Supplementary Materials for *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]✉
✉ indicates corresponding authors
[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, Jiaxing

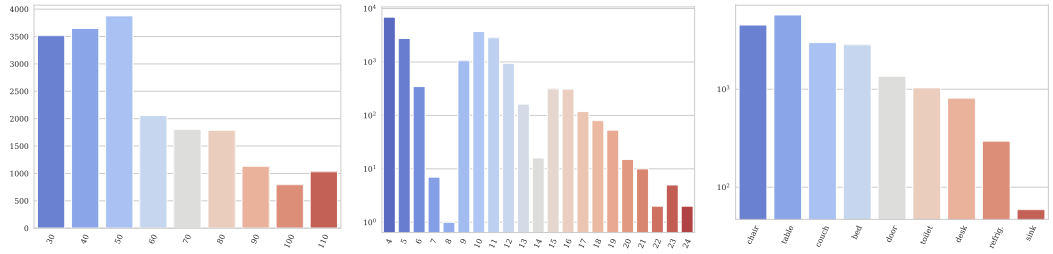
https://silverster98.github.io/HUMANISE/

We present additional statistics of *HUMANISE* and detailed quantitative comparisons between *HUMANISE* and PROX in Appendices A and B, respectively. In Appendix C, we show that our motion alignment pipeline can be easily generalized to various 3D scene datasets and actions, resulting in a vast amount of synthetic data to be used for Human-Scene Interaction (HSI) tasks. Appendix D provides additional qualitative results for both reconstruction and generation, and Appendix E presents some examples of failure cases. In Appendix F, we describe additional ablative and proof-of-concept experiments.

A Statistics of *HUMANISE*

We report additional statistics of the *HUMANISE* dataset. Fig. A1a shows the distribution of frame length. *HUMANISE* contains motions ranging from 30 to 120 frames. Fig. A1b shows the sentence lengths distribution of the language description. The average length of language descriptions in *HUMANISE* is 7.63. Fig. A1c shows the frequencies of the 9 interactive object categories. Since we focus on indoor scene activities, the most frequently interactive objects are "chair," "table," and "couch."

We also report the motion diversity when synthesizing the *HUMANISE* dataset. 1305 different motions from AMASS are selected to synthesize 19.6k motion sequences in 643 3D scenes. Among all the synthesized motion sequences 19.6k [1305], the number of the synthesized motion sequences and the number of selected motions from AMASS for each action are: 8264 [717] for walk, 5578[365] for sit, 3463 [190] for stand-up, and 2343 [33] for lie-down.



(a) Frame length distribution. (b) Sentence length distribution (c) Interactive object frequency.

Figure A1: **Statistics of the *HUMANISE* dataset.**

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

B Quantitative comparison between *HUMANISE* and PROX

In this section, we present quantitative metrics to compare the dataset quality between *HUMANISE* and PROX [Hassan et al., 2019]. We first introduce a collision distance to evaluate the quality of physical implausibility between the human body and the 3D scene. More specifically, it is calculated as the average distance (in meters) between the human body surface and the scene points that penetrate the human body. We also perform human perceptual studies to evaluate the dataset quality in terms of the overall quality, motion smoothness, and quality of human-scene interaction. A worker is asked to score from 1 to 5 for (i) the overall *quality*, (ii) the *smoothness* of the motion, and (iii) whether the human bodies have physically plausible *HSI* with the 3D scenes. A higher value indicates higher quality. We randomly select 100 motion segments each from PROX and *HUMANISE*, and three workers score each sample. As can be seen from Tab. A1, motions in *HUMANISE* have higher quality in terms of collision, smoothness, HSI, and overall quality, which further validates the significance of our proposed dataset in HSI research.

Table A1: **Comparison between HUAMNISE and existing HSI datasets.**

Dataset	Collision Distance↓	Smoothness Score↑	HSI Score↑	Quality Score↑
PROX-Q [Hassan et al., 2019]	0.024	3.13±1.29	3.54±1.35	3.32±1.20
HUMANISE	0.012	4.49±0.84	4.21±1.01	4.14±0.93

C Generalization of motion alignment pipeline

Our dataset synthesis pipeline can generalize to other 3D scene datasets and more action types. This means we can generate a massive amount of synthetic data by aligning the motions with 3D scenes. Below we show some qualitative examples.

