π_0: Our First Generalist Policy

Published October 31, 2024
Email research@physicalintelligence.company
Paper π_0.pdf

We are living through an AI revolution: the past decade witnessed practically useful AI assistants, AI systems that can generate photorealistic images and videos, and even models that can predict the structure of proteins. But in spite of all these advances, human intelligence dramatically outpaces AI when it comes to the physical world. To paraphrase Moravec's paradox, winning a game of chess or discovering a new drug represent "easy" problems for AI to solve, but folding a shirt or cleaning up a table requires solving some of the most difficult engineering problems ever conceived. To build AI systems that have the kind of physically situated versatility that people possess, we need a new approach — we need to make AI systems embodied so that they can acquire physical intelligence.

Over the past eight months, we've developed a general-purpose robot foundation model that we call π_0 (pi-zero). We believe this is a first step toward our long-term goal of developing artificial physical intelligence, so that users can simply ask robots to perform any task they want, just like they can ask large language models (LLMs) and chatbot assistants. Like LLMs, our model is trained on broad and diverse data and can follow various text instructions. Unlike LLMs, it spans images, text, and actions and acquires physical intelligence by training on embodied experience from robots, learning to directly output low-level motor commands via a novel architecture. It can control a variety of different robots, and can either be prompted to carry out the desired task, or fine-tuned to specialize it to challenging application scenarios. An extended article on our work can be found here.

The promise of generalist robot policies

Today's robots are narrow specialists. Industrial robots are programmed for repetitive motions in choreographed settings, repeatedly making the same weld in the same spot on an assembly line or dropping the same item into the same box. Even such simple behaviors require extensive manual engineering, and more complex behaviors in messy real-world environments such as homes are simply infeasible. AI could change that, allowing robots to *learn* and follow user instructions, so that programming a new behavior is as simple as telling the robot what you want done, and the robot can itself figure out how to adapt its behavior to its environment. But this requires data. Language models and other foundation models mine data from the web, utilizing a significant fraction of all available documents. There is no such treasure trove of robot data, so to enable a robot to learn a new skill, large amounts of data need to be collected *with that particular robot and for that particular application*.

If we could train a single *generalist* robot policy that can perform a wide range of different skills and control a wide range of different robots, we would overcome this challenge: such a model would need only a little bit of data from each robot and each application. Just as a person can learn a new skill quickly by drawing on a lifetime's worth of experience, such a generalist robot policy could be specialized to new tasks with only modest amounts of data. This would not be the first time that a generalist model beat a specialist at the specialist's own task: language models have superseded more specialized language processing systems precisely because they can better solve those downstream specialist tasks by drawing on their diverse and general purpose pretraining. In the same way that LLMs provide a *foundation model* for language, these generalist robot policies will provide a robot foundation model for physical intelligence.

To get there, we will need to solve major technical challenges. Our first step is π_0, a prototype model that combines large-scale multi-task and multi-robot data collection with a new network architecture to enable the most capable and dexterous generalist robot policy to date. While we believe this is only a small early step toward developing truly general-purpose robot models, we think it represents an exciting step that provides a glimpse of what is to come.

A cross-embodiment training mixture

π0 uses Internet-scale vision-language pre-pretraining, open-source robot manipulation datasets, and our own datasets consisting of dexterous tasks from 8 distinct robots as inputs.

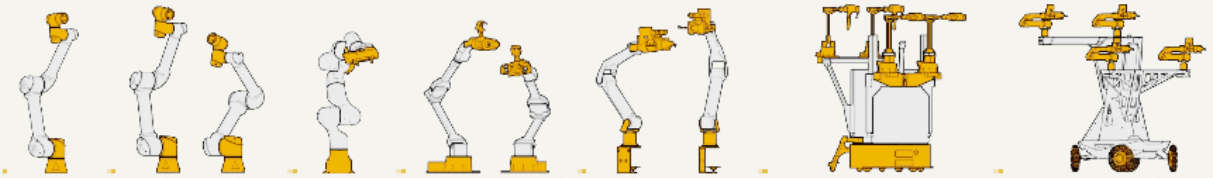
INPUTS

1 Open X Embodiment Dataset.

2 Internet-Scale pre-training.

3 π Dataset: Multiple dexterous robots.

