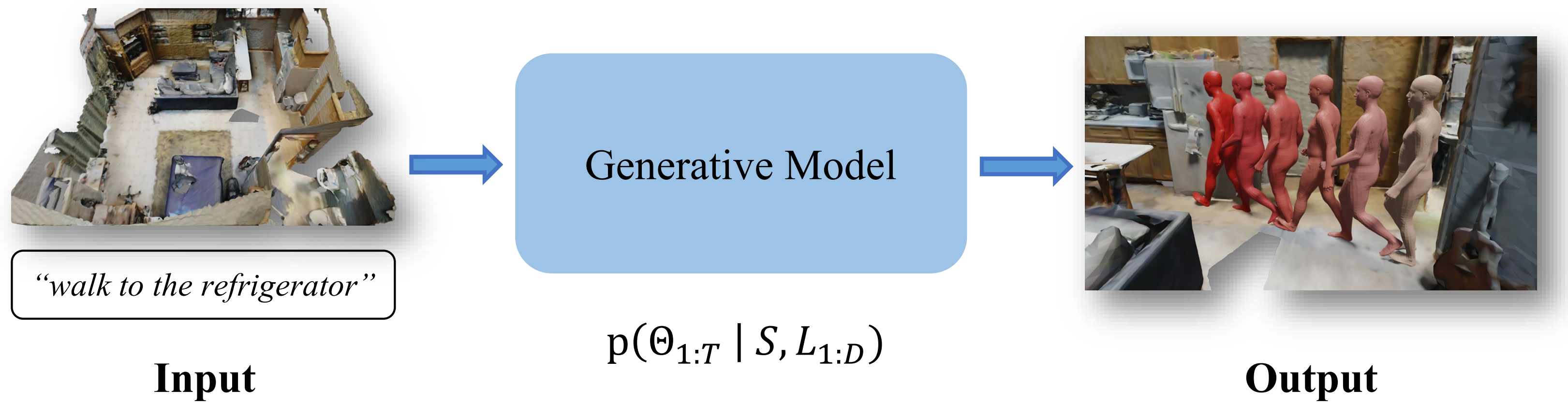
HUMANISE: Language-conditioned Human Motion Generation in 3D Scenes

Zan Wang[1,2], Yixin Chen[2], Tengyu Liu[2], Yixin Zhu[3*], Wei Liang[1,4*], Siyuan Huang[2*]

[1] School of Computer Science & Technology, Beijing Institute of Technology
[2] Beijing Institute for General Artificial Intelligence (BIGAI)
[3] Institute for Artificial Intelligence, Peking University
[4] Yangtze Delta Region Academy of Beijing Institute of Technology

Language-conditioned Human Motion Generation in 3D Scene

We propose a new generation task, language-conditioned human motion generation in 3D scenes. Our goal is to generate a human motion sequence that is both semantically consistent with the language description and physically plausible in interacting with the scene.



The proposed task is more challenging than previous motion generation tasks in three aspects:

- The motion generation is conditioned on the multi-modal information including both the 3D scene and the language description.
- The generated human motions should perform the correct action and be precisely grounded near the target location according to the language descriptions.
- The generated human motions should be realistic and physically plausible within the 3D scenes.

HUMANISE Dataset

Learning to generate diverse scene-aware and goal-oriented human motions in 3D scenes remains challenging due to the limitations in existing HSI dataset, *i.e.*,

- **limited scale and quality;**
- **absence of scene and interaction semantics.**

To solve the above issues, we propose a **large-scale** and **semantic-rich** synthetic HSI dataset, *HUMANISE*, by aligning captured human motion sequences with the scanned indoor scenes. To automatically generate language descriptions for synthesized motions, we also design a compositional template:

