HomeArena: Synthetic Data Generation for Fine-Tuning Vision-Language-Action Models in Service Robotics

Cris L. Froes[0009−0007−4475−4874], Gabriel A. Dorneles[0009−0001−6554−5832], Junior D. Jesus[0000−0002−8819−1168], Kristofer S. Kappel[0000−0002−9124−8540], Richard J. C. Assis[0009−0001−0911−7382], Gabriel R. de Souza[0009−0006−4704−0407], Rodrigo S. Guerra[0000−0003−4011−0901], and Paulo L.J. Drews-Jr[0000−0002−7519−0502]

NAUTEC, Center for Computational Sciences - C3,
Universidade Federal do Rio Grande - FURG, Brazil
cris.lima.froes@gmail.com
http://www.c3.furg.br

Abstract. The application of Vision-Language-Action (VLA) models to robotic manipulation in RoboCup@Home competitions faces the bottleneck of intensive data requirements of real-world teleoperation tasks. This work introduces HomeArena, a simulation framework that was designed to generate large-scale synthetic datasets for fine-tuning VLA models on object picking and articulated object manipulation. To address the challenge of generalization, PoseCondVLA was proposed as a method that concatenates the position and quaternion rotation vectors of the target object into the pretrained VLA policies. Simulated experimental results demonstrate that while standard VLA baselines struggle with spatial variations, PoseCondVLA improves success rates up to 90% and 78% in randomized pose settings for scenarios where baseline fails. Our findings validate the use of synthetic data for VLA post-training and offer a scalable path for sim-to-real deployment in service robotics.

Keywords: Vision-language-action model · Robotic manipulation · Imitation learning · Robot learning.

1 Introduction

RoboCup@Home is the largest international annual competition for autonomous service robots, focusing on developing service and assistive robot technology for future domestic applications [1]. A core challenge in the league is mobile manipulation, where robots use a software framework to autonomously detect and interact with various household objects and furniture [2]. However, applying state-of-the-art methods for learning manipulation policies faces a data bottleneck due to the time spent manually collecting teleoperated data for each new environment and task. This challenge is difficult because the ability to manipulate articulated objects, such as drawers, doors, and household devices, is highly relevant for various tasks performed during competitions.

To address this, applying methods like MimicGen [3] for automated demonstration generation has the potential to reduce the time spent collecting data with the real robot. However, it requires simulated environments that are faithful to the robot's real environment to reduce the Sim2Real gap between the data used to train the model and the data the model will encounter during inference [4, 5].

Parallel to these developments, in recent years, there has been a growing interest in applying Large Language Models (LLMs) and their multimodal variants, Vision-Language Models (VLMs), in the field of robotic learning [6]. In the past, these models were limited to text tokens as their primary output modality, being unable to generate high-dimensional continuous actions that are necessary for end-to-end robotic control, such as commanding the end-effector pose of a robotic arm. This gap has been bridged by a new category of models known as Vision-Language-Action (VLA) models. VLAs extend vision-language models to generate robotic arm control actions, typically through supervised fine-tuning on datasets of trajectories collected via teleoperation [7, 8].

The relative scarcity of robotic data composed of action trajectories collected by an expert, compared to the wide availability of datasets for vision and language modalities, creates an opportunity to develop tools that use simulated data to ease the requirements for real data introduced by VLA models. Associated with this factor, the availability of pre-existing public datasets with articulated 3D models of various domestically relevant categories, such as PartNet Mobility [9] and ArtVIP [10], enables the use of these asset repositories as a data source for training VLA models for articulated object manipulation.

The two main contributions of the work are as follows:

- **HomeArena**: A simulation framework for generating large-scale datasets for supervised fine-tuning of VLA models in service robotics, more specific for object picking and articulated object manipulation tasks.
- **PoseCondVLA**: A simple but effective method for injecting object-centric pose observations into a pretrained VLA policy. We demonstrate that this approach improves the policy's capability to learn from and generalize to novel object poses in the workspace of the robotic arm.

