

ICCV 2025 Tutorial
Time: 2025-10-20
Location: 306B

Foundation Models Meet Embodied Agents



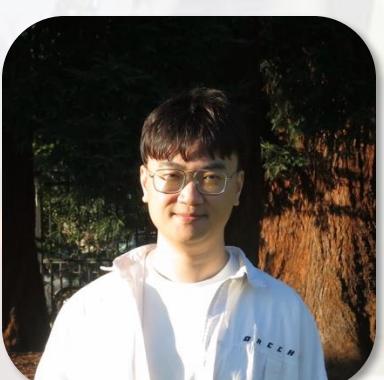
Manling Li
Northwestern



Yunzhu Li
Columbia



Jiayuan Mao
Amazon FAR and UPenn



Wenlong Huang
Stanford



Northwestern
University

COLUMBIA

Penn
UNIVERSITY OF PENNSYLVANIA

Stanford
University

ICCV 2025 Tutorial
Time: 2025-10-20
Location: 306B

Robotic Foundation Models

ICCV Tutorial: Foundation Models Meet Embodied Agents



Northwestern
University





World ▾ Business ▾ Markets ▾ Sustainability ▾ More ▾

Robot AI startup Physical Intelligence raises \$400 mln from Bezos, OpenAI

By Reuters

November 4, 2024 12:38 PM EST · Updated 3 months ago



Series B: 1X Secures \$100M Funding

January 11, 2024

Author: 1X

TECH

Nvidia, Bill Gates-backed robotics startup Field AI hits \$2 billion valuation after recent raise

PUBLISHED WED, AUG 20 2025 10:03 AM EDT | UPDATED WED, AUG 20 2025 4:53 PM EDT

Skild AI grabs \$300M to build foundation model for robotics

By Mike Oitzman | July 10, 2024

From self-driving cars to chore-battling bots: Robot Guru Kyle Vogt raises \$150M for The Bot Company



BY VIVEK CHHETRI · MAY 14, 2024 · 2 MINUTE READ



World ▾ Business ▾ Markets ▾ Sustainability ▾ More ▾

Robotics startup Figure raises \$675 mln from Microsoft, Nvidia, OpenAI

By Harshita Mary Varghese and Krystal Hu

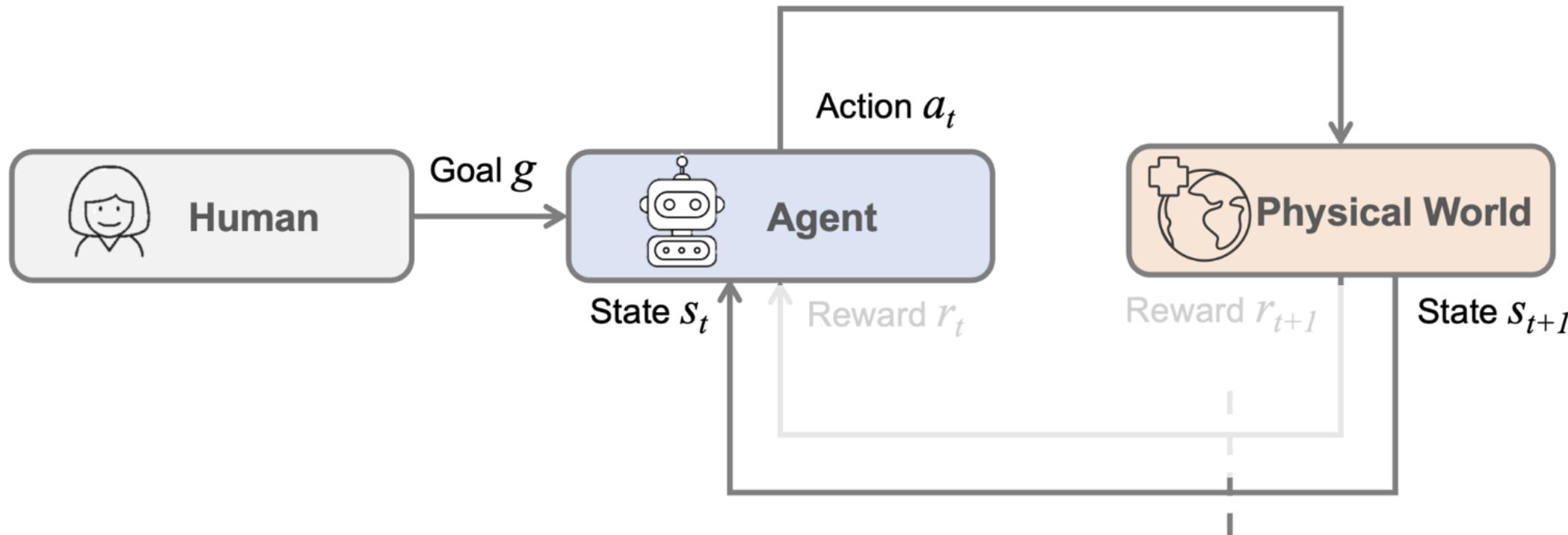
February 29, 2024 11:20 AM EST · Updated a year ago



Robotic Foundation Models



- What is a Robotic Foundation Model?
 - No explicit representation of states / transition functions
 - A policy that maps (observation/state, goal) to action



Robotic Foundation Models



- What is a Robotic Foundation Model?
 - No explicit representation of states / transition functions
 - A policy that maps (observation/state, goal) to action

Imitation Learning
(Chi et al., Diffusion Policy)



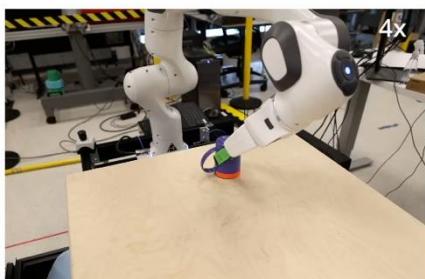
Diffusion Policy

LSTM-GMM

BET

IBC

Diffusion Policy learns multi-modal behavior and commits to only one mode within each rollout. LSTM-GMM and IBC are biased toward one mode, while BET failed to commit.



Reinforcement Learning
(OpenAI, Solving Rubik's Cube)



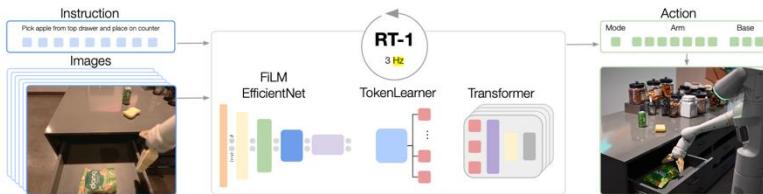
- ❑ What is a Robotic Foundation Model?
 - ❑ No explicit representation of states / transition functions
 - ❑ A policy that maps (observation/state, goal) to action
- ❑ Current Foundational Vision-and-Language Models
 - ❑ The output may **not** always be **perfect**.
 - ❑ It will always generate something **reasonable**.
- ❑ Robotic Foundation Models
 - ❑ The synthesized action may **not** always be **optimal**.
 - ❑ The generated trajectory will always be **beautiful** and **reasonable**.
- ❑ Different names
 - ❑ Vision-Language-Action Models (VLAs), Large behavior models (LBMs)

Robotic Foundation Models

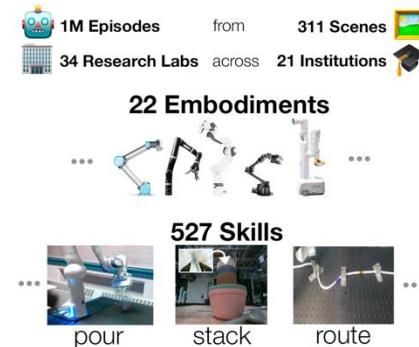


□ What is a Robotic Foundation Model?