Generalize to other 3D scene datasets We employ our motion alignment pipeline to generate human motions in completely hand-craft 3D scenes, *i.e.*, Replica [Straub et al., 2019]. Some aligned results are shown in Fig. A2.

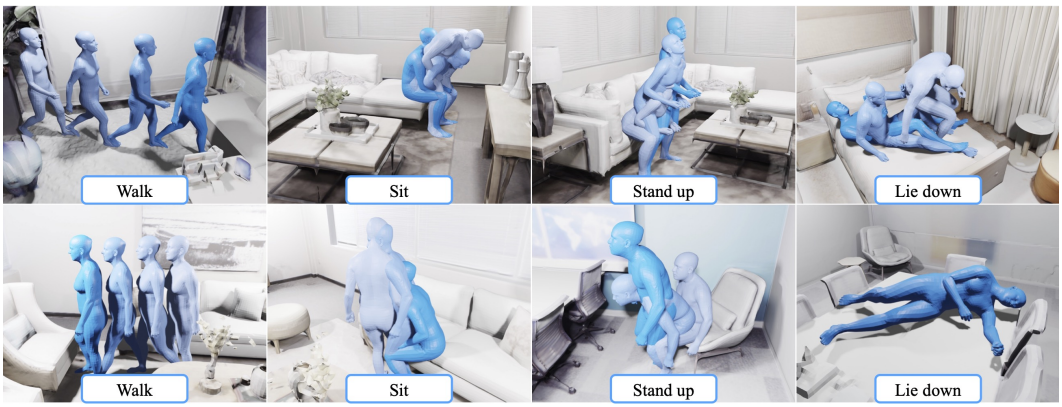


Figure A2: **Motions aligned with scenes in Replica [Straub et al., 2019].**

Generalize to other actions We additionally align motions of extra action types (*i.e.*, *jump up*, *turn to*, *open*, and *place something*) with scanned scenes. Some selected results are shown in Fig. A3.

D Additional qualitative results

Qualitative results of reconstruction More qualitative reconstruction results are shown in Fig. A4.

More qualitative results of generation More qualitative generation results are shown in Fig. A5.

E Typical examples of failure cases

Fig. A6 and Fig. A7 shows that our model sometimes fails to ground the correct target object for interaction and fails to generate correct Human-Object Interaction (HOI) relation.

Figure A3: **Examples of motions with action types *jump up*, *turn to*, *open*, and *place something*.**

F Additional ablations

F.1 Ablations of the *lie-down* action

To diagnose why the *lie down* action has smaller *goal dist.* in the action-specific setting, we conduct additional ablative experiments by using different grounding loss weight, *i.e.*, α_o, and varying the amount of data during training. Specifically, we try $\alpha_o = 0.01$, $\alpha_o = 0.001$, and $\alpha_o = 0$, while the original $\alpha_o = 0.1$. We additionally use 10% and 50% of the original dataset to train the action-specific model of *lie down* action. The quantitative results are shown in Tab. A2.

From the table, we can find that (1) the *goal dist.* increases as α_o decreases, but the values are still smaller than other actions (see Tab. 2). (2) the models trained with fewer data reach approximately the same performance in *goal dist.* as the model trained with all *lie down* data, but the reconstruction metrics drop significantly. These two observations indicate that the grounding task for *lie down* action is easier than other actions. We hypothesize this is because most *lie down* actions interact with *bed* or *table*, which has a flat surface and occupies a much larger area compared to other types of furniture.

