Generating Human Motion from Textual Descriptions with Discrete Representations
Supplementary Material

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

[*]Equal contribution [†]Corresponding author

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

Project Page: https://mael-zys.github.io/T2M-GPT/

In this supplementary material, we present:

- Section 1: ablation study of T2M-GPT architecture.
- Section 2: ablation study of the reconstruction loss ($\mathcal{L}_{re}$ in Equation [3]) for motion VQ-VAE.
- Section 3: ablation study of τ for the corruption strategy in T2M-GPT training.
- Section 4: ablation study of the number of codes in VQ-VAE.
- Section 5: more details on the evaluation metrics and the motion representations.
- Section 6: the detail of the Motion VQ-VAE architecture.
- Section 7: limitations of our proposed approach.
- Section 8: more funding information.

1. 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 1. 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.

2. 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 [3]) and the hyper-parameter α (Equation [3]). The results are presented in Table 2. 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.

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

In this section, we study τ, which is used for corrupting sequences during the training of T2M-GPT. The results are provided in Table 3. 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 [1], as it achieves comparable Top-1 accuracy compared to $\tau \in \mathcal{U}[0, 1]$ but with much better FID.

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

We investigate the number of codes in the codebook in Table 4. 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.

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

5.1. Evaluation metrics

We detail the calculation of several evaluation metrics, which are proposed in [1]. 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 [1].

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

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

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 1. **Ablation study of T2M-GPT architecture on HumanML3D [1] 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.

$\mathcal{L}_{cons}$	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 2. **Ablation of losses for VQ-VAE on HumanML3D [1] test set.** We report FID and Top1 metric for the models trained 300K iterations.

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 3. **Analysis of τ on HumanML3D [1] test set.**

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 4. **Study on the number of code in codebook on HumanML3D [1] test set.**

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:

$$\text{MM-Dist} = \frac{1}{N}\sum_{i=1}^{N}\|f_{pred,i} - f_{text,i}\| \tag{2}$$

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

$$\text{Diversity} = \frac{1}{S_{dis}}\sum_{i=1}^{S_{dis}}||f_{pred,i} - f'_{pred,i}|| \tag{3}$$

In our experiments, we set S_{dis} to 300 as [1].

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:

$$\text{MModality} = \frac{1}{10N}\sum_{i=1}^{N}\sum_{j=1}^{10}\|f_{pred,i,j} - f'_{pred,i,j}\| \tag{4}$$

5.2. Motion representations

We use the same motion representations as [1]. 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; $j^p \in \mathbb{R}^{3j}, j^v \in$

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 5. **Ablation study of different dilation rate in VQ-VAE on HumanML3D [1] test set.**

$\mathbb{R}^{3j}, j^r \in \mathbb{R}^{6j}$ 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.

6. VQ-VAE Architecture

We illustrate the detailed architecture of VQ-VAE in Table 6. The dimensions of the HumanML3D [1] and KIT-ML [2] 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 5 for reconstruction. We notice that setting the dilation rate as (9, 3, 1) gives the most effective and stable performance.

7. Limitations

Our approach has two limitations: *i)* for excessively long texts, the generated motion might miss some details of the textual description. Note that this typical failure case exists for all competitive approaches. *ii)* some generated motion sequences slightly jitter on the legs and hands movement, this can be seen from the visual results provided in the supplementary material. We think the problem comes from the VQ-VAE architecture, with a better-designed architecture, the problem might be alleviated. For a real application, the jittering problem could be addressed using a temporal smoothing filter as a post-processing step.

8. Funding Support

This work is supported by:

- Natural Science Foundation of China (No. 62176155).
- Natural Science Foundation of Jilin Province (20200201037JC).
- Provincial Science and Technology Innovation Special Fund Project of Jilin Province (20190302026GX)
- Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102)

References

[1] 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

[2] Matthias Plappert, Christian Mandery, and Tamim Asfour. The kit motion-language dataset. *Big data*, 2016. 3

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 6. **Architecture of our Motion VQ-VAE.**