Supplemental Material

We show in this supplemental material additional qualitative (Sec. A) and quantitative (Sec. B) results, detail our baseline evaluation protocol (Sec. C), elaborate on the metrics used in the main paper (Sec. D), show the architecture used in our approach (Sec. E), and provide additional details regarding the data (Sec. F).

A. Additional Qualitative Results

A.1. Additional Interactions

We show additional generated 3D human-object interactions of our method in Fig. 2, with object geometry and text condition on the left, and our generated sequence on the right.

A.2. Same Prompt, Different Interactions

We evaluate the ability of our method to generate diverse interactions for a fixed text condition visually in Fig. 1, for text prompt "Move a stool" and "Sit on a stool". In the ground truth training data, move is only done with one or two hands, and feet; moving with the butt sometimes occurs for the text description "Sit on a stool".

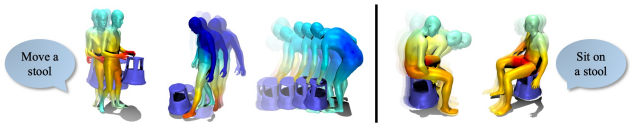

Figure 1. Our method is able to generate diverse human-object interactions for the same prompts.

B. Additional Quantitative Results

B.1. Evaluating Penetrations and Floating

Our method discourages penetration and floating implicitly, by enforcing correct contact distances as a soft constraint at train and test time. However, the exact fidelity and diversity of our results is hard to capture with any single metric. Thus, we evaluate multiple such metrics in the main paper (R-Precision, FID, Diversity, MultiModality), and conduct a perceptual user study to verify the metrics' expressiveness.

Here, we provide an additional evaluation based on intuitive physics-based metrics: Tab. 1 evaluates the mean ratio of frames with some penetration as well as the ratio of penetrating vertices overall, showing that penetrations typically happens with small body parts (e.g., hands, which also occurs in the ground-truth data). We also evaluate the ratio of frames and vertices with human and object not in contact, including floating and stationary objects, which is expected to be close to the dataset ratio.

Results show similar penetration and floating between our generations and ground-truth training data.

	BEHAVE				CHAIRS			
	Penetration		Non-Contact		Penetration		Non-Contact	
	Frames	Vertices	Frames	Vertices	Frames	Vertices	Frames	Vertices
Dataset	32.9%	4.1%	21.4%	86.2%	26.9%	1.1%	11.9%	70.4%
Ours	31.3%	3.0%	17.8%	93.3%	35.8%	4.2%	14.1%	74.3%

Table 1. Penetration and non-contact (including floating) ratios in terms of frames as well as overall vertices vs ground-truth data.

B.2. Evaluating Contact

Tab. 2 evaluates our contact predictions using precision/recall and distance metrics. We follow [4, 11, 12] to define contact if $\leq$5cm from object. We also report mean ℓ_1 error in contact distance predictions. All metrics are reported for body parts $\leq 1m$ of the object, to focus on contact scenarios. Better contact prediction corresponds with better HOI generations. Note that none of our baselines predict contact distances.

	BEHAVE			CHAIRS		
Approach	Precision ↑	Recall ↑	Distance ↓	Precision ↑	Recall ↑	Distance ↓
Separate contact pred.	23.4%	25.6%	0.53	58.6%	49.1%	0.24
No contact weighting	29.5%	33.5%	0.34	60.6%	63.4%	0.10
No contact guidance	46.3%	39.2%	0.31	64.2%	70.2%	0.12
Ours	**63.6%**	**59.5%**	**0.07**	**78.3%**	**84.5%**	**0.04**

Table 2. Evaluation of predicted contact distances, in terms on precision and recall ($\leq 5cm$ distance), as well as mean contact ℓ_1 error in meters.

B.3. Novelty of Generated Interactions

We perform an additional interaction novelty analysis to verify that our method does not simply retrieve memorized train sequences but is indeed able to generate novel human-object interactions. To do so, we generate ≈ 500 sequences from both datasets and retrieve the top-3 most similar train sequences, as measured by the l_2 distance in human body and object transformation parameter space.