UR5e | Bimanual UR5e | Franka | Bimanual Trossen | Bimanual Arx | Mobile Trossen | Mobile Fibocom

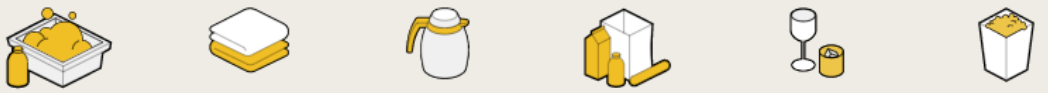


4 π_0

The model can then perform a wide variety of tasks, via either direct prompting or fi[illegible]g.

OUTPUT

Laundry | Folding clothes | Make coffee | Bag groceries | Bus table | Open popcorn



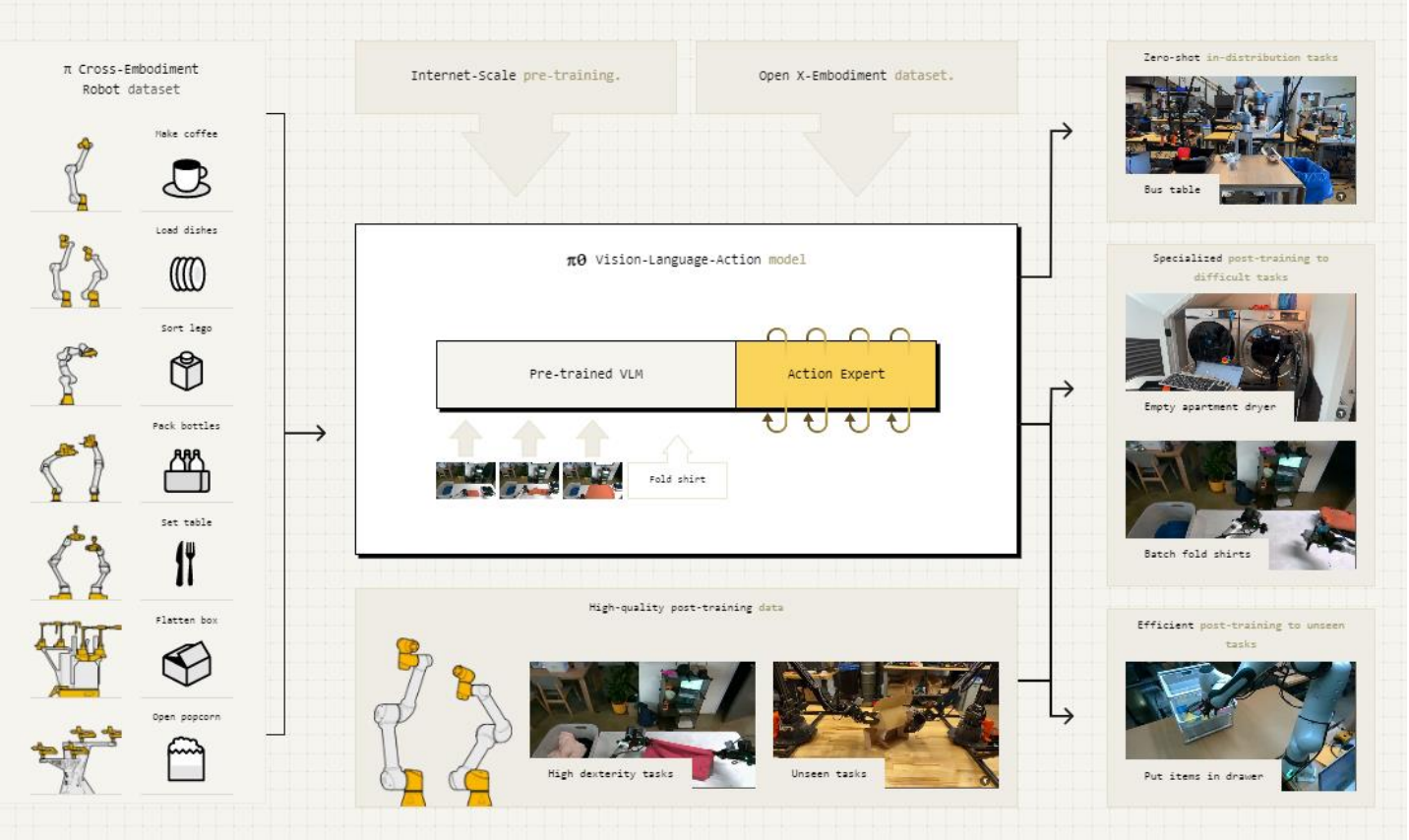
And many more

The tasks in this dataset exercise different dimensions of robot dexterity while covering the range of real tasks that these robots might be asked to perform, from bussing dishes to packing items into envelopes, folding clothing, routing cables, assembling boxes, plugging in power plugs, packing food into to-go boxes, and picking up and throwing out trash. Our goal in selecting these tasks is not to solve any particular application, but to start to provide our model with a general understanding of physical interactions — an initial foundation for physical intelligence.

Inheriting Internet-scale semantic understanding

Beyond training on many different robots, π_0 inherits semantic knowledge and visual understanding from Internet-scale pretraining by starting from a pre-trained vision-language model (VLM). VLMs are trained to model text and images on the web — widely used VLMs include GPT-4V and Gemini. We use a smaller 3 billion parameter VLM as a starting point, and adapt it for real-time dexterous robot control.

VLMs effectively transfer semantic knowledge from the web, but they are trained to output only discrete language tokens. Dexterous robot manipulation requires π_0 to output motor commands at a high frequency, up to 50 times per second. To provide this level of dexterity, we developed a novel method to augment