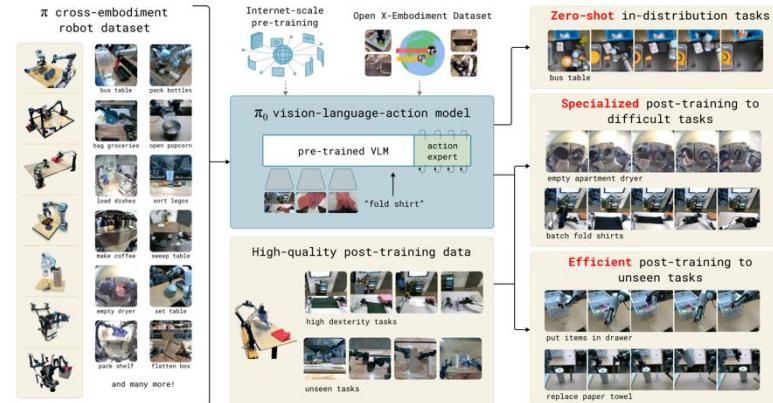
- No explicit representation of states / transition functions
- A policy that maps (observation/state, goal) to action



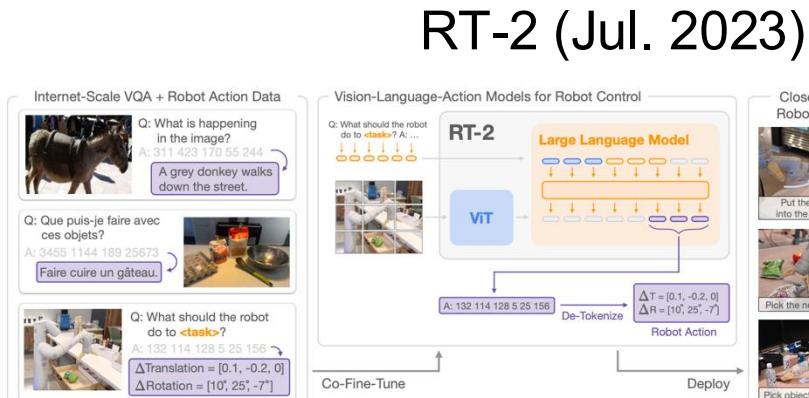
RT-1 (Dec. 2022)



RT-X (Oct. 2023)

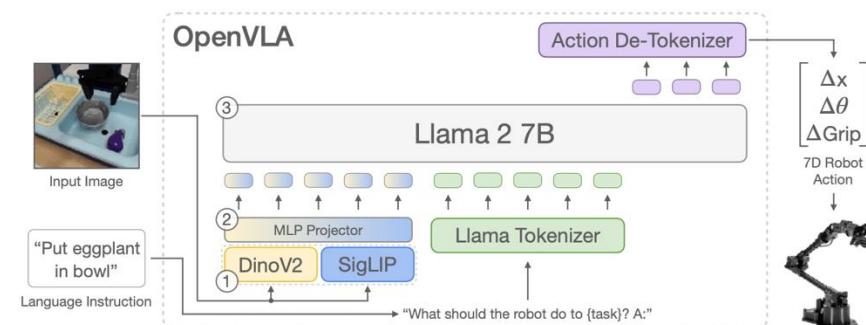


Pi-Zero (Oct. 2024)



RT-2 (Jul. 2023)

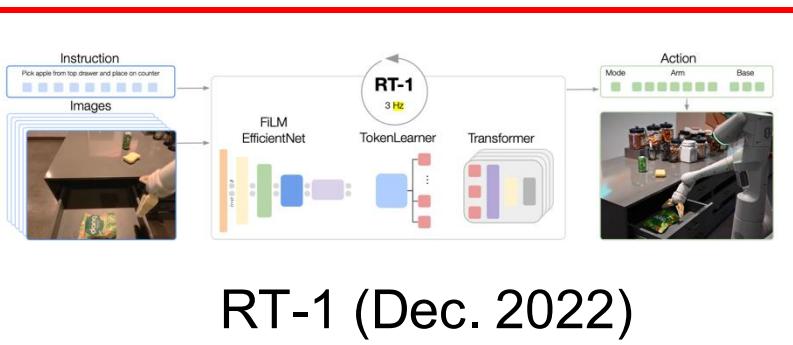
OpenVLA (Jun. 2024)



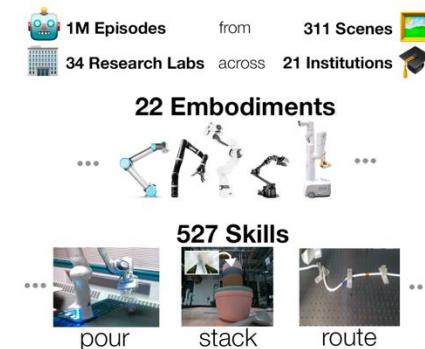
Robotic Foundation Models



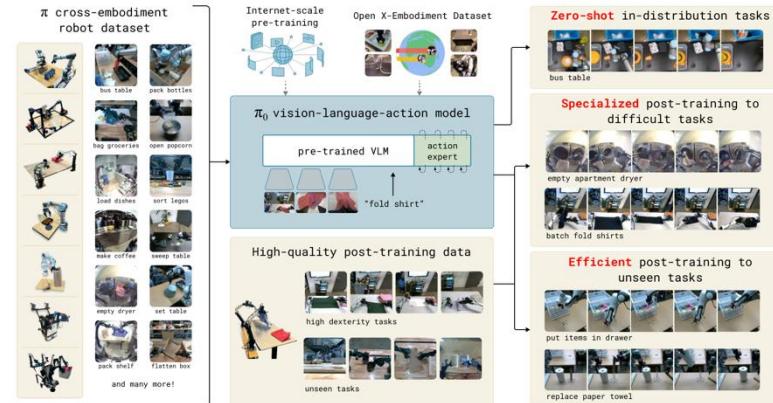
- What is a Robotic Foundation Model?
 - No explicit representation of states / transition functions
 - A policy that maps (observation/state, goal) to action



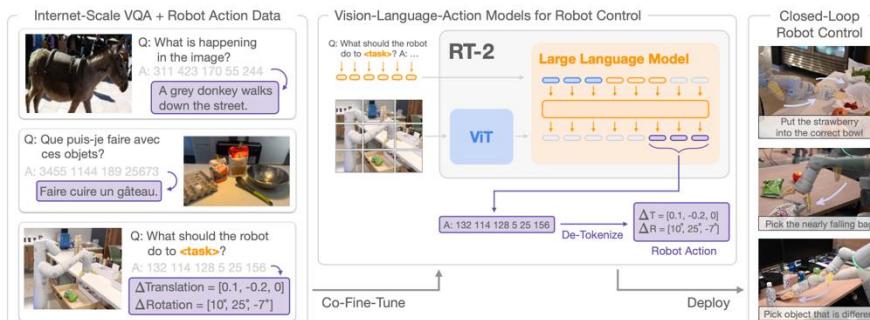
RT-1 (Dec. 2022)



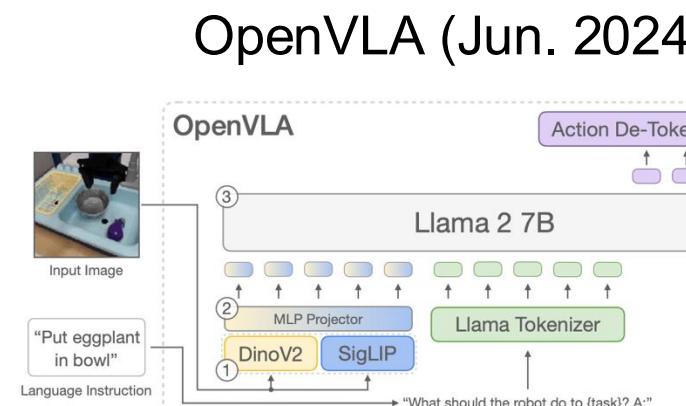
RT-X (Oct. 2023)



Pi-Zero (Oct. 2024)



RT-2 (Jul. 2023)



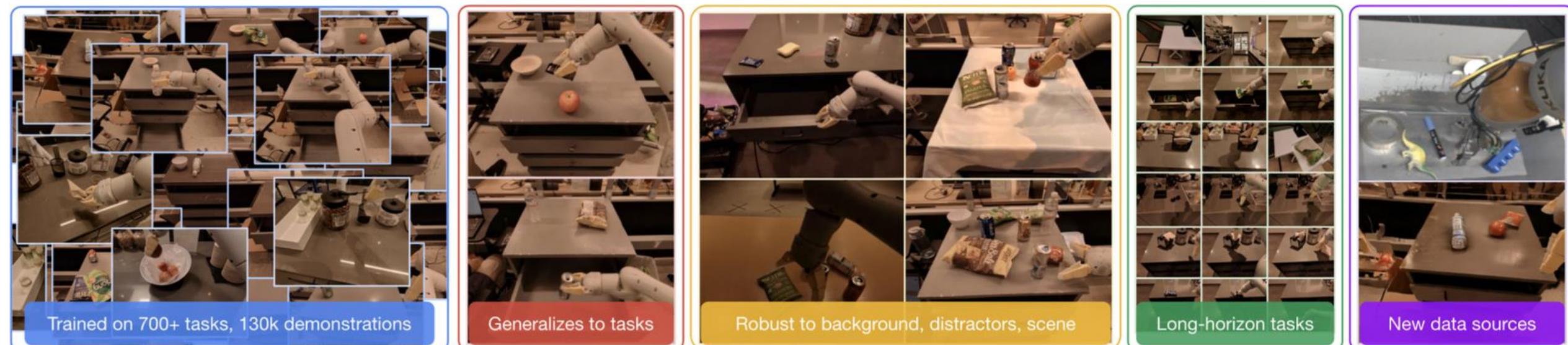
OpenVLA (Jun. 2024)

