INHABIT: SCALING 3D HUMAN-SCENE INTERACTION DATA WITH IMAGE FOUNDATION MODELS

Anonymous authors
Paper under double-blind review



Figure 1: **InHabit** generates diverse, scene-aware 3D human placement across varied environments and actions. In this craft shop, it produces realistic behaviors such as browsing, leaning, crouching, and reaching. Such interactions are specific to this environment and are hard to reenact in studio capture. We urge the reader to watch the supplementary video and to take a look at the attached webpage to see 3D visualization of our data.

ABSTRACT

Embodied AI and human-centric 3D scene understanding require data of people meaningfully interacting with diverse environments, yet such data remains scarce. Real-world capture has steadily pushed toward more scenes, but is bound to studios and at most around a hundred scenes, while synthetic datasets scale easily but place people without regard to what the scene affords. We argue that useful human–scene interaction data requires diversity in both scenes and interactions, and we show this with our proposed data engine InHabit. InHabit is a fully automatic engine that renders an existing 3D scene, uses 2D foundation models to propose and depict context-appropriate interactions, and lifts them into SMPL-X bodies, where the known scene geometry makes lifting well-constrained. Applied to Habitat-Matterport3D, InHabit produces *InHabitants*, which pairs photorealistic images with SMPL-X bodies and complete scene geometry across $\sim$6.6k rooms, about 60$\times$ more scenes than the largest captured dataset. We illustrate the quality and value of such large-scale human–scene interaction data through a variety of experiments. In a perceptual study, our interactions are preferred over two prior placement methods in 78% of cases. Across three tasks—1) contact estimation, 2) human–scene reconstruction, and 3) generating humans in scenes—training with *InHabitants* outperforms training on synthetic data with exact ground truth (BEDLAM), motion-capture data (TRUMANS), and real RGB-D captures (PROX); e.g., GRAFT trained on *InHabitants* reaches a contact F1 of 0.594 on PROX, versus 0.492 when trained on PROX itself.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1 INTRODUCTION

Understanding 3D scenes from a human-centric perspective is essential for robotics and embodied AI (Puig et al., 2024; Wang et al., 2024; Zhen et al., 2024). Large collections of 3D scenes (Dai et al., 2017; Ramakrishnan et al., 2021) and of captured human motion (Mahmood et al., 2019) exist separately, but datasets of people interacting with 3D environments remain small: capture-based efforts (Hassan et al., 2019; Huang et al., 2022; Guzov et al., 2021; Araujo et al., 2023) cover fewer than 30 scenes each. Synthetic datasets avoid the cost of capture by placing pre-recorded motions into scenes (Black et al., 2023; Wang et al., 2022), but yield people who sit where there is no chair or hover above the furniture they should rest on, rather than engaging with it.

More limiting than their size is what these datasets contain. Capture setups are typically built from generic furniture, so the recorded interactions are those that any room allows: sitting on a horizontal surface, lying on a large one, reaching toward a point. The interactions that make a space what it is are instead defined by what its objects are *for*: playing pool, riding a gym bike, measuring a distance on a wall map, standing on a chair to change a light bulb, or carrying a table together with someone. Recording such behavior would require capture sessions in pool halls, gyms, workshops, and every other type of environment, so it is essentially absent from existing data. Existing data therefore spans only a narrow band of the interaction space, and models trained on it inherit that narrowness, generalizing poorly to how people actually occupy a scene. Covering it requires a way to produce human–scene interaction (HSI) data in arbitrary scenes, without capture.

To produce such data, we build on the observation that 2D foundation models have seen millions of images of exactly these activities, and that this knowledge can be transferred into 3D by rendering a scene we already have in 3D. We introduce InHabit, a fully automatic engine that follows a *render–generate–lift* principle: we render a view of an existing 3D scene, a vision–language model (Google DeepMind, 2025a) proposes interactions the visible environment affords, an image-editing model (Google DeepMind, 2025a) inserts one or more people performing them, and an optimization-based step (Yalandur Muralidhar et al., 2025) recovers SMPL-X (Pavlakos et al., 2019) bodies in the scene. Two obstacles stand in the way. Lifting a single edited view is ill-posed, which we address by optimizing against the complete scene geometry with an explicit human scale, rather than against partial depth. And naively asking an editing model for a person yields generic standing poses, which we address by letting the VLM propose both scene-afforded and action-driven interactions, preserving the long tail.

Applied to Habitat-Matterport 3D (Ramakrishnan et al., 2021), InHabit produces *InHabitants*, spanning $\sim$6.6k rooms, about 60$\times$ more scenes than the largest captured dataset. Every sample pairs a photorealistic image with SMPL-X bodies and the complete scene geometry, and includes single- and multi-person interactions, all without manual annotation. In a perceptual study, our placements are preferred over prior work in 78% of cases.

Data generated this way is not only plausible, it is better training data than what capture provides. On established benchmarks for contact estimation, human–scene reconstruction and generating humans in scenes, training with *InHabitants* outperforms training on motion-capture data (TRUMANS (Jiang et al., 2024)), real RGB-D captures (PROX (Hassan et al., 2019)), and synthetic data with exact ground truth (BEDLAM (Black et al., 2023)), improving SOTA methods for contact estimation (Tripathi et al., 2023a) and HSI reconstruction (YM et al., 2026; Chen et al., 2026). In an ablation over size-matched subsets (Sec. F), the breadth of interactions contributes most to these gains. Notably, GRAFT (YM et al., 2026) trained on *InHabitants* reaches a contact F1 of 0.594 on PROX, against 0.492 when trained on PROX itself.

In summary, our contributions are:

1. InHabit, a fully automatic engine that populates any indoor 3D scene with diverse, scene-appropriate human interactions, requiring no manual annotation or human in the loop.
2. *InHabitants*, an HSI dataset spanning $\sim$6.6k rooms with complete scene geometry, SMPL-X bodies and photorealistic images, including multi-person interactions.
3. Evidence that fully automatically generated data competes with and surpasses motion-capture, real-world and exactly-labeled synthetic training data across three tasks.

We will release our dataset and code publicly.

2 Related Work

Human–Scene Interaction Datasets. Pioneering work on datasets, such as PiGraphs (Savva et al., 2016) and PROX (Hassan et al., 2019), obtains human–scene data from real RGB-D recordings in only a handful of scenes. Later works (Huang et al., 2022; Guzov et al., 2021; 2024; Zhang et al., 2022) use MoCap within pre-scanned scenes, but their dependency on complex MoCap and scanning infrastructure still limits their scalability, reaching at most 30 scenes. TRUMANS (Jiang et al., 2024) covers 100 scenes by procedurally rearranging objects, but remains MoCap-dependent for human poses and is semantically constrained. Synthetic datasets (Black et al., 2023; Wang et al., 2022; Cao et al., 2020) are scalable but place prerecorded human motions into scenes, producing humans that are merely physically valid in the scene, but are not truly engaging with it. CIRCLE (Araujo et al., 2023) combines VR and MoCap to capture richer behaviors, but contains only 9 scenes. Across these datasets the same tradeoff recurs: data is either grounded in real interactions but confined to a few scenes, or scalable but populated with people who ignore their surroundings. Using InHabit, we introduce *InHabitants*, spanning $\sim$6.6k rooms, where every interaction is proposed for the specific environment it occurs in and lifted against the complete scene geometry. An ablation in Sec. F indicates that this breadth contributes most to its value as training data.

Human Placement in 3D Scenes. The dominant approach to placing humans in 3D scenes relies on geometric features and optimization for physical constraints such as contact and penetration (Zhang et al., 2020). POSA (Hassan et al., 2021) learns a data-driven contact prior from captured interactions, but requires a pre-specified input pose, and only processes scene semantics through coarse-grained labels. GenZI (Li & Dai, 2024) generates 3D human-scene interactions from text but requires the user to specify the location of the interaction and relies on a brittle multi-view consensus optimization. While these methods produce physically plausible configurations, they neither reason about what a scene affords nor run without per-interaction human input, so none of them scales to generating a dataset. InHabit addresses both by letting 2D foundation models decide what people do where, removing the human from the loop entirely and populating $\sim$6.6k rooms automatically.

Leveraging 2D Image Generation for 3D Tasks. A recent line of work uses 2D image or video generation models to solve 3D tasks. HumanWild (Ge et al., 2025) generates humans in scenes using diffusion models and recovers 3D annotations from the generated images, showing that 2D generation can substitute for expensive real-world capture. ComA (Kim et al., 2024) uses 2D inpainting to synthesize human–object interaction images, then lifts these to 3D affordance maps, showing that interaction knowledge latent in 2D models can be grounded in 3D. DAViD (Kim et al., 2025) generates 2D human–object interaction videos using video diffusion and lifts them to 4D interaction samples. While these works are promising and show that 2D generative models can be a source of 3D human data, they operate only on isolated human–object pairs or single tasks without reasoning about the broader scene context, and none can generate HSI data at full building scale. We take this principle further, to entire indoor environments, producing scene-appropriate interaction data at unprecedented scale fully automatically.

3 Method

Our goal is to generate large-scale 3D human–scene interaction (HSI) data with natural, diverse, and context-aware behaviors. We formulate data generation as a *render–generate–lift loop* (Fig. 2): we render an existing 3D scene, use image foundation models to generate realistic humans in image space, and lift the results back into physically grounded 3D bodies aligned with the known scene. Rendering is what connects image foundation models to 3D: since each image comes from a known scene and camera, the diverse, context-aware humans these models insert are registered to metric geometry, turning ill-posed monocular reconstruction into a well-constrained fit.