pre-trained VLMs with continuous action outputs via flow matching, a variant of diffusion models. Starting from diverse robot data and a VLM pre-trained on Internet-scale data, we train our vision-language-action flow matching model, which we can then post-train on high-quality robot data to solve a range of downstream tasks.



Post-training for dexterous manipulation

More complex and dexterous tasks may require the model to be fine-tuned to specialize it to downstream challenges. Fine-tuning the model with high-quality data for a challenging task, such as folding laundry, is analogous to the post-training process employed by LLM designers. Pre-training teaches the model about the physical world, while fine-tuning forces it to perform a particular task *well*. Let's take a look at some of these tasks.

After post-training, the robot can unload the dryer, bring the clothes over to the table, and fold the clothes into a stack. The video is uncut, from a single policy operating fully autonomously.

Laundry. We fine-tuned π_0 to fold laundry, using either a mobile robot or a fixed pair of arms. The goal is to get the clothing into a neat stack. This task is exceptionally difficult for robots (...and some humans): while a single t-shirt laid flat on the table can sometimes be folded just by repeating a pre-scripted set of motions, a pile of tangled laundry can be crumpled in many different ways, so it is not enough to simply move the arms through the same motion. To our knowledge, no prior robot system has been demonstrated to perform this task at this level of complexity.

Notably, by training on diverse data, we find that the robot is able to recover when someone tries to intervene in a variety of different ways.

