WED-PM-227

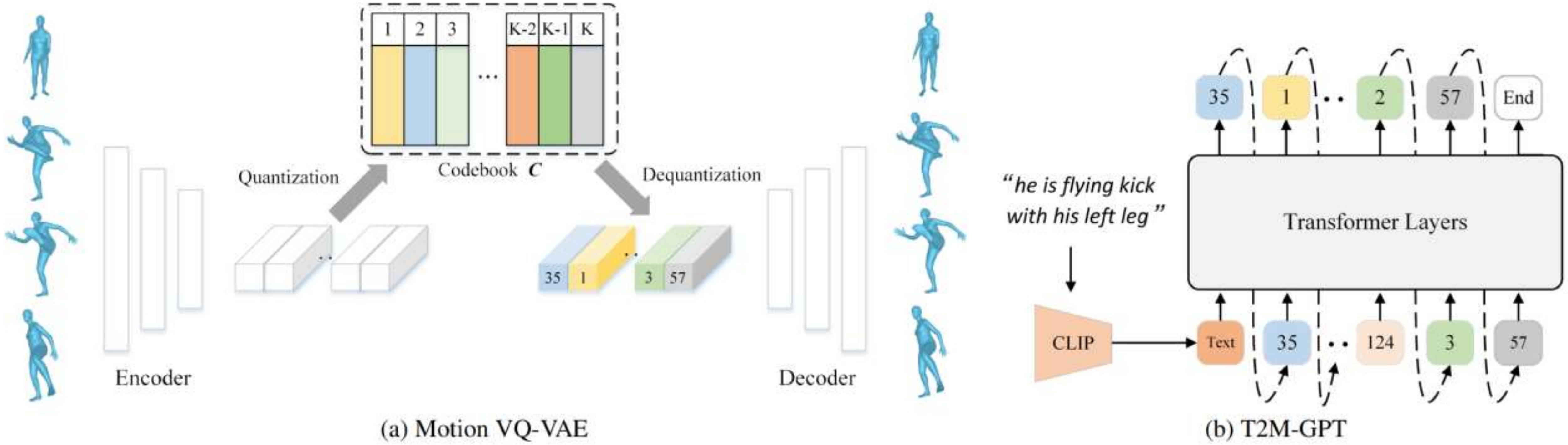
Generating Human Motion from Textual Descriptions with Discrete Representations

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

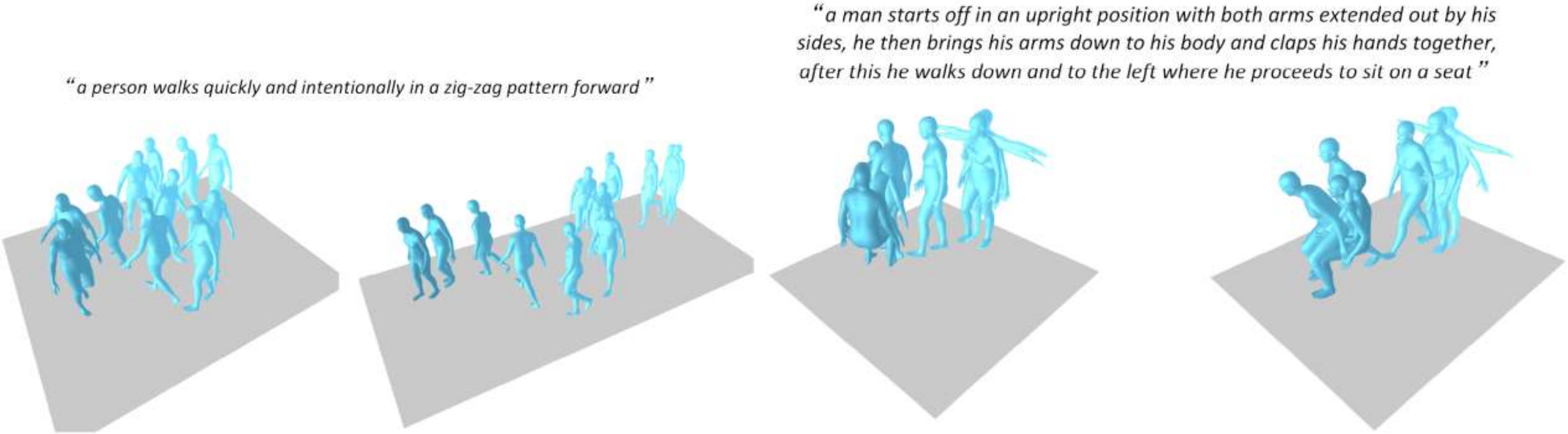
[1] Jilin University [2] Shanghai Jiao Tong University [3] Tencent AI Lab