Table A2: **Ablations about *lie down* action.**

Model	Reconstruction					Generation	
	transl.↓	orient.↓	pose↓	MPJPE↓	MPVPE↓	goal dist.↓	APD↑
$\alpha_o = 0$	6.67	3.06	0.79	142.94	142.08	0.444	11.72
$\alpha_o = 0.001$	6.76	2.78	0.82	142.30	141.77	0.306	10.97
$\alpha_o = 0.01$	6.77	3.31	0.77	144.15	143.43	0.211	9.50
50%	8.04	4.24	0.96	170.50	169.96	0.177	8.56
10%	10.44	9.15	5.27	248.82	243.99	0.207	8.05
Ours	6.46	3.09	0.76	136.87	136.20	0.196	9.18

F.2 Multi-stage pipeline and end-to-end pipeline

As an initial attempt to explore the difference between an end-to-end approach and a cascaded approach for our task, we conduct experiments to compare against baselines that directly use action

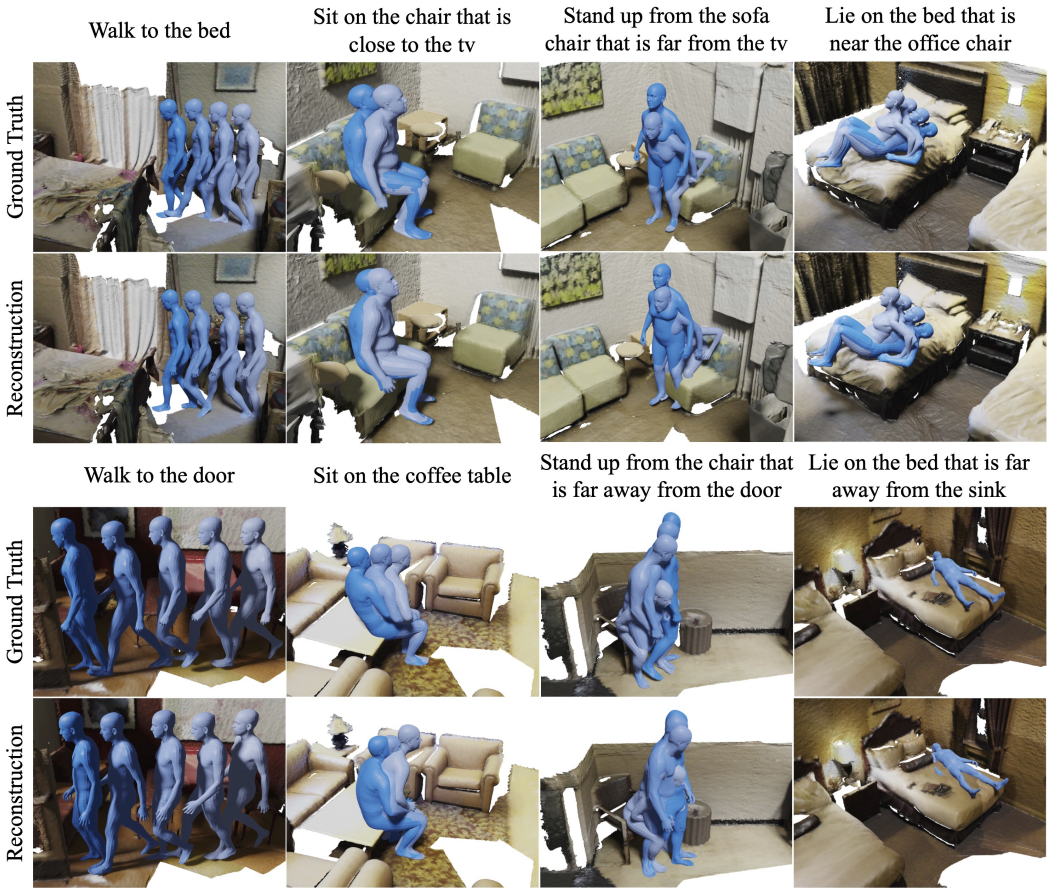


Figure A4: **Qualitative reconstruction results of the action-specific models.**

or target object as conditions. More specifically, We modify our conditional variational auto-encoder (cVAE) model by replacing the condition with the concatenation of the global scene feature, the target object center, and the action category. We use the point cloud feature extracted from PointNet++ as the global scene feature, and we adopt the one-hot embedding to represent the GT action categories. For the target object center, we use a 3D grounding model pre-trained on ScanNet, *i.e.*, ScanRefer [Chen et al., 2020], to estimate the target object center. We denote this baseline as $\mathrm{GT}_{\mathrm{action}}$. We also test another baseline that directly uses the ground truth position instead of the predicted target object position. We denote this baseline as $\mathrm{GT}_{\mathrm{action+target}}$. The quantitative results are shown in Tab. A3.

