

GenZI: Zero-Shot 3D Human-Scene Interaction Generation

Supplementary Material

In this supplementary material, we first provide more experimental results in Appendix A and then describe more implementation details in Appendix B.

A. More Results

Generation Variations. In Fig. 7, we show different synthesized 3D interaction variations, given the same 3D scene, text prompt, and location specification input. By using a different collection of multi-view interaction hypotheses, produced by latent diffusion inpainting, our approach can generate various plausible 3D human-scene interactions from the same input.

Multi-view Human Inpaintings. In Fig. 8, we show the multi-view human inpaintings used in our robust 3D lifting optimization. In the initial synthesis stage, the images are obtained with our dynamically-masked inpainting. In the refinement stage, the images are generated with the silhouette masks of the posed 3D human from the initial stage. The optimized view consistency score is shown next to each inpainted image (as a blue bar). It is observed that through iterative refinement, the quality of both the multi-view human inpaintings and the synthesized 3D interactions gradu-

Method	Semantics		Diversity		Physical Plausibility	
	CLIP \uparrow	Entropy \uparrow	Cluster Size \uparrow	Non-collision \uparrow	Contact \uparrow	
HUMANISE	0.2537	2.8188	0.6954	0.8882	0.7414	
Ours	0.2710	2.7304	1.0500	0.9767	0.9868	

Table 3. More quantitative comparisons on Sketchfab. Our zero-shot approach outperforms HUMANISE, a recent supervised method generating human motions in 3D scenes from text inputs.

ally improves.

Analysis of Dynamically Generated Masks. We find that for each interaction prompt, our dynamic masking yields on average 13.4 out of 16 rendered views with inpainted humans detected by our 2D pose estimator (Sec. 3.3). This greatly surpasses the minimum 3-views in Eq. 3 of robust 3D lifting, providing strong evidence for the high quality of the dynamically generated masks.

More Comparisons on the Sketchfab Dataset. We additionally compare with HUMANISE [3], which generates human motions in 3D scenes from text inputs. As HUMANISE outputs sequences, we select the frame with the best CLIP score (Sec. 4) to the input prompt for comparison. Tab. 3 and Fig. 9 show that our approach outperforms HUMANISE (especially in the more reliable CLIP

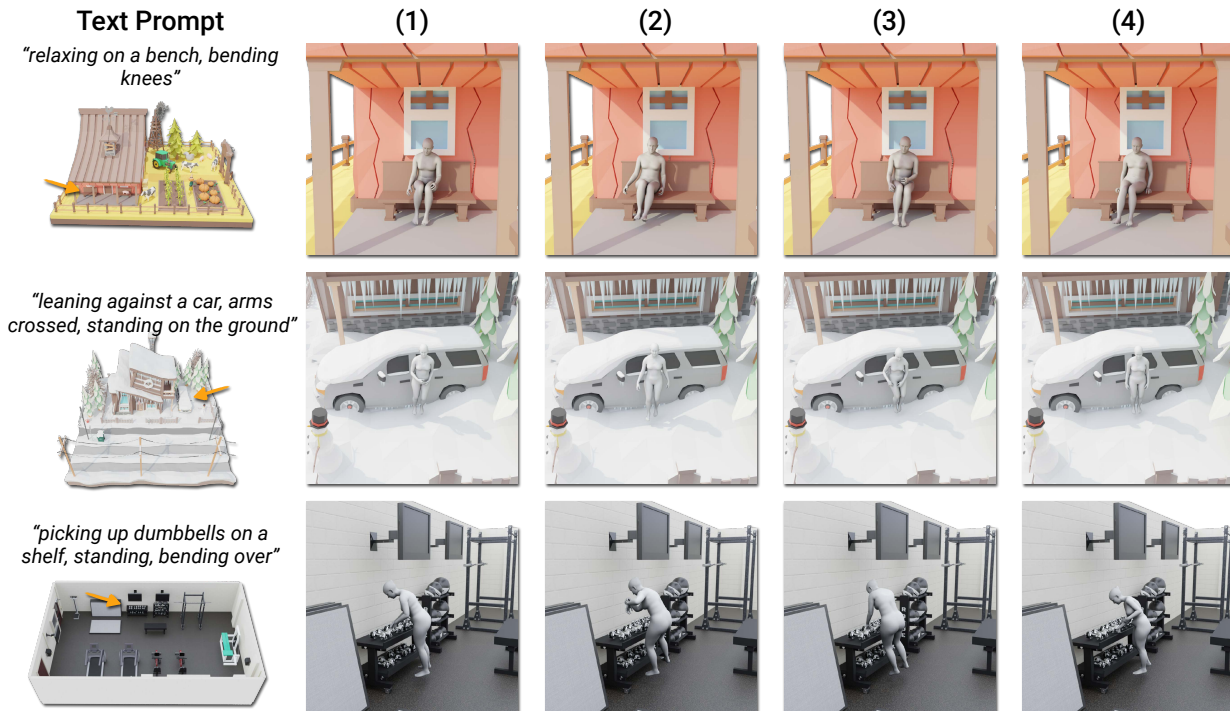


Figure 7. Our approach can generate different plausible human interactions in a 3D scene, from (1) to (4), given the same text prompt and location specification.

