NEURAL INFORMATION PROCESSING SYSTEMS

Alibaba Group

MoGenTS: Motion Generation based on Spatial-Temporal Joint Modeling

Weihao Yuan, Yisheng He, Weichao Shen, Yuan Dong, Xiaodong Gu, Zilong Dong, Liefeng Bo, Qixing Huang

Alibaba Group
The University of Texas at Austin

Background: Human motion generation

"A person walks in a circle"

Motivation: Continuous or Discrete?

- **Continuous:** Directly regress the continuous human motions from the text inputs

  - Pros: Directly optimizing towards ground-truth data and does not lose the numerical precision

  - Cons: Regressing continuous motion that encompasses complex skeletal joint information and is limited by the quality and scale of current text-to-motion datasets

- **Discrete:** Leverages vector quantization (VQ) to convert continuous motion to discrete tokens

  - Pros: Transform the regression problem into a classification problem, such that the difficulty of motion generation could be greatly reduced.

  - Cons: The VQ process inevitably introduces approximation errors, which impose undesirable limits on the quality of the generated motions.

Motivation: Accuracy of the VQ approximation

Key: improving the accuracy of the VQ approximation

Most previous methods quantize all joints of one frame into one vector and approximate this vector with one code from the codebook.

- ➢ makes the encoding process difficult, as each code within the codebook is tasked with encapsulating the comprehensive information of all joints, making the quantization fundamentally more complex
- ➢ loss of spatial relationships between the individual joints, hence the subsequent network could not capture and aggregate the spatial information.

Motivation: Quantize each joint

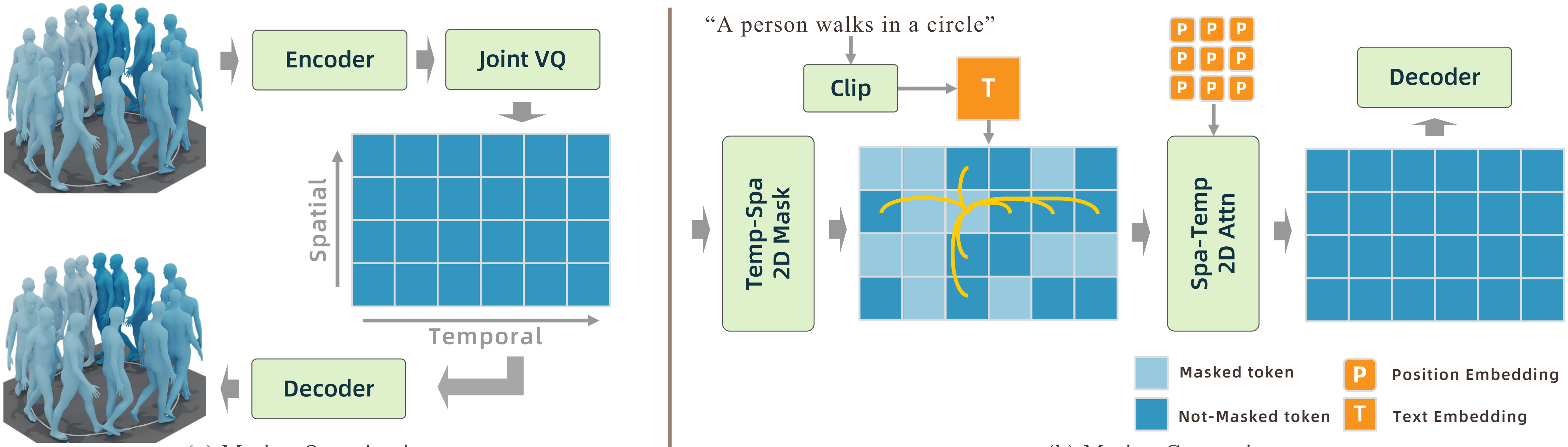
NEURAL INFORMATION PROCESSING SYSTEMS

Quantize each joint rather than the whole-body pose into one vector.

- First, encoding at the joint level significantly simplifies the quantization process, as the complexity associated with representing the information of a single joint is markedly lower than that of the entire pose.
- Second, with each joint encoded separately, the resulting tokens maintain a spatial-temporal distribution that preserves both the spatial relationships among joints and the temporal dynamics of their movements.
- Third, the spatial-temporal distribution of these tokens naturally organizes into a 2D structure, akin to that of 2D images. This similarity enables the application of various 2D operations, such as 2D convolution, 2D positional encoding, and 2D attention mechanisms.

