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

Maximilian Xiling Li, Paul Mattes, Nils Blank, Korbinian Rudolf, Paul Lödige, Rudolf Lioutikov
Intuitive Robots Lab, Karlsruhe Institute of Technology, Germany
{maximilian.li, lioutikov}@kit.edu

***Abstract*—We introduce the Multi-Object Photoreal Simulation (MOPS) dataset, addressing the lack of computer vision datasets specifically designed for robot manipulation. MOPS provides photorealistic simulated environments with comprehensive ground truth annotations, using a zero-shot asset augmentation pipeline based on large language models. This pipeline annotates 3D assets at the part level and normalizes assets across libraries. The dataset delivers pixel-level segmentation for critical robotics tasks including part segmentation and affordance prediction. By combining detailed annotations with photorealistic simulation, MOPS generates diverse indoor scenes to accelerate progress in robot perception, manipulation, and interaction with real-world environments. The dataset and generation framework will be made publicly available.**

I. Introduction

Machine learning methods in computer vision rely on task-specific datasets spanning various applications from affordance segmentation [16], 3D Part Segmentation [3], Scene Graph Generation (SGG) [29], to 6D pose estimation [26]. While these datasets have driven significant advancements, they predominantly feature static scenes without temporal interaction sequences. Additionally, the robotics domain remains critically underrepresented despite embodied agents requiring robust environmental perception for effective autonomous operation.

Ideally, datasets for learning vision for robotic manipulation should fulfill several key requirements:

Req Obj: Manipulation relevant *Objects*: Common household items found in living spaces.

Req Ann: Manipulation relevant *Annotations*: High-resolution labels including *part information*, *affordance labels 6D poses*.

Req Rep: Manipulation relevant *Representations*: Beyond images, incorporating other scene representations such as pointclouds or scene graphs.

Req Env: Manipulation relevant and realistic *Environments*: Photorealistic rendering or real-world setups with natural clutter, beyond controlled laboratory conditions.

Req Int: Manipulation relevant *Interactions*: Capturing agent-agent, agent-object and object-object interactions over time, ideally supporting direct policy evaluation.

We propose a new dataset generation framework with pixel-level ground truth for **M**ulti-**O**bjective **P**hotoreal **S**imulation (MOPS) addresses all of these requirements, unlike existing datasets which satisfy only partial subsets (see Table I). The MOPS dataset generator bridges the gap between interactive robotics datasets and high-quality vision annotations. MOPS uses assets from PartNet-Mobility [25] and RoboCasa [17] to create scenes with articulated household objects (Req Obj) in photorealistically rendered scenes (Req Env). MOPS leverages a zero-shot asset augmentation pipeline built on GPT-4o [18] to normalize assets, and to provide manipulation relevant annotations such as affordances (Req Ann). MOPS provides pixel-level ground truth for class, part and instance segmentations alongside affordance labels (see Figure 2), geometric information (normal maps, 6D poses), multiple sensor modalities (RGB-D, pointclouds) and can generate scene graphs (Req Rep) with an LLM on demand. MOPS uses the Maniskill3 [22] simulator to enable dynamic, interactive scenes suitable for evaluating learned robot behavior or recording teleoperated demonstrations (Req Int) - making it valuable for both vision and robotics communities.

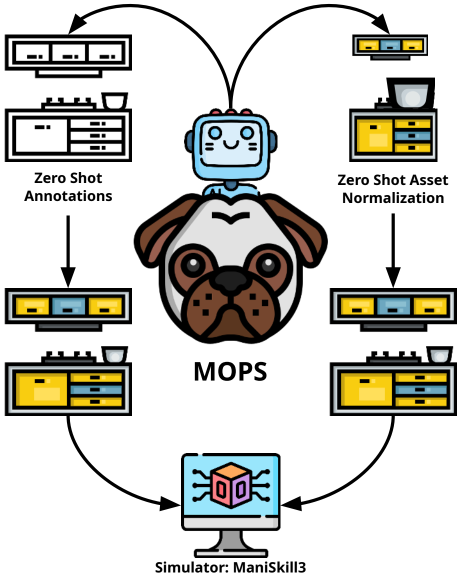


Fig. 1: MOPS provides labeled, realistic data for robotics and vision tasks through Large Language Models-enabled zero-shot annotation and normalization of 3D assets, which are then used to create new indoor scenarios for data collection.