Overview



1
2
3
K-2
K-1
K
Codebook C
Quantization
Dequantization
35
1
3
57
Encoder
Decoder
(a) Motion VQ-VAE
"he is flying kick with his left leg"
CLIP
Text
35
124
3
57
Transformer Layers
35
1
2
57
End
(b) T2M-GPT

Motivation



Ground-truth　Our generation　Ground-truth　Our generation

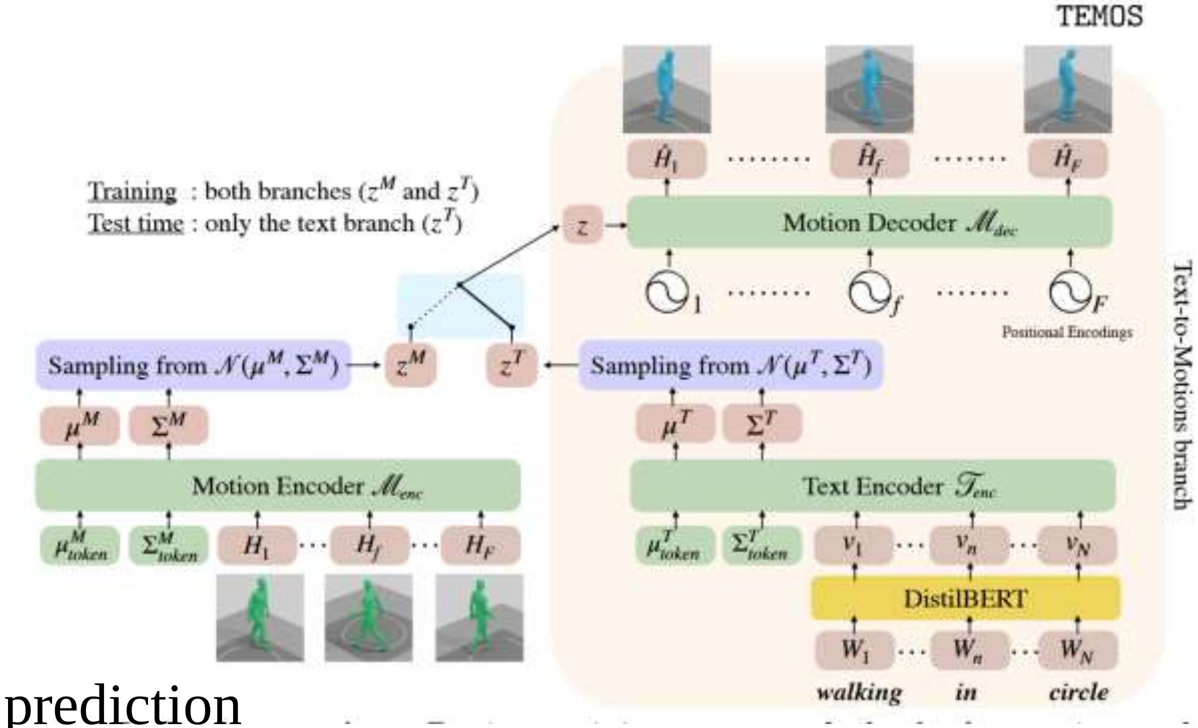
Contributions:

1. We present a simple yet effective approach for motion generation from textual descriptions

2. We show that GPT-like models incorporating discrete representations still remain a very competitive approach for motion generation

