Evaluating pi0 in the Wild: Strengths, Problems, and the Future of Generalist Robot Policies

Jie Wang*, Matthew Leonard, Kostas Daniilidis, Dinesh Jayaraman, Edward S. Hu
GRASP Lab, University of Pennsylvania

Abstract

https://penn-pal-lab.github.io/Pi0-Experiment-in-the-Wild/

Historically, deploying robotic manipulation models required tedious engineering and careful calibration, yet such systems are often **prone to fail with minor environment changes.** Recent vision-language-action (VLA) models promise **out-of-the-box generalization across diverse tasks and settings**. In this work, we present an observational study of the **pi0-fast-droid** model in a mock kitchen environment with a Franka Panda robot. Inspired by "vibe-checking" style evaluations in NLP, we conducted over **300 trials** across tasks such as pick-and-place, pouring, articulated-object manipulation, fabric folding, human–robot interaction, and household appliances (e.g., coffee machine operation). Our observations reveal that pi0 demonstrates *strong priors for sensible behaviors, robustness to transparent and camouflaged objects, and emergent sequential action behaviors without explicit memory.* However, performance remains highly sensitive to prompt design and camera viewpoints, with common failure modes including *early stopping, weak spatial reasoning, and poor out-of-distribution generalization to unseen appliances and novel backgrounds*. Despite these limitations, achieving even partial success **without manual calibration** represents an important step toward deployable generalist robot policies. To further advance evaluation diversity and rigor, we extend this effort through **RoboArena**, a distributed real-world benchmarking framework.

Keyword:- ***Generalist robot policies; vision-language-action models; robotic manipulation; evaluation;***

Scan QR to read full blog

Introduction

If you had access to a robot foundation model, how would you play with it?

In GRASP Lab, we were fortunate to be early testers of Physical Intelligence's pi0. We found it hard to rigorously benchmark all feasible tasks for a generalist policy.

Instead of trying every possible task, we took a leaf from the LLM playbook and adopted the *"vibe-checking"* approach, where we ask pi0 to perform whatever relevant tasks we were interested in, rather than relying on a standard benchmark.

Across 300+ trials, we discovered pi0's strengths, problems, and even some interesting quirks.

figure 1:We run evaluation on Franka Panda Arm, with DROID system in a mock kitchen environment.

Strength & Problems

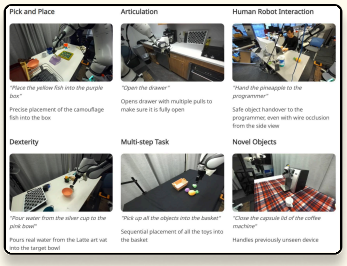


figure 2: Success cases and example tasks

What we are amazed by pi0:

It can do sensible behaviors across a wide variety of tasks, zero-shot. As shown in the left, it can precisely pick-and-place a camouflage fish into a specific box. Which is hard to distinguish from just RGB cameras.

Problems with pi0

Pi0 can recover from failures, and handle moving humans in the scene, but it struggles with mid-task freezing, collision avoidance, and fine-grained manipulation. For example, pi0 often early stops in the air during middle of a task.

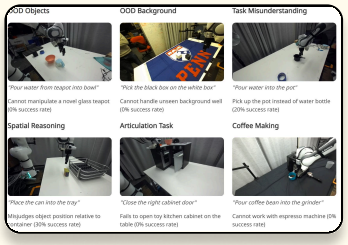


figure 3: Failure cases and common reasons

Word clouds & Results

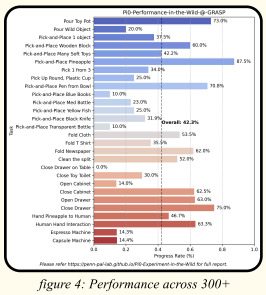


figure 4: Performance across 300+ trials, average progress 42.3%.

Pi0 achieves 42.3% average progress across diverse, hard tasks, even in unseen environments and objects. However, it also fails in some seemingly simple tasks. During evaluation, we assign a continuous score from 0 to 100 to each rollout. Notice it's not binary accuracy, please refer Appendix B.

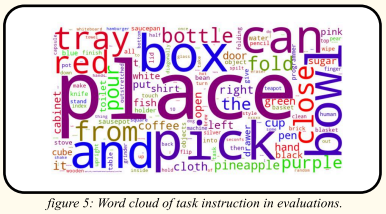


figure 5: Word cloud of task instruction in evaluations.

Conclusion

Our evaluations show that is a promising generalist policy: For the first time, we can download a policy and get 20–50 % success out of the box. However, to deliver a robot working at everyone's home, pi0 needs more foundational improvements: better perception, memory and safety. But it's already a large positive gradient toward generalist robots, and we're optimistic on that future.