Given any static 3D indoor scene (e.g., large-scale scanned environments such as HM3D (Ramakrishnan et al., 2021)), InHabit has four stages. 1) We automatically sample camera viewpoints, render RGB/depth views, and perform scene-aware human insertion in 2D image space (Sec. 3.1), where foundation models generate interaction proposals and edited images. 2) We reconstruct and place inserted humans in 3D with geometric and physical constraints from the full scene mesh (Sec. 3.2). 3) We scale this process through automated multi-view generation and batching (Sec. 3.3). 4) Because each stage can introduce artifacts, we apply post-reconstruction quality control to retain only

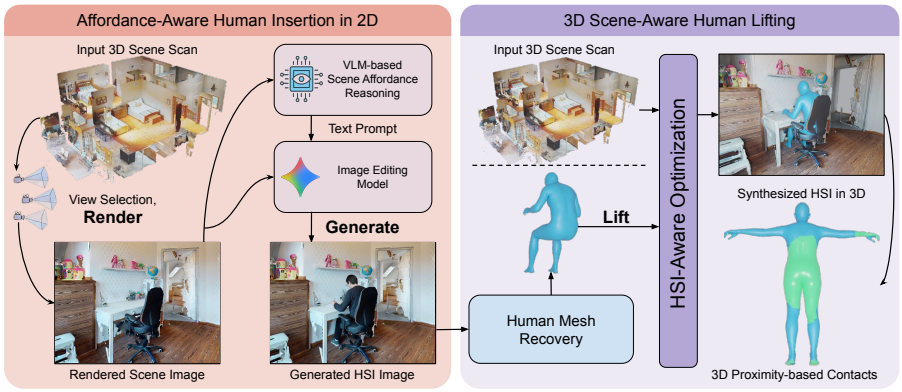

Figure 2: **Scalable 2D-to-3D human–scene interaction generation.** Overview of InHabit. *Left*: affordance-aware 2D insertion—from rendered scene views, a VLM proposes context-consistent interactions and an image-editing model synthesizes humans in the image. *Right*: 3D scene-aware lifting—the inserted humans are recovered as SMPL-X meshes and optimized with scene geometry and proximity-based contact constraints for physically plausible 3D HSI.

physically plausible samples (Sec. 3.3). The final output is a dataset of RGB-D images paired with complete 3D scene geometry and aligned SMPL-X human meshes, ready to be used as training data in downstream tasks.

3.1 Generating Scene-aware Humans in 2D

The first stage provides the behavior: it populates rendered scene views with humans performing diverse, contextually meaningful interactions. This requires understanding both *what* interactions a scene affords and *how* they would look, which we obtain from two foundation models, a vision–language model (VLM) and an image-editing model. A naive idea would be to instruct the editing model to "Add a person to this image." However, this yields mostly generic standing poses, ignoring rich affordances visible in the scene—a kitchen with a stove affords cooking, a living room with a sofa facing a TV affords watching, yet none of this is captured by a generic instruction.

Instead, we first query a VLM to analyze the rendered view and propose plausible activities for the visible scene. The VLM's world knowledge lets it reason about the functions of objects and typical human behaviors—e.g., that a kitchen with a microwave affords operating it, or that a bedroom affords lying down. We call this *affordance-driven prompting* as it produces semantically appropriate interactions for the scene without predefined text prompts or manual supervision. However, even affordance-driven proposals tend to converge on the most common activities and miss long-tail, rare ones. We therefore also use an *action-driven* strategy: given a set of generic action categories (e.g., reaching, leaning, pushing), we ask the VLM to tailor each to the current scene. For example, "reaching" in a room with a ceiling light may become "standing on a chair to replace the light bulb." These categories apply to any indoor scene without forcing implausible interactions, and yield more diverse interactions. We also prompt for multi-person interactions, e.g., a couple dancing or moving furniture (Fig. 12).

Given the VLM's textual interaction proposals, the image-editing model serves as the visual realization engine, generating body poses, proportions, and spatial placements adapted to the local scene context. As seen from Fig. 3, a person washing their hands is placed in front of the sink with hands below the tap. These scene-adapted pose variations and multi-person interactions would be impossible to generate with pure optimization-based methods (Li & Dai, 2024; Hassan et al., 2021; Zhang et al., 2020) that rely on geometric features without visual scene understanding (Fig. 6). Prompt and design details are provided in Sec. C.2.

3.2 Lifting Inserted Humans to 3D

The second stage provides the geometry, building on physically-grounded human–scene reconstruction such as PhySIC (Yalandur Muralidhar et al., 2025) and PROX (Hassan et al., 2019). PhySIC estimates both 3D humans and scenes from RGB images by optimizing the human pose to satisfy geometric constraints with the predicted depth map, whereas PROX fits humans to RGB-D frames



Figure 3: *Left*: representative samples from *InHabitants*. Each pair shows the edited RGB image (left) and the corresponding lifted 3D human in the 3D scene (right). Our dataset contains diverse, context-aware interactions across varied indoor environments. *Right*: distribution of *InHabitants* by action type (*top*) and by interacted object category (*bottom*), showing broad coverage of both diverse activities and scene elements.

using physical constraints. Since we take a complete known 3D scene geometry as input, our lifting task is better constrained, allowing more precise human placement and alignment with scene geometry. In Secs. D and E we ablate both choices; without the complete scene geometry, for example, the contact accuracy of the lifted bodies drops substantially (Tab. 6).

We optimize the human parameters of pose $\boldsymbol{\theta}$, shape $\boldsymbol{\beta}$, camera-relative translation $\mathbf{t}$, and scale s with the following objective inspired by PhySIC:

$$\begin{aligned}\text{minimize}_{\boldsymbol{\theta},\boldsymbol{\beta},\mathbf{t},s} \quad & \lambda_{\text{proj}}\mathcal{L}_{\text{proj}} + \lambda_{\text{depth}}\mathcal{L}_{\text{depth}} + \lambda_{\text{con}}\mathcal{L}_{\text{contact}} \\ & + \lambda_{\text{pen}}\mathcal{L}_{\text{penetration}} + \lambda_{\text{reg}}\mathcal{L}_{\text{reg}},\end{aligned} \tag{1}$$

where $\mathcal{L}_{\text{proj}}$ enforces 2D reprojection consistency, $\mathcal{L}_{\text{depth}}$ aligns the visible human surface with the estimated depth, $\mathcal{L}_{\text{contact}}$ encourages plausible contact with scene surfaces, and $\mathcal{L}_{\text{penetration}}$ penalizes intersections with the scene geometry. Full loss definitions and weights are provided in Sec. C.3.

3.3 Generating the InHabitants Dataset

InHabit is agnostic to the source of 3D scenes and can operate on any textured scene mesh. In this work, we instantiate it on the Habitat-Matterport 3D dataset (HM3D) (Ramakrishnan et al., 2021), containing $\sim$800 building-scale scans spanning residential, commercial, and institutional interiors with multiple floors and rooms, containing kitchens, bedrooms, offices, hallways, stairwells, etc. This enables a diverse distribution of human–scene interactions in the resulting dataset (Fig. 3). As detailed in Sec. C.6, the full dataset is generated in about 19 hours on 32 consumer GPUs.

For scalability, we need automatic viewpoint selection, but random camera poses would often look directly at a nearby wall, be obstructed by furniture or show only ceilings or narrow spaces where inserting a human is infeasible. We therefore filter viewpoints using a depth heuristic: we require 1) floor visibility in the bottom of the image, so that humans can be added without truncation, and 2) sufficient free space ahead, allowing an unobstructed view. Both criteria are evaluated by threshold checks on the depth map, with details in Sec. C.1.

Since some rooms can be largely empty and offer limited affordances (e.g., bare corridors, unfurnished storage rooms), the VLM's affordance-driven prompting (Sec. 3.1) would only yield a narrow set of generic activities. To achieve pose diversity even then, we augment our prompting with fine-grained text descriptions of body poses from the PoseScript dataset (Delmas et al., 2022).

Post-reconstruction Filtering. Image editing models are not perfect and the output is not always physically consistent with the underlying 3D scene geometry. To remove implausible or incomplete

Table 1: **Dataset diversity.** We compare the diversity of interactions and objects with other datasets. We extract actions and target objects with a VLM and compute average cosine distance to measure their diversity. Despite containing the fewest images, *InHabitants* has the most diverse interactions, target objects and poses, and covers about 60× more scenes than TRUMANS.

Dataset	Images	Scenes	Action Diversity ↑	Target Obj. Div. ↑	APD ↑
TRUMANS	1.6M	100	0.406	0.452	25.71
PROX-D	88k	12	0.410	0.445	32.81
InHabitants (Ours)	78k	**6.6k**	**0.485**	**0.493**	**35.42**

Table 2: ***InHabitants* improves human–scene reconstruction methods substantially**. Best result per model in bold, second best underlined.

Model	Training Data	PROX Quantitative			RICH-100		
		F1 score ↑	V2S ↓	PA-MPJPE ↓	F1 score ↑	V2S ↓	PA-MPJPE ↓
Human3R	BEDLAM	0.298	199.0	<u>57.8</u>	<u>0.402</u>	272.6	<u>47.0</u>
	PROX	0.215	250.1	66.6	0.279	302.1	73.5
	InHabitants (Ours)	<u>0.373</u>	**177.0**	59.5	<u>0.402</u>	<u>257.9</u>	48.1
	+ BEDLAM	**0.415**	<u>179.4</u>	**57.2**	**0.459**	**249.2**	**46.0**
GRAFT	BEDLAM	0.396	<u>183.79</u>	84.0	0.469	230.14	65.2
	PROX	0.492	**166.51**	72.5	0.378	270.14	84.4
	InHabitants (Ours)	**0.594**	184.36	**49.7**	**0.578**	<u>224.66</u>	<u>46.3</u>
	+ BEDLAM	<u>0.564</u>	185.75	<u>51.0</u>	<u>0.561</u>	**221.52**	**45.0**

human insertions and ensure high-quality data, we apply three complementary, automated post-reconstruction filters. We discard 1) examples with strongly truncated humans, 2) generations where the edited image contains unintended scene changes besides inserting the human, and 3) human mesh reconstructions with implausibly small volumes. Details can be found in Sec. C.5; the resulting poses are also physically more stable than those of captured datasets (Sec. G).

Dataset Diversity. To characterize the dataset, we use a VLM (Google DeepMind, 2025b) to classify each generated sample by its action and the primary object involved, in an open-ended manner. Fig. 3 shows the distribution over the most frequent actions and objects, illustrating the breadth of interactions in *InHabitants*. We further quantify our diversity in Tab. 1, showing higher interaction diversity, target-object diversity, and Average Pose Distance (APD) than TRUMANS and PROX-D.

4 Experiments

In this section we show how *InHabitants* contributes to advances of state-of-the-art methods on established benchmarks for three different tasks: 1) image-based human–scene reconstruction, 2) generating 3D humans in scenes, and 3) body contact surface estimation. The scale and diversity of *InHabitants* are well suited to these tasks, as they would require expensive motion capturing or tedious manual annotations otherwise. Furthermore, we demonstrate that our generation framework creates poses that are both more natural and diverse than other 3D human placement methods.