Ultimately, the work is divided into the following sections: Section 2 presents the related works in the area. Section 3 describes the methodology applied. Section 4 shows the collected results. Section 5 discusses the results. Finally, Section 6 draws conclusions and suggests future work.

2 Related Work

Several works have studied articulated object manipulation using more specialized models rather than VLAs. In ArticuBot [11], the development of a generalist policy for manipulation of any type of articulated object was studied. This work resembles ours by using trajectories generated in simulation as a data source

for training. However, ArticuBot differs by using 3D point clouds as network observation, instead of using monocular RGB images.

YOEO [12] demonstrates that the training of a real-time perception model specialized for estimating the kinematic parameters of the joints that describes a given articulated model. In comparison with our work, it uses datasets of 3D models of articulated objects as a training data source, but differs by training a 3D perception model, instead of an action-generating policy, and by using point clouds as the network's input modality, rather than only monocular RGB images.

A3VLM [13] already investigates the training of an articulated object perception model based on a vision-language model as a base. It resembles our work by also using datasets of 3D models of articulated objects as a data source, and by using RGB images as the input modality, but differs by being a perception model, and not an action-generating policy.

SAGE [14] presents a modular generalist framework for language-conditioned manipulation of any type of articulated object. While it shares our foundation in vision-language models, it differs from us by employing a modular architecture. Its approach to action generation through the usage of smaller expert models in combination with the vision-language model in order to ground language instructions to general part affordances, which makes the usage of a classical motion planner for generating trajectories. Furthermore, SAGE relies on the depth information of RGB-D observations, where our work uses only monocular RGB images.

Kinematics-aware Prompting [15] introduces the framework that enables pretrained LLMs and VLMs to generate low-level action trajectories for manipulating articulated objects through prompting. It resembles our work by generating the low-level executable action trajectories required for manipulating objects with a robotic arm. It diverges from our work by relying on a model with textual kinematic description of the articulated object. In contrast, our approach learns to generate these trajectories from demonstration data, removing the need for manual description of the object's structure during the training.

Cart [16] studies the training of a neural-network model for applying movements to the point clouds of articulated objects, given language instructions. It resembles our work by also studying articulated object manipulation conditioned on language commands, but differs from our work by training a model to predict the part-level motion of the articulated object point cloud, instead of training a model to predict the action trajectory of the robotic arm in order to manipulate the same object.

Mittal et al [17] studies a classical planning-based system for mobile manipulation of articulated objects. It resembles our work by also studying how to open doors and drawers with a robotic manipulator, but differs from our work by employing a classical pipeline instead of an end-to-end learned model.

ControlVLA [18] leverages a ControlNet style architecture in order to inject object-centric representations into a pretrained VLA policy during finetuning. It resembles our work by studying how to inject object-centric representations into

a pretrained VLA model in order to improve generalization, but differs from our work by leveraging a more complex ControlNet inspired technique for injecting the object-centric features into the VLA, as well as by using a learned extractor for acquiring object-centric features from monocular RGB, instead of directly leveraging target object pose as an object-centric feature itself.

Contact-Anchored Policies [19] trains a modular skill library of visuomotor manipulation policies conditioned on contact points in euclidean space. It shares some similarities with our work by conditioning visuo-motor policies on 3D information relevant to each task, but differs from our work by implementing contact point conditioning upon smaller scale single-task expert policies, instead of leveraging a pretrained VLA model.

The related works presented so far provide evidence of a research gap in the state-of-the-art of articulated object manipulation research: while several methods have been proposed so far in order to tackle articulated object manipulation tasks, most of these methods are either modular pipelines, or leverage point cloud sensor readings as the primary modality, or both. The lack of end-to-end vision language action models specifically trained for articulated object manipulation tasks highlights the need for more evaluation benchmarks and imitation learning datasets developed specifically for this particular problem.