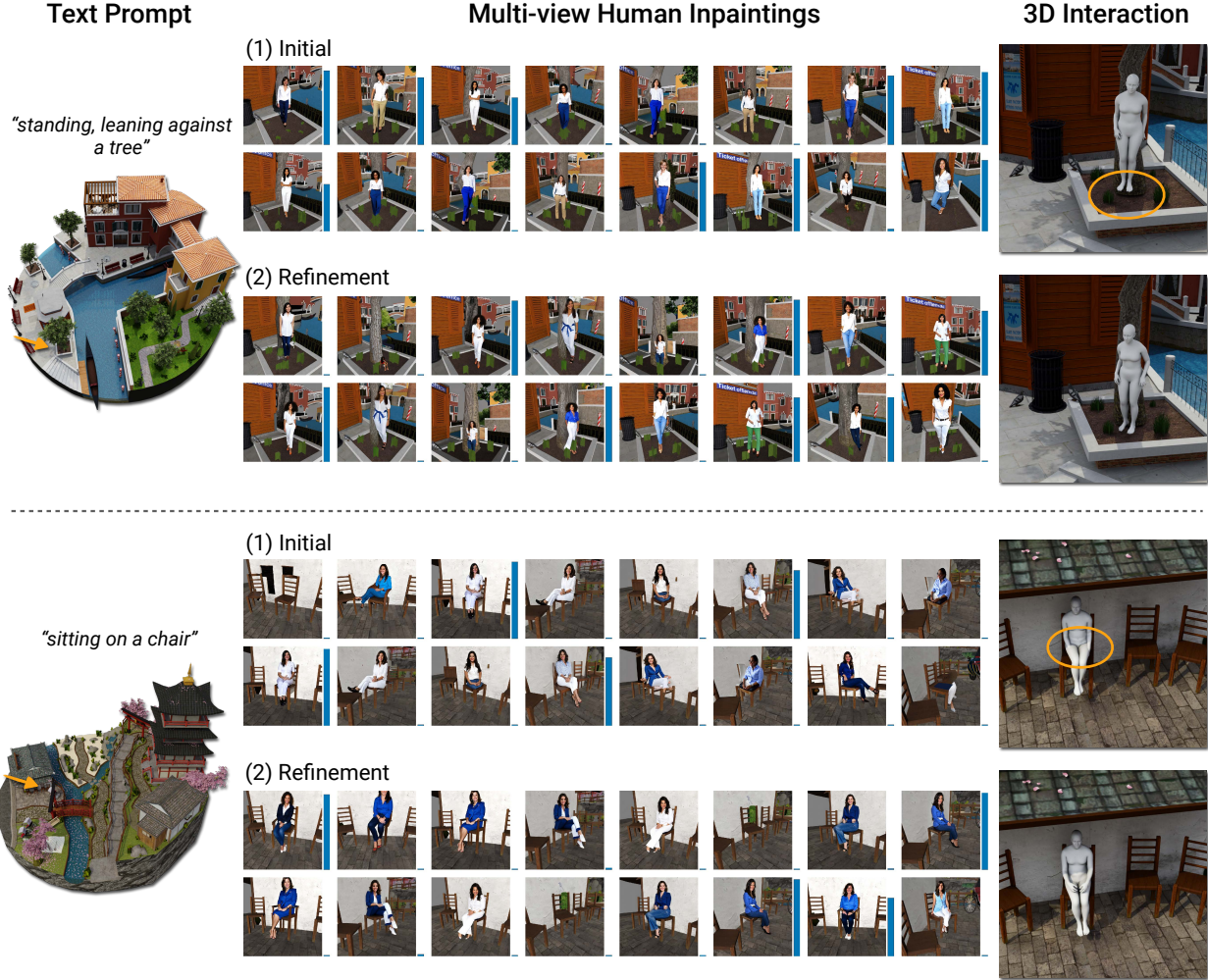


Figure 8. Multi-view human inpaintings used in our robust 3D lifting optimization. (1) Initial synthesis stage: images resulting from dynamically-masked inpainting; (2) Refinement stage: inpaintings with the silhouette masks of the posed 3D human from the initial stage. Without refinement, the person floats above the ground (top), or has self-penetration (bottom). The blue bar next to each inpainted image represents its optimized view consistency score.