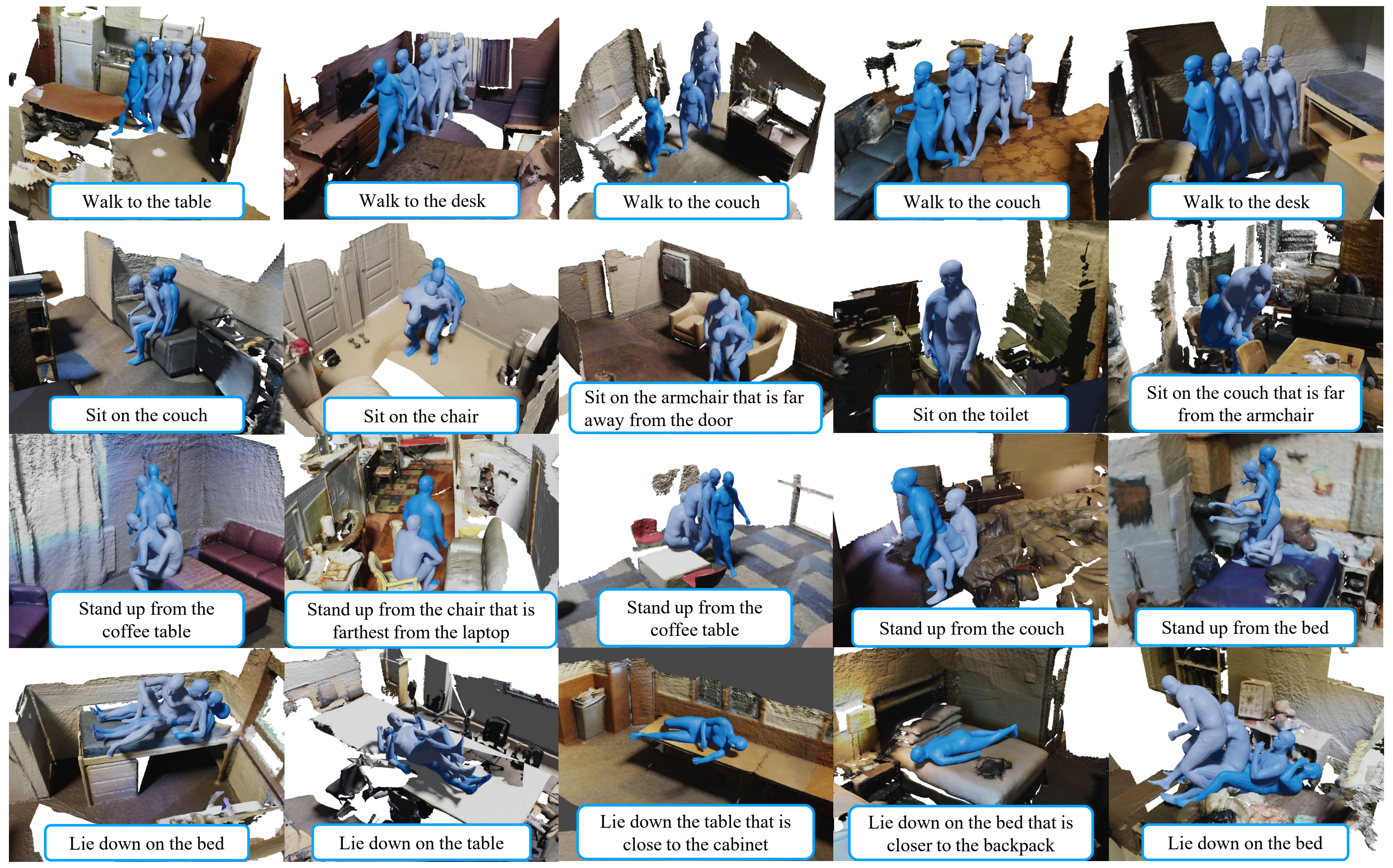
< action >< target-class > [< spatial-relation >< anchor-class(es) >]

For instance, sit on the armchair near the desk.

HUMANISE shows advantages in scene variety, clip number, frame number, and human pose quality. Of note, *HUMANISE* is the only dataset with rich semantic information of HSIs.