3 Methodology

A set of simulated environments, shown in Fig. 1 was developed using the IsaacLab Arena framework [20] to train relevant skills for @Home category domestic robots: opening a cabinet drawer, opening a microwave door, and picking up a subset of the objects from the YCB [21] object set. The simulation was developed using 3D assets describing the kinematic structure of a fixed base manipulator robot, the UFactory xArm6 model with 6 degrees-of-freedom. Assets of the rigid and articulated objects, as well as the background, were sourced from the existing IsaacLab Arena and IsaacSim assets.

The following simulation tasks were developed for collecting training data:

- **Pick Tomato Soup Can**: the robot must grasp a tomato soup can placed on a countertop in front of it and lift it by at least 20 cm.
- **Pick Mustard Bottle**: the robot must grasp a mustard bottle placed on a countertop in front of it and lift it by at least 20 cm.
- **Pick Sugar Box**: the robot must grasp a sugar box placed on a countertop in front of it and lift it by at least 20 cm.
- **Pick Cracker Box**: the robot must grasp a cracker box placed on a countertop in front of it and lift it by at least 20 cm.
- **Open Cabinet Drawer**: the robot must grasp the handle of a cabinet drawer placed in front of it and pull it along the cabinet joint axis in order to open the drawer by at least 20 cm.
- **Open Microwave Door**: the robot must grasp the handle of a microwave door placed in front of it and pull it along the joint axis of the door in order to open it by at least 15%.



(a) Open Cabinet Drawer (b) Pick Cracker Box (c) Open Microwave Door



(d) Pick Mustard Bottle (e) Pick Tomato Soup Can (f) Pick Sugar Box

Fig. 1: Manipulation tasks in a simulated kitchen environment.

The following fields were used as part of the observation space:

- **Top Camera**: a monocular RGB camera mounted in a fixed pose relative to the base link of the robotic arm, facing forward with a downward inclination of 45 degrees. The camera has a resolution of 256x256 pixels.
- **Wrist Camera**: a monocular RGB camera mounted on the gripper link of the robotic arm. The camera has a resolution of 256x256 pixels.
- **Proprioceptive State**: the end-effector position and end-effector rotation as quaternion vectors, concatenated with the joint state of the active gripper joint.

For the action space, the relative delta of the position and orientation of the end-effector was used as target for an inverse kinematics controller, while the gripper was controlled with a binary open/close action.

In order to collect demonstration trajectories, scripted oracle policies were developed for the mentioned tasks. For object pick-up tasks, the oracle policy uses the cuRobo [22] motion planning library to plan a grasping trajectory given a set of grasp poses previously sampled from the mesh of the object. At each motion planning attempt, the grasp pose with the smaller euclidean distance in position to the robotic arm end effector pose is selected and, in case of motion

planning failure due to collision or kinematic constraints, the next closest grasp pose is selected.

For the drawer opening task, the oracle policy plans a grasping motion using a manually annotated grasp pose on the handle of the cabinet drawer. After grasping the handle, the policy uses a screw motion to move in a linear path along the axis of the drawer's joint in order to open it.

For the microwave door opening task, the oracle policy plans a grasping motion to a manually annotated grasp pose on the handle of the microwave door and then moves in an arc around the joint axis of the door while keeping the end effector orientation constant.

After the datasets were collected, fine-tuning experiments were conducted on models available in the LeRobot [23] robotic imitation learning library. In this stage, experiments were performed with the following models:

- **SmolVLA [24]**: A state-of-the-art vision-language-action model, with action generation via flow matching and pre-training on a large-scale dataset collected by the LeRobot library user community. The batch size used for training was 64, and 20,000 gradient descent steps were used during training, with the wrist and top camera images resized to a width and height of 256 pixels. Moreover, all default training hyperparameters from the library were used, without manual tuning.
- **SmolVLA-NoPretrain [24]**: A variant of SmolVLA without pretraining on the large-scale LeRobot community dataset. We train this variant in order to isolate the impact of the LeRobot pretraining dataset on the success rate of our finetuned models.
- **PoseCondVLA**: Our proposed method for improved object pose generalization, which consists of concatenating the position and quaternion rotation vectors of the target object with the proprioceptive state vector of the end-effector of the robotic arm before projecting those into the token embedding space of the finetuned SmolVLA model. This contribution is illustrated in Fig. 2.

