Supplementary for Compositional Human-Scene Interaction Synthesis with Semantic Control

Kaifeng Zhao[1], Shaofei Wang[1], Yan Zhang[1], Thabo Beeler[2], and Siyu Tang[1]

[1] ETH Zürich
{kaifeng.zhao, shaofei.wang, yan.zhang, siyu.tang}@inf.ethz.ch
[2] Google
thabo.beeler@gmail.com

In the supplementary, we first provide method details, including architecture illustrations, training losses, and compositional body generation in Sec. A. We then elaborate on experiment details in Sec. B. In Sec. C, we present ablation studies on different body representations and the two-stage generation framework. In Sec. D, we show more qualitative results and discuss typical failure cases and limitations.

A Method Details

A.1 Architecture Details

PelvisVAE. We illustrate the detailed architecture of PelvisVAE in Fig. S1. The PelvisVAE encoder and decoder use a stack of 2 transformer layers with an embedding dimension of 64. PelvisVAE represents a human as a pelvis frame of location and orientation. At the decoder, PelvisVAE takes a zero vector as the body token.

PointNet++. We train a PointNet++ [12] network to extract sparse key points from point clouds, which reduces the number of object nodes to a suitable level for transformers. We show the architecture of the PointNet++ object encoder in Fig. S2. The PointNet++ module comprises two set abstraction layers to extract 256 key points for each object. The output points features are projected to vectors of dimension 128 for BodyVAE and dimension 64 for PelvisVAE using a linear layer.

SMPL-X Regressor We train a multi-layer perceptron (MLP) to regress SMPL-X parameters from body meshes as shown in Fig. S3. The MLP has six linear layers, and we employ residual connections. The MLP outputs the SMPL-X body poses using the 6D continuous representation [18], body shape parameters, and the first six hand pose PCA components for each hand. Following [8], We detach the gradients of the reconstructed body from the computational graph and use the concatenation of the reconstructed body and a template T-pose body as inputs to the MLP.

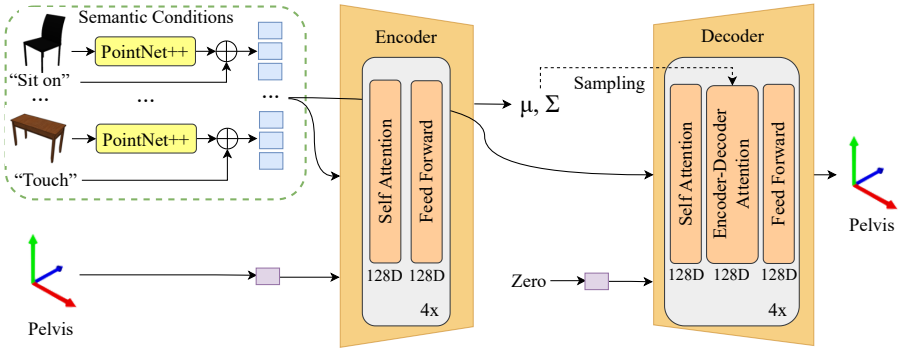


Fig. S1: Detailed architecture of PelvisVAE.

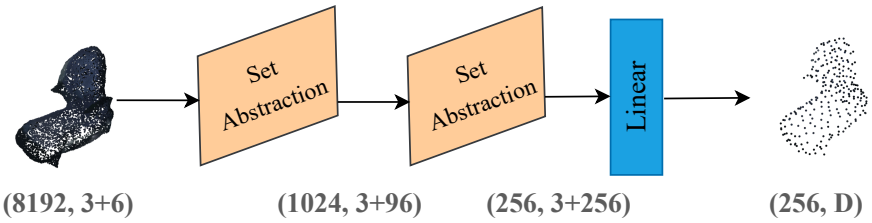


Fig. S2: The architecture of the PointNet++ object encoder, where D denotes the embedding dimension used by transformers.