4.1 Human–Scene Reconstruction

Human–scene reconstruction (HSR) jointly recovers, from a single image, the 3D body and the surrounding scene geometry in a spatially consistent way (e.g., feet on the floor, hand on the table). Diverse real-world training data with both signals paired is scarce, exactly the gap *InHabitants* fills.

We train two recent image-based HSR methods, GRAFT (YM et al., 2026) and Human3R (Chen et al., 2026) on *InHabitants* and evaluate on PROX Quant. (Hassan et al., 2019) and RICH-100 (Yalandur Muralidhar et al., 2025; Huang et al., 2022). We compare *InHabitants* against two commonly used datasets: BEDLAM (Black et al., 2023) and PROX (Hassan et al., 2019). BEDLAM is a synthetic dataset that provides images, 3D body annotations and depth maps. However, it places human motion randomly into the scene, resulting in mostly walking poses and interactions that are not supported by the scene, such as sitting in the air. PROX, on the other hand, records people interacting in various ways (e.g., sitting or lying), but captures only 12 small scenes. Our HSR metrics are

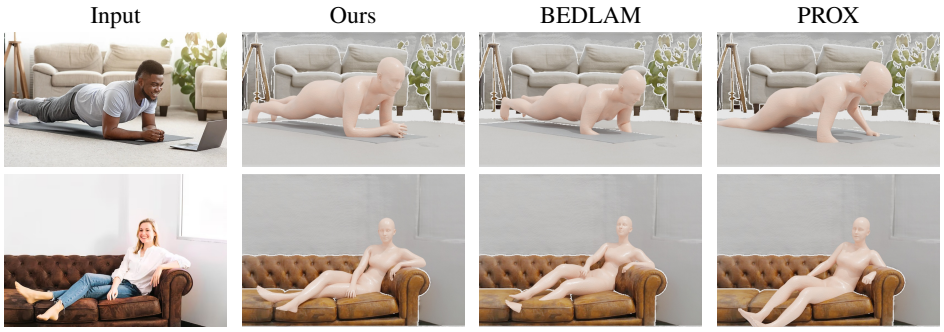

Figure 4: **GRAFT trained on *InHabitants* reconstructs bodies with more realistic scene interactions.** We apply GRAFT trained on *InHabitants*, BEDLAM and PROX, to unconstrained internet images. Training on *InHabitants* yields bodies that rest on the surfaces that support them, while others more often intersect or float relative to them, such as the legs sinking into the sofa for BEDLAM.

Table 3: **SceneDiffuser achieves improved results on ScanNet when trained on *InHabitants*.**

Training Data	Contact ↑	Penetration (in L) ↓	Foot-Contact ↑	Physical Stability ↑	Diversity ↑
PROX	0.570	**5.2**	0.04	0.14	2.76
TRUMANS	0.733	7.1	0.21	0.27	3.70
InHabitants (Ours)	**0.795**	6.8	**0.29**	**0.32**	**4.02**

F1-score for scene contacts and V2S (YM et al., 2026) to evaluate the alignment of the body with the scene. We also report PA-MPJPE, measuring pose accuracy.

Training on *InHabitants* improve HSR metrics regardless of the underlying method used and outperforms using existing datasets overall when compared on both generalization benchmarks (see Tab. 2). This is due to the larger number of scenes and the diverse interactions our dataset provides, which PROX and BEDLAM lack. Human3R trained exclusively on our data is on par with BEDLAM on RICH, since RICH sequences are restricted to feet and hand contacts. The advantage is more apparent on PROX-Quant., which contains more diverse scene interactions. However, combining both BEDLAM and *InHabitants* results in the best performance for Human3R, highlighting that our interactions are complementary to BEDLAM's. GRAFT, a method that explicitly reasons about contacts in 3D space, trained on our data alone improves substantially in HSR and pose metrics compared to the other datasets. It also generalizes better to in-the-wild images as can be seen in Fig. 4. Notice how the bodies rest on the surface with minimal penetration. *InHabitants* is well suited for this approach and advances the state of HSR.

4.2 Learning to Generate Humans in Scenes

InHabitants is generated with a complex pipeline involving 2D generation and 3D reconstruction. However, once generated, this data can be used to train a much faster model that generates SMPL bodies given a 3D scene as input, distilling the entire InHabit pipeline into a single diffusion model. We adopt SceneDiffuser (Huang et al., 2023) for this, a pose diffusion model conditioned on the input point cloud of the scene. We train it on *InHabitants* and compare against models trained on PROX and TRUMANS (Jiang et al., 2024). We evaluate all models on 638 ScanNet (Dai et al., 2017) scenes, an out-of-distribution setting for all trained models. Following prior work (Huang et al., 2023), we measure scene contact and pose diversity, and additionally report penetration volume (SMPL–scene intersection), foot contact, and the percentage of physically stable poses (Sec. G).

The model trained on *InHabitants* outperforms both baselines in terms of contact, foot contact, and physical stability, while still providing a higher level of diversity (see Tab. 3 and Fig. 5). The high diversity of scenes and interactions allows it to generalize better to out-of-distribution data. The models trained on PROX and TRUMANS tend to generate more floating bodies as indicated by the lower contact and foot contact numbers. Especially the model trained on PROX struggles to create realistic contact, leading to its lower penetration volume. Conversely, our higher contact comes with some penetration, e.g., bodies sinking into the bed (Fig. 5), which may partly stem from our reconstruction deliberately allowing penetration to account for soft surfaces such as sofas and beds. Overall, our data provides high quality, diverse interactions to help further research.



Figure 5: **SceneDiffuser trained on *InHabitants* generates stable, diverse poses with realistic scene-contact compared to baselines.** Models trained on TRUMANS and PROX produce more floating bodies and interactions that are physically unstable. Notice the floating feet for TRUMANS and PROX, whereas ours are in close contact to the floor, highlighting the quality of our data.

Table 4: ***InHabitants* improves contact estimation** performance of DECO on DAMON-test.

Training Data	Prec. ↑	Rec. ↑	F1 ↑	geo. err ↓
DAMON	0.639	0.523	0.512	25.06
DAMON+TRUMANS	0.670	0.481	0.497	18.87
DAMON+*InHabitants*	**0.670**	**0.532**	**0.534**	**17.60**

Table 5: **Placement evaluation** regarding semantic prompt alignment and contacts.

Method	CLIP ↑	Prec. ↑	Rec. ↑	F1 ↑
GenZI	0.252	0.500	0.052	0.060
POSA++	0.249	0.715	0.134	0.195
InHabit (Ours)	**0.266**	**0.720**	**0.229**	**0.267**

4.3 Human–Scene Contact Estimation

Body contact estimation predicts, from a single RGB image, which body-surface vertices touch the scene. It is a key signal for human–scene reasoning, but is bottlenecked by the cost of per-vertex annotation on real images. We show that *InHabitants* provides high-quality contact supervision that directly improves a SOTA estimator. Since we recover both the human and the surrounding geometry, we read off per-vertex contacts directly from spatial proximity, yielding image-to-contact pairs without manual labeling. We use these to train DECO (Tripathi et al., 2023a), a SOTA model for in-the-wild body contact estimation, and evaluate on the DAMON benchmark.

We follow the standard DECO protocol for both training and evaluation, including its metrics. Following TRUMANS (Jiang et al., 2024), the closest prior dataset providing human–scene contact supervision, we train on the DAMON+*InHabitants* combination; results are reported in Tab. 4, where the DAMON and DAMON+TRUMANS results are those reported by TRUMANS. DAMON+*InHabitants* outperforms DECO trained on DAMON alone, and also outperforms training on DAMON+TRUMANS. This is remarkable, as TRUMANS is a labor-intensive real-world motion capture dataset, whereas our dataset is generated fully automatically. One reason is that TRUMANS is limited to 100 scenes, offers narrower action coverage (see Tab. 1), and contains images that are less realistic than ours. This shows that our images are aligned with the 3D scene well enough to provide a high-quality supervision signal for fine details such as contacts, which generalizes to the manually annotated labels of DAMON.

4.4 Comparing InHabit with Human Placement Baselines

While InHabit is fully automated and does not require user-specified text prompts, for comparability with prior work, we evaluate the text-to-interaction step of our pipeline against the state of the art. *GenZI* (Li & Dai, 2024) lifts multi-view Stable Diffusion inpaintings to 3D, given a text prompt and a target location. *POSA* (Hassan et al., 2021) predicts per-vertex semantic contacts and places the body by contact optimization with the scene geometry. As POSA does not support text input, we extend it to POSA++ by adding text-based pose retrieval (Patel et al., 2021; Delmas et al., 2024). We use 200 semantically annotated scenes from HM3D (Puig et al., 2024), to accommodate POSA++ and sample 150 diverse prompt-location pairs from our generated affordance-aware prompts. To evaluate prompt consistency we follow the procedure of Li & Dai (2024) and compute CLIP score. To assess whether contacts are both physically plausible and semantically meaningful, we introduce

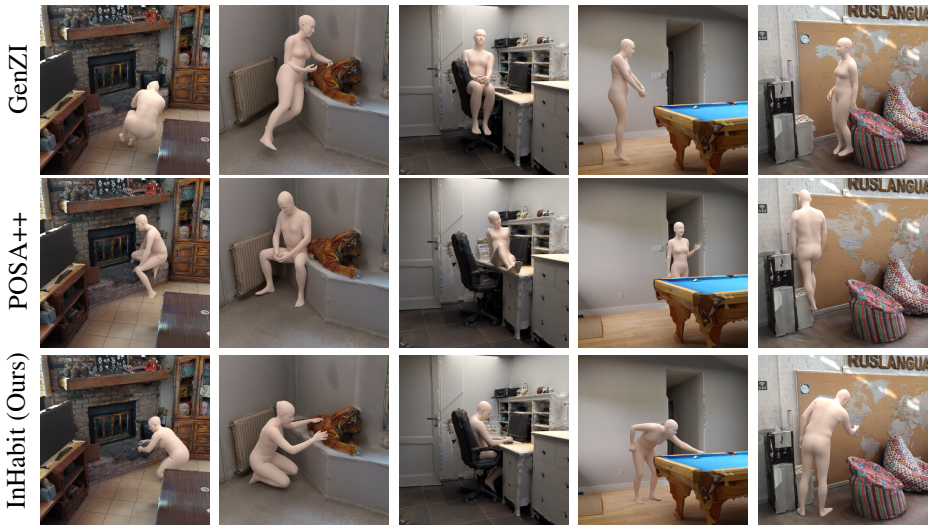

Figure 6: Qualitative comparison. Our method generates diverse interactions, including those with rare objects (tiger, fireplace) and touch-less interactions (map). It also produces natural poses fitting for the context of the object (office chair, pool table).

semantic contact metrics: an LLM predicts the body parts expected to be involved in each interaction (e.g., feet and hands for *leaning on a counter*), which serve as pseudo-GT for computing Precision, Recall and F1-score of the body-scene contacts.

