Move as You Say, Interact as You Can:
Language-guided Human Motion Generation with Scene Affordance

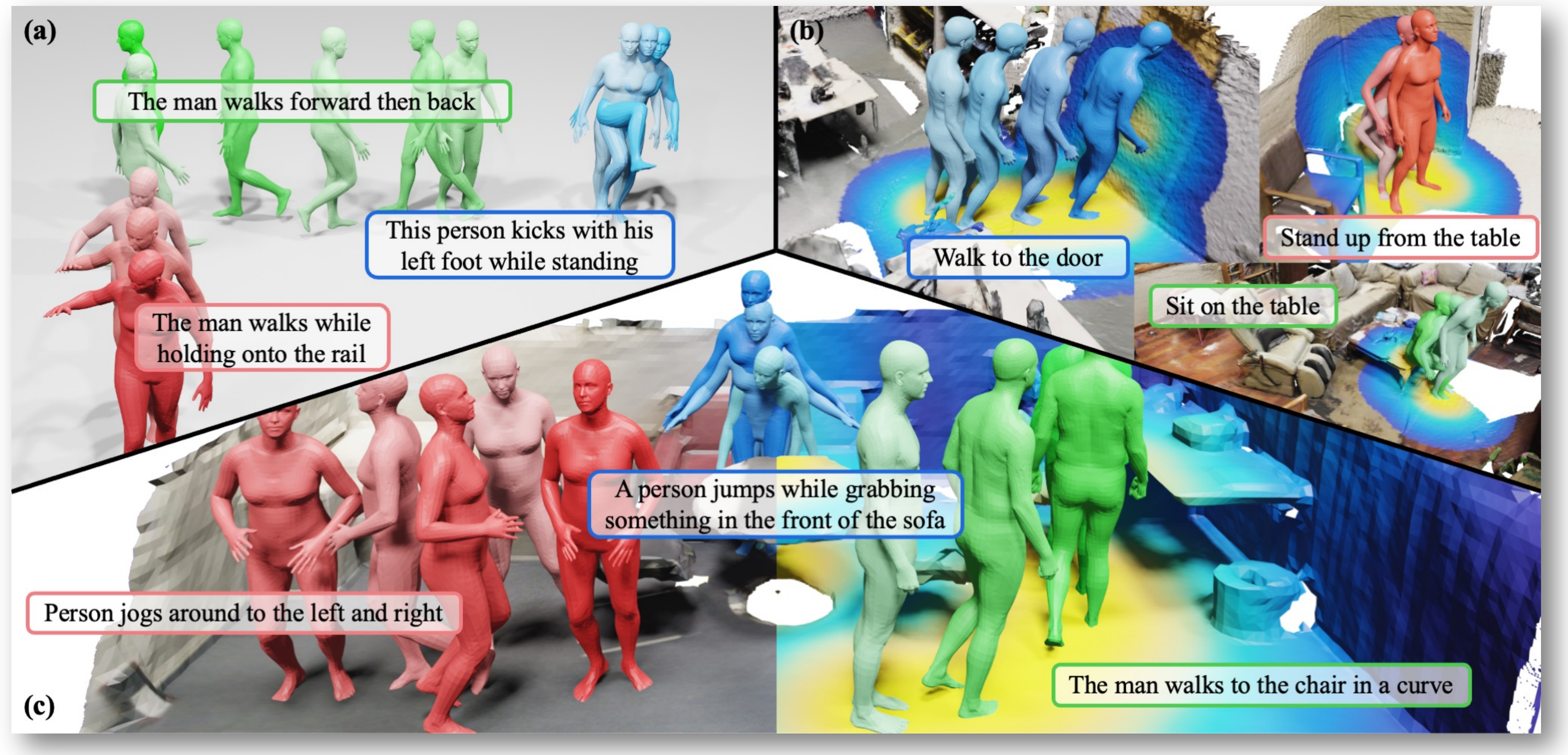
Zan Wang[1,2], Yixin Chen[2], Baoxiong Jia[2], Puhao Li[2,3], Jinlu Zhang[2,4], Jingze Zhang[2,3], Tengyu Liu[2], Yixin Zhu[5]✉, Wei Liang[1,6]✉, Siyuan Huang[2]✉

[1]Beijing Institute of Technology [2]State Key Laboratory of General Artificial Intelligence, BIGAI [3]Tsinghua University
[4]CFCS, School of Computer Science, Peking University [5]Institute for AI, Peking University [6]Yangtze Delta Region Academy of Beijing Institute of Technology, Jiaxing