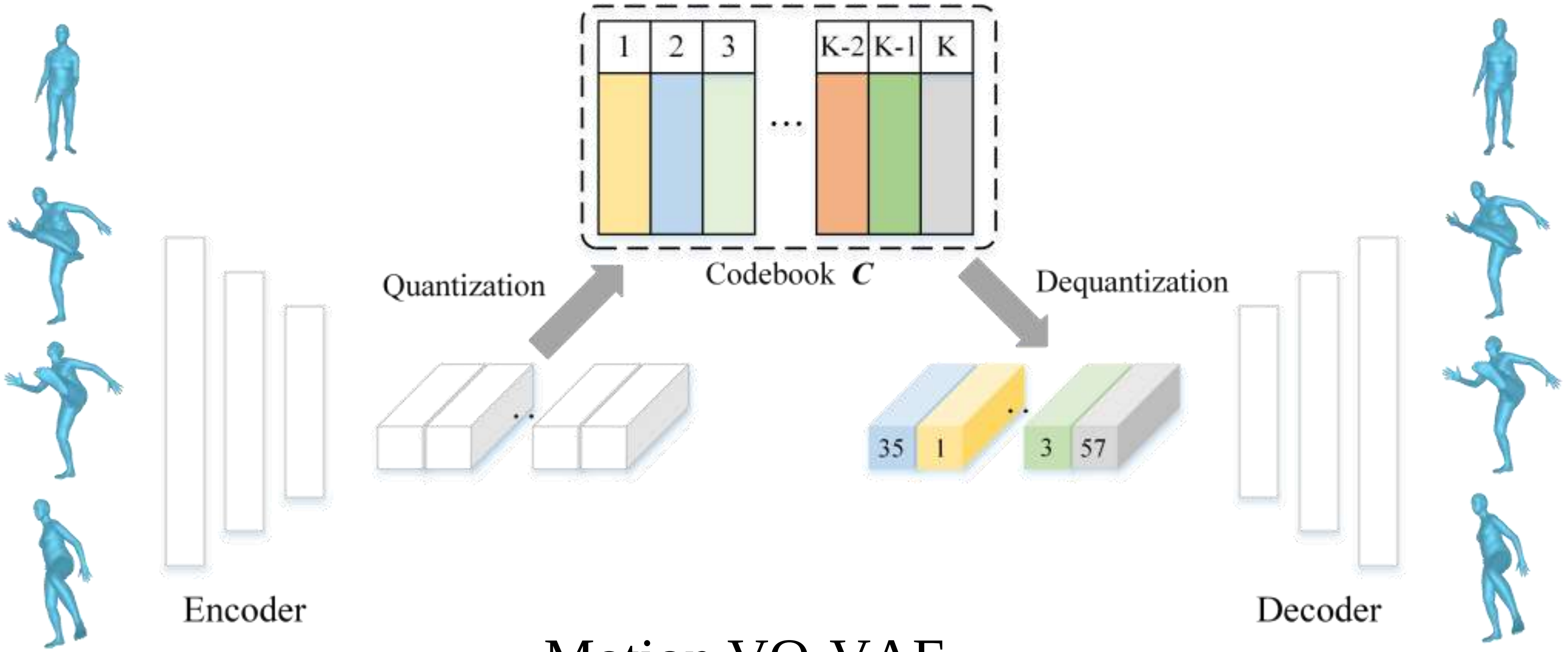
3. We provide a detailed analysis of the impact of quantization strategies and dataset size

Previous work:



1. T2M (Guo et al. 2022, CVPR): motion length prediction
2. TM2T (Guo et al. 2022, ECCV): text-to-motion and motion-to-text tasks
3. TEMOS (Petrovich et al. 2022, ECCV): transformer-based VAE
4. MotionDiffuse (Zhang et al. 2022, Arxiv): diffusion-based models
5. MDM (Tevet et al. 2023, ICLR): diffusion-based models