As shown in Tab. 5, InHabit achieves the highest CLIP score as well as the best semantic contact metrics. POSA++ pure optimization based approach cannot adapt to the scene, and interactions without scene contact, such as playing pool or pointing at a map, cannot be modeled (Fig. 6). GenZI, in turn, suffers from inconsistencies across its generated views, which lead to inaccurate lifting and implausible configurations (e.g., a foot attached to a wall), reflected in its lower contact recall. InHabit generates interactions in image space and reasons about the scene prior to lifting, yielding poses physically and semantically consistent with the scene.

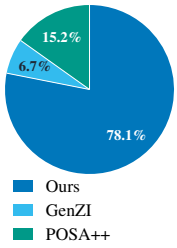

Figure 7: InHabit's poses are preferred 78% of the time.

We conducted a perceptual user study, where 41 participants were shown triplets of generations made with InHabit, GenZI and POSA++ at corresponding scene locations, in random order, as interactive 3D visualizations (interface shown in Fig. 8). They were asked to select the single best result, considering the realism of human–scene interaction, physical feasibility and pose naturalness. Across 726 total responses, InHabit was preferred 78.1% of the time (Fig. 7).

5 Conclusion

We have presented InHabit, a fully automatic engine that populates 3D scenes with diverse, scene-appropriate human interactions by transferring the interaction knowledge of 2D foundation models into 3D. Applied to Habitat-Matterport 3D, it produces *InHabitants*, spanning ∼6.6k rooms, where every sample pairs a photorealistic image with SMPL-X bodies and the complete scene geometry. Training with *InHabitants* improves contact estimation, human–scene reconstruction and generative models of humans in scenes over motion-capture, real captured and exactly-labeled synthetic training data. Currently, InHabit generates static poses within static 3D scenes, but the principle extends to video generation and dynamic scenes as the underlying generative models continue to evolve. More broadly, our results show that the coverage of human–scene interactions that capture cannot reach can now be produced automatically, at a scale that makes it useful for training.

AI USE STATEMENT

In this work, we used generative AI tools to aid or polish writing, to draft sections of the paper, for retrieval and discovery of related work, for execution, and for generating synthetic datasets. We did not use AI for ideation of our novel contributions. We did not use generative AI tools to derive or verify mathematical claims, and this work contains no theoretical results. We have reviewed all AI-assisted work and take responsibility for the final content of this paper, including text, claims or artifacts produced with the aid of generative AI.

Our use of these tools for the paper itself was primarily transformational rather than generative: the authors wrote the content of each section as drafts and outlines, and used LLMs to suggest prose variations, to condense text, and to improve clarity and grammar. No section was produced from a prompt alone. All AI-assisted text was reviewed and edited by the authors.

The dataset introduced in this paper, *InHabitants*, is itself generated with foundation models: a vision–language model proposes scene-appropriate interactions and an image-editing model synthesizes the corresponding humans in rendered views, as described in Sec. 3.1. All generated samples pass the automatic quality control of Sec. C.5, and we show failure cases in Sec. I. The dataset is a research artifact produced by our method, and we take responsibility for its content.

REPRODUCIBILITY STATEMENT

We will publicly release our code and the full *InHabitants* dataset, including the generated images and the lifted SMPL-X bodies with their contacts. The 3D scenes are taken from the publicly available HM3D dataset. Beyond the method description in Secs. 3.1 and 3.2, the appendix documents every stage of the pipeline: the viewpoint selection heuristic (Sec. C.1), all prompts given to the vision–language and image-editing models (Sec. C.2), the reconstruction objective with all loss terms, weights and off-the-shelf components (Sec. C.3), and the quality-control filters with their thresholds (Sec. C.5). Sec. C.6 reports the runtime and hardware used to generate the dataset. For all downstream tasks we follow the standard training and evaluation protocols of the respective benchmarks, as described in the setup of each experiment. Since the pipeline builds on foundation models that are updated over time, we release the generated data itself, so that all downstream results can be reproduced from the released dataset.

REFERENCES

Joao Pedro Araujo, Jiaman Li, Karthik Vetrivel, Rishi Agarwal, Deepak Gopinath, Jiajun Wu, Alexander Clegg, and C. Karen Liu. CIRCLE: Capture in rich contextual environments. In *CVPR*, 2023.

Fabien Baradel, Matthieu Armando, Salma Galaaoui, Romain Brégier, Philippe Weinzaepfel, Grégory Rogez, and Thomas Lucas. Multi-HMR: Multi-person whole-body human mesh recovery in a single shot. In *ECCV*, 2024.

Michael J. Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. BEDLAM: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In *CVPR*, 2023.

Zhe Cao, Hang Gao, Karttikeya Mangalam, Qi-Zhi Cai, Minh Vo, and Jitendra Malik. Long-term human motion prediction with scene context. In *ECCV*, 2020.

Yue Chen, Xingyu Chen, Yuxuan Xue, Anpei Chen, Yuliang Xiu, and Gerard Pons-Moll. Human3R: Everyone everywhere all at once. In *ICLR*, 2026.

Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. ScanNet: Richly-annotated 3D reconstructions of indoor scenes. In *CVPR*, 2017.

Ginger Delmas, Philippe Weinzaepfel, Thomas Lucas, Francesc Moreno-Noguer, and Grégory Rogez. PoseScript: 3D human poses from natural language. In *ECCV*, 2022.

Ginger Delmas, Philippe Weinzaepfel, Francesc Moreno-Noguer, and Grégory Rogez. PoseEmbroider: Towards a 3D, visual, semantic-aware human pose representation. In *ECCV*, 2024.

Yongtao Ge, Wenjia Wang, Yongfan Chen, Hao Chen, and Chunhua Shen. 3D human reconstruction in the wild with synthetic data using generative models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Google DeepMind. Gemini 2.5: Our most intelligent AI model. *arXiv:2507.06261*, 2025a.

Google DeepMind. Gemini 3. `https://deepmind.google/models/gemini/`, 2025b.

Vladimir Guzov, Aymen Mir, Torsten Sattler, and Gerard Pons-Moll. Human POSEitioning system (HPS): 3D human pose estimation and self-localization in large scenes from body-mounted sensors. In *CVPR*, 2021.

Vladimir Guzov, Julian Chibane, Riccardo Marin, Yannan He, Yunus Saracoglu, Torsten Sattler, and Gerard Pons-Moll. Interaction replica: Tracking human-object interaction and scene changes from human motion. In *3DV*, 2024.

Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J. Black. Resolving 3D human pose ambiguities with 3D scene constraints. In *ICCV*, 2019.

Mohamed Hassan, Partha Ghosh, Joachim Tesch, Dimitrios Tzionas, and Michael J. Black. Populating 3D scenes by learning human-scene interaction. In *CVPR*, 2021.

Chun-Hao P. Huang, Hongwei Yi, Markus Höschle, Matvey Safroshkin, Tsvetelina Alexiadis, Senya Polikovsky, Daniel Scharstein, and Michael J. Black. Capturing and inferring dense full-body human-scene contact. In *CVPR*, 2022.

Siyuan Huang, Zan Wang, Puhao Li, Baoxiong Jia, Tengyu Liu, Yixin Zhu, Wei Liang, and Song-Chun Zhu. Diffusion-based generation, optimization, and planning in 3D scenes. In *CVPR*, 2023.

Nan Jiang, Zhiyuan Zhang, Hongjie Li, Xiaoxuan Ma, Zan Wang, Yixin Chen, Tengyu Liu, Yixin Zhu, and Siyuan Huang. Scaling up dynamic human-scene interaction modeling. In *CVPR*, 2024.

Hyeonwoo Kim, Sookwan Han, Patrick Kwon, and Hanbyul Joo. Beyond the contact: Discovering comprehensive affordance for 3D objects from pre-trained 2D diffusion models. In *ECCV*, 2024.

Hyeonwoo Kim, Sangwon Baik, and Hanbyul Joo. DAViD: Modeling dynamic affordance of 3D objects using pre-trained video diffusion models. In *ICCV*, 2025.

Lei Li and Angela Dai. GenZI: Zero-shot 3D human-scene interaction generation. In *CVPR*, 2024.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In *ECCV*, 2024.

Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *ICCV*, 2019.

Priyanka Patel, Chun-Hao P. Huang, Joachim Tesch, David T. Hoffmann, Shashank Tripathi, and Michael J. Black. AGORA: Avatars in geography optimized for regression analysis. In *CVPR*, 2021.

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *CVPR*, 2019.

Xavier Puig, Eric Undersander, Andrew Szot, Mikael Dallaire Cote, Dhruv Batra, and Roozbeh Mottaghi. Habitat 3.0: A co-habitat for humans, avatars and robots. In *ICLR*, 2024.

Santhosh Kumar Ramakrishnan, Aaron Gokaslan, Erik Wijmans, Oleksandr Maksymets, Alexander Clegg, John M. Turner, Eric Undersander, Wojciech Galuba, Andrew Westbury, Angel X. Chang, Manolis Savva, Yili Zhao, and Dhruv Batra. Habitat-Matterport 3D dataset (HM3D): 1000 large-scale 3D environments for embodied AI. In *NeurIPS Datasets and Benchmarks Track*, 2021.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. SAM 2: Segment anything in images and videos. In *ICLR*, 2025.

István Sárándi and Gerard Pons-Moll. Neural localizer fields for continuous 3D human pose and shape estimation. In *NeurIPS*, 2024.

Manolis Savva, Angel X. Chang, Pat Hanrahan, Matthew Fisher, and Matthias Nießner. PiGraphs: Learning interaction snapshots from observations. *ACM Transactions on Graphics (TOG)*, 35(4), 2016.