Project Page https://afford-motion.github.io/

CVPR SEATTLE, WA JUNE 17-21, 2024

Introduction


(a)
The man walks forward then back
This person kicks with his left foot while standing
The man walks while holding onto the rail
(b)
Walk to the door
Stand up from the table
Sit on the table
A person jumps while grabbing something in the front of the sofa
Person jogs around to the left and right
The man walks to the chair in a curve
(c)

We propose to ***leverage scene affordance as an intermediate representation*** *to facilitate language-guided human motion generation in 3D scenes.*

Challenges:

- **The inherent complexity of marrying 3D scene grounding and conditional motion generation**
  - Impede the model's ability *to* generalize to novel scenarios.
- **The generative models' dependency on large volumes of high-quality paired data**
  - Lack large-scale, motion-diverse, and semantic-rich HSI.

Contributions

- We introduce a novel two-stage model that *incorporates scene affordance as an intermediate representation*, facilitating language-guided human motion synthesis in 3D environments.
- We demonstrate our method's *superiority* over existing motion generation models on HumanML3D and HUMANISE benchmarks.
- Our model showcases remarkable **generalization capabilities**, performing impressively in generating human motions within **unseen** scenarios.

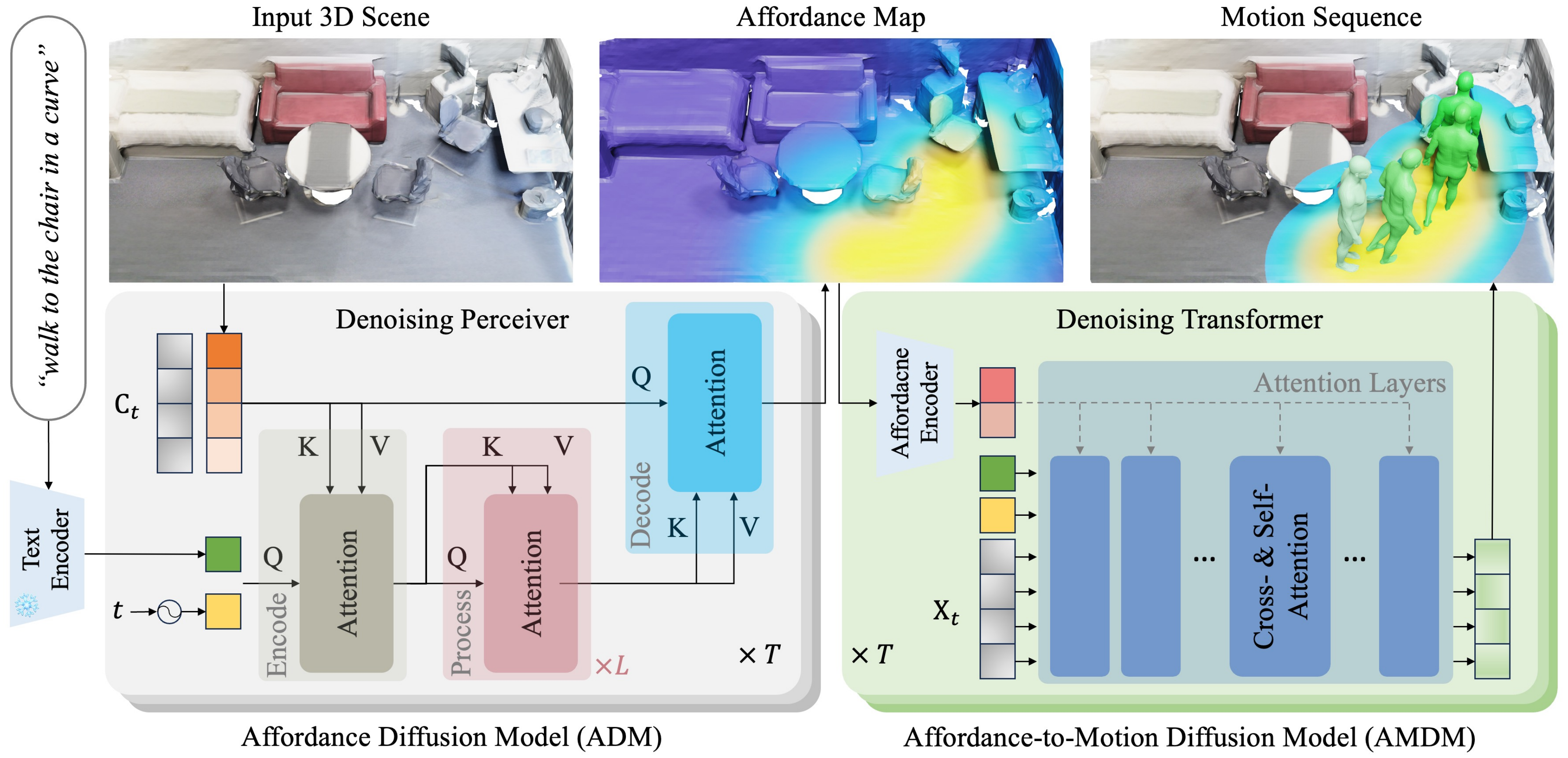
Method

Affordance Map

- We derive the scene affordance map from the distance field between scene points and human skeleton joints

$$\mathbf{c}(n,j) = \exp\left(-\frac{1}{2}\frac{\mathbf{d}(n,j)}{\sigma^2}\right) \quad \mathbf{C} = \texttt{max-pool}(\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_F)$$

  - Enhance the 3D grounding
  - Provide a nuanced understanding of the geometric



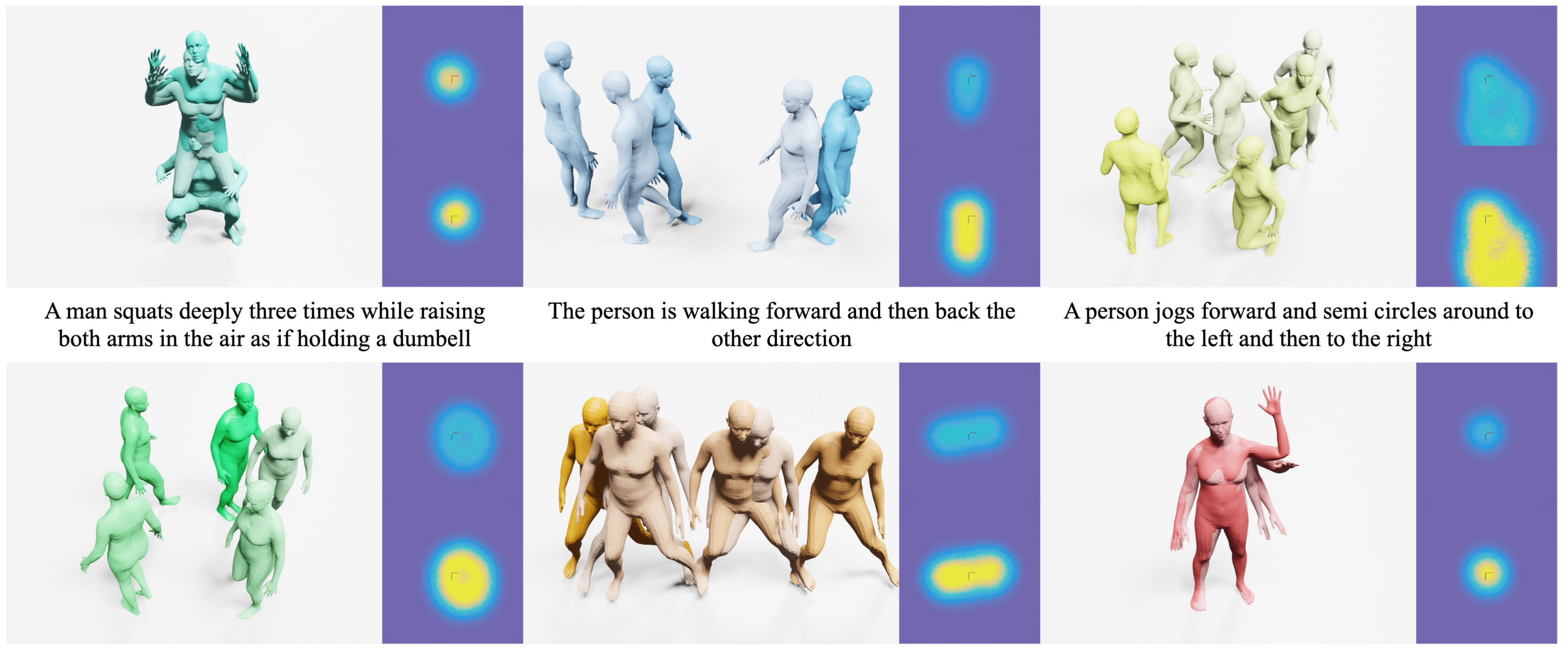
Results

Results on HumanML3D [Guo et al., CVPR 2022]