Methods: stage 1



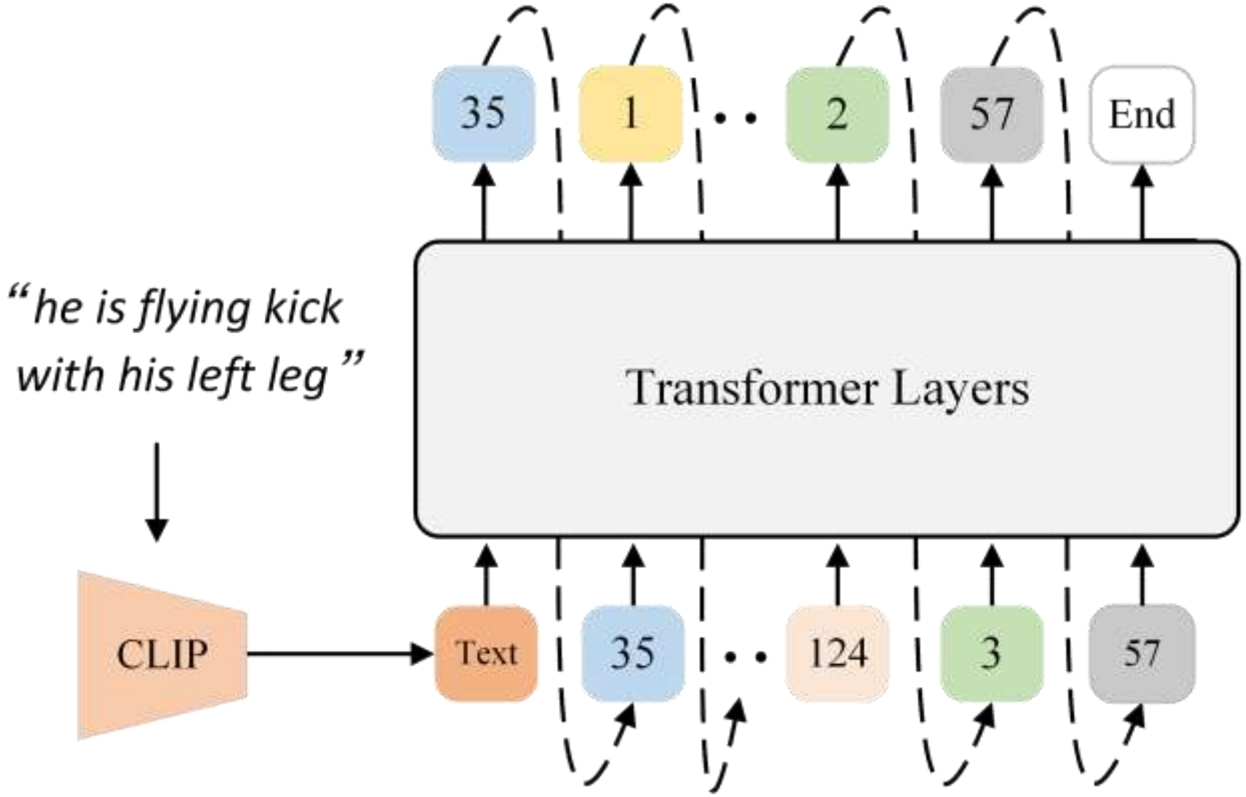
Motion VQ-VAE

Quantization strategy:

EMA: $C^t \leftarrow \lambda C^{t-1} + (1-\lambda)C^t$

Code reset: reassigns inactivate codes

Methods: stage 2



T2M-GPT

Results

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 [42]	$0.180^{\pm.002}$	$0.300^{\pm.002}$	$0.396^{\pm.002}$	$11.75^{\pm.035}$	$5.529^{\pm.007}$	$6.223^{\pm.061}$	-
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 [21]	$0.301^{\pm.002}$	$0.425^{\pm.002}$	$0.552^{\pm.004}$	$6.532^{\pm.024}$	$5.012^{\pm.018}$	$8.332^{\pm.042}$	-
MoCoGAN [67]	$0.037^{\pm.000}$	$0.072^{\pm.001}$	$0.106^{\pm.001}$	$94.41^{\pm.021}$	$9.643^{\pm.006}$	$0.462^{\pm.008}$	$0.019^{\pm.000}$
Dance2Music [37]	$0.033^{\pm.000}$	$0.065^{\pm.001}$	$0.097^{\pm.001}$	$66.98^{\pm.016}$	$8.116^{\pm.006}$	$0.725^{\pm.011}$	$0.043^{\pm.001}$
TEMOS§ [53]	$0.424^{\pm.002}$	$0.612^{\pm.002}$	$0.722^{\pm.002}$	$3.734^{\pm.028}$	$3.703^{\pm.008}$	$8.973^{\pm.071}$	$0.368^{\pm.018}$
TM2T [23]	$0.424^{\pm.003}$	$0.618^{\pm.003}$	$0.729^{\pm.002}$	$1.501^{\pm.017}$	$3.467^{\pm.011}$	$8.589^{\pm.076}$	$2.424^{\pm.093}$
Guo *et al.* [22]	$0.455^{\pm.003}$	$0.636^{\pm.003}$	$0.736^{\pm.002}$	$1.087^{\pm.021}$	$3.347^{\pm.008}$	$9.175^{\pm.083}$	$2.219^{\pm.074}$
MLD§ [71]	$0.481^{\pm.003}$	$0.673^{\pm.003}$	$0.772^{\pm.002}$	$0.473^{\pm.013}$	$3.196^{\pm.010}$	$9.724^{\pm.082}$	$2.413^{\pm.079}$
MDM [66]§	-	-	$0.611^{\pm.007}$	$0.544^{\pm.044}$	$5.566^{\pm.027}$	$9.559^{\pm.086}$	$2.799^{\pm.072}$
MotionDiffuse [74]§	$0.491^{\pm.001}$	$0.681^{\pm.001}$	$0.782^{\pm.001}$	$0.630^{\pm.001}$	$3.113^{\pm.001}$	$9.410^{\pm.049}$	$1.553^{\pm.042}$
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 ($\tau \in \mathcal{U}[0,1]$)	$0.492^{\pm.003}$	$0.679^{\pm.002}$	$0.775^{\pm.002}$	$0.141^{\pm.005}$	$3.121^{\pm.009}$	$9.722^{\pm.082}$	$1.831^{\pm.048}$