However, testing pi0 only at a research lab is far not enough. To make general purpose robots working in everyone's home, we need more people's joint effort. The spirit is similar to LLMs: by **making capabilities visible and engaging the broader community,** we can iterate faster and attract new contributors. Let's try pi0 and share your vibe-checks!

Pi0's Behavior & Quirks

figure 6: instruction-'fold the newspaper', pi0 folds up the Daily Penn newspaper

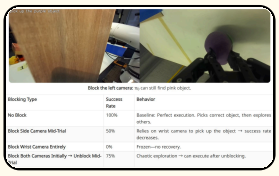
Generalization Behavior:

With large-scale pre-training foundation Vision-Language-Models(VLMs) as backbone, Physical Intelligence trained a Vision-Language-Action Models(VLAs). pi0 can generalize on new object according to the language command. This is very interesting because we usually don't expect a policy works zero-shot in robotic manipulation.

Emergent sequential behavior:

Without any memory or hard-coded logic, the policy imitates multi-step demonstrations across tasks. Very interesting, pi0 can know when the task is finished and reset, which is sometimes better and safer than hard-code reset! (it collides less with env!)

figure 7(right): example of sequential execution



PI0 is secretly an FPV player:

Here is a quirk of pi0 --- It mainly relies on wrist camera. In experiments, it still works without third camera view. Our assumption is its neural network may have more 'attention' on what the gripper see to execute tasks.

figure 8 (left): Camera Blocking Experiments
figure 9 (right): Object Blocking Experiments

Prompt Engineering Matters:

We observed pi0's success rate for the same task can change greatly based on different instructions. To make pi0 do a task, you need to try multiple prompts, and be very straightforward like: *Do something from A To B.*

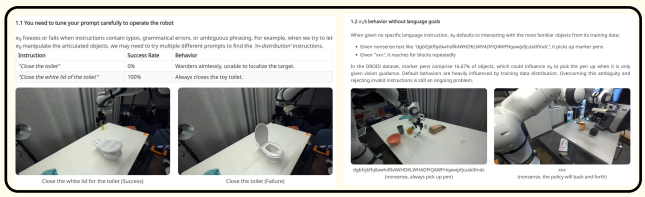


figure 10(left) : find best prompt to execute task

figure 11(right): without language goal, pi0 can still execute

Experiments Setup

Hardware:
- Franka Research 3 Arm
- Robotiq 2F-85 gripper
- Cameras:
  - Side-view: ZED 2 stereo camera x2
  - Wrist-mounted: ZED Mini

GPU Server for model inference:
- NVIDIA RTX A6000 (48GB VRAM)

Workstation for robot control:
- NVIDIA GeForce RTX 3080 (16GB VRAM)

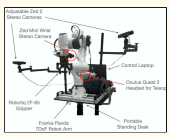


figure 12: DROID setup

Model: pi0-FAST-DROID
- *Vision-Language Model*: PaliGemma 3B
- *FAST+*: Frequency-space Action Sequence Tokenization (FAST)
- *Training Data:* Pretrained on π cross-embodiment robot dataset & Open X-Embodiment, fine tuned on DROID dataset.

Model is open-source available at:
https://github.com/Physical-Intelligence/openpi/

Extension: RoboArena, our CoRL2025 Oral Paper

Collaborating researchers from UC Berkeley, Stanford etc, we introduce **RoboArena,** a distributed, real-world benchmark to evaluate generalist robot policies like pi0. Similar to the *'vibe-check'* spirit above, we don't standardize the set of tasks to evaluate. Instead, we encourage evaluators to pick any environment and try any tasks. By aggregating a large number of **double-blind, pairwise comparisons** across a network of institutions, we considerably expand the diversity of evaluations. This work has been accepted as an **Oral Presentation at CoRL 2025**. Scan the QR code to learn more!

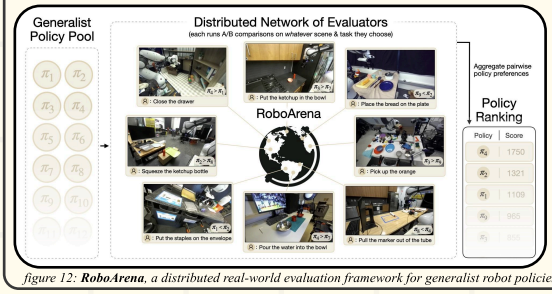


*figure 12: **RoboArena**, a distributed real-world evaluation framework for generalist robot policies.*

RoboArena