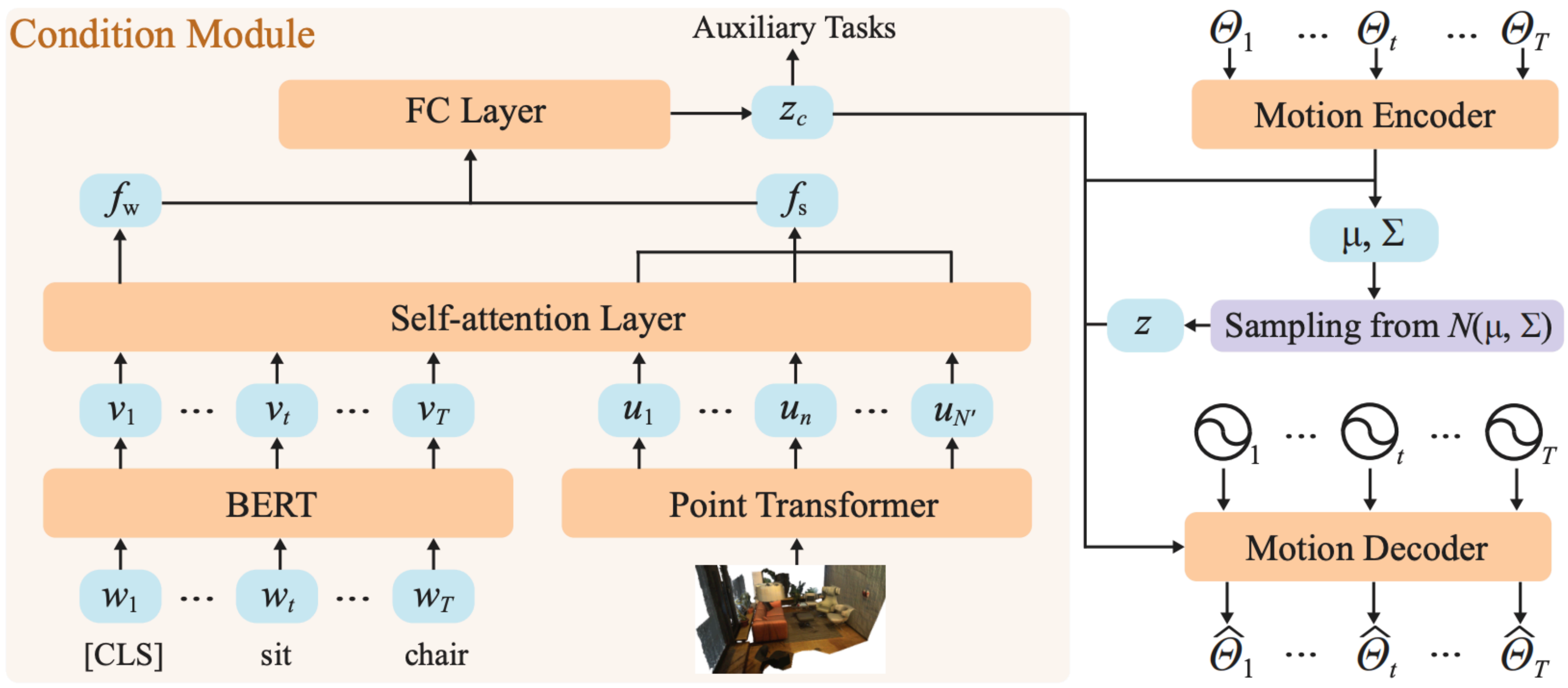
Dataset	#Scenes	#Clips	#Frames	Human Representation	Pose Jittering	Semantics
PiGraph [Savva et al., 2016]	30	63	0.1M	Skeleton	✓	✗
PROX-Q [Hassan et al., 2019]	12	60	0.1M	Shape	✓	✗
GTA-IM [Cao et al., 2020]	49	119	1.0M	Skeleton	✗	✗
HUMANISE	643	19.6k	1.2M	Shape	✗	✓

HUMANISE Dataset Preview


Walk to the table
Walk to the desk
Walk to the couch
Walk to the couch
Walk to the desk
Sit on the couch
Sit on the chair
Sit on the armchair that is far away from the door
Sit on the toilet
Sit on the couch that is far from the armchair
Stand up from the coffee table
Stand up from the chair that is farthest from the laptop
Stand up from the coffee table
Stand up from the couch
Stand up from the bed
Lie down on the bed
Lie down on the table
Lie down the table that is close to the cabinet
Lie down on the bed that is closer to the backpack
Lie down on the bed

Framework

Based on the conditional variational auto-encoder framework, we present a novel generative model to generate human motions conditioned on the given scene and language description.