A.2 Training Loss

Our BodyVAE is trained to minimize the $\mathcal{L}_{body}$ loss:

$$\mathcal{L}_{body} = \mathcal{L}_{interaction} + \mathcal{L}_{mesh} + \mathcal{L}_{feature} + \mathcal{L}_{KL} + \mathcal{L}_{regress}, \tag{1}$$

where the terms represent the interaction loss, body mesh reconstruction loss, contact feature reconstruction loss, the Kullback-Leibler divergence, and the SMPL-X regression loss. Weights for each term are omitted for simplicity. The term $\mathcal{L}_{interaction}$ encourages the vertices with predicted positive contact feature to have zero distance to the input objects:

$$\mathcal{L}_{interaction} = \sum_{i=1}^{655} \hat{c}^i \cdot \min_{v_o \in V_o} ||\hat{v}^i - v_o||_2, \tag{2}$$

where $\hat{c}^i$ and $\hat{v}^i$ denote the predicted contact feature and location of body vertex i respectively, and $V_o = \bigcup_{i=1}^{M} o^i$ denotes the set of all points of input interaction objects.

The body mesh reconstruction loss consists of the vertex coordinate loss $\mathcal{L}_{vertex}$, the surface normal loss $\mathcal{L}_{normal}$, the edge length loss $\mathcal{L}_{edge}$ following

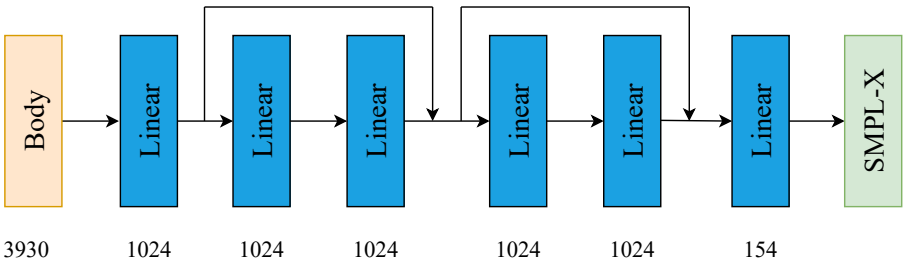


Fig. S3: The architecture of the SMPL-X regressor.

[4, 13], and the normal consistency loss $\mathcal{L}_{consistency}$ that regularizes the normal of adjacent faces to change smoothly, which helps in particular with details of the hands. The loss terms are defined by:

$$\mathcal{L}_{mesh} = \mathcal{L}_{vertex} + \mathcal{L}_{normal} + \mathcal{L}_{edge} + \mathcal{L}_{consistency}. \tag{3}$$

$$\mathcal{L}_{vertex} = ||\hat{V} - V||_1, \tag{4}$$

$$\mathcal{L}_{normal} = \sum_{f \in F} \sum_{(i,j) \in f} \left| \langle n_f, \frac{\hat{v^i} - \hat{v^j}}{||\hat{v^i} - \hat{v^j}||_2} \rangle \right|, \tag{5}$$

$$\mathcal{L}_{edge} = \sum_{f \in F} \sum_{(i,j) \in f} \left| 1 - \frac{||\hat{v^i} - \hat{v^j}||_2}{||v^i - v^j||_2} \right|, \tag{6}$$

$$\mathcal{L}_{consistency} = \sum_{f^i, f^j \in F, f^i \cap f^j \neq \emptyset} 1 - \langle \hat{n}_{f^i}, \hat{n}_{f^j} \rangle, \tag{7}$$

where V, F denote the vertices and faces of input body mesh, n_f denotes the normal of triangle $f \in F$ and $\hat{(\cdot)}$ denotes the corresponding reconstructions.

The contact feature reconstruction loss is calculated as the binary cross-entropy loss (BCE) between reconstructed $\hat{C}$ and ground truth C contact features:

$$\mathcal{L}_{feature} = BCE(\hat{C}, C). \tag{8}$$

We use the robust Kullback-Leibler divergence (KL) [15, 16] to avoid posterior collapse:

$$\mathcal{L}_{KL} = \Psi(KL(q(z|\mathcal{I})||\mathcal{N}(0, I))), \tag{9}$$

where Ψ is the Charbonnier function [2] $\Psi(s) = \sqrt{1 + s^2} - 1$. The SMPL-X parameter regression loss $\mathcal{L}_{regress}$ comprises parameter error and vertex error:

$$\mathcal{L}_{regress} = ||\theta - \hat{\theta}||_2 + |\mathcal{M}(\theta) - \mathcal{M}(\hat{\theta})|, \tag{10}$$

where θ and $\hat{\theta}$ are GT and predicted SMPL-X body parameters respectively, and $\mathcal{M}$ denotes the SMPL-X model mapping parameters to mesh vertices.