- Quantative Results

Model	R-Precision ↑			FID ↓	MultiModal Dist. ↓	Diversity →	MultiModality ↑
	Top 1	Top 2	Top 3				
Real	$0.511^{\pm.003}$	$0.703^{\pm.003}$	$0.797^{\pm.002}$	$0.002^{\pm.000}$	$2.974^{\pm.008}$	$9.503^{\pm.065}$	-
Language2Pose [3]	$0.246^{\pm.002}$	$0.387^{\pm.002}$	$0.486^{\pm.002}$	$11.02^{\pm.046}$	$5.296^{\pm.008}$	$7.676^{\pm.058}$	-
T2M [29]	$\mathbf{0.457}^{\pm\mathbf{.002}}$	$\mathbf{0.639}^{\pm\mathbf{.003}}$	$\mathbf{0.740}^{\pm\mathbf{.003}}$	$1.067^{\pm.002}$	$\mathbf{3.340}^{\pm\mathbf{.008}}$	$9.188^{\pm.002}$	$2.090^{\pm.083}$
MDM [76]	$0.319^{\pm.005}$	$0.498^{\pm.004}$	$0.611^{\pm.007}$	$0.544^{\pm.044}$	$5.566^{\pm.027}$	$\mathbf{9.559}^{\pm\mathbf{.086}}$	$2.799^{\pm.072}$
Ours	$0.341^{\pm.010}$	$0.514^{\pm.016}$	$0.625^{\pm.011}$	$\mathbf{0.352}^{\pm\mathbf{.109}}$	$5.455^{\pm.073}$	$9.772^{\pm.117}$	$\mathbf{2.835}^{\pm\mathbf{.075}}$
MDM† [76]	$0.418^{\pm.005}$	$0.604^{\pm.005}$	$0.707^{\pm.004}$	$0.489^{\pm.025}$	$3.631^{\pm.023}$	$\mathbf{9.449}^{\pm\mathbf{.066}}$	$\mathbf{2.873}^{\pm\mathbf{.111}}$
Ours†	$\mathbf{0.432}^{\pm\mathbf{.007}}$	$\mathbf{0.629}^{\pm\mathbf{.007}}$	$\mathbf{0.733}^{\pm\mathbf{.006}}$	$\mathbf{0.352}^{\pm\mathbf{.109}}$	$\mathbf{3.430}^{\pm\mathbf{.061}}$	$9.825^{\pm.159}$	$2.835^{\pm.075}$

- Qualitative Results



A man squats deeply three times while raising both arms in the air as if holding a dumbell
The person is walking forward and then back the other direction
A person jogs forward and semi circles around to the left and then to the right
The person walks in a clockwise circle
A person jumps from side to side right to left
A person waves with his left hand

Results on HUMANISE [Wang et al., NeurIPS 2022]

- Quantative Results

Model	goal dist.↓	APD↑	contact↑	non-collision↑	quality score↑	action score↑
cVAE [84]	$0.422^{\pm.011}$	$4.094^{\pm.013}$	$84.06^{\pm.716}$	$\mathbf{99.77}^{\pm\mathbf{.004}}$	2.25 ± 1.26	3.66 ± 1.38
one-stage @ Enc	$0.326^{\pm.013}$	$\mathbf{5.510}^{\pm\mathbf{.019}}$	$76.11^{\pm.684}$	$99.71^{\pm.014}$	2.60 ± 1.24	3.88 ± 1.32
one-stage @ Dec	$0.185^{\pm.014}$	$4.063^{\pm.020}$	$86.43^{\pm.845}$	$99.76^{\pm.006}$	3.09 ± 1.34	4.18 ± 1.16
Ours @ Enc	$\mathbf{0.156}^{\pm\mathbf{.006}}$	$2.597^{\pm.008}$	$95.86^{\pm.323}$	$99.69^{\pm.007}$	3.46 ± 1.15	$\mathbf{4.47 \pm 0.84}$
Ours @ Dec	$\mathbf{0.156}^{\pm\mathbf{.006}}$	$2.459^{\pm.009}$	$\mathbf{96.04}^{\pm\mathbf{.298}}$	$99.70^{\pm.005}$	$\mathbf{3.55 \pm 1.19}$	4.44 ± 0.85

- Qualitative Results

"Lie down on the bed" *"sit on the chair"* *"stand up from the toilet"* *"walk to the desk"*

Results on Novel Evaluation Set

"A person wanders in the room around the table." *"A man dances on the bed happily."*

Please refer to our project page for the animation videos and more results.