Robotic Transformer 1 (RT-1)



- ❑ First released in December 2022
- ❑ Huge success in large-scale training for CV and NLP
- ❑ Can these lessons be applied to robotics?
- ❑ Large-scale data collection efforts from Google

17 months with a fleet of 13 robots, containing ~130k episodes and over 700 tasks



Robotic Transformer 1 (RT-1)



- ❑ First released in December 2022
- ❑ Huge success in large-scale training for CV and NLP
- ❑ Can these lessons be applied to robotics?
- ❑ Large-scale data collection efforts from Google



(a)



(b)



(c)



(d)



(e)

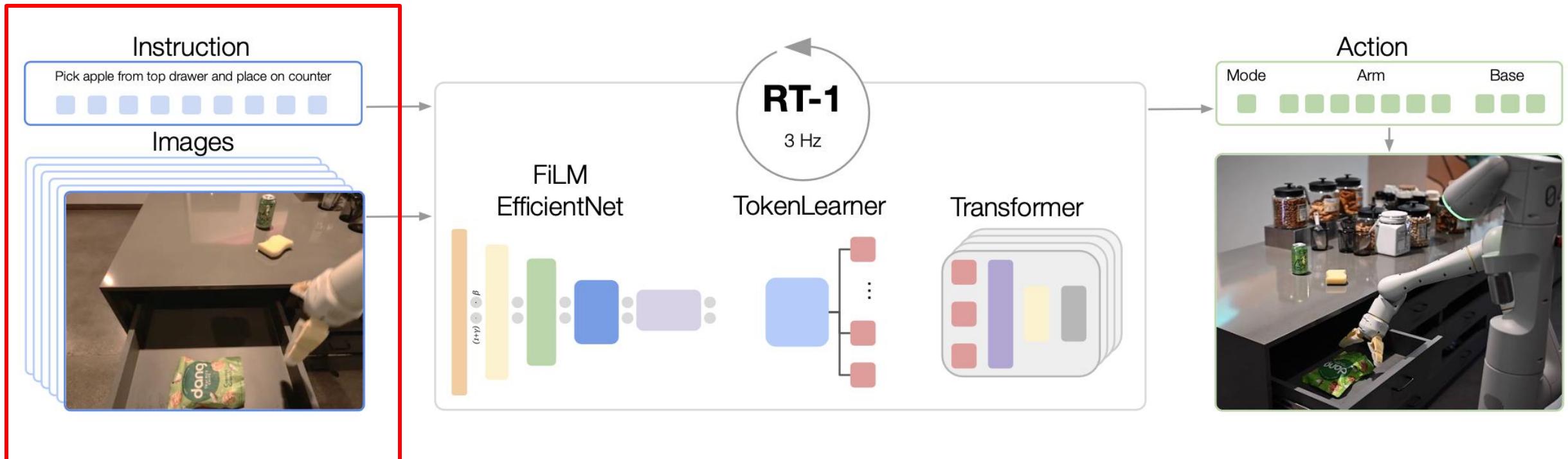


(f)

Robotic Transformer 1 (RT-1)



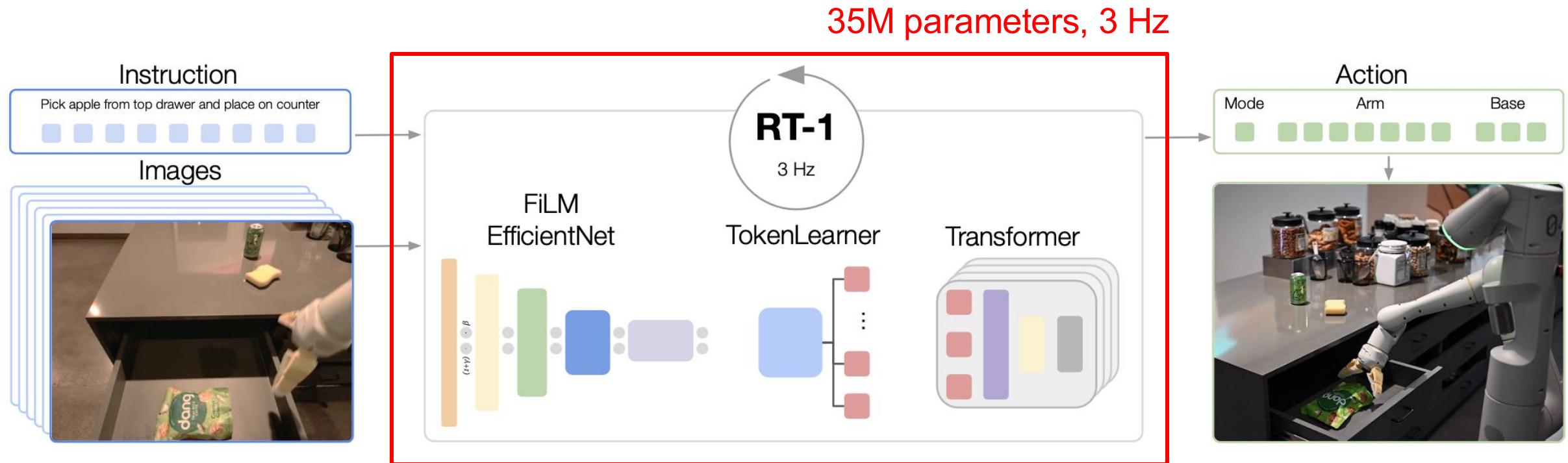
- Large-scale imitation learning
 - A policy that maps (observation/state, goal) to action



Robotic Transformer 1 (RT-1)



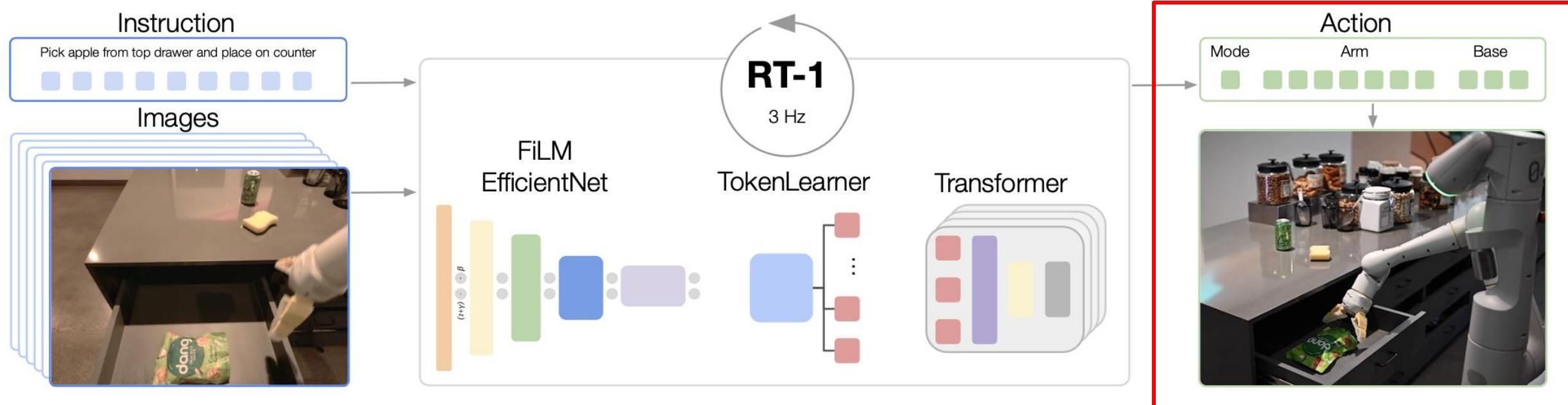
- Large-scale imitation learning
 - A policy that maps (observation/state, goal) to action



Robotic Transformer 1 (RT-1)



- Large-scale imitation learning
 - A policy that maps (observation/state, goal) to action