From the table, we can see that the baseline that directly uses GT action-conditioning reaches approximately the same performance as our model, while utilizing the GT action and the GT target position can significantly improve the motion generation metrics. We hypothesize this is because the action can be easily parsed from the instruction and 3D object grounding is significantly more challenging than other subtasks, which affects the performance most. The results also justify the multi-stage pipeline could achieve similar performance as the end-to-end pipeline in the current setting.

Table A3: **Comparison between multi-stage pipeline and end-to-end pipeline.**

Model	Reconstruction					Generation			
	transl.↓	orient.↓	pose↓	MPJPE↓	MPVPE↓	goal dist.↓	APD↑	quality score↑	action score↑
$\mathrm{GT}_{\mathrm{action}}$	4.23	2.90	1.90	98.67	97.17	1.406	15.98	2.56±1.02	3.93±1.11
$\mathrm{GT}_{\mathrm{action+target}}$	4.13	2.84	1.92	96.02	94.57	0.207	9.37	2.78±1.33	3.86±1.31
Ours	4.20	2.91	1.96	98.01	96.53	1.008	11.83	2.57±1.20	3.59±1.38

G Use of existing assets

Our dataset *HUMANISE* is build on data from AMASS [Mahmood et al., 2019], ScanNet [Dai et al., 2017], BABEL [Punnakkal et al., 2021] and Sr3D [Achlioptas et al., 2020]. AMASS is licensed

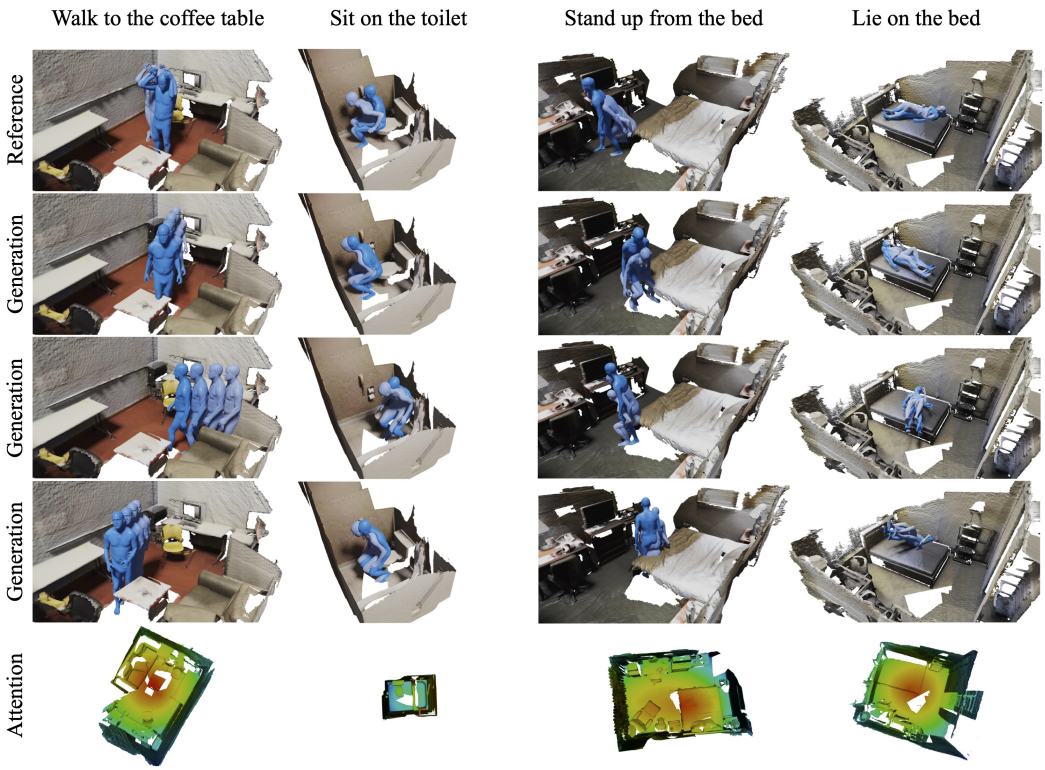
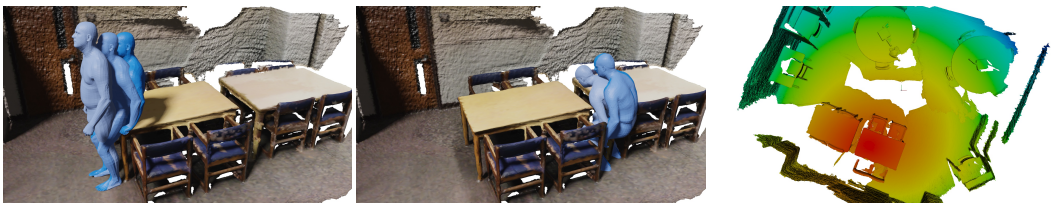


