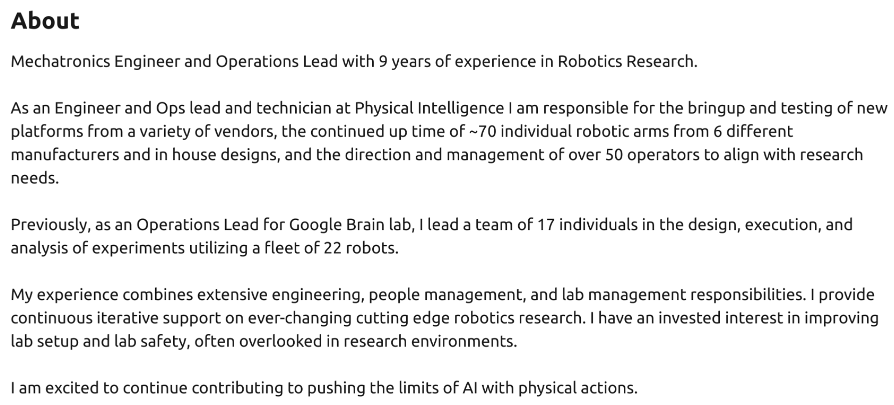
[PI'24] π_0: A Vision-Language-Action Flow Model for General Robot Control

1. Link: https://www.physicalintelligence.company/blog/pi0
2. Arthurs and institution: Physical Intelligence

    The authors contributed to the following areas (listed alphabetically):
    Data and operations: Noah Brown, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Liyiming Ke, Suraj Nair, Lucy Shi, and Anna Walling.
    Evaluation experiments: Kevin Black, Michael Equi, Chelsea Finn, Brian Ichter, Liyiming Ke, Adrian Li-Bell, Suraj Nair, Karl Pertsch, and Lucy Shi.
    Model design: Kevin Black, Brian Ichter, Sergey Levine, Karl Pertsch, Lucy Shi, and Quan Vuong.
    Post-training: Michael Equi, Chelsea Finn, Liyiming Ke, Adrian Li-Bell, Suraj Nair, and Lucy Shi.
    Pre-training: Kevin Black, Danny Driess, Brian Ichter, Sergey Levine, Karl Pertsch, Lucy Shi, and Quan Vuong.
    Robot hardware: Noah Brown, Adnan Esmail, Chelsea Finn, Tim Jones, and Mohith Mothukuri.
    Robot software: Karol Hausman, Szymon Jakubczak, Sergey Levine, James Tanner, and Haohuan Wang.
    Training infrastructure: Kevin Black, Michael Equi, Sergey Levine, Adrian Li-Bell, Suraj Nair, Quan Vuong, Haohuan Wang, and Ury Zhilinsky.
    Writing and illustration: Kevin Black, Chelsea Finn, Lachy Groom, Karol Hausman, Brian Ichter, Sergey Levine, and Quan Vuong.

    { width=50% }

    1. Kevin Black, Karl Pertsch, Suraj Nair, Lucy Xiaoyang Shi: Sergey and Chelsea's student
    2. Noah Brown: testing engineer--> robot data collection--> google brain operation


About

Mechatronics Engineer and Operations Lead with 9 years of experience in Robotics Research.

As an Engineer and Ops lead and technician at Physical Intelligence I am responsible for the bringup and testing of new platforms from a variety of vendors, the continued up time of ~70 individual robotic arms from 6 different manufacturers and in house designs, and the direction and management of over 50 operators to align with research needs.

Previously, as an Operations Lead for Google Brain lab, I lead a team of 17 individuals in the design, execution, and analysis of experiments utilizing a fleet of 22 robots.

My experience combines extensive engineering, people management, and lab management responsibilities. I provide continuous iterative support on ever-changing cutting edge robotics research. I have an invested interest in improving lab setup and lab safety, often overlooked in research environments.

I am excited to continue contributing to pushing the limits of AI with physical actions.