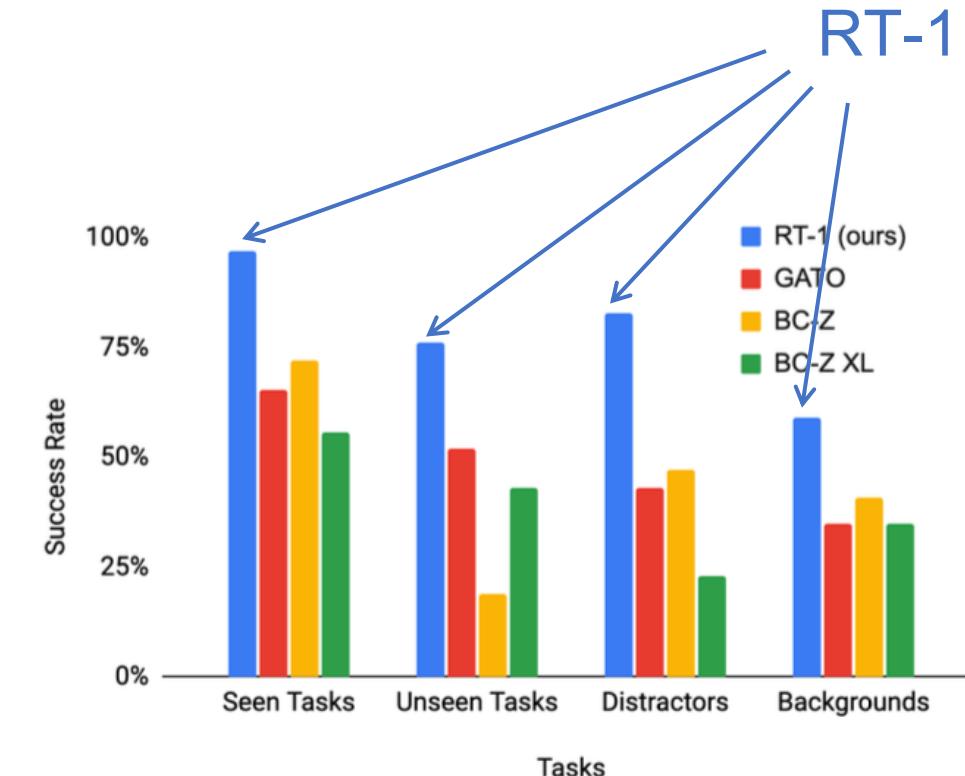


Robotic Transformer 1 (RT-1)



- Question #1: Can an RT-1 learn to perform language-conditioned tasks?

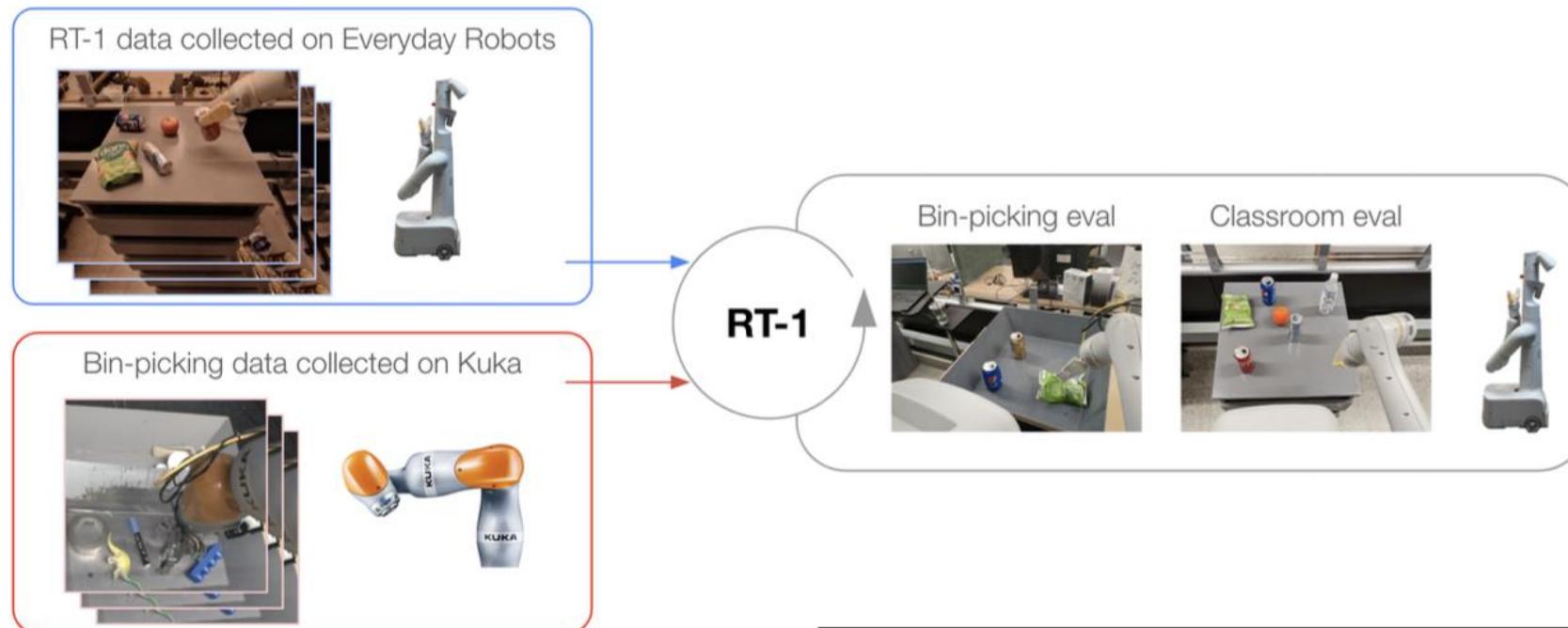
Model	Seen Tasks	Unseen Tasks	Distractors	Backgrounds
Gato (Reed et al., 2022)	65	52	43	35
BC-Z (Jang et al., 2021)	72	19	47	41
BC-Z XL	56	43	23	35
RT-1 (ours)	97	76	83	59



Robotic Transformer 1 (RT-1)



□ Question #2: Data from different robot?



Bin-picking data from
a different robot also helps

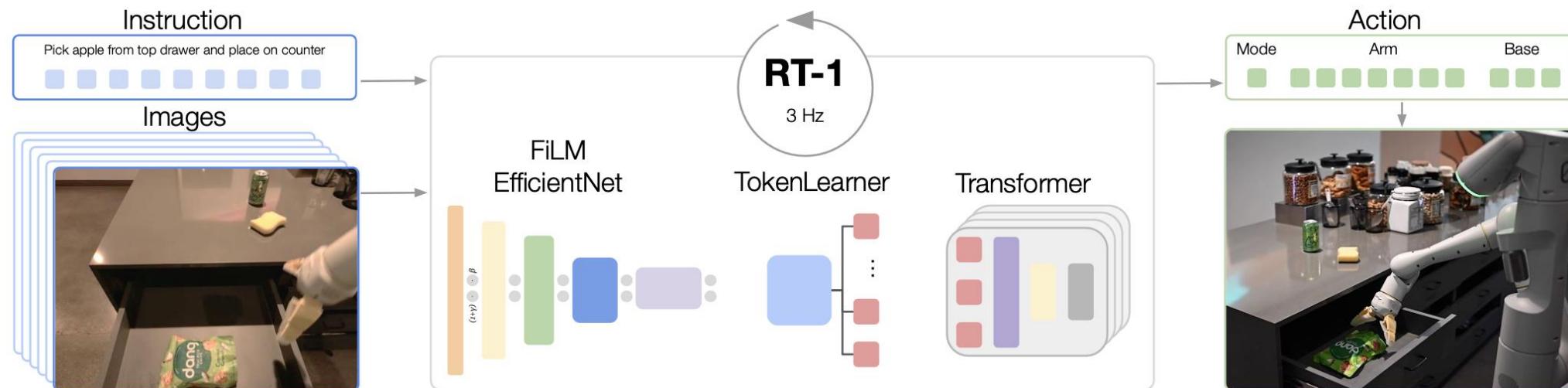
Models	Training Data	Classroom eval	Bin-picking eval
RT-1	Kuka bin-picking data + EDR data	90(-2)	39(+17)
RT-1	EDR only data	92	22
RT-1	Kuka bin-picking only data	0	0

Robotic Transformer 1 (RT-1)



- ❑ Large-scale language-conditioned imitation learning.
- ❑ Significant data collection and engineering efforts.
- ❑ Among the initial investigations: (1) how to scale up and (2) what to expect.

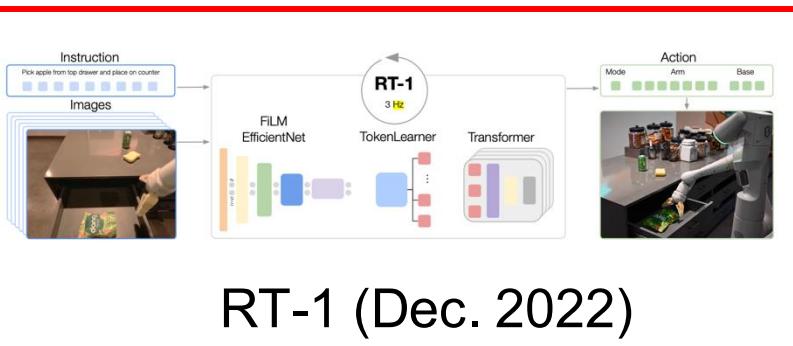
- ❑ Haven't leveraged larger-scale internet data.
- ❑ Cannot generalize to new skills.
- ❑ Efficiency limited to simple and quasi-static tasks.