Evaluation was then performed by running the models 50 consecutive times in each environment, and the success rate over the 50 attempts was reported. We executed two different experiments that aim to answer the following research questions:

- **Q1: Does cross-embodiment pretraining improve the success rate in our proposed simulation environments?** To answer this question, we execute training experiments with both a base vision-language-action model pretrained with large-scale real-world data from a different embodiment than our robot (SO-100 vs xArm6) and the same model architecture without the large-scale cross-embodiment pretraining. During those experiments, we collect 50 demonstrations per environment for all six aforementioned environments, with no randomization in the pose of the objects.

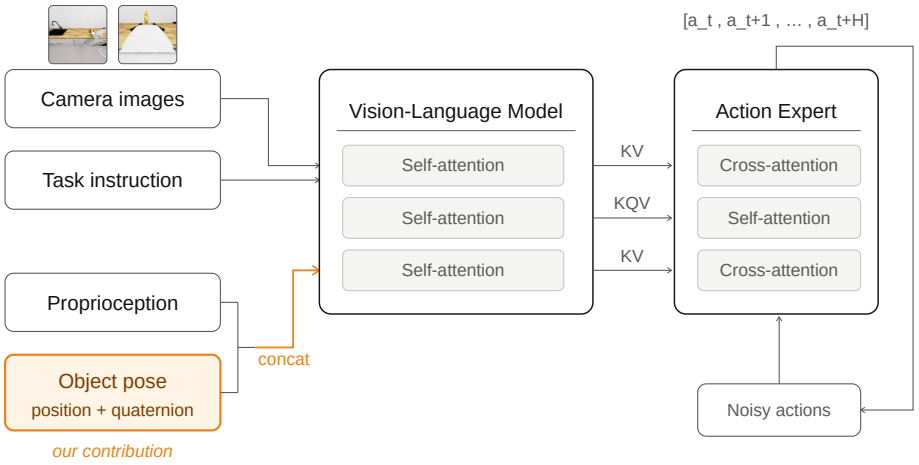


Fig. 2: SmolVLA Architecture. Our contribution (PoseCondVLA) is highlighted in orange.

– **Q2: Does PoseCondVLA improve the success rate and generalization in one of our simulated environments when the target object pose is randomized during data generation and inference?** To answer this question, we execute finetuning experiments with both a pretrained vision-language-action model baseline and our proposed PoseCondVLA method instantiated on top of the same vision-language-action model. During those experiments, we collect 1,000 demonstration trajectories for the *PickMustardBottle* task in our benchmark, while randomizing the target bottle position on the table in a 0.4m × 0.4m area around the initial pose. To evaluate zero-shot generalization to unseen object categories, we additionally evaluate the trained model on the *PickSugarBox* task without any further training.

All fine-tuning experiments start from the same pretrained SmolVLA base weights. Q1 and Q2 are independent runs and share no fine-tuning gradients. In particular, the PoseCondVLA model evaluated in Q2 has never seen the sugar box during its own fine-tuning, making PickSugarBoxPoseRand a zero-shot evaluation on an unseen object category.

4 Results

Table 1 contains the results obtained from the experiments conducted in the previously described multi-task setting with the goal of answering **Q1**. As can be observed, the model without pre-training achieves parity with the model pre-trained on the LeRobot SO-100 dataset on 3/6 tasks, but fails to achieve any success in the *PickTomatoSoupCan* and *OpenMicrowaveDoor* tasks. The

model with pretraining achieves a perfect success rate on the *PickMustardBottle*, *PickSugarBox* and *PickCrackerBox* tasks, while also achieving a considerable success rate in the *PickTomatoSoupCan* and *OpenMicrowaveDoor* tasks. Both models fail to achieve any success on the *OpenCabinetDrawer* task.

