SO-101 Bench: Measuring the Gap Between Semantic and Geometric Competence in Vision-Language-Action Models

Truman Hickok
Independent
thicko18@yahoo.com

Abstract

Vision-language-action (VLA) models have become increasingly effective general-purpose robot policies, but it remains unclear how well their semantic competence transfers to precise geometric manipulation under novel objects and embodiments. We present SO-101 Bench, a compact benchmark for language-conditioned tabletop manipulation with an SO-101 arm. The benchmark evaluates four tasks across 56 household objects spanning seen, unseen/seen-class, and unseen/unseen-class regimes. We collect over 4,000 teleoperated demonstrations (∼20 hours), fine-tune GR00T-N1.6-3B, and evaluate the resulting policy in the real world. While the policy reaches 97.4% success on the simplest one-object bin task, performance degrades sharply under spatial constraints, clutter, and novelty, falling to 6.5% on the four-object bin task and 6.3% on the between *task for fully unseen objects. Failure analysis suggests the main bottleneck is not instruction grounding, but geometric execution: imprecise grasps, poor grasp strategies for unfamiliar shapes, weak recovery, and fragile spatial composition. We also preview the benchmark's Isaac Lab extension for scalable real-to-sim evaluation.*

1. Introduction

Vision-language-action (VLA) models offer a practical route to general-purpose robot manipulation by combining vision-language pretraining with action generation. Systems such as RT-2, Open X-Embodiment, π_0, and GR00T make this direction increasingly compelling [1, 8, 9, 11]. Yet evaluation remains difficult: real rollouts are costly, simulation is valuable only when it tracks real behavior, and benchmarks such as LIBERO, BEHAVIOR-1K, VLABench, VLA-eval, and SIMPLER expose the tradeoff between scale, realism, and diagnostic control [2, 4–6, 10].

We introduce SO-101 Bench, a compact real-world benchmark for testing whether VLAs can convert semantic understanding into precise geometric manipulation with both seen and unseen objects. Across bin placement, next-to placement, between placement, and directional movement, a fine-tuned frontier VLA often identifies the right object and coarse action, but remains brittle in grasp precision, relational geometry, and recovery behavior.

2. Benchmark and Policy

SO-101 Bench uses an SO-101 arm in a tabletop workspace with overhead and wrist RGB cameras recorded at 640×480. The tasks isolate distinct demands: bin placement tests grasping; "next-to" placement tests basic relational placement; "between" placement requires centering between referents; and directional movement requires moving an object to a scene-induced boundary.

We collect approximately 4,400 teleoperated demonstrations, or roughly 20 hours, over 30 household objects. About half cover bin placement and the rest cover spatial tasks. Scenes include same-class and same-color distractors, and instructions do not always include color terms, forcing reliance on language, object identity, and spatial context.

We fine-tune GR00T-N1.6-3B with the official GR00T codebase [8]. To mitigate arm occlusion, the policy receives the current overhead and wrist frames plus the episode's first overhead frame as a simple memory cue. We cross-validate 20k, 35k, 55k, and 75k checkpoints and use the 55k checkpoint for evaluation.

Evaluation covers 56 objects: 24 seen objects used in fine-tuning, 17 unseen objects from seen classes, and 15 unseen objects from novel classes. All unseen objects are teleoperation-verified as graspable. Each target receives at most three grasp attempts; trials fail if the bin moves more than one inch, a non-target object moves more than 0.5 inches, or the task-specific geometric constraint is violated.

3. Results and Analysis

Figure 1 summarizes the main results. The policy succeeds in easy familiar-object settings, reaching 97.4% one-object

Figure 1. Task success rates across object generalization regimes. ”Bin” = ”Place each object in the plastic bin”, ”Next to” = ”Place [object 1] next to [object 2]”, ”Between” = ”Place [object 1] between [object 2] and [object 3]”, ”Move” = ”Move [object 1] [direction]”.