Dataset	*Objects*	*Annotations*	*Representations*	*Environment*	*Interaction*	*Robot Trajectories*
Vision Datasets						
CUB-200-2011 [23]		P*	R			
CityScapes [4]		S+I	R			
SemanticKITTI [2]		S+I	R+D+M			
Visual Genome [12]		A	R+G			
PSG [27]		S+A	R+S+G			
ScanNet++ [28]	✓	S+I	R+D+M	✓		
HyperSim [20]	✓	S+I	R+D+M	✓		
RGB-D Part Aff. [16]	✓	A	R+D			
3D AffordanceNet [5]	✓	A	M			
PartNet-Mobility [25]	✓	P	M			
Robotics Datasets						
Open-X [19]	✓		R+D	✓		✓
DROID [10]	✓		R+D	✓		✓
AI2-THOR [11]	✓	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+G**	✓	✓	(✓)

TABLE I: Comparison of different computer vision and robotics datasets and their relevance to robot manipulation.
S: Semantic Segmentation, I: Instance Segmentation, P: Part Segmentation, P*: Part Center Points, A: Affordance Segmentation, R: RGB, D: Depth, M: 3D Meshes or Pointclouds, G: Scene Graphs. MOPS is compatible with demonstrations by RoboCasa, but does not provide new robot trajectories.

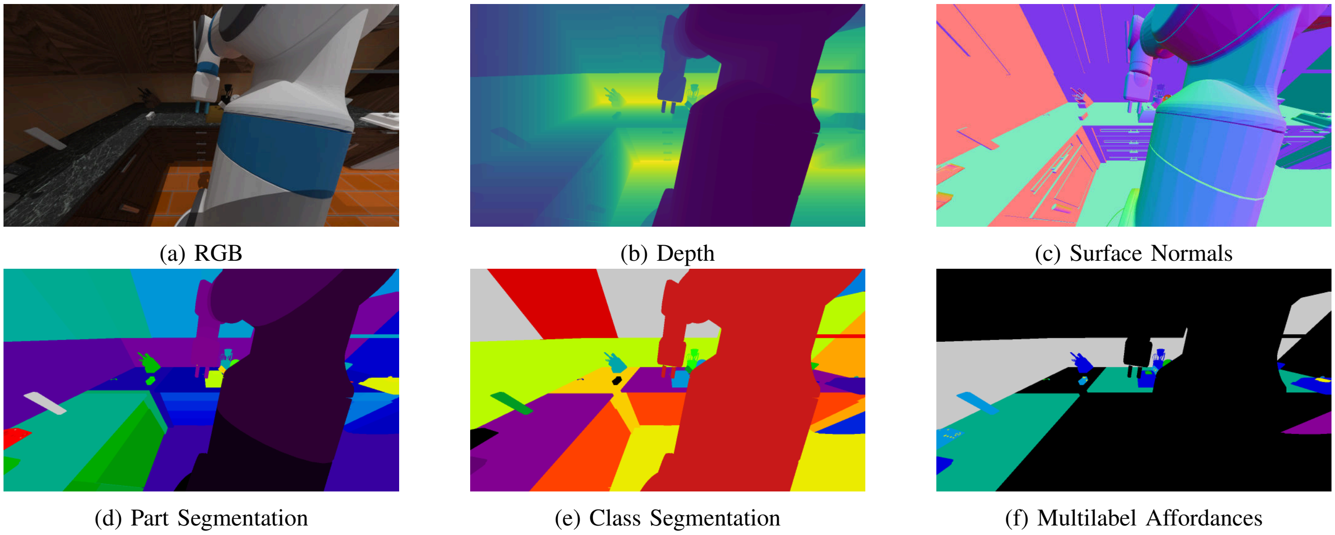
II. Related Work

Vision Datasets: The computer vision community has developed specialized datasets for tasks like image classification (CUB-200-2011 [23]), semantic segmentation (Cityscapes [4], SemanticKITTI [2]), and scene graph generation (Visual Genome (VG) [12], Panoptic Scene Graphs (PSG) [27]). However, these datasets are less relevant to robot manipulation (Req Obj). Indoor scene datasets like ScanNet++ [28] and Hypersim [20] provide RGB-D data with semantic annotations but lack affordances and 6D poses (Req Ann). The RGB-D Part Affordance dataset [16] offers affordance annotations but lacks environmental realism (Req Env) and interactivity (Req Int). Our MOPS dataset addresses these issues by providing procedural, synthetic scenes of cluttered indoor environments with multiple, pixel-wise ground truth annotations.