Table 1. **Comparison with the state-of-the-art methods on HumanML3D [22] test set.** We compute standard metrics following Guo *et al.* [22]. 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. § reports results using ground-truth motion length.

Results

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 [42]	$0.103^{\pm.003}$	$0.178^{\pm.005}$	$0.241^{\pm.006}$	$24.86^{\pm.348}$	$7.960^{\pm.031}$	$6.744^{\pm.106}$	-
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 [21]	$0.255^{\pm.006}$	$0.432^{\pm.007}$	$0.531^{\pm.007}$	$5.203^{\pm.107}$	$4.986^{\pm.027}$	$9.563^{\pm.072}$	-
MoCoGAN [67]	$0.022^{\pm.002}$	$0.042^{\pm.003}$	$0.063^{\pm.003}$	$82.69^{\pm.242}$	$10.47^{\pm.012}$	$3.091^{\pm.043}$	$0.250^{\pm.009}$
Dance2Music [37]	$0.031^{\pm.002}$	$0.058^{\pm.002}$	$0.086^{\pm.003}$	$115.4^{\pm.240}$	$10.40^{\pm.016}$	$0.241^{\pm.004}$	$0.062^{\pm.002}$
TEMOS§ [53,71]	$0.353^{\pm.002}$	$0.561^{\pm.002}$	$0.687^{\pm.002}$	$3.717^{\pm.028}$	$3.417^{\pm.008}$	$10.84^{\pm.071}$	$0.532^{\pm.018}$
TM2T [23]	$0.280^{\pm.006}$	$0.463^{\pm.007}$	$0.587^{\pm.005}$	$3.599^{\pm.051}$	$4.591^{\pm.019}$	$9.473^{\pm.100}$	$3.292^{\pm.034}$
Guo *et al.* [22]	$0.361^{\pm.006}$	$0.559^{\pm.007}$	$0.681^{\pm.007}$	$3.022^{\pm.107}$	$3.488^{\pm.028}$	$10.72^{\pm.145}$	$2.052^{\pm.107}$
MLD§ [71]	$0.390^{\pm.008}$	$0.609^{\pm.008}$	$0.734^{\pm.007}$	$0.404^{\pm.027}$	$3.204^{\pm.027}$	$10.80^{\pm.117}$	$2.192^{\pm.071}$
MDM [66]§	-	-	$0.396^{\pm.004}$	$0.497^{\pm.021}$	$9.191^{\pm.022}$	$10.847^{\pm.109}$	$1.907^{\pm.214}$
MotionDiffuse [74]§	$0.417^{\pm.004}$	$0.621^{\pm.004}$	$0.739^{\pm.004}$	$1.954^{\pm.062}$	$2.958^{\pm.005}$	$11.10^{\pm.143}$	$0.730^{\pm.013}$
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 ($\tau \in \mathcal{U}[0,1]$)	$0.416^{\pm.006}$	$0.627^{\pm.006}$	$0.745^{\pm.006}$	$0.514^{\pm.029}$	$3.007^{\pm.023}$	$10.921^{\pm.108}$	$1.570^{\pm.039}$