{ width=70% }

3. Danny Driess: Marco Toussaint's stutdent
4. Adnan Esmail: SVP hardware Engineering from tesla
5. Lachy Groom: CEO of PI, Stripe --> angel investor
6. Karol Hausman: CEO of PI, Standford+deepmind
7. Brian Ichter: co-founder, deepmind
8. Liyiming Ke: phd from UofW
9. Adrian Li-Bell: phd from Cambridge
10. Mohith Mothukuri: 3D product designer
11. Quan Vuong: phd from UCSD(Su Hao)
12. Anna Walling: operation guy
13. Ury Zhilinsky: manager?
14. rest of them are unkown.

TL;DR A generalist robot policy uses a pre-trained vision-language model (VLM) backbone, as well as a diverse crossembodiment dataset with a variety of dexterous manipulation tasks. The model is adapted to robot control by adding a separate action expert that produces continuous actions via flow matching, enabling complex multi-stage tasks with precise and fluent manipulation skills.

TODOs

1. read other VLA papers
2. read flow-matching, conditional flow-matching paper
    1. Flow matching for generative modeling
    2. Rectified flow: A marginal preserving approach to optimal transport
3. read transfussion
    1. Transfusion: Predict the next token and diffuse images with one multi-modal model
    2. Diffusion Forcing: Next-token Prediction Meets Full-Sequence Diffusion

Thoughts and critisims

1. Compared to a clear instruction in model architecture, the arthors did not fully reveal the training process.
2. The model split the inputs into a 'blockwise' attention mechanism. Will the block of states be attended to the image and task instruction?
3. Comparing to the other LLM, does the choice of 3B VLM backbone indicate we need less knowledge for our daily routines than an almighty assistant?

Related works

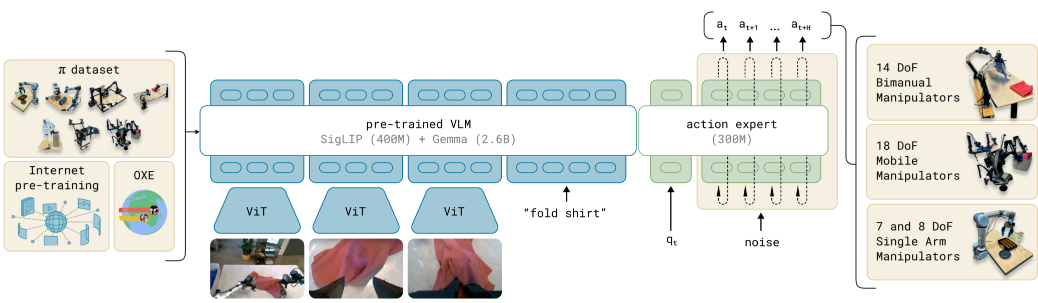
VLA

1. Definition: use pre-trained VLMs that are fine-tuned for robot control, the model employ autoregressive discretization to represent actions in a manner analogous to text tokens.
2. works
    1. RT-2 (previous work)
    2. Transfusion: Predict the next token and diffuse images with one multi-modal model (hybridize diffusion and autoregressive largelanguage models)

Contributions

1. a novel generalist robot policy architecture based on VLM pre-training and flow matching
2. An empirical investigation of pre-training/posttraining recipes for such robot foundation models.
3. We evaluate our model out of the box with language commands, with fine-tuning to downstream tasks.

Key concepts



The philosophys of paper

1. If VLA are to make tangible progress toward AI systems that exhibit the kind of physically situated versatility that people possess, we will need to train them on physically situated data
2. general-purpose foundation models that are pre-trained on diverse multi-task data tend to outperform narrowly tailored and specialized solutions
    1. resolve data scarcity
    2. resolve robustness and generalization challenges
3. VLA should
    1. be done on large scale data
    2. get the right model architecture
    3. get a right training recipe

Model

1. inputs
    1. images (supports multiview) at time t
    2. language instruction at time t
    3. robot states at time t
    4. noisy actions
2. outputs
    1. vector field v_t^τ (use to get H-step action chunk)
3. MoE architecture
    1. VLM backbone to process image and text
        1. PaliGemma 3B
        2. blockwise causal attention mask
    2. action backbone to generate actions
        1. 0.3B