The PelvisVAE is trained with the following losses:

$$\mathcal{L}_{pelvis} = \mathcal{L}_{transl} + \mathcal{L}_{orient} + \mathcal{L}_{KL}, \tag{11}$$

where $\mathcal{L}_{transl}$ and $\mathcal{L}_{orient}$ denote the reconstruction loss of pelvis joint location and orientation, and $\mathcal{L}_{KL}$ is the robust Kullback-Leibler divergence loss from Eq. (9).

A.3 Interaction-Based Optimization.

We use the sampled SMPL-X parameters as initialization and optimize body translation t, global orientation R and pose θ. The optimization objective is given by:

$$E(t, R, \theta) = \mathcal{L}_{interaction} + \mathcal{L}_{coll} + \mathcal{L}_{reg}, \tag{12}$$

The interaction term $\mathcal{L}_{interaction}$ is defined in Eq. (2). The scene collision term is defined as $\mathcal{L}_{coll} = \sum_{i=1}^{655} \Psi(v^i)$, with $\Psi(v^i)$ denoting the signed distance of vertex i to the scene. For computational efficiency, we use a precomputed SDF grid for each scene and use interpolation to get SDF value for body vertices. The regularization term $\mathcal{L}_{reg} = |t - t_{init}| + |R - R_{init}| + ||\theta - \theta_{init}||_2$ penalizes t, R and θ deviating from their initialization.

A.4 Implementation Details

Our implementation is based on PyTorch [11]. We use the Adam optimizer [5] with the learning rate 3e-4 and batch size of 8 for training all models. For PelvisVAE, we use weights of 3, 1, 1 for $\{\mathcal{L}_{transl}, \mathcal{L}_{orient}, \mathcal{L}_{KL}\}$. For BodyVAE, we use weights of 1, 1, 0.1, 0.2, 0.05 for $\{\mathcal{L}_{interaction}$, $\mathcal{L}_{vertex}$, $\mathcal{L}_{normal}$, $\mathcal{L}_{edge}$, $\mathcal{L}_{consistency}\}$. We use weight 1 for $\mathcal{L}_{KL}$ and apply the weight annealing scheme [1]. For SMPL-X regressor, we use weights of 1 for vertex and body pose reconstruction and 0.1 for shape parameter and hand PCA components reconstruction. our contact features used a threshold object distance of 5cm. We use latent dimensions of 6 and 128 for PelvisVAE and BodyVAE, respectively.

For the interaction-based optimization, we respectively use weights of 1 for interaction term $\mathcal{L}_{interaction}$, 32 for collision term $\mathcal{L}_{coll}$, 0.1 for translation and orientation regularization and 32 for pose regularization.

For composite pelvis sampling, we use the Adam optimizer with a learning rate of 0.1 and 100 optimization steps. We scale the sum of log probability of latent codes with the weight of 0.05 to balance the influence of pelvis frame difference and probability.

B Experiment Details

B.1 Dataset Collection

We extend the PROX-S dataset based on PROX [7], which contains 3D reconstructions of 12 static scenes and RGB-D recordings of natural human-scene interactions captured using a Kinect sensor. The pseudo-ground truth body fittings of PROX recordings are estimated using [7, 14]. To obtain object instance segmentation and interaction semantics, we further process the PROX dataset to get: (1) 3D instance segmentation in all the PROX scenes and (2) per-frame interaction semantic labels in the form of action-object pairs.

We first conduct instance segmentation for the 12 scenes based on the scene semantic annotation provided in the PROX-E [17] dataset. Specifically, we split the scene into multiple possible over-segmented instances using connected components analysis. Then we manually annotate the instances in the scenes. The instance segmentation results of the 4 test scenes are visualized in Fig. S4.

To obtain interaction semantics, we densely annotated the PROX dataset using the VIA video annotation tool [6]. The annotators are asked to label all intervals containing interactions specified by the action-noun pairs. The annotation tool is illustrated in Fig. S5. Note that multiple interaction labels can exist in a single frame if the human interacts with multiple objects. We further retrieve interaction object instance ID using the scene segmentation and object category annotation. When there is more than one object instance of the annotated category in one scene, we assign the object label to the instance with the closest distance to action-related body parts.

