Multi-Objective Photoreal Simulation (MOPS) Dataset for Computer Vision in Robot Manipulation

Maximilian Xiling Li[1] Paul Mattes[1] Nils Blank[1] Rudolf Lioutikov[1,2]
[1]Intuitive Robots Lab, Karlsruhe Institute of Technology, Germany
[2]Robotics Institute Germany (RIG)
{m.li, lioutikov}@kit.edu

Abstract

Datasets providing high-quality visual annotations in manipulation-relevant scenes remain limited. We introduce the Multi-Object Photoreal Simulation (MOPS) dataset, offering comprehensive ground truth annotations in photorealistic simulated environments. MOPS uses an LLM-based zero-shot pipeline to normalize 3D object scale and generate part-level affordances, providing pixel-level segmentations for fine-grained parts and affordance prediction (e.g., "graspable", "pushable"). We validate MOPS through vision and robot learning benchmarks. The dataset and framework will be made publicly available.

1. Introduction

Computer vision methods depend on task-specific datasets for training and evaluation, yet creating comprehensive datasets with detailed annotations requires substantial human effort, limiting scale and granularity. While numerous benchmarks exist for tasks like affordance segmentation [16], part segmentation [4], scene graph generation [30], and 6D pose estimation [27], the robotics domain remains underrepresented. Embodied agents require robust perception of actionable parts and manipulation opportunities, yet few vision datasets address these needs. We argue that datasets for learning visual perception for robot manipulation should fulfill five key requirements:

1. **Manipulation-relevant Objects** (REQ OBJ): Common household items found in living spaces.
2. **Manipulation-relevant Annotations** (REQ ANN): High-resolution labels including part information, affordances, and 6D poses.
3. **Manipulation-relevant Representations** (REQ REP): Multiple scene representations such as pointclouds or scene graphs beyond images.
4. **Manipulation-relevant Environments** (REQ ENV): Photorealistic rendering of real-world scenes with natural clutter.
5. **Manipulation-relevant Interactions** (REQ INT): Agent-object and object-object interactions over time, supporting direct policy evaluation.

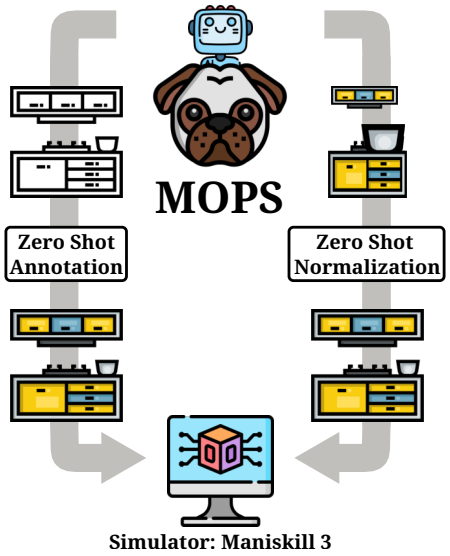

Figure 1. MOPS provides labeled, realistic data for robotics and vision tasks through Large Language Models-enabled zero-shot annotation and normalization of 3D assets, used to create indoor scenarios for data collection.

Existing datasets fulfill only subsets of these requirements. Human-centric video datasets [22] lack suitability for robot policy training. Affordance [16] and pose estimation [27] datasets lack environmental realism and dynamics. Robot datasets [11, 19] feature realistic scenes but lack vision-quality ground truth, even with annotation frameworks like NILS [2]. We introduce MOPS, a dataset generation framework with pixel-level ground truth addressing all five requirements (Figure 2). MOPS uses assets from PartNet-Mobility [26] and RoboCasa [17] to create articulated household scenes (REQ OBJ) in photorealistic envi-