Fig. 3 shows the top-3 closest train sequences, along with a histogram of l_2 distances computed on our test set of ≈ 500 generated sequences. In red, we mark the intra-trainset distance between samples in the train set. We observe that the distance between our generated sequences and the closest train sequence is mostly larger than the intra-train distance. Thus, our method is able to produce samples that are novel and not simply retrieved train sequences.

B.4. SMPL Bodies vs. HumanML3D Skeletons

We observe slight pose jitter and foot skating in our ground-truth training data (especially BEHAVE, captured with Kinect sensors). As a result, our model reflects some of these effects. Skeleton representations such as HumanML3D [3] could tackle these artifacts, but do not work with contact as effectively as SMPL bodies. Nevertheless, we train ours with HumanML3D parameters for comparison in Tab. 3 (fitting SMPL after for comparable evaluation) which leads to degraded performance due to less effective contact guidance.

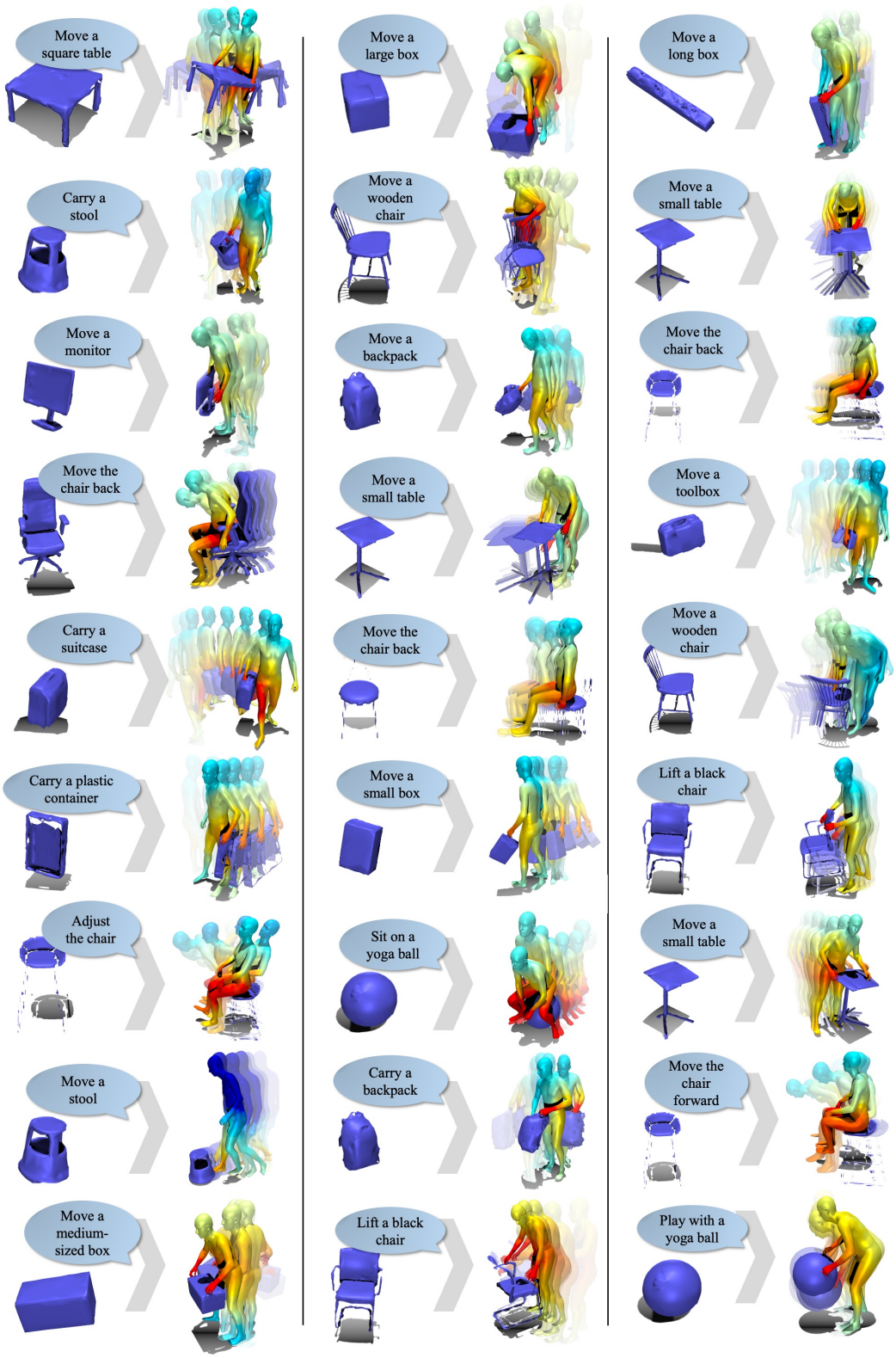

Figure 2. Additional qualitative evaluation. Our method produces diverse and realistic 3D human-object interaction sequences, given object geometry and short text description of the action. The sequences depict high-quality human-object interactions by modeling contact, mitigating floating and penetration artifacts.

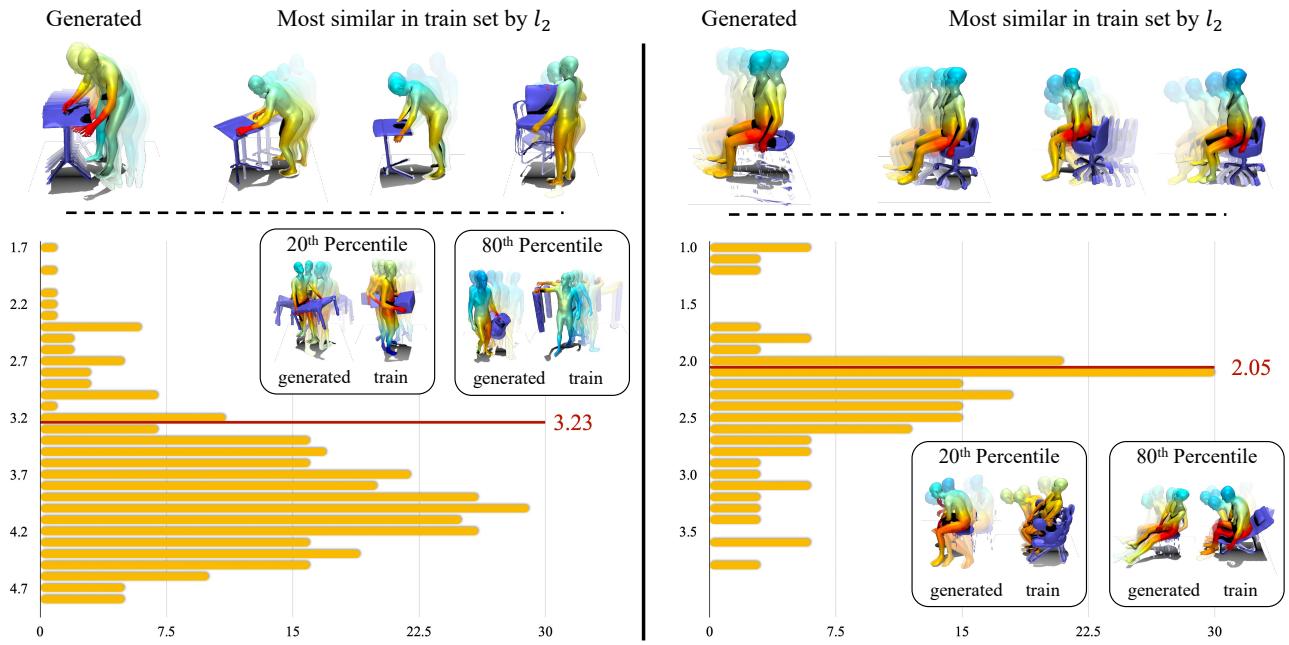

Figure 3. Human-Object Interaction Sequence Novelty Analysis. Performed on BEHAVE [1] (left) and CHAIRS [5] (right). We retrieve top-3 most similar sequences from the train set, and plot a histogram of distances to the closest train sample. While sequences at the 20th percentile still resemble the generated interactions, there is a large gap in the 80th percentile. We show the intra-trainset distance in red. Our approach generates novel shapes, not simply retrieving memorized train samples.