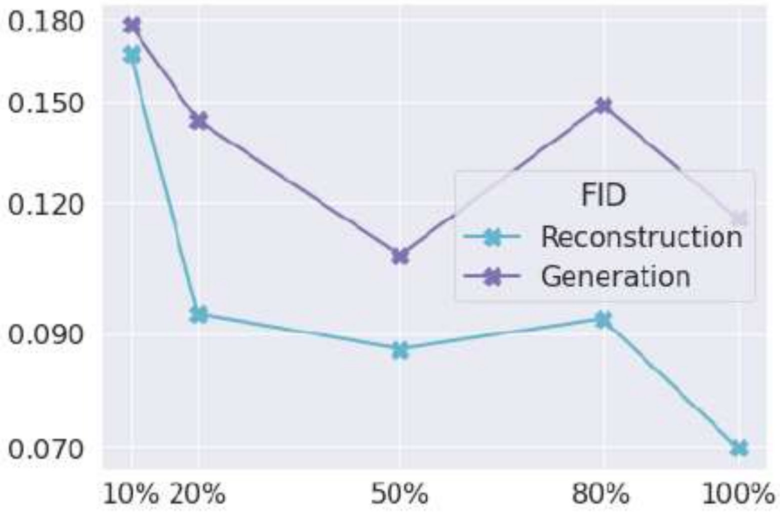
Table 2. **Comparison with the state-of-the-art methods on KIT-ML [54] test set.** We compute standard metrics following Guo *et al.* [22]. For each metric, we repeat the evaluation 20 times and report the average with 95% confidence interval. Red and Blue indicate the best and the second best result. § reports results using ground-truth motion length.

Ablation Study

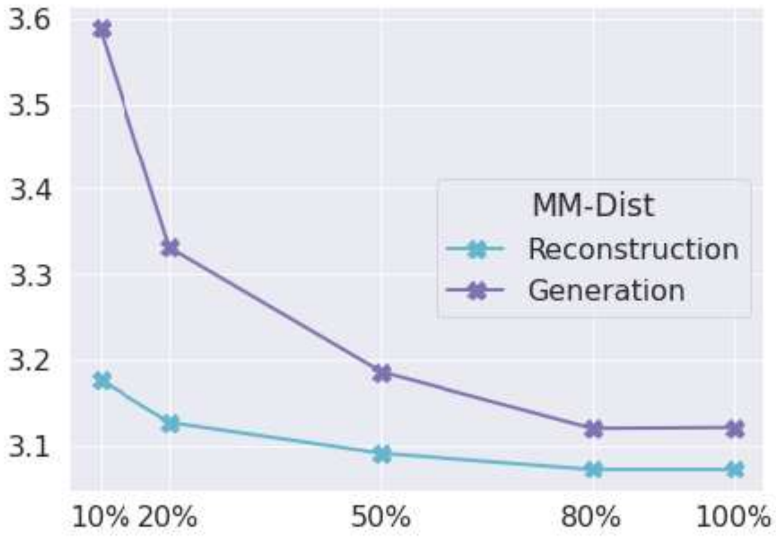
Analysis of VQ-VAE quantizers on HumanML3D

