512, kernel_size=(1,), stride=(1,)))) (2): Upsample(scale_factor=2.0, mode=nearest) (3): Conv1D(512, 512, kernel_size=(3,), stride=(1,), padding=(1,)) (1): ReLU() (2): Conv1D(512, D_{in}, kernel_size=(3,), stride=(1,), padding=(1,))

Table 9. **Architecture of our Motion VQ-VAE.**