3D Datasets: Datasets with 3D models could be used for creating an interactive simulation. 3D AffordanceNet [5] provides 3D point clouds with affordance labels (Req Obj& Req Ann) but lacks material information for photorealistic rendering (Req Env). PartNet-Mobility [25] includes material information and articulation but lacks affordances. A critical limitation of these 3D assets is that they are not modeled to a common reference scale, appearing disproportionate relative to robotic manipulators. We propose a zero-shot asset augmentation pipeline based on an LLM to enrich the PartNet-Mobility assets with affordance annotations and realistic scale ranges (where 1.0 simulation units equal 1.0 meters) to prepare them for simulation in ManiSkill3 [22].

Robotics Datasets: Real-world robotics datasets like Open X-Embodiment [19] and DROID [10] pair sensor inputs with robot trajectories but exhibit inconsistent data quality and prohibitive scaling costs. Simulation frameworks like AI2-THOR [11], OmniGibson [13], and RoboCasa [17] provide photorealistic environments (Req Env) with advanced physics capabilities. While generative frameworks such as RoboGen [24] and Genesis [1] generate diverse scenes, they are limited in generating cluttered environments. Despite simulated datasets offering ground truth annotations (Req Ann) and multiple representations (Req Rep), they generally lack comprehensive pixel-wise annotations for affordances, semantic concepts, and scene graphs.

Our MOPS dataset addresses these limitations by combining RoboCasa's scene variety with zero-shot augmented assets from PartNet-Mobility, providing unlimited realistic scenes with pixel-wise ground truth annotations (including affordances) in cluttered environments across multiple representations. Built on ManiSkill3 [22], MOPS improves visual quality and generation speed through raytracing and GPU parallelization. Table I compares current datasets against manipulation-relevant dataset requirements.


(a) RGB

(b) Depth

(c) Surface Normals

(d) Part Segmentation

(e) Class Segmentation

(f) Multilabel Affordances

Fig. 2: MOPS provides multiple, pixelwise ground truth maps in addition to RGB-D perception.

III. Zero-Shot 3D Asset Augmentation

MOPS creates photorealistic indoor environments for robot manipulation (Req Env) using the ManiSkill3 [22] simulator and renderer. To populate the scenes with realistic and interactive objects (Req Obj) MOPS uses 3D assets from RoboCasa [17] and PartNet-Mobility [25]. However, these assets require augmentation to address two key issues: inconsistent reference scales that compromise realism, and the need for affordance annotations to generate pixel-wise ground truth masks. MOPS resolves both challenges through a zero-shot augmentation pipeline powered by GPT-4o [18].

*A. Zero-Shot Asset Normalization (*Req Obj*)*

Most 3D assets, even from a singular collection like Partnet-Mobility [25], are not modeled on the same reference scale, thus creating unrealistic size relationships between objects or with simulated robots (see Figure 3). We address this through zero-shot asset normalization. First we confirm the simulations reference scale of 1.0 simulation units equaling 1.0 meters by verifying the Franka Emika 7-DoF arm's Denavit-Hartenberg parameters [6]. For each PartNet-Mobility asset (which includes XYZ-bounding boxes and category labels), we leverage GPT-4o's common-sense knowledge to obtain realistic minimum and maximum Width × Height × Depth dimensions. Since asset orientation is unknown, we calculate scaling factors by dividing the largest bounding box dimension by the largest WHD dimension. When loading assets, we randomly sample from a uniform distribution within the realistic range, increasing object variety.

*B. Zero-Shot Affordance Annotation (*Req Ann*)*

MOPS again leverages LLM common-sense reasoning to generate multi-label affordances at both part and object levels. GPT-4o outputs affordance lists for each object and, if available, object part (see Figure 3). To improve label quality, we cluster the affordance lists using sentence embeddings to eliminate duplicates (e.g., *closable* / *close*) and align semantic clusters (e.g., *heatable* / *warmup-able*). While future work could extend this to region-level affordances similar to the manually labeled 3D AffordanceNet [5], our zero-shot annotation approach already significantly reduces human labeling efforts while also providing one of the largest affordance datasets with a total of **56** affordance labels for **23,048** 3D parts of **2,346** objects across **46** object categories (PartNet-Mobility [25]) and additional object level annotations for **1,008** objects from **101** categories (RoboCasa [17]).