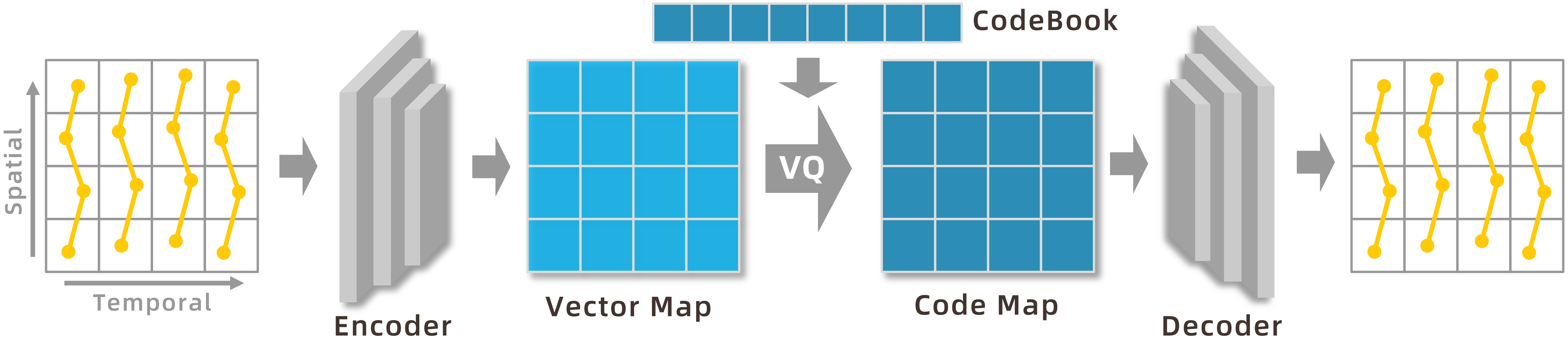
Framework

Starting from the 2D motion quantization, we propose a spatial-temporal modeling framework for human motion generation.



(a) Motion Quantization (b) Motion Generation

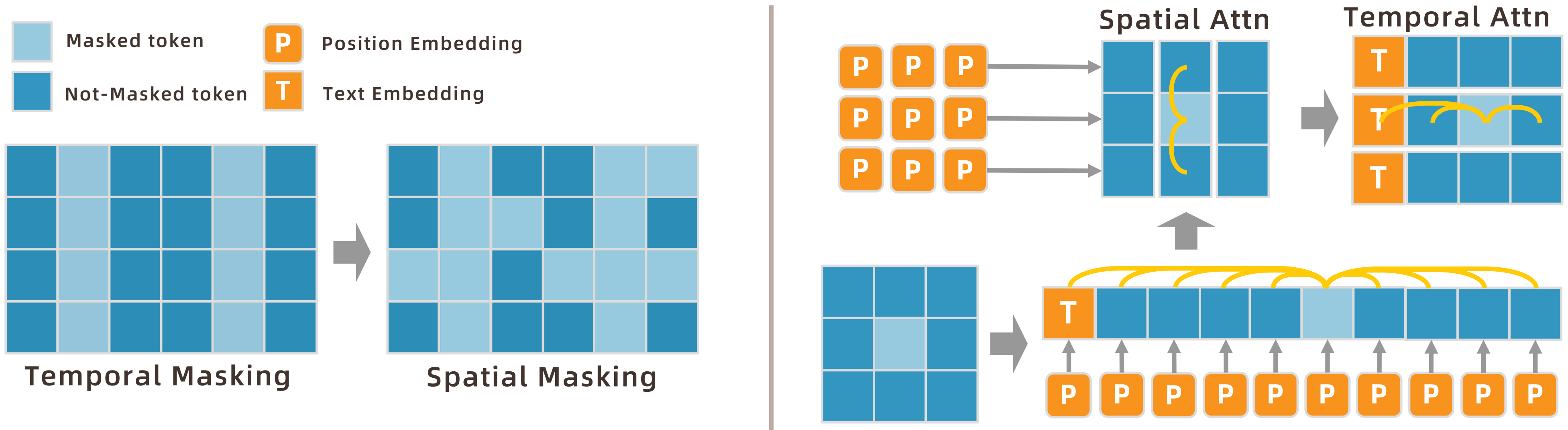
Motion Quantization



Spatial-temporal 2D Joint Quantization of Motion

We first represent the input motion sequence in the joint-time structure, and encode it into a 2D vector map. Subsequently, the vectors are quantized into codes of a codebook, represented by the indices of the selected codebook entries, i.e., the joint tokens. Therefore, after the quantization, the input motion is converted to tokens arranged in a 2D structure, where one dimension is spatial while the other one is temporal. This results in a 2D motion map, which is similar to a 2D image.