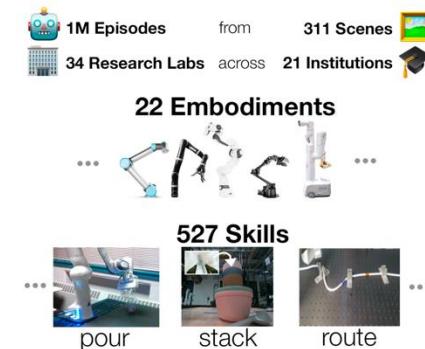
Robotic Foundation Models



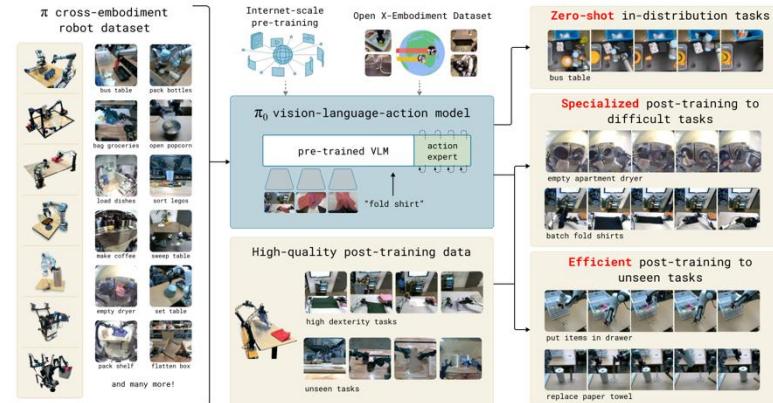
- What is a Robotic Foundation Model?
 - No explicit representation of states / transition functions
 - A policy that maps (observation/state, goal) to action



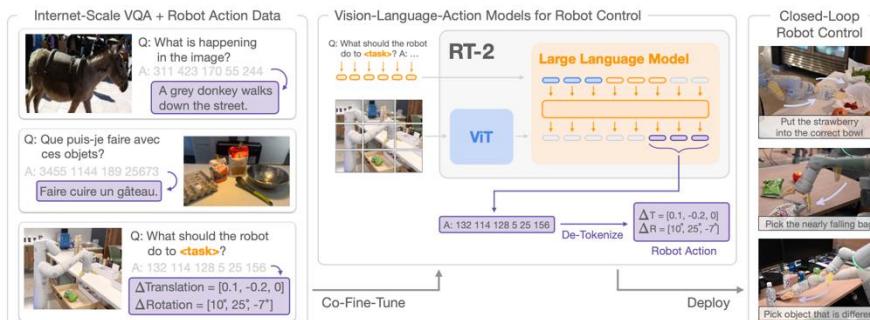
RT-1 (Dec. 2022)



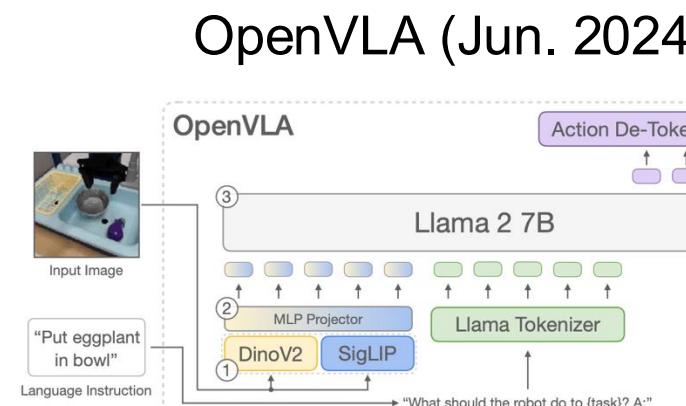
RT-X (Oct. 2023)



Pi-Zero (Oct. 2024)



RT-2 (Jul. 2023)

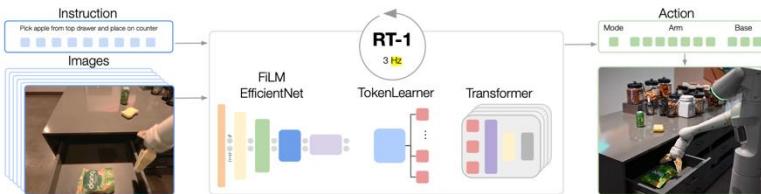


OpenVLA (Jun. 2024)

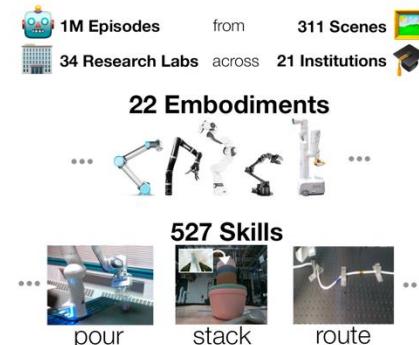
Robotic Foundation Models



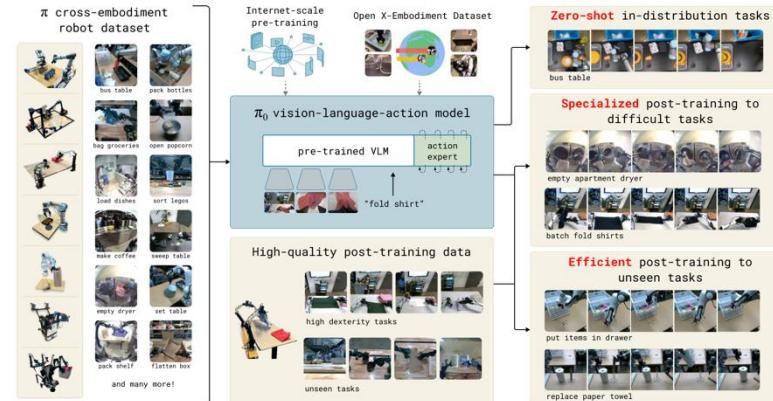
- What is a Robotic Foundation Model?
 - No explicit representation of states / transition functions
 - A policy that maps (observation/state, goal) to action



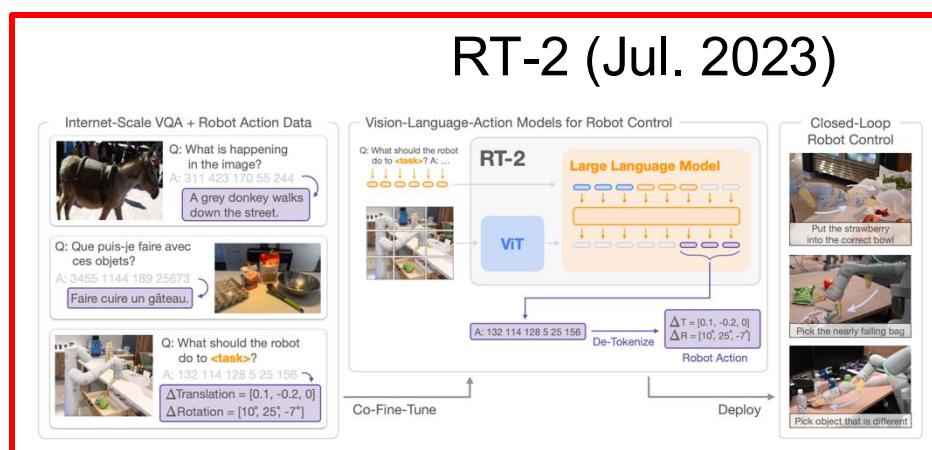
RT-1 (Dec. 2022)



RT-X (Oct. 2023)

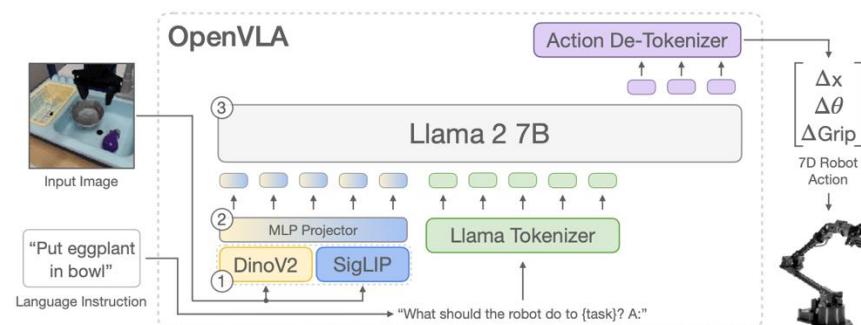


Pi-Zero (Oct. 2024)



RT-2 (Jul. 2023)

OpenVLA (Jun. 2024)