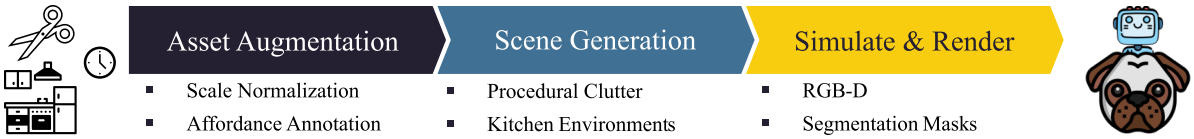

Figure 2. The three stages of the MOPS dataset creation process, using 3D assets from RoboCasa and PartNet-Mobility.

ronments (REQ ENV). A zero-shot pipeline built on GPT-4o [18] normalizes asset scale and generates affordance annotations (REQ ANN). The framework provides pixel-level ground truth for class, part, and instance segmentation, affordance labels, geometric information, and multiple modalities including RGB-D and pointclouds (REQ REP). Built on ManiSkill3 [23], MOPS enables interactive scenes for evaluating robot behavior (REQ INT). Our contributions are:

- A zero-shot LLM pipeline for automatic scale normalization and affordance annotation of 3D assets
- A framework producing unlimited photorealistic scenes with pixel-perfect ground truth for 56 affordances across 137 object categories
- Three benchmark datasets (MOPS-Object, MOPS-Clutter, MOPS-Kitchen) with vision and robot learning baselines

2. Related Work

Vision Datasets: Fine-grained classification datasets like CUB-200-2011 [24] and driving-focused segmentation benchmarks [1, 6] lack relevance to manipulation (REQ OBJ). Indoor scene datasets such as ScanNet++ [29] and Hypersim [20] are closer to robotics needs but lack annotations like 6D poses or affordances (REQ ANN). The RGB-D Part Affordance dataset [16] provides affordance labels but with limited scene realism (REQ ENV). None of these datasets offer interactivity (REQ INT).

3D Asset Datasets: 3D AffordanceNet [7] provides affordance-annotated point clouds (REQ OBJ& REQ ANN) but lacks material information for photorealistic rendering (REQ ENV). PartNet-Mobility [26] adds materials and articulation for simulation (REQ INT) but lacks affordance labels (REQ ANN). Both suffer from inconsistent object scaling relative to robotic manipulators. MOPS addresses these gaps via a zero-shot LLM-based pipeline that augments PartNet-Mobility assets with affordance annotations and normalized scaling.

Robotics Datasets: Large real-world collections like Open X-Embodiment [19] and DROID [11] face consistency and scalability challenges, even with zero-shot labeling [2]. Simulation frameworks offer scalable alternatives: AI2-THOR [12] provides procedural scene variety, OmniGibson [13] complex physics, and RoboCasa [17] over 100k trajectories via demonstrations and MimicGen [14]. Generative approaches like RoboGen [25] and Genesis [8] enable broad scene variety but struggle with cluttered environments (REQ ENV). However, to our knowledge, no existing framework provides comprehensive pixel-wise ground truth for semantic concepts and affordances out of the box.

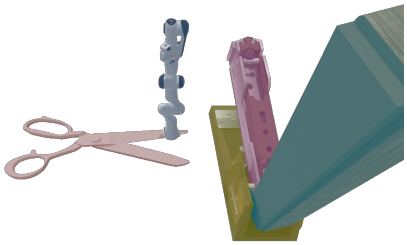

Figure 3. The MOPS pipeline scales 3D assets to realistic ranges relative to the robot using LLM-based scale estimates. Part-level affordances are also generated by the LLM (colored overlays).

MOPS combines RoboCasa's scene variety with augmented PartNet-Mobility assets, generating unlimited realistic scenes with full pixel-wise annotations including affordances. Built on ManiSkill3 [23], it leverages raytracing and GPU parallelization akin to Isaac Lab [15]. Table 1 provides a detailed comparison.

3. Zero-Shot 3D Asset Augmentation

The MOPS pipeline consists of three stages (Figure 2): (1) zero-shot asset augmentation, (2) procedural scene generation, and (3) photorealistic rendering with pixel-perfect annotations.