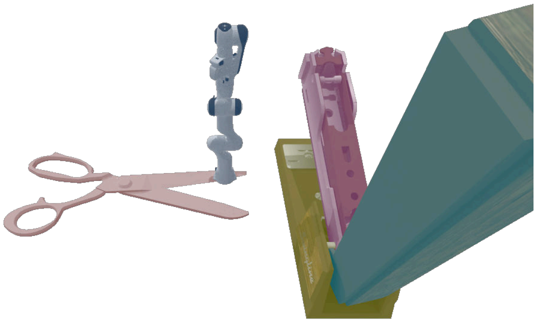


Fig. 3: MOPS uses an LLM to generate part-level and object-level affordances for 3D assets. The assets are not normalized in scale for better visibility.

IV. MOPS Dataset Generation

MOPS generates virtually unlimited simulation scenes by combining RoboCasa's 120 realistic indoor environments with 2,300 PartNet-Mobility articulated objects and zero-shot asset augmentations. To increase the relevance for robot manipulation, MOPS provides the following technical enhancements.

*A. MOPS Ground Truth Masks (*Req Ann*)*

MOPS provides multiple scene representations for robot manipulation vision tasks. Various camera positions include birdseye views for SLAM reconstruction. External over-the-shoulder, ego, and in-hand cameras mimick typical robot setups. The virtual cameras deliver standard image modalities: raytraced RGB for realism or rasterized rendering for speed, depth images in millimeters for RGB-D inputs or depth estimation training, surface normal maps and part-level segmentation.

Additional ground truth includes 6D object poses, instance and semantic segmentation masks (generated via look-up tables populated during scene creation), and pixel-perfect affordance annotations derived from the zero-shot asset augmentation pipeline. Figure 2 presents the different camera modalities from an ego camera in a RoboCasa scene.

*B. MOPS Extended Representations (*Req Rep*)*

MOPS generates scene graphs on demand via LLM queries, leveraging pixel-wise affordance annotations and object data. The ManiSkill3 simulation also creates point clouds by merging multi-camera 2.5D images. Camera parameters and lighting configurations are fully customizable. The SAPIEN renderer in ManiSkill3 provides realistic stereo depth sensors with active IR lighting and simulated noise [21]. Though not yet available in the current ManiSkill3 release[1], integrating these sensors into MOPS will be straightforward based on experience with ManiSkill2.

*C. MOPS Realistic Environments (*Req Env*)*

MOPS generates a vision dataset of realistic environments relevant to robot manipulation using RoboCasa assets and ManiSkill3's photorealistic raytraced rendering. To create cluttered scenes with object overlap and distractors (see Figure 4), MOPS procedurally places augmented PartNet-Mobility assets in kitchen scenes by: (1) identifying countertop locations, (2) computing available space from collision mesh bounding boxes, and (3) randomly positioning objects within this space. While more heuristic than trained approaches such as ClutterGen [8], this method can be easily applied to novel environments without requiring any training.

[1] ManiSkill v3.0.0b20, retrieved 2025-04-28 from GitHub, Commit *256343*

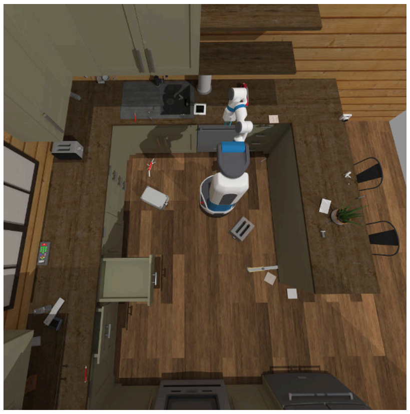


Fig. 4: A RoboCasa kitchen filled with PartNet-Mobility clutter.

*D. Interactive Simulation (*Req Int*)*

MOPS leverages the ManiSkill3 simulation to provide full interactivity for testing robot policies. New demonstrations can be recorded using a teleoperation interfaces based on keyboard and mouse, or rudimentary tracking of VR controllers using the Meta Quest 3 (see Figure 5). To further increase compatibility with the existing RoboCasa demonstrations, we added the same augmentations to the RoboSuite / MuJoCo Simulation underlying RoboCasa [17].

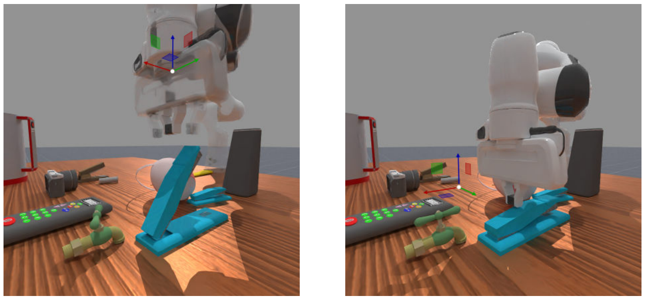


Fig. 5: MOPS offers mouse-and-keyboard and a simple VR controller teleoperation interface. The semitransparent robot arm indicates to the user the new target position.