C. Baseline Evaluation Setup

There is no previous approach to modeling 3D human-object interactions from text and object geometry for direct comparison. Thus, we compare to the two closest methods, and compare to them in multiple settings, for a fair comparison.

The most related approach is InterDiff [10]. Their setting is to generate a short sequence of human-object interactions, from an observed such sequence as condition, with geometry but no text input. Their goal is to generate one, the most likely, sequence continuing the observation. We use their full approach, including the main diffusion training together with the post-processing refinement step. We compare in two different settings: First, in their native setup, running their method unchanged and modifying ours to take in geometry and past sequence observation instead of text (Motion-Cond. HOI in Tab. 1 main). Then, we modify their approach to take in geometry and text, replacing their past motion encoder with our CLIP-based text encoder (Text-Cond. HOI in Tab. 1 main). We observe that our method is able to outperform InterDiff in both scenarios, for both datasets.

We additionally compare to MDM [9], a recent diffusion-based state-of-the-art human motion generation approach. Their approach is based on a transformer encoder formulation, using each human body as a token in the attention. We run their method on SMPL parameters and first compare in their native setting, only predicting human motion. We compare to the human motion generated by our method which is trained to generate full human-object interactions (Text-Cond. Human Only in Tab. 1 main). We also compare to human motion sequences generated by InterDiff in this setting. We see that our method is able to outperform both baselines even in this setting, demonstrating the added benefit of learning interdependencies of human and object motion. For the comparison in our setting, we modify MDM by adding additional tokens for the objects to the attention formulation. Our approach performs more realistic and diverse sequences in both settings which better follow the text condition.

	BEHAVE				CHAIRS			
Representation	R-Prec. (top-3) ↑	FID ↓	Diversity →	MModality →	R-Prec. (top-3) ↑	FID ↓	Diversity →	MModality →
Ours (HumanML3D)	0.33	11.94	2.15	3.75	0.48	12.83	4.39	5.11
Ours	**0.62**	**6.31**	**6.63**	**5.47**	**0.74**	**6.45**	**8.91**	**5.94**

Table 3. Ours (using SMPL bodies) vs. using HumanML3D [3] skeletons and fitting SMPL bodies afterwards. While HumanML3D is designed to reduce jitter and foot skating, it leads to degraded performance in our scenario due to less effective contact guidance.

D. Fidelity and Diversity Metrics

We base our fidelity and diversity metrics R-Precision, FID score, Diversity, and MultiModality on practices established for human motion generation [2, 3, 9], with minor modifications: We use the same networks used by these previous approaches, and adapt the input dimensions to fit our feature lengths, $F = 79$ when evaluating human body motion only, and $F = 79 + 128 + 9 = 216$ (SMPL parameters, contact distances, object transformations) for full evaluation in the human-object interaction scenario.

E. Architecture Details

Fig. 4 shows our detailed network architecture, including encoder, bottleneck, and decoder formulations.

F. Data Details

F.1. Datasets

CHAIRS [5] captures 46 subjects as their SMPL-X [6] parameters using a mocap suit, in various settings interacting with a total of 81 different types of chairs and sofas, from office chairs over simple wooden chairs to more complex models like suspended seating structures. Each captured sequence consists of 6 actions and a given script; the exact separation into corresponding textual descriptions was manually annotated by the authors of this paper. In total, this yields ≈ 1300 sequences of human and object motion, together with a textual description. Every object geometry is provided as their canonical mesh; we additionally generate ground-truth contact and distance labels based on posed human and object meshes per-frame for each sequence. We use a random 80/10/10 split along object types, making sure that test objects are not seen during training.

BEHAVE [1] captures 8 participants as their SMPL-H [8] parameters captured in a multi-Kinect setup, along with the per-frame transformations and canonical geometries of 20 different object with a wide range, including yoga mats and tables. This yields ≈ 130 longer sequences. We use their original train/test split.

F.2. Object Geometry Representation

We represent object geometry as a point cloud, to be processed by a PointNet [7] encoder. For this, we sample $N = 256$ points uniformly at random on the surface of an object mesh. Each object category is sampled once as a pre-processing step and kept same for train and inference.