The PROX-S dataset contains around 32K frames of human-scene interactions from 43 sequences recorded in 12 indoor scenes. We follow the dataset split in [17] to use 'MPH16', 'MPH1Library', 'N0SittingBooth', and 'N3OpenArea' as test scenes and the remaining eight scenes for training. The training data comprises a total of 17 different actions and 42 categories of interactions defined as action-object pairs. We evaluate interaction synthesis with semantic control on about 150 different combinations of action and object instances in the four test scenes.

B.2 Perceptual Study

We conduct binary perceptual studies to evaluate the interaction naturalness and unary perceptual studies to evaluate the semantic accuracy of the generated interactions on Amazon Mechanical Turk (AMT). The AMT interfaces of the two perceptual studies are illustrated in Fig. S6. For the binary perceptual studies, we uniformly select 160 samples from PROX pseudo ground truth with varying semantic labels and respectively generate 160 random samples with the same semantic labels using our method and two baselines. We render each interaction with two different views and compare out method against the two baselines and pseudo ground truth. During the study, participants are instructed to select one sample that they think is more realistic from two samples generated with

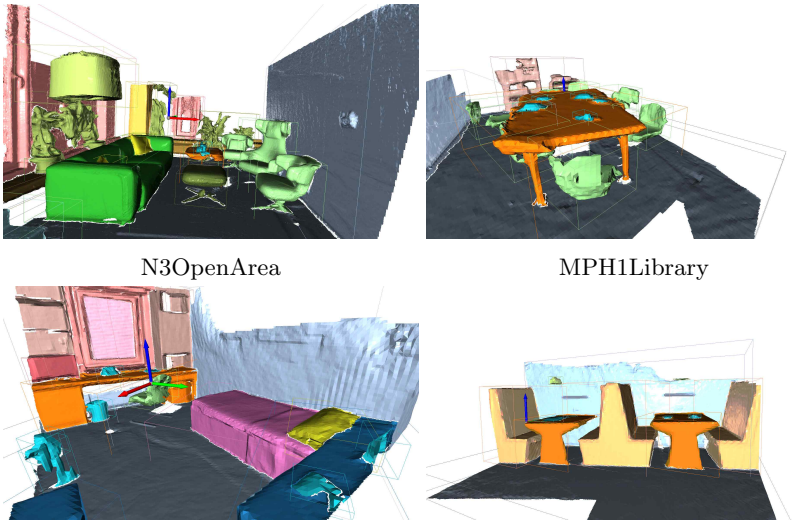


Fig. S4: Visualization of instance segmentation results in the 4 test scenes. The objects are colored according to their semantic categories.

different methods. Each comparison task is distributed to three participants for evaluation.

For the unary perceptual studies, we sample and render one interaction for the 155 combinations of actions and objects in our test scenes. These interaction samples are shown to the participants with the interaction semantic labels and the participants are instructed to rate the semantic accuracy from 1 (strongly disagree) to 5 (strongly agree).

C Ablation Study

We investigate the influence of body representations and the two-stage generation design.

We compare three body representations of joint locations (JL), joints locations and orientations (JLO), and mesh vertices locations (VL) in BodyVAE by evaluating the semantic contact score of sampled interactions without optimization, and the consistency error between generated body and corresponding SMPL-X body defined as:

$$\mathcal{L}_{body} = |\mathbf{B} - \mathcal{M}(\theta, \beta)|, \tag{13}$$

where $\mathbf{B}$ denotes the generated body joints (JL and JLO) or vertices locations(VL), and $\mathcal{M}(\cdot)$ denotes the SMPL-X body model that yields body joints and vertices given body pose θ and shape parameter β. We use the pose and