Shashank Tripathi, Agniv Chatterjee, Jean-Claude Passy, Hongwei Yi, Dimitrios Tzionas, and Michael J. Black. DECO: Dense estimation of 3D human-scene contact in the wild. In *ICCV*, 2023a.

Shashank Tripathi, Lea Müller, Chun-Hao P. Huang, Omid Taheri, Michael J. Black, and Dimitrios Tzionas. 3D human pose estimation via intuitive physics. In *CVPR*, 2023b.

Ruicheng Wang, Sicheng Xu, Yue Dong, Yu Deng, Jianfeng Xiang, Zelong Lv, Guangzhong Sun, Xin Tong, and Jiaolong Yang. MoGe-2: Accurate monocular geometry with metric scale and sharp details. In *NeurIPS*, 2025.

Tai Wang, Xiaohan Mao, Chenming Zhu, Runsen Xu, Ruiyuan Lyu, Peisen Li, Xiao Chen, Wenwei Zhang, Kai Chen, Tianfan Xue, et al. EmbodiedScan: A holistic multi-modal 3D perception suite towards embodied AI. In *CVPR*, 2024.

Zan Wang, Yixin Chen, Tengyu Liu, Yixin Zhu, Wei Liang, and Siyuan Huang. HUMANISE: Language-conditioned human motion generation in 3D scenes. In *NeurIPS*, 2022.

Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. ViTPose: Simple vision transformer baselines for human pose estimation. In *NeurIPS*, 2022.

Pradyumna Yalandur Muralidhar, Yuxuan Xue, Xianghui Xie, Margaret Kostyrko, and Gerard Pons-Moll. PhySIC: Physically plausible 3D human-scene interaction and contact from a single image. In *SIGGRAPH Asia*, 2025.

Pradyumna YM, Yuxuan Xue, Yue Chen, Nikita Kister, István Sárándi, and Gerard Pons-Moll. GRAFT: Geometric refinement and fitting transformer for human scene reconstruction. In *ECCV*, 2026.

Siwei Zhang, Yan Zhang, Qianli Ma, Michael J. Black, and Siyu Tang. PLACE: Proximity learning of articulation and contact in 3D environments. In *3DV*, 2020.

Xiaohan Zhang, Bharat Lal Bhatnagar, Vladimir Guzov, Sebastian Starke, and Gerard Pons-Moll. COUCH: Towards controllable human-chair interactions. In *ECCV*, 2022.

Haoyu Zhen, Xiaowen Qiu, Peihao Chen, Jincheng Yang, Xin Yan, Yilun Du, Yining Hong, and Chuang Gan. 3D-VLA: A 3D vision-language-action generative world model. In *ICML*, 2024.

A Appendix Summary

In this appendix we provide additional implementation details of InHabit in Sec. C and discuss limitations in Sec. B. In Sec. D we ablate the use of the complete 3D scene during the lifting step. In Sec. E we ablate the explicit scale parameter in the lifting stage. In Sec. F we ablate the interaction and scene diversity of the training data. In Sec. G we compare the physical stability of *InHabitants* poses with existing datasets. We show more qualitative results in Sec. H. Finally, in Sec. I we show failure cases of InHabit and how they can be addressed. We also provide a video showcasing the population of a scene and a webpage for interactive visualization of our generations.

B Limitations

In this work we create scene interaction in 3D by first generating them in 2D. Our goal was to create a dataset of 3D scenes with corresponding realistic images. To make this tractable, we try to limit the scene changes the image generative models introduce to the image and use filtering as a post-processing step. This discards a lot of useful interactions, e.g., opening shelves, moving chairs. Extending InHabit to model changes to the scene is an interesting future direction and would be a valuable addition to the dataset. Furthermore, we focus on static interactions, so our people are frozen in time, but extending the dataset with dynamic interactions and moving people would create valuable data for applications that need to perceive people in scenes.

C Implementation Details

In this section we provide additional implementation details about view selection, prompting, reconstruction, the user study interface and quality control.

C.1 Automatic View Selection Heuristic

The goal of this heuristic is to select views that show the floor and see sufficiently deep into the scene. Therefore, two regions of the image are important: the bottom strip and the central part. We define two masks, M_c and M_b. The first, M_c, covers a circular region in the center; its objective is to reject views with shallow depth. The second, M_b, covers a strip at the bottom of the image to reject views where the floor is occluded. In both cases, we compare the depth within these masks against thresholds τ_c and τ_b, respectively. A rejection score S_i is calculated based on the proportion of pixels below these thresholds.

More formally:

$$S_i = \frac{1}{|M_i|} \sum_{p \in M_i} \mathbf{1}_{D(p) < \tau_i}, \quad i \in \{c, b\} \tag{2}$$

where $\mathbf{1}$ is the indicator function and $|M_i|$ is the pixel count of the mask. A candidate view is then rejected if the score for either region exceeds a predefined tolerance ρ_i:

$$\text{Reject if: } (S_c > \rho_c) \vee (S_b > \rho_b) \tag{3}$$

We set the radius of the circle to a quarter of the image height, $\tau_c = 2\,\text{m}$, τ_b is set to the height of the camera, $\rho_c = 0.5$ and $\rho_b = 0.3$.

C.2 Prompting

Our affordance driven prompt is: "*Give a list of human actions possible within this scene. Only suggest interactions that involve dynamic interactions and or touch. Avoid only standing or walking. Only consider visible elements that are closer to the camera. First think about the scene then generate the list. Each action should include a description of the object and also the description of the body posture. Combine them, do not split them up between different keys. Generate 4 responses*

with their corresponding probabilities, sampled from the full distribution. Return only VALID JSON. Use only the keys 'action' and 'probability'. Separate the json part from the rest with BEGIN."

For our action-driven prompt we first sample three actions from a list of selected actions and include them in the following prompt: "*Use these actions: 'ACTION-1', 'ACTION-2', 'ACTION-3' and suggest what an action in the scene depicted in the image that involve each of these actions. Only involve visible elements that are closer to the camera. First think about the scene then generate the list. Each action should include one of the given words, a description of the object and also the description of the body posture. Combine them, do not split them up between different keys. Generate 3 responses with their corresponding probabilities, sampled from the full distribution, each response should use only one of the given verbs. Return only VALID JSON. Use only the keys 'action' and 'probability'. Separate the json part from the rest with BEGIN.*"

For multi person interactions we use: "*Give a list of human actions possible within this scene. Only suggest interactions that involve dynamic interactions and or touch and multiple people (but at most 4). Avoid only standing or walking. Only consider visible elements that are closer to the camera. First think about the scene then generate the list. Each action should include a description of the object and also the description of the body posture, and the number of people. Combine them, do not split them up between different keys. Generate 2 responses with their corresponding probabilities, sampled from the full distribution. Each response should have only a single interaction. Return only VALID JSON. Use only the keys 'action' and 'probability'. Separate the json part from the rest with BEGIN.*"

We extract the interactions from the JSON string and use them within the following prompt template as the input to the image editing model: "*Using this provided image of this scene please add* {*Interaction*}*. Keep the rest of the room unchanged. Do not move any furniture. Keep everything closed. Do not add any other new objects besides what is explicitly requested.*"

C.3 Reconstruction Optimization

Off-the-shelf components. Humans in the edited image are segmented with SAM 2 (Ravi et al., 2025) prompted by Grounding DINO (Liu et al., 2024) detections (the LangSAM implementation). 2D keypoints are estimated with ViTPose (Xu et al., 2022), the initial SMPL-X body with NLF (Sárándi & Pons-Moll, 2024), and depth with MoGe-2 (Wang et al., 2025).

Our lifting objective follows standard physically grounded human–scene fitting, similar to PhySIC (Yalandur Muralidhar et al., 2025) and PROX (Hassan et al., 2019). We optimize pose $\boldsymbol{\theta}$, shape $\boldsymbol{\beta}$, camera-relative translation $\mathbf{t}$, and scale s:

$$\begin{aligned}\min_{\boldsymbol{\theta},\boldsymbol{\beta},\mathbf{t},s} \quad & \lambda_{\text{proj}}\mathcal{L}_{\text{proj}} + \lambda_{\text{depth}}\mathcal{L}_{\text{depth}} + \lambda_{\text{con}}\mathcal{L}_{\text{contact}} \\ & + \lambda_{\text{pen}}\mathcal{L}_{\text{penetration}} + \lambda_{\text{reg}}\mathcal{L}_{\text{reg}}.\end{aligned} \tag{4}$$

Let $\mathcal{M}(\boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{t}, s)$ denote the posed SMPL-X mesh (Pavlakos et al., 2019), let $\phi(\cdot)$ denote the signed distance field of the scene, let $\mathcal{C}$ be the set of contact-prone body vertices obtained from the DECO contact labels, and let $\mathcal{P}_{\text{depth}}$ be the partial human point cloud obtained from the estimated human depth map. We denote by $\mathcal{V}_{\text{vis}}$ the subset of mesh vertices visible under ray casting, and by $\mathbf{J}$ and $\widehat{\mathbf{J}}$ the projected and detected 2D joints, respectively.

We define the losses as follows. The contact loss encourages contact-prone vertices to lie close to the scene surface,

$$\mathcal{L}_{\text{contact}} = \frac{1}{|\mathcal{C}|}\sum_{\mathbf{v}\in\mathcal{C}} \left|\phi(\mathbf{v})\right|, \tag{5}$$

while the penetration loss penalizes vertices that lie inside the scene,

$$\mathcal{L}_{\text{penetration}} = \frac{1}{|\mathcal{V}|}\sum_{\mathbf{v}\in\mathcal{V}} \left|\min\left(\phi(\mathbf{v}), 0\right)\right|, \tag{6}$$

where $\mathcal{V}$ denotes all mesh vertices. Both terms use an L_1 penalty on the scene SDF.