Table 1: Multitask success rate with no pose randomization

Task	**SmolVLA**	**SmolVLA-NoPretrain**
PickTomatoSoupCan	64%	0%
PickMustardBottle	100%	100%
PickSugarBox	100%	100%
PickCrackerBox	100%	100%
OpenCabinetDrawer	0%	0%
OpenMicrowaveDoor	84%	0%

Table 2 contains the results of the experiments executed in the pose generalization setting in order to answer **Q2**. As it can be observed, the SmolVLA model trained with the proposed PoseCondVLA method not only achieves a significantly higher success rate on the trained *PickMustardBottlePoseRand* task, but also generalizes with non-trivial success rate to the *PickSugarBoxPoseRand* task. On the other hand, the model trained without pose conditioning fails to achieve any successful execution on both tasks.

Table 2: Pose generalization success rate

Task	**SmolVLA-PoseCond**	**SmolVLA**
PickMustardBottlePoseRand	90%	0%
PickSugarBoxPoseRand	78%	0%

5 Discussion

The improved success rate of the SmolVLA model with large-scale manipulation data pretraining suggests that the pretraining of SmolVLA on the LeRobot community dataset might allow the model to learn some object manipulation related features that seem to be enough related to the novel embodiment that it allows for more sample-efficient learning than the same baseline model trained from scratch.

The complete failure on *OpenCabinetDrawer*, for both pretrained and non-pretrained variants, is the most informative negative result, particularly in contrast with the 84% success on *OpenMicrowaveDoor*. Plausible contributing factors include the prismatic joint's tighter tolerance to end-effector drift relative to

the microwave's revolute joint, the underrepresentation of sustained linear-pull trajectories in the pretraining dataset, and the fact that 50 demonstrations may simply be below the threshold for contact-rich tasks of this type. Prior work on imitation learning for articulated manipulation uses hundreds to thousands of trajectories per task. Further analysis on these factors would require an ablation experiment on the drawer task alone, which we leave to future work.

For the pose generalization experiments, the baseline model collapse indicates that, at this level of pose variability, the visual and proprioceptive inputs alone are not sufficient for the fine-tuning dataset to ground the action distribution to the object's spatial location.

Concatenating the object's position and quaternion into the state vector closes this gap. The transfer to the sugar box is particularly informative: because the Q2 fine-tuning run is independent of Q1 and uses only mustard-bottle demonstrations, the sugar box is an object category unseen during this network's fine-tuning. The 78% success rate under both pose randomization and category shift suggests that this simple, architecturally cheap form of state augmentation is sufficient to unlock pose generalization that the baseline fails to learn from pixels alone.

The most important caveat is that PoseCondVLA uses privileged ground-truth object pose provided by the simulator. For deployment on a physical robot, this signal must be supplied by a perception pipeline whose error will degrade the policy in a manner we have not yet characterized. Integrating an open-vocabulary 6D pose estimator into the inference loop, and measuring the resulting success rate as a function of estimator error, is the natural next experiment.

6 Conclusion and Future Work

This work presents HomeArena, a synthetic data generation tool for finetuning VLA models on articulated and rigid object manipulation tasks relevant for the RoboCup@Home competition. Finetuning experiments in simulation confirm the validity of the generated training data for object pickup, door and drawer opening tasks, but results remain mostly limited to simulation. Further work will explore sim-to-real transfer of the learned VLA policies into a physical xArm6 robot arm.

The proposed PoseCondVLA method significantly improves generalization to novel object poses, but we acknowledge effective sim-to-real transfer will require a separate perception pipeline for the estimation of object poses relative to the base frame of the robotic arm. Our future work will integrate pretrained foundation models for open-vocabulary perception into the inference pipeline of the VLA model, aiming for effective sim-to-real transfer of the finetuned policies.