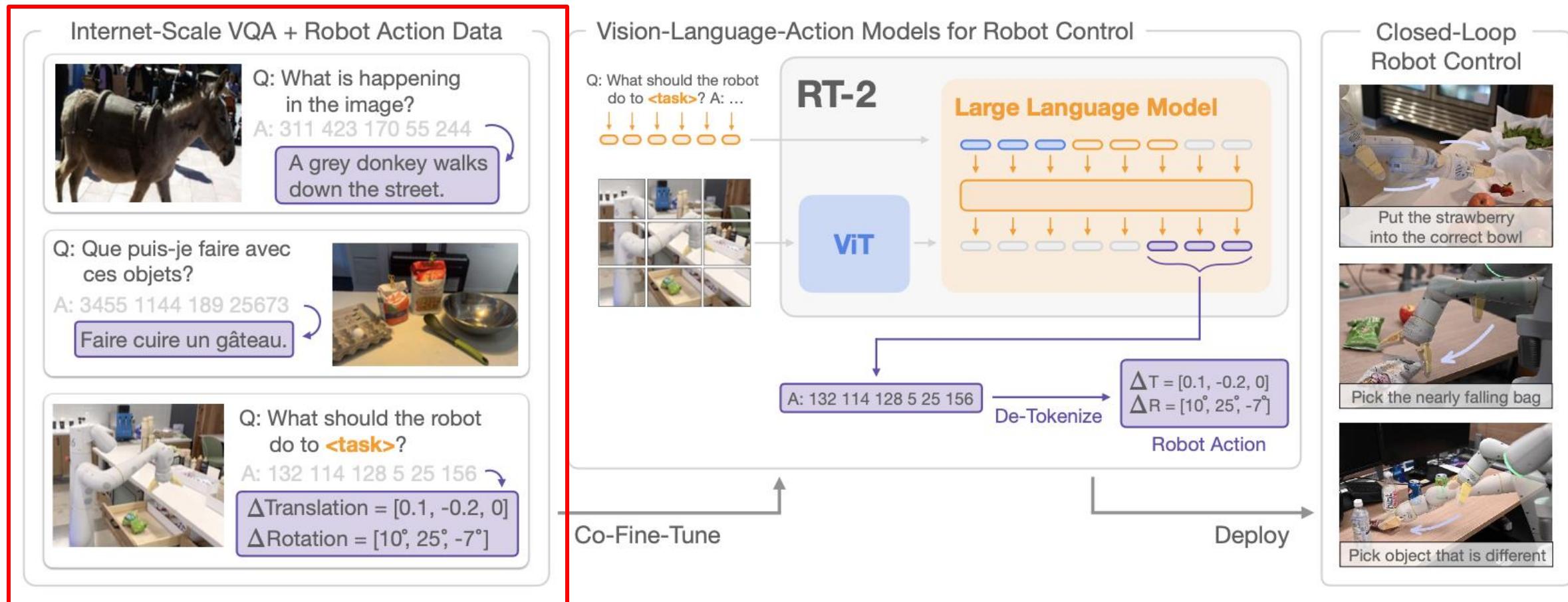


Robotic Transformer 2 (RT-2)



- ❑ First released in July 2023
- ❑ How VLMs can be incorporated into Robotic Foundation Models?
- ❑ Key idea: co-fine-tune VLMs on both
 - ❑ (1) robot data
 - ❑ (2) Internet-scale vision-language tasks (e.g., VQA)
- ❑ Introduced the name: Vision-Language-Action Models (VLA)

Robotic Transformer 2 (RT-2)



Tokenize the robot actions

Robotic Transformer 2 (RT-2)

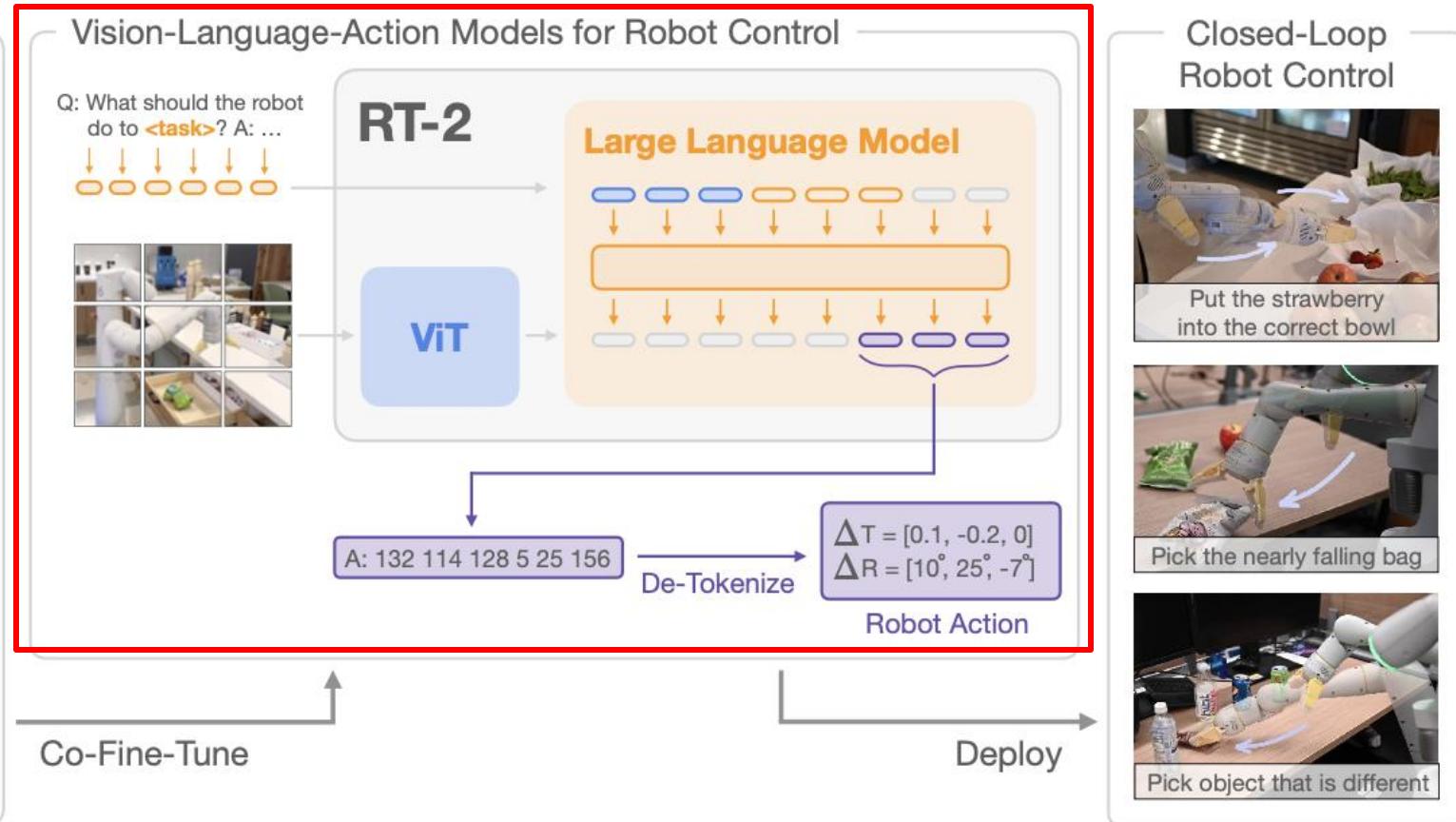


Internet-Scale VQA + Robot Action Data

Q: What is happening in the image?
A: 311 423 170 55 244
A grey donkey walks down the street.

Q: Que puis-je faire avec ces objets?
A: 3455 1144 189 25673
Faire cuire un gâteau.