MOPS populates scenes with 3D assets from RoboCasa [17] and PartNet-Mobility [26]. However, these assets require augmentation: objects lack a common reference scale, breaking realism (REQ OBJ & REQ ENV), and miss affordance annotations for pixel-wise ground truth (REQ ANN). MOPS addresses both issues through a zero-shot pipeline based on GPT-4o [18].

Scale Normalization (REQ OBJ). We query GPT-4o for realistic minimum and maximum Width×Height×Depth dimensions per object category. Scaling factors $s_{\min}$ and $s_{\max}$ are computed by dividing the largest bounding box dimension by the largest WHD dimension. At load time, a random factor is sampled for $[s_{\min}, s_{\max}]$ to increase visual variety (Figure 3).

Table 1. **Related Work Overview.** Comparison of computer vision and robotics datasets. MOPS bridges the gap by providing comprehensive part-level affordance annotations (S+I+P+A) across multiple modalities (R+D+M).

Dataset	Objects	Annotations	Represent.	Env.	Interact.	Trajectory
Vision Datasets						
CUB-200-2011 [24]	—	P*	R	—	—	—
CityScapes [6]	—	S+I	R	—	—	—
SemanticKITTI [1]	—	S+I	R+D+M	—	—	—
ScanNet++ [29]	✓	S+I	R+D+M	✓	—	—
HyperSim [20]	✓	S+I	R+D+M	✓	—	—
RGB-D Part Aff. [16]	✓	A	R+D	—	—	—
3D AffordanceNet [7]	✓	A	M	—	—	—
PartNet-Mobility [26]	✓	P	M	—	—	—
Robotics Datasets						
2-Handed [9]	✓	S+I	R+D	✓	—	—
Open-X [19]	✓	—	R+D	✓	—	✓
DROID [11]	✓	—	R+D	✓	—	✓
AI2-THOR [12]	✓	S+I	R+D+M	✓	✓	—
OmniGibson [13]	✓	S+I	R+D+M	✓	✓	—
RoboCasa [17]	✓	S+I	R+D+M	✓	✓	✓
MOPS (Ours)	✓	**S+I+P+A**	**R+D+M**	✓	✓	(✓)

Abbreviations: S: Semantic Seg., **I**: Instance Seg., **P**: Part Seg., **P***: Part Centers, **A**: Affordance Seg., **R**: RGB, **D**: Depth, **M**: 3D (Mesh/PC).(✓) denotes compatibility with external trajectories (RoboCasa).

Affordance Annotation (REQ ANN). PartNet-Mobility provides articulated, part-based assets enabling rich interactions (REQ INT) but lacks affordance labels. GPT-4o generates multi-label affordances for each object and part; for assets without part annotations (e.g., RoboCasa), the entire object is labeled. A sentence embedding model clusters the resulting labels to merge duplicates (e.g., *closable* / *close*) and align semantically equivalent terms.

4. Dataset Generation

MOPS generates unlimited scenes by combining 120 RoboCasa indoor environments with 2,300 articulated PartNet-Mobility objects augmented via the zero-shot pipeline. Our generation pipeline achieves a throughput of $\approx$ 1000 images/hour on an Intel 12700 20 Core CPU and NVidia 3070 8GB GPU desktop computer.

Ground Truth Annotations (REQ ANN). MOPS provides multiple camera views (birds-eye, over-the-shoulder, ego, and in-hand) with raytraced or rasterized RGB, depth in millimeters, and surface normal maps. From the simulation state, MOPS generates pixel-perfect ground truth for part, class, and instance segmentation via look-up tables populated during scene creation. Combined with the zero-shot affordance pipeline, this yields multi-label per-pixel affordance annotations, represented as binary vectors indicating presence of each affordance type. Additionally, 6D object poses are available for each camera frame. Figure 4 illustrates these modalities.