Motion Generation



(a) Spatial-temporal Masking

(b) Spatial-temporal Attention

Spatial-temporal 2D Motion Generation

We mask the 2D token map with a temporal-spatial 2D masking strategy, and then use a 2D transformer to predict the masked tokens, conditioned on the embedding T of the given text prompt. The 2D motion transformer considers both spatial attention and temporal attention between different 2D tokens. The 2D position embedding P is also used to convey the spatial and temporal locations of each token.

Experiments

Methods	HumanML3D		KIT-ML	
	FID ↓	MPJPE ↓	FID ↓	MPJPE ↓
TM2T [13]	0.307	230.1	-	-
M2DM [15]	0.063	-	0.413	-
T2M-GPT [14]	0.070	58.0	0.472	-
MoMask [17]	$0.019^{\pm.000}$	$29.5^{\pm.0}$	$0.112^{\pm.002}$	$37.2^{\pm.1}$
Ours	$\mathbf{0.005}^{\pm.000}$	$\mathbf{13.8}^{\pm.0}$	$\mathbf{0.019}^{\pm.001}$	$\mathbf{17.4}^{\pm.1}$

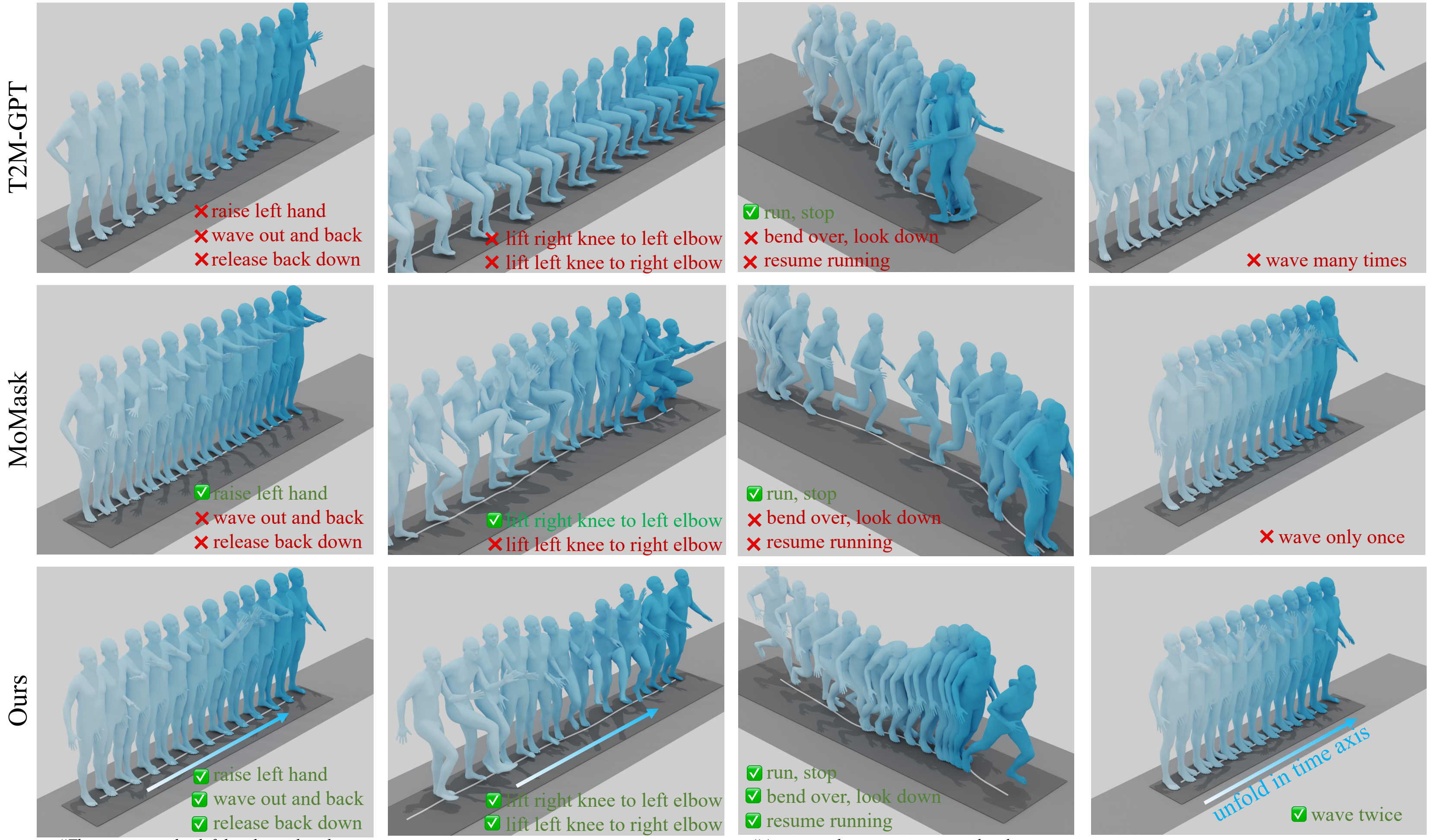
Table 2: Evaluation of motion quantization on HumanML3D dataset and KIT-ML dataset. MPJPE is measured in millimeters.