	Next to			Between			Move [direction]		
Failure type	**S**	**U / S**	**U / U**	**S**	**U / S**	**U / U**	**S**	**U / S**	**U / U**
placed on top of object(s)	14	6	3	19	11	2	9	7	4
semantic error	6	11	13	3	10	17	3	3	5
imprecise grasp	8	7	7	10	8	9	8	5	7
not close enough	14	12	2	—	—	—	4	3	1
not centered enough	—	—	—	18	7	5	—	—	—
moved object(s)	5	4	2	2	6	1	4	3	—
bad grasp strategy	1	1	4	2	4	6	2	2	4
other	2	1	5	3	3	2	1	1	8
not close	—	—	1	10	10	2	—	—	—
too close to referent	—	—	—	5	—	1	—	—	—
moved past boundary	—	—	—	—	—	—	1	—	1
Total failures	**50**	**42**	**37**	**72**	**59**	**45**	**32**	**24**	**30**

Table 1. Failure frequencies for the three spatial instruction-following tasks. ”S” indicates ”Seen” while ”U” indicates ”Unseen” (e.g ”U/S” means unseen object, seen class).

bin-placement success on seen objects and 92.5% on unseen objects from seen classes. Fully unseen objects fall to 57.5%, showing that object novelty remains a major bottleneck even for simple grasping.

Clutter and spatial precision widen this gap. Four-object bin placement drops from 78.5% on seen objects to 48.4% on unseen/seen-class objects and 6.5% on unseen/unseen-class objects. Spatial tasks are harder still: next-to placement reaches 55.0%, 46.3%, and 25.0%; between placement reaches 43.9%, 36.4%, and 6.3%; and directional movement reaches 56.9%, 51.1%, and 33.3% across the three regimes. Directional movement outperforming between placement suggests that the policy often selects the right coarse action but struggles with centering, orientation, and object-extent reasoning.

Table 1 shows that familiar-object failures are mostly geometric or physical: placing targets on referents, missing distance or centering constraints, imprecise grasps, and unintended object motion. Under full object novelty, semantic errors and bad grasp strategies increase, but many failures still reflect poor closed-loop visuospatial control, occlusion handling, object reorientation, and recovery after partial failure.

Figure 2. Real image from the dataset (left) and an image from our work-in-progress Isaac Lab environment (right).

4. Discussion

SO-101 Bench is small in hardware footprint but diagnostic in structure, combining language grounding, distractors, clutter, relational constraints, and explicit object-generalization splits. The results show that current VLAs can appear semantically competent while failing at the geometric and physical details needed for reliable manipulation.

The benchmark is the first stage of a larger real-to-sim evaluation pipeline. Figure 2 shows the current real-sim comparison between a dataset image and our work-in-progress Isaac Lab reconstruction, which is the central product of the broader project. Using Gaussian-splat-based reconstruction and scanned object assets, we aim to evaluate the same policy in Isaac Lab, measure real-to-sim correspondence, and scale SO-101 Bench to larger object sets [3, 7]. The failure modes in Table 1 motivate this direction directly: future VLAs need stronger memory under occlusion, precise perception, visuospatial planning, tactile grounding, and robust recovery behaviors.

References

[1] Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Lucy Xiaoyang Shi, James Tanner, Quan Vuong, Anna Walling, Haohuan Wang, and Ury Zhilinsky. π_0: A Vision-Language-Action Flow Model for General Robot Control. *arXiv preprint arXiv:2410.24164*, 2024. 1

[2] Suhwan Choi, Yunsung Lee, Yubeen Park, Chris Dongjoo Kim, Ranjay Krishna, Dieter Fox, and Youngjae Yu. vla-eval: A Unified Evaluation Harness for Vision-Language-Action Models. *arXiv preprint arXiv:2603.13966*, 2026. 1

[3] Arhan Jain, Mingtong Zhang, Kanav Arora, William Chen, Marcel Torne, Muhammad Zubair Irshad, Sergey Zakharov, Yue Wang, Sergey Levine, Chelsea Finn, Wei-Chiu Ma, Dhruv Shah, Abhishek Gupta, and Karl Pertsch. PolaRiS: Scalable Real-to-Sim Evaluations for Generalist Robot Policies. *arXiv preprint arXiv:2512.16881*, 2025. 2

[4] Chengshu Li, Ruohan Zhang, Josiah Wong, Cem Gokmen, Sanjana Srivastava, Roberto Martín-Martín, Chen Wang, Gabrael Levine, Michael Lingelbach, Jiankai Sun, Mona Anvari, Minjune Hwang, Manasi Sharma, Arman Aydin, Dhruva Bansal, Samuel Hunter, Kyu-Young Kim, Alan Lou, Caleb R. Matthews, Ivan Villa-Renteria, Jerry Huayang Tang, Claire Tang, Fei Xia, Silvio Savarese, Hyowon Gweon, C. Karen Liu, Jiajun Wu, and Li Fei-Fei. BEHAVIOR-1K: A Benchmark for Embodied AI with 1,000 Everyday Activities and Realistic Simulation. In *Proceedings of The 6th Conference on Robot Learning*, pages 80–93. PMLR, 2023. 1