V. Conclusion

We introduce MOPS, a dataset generation pipeline for robot vision learning that provides photorealistic renderings of household objects with comprehensive pixel-level annotations for robotics-relevant tasks (segmentation, affordances, 6D poses). By combining PartNet-Mobility and RoboCasa assets, MOPS generates cluttered objects in realistic kitchen scenes with full ManiSkill3 interaction capabilities. This combination of high-quality visual data and rich annotations bridges the gap between computer vision datasets and robotics requirements.

Our zero-shot asset augmentation pipeline leverages GPT-4o to create one of the most diverse affordance datasets (50+ labels across 100+ object categories), easily extendable to new assets or semantic labels. This automated approach significantly reduces the manual annotation burden typically associated with creating robotics-relevant datasets, enabling rapid scaling to new domains and tasks.

Limitations. MOPS inherits ManiSkill3's constraints, notably the inability to use raytraced rendering with GPU parallelization. Future work could adapt MOPS to frameworks like Isaac Lab [14], Genesis [1] or RoboVerse [7] for further improved rendering and parallelization.

Future Work. Extending our augmentation pipeline to generate materials and textures would enable utilization of additinal asset libraries such as PartNet [15] and ShapeNet [3]. Additionally, future work will explore full teleoperation interfaces like IRIS [9] for recording new robot trajectories in AR/VR.

Acknowledgement. The research presented in this paper was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – 448648559.

References

[1] Genesis Authors. Genesis: A universal and generative physics engine for robotics and beyond, December 2024. URL https://github.com/Genesis-Embodied-AI/Genesis.

[2] 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.

[3] 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.

[4] 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)*, June 2016.

[5] 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.

[6] Franka Robotics. Franka Denavit–Hartenberg parameters. URL https://frankaemika.github.io/docs/control_parameters.html#denavithartenberg-parameters. Franka FCI Documentation.

[7] Haoran Geng, Feishi Wang, Songlin Wei, Yuyang Li, Bangjun Wang, Boshi An, Charlie Tianyue Cheng, Haozhe Lou, Peihao Li, Yen-Jen Wang, Yutong Liang, Dylan Goetting, Chaoyi Xu, Haozhe Chen, Yuxi Qian, Yiran Geng, Jiageng Mao, Weikang Wan, Mingtong Zhang, Jiangran Lyu, Siheng Zhao, Jiazhao Zhang, Jialiang Zhang, Chengyang Zhao, Haoran Lu, Yufei Ding, Ran Gong, Yuran Wang, Yuxuan Kuang, Ruihai Wu, Baoxiong Jia, Carlo Sferrazza, Hao Dong, Siyuan Huang, Koushil Sreenath, Yue Wang, Jitendra Malik, and Pieter Abbeel. Roboverse: Towards a unified platform, dataset and benchmark for scalable and generalizable robot learning, April 2025. URL https://github.com/RoboVerseOrg/RoboVerse.

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

[9] Xinkai Jiang, Qihao Yuan, Enes Ulas Dincer, Hongyi Zhou, Ge Li, Xueyin Li, Julius Haag, Nicolas Schreiber, Kailai Li, Gerhard Neumann, and Rudolf Lioutikov. Iris: An immersive robot interaction system. *arXiv preprint arXiv:2502.03297*, 2025.

[10] 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.

[11] 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.

[12] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017.

[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.

[14] 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. doi: 10.1109/LRA.2023.3270034.

[15] Kaichun Mo, Shilin Zhu, Angel X. Chang, Li Yi, Subarna Tripathi, Leonidas J. Guibas, and Hao Su. PartNet: A large-scale benchmark for fine-grained and hierarchical part-level 3D object understanding. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

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

[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.

[18] OpenAI. Gpt-4o system card, 2024. URL https://arxiv.org/abs/2410.21276.

[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.

[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.

[21] Sapien Developers. Sapien 3.0 documentation, 2024. URL https://sapien-sim.github.io/docs/user_guide/rendering/depth_sensor.html. Sapien StereoDepthSensor Documentation.

[22] 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.

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

[24] 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.

[25] 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)*, June 2020.

[26] 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.