Training

training

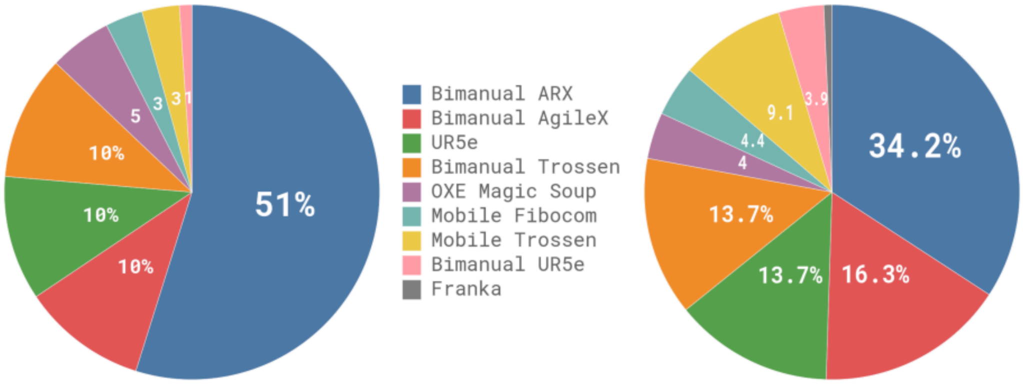


Fig. 4: **Overview of our dataset**: The pre-training mixture consists of a subset of OXE [10] and the π dataset. We use a subset of OXE, which we refer to as OXE Magic Soup [24]. The right figure illustrates the weight of the different datasets in the pre-training mixture. The left figure illustrates their relative sizes as measured by the number of steps.

{ width=70% }

1. dataset
    1. OXE Magic Soup and π dataset
    2. 1-2 cameras
    3. 2-10hz control
    4. π data consists of 106M steps single-arm robot and 797M dual-arm, with 68 tasks with complex behaviors
    5. unbalanced dataset
    6. add paddings to those robot with lower dim.
2. instruction processing
    1. use Saycan to decompose complex task

Hardwares

Fig. 5: **The robots used in our experiments.** These include single and dual-arm manipulators with 6-DoF and 7-DoF arms, as well as holonomic and nonholonomic mobile manipulators. π_0 is trained jointly on all of these platforms.

{ width=70% }

UR5e. An arm with a parallel jaw gripper, with a wrist-mounted and over-the-shoulder camera, for a total of two camera images and a 7-dimensional configuration and action space.

Bimanual UR5e. Two UR5e setups, for a total of three camera images and a 14-dimensional configuration and action space.

Franka. The Franka setup has two cameras and an 8-dimensional configuration and action space.

Bimanual Trossen. This setup has two 6-DoF Trossen ViperX arms in a configuration based on the ALOHA setup [4, 57], with two wrist cameras and a base camera, and a 14-dimensional configuration and action space.

Bimanual ARX & bimanual AgileX. This setup uses two 6-DoF arms, and supports either ARX or AgileX arms, with three cameras (two wrist and one base) and a 14-dimensional configuration and action space. This class encompasses two distinct platforms, but we categorize them together because of their similar kinematic properties.

Mobile Trossen & mobile ARX. This setup is based on the Mobile ALOHA [57] platform, with two 6-DoF arms on a mobile base, which are either ARX arms or Trossen ViperX arms. The nonholonomic base adds two action dimensions, for a 14-dimensional configuration and 16-dimensional action space. There are two wrist cameras and a base camera. This class encompasses two distinct platforms, but we categorize them together because of their similar kinematic properties.

Mobile Fibocom. Two 6-DoF ARX arms on a holonomic base. The base adds three action dimensions (two for translation and one for orientation), for a 14-dimensional configuration and 17-dimensional action space.

We summarize the proportion of our dataset from each robot in Figure 4.

{ width=70% }

Experiments

questions

1. How well does π_0 perform after pre-training on a variety of tasks that are present in the pre-training data
2. How well does the model follow language commands?
3. How does the model compare to methods that have been proposed specifically for addressing dexterous manipulation tasks
4. Can the model be adapted to complex, multi-stage tasks?