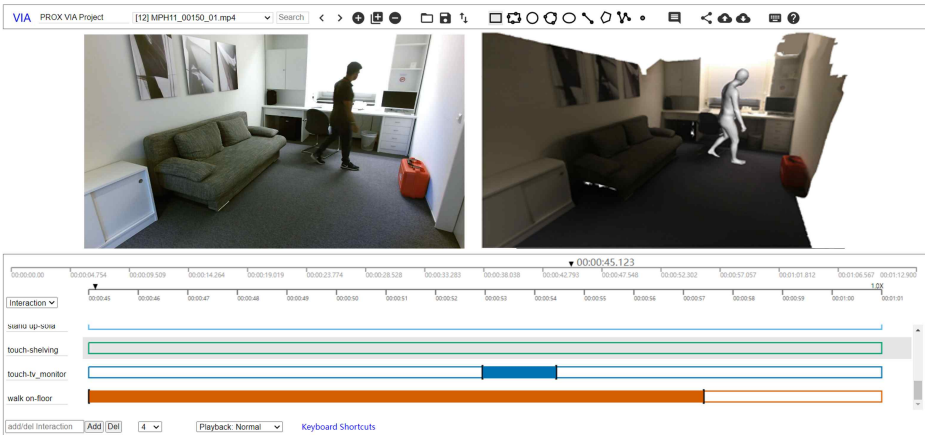


Fig. S5: Our annotation tool for annotating interaction semantics. RGB recordings and the visualization of SMPL-X fitting are displayed side-by-side. Annotators are instructed to label the intervals containing interactions using action-noun pairs.

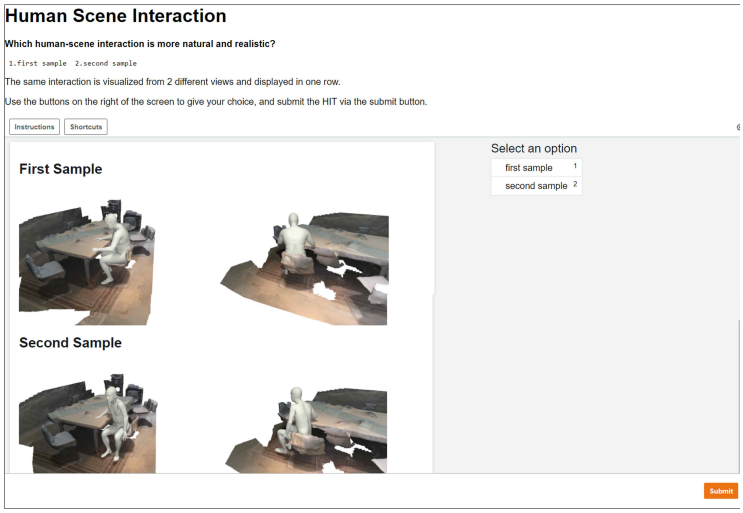
Table 1: Evaluation of body representation choices.

	Semantic Contact ↑	Body Consistency (m) ↓
Vertex Location	**0.72**	**0.01**
Joint Location	0.71	0.02
Joint Location and Rotation	0.69	0.04

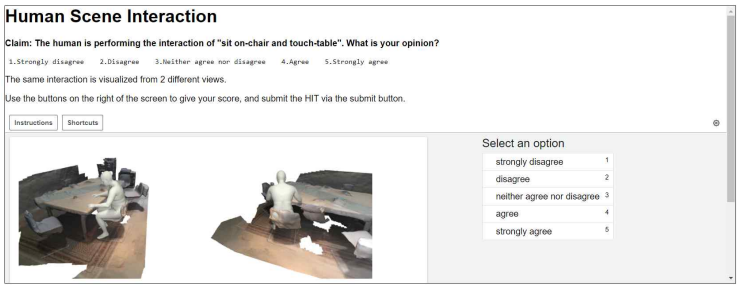
shape parameters predicted by the SMPL-X regressor for JL and VL. We do not train a regressor for JLO and directly use the input shape parameter of the template body and the generated joint orientations to pose the SMPL-X body.

Directly generating joint locations with orientations in JLO leads to a lower semantic contact score of 0.69 and a significantly worse body consistency error of 0.04m. This is because generating joint locations and orientations together by networks does not directly yield valid SMPL-X bodies, as the bone lengths defined by the generated joint locations may not correspond to valid human skeletons. Regressing joint rotations from joint or vertex locations leads to better performance and vertex location representation achieves the best semantic contact score of 0.72 and body consistency error of 0.01m. Our result indicates that regressing SMPL-X body parameters from body mesh vertices is easier than from a skeleton of joints and generates slightly better human-object contact.