Q: What should the robot do to <task>?
A: 132 114 128 5 25 156
 $\Delta\text{Translation} = [0.1, -0.2, 0]$
 $\Delta\text{Rotation} = [10^\circ, 25^\circ, -7^\circ]$

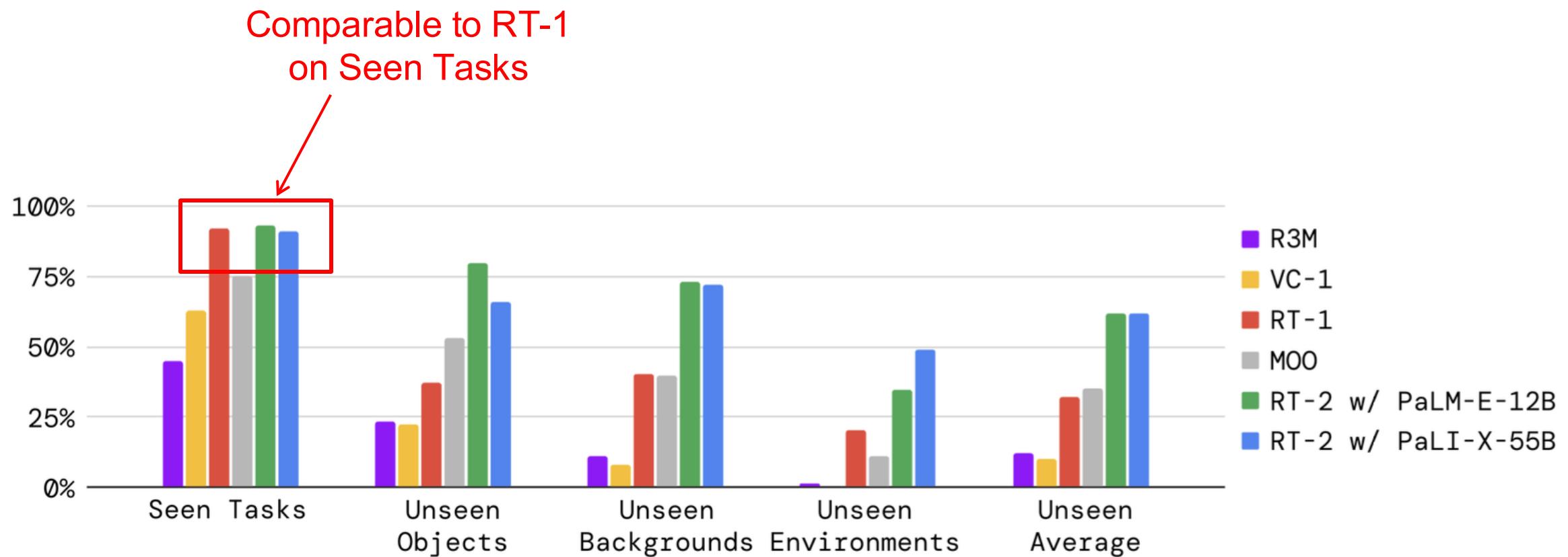


- Co-Fine-Tuning
- 55B (1~3 Hz), 5B (5Hz)
- Cannot run locally, developed a multi-TPU cloud service
- Querying this service over the network

Robotic Transformer 2 (RT-2)



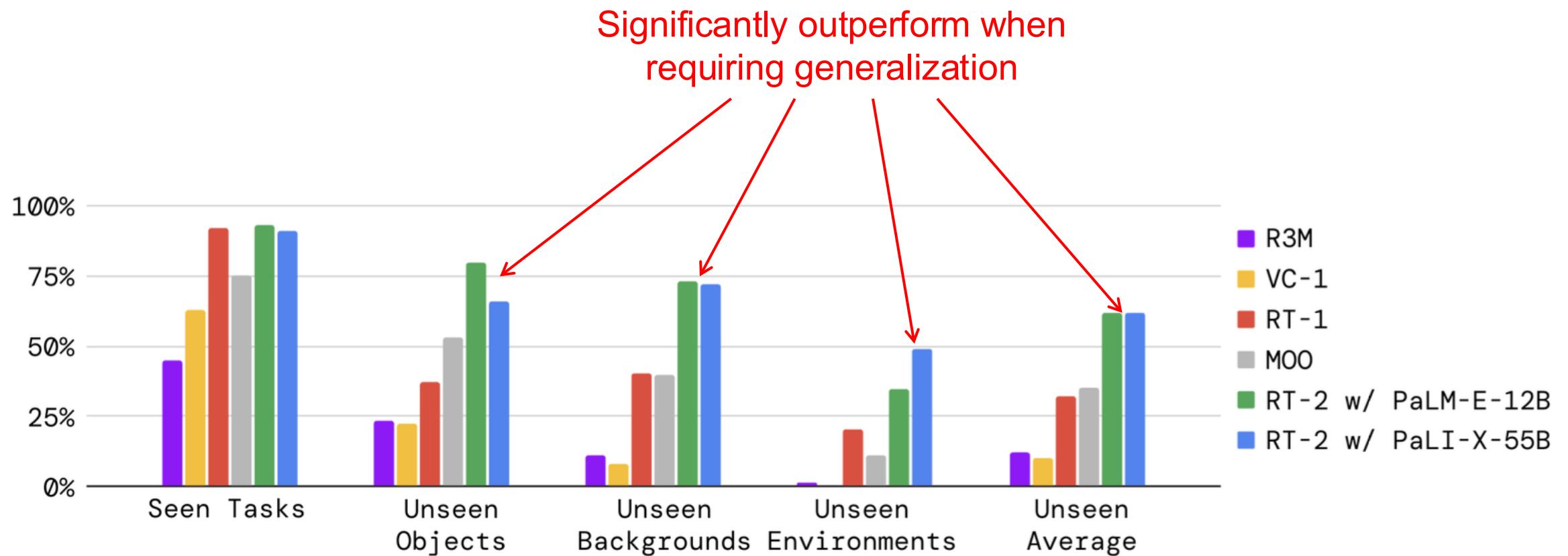
- How well does it work?



Robotic Transformer 2 (RT-2)



- How well does it work?



Robotic Transformer 2 (RT-2)



put strawberry
into the correct
bowl



pick up the bag
about to fall
off the table



move apple to
Denver Nuggets



pick robot



place orange in
matching bowl



move redbull can
to H



move soccer ball
to basketball



move banana to
Germany



move cup to the
wine bottle



pick animal with
different colour



move coke can to
Taylor Swift



move coke can to
X



move bag to
Google



move banana to
the sum of two
plus one



pick land animal

Robotic Transformer 2 (RT-2)



Robotics

Google's DeepMind team highlights new system for teaching robots novel tasks

ARTIFICIAL INTELLIGENCE / TECH

Brian Heater @bheav

Google is training robots the way it trains AI chatbots



/ Google's new robots don't need complex instructions now that they can access large language

WILL KNIGHT BUSINESS AUG 16, 2022 10:00 AM

Google's New Robot Learned to Take Orders by Scraping the Web

The machine learning technique that taught notorious text generator GPT-3 to write can also help robots make sense of spoken commands.



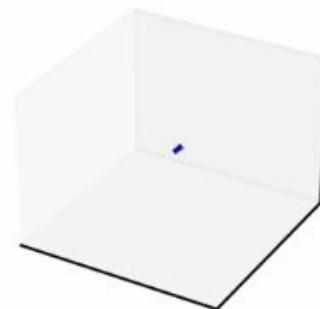
COURTESY OF GOOGLE

Google's RT-2 AI Model: A Step Closer To Robots That Can Learn Like Humans

Janakiram MSV Senior Contributor

I cover emerging technologies with a focus on infrastructure and AI

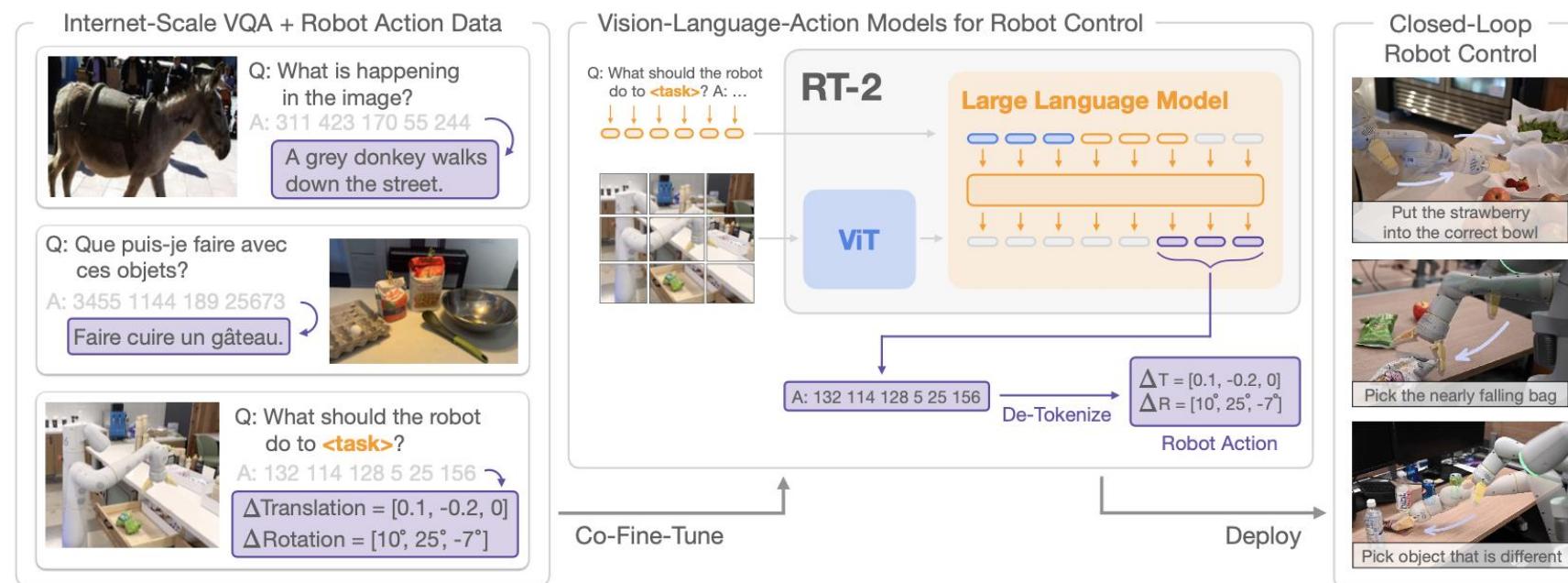
Follow



Robotic Transformer 2 (RT-2)



