Training SO-100 robot to arrange a table

Task:

We train a policy for SO100 robots for the task of arranging objects on a desk. Specifically, we tackled the task of collecting cluttered pens from a desk-top environment and depositing them precisely into a designated bin. This seemingly straightforward task presents significant challenges for autonomous systems, requiring robust perception, sophisticated motion planning, and precise grasping capabilities. The arm needs to approach the pens one by one, grab them in correct form and once grabbed, move towards a small bin and put the pin inside the bin. The task is accomplished when there is no pen remaining on the desk and all are deposited in the bin.

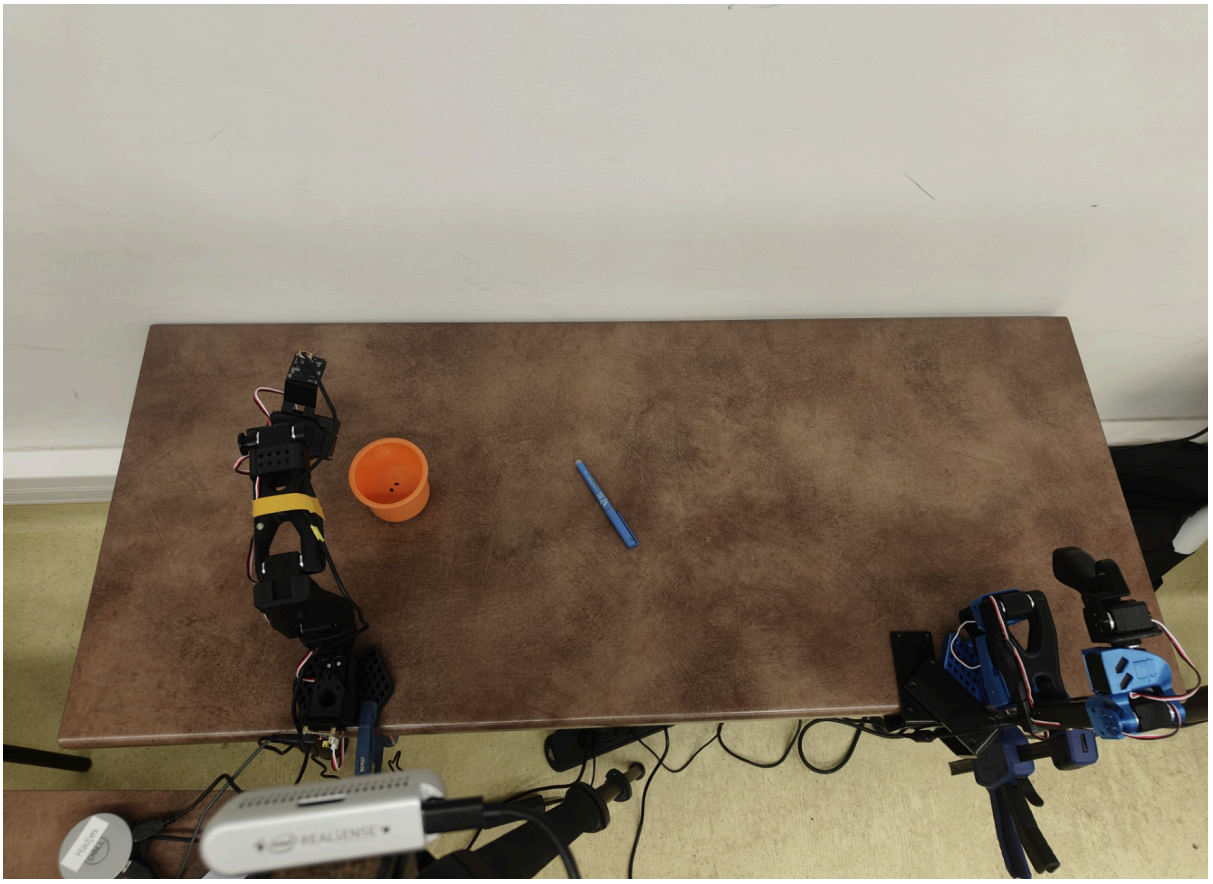
Setup and configuration:

The SO-100 robot (Standard Open Arm 100) is a robotic arm with 5 “body” joints and a “gripper” joint making it a 6 DoF arm. The setup includes the leader and follower robots. The leader-arm was assembled using 3D printed pieces.
A significant undertaking of this project involved the ground-up assembly of the leader arm. This engineering process, conducted entirely within the hackathon timeframe, ensured complete control over its mechanical and electrical integration. The leader arm is not merely a control device; it's an intuitive haptic interface. By physically manipulating the leader arm, an operator can directly teleoperate the main follower arm, translating their movements with high fidelity and minimal latency. The follower arm is equipped with a camera attached to the grabber, and there is also another Intel Realsense camera giving a top-view of the scene.

Environment:

The environment includes a table, the two arms with fixed positions, a small bin, and pens. The tools in the scene include a small yellow bin and multiple pens of different shapes, colors and sizes. Within this environment, the two SO100 robotic arms were securely fixed in predetermined positions. This fixed configuration ensured consistent kinematic relationships and repeatable experimental conditions, crucial for both teleoperation training and policy deployment. The placement of the leader arm was optimized for ergonomic operator control, while the follower arm was positioned to maximize its reach over the designated task area on the table.

Key tools and objects within the scene included a small yellow bin, serving as the target receptacle for the pen collection task. We incorporated multiple pens varying in shape, color, and size. A top-view of the setup is provided below:



Data collection:

We collected data using teleoperation via a leader arm to perform a grasping task. Each episode involves placing 1, 2, or 3 pens into a pen holder. On the first day, we recorded 91 episodes under the dataset name: Alexisbo/training-dataset-pen-to-cup

On the following day, we added 84 more episodes. These were merged with the initial dataset to create a consolidated dataset named: Alexisbo/full_dataset_grasping

Each episode takes approximately 1 minute to collect, and is recorded at 30 fps rate.

You can visualize the dataset at the following link:
https://lerobot-visualize-dataset.hf.space/Alexisbo/full_dataset_grasping/episode_0

Training:

We trained 2 different policies offline on our collected dataset using the LeRobot framework:

- Action Chunking Transformer (ACT) [Zhao T. Z. et al., 2023, *Learning Fine-Grained Bimanual Manipulation with Low-Cost Hardware*], an imitation learning policy which learns a generative model for action sequences. We trained it for 100 000 steps

(batch size = 8) on a 90 episodes dataset, and then we continued the training for 10000 steps (batch size of 64) on an extended dataset with 90 extra episodes.

- SmolVLA, [Shukor M. et al., 2025, *SmolVLA: A Vision-Language-Action Model for Affordable and Efficient Robotics*], a Vision Language Action model trained by HuggingFace using community datasets of the SO-100 robot.

Results:

https://www.youtube.com/shorts/DeBhIQQYkkI (a success amongst failures)

Model	**GPU used**	**Total Steps**	**Time Taken/500 steps**	**Total time**
SmolVLA	RTXA6000	20 000	60mins	
ACT	RTX A6000	100 000	10mins	200 mins

SmolVLA was fine-tuned for 1000 steps and did not perform well: the robot arm was unstable and shaking. Another test was performed by fine tuning SmolVLA for another 1000 steps and it seemed to perform better than the previous test but it still did not perform satisfactorily on the task. The shaking went down compared to the previous test but it still failed to grasp the pens in most scenarios. A third and final test was performed with SmolVLA fine-tuned for 2500 steps in total and it showed slightly better performance than the one tuned for 2000 steps. This behaviour can be accounted for by the recommended training steps for smolVLA, which are 20,000 steps but due to time and GPU constraints, we could not train it for the recommended amount. Consequently, the performance of the robot suffers but it does show potential to improve by trying to head towards the object of interest (the pen). We believe that if the SmolVLA is trained for 20,000 steps, it will be able to demonstrate decent performance.

On the other hand, ACT performs considerably well when compared to testing with SmolVLA. It achieved a success rate of about 10%.