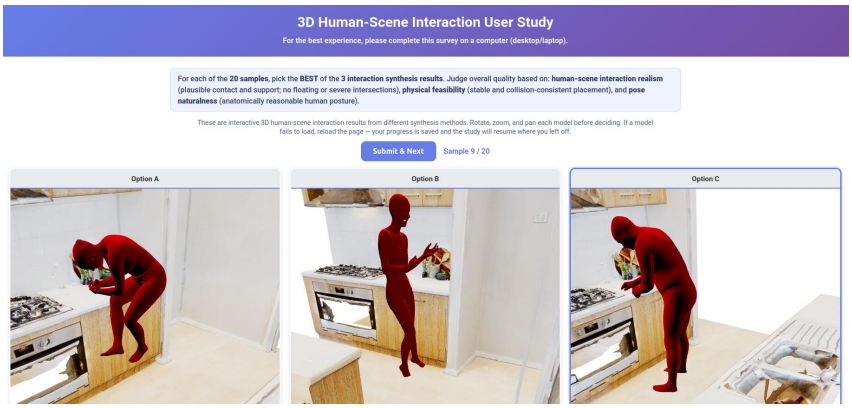

Figure 8: User interface for our user study. The camera automatically orbits around the center of the human bounding box. Additionally, participants can rotate, translate and zoom each scene individually.



Figure 9: Filtered examples from our quality-control stage. We discard generations that provide too little visible evidence for reliable lifting, alter nearby scene content, or lead to distorted body fits.

To align the visible body surface with the human depth estimate, we use a standard Chamfer distance between the ray-casting-visible mesh vertices and the depth-derived point cloud:

$$\mathcal{L}_{\text{depth}} = \text{CD}(\mathcal{V}_{\text{vis}}, \mathcal{P}_{\text{depth}}). \tag{7}$$

For the 2D term, we project joints into normalized pixel coordinates in $[0, 1]$ and penalize reprojection error as

$$\mathcal{L}_{\text{proj}} = \frac{1}{|\mathbf{J}|} \sum_i \left\| \mathbf{J}_i - \widehat{\mathbf{J}}_i \right\|_2^2. \tag{8}$$

Finally, we regularize the optimized parameters toward their initialization $(\boldsymbol{\theta}_0, \boldsymbol{\beta}_0, \mathbf{t}_0, s_0)$ using an L_2 penalty:

$$\mathcal{L}_{\text{reg}} = \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_2^2 + \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_2^2 + \|\mathbf{t} - \mathbf{t}_0\|_2^2 + |s - s_0|^2. \tag{9}$$

In all experiments, we use $\lambda_{\text{con}} = 0.7$, $\lambda_{\text{pen}} = 1.1$, $\lambda_{\text{depth}} = 0.005$, $\lambda_{\text{reg}} = 13$, and $\lambda_{\text{proj}} = 3000$.

C.4 User Study

We present the user interface of our user study in Fig. 8.

C.5 Quality Control

Some edited images are not suitable for lifting into clean 3D human–scene interactions. We therefore apply three complementary post-reconstruction filters before accepting a sample into the dataset: 1) a *visibility filter* that rejects strongly truncated insertions, 2) a *depth-based filter* that rejects edits which alter local scene structure, and 3) a *volume filter* that rejects distorted SMPL-X fits. We show examples in Fig. 9.

Visibility Filter. The editing model can place humans in reflective surfaces (e.g., mirrors or glass) without a corresponding physical person in the scene, or produce severely truncated insertions (e.g., only a hand). To detect these cases, let $\mathcal{H}_i$ denote the reconstructed human mesh for human i in an image. We uniformly sample a set of $N = 20{,}000$ points $\{\mathbf{p}_j\}_{j=1}^{N}$ on $\mathcal{H}_i$ and project them onto the image plane using the camera projection $\pi(\cdot)$. Given a human segmentation mask M_i in the image, we collect the *mask-consistent* points whose projections fall inside M_i:

$$V_i = \{\mathbf{p}_j : \pi(\mathbf{p}_j) \in M_i\},$$

and define the *visibility ratio* as

$$\mathrm{vis}(\mathcal{H}_i) = \frac{|V_i|}{N}.$$

Humans with $\mathrm{vis}(\mathcal{H}_i) < \tau_{\mathrm{vis}} = 0.5$ are discarded. If no humans remain in an image after this step, the image is dropped entirely, indicating the editing model failed to produce a valid insertion.

Depth-Based Filter. The editing model can also introduce unintended local geometry changes near the inserted human, such as hallucinated or displaced furniture. To detect these cases, let D_{edit} denote the monocular depth map estimated from the edited image, aligned to metric scale using the known scene geometry, and let $M = \bigcup_i M_i$ be the combined binary mask of all humans. We define a thin boundary region around the humans as $M^{\mathrm{b}} = \mathrm{dilate}(M) - M$, where the dilation is performed using OpenCV with a 5×5 kernel for 3 iterations and then a 9×9 kernel for another 3 iterations. We focus on this near-human band because edits are typically local to inserted people; distant regions can be preserved by compositing the near-human edited patch back into the original image. For each pixel $\mathbf{q}$ in this boundary, we compare D_{edit} against the known scene depth D_{scene} and compute the fraction of pixels whose *relative* depth discrepancy (normalized by the edited-image boundary depth, so the criterion measures a relative mismatch rather than an absolute metric offset) exceeds a per-pixel tolerance $\tau_{\mathrm{depth}} = 0.2$:

$$\Delta = \frac{\left|\left\{\mathbf{q} \in M^{\mathrm{b}} : \frac{|D_{\text{edit}}(\mathbf{q}) - D_{\text{scene}}(\mathbf{q})|}{D_{\text{edit}}(\mathbf{q})} > \tau_{\mathrm{depth}}\right\}\right|}{|M^{\mathrm{b}}|}.$$

If the violation ratio Δ exceeds an image-level threshold $\tau_{\mathrm{border}} = 0.35$, the reconstruction is discarded. This operates as a joint, image-level filter, since geometry edits can extend beyond any single person's boundary.

Volume Filter. The image editing model can occasionally insert children, which fall outside the adult body model we employ (SMPL-X), and the human detector can misfire on depictions of people in portraits, posters, or sculptures. In both cases, fitting an adult SMPL-X mesh to these detections produces highly deformed body shapes: either unnaturally compressed to match a child's proportions or distorted to conform to a non-human surface. We observe that these failure modes consistently yield meshes with anomalously low body volume. Motivated by this, we compute the volume $\mathcal{V}(\mathcal{H}_i)$ of each reconstructed SMPL-X mesh $\mathcal{H}_i$ and discard any instance with $\mathcal{V}(\mathcal{H}_i) < \tau_{\mathrm{vol}} = 0.0375\,\mathrm{m}^3$. Despite its simplicity, this filter is highly effective: it removes the vast majority of problematic cases and, in practice, performs comparably to more expensive alternatives such as dedicated human detectors (e.g., Multi-HMR (Baradel et al., 2024)) or pose-prior-based filtering (e.g., VPoser (Pavlakos et al., 2019)).

Together, these three filters ensure that the final dataset contains only humans that are sufficiently visible, properly aligned, plausibly shaped, and do not introduce implausible scene changes.

C.6 Runtime and Scalability

InHabit is designed for dataset-scale inference. All pre-optimization stages—segmentation, 2D keypoints, 3D pose, contact prediction, and depth—are batched across images and across humans within an image, while the optimization stage processes all humans in a given image in a single forward pass. Each worker fits within 5–8 GB of VRAM, so the pipeline runs on consumer-grade GPUs (we use 32 RTX 2080s on a shared cluster). Lifting takes ≈25 GPU-seconds per image, with image generation and rendering adding 3–4 s. Interaction proposals and image edits are obtained through the Gemini API, using Gemini 2.5 Flash (Google DeepMind, 2025a) both to propose interactions and,

Table 6: **Ablation of the reconstruction.** Contact-based Precision, Recall, and F1 of InHabit's lifting stage when ablating 1) the explicit human-scale parameter (*w/o human scale*), and 2) the use of complete scene geometry, falling back to partial depth (*w/o gt scene*). Since generated images have no ground-truth contacts, we evaluate against coarse, region-level contact labels derived from a VLM; the values are therefore meaningful relative to each other rather than in absolute terms. Both components contribute to final reconstruction quality.

Reconstruction	Prec. ↑	Rec. ↑	F1 ↑
InHabit (Ours)	**0.424**	**0.393**	**0.368**
w/o human scale	0.396	0.359	0.334
w/o gt scene	0.354	0.337	0.297

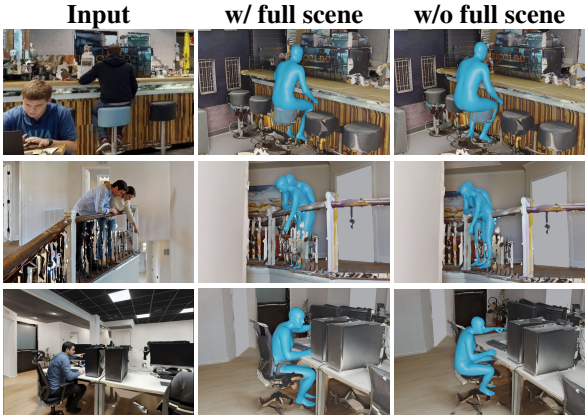

Figure 10: Ablation of complete-scene versus partial-geometry lifting. Starting from the same edited image, we run lifting either with the complete scene mesh or with partial depth only. When hidden scene geometry is unavailable, the signed-distance field becomes ambiguous in occluded regions, leading to misplaced support and incorrect contacts. Using the full mesh provides consistent geometric constraints and yields cleaner reconstructions.

in its image variant (Gemini 2.5 Flash Image), to edit the images. Reconstructing the full ∼78k-image dataset takes ≈19 h of wall-clock time. The low memory footprint and horizontal scaling let InHabit scale on cheap, abundant hardware.

D Full Scene vs Partial Depth

We ablate the use of the complete 3D scene during the lifting step against a variant that relies only on partial depth maps. We present the results in Fig. 10. When only partial depth is available, occluded scene surfaces are absent from the geometry used during optimization. This makes the signed-distance field incomplete exactly where it is needed most, so the contact and penetration terms can guide the body toward incorrect placements, as seen for the chairs (rows 1 and 3) and the railing (row 2). Using the full scene mesh restores those missing constraints and lets the optimizer reason about nearby but unseen support surfaces. This reduces interpenetration and yields more faithful contact placement, especially in challenging cases such as the chair in row 3. Quantitatively, this is reflected in Tab. 6: removing the complete scene (*w/o gt scene*) drops contact F1 from 0.368 to 0.297, with corresponding decreases in precision (0.424→0.354) and recall (0.393→0.337), confirming that the missing geometric constraints translate directly into worse contact placement.