Acknowledgments. This work was partly supported by FAURG and CNPQ.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. M. Matamoros, V. Seib, R. Memmesheimer, and D. Paulus, "Robocup@ home: Summarizing achievements in over eleven years of competition," in *2018 IEEE International Conference on Autonomous Robot Systems and Competitions (ICARSC)*. IEEE, 2018, pp. 186–191.
2. J.-B. Yi, T. Kang, D. Song, and S.-J. Yi, "Unified software platform for intelligent home service robots," *Applied Sciences*, vol. 10, no. 17, p. 5874, 2020.
3. A. Mandlekar, S. Nasiriany, B. Wen, I. Akinola, Y. Narang, L. Fan, Y. Zhu, and D. Fox, "MimicGen: A Data Generation System for Scalable Robot Learning using Human Demonstrations," 2023. [Online]. Available: https://arxiv.org/abs/2310.17596
4. G. A. Dorneles, J. F. S. S. Lemos, K. S. Kappel, R. S. Guerra, and P. L. J. Drews, "Affordgen: Affordance-based dataset generator for robot manipulation learning," in *2025 IEEE International Conference on Advanced Robotics (ICAR)*, 2025.
5. G. A. Dorneles, K. S. Kappel, S. L. Briao, J. F. de Souza Santos Lemos, R. da Silva Guerra, and P. L. J. D. Junior, "A review on dataset collection strategies for learning methods in robotic manipulation," in *RoboCup 2025: Robot World Cup XXVIII*, ser. Lecture Notes in Computer Science, vol. 16460. Springer, Cham, 2026, to appear, due May 2026.
6. K. Kawaharazuka, J. Oh, J. Yamada, I. Posner, and Y. Zhu, "Vision-language-action models for robotics: A review towards real-world applications," *IEEE Access*, 2025.
7. A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, J. Dabis, C. Finn, K. Gopalakrishnan, K. Hausman, A. Herzog, J. Hsu *et al.*, "Rt-1: Robotics transformer for real-world control at scale," in *Proceedings of Robotics: Science and Systems (RSS)*, Daegu, Republic of Korea, Jul. 2023.
8. K. Black, N. Brown, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, L. Groom, K. Hausman, B. Ichter, S. Jakubczak, T. Jones, L. Ke, S. Levine, A. Li-Bell, M. Mothukuri, S. Nair, K. Pertsch, L. X. Shi, J. Tanner, Q. Vuong, A. Walling, H. Wang, and U. Zhilinsky, "π_0: A vision-language-action flow model for general robot control," 2024. [Online]. Available: https://arxiv.org/abs/2410.24164
9. F. Xiang, Y. Qin, K. Mo, Y. Xia, H. Zhu, F. Liu, M. Liu, H. Jiang, Y. Yuan, H. Wang, L. Yi, A. X. Chang, L. J. Guibas, and H. Su, "SAPIEN: A SimulAted Part-based Interactive ENvironment," 2020. [Online]. Available: https://arxiv.org/abs/2003.08515
10. Z. Jin, Z. Che, Z. Zhao, K. Wu, Y. Zhang, Y. Zhao, Z. Liu, Q. Zhang, X. Ju, J. Tian, Y. Xue, and J. Tang, "ArtVIP: Articulated Digital Assets of Visual Realism, Modular Interaction, and Physical Fidelity for Robot Learning," 2025. [Online]. Available: https://arxiv.org/abs/2506.04941
11. Y. Wang, Z. Wang, M. Nakura, P. Bhowal, C.-L. Kuo, Y.-T. Chen, Z. Erickson, and D. Held, "ArticuBot: Learning Universal Articulated Object Manipulation Policy via Large Scale Simulation," 2025. [Online]. Available: https://arxiv.org/abs/2503.03045
12. J. Huang, H. Lin, T. Wang, Y. Fu, Y.-G. Jiang, and X. Xue, "You Only Estimate Once: Unified, One-stage, Real-Time Category-level Articulated Object 6D Pose Estimation for Robotic Grasping," 2025. [Online]. Available: https://arxiv.org/abs/2506.05719
13. S. Huang, H. Chang, Y. Liu, Y. Zhu, H. Dong, P. Gao, A. Boularias, and H. Li, "A3VLM: Actionable Articulation-Aware Vision Language Model," 2024. [Online]. Available: https://arxiv.org/abs/2406.07549