To quantify the importance of the two-stage design in interaction generation, we train two models for one-stage generation which directly predicts the human body given objects in the original scene coordinate system and re-centered scene coordinate system, respectively. The re-centered scene coordinate system translates the origin to the average of interaction object points. The semantic contact score of interactions generated from the one-stage models using original and re-



(a) Binary perceptual study



(b) Unary perceptual study

Fig. S6: AMT User interfaces for the perceptual studies.

centered scene coordinates rapidly drops to 0.20 and 0.31 respectively, compared to 0.72 of the two-stage method. Our result shows that learning to jointly predict global interaction location, orientation, and body pose is much more difficult and the two-stage design is necessary for generating interactions with natural human-scene contact.

D More Qualitative Results

D.1 Retarget Novel Objects.

Our method can naturally retarget learned interactions to unseen objects with similar geometry and affordance because we use the point cloud object representation which encodes object shapes, instead of the object category in previous works. Figure S7 shows some created novel interactions that are not seen in

"lie on the floor" "walk on the table" "step up the bed" "sit on the monitor"

Fig. S7: Novel combinations of actions and objects generated by our method that were not part of the training data.

Fig. S8: Interaction synthesis results with explicitly controlled varying human body shapes, where the extremely thin and heavy bodies are not seen during training.

training. It demonstrates the generalization capability of our method and the potential for synthesizing interactions with open-set objects beyond predefined object categories. Moreover, our method also creates some interesting interactions that are less possible in the real world, e.g., sitting on a monitor.

D.2 Explicit Body Shape Control.

Our method features explicit body shape control in interaction generation, which is achieved by using personalized body templates. Figure S8 shows interaction synthesis results where we control the SMPL-X body shape parameters changing from -3 to 3. Note that the extremely thin and heavy bodies are not seen during training and our method generalizes to such extreme shapes.

D.3 Random Interaction Samples

We show more random interaction samples from our method in Fig. S9. Our method generates natural and diverse human-scene interactions.

D.4 Interaction Refinement

We demonstrate how the interaction-based optimization improves human-scene penetration and contact in Fig. S10.

D.5 Compositional Interaction Generation

We show composite interactions randomly generated by composing atomic interactions in Fig. S11. Our method is capable of generating composite interactions without corresponding training data.

D.6 Synthesis in Scenes with Noisy Segmentation

To show the possibility of generating interactions in scenes without ground truth segmentation using our method, we generate interactions in scenes with noisy segmentation obtained using off-the-shelf segmentation methods [3,10] where the object geometry can be incomplete and noisier than the scene segmentation we use. We show the generation results on noisily segmented PROX test scenes in Fig. S12. Our method can generate reasonable interactions given noisy objects as long as the object shape is not significantly different from training objects.

Regarding training with such noisy scene segmentation, we find it demands prohibitively more effort in collecting interaction semantic annotation with noisy segmentation and conclude that a clean scene segmentation that is consistent with human perception is necessary.

D.7 Failure Cases and Limitation

We show typical failure cases of generating interactions from semantic specifications in Fig. S13 and failure cases of composing atomic interactions in Fig. S14.

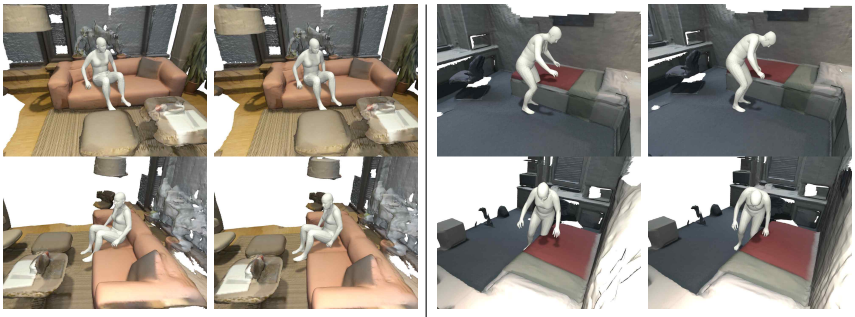
Our method has some limitations. Firstly, our generative models currently ignore scene objects that are not explicitly specified in the input, which can lead to penetration with unspecified objects. We currently solve such penetration using post-processing based on pre-computed scene SDF grids. It is possible to get rid of the demand for scene SDF grid if we use recent human-occupancy methods [9], and learning obstacle-aware generative models is an interesting future direction.