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

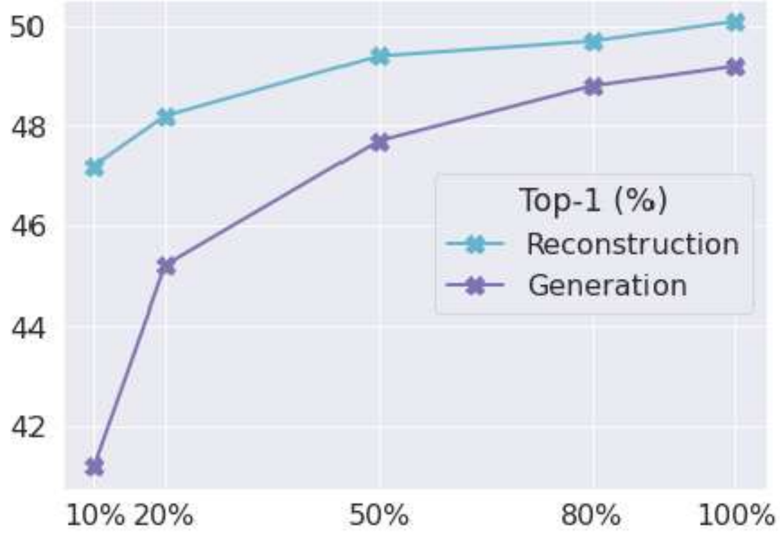
Impact of dataset size on HumanML3D



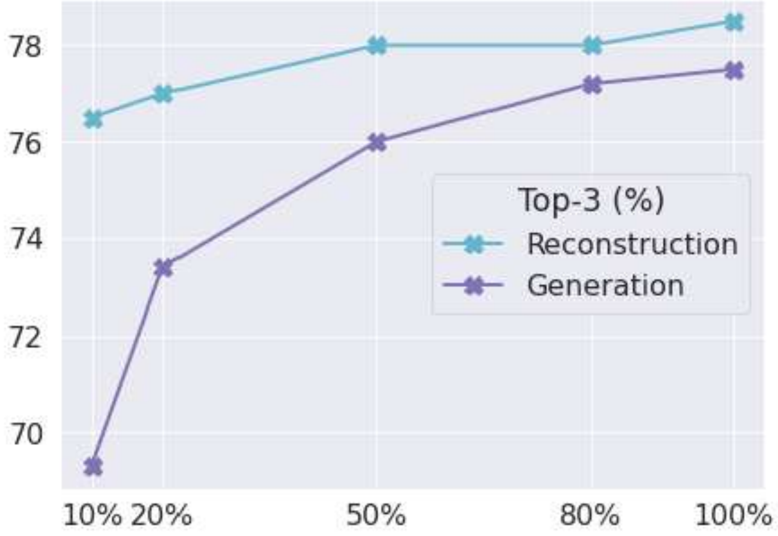
(a) FID



(b) MM-Dist



(c) Top-1 accuracy

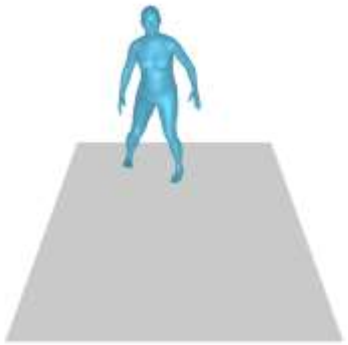


(d) Top-3 accuracy

Visualization

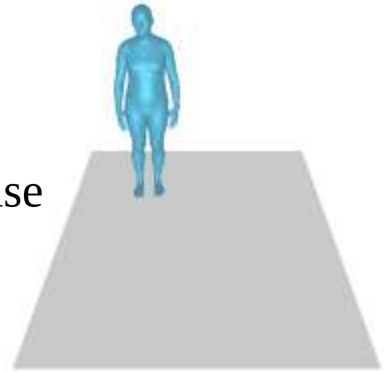
Text: a man steps forward and does a handstand.

T2M

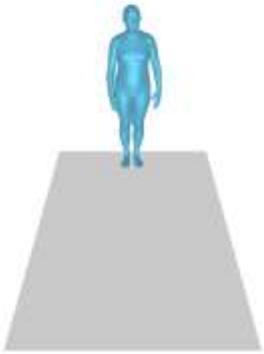


MDM

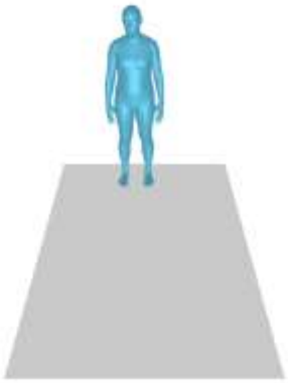
MotionDiffuse



Ground-truth



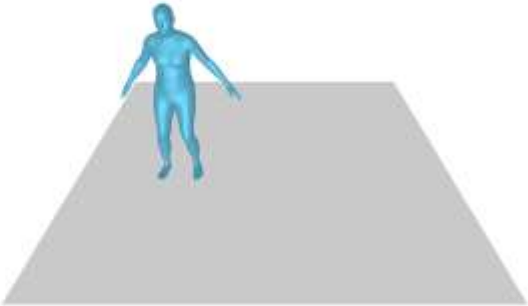
Ours



Visualization

Text: A man rises from the ground, walks in a circle and sits back down on the ground.