14. H. Geng, S. Wei, C. Deng, B. Shen, H. Wang, and L. Guibas, "SAGE: Bridging Semantic and Actionable Parts for GEneralizable Articulated-Object Manipulation under Language Instructions," in *Robotics: Science and Systems XX*. Robotics: Science and Systems Foundation, Jul. 2024.
15. W. Xia, D. Wang, X. Pang, Z. Wang, B. Zhao, D. Hu, and X. Li, "Kinematic-aware Prompting for Generalizable Articulated Object Manipulation with LLMs," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. Yokohama, Japan: IEEE, May 2024, pp. 2073–2080.
16. R. Chu, Z. Liu, X. Ye, X. Tan, X. Qi, C.-W. Fu, and J. Jia, "Command-driven Articulated Object Understanding and Manipulation," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Vancouver, BC, Canada: IEEE, Jun. 2023, pp. 8813–8823.
17. M. Mittal, D. Hoeller, F. Farshidian, M. Hutter, and A. Garg, "Articulated Object Interaction in Unknown Scenes with Whole-Body Mobile Manipulation," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. Kyoto, Japan: IEEE, Oct. 2022, pp. 1647–1654.
18. P. Li, Y. Wu, Z. Xi, W. Li, Y. Huang, Z. Zhang, Y. Chen, J. Wang, S.-C. Zhu, T. Liu, and S. Huang, "Controlvla: Few-shot object-centric adaptation for pre-trained vision-language-action models," *ArXiv*, vol. abs/2506.16211, 2025. [Online]. Available: https://api.semanticscholar.org/CorpusID:279465293
19. Z. J. Cui, O. Rayyan, H. Etukuru, B. Tan, Z. Andrianarivo, Z. Teng, Y. Zhou, K. Mehta, N. Wojno, K. C.-K. Wu, M. Anjaria, Z. Wu, M. Mao, G. Zhang, B. Shah, Y. Kim, S. Chintala, L. Pinto, and N. M. M. Shafiullah, "Contact-anchored policies: Contact conditioning creates strong robot utility models," *ArXiv*, vol. abs/2602.09017, 2026. [Online]. Available: https://api.semanticscholar.org/CorpusID:285452819
20. NVIDIA, M. Mittal, P. Roth, J. Tigue, A. Richard, O. Zhang, P. Du *et al.*, "Isaac Lab: A GPU-accelerated simulation framework for multi-modal robot learning," 2025. [Online]. Available: https://arxiv.org/abs/2511.04831
21. B. Calli, A. Walsman, A. Singh, S. Srinivasa, P. Abbeel, and A. M. Dollar, "Benchmarking in manipulation research: Using the yale-cmu-berkeley object and model set," *IEEE Robotics Automation Magazine*, vol. 22, no. 3, pp. 36–52, 2015.
22. B. Sundaralingam, S. K. S. Hari, A. Fishman, C. R. Garrett, K. V. Wyk, V. Blukis, A. Millane, H. Oleynikova, A. Handa, F. T. Ramos, N. D. Ratliff, and D. Fox, "Curobo: Parallelized collision-free robot motion generation," *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 8112–8119, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:259337127
23. "huggingface/lerobot," Nov. 2025, original-date: 2024-01-26T15:50:41Z. [Online]. Available: https://github.com/huggingface/lerobot
24. M. Shukor, D. Aubakirova, F. Capuano, P. Kooijmans, S. Palma, A. Zouitine, M. Aractingi, C. Pascal, M. Russi, A. Marafioti, S. Alibert, M. Cord, T. Wolf, and R. Cadene, "SmolVLA: A Vision-Language-Action Model for Affordable and Efficient Robotics," Jun. 2025, arXiv:2506.01844. [Online]. Available: http://arxiv.org/abs/2506.01844