Besides, we observe that hand-object contact in generated results of the touching action are not accurate enough, an issue caused by the low-quality hand estimation in used pseudo-ground truth data. Using hand-object interaction data with high-quality hand estimation can be future work to improve hand-object contact in synthesis results.

In addition, the action semantics considered in this paper is relatively coarse-grained due to the limited scale of interaction data in PROX. We do not distinguish left and right limbs in annotation due to limited data and can not generate fine-grained composite semantics such as touching the chair with the left hand while touching the table with the right hand. Given larger scale interaction data, we expect to model more expressive interaction semantics and compose actions corresponding to finer-grained body parts segmentation, e.g., put the left palm on the table and lean on the right elbow on the table.

Moreover, we observe our mask-based composition method fails in two cases as shown in Figure S14: 1) when physically impossible interaction combinations are specified as input, such as sitting on a cabinet while touching a bed 10 meters away. 2) when the data distributions of atomic actions have no intersection in the training data, such as the combination of lying and touching since none of the touching poses are simultaneously lying in our training data.

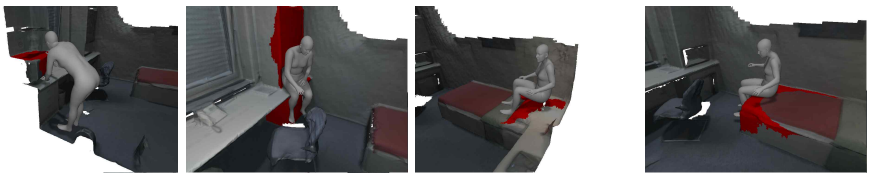
Fig. S9: Randomly sampled interactions from our method. Each interaction is rendered with two views.



(a) before (b) after (c) before (d) after

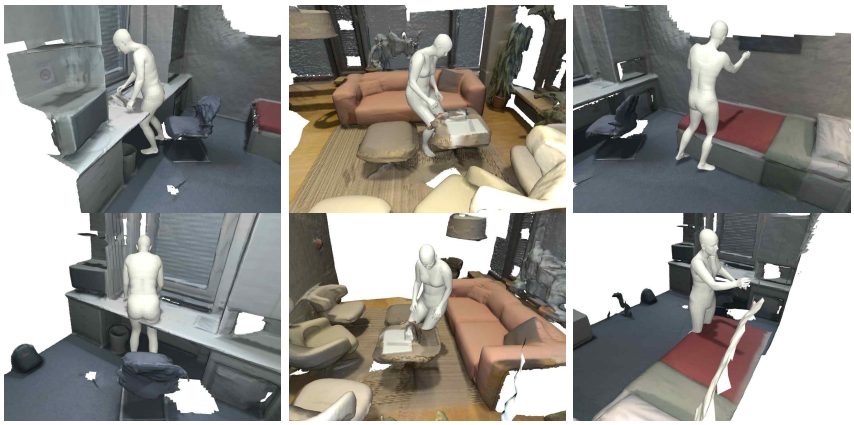
Fig. S10: Illustration of interaction-based optimization where (b) and (d) are the optimized results of (a) and (c). The human-scene penetration and contact are improved after the optimization.

Fig. S11: Randomly sampled composite interaction from our method. Note that the model is not trained with the corresponding training data.



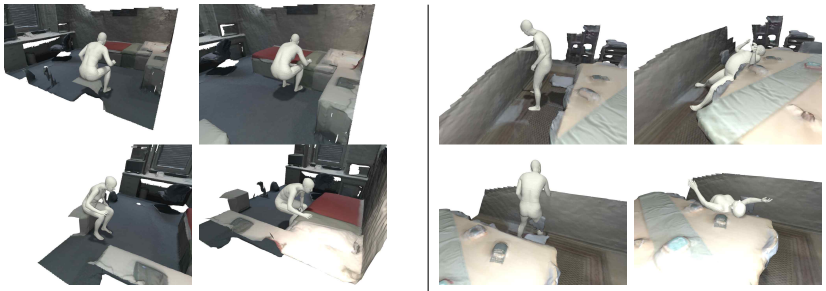
touch sit on sit on sit on

Fig. S12: Synthesized interactions in PROX scenes with noisy object instance segmentation. The noisy interaction objects are highlighted as red.