[5] Xuanlin Li, Kyle Hsu, Jiayuan Gu, Oier Mees, Karl Pertsch, Homer Rich Walke, Chuyuan Fu, Ishikaa Lunawat, Isabel Sieh, Sean Kirmani, Sergey Levine, Jiajun Wu, Chelsea Finn, Hao Su, Quan Vuong, and Ted Xiao. Evaluating Real-World Robot Manipulation Policies in Simulation. In *Proceedings of The 8th Conference on Robot Learning*, pages 3705–3728. PMLR, 2025.

[6] Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. LIBERO: Benchmarking Knowledge Transfer for Lifelong Robot Learning. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. 1

[7] Mayank Mittal, Pascal Roth, James Tigue, Antoine Richard, Octi Zhang, Peter Du, Antonio Serrano-Muñoz, Xinjie Yao, René Zurbrügg, Nikita Rudin, Lukasz Wawrzyniak, Milad Rakhsha, Alain Denzler, Eric Heiden, Ales Borovicka, Ossama Ahmed, Iretiayo Akinola, Abrar Anwar, Mark T. Carlson, Ji Yuan Feng, Animesh Garg, Renato Gasoto, Lionel Gulich, Yijie Guo, et al. Isaac Lab: A GPU-Accelerated Simulation Framework for Multi-Modal Robot Learning. *arXiv preprint arXiv:2511.04831*, 2025. 2

[8] NVIDIA GEAR Team, Allison Azzolini, Johan Bjorck, Valts Blukis, Fernando Castañeda, Rahul Chand, Yan Chang, Danyi Chen, Nikita Cherniadev, Xingye Da, Runyu Ding, Shunjia Ding, Hassan Eslami, Linxi "Jim" Fan, Yu Fang, Max Fu, Shenyuan Gao, Yunhao Ge, Fengyuan Hu, Spencer Huang, Joel Jang, Xiaowei Jiang, Yunfan Jiang, Ryan Julian, Kaushil Kundalia, Jan Kautz, Zhiqi Li, Kevin Lin, Wei Liu, Runyu Lu, Zhengyi Luo, Loic Magne, Yunze Man, Ajay Mandlekar, Abhishek Mishra, Avnish Narayan, Connor Pederson, Nadun Ranawaka, Scott Reed, Sunil Srinivasa, You Liang Tan, Guanzhi Wang, Jing Wang, Qi Wang, Shihao Wang, Jimmy Wu, Yubo Wu, Yuqi Xie, Tianyi Xiong, Mengda Xu, Yinzhen Xu, Fu-En Yang, Seonghyeon Ye, Zhiding Yu, et al. GR00T N1.6. https://research.nvidia.com/labs/gear/gr00t-n1_6/, 2025. Official NVIDIA research page. 1

[9] Abby O'Neill, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, et al. Open X-Embodiment: Robotic Learning Datasets and RT-X Models. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6892–6903. IEEE, 2024. 1

[10] Shiduo Zhang, Zhe Xu, Peiju Liu, Xiaopeng Yu, Yuan Li, Qinghui Gao, Zhaoye Fei, Zhangyue Yin, Zuxuan Wu, Yu-Gang Jiang, and Xipeng Qiu. VLABench: A Large-Scale Benchmark for Language-Conditioned Robotics Manipulation with Long-Horizon Reasoning Tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11142–11152, 2025. 1

[11] Brianna Zitkovich, Nikhil Joshi, Alex Irpan, Brian Ichter, Jasmine Hsu, Alexander Herzog, Karol Hausman, Keerthana Gopalakrishnan, Chuyuan Fu, Pete Florence, Chelsea Finn, Kumar Avinava Dubey, Danny Driess, Tianli Ding, Krzysztof Marcin Choromanski, Xi Chen, Yevgen Chebotar, Justice Carbajal, Noah Brown, Anthony Brohan, Montserrat Gonzalez Arenas, and Kehang Han. RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control. In *Proceedings of The 7th Conference on Robot Learning*, pages 2165–2183. PMLR, 2023. 1