Condition Module
Auxiliary Tasks
FC Layer
z_c
f_w
f_s
Self-attention Layer
v_1 ··· v_t ··· v_T
u_1 ··· u_n ··· $u_{N'}$
BERT
Point Transformer
w_1 ··· w_t ··· w_T
[CLS]
sit
chair
Θ_1 ··· Θ_t ··· Θ_T
Motion Encoder
μ, Σ
z
Sampling from $N(\mu, \Sigma)$
Motion Decoder
$\hat{\Theta}_1$ ··· $\hat{\Theta}_t$ ··· $\hat{\Theta}_T$

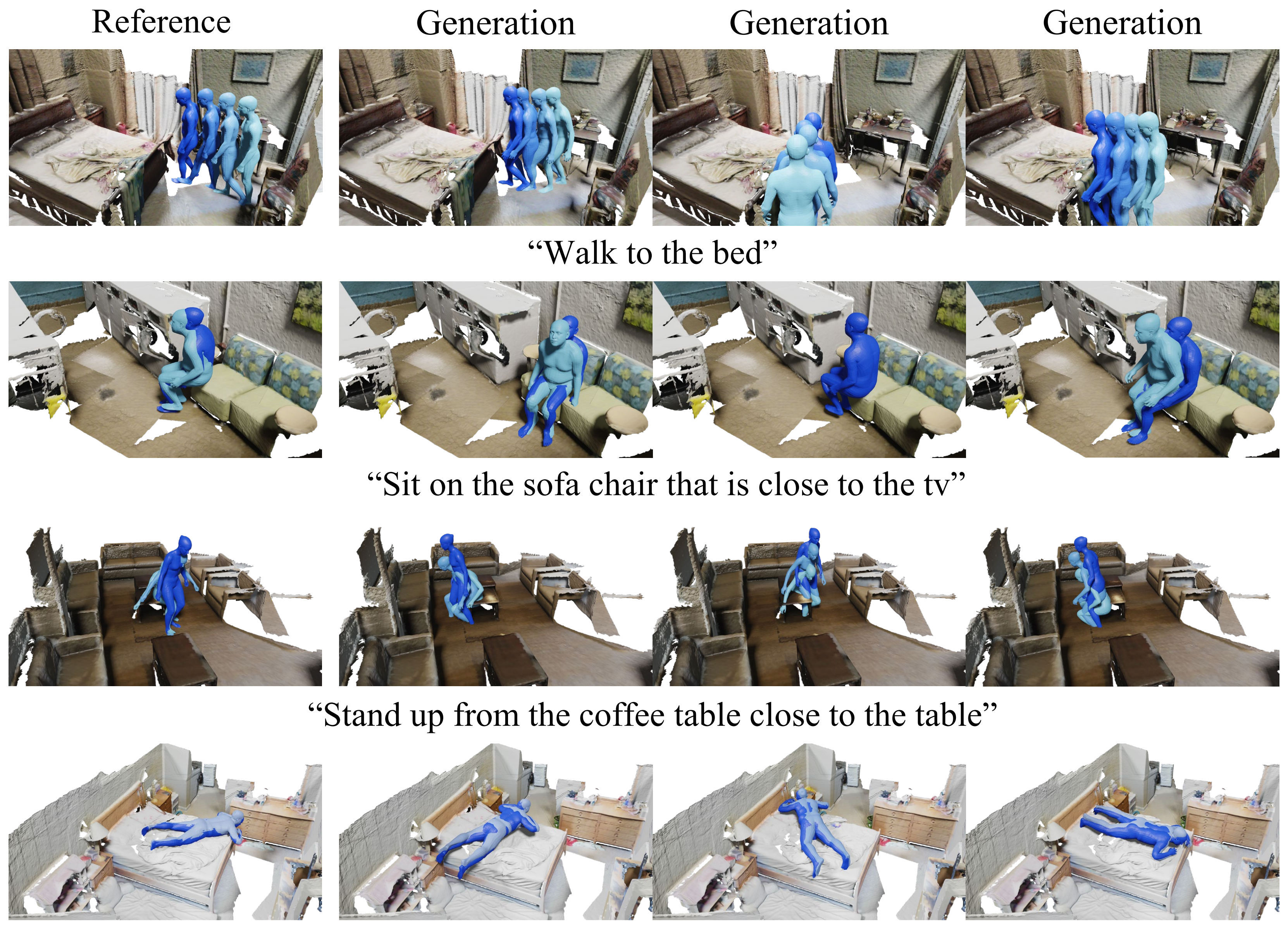
Quantitative Results

The performance advantage of our model, especially on the generation metrics, validates our model designs and the importance of the proposed auxiliary task to the generative model.