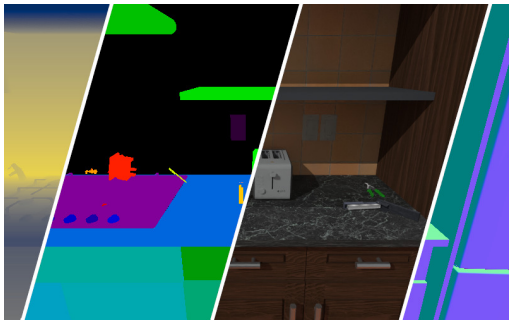

Figure 4. Ground Truth Rendering in MOPS: Depth, Affordance Segmentation, RGB and Surface Normals.

Extended Representations (REQ REP). ManiSkill3 supports point cloud generation by merging 2.5D images across cameras, with configurable intrinsics, extrinsics, and lighting.

Realistic Environments (REQ ENV). To create realistically cluttered scenes, MOPS procedurally places aug-

mented assets on kitchen countertops by computing available space from collision mesh bounding boxes and dropping objects at random positions. While heuristic compared to trained approaches like ClutterGen [10], this generalizes to novel environments without training.

Interactive Simulation (REQ INT). MOPS leverages ManiSkill3 for full physics-enabled interactivity. Existing robot trajectories, such as RoboCasa demonstrations, can be rolled out to record observations capturing object-object and robot-object interactions. The framework also supports recording new demonstrations via ManiSkill3's teleoperation interface (Figure 5). Recorded trajectories can be replayed to generate MOPS's comprehensive ground truth annotations.

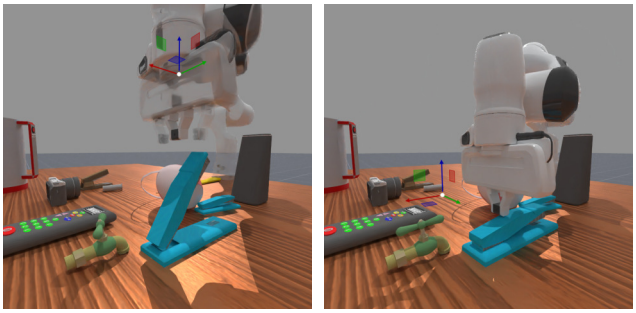

Figure 5. **M**ulti-**O**bjective **P**hotoreal **S**imulation (MOPS) uses ManiSkill3's teleoperation interface. The semitransparent arm previews the target position.

5. Affordance Analysis

Multi-**O**bjective **P**hotoreal **S**imulation (MOPS) provides affordance annotations at multiple granularities: part-level annotations from PartNet-Mobility [26] (24 affordance types for 14,100 parts across 2,345 objects) and object-level annotations for RoboCasa assets (44 types across 101 categories), yielding **56 affordance labels for 137 object categories** in total. Compared to 3D-AffordanceNet [7], MOPS offers 3.5× more affordance types and 6× more object categories (Table 2), prioritizing taxonomic diversity over instance-level repetition.

Table 2. Dataset diversity comparison. MOPS provides significantly higher taxonomic coverage despite fewer individual instances than 3D-AffordanceNet.

Dataset	Level	Aff. Labels	Obj. Cat.	Objects
RGB-D Part	Part	7	17	105
3D-AffNet	Part	16	23	**22,949**
MOPS-Partnet	Part	24	46	2,345
MOPS-Robocasa	Object	44	101	1,008
MOPS (Total)	**Mixed**	**56**	**137**	3,353

Table 3. **Multi-Label Segmentation Baselines.** Models are evaluated on 56 affordance labels using macro-averaged mIoU and F1 scores.

		Clutter Setting		Kitchen Setting	
Model	**Training**	**mIoU**	**F1**	**mIoU**	**F1**
DeepLabV3	Full	15.2	16.5	17.9	20.3
Segformer	Full	**16.2**	**17.9**	22.1	23.9
DINOv2	Frozen	13.5	14.3	**34.9**	**38.9**

Table 4. **Robot Manipulation Performance.** Imitation learning results on 24 RoboCasa tasks, evaluated over 10 environment seeds each.