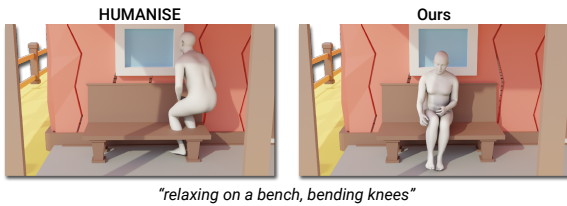


Figure 9. Qualitative comparison with HUMANISE on Sketchfab. Our zero-shot approach generates more realistic 3D interactions and generalizes better than the supervised HUMANISE method.

metric), which requires training on indoor interaction data and struggles to generalize on Sketchfab with more out-of-distribution objects.

Evaluation on the PROX-S Dataset. We perform further comparisons on a recent indoor scene dataset PROX-S [4]

for 3D interaction synthesis. PROX-S consists of 12 indoor scenes with 3D instance segmentations and interaction annotations in the form of $\langle \text{action}, \text{object} \rangle$ pairs. Four of the scenes are used for testing. The interaction synthesis is evaluated on about 150 different combinations of action and object instances in the test set.

To adapt our approach GenZI to PROX-S, we map the provided interaction labels, *e.g.*, $\langle \text{sit on}, \text{sofa} \rangle$, to natural language descriptions, *e.g.*, “sitting on the sofa”. We use the bounding box centers of object instances as the approximate 3D location input. We stress that in a general synthesis scenario (*e.g.*, scenes from the Sketchfab dataset), our approach does *not* require any 3D scene segmentations.

Several baseline methods focused on indoor 3D interaction synthesis are compared, including PiGraph-X [2],

Method	Zero-Shot	Semantics		Diversity		Physical Plausibility	
		CLIP \uparrow	Entropy \uparrow	Cluster Size \uparrow	Non-collision \uparrow	Contact \uparrow	
PiGraph-X	\times	0.2562	3.719	1.019	0.861	0.981	
POSA-I	\times	0.2594	3.680	1.061	0.974	0.941	
COINS	\times	0.2617	3.685	0.782	0.981	0.969	
Ours	\checkmark	0.2544	3.748	0.869	0.992	0.961	

Table 4. Quantitative comparisons on the PROX-S dataset. Our zero-shot approach achieves comparable performance, compared to the baselines that *learn* from the PROX-S training set with 3D scene segmentations, captured 3D human poses, and interaction annotations.

POSA-I [1], and COINS [4]. Note that all the baselines require learning from the PROX-S training set that has 3D scene segmentations, captured 3D human poses, and interaction labellings. In contrast, our approach does *not* require any 3D learning or captured 3D interaction data.

Tab. 4 shows the quantitative evaluation of semantic consistency, diversity, and physical plausibility on PROX-S, and Fig. 10 presents the qualitative comparisons. Our zero-shot approach achieves competitive synthesis performance, when compared to the baselines that are specifically trained on PROX-S. We note that the scenes in PROX-S have noisy, incomplete 3D geometry and very low-quality texture details, as shown in Fig. 10, thus their rendered images can be challenging for latent diffusion to inpaint 2D interaction hypotheses. Nevertheless, our approach has the best diversity entropy and non-collision scores, and can generate plausible 3D interactions in indoor scenes without relying on any captured indoor interaction data.

B. Implementation Details

Multi-view Camera Setup. In Sec. 3.2 of the main text, we create a multi-view representation of the scene context at the location \mathbf{p} by rendering the 3D scene \mathcal{S} from k virtual cameras looking at \mathbf{p} . To determine the camera positions, we first randomly sample a set of 3D points on the $+z$ hemisphere centered at \mathbf{p} with a radius of $d = 2.0\text{m}$, assuming $+z$ as the upward direction for the cameras. We then filter these sampled viewpoints according to the visibility of \mathbf{p} via depth testing. For more robust visibility testing, we opt to crop a local surface patch of \mathcal{S} at \mathbf{p} within a radius of $r = 0.15\text{m}$, and compute the ratio of this patch’s visible area from each viewpoint, based on which the top- k viewpoints are selected.