Figure A5: **Qualitative generation results of the action-specific models.**

(a) Reference (b) Failure Sample (c) Attention

Figure A6: **Failure case with incorrect grounding.** The language description, in this case, is *walk to the table that is in the middle of the box and the trash can*.



(a) Reference (b) Failure Sample (c) Attention

Figure A7: **Failure case with incorrect HOI relation.** The language description, in this case, is *sit on the table that is close to the door*.

under the terms of the Dataset Copyright License for non-commercial scientific research purposes. The ScanNet dataset is released under the ScanNet Terms of Use, and the code is licensed under the terms of the MIT license. BABEL is licensed under the terms of the Software Copyright License for non-commercial scientific research purposes. Sr3d is licensed under the terms of the MIT license.

H Human study instructions

Here we include the instructions given to participants of the human perceptual studies for evaluating the generation results in *HUMANISE*.

Instruction You are required to score the given human motion in the scene from the following two aspects: (i) overall quality and (ii) action consistency. For overall quality, you need to consider all the factors that may influence the quality of human motion, including the naturalness, smoothness, physical plausibility, and consistency with the corresponding language descriptions, *etc*. Score the motion from 1 to 5. A higher score means the motion has a higher quality. For action consistency, you should only consider if the action of the motion is consistent with the given language description. Score the motion from 1 to 5. A higher score means the action of the motion is more consistent with the language description.

I Societal impacts

We firmly believe that our work will promote the understanding of HSI, which facilitates real-world applications including VR/AR, human avatars, video game animation, assistant robots, and metaverse. On the other hand, scene and HSI understanding can be extended to activity tracking, which might raise privacy issues to some extent.

References

Achlioptas, P., Abdelreheem, A., Xia, F., Elhoseiny, M., and Guibas, L. (2020). Referit3d: Neural listeners for fine-grained 3d object identification in real-world scenes. In *European Conference on Computer Vision (ECCV)*. 4

Chen, D. Z., Chang, A. X., and Nießner, M. (2020). Scanrefer: 3d object localization in rgb-d scans using natural language. In *European Conference on Computer Vision (ECCV)*. 4

Dai, A., Chang, A. X., Savva, M., Halber, M., Funkhouser, T., and Nießner, M. (2017). Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. 4

Hassan, M., Choutas, V., Tzionas, D., and Black, M. J. (2019). Resolving 3d human pose ambiguities with 3d scene constraints. In *International Conference on Computer Vision (ICCV)*. 2

Mahmood, N., Ghorbani, N., Troje, N. F., Pons-Moll, G., and Black, M. J. (2019). Amass: Archive of motion capture as surface shapes. In *International Conference on Computer Vision (ICCV)*. 4

Punnakkal, A. R., Chandrasekaran, A., Athanasiou, N., Quiros-Ramirez, A., and Black, M. J. (2021). Babel: Bodies, action and behavior with english labels. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. 4

Straub, J., Whelan, T., Ma, L., Chen, Y., Wijmans, E., Green, S., Engel, J. J., Mur-Artal, R., Ren, C., Verma, S., et al. (2019). The replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*. 2