[27] Jingkang Yang, Yi Zhe Ang, Zujin Guo, Kaiyang Zhou, Wayne Zhang, and Ziwei Liu. Panoptic scene graph generation, 2022. URL https://arxiv.org/abs/2207.11247.

[28] 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.

[29] 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.

Appendix: Example Single Object Images

Figure 6 shows example observations from the single object configuration, which mimics single object datasets with uniform backgrounds like RGB-D Part Affordance [16].

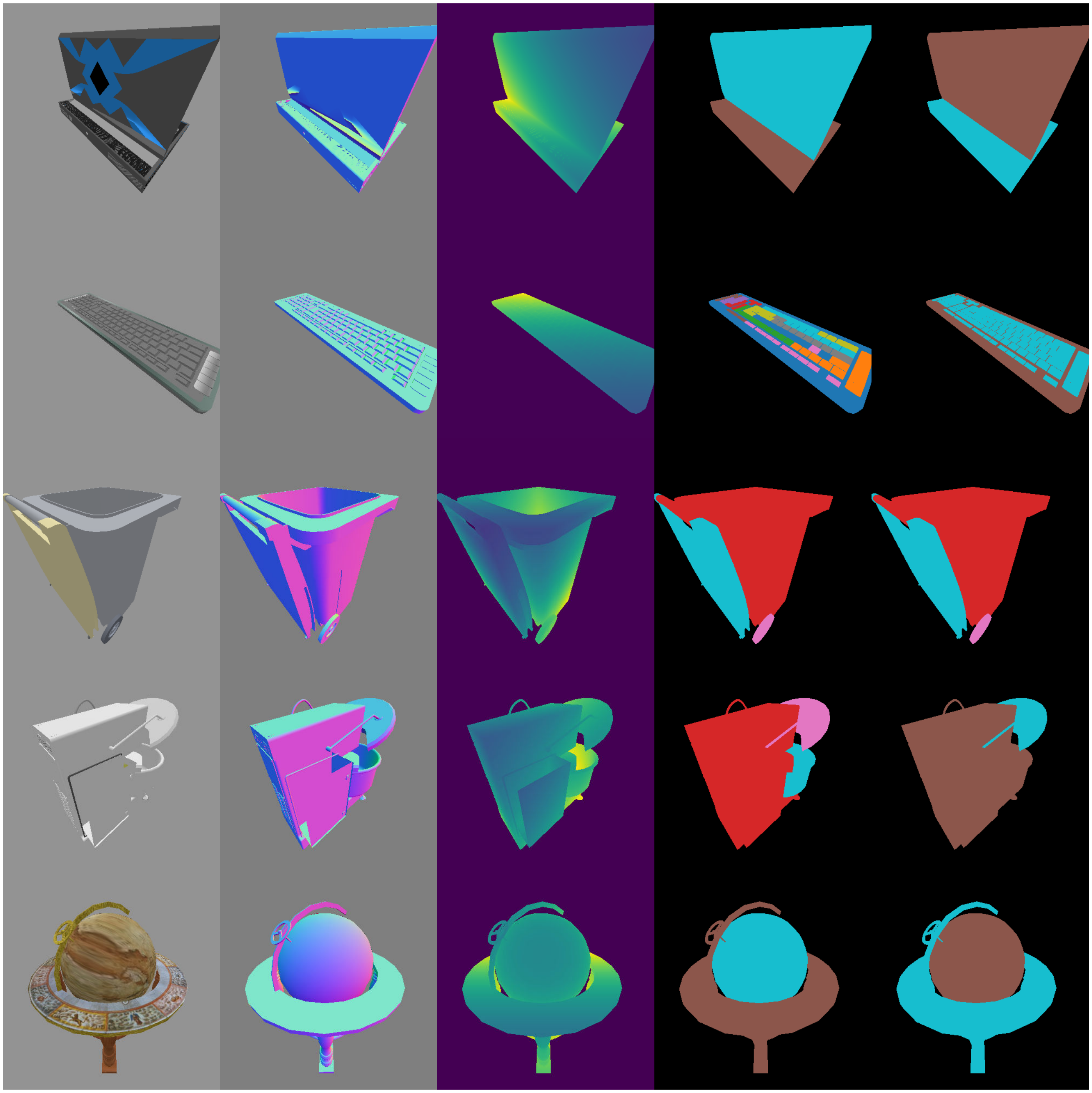


Fig. 6: Single Object images. From Left to Right: RGB Image, Normal Map, Depth Image, Part Segmentation, Affordance Segmentation. Please note that the colors for Affordance Segmentation visualizaton only provide contrast and do not share meaning across images.