Policy Inputs	Success Rate	Gain
RGB only	13.33%	–
RGB + MOPS Affordances	**21.25%**	+7.92

6. Experiments

We construct two benchmarks with no 3D model overlap between train/test splits. Each 512 × 512 image uses randomized lighting, object poses, and camera positions, captured after physics settling. **MOPS-Clutter**: unstructured tabletop scenes (3k train, 1k test). **MOPS-Kitchen**: assets within complex RoboCasa kitchens with distractors (3k train, 1k test).

Vision Benchmarks. We evaluate multi-label affordance segmentation on MOPS-Clutter/Kitchen (Table 3) using DeepLabV3 [5], Segformer [28], and DINOv2.

Robot Manipulation. We train DiTFlow [21] imitation learning policies within LeRobot [3] on 24 RoboCasa tasks using 50 demonstrations each (see Table 4.

7. Conclusion

We introduce MOPS, a dataset generation pipeline providing photorealistic renderings of manipulation-relevant scenes with pixel-level ground truth for segmentation, affordances, and 6D poses. A zero-shot LLM-based augmentation pipeline yields 56 affordance labels across 137 object categories, easily extensible to new assets or semantic labels. Vision benchmarks confirm dataset difficulty, and ground-truth affordances improve robot manipulation success by +6.25pp over RGB-only baselines. Future work includes extending the augmentation pipeline to generate materials and textures for broader asset coverage, and training robust affordance prediction models for end-to-end manipulation.

Acknowledgements

The work was funded by the German Research Foundation (DFG) – 448648559. This work is supported by the Helmholtz Association Initiative and Networking Fund under the KiKIT Pilot Program Core-Informatics. The authors gratefully acknowledge the support of the Robotics Institute Germany (RIG).

References

[1] J. Behley, M. Garbade, A. Milioto, J. Quenzel, S. Behnke, C. Stachniss, and J. Gall. SemanticKITTI: A Dataset for Semantic Scene Understanding of LiDAR Sequences. In *Proc. of the IEEE/CVF International Conf. on Computer Vision (ICCV)*, 2019. 2, 3

[2] Nils Blank, Moritz Reuss, Marcel Rühle, Ömer Erdinç Yağmurlu, Fabian Wenzel, Oier Mees, and Rudolf Lioutikov. Scaling robot policy learning via zero-shot labeling with foundation models. In *8th Annual Conference on Robot Learning*, 2024. 1, 2

[3] Remi Cadene, Simon Alibert, Alexander Soare, Quentin Gallouedec, Adil Zouitine, Steven Palma, Pepijn Kooijmans, Michel Aractingi, Mustafa Shukor, Dana Aubakirova, Martino Russi, Francesco Capuano, Caroline Pascal, Jade Choghari, Jess Moss, and Thomas Wolf. Lerobot: State-of-the-art machine learning for real-world robotics in pytorch. https://github.com/huggingface/lerobot, 2024. 4

[4] Angel X. Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 1

[5] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation, 2017. 4

[6] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2, 3

[7] Shengheng Deng, Xun Xu, Chaozheng Wu, Ke Chen, and Kui Jia. 3d affordancenet: A benchmark for visual object affordance understanding. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1778–1787, 2021. 2, 3, 4

[8] Genesis. Genesis: A universal and generative physics engine for robotics and beyond, 2024. 2

[9] Marvin Heidinger, Snehal Jauhri, Vignesh Prasad, and Georgia Chalvatzaki. 2handedafforder: Learning precise actionable bimanual affordances from human videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14743–14753, 2025. 3

[10] Yinsen Jia and Boyuan Chen. Cluttergen: A cluttered scene generator for robot learning. In *8th Annual Conference on Robot Learning*, 2024. 4

[11] Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yunliang Chen, Kirsty Ellis, et al. Droid: A large-scale in-the-wild robot manipulation dataset, 2024. 1, 2, 3

[12] Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Matt Deitke, Kiana Ehsani, Daniel Gordon, Yuke Zhu, et al. Ai2-thor: An interactive 3d environment for visual ai. *arXiv preprint arXiv:1712.05474*, 2017. 2, 3