E Explicit Scale Parameter

We ablate the explicit scale parameter introduced in the lifting stage and compare it to a variant that keeps scale fixed. We present the results in Fig. 11. Without a dedicated scale variable, any mismatch between the pose estimator's scale and the scene must be compensated by the remaining optimization variables. In practice this is awkward: the SMPL-X shape coefficients β affect the body nonlinearly and change its outline, while translation has to move in tandem to preserve reprojection agreement. That coupling makes the correct metric size hard to recover. Allowing the optimizer to

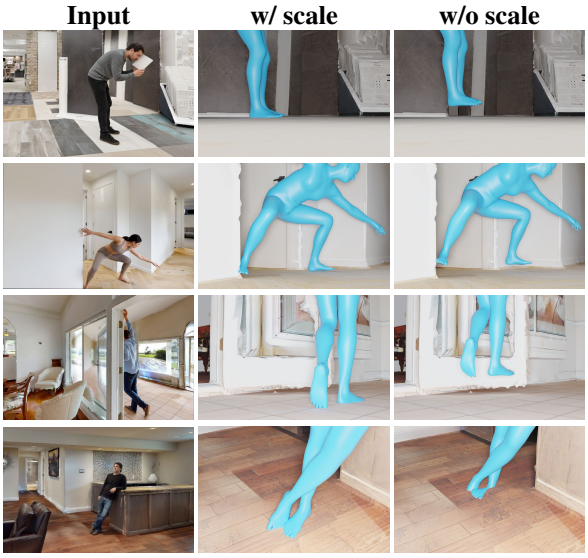

Figure 11: Ablation of the scale variable used during lifting. We apply the same procedure to the edited image, once optimizing the scale term s and once keeping scale fixed. Without a dedicated scale variable, metric mismatch must be absorbed indirectly through coupled changes in body shape and position while still matching the image silhouette. Optimizing s provides a direct correction for that mismatch and improves grounding, support, and contact with the scene.

Table 7: **Ablation of interaction and scene diversity.** GRAFT (YM et al., 2026) trained on size-matched subsets of *InHabitants* (∼22k images each; scene-view subsets with 2,110, 4,219, and 8,438 views) that differ in a single factor, evaluated on PROX Quant.

	Interactions		Scene views			Captured
	feet-only	all	25%	50%	100%	PROX
PROX F1 ↑	0.468	**0.577**	0.546	0.551	**0.578**	0.492

adjust s directly separates scene-level sizing from body shape, making the fit more stable. Qualitatively, this leads to better floor support, more accurate placement relative to nearby surfaces, and fewer floating or interpenetrating bodies. This effect is also visible in Tab. 6: removing the scale parameter (*w/o human scale*) drops contact F1 from 0.368 to 0.334 (precision 0.424→0.396, recall 0.393→0.359), showing that the indirect compensation through β and translation is unable to recover the same quality of contact placement.

F Ablation of Interaction and Scene Diversity

To understand where the downstream gains come from, we train GRAFT on subsets of *InHabitants* that are matched in size (∼22k images each) and differ in a single factor: interaction diversity (feet-only contacts vs. all interactions) or scene diversity (25%, 50%, or 100% of all views). As shown in Tab. 7, interaction diversity is the dominant factor: at equal size, F1 on PROX rises from 0.468 to 0.577. Scene diversity contributes a smaller but consistent improvement (0.546 → 0.578 from 25% to 100%). Note that even our smallest subset spans 2,110 scene views, an order of magnitude beyond captured datasets, so these gains are measured in a regime of scene coverage that capture cannot reach. Notably, the feet-only subset underperforms training on captured PROX data (0.492), while the diverse subset clearly surpasses it, despite containing only about a quarter as many training images (∼22k vs. ∼88k). This indicates that the benefit of *InHabitants* stems mainly from the breadth of interactions it contains. The same trend holds against TRUMANS (Tab. 4), which contains far more frames than *InHabitants* but narrower interactions. Compared with Tab. 2, GRAFT trained on BEDLAM (0.396) or on our feet-only subset (0.468) performs worse than GRAFT trained on PROX (0.492), whereas training on our subset with all interactions exceeds it (0.577).

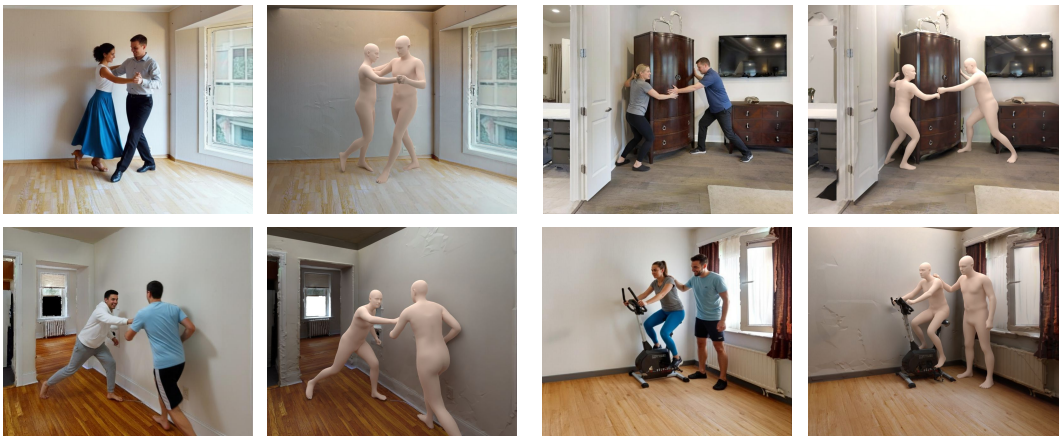

Figure 12: InHabit can naturally generate interactions involving multiple people, which was not supported by prior work.

G Physical Stability of *InHabitants*

Table 8: **Physical stability of dataset poses.** Percentage of poses whose center of mass projects inside the base of support, following IPMAN (Tripathi et al., 2023b).

Dataset	Stable poses (%) ↑
PROX-D	65.40
TRUMANS	75.88
InHabitants (Ours)	**83.21**

Our lifting stage (Sec. 3.2) does not explicitly enforce physical balance, so we verify that the resulting poses are physically stable. Following IPMAN (Tripathi et al., 2023b), we consider a pose stable if the projection of its center of mass onto the ground plane lies within its base of support (BoS). The BoS is the convex hull of the body's contact points, projected onto the ground plane, where contact points are obtained with a simple distance threshold to the scene geometry. As shown in Tab. 8, 83.21% of *InHabitants* poses are stable, compared to 75.88% for TRUMANS and 65.40% for PROX-D.

H Additional Qualitative Results

Fig. 12 shows interactions involving several people in the same scene, which InHabit produces from a single prompt to the image-editing model. Fig. 13 compares InHabit with GenZI and POSA++ on further scenes and prompts.

We show more examples from our dataset *InHabitants* in Figs. 14 to 16. Our data contains diverse scene interactions such as riding a gym bike (Fig. 16), looking for a clothing item (Fig. 14), measuring distances on a map (Fig. 15) and playing pinball (Fig. 15).

I Failure Cases

Fig. 17 shows representative failure cases of InHabit. Many originate in the image editor: lifting faithfully reconstructs what the edited image shows, including a person without physical support (a). The other cases are local, and most are handled by the quality control of Sec. C.5. The visibility filter removes most strongly truncated insertions; rare cases such as (c) pass it, and more aggressive filtering can remove them. Scene changes far from the person (d) can be removed by compositing only the edited patch into the rendered scene, while changes close to the person are mostly caught by the depth-based filter. When the edit hides a body part (b), the lifted body can penetrate the scene by a few centimeters, and holes in the scanned meshes (e) stem from the source scans. Since InHabit is modular, its quality will improve with advances in reconstruction and image generation models.