Appendix: Example Tabletop Clutter Images

Figure 7 shows example observations from randomly generated, cluttered tabletop scenes. These images show interactive scenes including the robot base, ready for learning robot behavior. For a purely vision-based dataset, the environment geometry with table, floor, robot and background could be easily disabled.

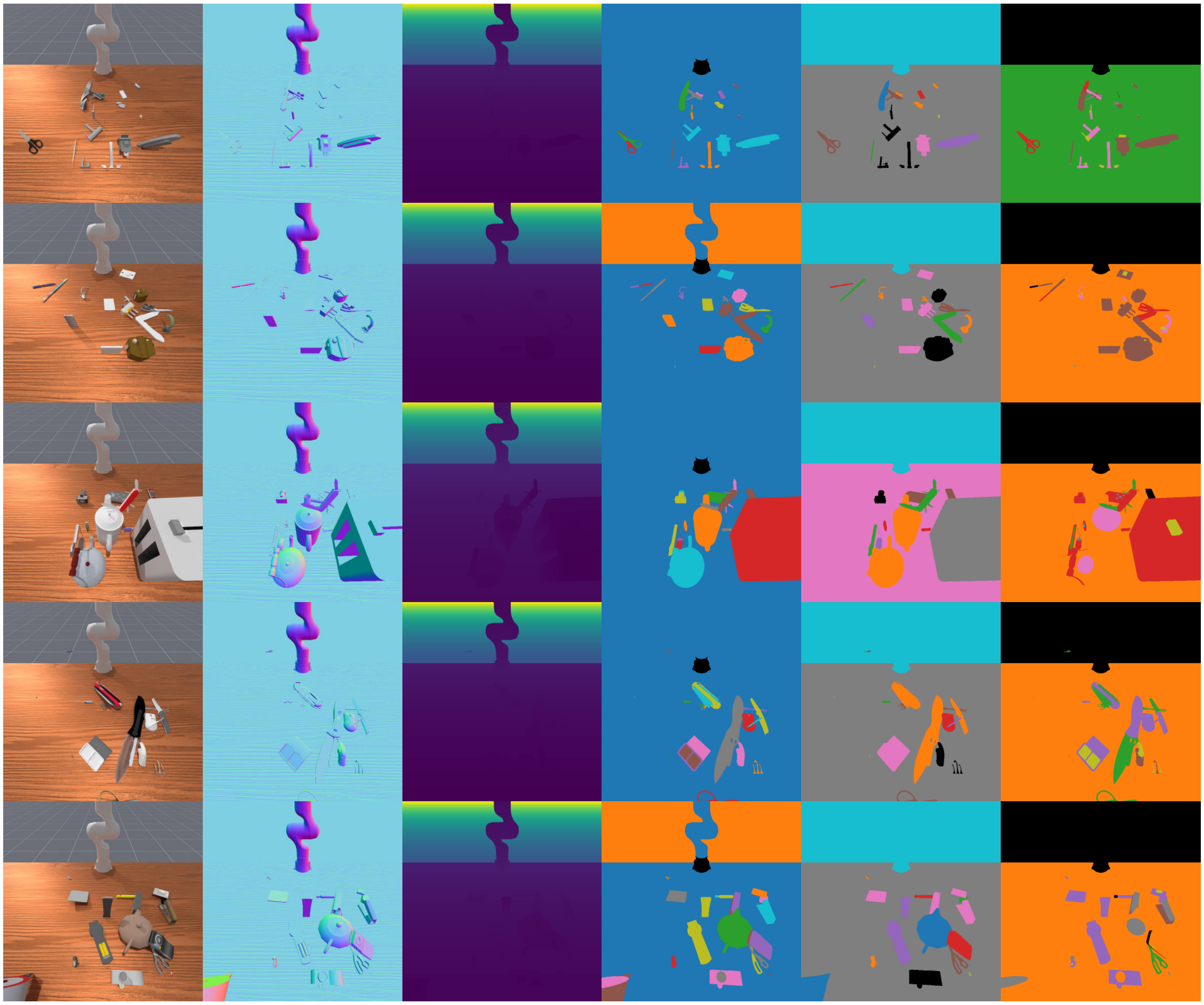


Fig. 7: Cluttered Tabletop. From Left to Right: RGB Image, Normal Map, Depth Image, Part Segmentation, Class Segmentation, Affordance Segmentation. Please note that the colors for segmentation visualizations only provide contrast and do not share meaning across images. The depth images lack visual detail in this illustration, due to the floor in the background going towards infinity.