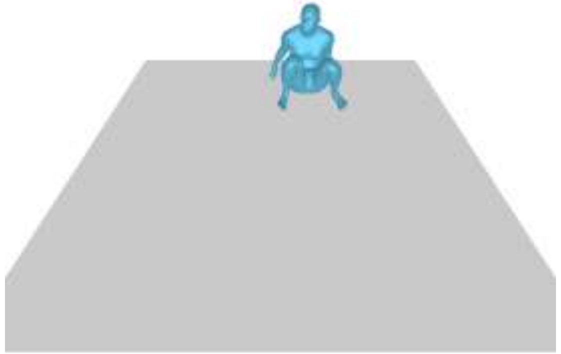
T2M



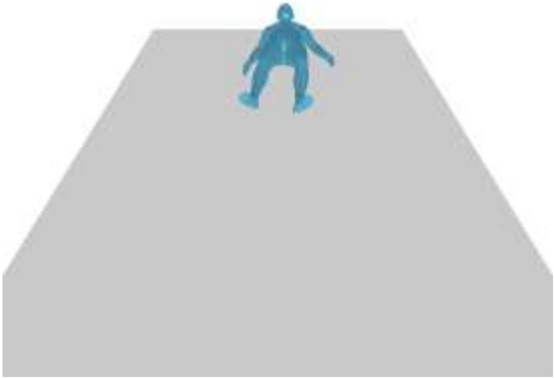
MDM

MotionDiffuse

Ground-truth



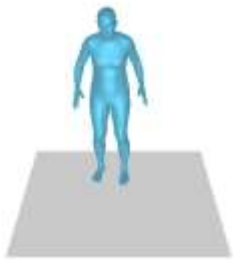
Ours



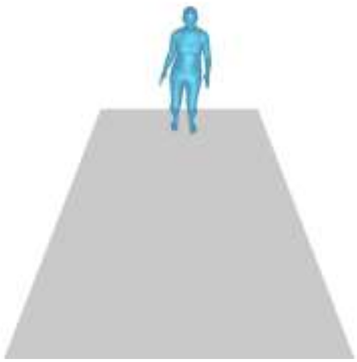
Visualization

Text: a person jogs in place, slowly at first, then increases speed. they then back up and squat down.

T2M

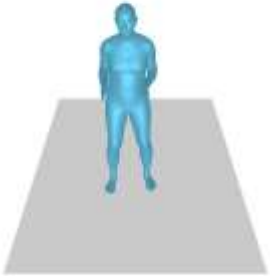


MDM

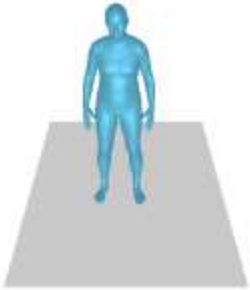


MotionDiffuse

Ground-truth



Ours



Visualization

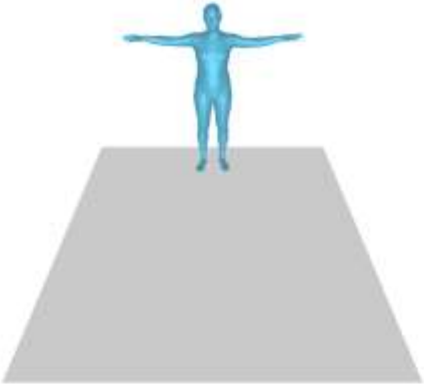
Text: a man starts off in an up right position with botg arms extended out by his sides, he then brings his arms down to his body and claps his hands together. after this he wals down amd the the left where he proceeds to sit on a seat

T2M

MDM

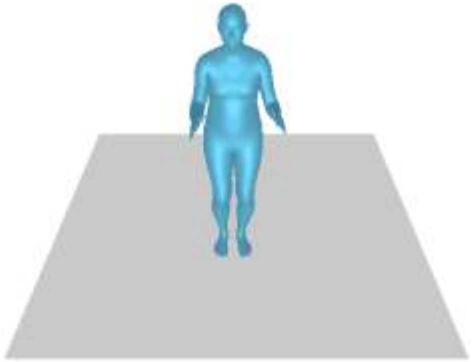
MotionDiffuse

Ground-truth



Ours

More visual results are provided in the project page



Conclusion

- We investigated a classic framework based on VQ-VAE and GPT to synthesize human motion
- Our method achieved comparable or even better performances than concurrent diffusion-based approaches

Thank you!

- Project page: https://mael-zys.github.io/T2M-GPT/
- Code: https://github.com/Mael-zys/T2M-GPT