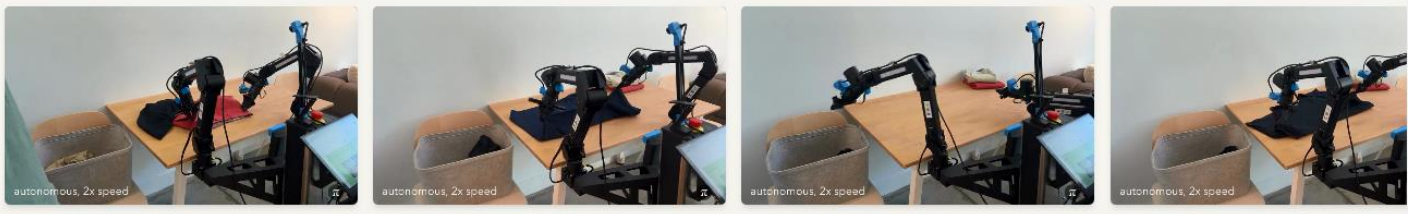
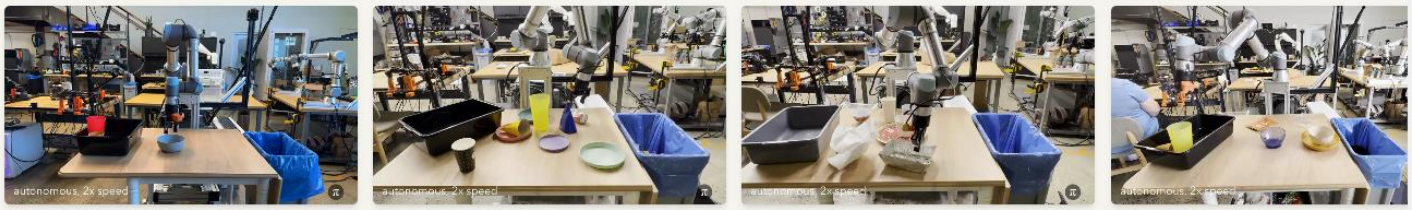
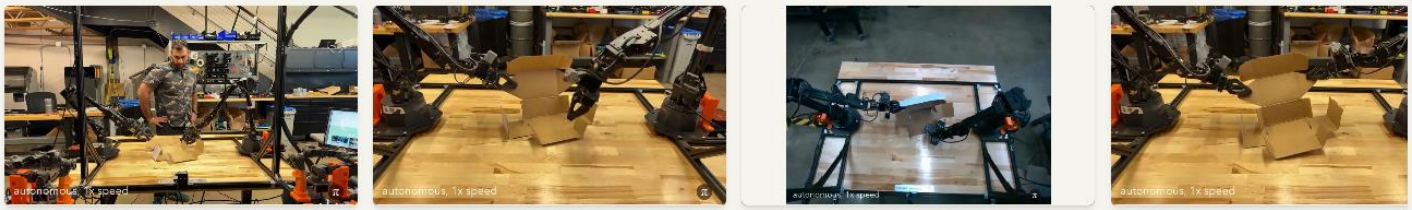


Table bussing. We also fine-tuned the model to bus a table. This requires the robot to pick up dishes and trash on the table, putting any dishes, cutlery, or cups into a bussing bin, and putting trash into the trash bin. This task requires the robot to handle a dizzying variety of items. One of the exciting consequences of training π_0 on large and diverse datasets was the range of emergent strategies that the robot employed: instead of simply grasping each item in turn, the model could stack multiple dishes to put them into the bin together, or shake off trash from a plate into the garbage before placing the plate into the bussing bin.

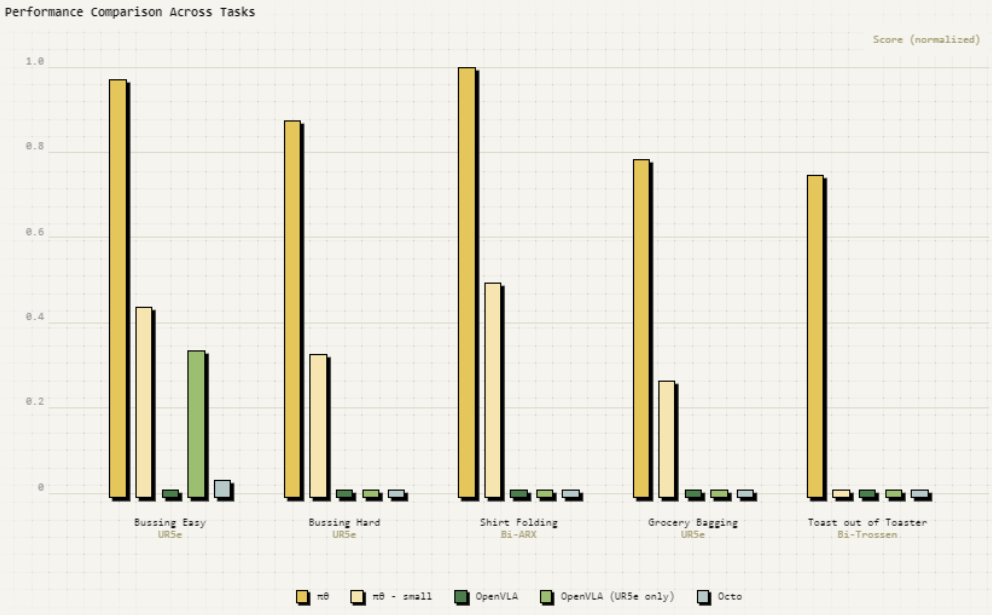


Assembling a box. Here, the robot has to take a flattened cardboard box and build it, folding the sides and then tucking in the flaps. This is very difficult, because each fold and tuck might fail in unexpected ways, so the robot needs to watch its progress and adjust as it goes. It also needs to brace the box with both arms, even using the table, so that the partially folded box doesn't come apart.