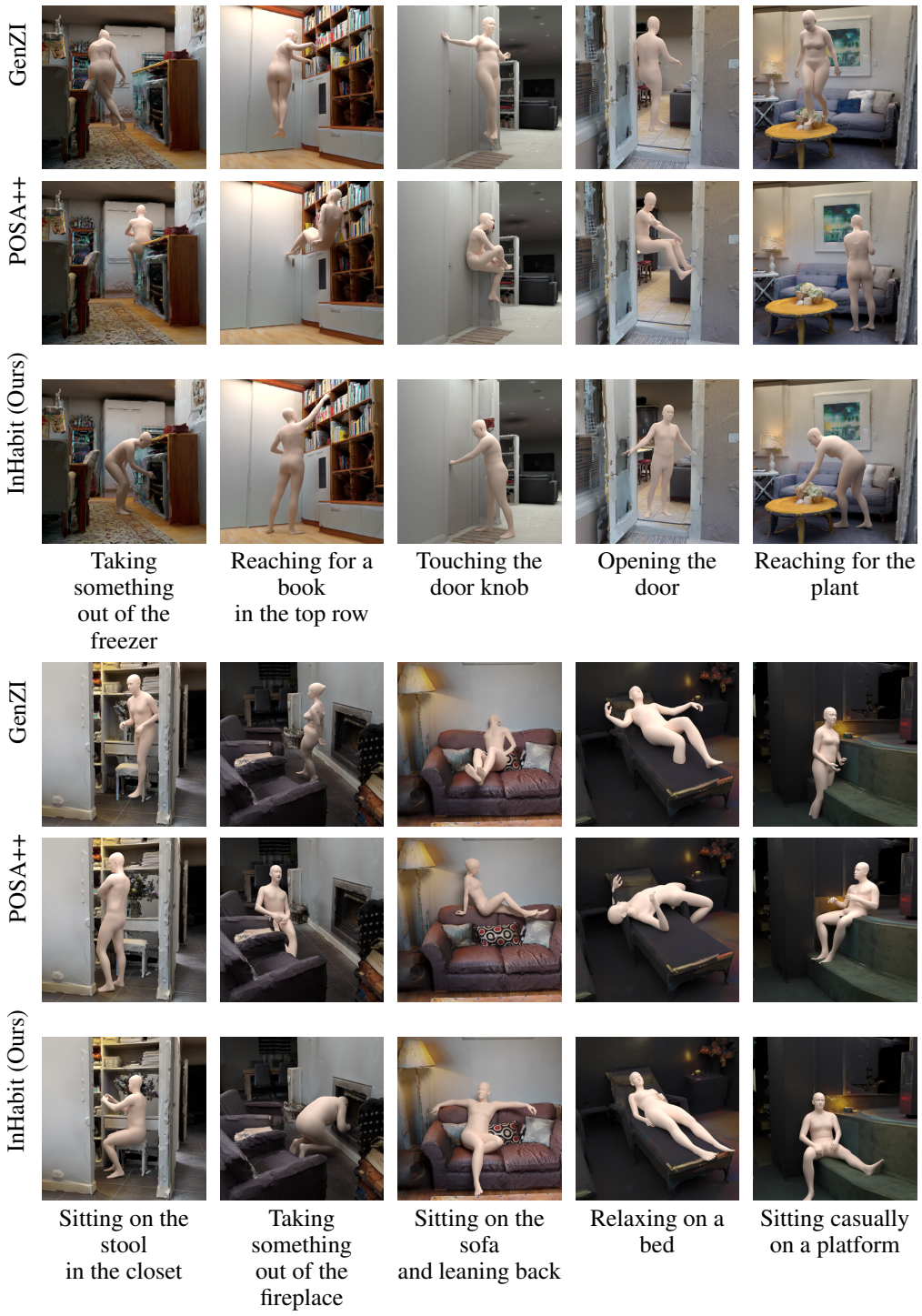

Figure 13: Additional qualitative comparisons of InHabit (Ours) with GenZI (Li & Dai, 2024) and POSA++ (Hassan et al., 2021). Both baselines fail at physical plausibility (meshes stick to walls, feet do not touch the floor) and meaningful interaction (e.g., reaching for a book, touching a door handle, reaching for flowers), while InHabit places people meaningfully without deep scene penetrations or floating bodies.

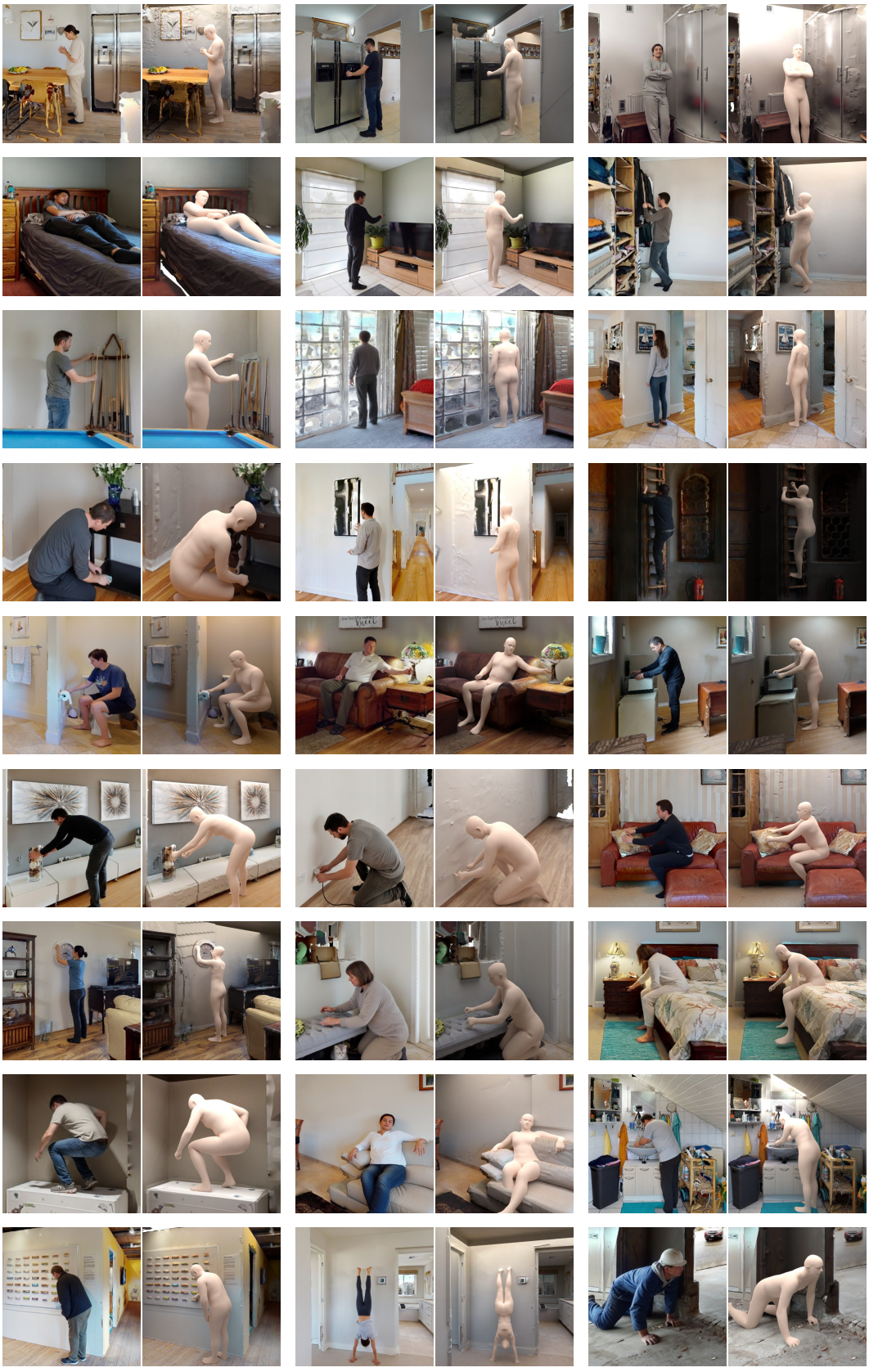

Figure 14: Additional samples from *InHabitants*. Each pair shows the edited RGB image (left) and the corresponding lifted 3D human in the 3D scene (right).



Figure 15: Additional samples from *InHabitants*. Each pair shows the edited RGB image (left) and the corresponding lifted 3D human in the 3D scene (right).



Figure 16: Additional samples from *InHabitants*. Each pair shows the edited RGB image (left) and the corresponding lifted 3D human in the 3D scene (right).

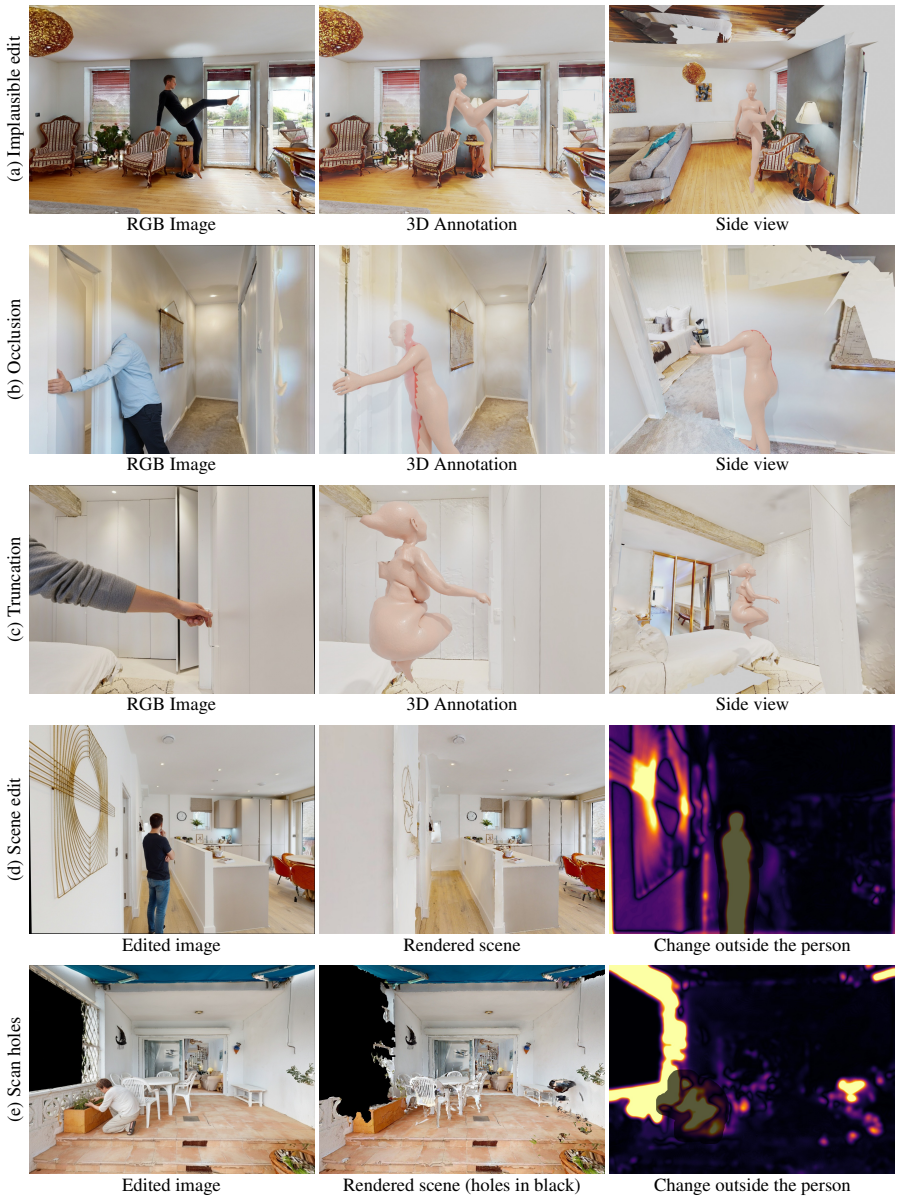

Figure 17: **Failure cases.** Many of these failure modes originate in the image editor and are expected to diminish as editing models improve. (a) The editor produces a person without physical support; lifting faithfully reproduces the floating pose. (b) In the edit the head is hidden behind a door frame; without evidence for it, the lifted body is embedded $\sim$7 cm in the wall (input view: occluded body parts shown semi-transparent, penetration in red). (c) Only an arm is visible; the lifted body is a crouched figure floating in front of the camera. In rare cases such as this one, a truncated insertion passes the visibility filter; more aggressive filtering can remove it. (d) The editor alters the scene far from the person; geometry changes near the person are mostly detected by our depth-based filter. (e) The editor paints over holes in the scanned mesh; a more complete scene dataset mitigates this.