[13] Chengshu Li, Ruohan Zhang, Josiah Wong, Cem Gokmen, Sanjana Srivastava, Roberto Martín-Martín, Chen Wang, Gabrael Levine, Michael Lingelbach, Jiankai Sun, et al. Behavior-1k: A benchmark for embodied ai with 1,000 everyday activities and realistic simulation. In *Conference on Robot Learning*, pages 80–93. PMLR, 2023. 2, 3

[14] Ajay Mandlekar, Soroush Nasiriany, Bowen Wen, Iretiayo Akinola, Yashraj Narang, Linxi Fan, Yuke Zhu, and Dieter Fox. Mimicgen: A data generation system for scalable robot learning using human demonstrations. In *Conference on Robot Learning*, pages 1820–1864. PMLR, 2023. 2

[15] Mayank Mittal, Calvin Yu, Qinxi Yu, Jingzhou Liu, Nikita Rudin, David Hoeller, Jia Lin Yuan, Ritvik Singh, Yunrong Guo, Hammad Mazhar, Ajay Mandlekar, Buck Babich, Gavriel State, Marco Hutter, and Animesh Garg. Orbit: A unified simulation framework for interactive robot learning environments. *IEEE Robotics and Automation Letters*, 8(6): 3740–3747, 2023. 2

[16] Austin Myers, Ching L. Teo, Cornelia Fermüller, and Yiannis Aloimonos. Affordance detection of tool parts from geometric features. In *ICRA*, 2015. 1, 2, 3

[17] Soroush Nasiriany, Abhiram Maddukuri, Lance Zhang, Adeet Parikh, Aaron Lo, Abhishek Joshi, Ajay Mandlekar, and Yuke Zhu. Robocasa: Large-scale simulation of everyday tasks for generalist robots. In *Robotics: Science and Systems*, 2024. 1, 2, 3

[18] OpenAI. Gpt-4o system card, 2024. 2

[19] Abby O'Neill, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, Ajinkya Jain, et al. Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6892–6903. IEEE, 2024. 1, 2, 3

[20] Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M. Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *International Conference on Computer Vision (ICCV) 2021*, 2021. 2, 3

[21] Nur Muhammad Shafiullah and Daniel San José Pro. DiT-Flow policy in LeRobot. GitHub Pull Request #680, https://github.com/huggingface/lerobot/pull/680, 2024. Accessed: 2026-01-26. 4

[22] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild, 2012. 1

[23] Stone Tao, Fanbo Xiang, Arth Shukla, Yuzhe Qin, Xander Hinrichsen, Xiaodi Yuan, Chen Bao, Xinsong Lin, Yulin Liu, Tse-kai Chan, Yuan Gao, Xuanlin Li, Tongzhou Mu, Nan Xiao, Arnav Gurha, Zhiao Huang, Roberto Calandra, Rui Chen, Shan Luo, and Hao Su. Maniskill3: Gpu parallelized robotics simulation and rendering for generalizable embodied ai. *arXiv preprint arXiv:2410.00425*, 2024. 2

[24] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. Caltech-ucsd birds-200-2011. Technical Report CNS-TR-2011-001, 2011. 2, 3

[25] Yufei Wang, Zhou Xian, Feng Chen, Tsun-Hsuan Wang, Yian Wang, Katerina Fragkiadaki, Zackory Erickson, David Held, and Chuang Gan. Robogen: Towards unleashing infinite data for automated robot learning via generative simulation, 2023. 2

[26] Fanbo Xiang, Yuzhe Qin, Kaichun Mo, Yikuan Xia, Hao Zhu, Fangchen Liu, Minghua Liu, Hanxiao Jiang, Yifu Yuan, He Wang, Li Yi, Angel X. Chang, Leonidas J. Guibas, and Hao Su. SAPIEN: A simulated part-based interactive environment. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1, 2, 3, 4

[27] Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. *arXiv preprint arXiv:1711.00199*, 2017. 1

[28] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34: 12077–12090, 2021. 4

[29] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2023. 2, 3

[30] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural motifs: Scene graph parsing with global context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5831–5840, 2018. 1