GHOST: Grounded Human Motion Generation with Open Vocabulary Scene-and-Text Contexts
Supplementary Material

Zoltán Á. Milacski[1] Koichiro Niinuma[2] Ryosuke Kawamura[2] Fernando de la Torre[1]
László A. Jeni[1]
[1] Robotics Institute, Carnegie Mellon University, Pittsburgh PA, USA
[2] Fujitsu Research of America, Pittsburgh PA, USA
srph25@gmail.com {kniinuma,rkawamura}@fujitsu.com {ftorre@cs.,laszlojeni@}cmu.edu

A. Additional Experiments

A.1. Additional Evaluation Metrics

In this section, we provide additional performance metrics. While the main paper primarily emphasizes the distance between the generated motions and the goal object, these metrics offer additional insights into those results.

Condition. Here, we compute 2 metrics for evaluating the grounding performance of the condition module. First, we calculate the cosine similarity of the encodings of the text prompt and the pooled cloud point that is nearest to the goal object center:

$$\cos\left(\mathcal{E}^{text}(\boldsymbol{L}), Pool\left(\mathcal{E}^{3D}(\boldsymbol{S})\right)_{goal\ center,4:}\right), \quad (1)$$

where $\mathcal{E}^{text}$ is the text encoder, and $Pool$ is either identity for the HUMANISE cVAE or the k-nearest neighbor downsampling module for our GHOST. We specifically employ goal object center indexing to ensure a fair comparison between all methods, as some of them lack a ground truth goal object mask at this level. Second, we report the $\mathcal{L}_{center}$ MSE regularization loss in meters from the main paper for regressing the goal object center point from both input modalities. We average both metrics across samples.

Reconstruction. This task is significantly easier than generation, given the availability of the ground truth motion location as an input. Yet, we present the respective performance results here in the supplementary material.

We assess the motion reconstruction capability by computing the MAE (ℓ_1 error) $\times 100$ between the ground truth and predicted SMPL-X parameters, specifically for global translation $\boldsymbol{t}$, global orientation $\boldsymbol{r}$, and body pose $\boldsymbol{\theta}$. Following [4, 5] to obtain more interpretable scores, we also calculate the Mean Per Vertex Position Error (MPVPE) and Mean Per Joint Position Error (MPJPE) [3] in millimeters. To handle sequences of various lengths, we average these results over the temporal dimension, and finally, across examples.

Generation. We also present standard deviations corresponding to the average goal object distances $d(\boldsymbol{L}, \boldsymbol{S})$ in the main paper.

Perceptual Study. We present comprehensive per-subject results for our perceptual experiment.

A.2. Additional Quantitative Results

Tab. 1 collects our more detailed results. In condition module evaluation, we found that our GHOST methods achieved significantly larger cosine similarities than the HUMANISE cVAE baseline, indicating better text-scene grounding. However, the goal object center regularization loss $\mathcal{L}_{center}$ correlated the best with the final goal object distance metric. Interestingly, our GHOST LSeg sometimes outperformed the OpenSeg variant in this regard, but the latter still achieved more reliable results. In reconstruction, our global orientation and pose errors were competitive, but the global translations, MPVPEs and MPJPEs were superior for the HUMANISE cVAE. This may be attributed to the impact of our additional regularization losses, which counteract reconstruction efforts, suggesting an area for future enhancement.

Tab. 2 shows the corresponding numbers for ablation. We observe that employing a closed vocabulary scene encoder resulted in strong text-goal cosine similarity. However, it still struggled to accurately regress the center of the goal object, potentially due to ambiguities between the embeddings of the goal object and the rest of the 3D scene. As expected, our regularization losses sometimes hampered reconstruction.

Table 1. Quantitative results of reconstruction and generation experiments on the HUMANISE dataset. The winning numbers are highlighted in bold for each action subset.