Evaluating and comparing π_0 to prior models

We compared π_0 to other robot foundation models that have been proposed in the academic literature on our tasks: OpenVLA, a 7B parameter VLA model that uses discretized actions, and Octo, a 93M parameter model that uses diffusion outputs. These tasks are very difficult compared to those that are typically used in academic experiments — for example, the tasks in the OpenVLA evaluation typically consist of single stage behaviors (e.g., "put eggplant into pot"), whereas our simplest bussing task consisting of sorting multiple objects into either a garbage bin or a bussing bin, and our more complex tasks might require multiple stages, manipulation of deformable objects, and the ability to deploy one of many possible strategies given the current configuration of the environment. These tasks are evaluated according to a scoring rubric that assigns a score of 1.0 for a fully successful completion, with "partial credit" for partially correct execution (e.g., bussing half the objects leads to a score of 0.5). The average scores across 5 evaluation tasks are shown below, comparing the full π_0 pre-trained model, π_0-small, which is a 470M parameter model that does not use VLM pre-training, OpenVLA, and Octo. Although OpenVLA and Octo can attain non-zero performance on the easiest of these tasks ("Bussing Easy"), π_0 is by far the best-performing model across all of the tasks. The small version, π_0-small, attains the second best performance, but there is more than a 2x improvement in performance from using our full-size architecture with VLM pre-training.



Average scores for π0, π0-small, OpenVLA, and Octo for evaluation on 5 test tasks. Across all of the tasks, π0 consistently attains good performance, and outperforms both the small variant and the prior models.

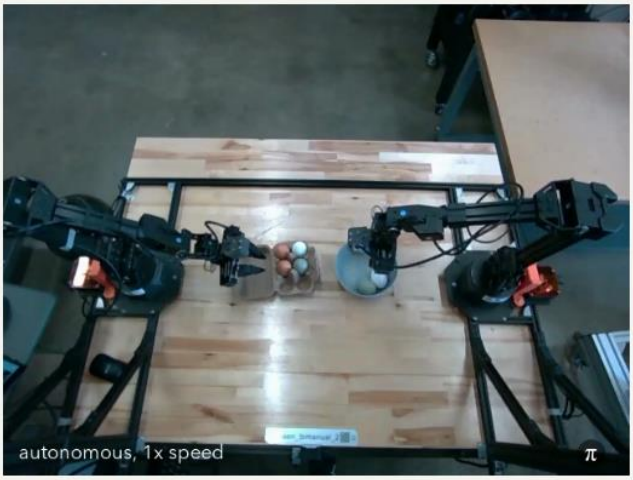
We include detailed videos from our rigorous empirical evaluation below, with examples of successful and failed episodes for both our direct prompting (out-of-box) experiments and the fine-tuning evaluation. Complete results from all experiments can be found in the full article.

Task

Eggs in carton

Status

Success | Fail



Where do we go from here?

Our mission at Physical Intelligence is to develop foundation models that can control any robot to perform any task. Our experiments so far show that such models can control a variety of robots and perform tasks that no prior robot learning system has done successfully, such as folding laundry from a hamper or assembling a cardboard box. But generalist robot policies are still in their infancy, and we have a long way to go. The frontiers of robot foundation model research include long-horizon reasoning and planning, autonomous self-improvement, robustness, and safety. We expect that the coming year will see major advances along all of these directions, but the initial results paint a promising picture for the future of robot foundation models: highly capable generalist policies that inherit semantic understanding from Internet-scale pretraining, incorporate data from many different tasks and robot platforms, and enable unprecedented dexterity and physical capability.

We also think that succeeding at this will require not only new technologies and more data, but a collective effort involving the entire robotics community. We already have collaborations underway with a number of companies and robotics labs, both to refine hardware designs for teleoperation and autonomy, and incorporate data from our partners into our pre-trained models so that we can provide access to models adapted to their specific platforms.

If you are interested in collaborating, please reach out. We are particularly excited to work with companies scaling up data collection with robots deployed for real-world applications, who are looking to collaborate on autonomy.

We are also hiring! If you'd be interested in joining us please get in touch.

For researchers interested in our work, collaborations, or other queries, please write to research@physicalintelligence.company.