APPENDIX: EXAMPLE CLUTTERED KITCHEN IMAGES

Figure 8 shows example observations from the cluttered RoboCasa kitchen configuration.

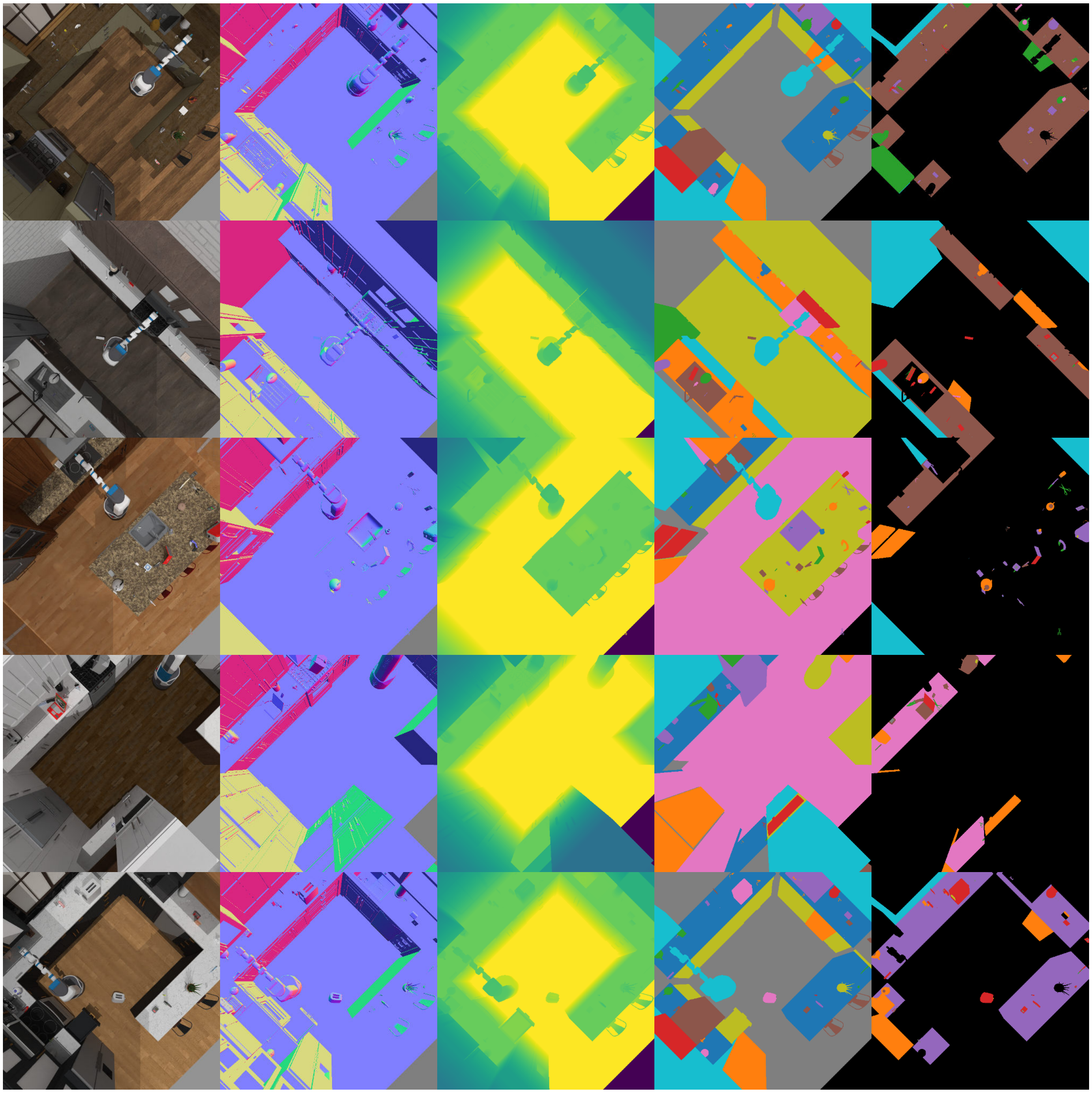


Fig. 8: RoboCasa Kitchens with clutter. From Left to Right: RGB Image, Normal Map, Depth Image, Class Segmentation, Affordance Segmentation. Please note that the colors for Affordance Segmentation visualizaton only provide contrast and do not share meaning across images.