(a) Penetration (b) Geometry Variance (c) Local Minimum

Fig. S13: Typical failure cases in our results. Example (a) shows penetration with thin-structure objects where the scene SDF is not effective in resolving penetration. Example (b) shows the synthesis result of touching a table with significantly different geometry from tables in training data, i.e., much lower and smaller in size. Example (c) shows a failure case of being stuck in local minimum in optimization where the human is blocked by the bed from touching the wall.



(a) (b)

Fig. S14: Typical failure cases in generating novel interactions by semantic composition. Example (a) shows the failed composition of sitting on a cabinet and touching a bed far away due to being physically impossible. Example (b) shows the failed composition of lying on the floor and touching the wall.

References

1. Bowman, S.R., Vilnis, L., Vinyals, O., Dai, A.M., Jozefowicz, R., Bengio, S.: Generating sentences from a continuous space. arXiv preprint arXiv:1511.06349 (2015)
2. Charbonnier, P., Blanc-Feraud, L., Aubert, G., Barlaud, M.: Two deterministic half-quadratic regularization algorithms for computed imaging. In: Proc. of ICIP. vol. 2. IEEE (1994)
3. Chen, S., Fang, J., Zhang, Q., Liu, W., Wang, X.: Hierarchical aggregation for 3d instance segmentation. In: Proc. of ICCV (2021)
4. Choi, H., Moon, G., Lee, K.M.: Pose2mesh: Graph convolutional network for 3d human pose and mesh recovery from a 2d human pose. In: Proc. of ECCV (2020)
5. Diederik, K., Jimmy, B., et al.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
6. Dutta, A., Zisserman, A.: The VIA annotation software for images, audio and video. In: Proc. of MM. MM '19, ACM, New York, NY, USA (2019). https://doi.org/10.1145/3343031.3350535
7. Hassan, M., Choutas, V., Tzionas, D., Black, M.J.: Resolving 3D human pose ambiguities with 3D scene constraints. In: Proc. of ICCV (2019)
8. Kolotouros, N., Pavlakos, G., Daniilidis, K.: Convolutional mesh regression for single-image human shape reconstruction. In: Proc. of CVPR (2019)
9. Mihajlovic, M., Saito, S., Bansal, A., Zollhoefer, M., Tang, S.: COAP: Compositional articulated occupancy of people. In: Proc. of CVPR (2022)
10. Nekrasov, A., Schult, J., Litany, O., Leibe, B., Engelmann, F.: Mix3D: Out-of-Context Data Augmentation for 3D Scenes. In: Proc. of 3DV (2021)
11. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Proc. of NeurIPS (2019)
12. Qi, C.R., Yi, L., Su, H., Guibas, L.J.: Pointnet++: Deep hierarchical feature learning on point sets in a metric space. arXiv preprint arXiv:1706.02413 (2017)
13. Wang, N., Zhang, Y., Li, Z., Fu, Y., Liu, W., Jiang, Y.G.: Pixel2mesh: Generating 3d mesh models from single rgb images. In: Proc. of ECCV (2018)
14. Zhang, S., Zhang, Y., Bogo, F., Marc, P., Tang, S.: Learning motion priors for 4d human body capture in 3d scenes. In: Proc. of ICCV (2021)
15. Zhang, Y., Black, M.J., Tang, S.: Perpetual motion: Generating unbounded human motion. arXiv preprint arXiv:2007.13886 (2020)
16. Zhang, Y., Black, M.J., Tang, S.: We are more than our joints: Predicting how 3d bodies move. In: Proc. of CVPR (2021)
17. Zhang, Y., Hassan, M., Neumann, H., Black, M.J., Tang, S.: Generating 3d people in scenes without people. In: Proc. of CVPR (2020)
18. Zhou, Y., Barnes, C., Lu, J., Yang, J., Li, H.: On the continuity of rotation representations in neural networks. In: Proc. of CVPR (2019)