		Condition		Reconstruction					Generation	
		Text-Goal obj.	MSE Goal obj.	MAE × 100			MPVPE	MPJPE	Goal obj.	
Action	Method	enc. cos sim. ↑	center reg. (m) ↓	trans. ↓	orient. ↓	pose ↓	(mm) ↓	(mm) ↓	dist.±std (m) ↓	APD ↓
walk	HUMANISE cVAE [5]	1.75	1.372	5.84	2.80	1.85	123.88	125.05	1.370±0.839	12.83
	GHOST LSeg (ours)	9.16	1.090	6.17	2.64	1.83	128.59	129.59	1.090±0.891	10.96
	GHOST OpenSeg (ours)	5.08	**0.990**	5.97	2.86	1.90	126.66	128.02	**0.952**±0.919	10.97
	GHOST OVSeg (ours)	**10.25**	1.101	6.45	2.88	1.87	137.38	138.43	1.027±0.945	**10.62**
sit	HUMANISE cVAE [5]	1.06	0.910	5.17	3.19	1.77	112.43	113.28	0.903±0.744	10.12
	GHOST LSeg (ours)	10.16	**0.621**	6.00	2.89	1.74	127.64	128.48	0.695±0.655	9.28
	GHOST OpenSeg (ours)	6.97	0.709	5.92	2.96	1.79	125.41	126.10	**0.668**±0.708	8.59
	GHOST OVSeg (ours)	**12.40**	0.735	6.10	3.17	1.77	129.72	130.37	0.680±0.743	**8.29**
stand up	HUMANISE cVAE [5]	-0.18	0.875	5.63	3.43	1.69	124.84	126.05	0.802±0.711	9.57
	GHOST LSeg (ours)	12.04	0.861	6.09	3.51	1.71	130.60	131.71	0.767±0.742	8.89
	GHOST OpenSeg (ours)	6.52	**0.595**	6.32	3.73	1.76	134.62	135.70	**0.600**±0.600	**8.45**
	GHOST OVSeg (ours)	**13.22**	0.674	6.91	3.58	1.74	148.29	149.25	0.626±0.681	8.59
lie	HUMANISE cVAE [5]	-3.64	0.397	6.46	3.09	0.76	136.20	136.87	0.196±0.476	9.18
	GHOST LSeg (ours)	**13.91**	**0.327**	7.84	3.04	0.76	169.87	170.54	**0.185**±0.425	8.87
	GHOST OpenSeg (ours)	4.83	0.410	6.99	3.01	0.88	150.64	151.45	0.200±0.468	**8.54**
	GHOST OVSeg (ours)	10.59	0.623	6.95	3.22	0.83	148.60	149.70	0.263±0.603	8.97
all	HUMANISE cVAE [5]	4.84	1.044	4.20	2.91	1.96	96.53	98.01	1.008±0.838	11.83
	GHOST LSeg (ours)	9.49	**0.754**	4.37	2.87	1.91	98.62	99.93	0.748±0.810	**9.54**
	GHOST OpenSeg (ours)	6.17	0.788	4.37	2.82	1.93	98.76	100.02	**0.732**±0.837	9.80
	GHOST OVSeg (ours)	**10.54**	0.823	4.08	2.92	1.90	93.15	94.54	0.767±0.829	10.08

Table 2. Quantitative results of ablation experiments on the walk action in the HUMANISE dataset. The winning numbers are highlighted in bold.

	Condition		Reconstruction					Generation	
	Text-Goal obj.	MSE Goal obj.	MAE × 100			MPVPE	MPJPE	Goal obj.	
Method	enc. cos sim. ↑	center reg. (m) ↓	trans. ↓	orient. ↓	pose ↓	(mm) ↓	(mm) ↓	dist.±std (m) ↓	APD ↓
GHOST OpenSeg w. BERT [2] text enc. (ours)	3.04	1.574	5.85	3.02	1.93	124.71	125.98	1.425±0.917	11.28
GHOST OpenSeg w. closed vocab. scene enc. [1,5] (ours)	**7.81**	1.230	5.95	2.96	1.88	125.35	126.53	1.021±1.032	10.38
GHOST OpenSeg w. $\lambda_{bbox} = 0$ (ours)	4.96	**0.990**	5.92	2.82	1.90	125.43	126.70	1.011±0.860	11.65
GHOST OpenSeg w. $\lambda_{class} = 0$ (ours)	4.91	1.028	6.07	2.75	1.87	128.67	129.85	0.982±0.925	11.09
GHOST OpenSeg w. $\lambda_{class} = 0.1$ (ours)	4.81	1.041	6.55	2.76	1.89	138.66	139.73	0.995±0.952	10.48
GHOST OpenSeg w. $\lambda_{class} = 1.0$ (ours)	4.70	1.021	6.32	2.75	1.88	133.17	134.41	0.970±0.979	**10.21**
GHOST OpenSeg (ours)	5.08	**0.990**	5.97	2.86	1.90	126.66	128.02	**0.952**±0.919	10.97

Table 3. Quantitative results of the perceptual study of agnostic all-actions models trained on the entire HUMANISE dataset. The winning numbers are highlighted in bold.

	Frequency of User Preference ↑								
Method	User 1	User 2	User 3	User 4	User 5	User 6	User 7	User 8	User 9
HUMANISE cVAE [5]	15	22	17	22	22	21	27	21	26
GHOST OpenSeg (ours)	**45**	**38**	**43**	**38**	**38**	**39**	**33**	**39**	**34**
	User 10	User 11	User 12	User 13	User 14	User 15	User 16	User 17	User 18
HUMANISE cVAE [5]	23	27	20	27	21	22	21	25	28
GHOST OpenSeg (ours)	**37**	**33**	**40**	**33**	**39**	**38**	**39**	**35**	**32**
	User 19	User 20	User 21	User 22	User 23	User 24	User 25	User 26	User 27
HUMANISE cVAE [5]	23	21	21	20	20	20	24	20	19
GHOST OpenSeg (ours)	**37**	**39**	**39**	**40**	**40**	**40**	**36**	**40**	**41**

Tab. 3 details our perceptual study results. All 27 subjects picked the samples generated by our GHOST method more frequently, with preferences up to 75%.

References

[1] Angela Dai, Angel X. Chang, Manolis Savva, et al. ScanNet: Richly-annotated 3D Reconstructions of Indoor Scenes. In *CVPR*, 2017. 2

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805*, 2018. 2

[3] Catalin Ionescu, Dragos Papava, Vlad Olaru, et al. Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments. *IEEE TPAMI*, 36(7):1325–1339, 2013. 1

[4] Jiashun Wang, Huazhe Xu, Jingwei Xu, et al. Synthesizing Long-Term 3D Human Motion and Interaction in 3D Scenes. In *CVPR*, pages 9401–9411, 2021. 1

[5] Zan Wang, Yixin Chen, Tengyu Liu, et al. HUMANISE: Language-conditioned Human Motion Generation in 3D Scenes. In *NeurIPS*, 2022. 1, 2