Methods	FID ↓	Top1 ↑	Top2 ↑	Top3 ↑	MM-Dist ↓	Diversity →
Ground Truth	$0.002^{\pm.000}$	$0.511^{\pm.003}$	$0.703^{\pm.003}$	$0.797^{\pm.002}$	$2.974^{\pm.008}$	$9.503^{\pm.065}$
TEMOS [4]	$3.734^{\pm.028}$	$0.424^{\pm.002}$	$0.612^{\pm.002}$	$0.722^{\pm.002}$	$3.703^{\pm.008}$	$8.973^{\pm.071}$
TM2T [13]	$1.501^{\pm.017}$	$0.424^{\pm.003}$	$0.618^{\pm.003}$	$0.729^{\pm.002}$	$3.467^{\pm.011}$	$8.589^{\pm.076}$
T2M [5]	$1.087^{\pm.021}$	$0.455^{\pm.003}$	$0.636^{\pm.003}$	$0.736^{\pm.002}$	$3.347^{\pm.008}$	$9.175^{\pm.083}$
MDM [7]	$0.544^{\pm.044}$	-	-	$0.611^{\pm.007}$	$5.566^{\pm.027}$	$9.559^{\pm.086}$
MLD [6]	$0.473^{\pm.013}$	$0.481^{\pm.003}$	$0.673^{\pm.003}$	$0.772^{\pm.002}$	$3.196^{\pm.010}$	$9.724^{\pm.082}$
MotionDiffuse [8]	$0.630^{\pm.001}$	$0.491^{\pm.001}$	$0.681^{\pm.001}$	$0.782^{\pm.001}$	$3.113^{\pm.001}$	$9.410^{\pm.049}$
PhysDiff [37]	0.433	-	-	0.631	-	-
MotionGPT [12]	0.567	-	-	-	3.775	9.006
T2M-GPT [14]	$0.141^{\pm.005}$	$0.492^{\pm.003}$	$0.679^{\pm.002}$	$0.775^{\pm.002}$	$3.121^{\pm.009}$	$9.761^{\pm.081}$
M2DM [15]	$0.352^{\pm.005}$	$0.497^{\pm.003}$	$0.682^{\pm.002}$	$0.763^{\pm.003}$	$3.134^{\pm.010}$	$9.926^{\pm.073}$
Fg-T2M [38]	$0.243^{\pm.019}$	$0.492^{\pm.002}$	$0.683^{\pm.003}$	$0.783^{\pm.002}$	$3.109^{\pm.007}$	$9.278^{\pm.072}$
AttT2M [16]	$0.112^{\pm.006}$	$0.499^{\pm.003}$	$0.690^{\pm.002}$	$0.786^{\pm.002}$	$3.038^{\pm.007}$	$9.700^{\pm.090}$
DiverseMotion [41]	$0.072^{\pm.004}$	$0.515^{\pm.003}$	$0.706^{\pm.002}$	$0.802^{\pm.002}$	$2.941^{\pm.007}$	$9.683^{\pm.102}$
ParCo [40]	$0.109^{\pm.005}$	$0.515^{\pm.003}$	$0.706^{\pm.003}$	$0.801^{\pm.002}$	$2.927^{\pm.008}$	$9.576^{\pm.088}$
MMM [21]	$0.080^{\pm.003}$	$0.504^{\pm.003}$	$0.696^{\pm.003}$	$0.794^{\pm.002}$	$2.998^{\pm.007}$	$9.411^{\pm.058}$
MoMask [17]	$0.045^{\pm.002}$	$0.521^{\pm.002}$	$0.713^{\pm.002}$	$0.807^{\pm.002}$	$2.958^{\pm.008}$	-
Ours	$\mathbf{0.033}^{\pm.001}$	$\mathbf{0.529}^{\pm.003}$	$\mathbf{0.719}^{\pm.002}$	$\mathbf{0.812}^{\pm.002}$	$\mathbf{2.867}^{\pm.006}$	$9.570^{\pm.077}$
Ground Truth	$0.031^{\pm.004}$	$0.424^{\pm.005}$	$0.649^{\pm.006}$	$0.779^{\pm.006}$	$2.788^{\pm.012}$	$11.080^{\pm.097}$
TEMOS [4]	$3.717^{\pm.028}$	$0.353^{\pm.002}$	$0.561^{\pm.002}$	$0.687^{\pm.002}$	$3.417^{\pm.008}$	$10.84^{\pm.100}$
TM2T [13]	$3.599^{\pm.153}$	$0.280^{\pm.005}$	$0.463^{\pm.006}$	$0.587^{\pm.005}$	$4.591^{\pm.026}$	$9.473^{\pm.117}$