Dynamic Masking. We summarize our dynamic masking scheme in Algorithm 1 using the same notation as in Sec. 3.2 of the main text. In practice, we use $T = 50$ denoising steps in latent diffusion inpainting. We set $T_{\min} = 25$ for updating the mask \mathbf{M}_t , and keep \mathbf{M}_t unchanged after $t < T_{\min}$ to stabilize the inpaintings.

Angle Prior. In Sec. 3.3 of the main text, we use an angle prior \mathcal{E}_{JA} to regularize extreme bending of the body joints

Algorithm 1 Inpainting with Dynamic Masking

- 1: **Input:** An image \mathbf{I} , a text prompt Γ , token indices h
 - 2: **Output:** An inpainted image $\bar{\mathbf{I}}$
 - 3: **Require:** A latent diffusion inpainting model Ω
 - 4: $\mathbf{z}_T \sim \mathcal{N}(0, 1)$ a Gaussian noise latent
 - 5: $\mathbf{M}_T = \mathbf{0}$
 - 6: **for** $t = T, T - 1, \dots, 1$ **do**
 - 7: $\mathbf{z}_{t-1}, \mathbf{A}_t \leftarrow \Omega(\mathbf{z}_t, \mathbf{M}_t, \mathbf{I}, \Gamma, t)$
 - 8: **if** $t \geq T_{\min}$ **then**
 - 9: $\mathbf{M}_{t-1} \leftarrow \text{binarize}(\text{sum}(\mathbf{A}_t[:, h]))$
 - 10: **end if**
 - 11: **end for**
 - 12: **return** \mathbf{z}_0
-

$\hat{\Theta} \in \mathbb{R}^{21 \times 3}$ represented in the axis-angle form:

$$\mathcal{E}_{\text{JA}} = \sum_{j,a \in \Lambda} |\hat{\Theta}_{j,a}| + \sum_{j,a,s \in \Delta} \max(s \cdot \hat{\Theta}_{j,a}, 0), \quad (7)$$

where $\hat{\Theta}_{j,a}$ denotes the angle of axis a of the j -th joint, and s denotes a sign (± 1). The 21 body joints are divided into two groups Λ and Δ for different angle regularizations in Eq. (7), where Λ consists of head, feet, and wrists, and the rest joints are included in Δ . More implementation details are available in our released code.

Acknowledgements. We thank the following artists for generously sharing their 3D scene designs on Sketchfab.com: “[a Food Truck Project](#)” by xanimvm, and “[WW2 Cityscene - Carentan inspired](#)” by Silkevdsmissen are licensed under [CC Attribution-NonCommercial-NoDerivs](#). “[Bangkok City Scene](#)” by ArneDC, “[Low Poly Farm V2](#)” by EdwiixGG, “[Low Poly Winter Scene](#)” by EdwiixGG, “[Modular Gym](#)” by Kristen Brown, “[1DAE10 Quintyn Glenn City Scene Kyoto](#)” by Glenn.Quintyn, and “[Venice city scene 1DAE08 Aaron Ongena](#)” by AaronOngena are licensed under [Creative Commons Attribution](#).

References

- [1] Mohamed Hassan, Partha Ghosh, Joachim Tesch, Dimitrios Tzionas, and Michael J. Black. Populating 3D scenes by learning human-scene interaction. In *CVPR*, 2021. 3
- [2] Manolis Savva, Angel X Chang, Pat Hanrahan, Matthew Fisher, and Matthias Nießner. PiGraphs: learning interaction snapshots from observations. *ACM TOG*, 35(4):1–12, 2016. 2
- [3] Zan Wang, Yixin Chen, Tengyu Liu, Yixin Zhu, Wei Liang, and Siyuan Huang. HUMANISE: Language-conditioned human motion generation in 3D scenes. *NeurIPS*, 2022. 1
- [4] Kaifeng Zhao, Shaofei Wang, Yan Zhang, Thabo Beeler, and Siyu Tang. Compositional human-scene interaction synthesis with semantic control. In *ECCV*, 2022. 2, 3

!"#\$%&' #() "
!, #-%&' (% &5. +

O(1%&2345

O+ -64!

* +!, -

+ . %/

!, -. ' *%&' (%&&1+

! -&2/0(% .)"\$+

!"#\$%&' (%) \$* +

!, #-%&' (% .)"\$+

! -&2/0(% O\$*3# 4+

!, #-%&' (%/O. #1+

Figure 10. Qualitative results on the PROX-S dataset. Our zero-shot approach can synthesize plausible 3D interactions in indoor scenes *without* relying on any captured indoor interaction data. In contrast, the baselines PiGraph-X, POSA-I, and COINS all require *learning* from the PROX-S training set with 3D scene segmentations, captured 3D human poses, and interaction annotations.