References

[1] Bharat Lal Bhatnagar, Xianghui Xie, Ilya A. Petrov, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. BEHAVE: dataset and method for tracking human object interactions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 15914–15925. IEEE, 2022. 3, 4

[2] Chuan Guo, Xinxin Zuo, Sen Wang, Shihao Zou, Qingyao Sun, Annan Deng, Minglun Gong, and Li Cheng. Action2motion: Conditioned generation of 3d human motions. In *MM '20: The 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA, October 12-16, 2020*, pages 2021–2029. ACM, 2020. 4

[3] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 5142–5151. IEEE, 2022. 1, 3, 4

[4] Mohamed Hassan, Partha Ghosh, Joachim Tesch, Dimitrios Tzionas, and Michael J. Black. Populating 3d scenes by learning human-scene interaction. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, vir-*

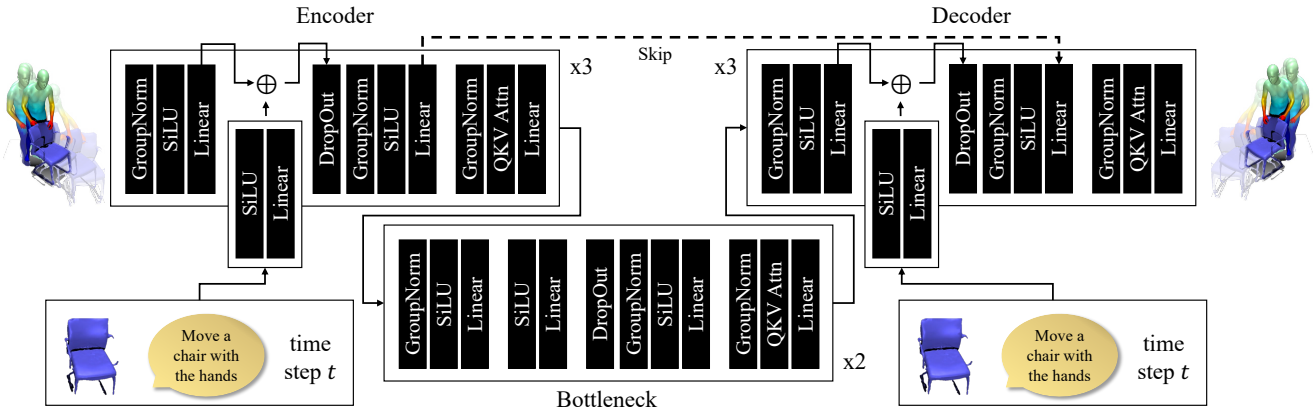

Figure 4. Network architecture specification.

tual, June 19-25, 2021, pages 14708–14718. Computer Vision Foundation / IEEE, 2021. 1

[5] Nan Jiang, Tengyu Liu, Zhexuan Cao, Jieming Cui, Zhiyuan Zhang, Yixin Chen, He Wang, Yixin Zhu, and Siyuan Huang. Full-body articulated human-object interaction. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 9331–9342. IEEE, 2023. 3, 4

[6] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 10975–10985. Computer Vision Foundation / IEEE, 2019. 4

[7] Charles Ruizhongtai Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 77–85. IEEE Computer Society, 2017. 4

[8] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: modeling and capturing hands and bodies together. *ACM Trans. Graph.*, 36(6):245:1–245:17, 2017. 4

[9] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit Haim Bermano. Human motion diffusion model. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. 3, 4

[10] Sirui Xu, Zhengyuan Li, Yu-Xiong Wang, and Liang-Yan Gui. Interdiff: Generating 3d human-object interactions with physics-informed diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14928–14940, 2023. 3

[11] Siwei Zhang, Yan Zhang, Qianli Ma, Michael J. Black, and Siyu Tang. PLACE: proximity learning of articulation and contact in 3d environments. In *8th International Conference on 3D Vision, 3DV 2020, Virtual Event, Japan, November 25-28, 2020*, pages 642–651. IEEE, 2020. 1

[12] Yan Zhang, Mohamed Hassan, Heiko Neumann, Michael J Black, and Siyu Tang. Generating 3d people in scenes without people. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6194–6204, 2020. 1