T2M [5]	$3.022^{\pm.107}$	$0.361^{\pm.005}$	$0.559^{\pm.007}$	$0.681^{\pm.007}$	$3.488^{\pm.028}$	$10.72^{\pm.145}$
MDM [7]	$0.497^{\pm.021}$	-	-	$0.396^{\pm.004}$	$9.191^{\pm.022}$	$10.85^{\pm.109}$
MLD [6]	$0.404^{\pm.027}$	$0.390^{\pm.008}$	$0.609^{\pm.008}$	$0.734^{\pm.007}$	$3.204^{\pm.027}$	$10.80^{\pm.117}$
MotionDiffuse [8]	$1.954^{\pm.062}$	$0.417^{\pm.004}$	$0.621^{\pm.004}$	$0.739^{\pm.004}$	$2.958^{\pm.005}$	$11.10^{\pm.143}$
MotionGPT [12]	0.597	-	-	-	3.394	10.54
T2M-GPT [14]	$0.514^{\pm.029}$	$0.416^{\pm.006}$	$0.627^{\pm.006}$	$0.745^{\pm.006}$	$3.007^{\pm.023}$	$10.86^{\pm.094}$
M2DM [15]	$0.515^{\pm.029}$	$0.416^{\pm.004}$	$0.628^{\pm.004}$	$0.743^{\pm.004}$	$3.015^{\pm.017}$	$11.417^{\pm.097}$
Fg-T2M [38]	$0.571^{\pm.047}$	$0.418^{\pm.005}$	$0.626^{\pm.004}$	$0.745^{\pm.004}$	$3.114^{\pm.015}$	$10.93^{\pm.083}$
AttT2M [16]	$0.870^{\pm.039}$	$0.413^{\pm.006}$	$0.632^{\pm.006}$	$0.751^{\pm.006}$	$3.039^{\pm.021}$	$10.96^{\pm.123}$
DiverseMotion [41]	$0.468^{\pm.098}$	$0.416^{\pm.005}$	$0.637^{\pm.008}$	$0.760^{\pm.011}$	$2.892^{\pm.041}$	$10.873^{\pm.101}$
ParCo [40]	$0.453^{\pm.027}$	$0.430^{\pm.004}$	$0.649^{\pm.007}$	$0.772^{\pm.006}$	$2.820^{\pm.028}$	$10.95^{\pm.094}$
MMM [21]	$0.429^{\pm.019}$	$0.381^{\pm.005}$	$0.590^{\pm.006}$	$0.718^{\pm.005}$	$3.146^{\pm.019}$	$10.633^{\pm.097}$
MoMask [17]	$0.204^{\pm.011}$	$0.433^{\pm.007}$	$0.656^{\pm.005}$	$0.781^{\pm.005}$	$2.779^{\pm.022}$	-
Ours	$\mathbf{0.143}^{\pm.004}$	$\mathbf{0.445}^{\pm.006}$	$\mathbf{0.671}^{\pm.006}$	$\mathbf{0.797}^{\pm.005}$	$\mathbf{2.711}^{\pm.024}$	$10.918^{\pm.090}$

Table 1: Evaluation on the HumanML3D dataset (upper half) and the KIT-ML dataset (lower half).

Experiments

NEURAL INFORMATION PROCESSING SYSTEMS



"The man raises his left hand over his chest, waves out and back in toward his body, then releases it back down toward his hip."

"A person lifts each knee towards the opposite elbow."

"A person who is running, stops, bends over and looks down while taking small steps, then resumes running."

"A person slowly waves with their right hand twice."

Experiments: Video

NEURAL INFORMATION PROCESSING SYSTEMS

Motion Generation based on Spatial-Temporal Joint Modeling

Supplementary Video

Thanks

NEURAL INFORMATION PROCESSING SYSTEMS

THANKS FOR YOUR WATCHING

Have a nice day.

Paper link: https://arxiv.org/abs/2409.17686
Project: https://aigc3d.github.io/mogents/
Code: https://github.com/weihaosky/mogents

Project QR

About Me QR