Model	Reconstruction					Generation			
	transl.↓	orient.↓	pose↓	MPJPE↓	MPVPE↓	goal dist.↓	APD↑	quality score↑	action score↑
sit	5.17	3.19	1.77	113.28	112.43	0.903	10.12	2.37±0.85	3.79±1.17
stand up	5.63	3.43	1.69	126.05	124.84	0.802	9.57	2.83±1.23	4.20±0.94
lie down	6.46	3.09	0.76	136.87	136.20	0.196	9.18	2.31±1.08	2.85±1.31
walk	5.84	2.80	1.85	125.05	123.88	1.370	12.83	2.91±1.27	3.88±1.26
w/o self-att.	5.72	2.65	1.85	122.19	120.81	1.500	13.28	2.88±1.14	3.80±1.09
PointNet++ Enc.	5.81	2.64	1.81	124.67	123.69	1.444	12.61	2.80±1.35	3.75±1.27
all actions	4.20	2.91	1.96	98.01	96.53	1.008	11.83	2.57±1.20	3.59±1.38
w/o $\mathcal{L}_o$	4.20	2.89	1.93	98.15	96.69	1.383	15.09	2.42±1.21	3.57±1.38
w/o $\mathcal{L}_a$	4.23	2.91	1.95	98.67	97.11	1.135	12.66	2.17±1.04	2.29±1.43
w/o aux. loss	4.28	2.99	1.92	99.30	97.80	1.361	15.18	1.97±0.98	2.44±1.38

Qualitative Results

The results show that our model can generate human motions that are diverse and semantically consistent with the language description.



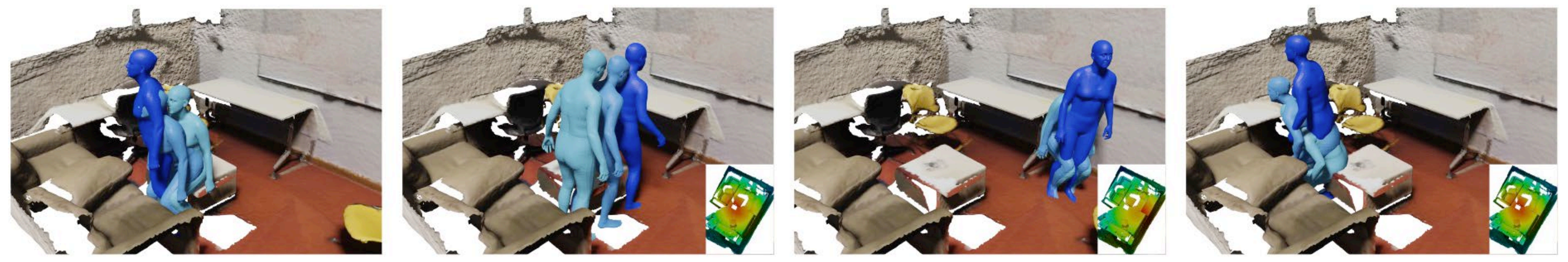
"Walk to the bed"

"Sit on the sofa chair that is close to the tv"

"Stand up from the coffee table close to the table"

"Lie on the bed"

The model *w/o aux. loss* struggles in (b) generating the action specified by the description or (c) locating the interacting object. (d) In comparison, our full model generates motions semantically consistent with the language description.



(a) Reference (b) *w/o aux. loss* (c) *w/o aux. loss* (d) Ours

Contributions

- We propose a **large-scale** and **semantic-rich** synthetic HSI dataset, ***HUMANISE***, that contains human motions aligned with 3D scenes and corresponding language descriptions.
- We introduce a new task of ***language-conditioned human motion generaiton in 3D scenes*** that requires a holistic and joint understanding of 3D scenes, human motions, and language.
- We develop a generative model that can produces **diverse** and **semantically consistent** human motions conditioned on the 3D scene and language description.

https://silverster98.github.io/HUMANISE

Page

Code