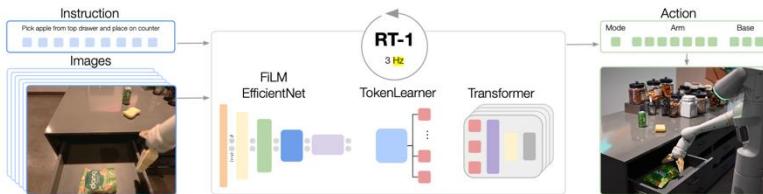
- ❑ Co-fine-tuning boosts generalization over semantic and visual concepts
- ❑ Limited to seen skills but can deploy them in new ways
- ❑ Efficiency is still an issue
- ❑ The absolute performance is still not ideal



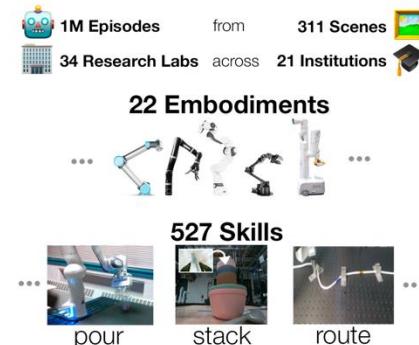
Robotic Foundation Models



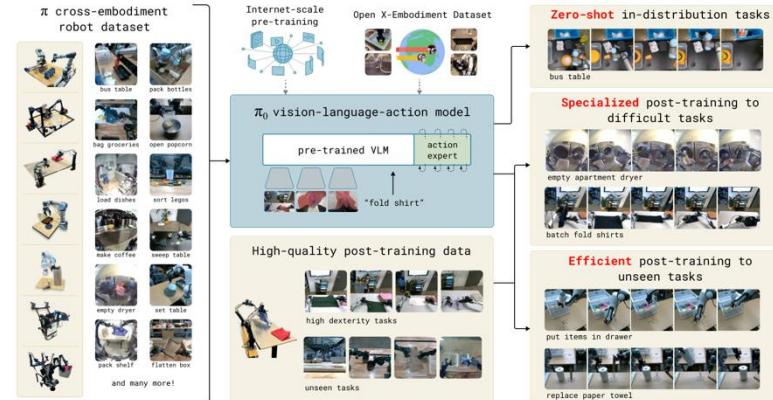
- What is a Robotic Foundation Model?
 - No explicit representation of states / transition functions
 - A policy that maps (observation/state, goal) to action



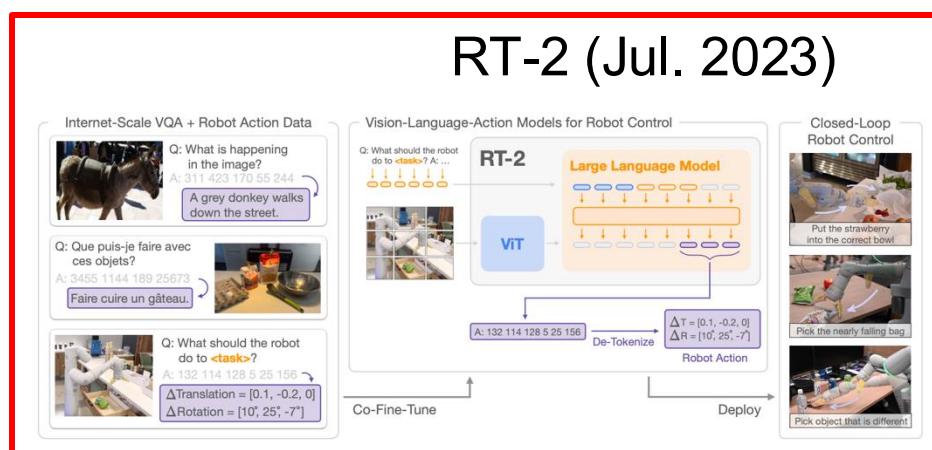
RT-1 (Dec. 2022)



RT-X (Oct. 2023)

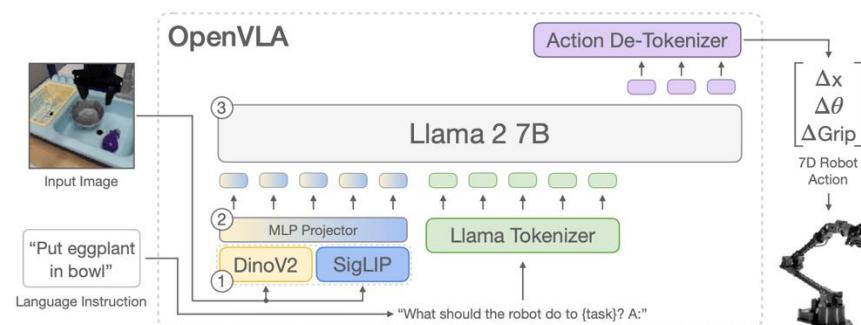


Pi-Zero (Oct. 2024)



RT-2 (Jul. 2023)

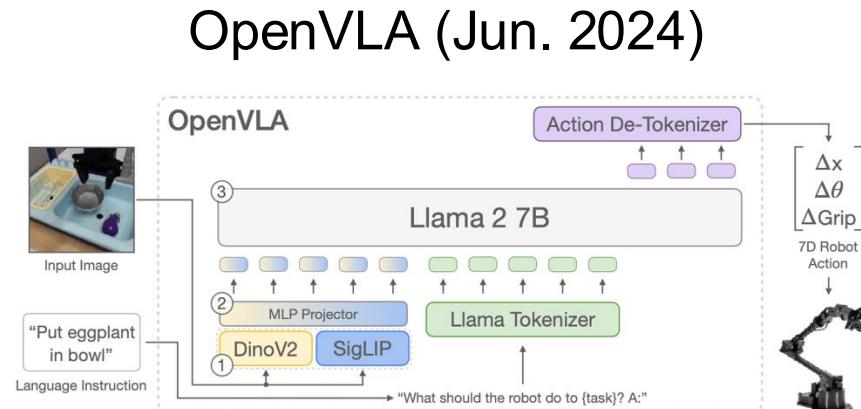
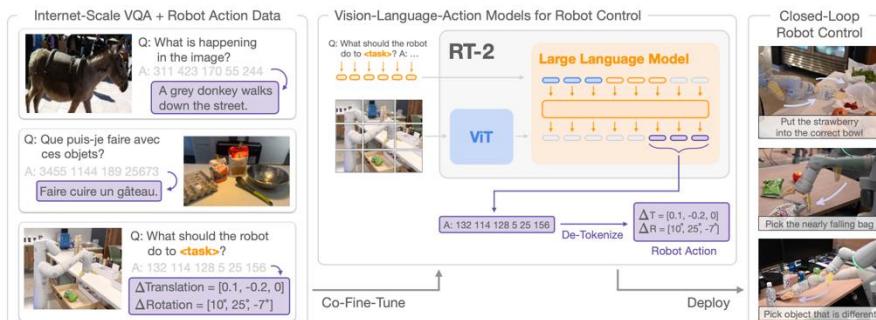
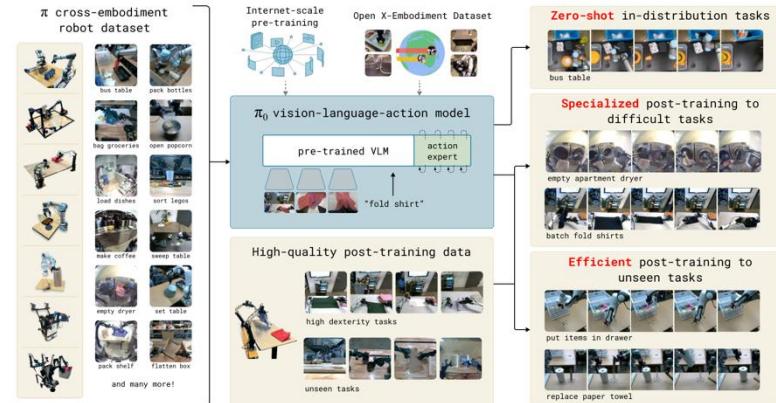
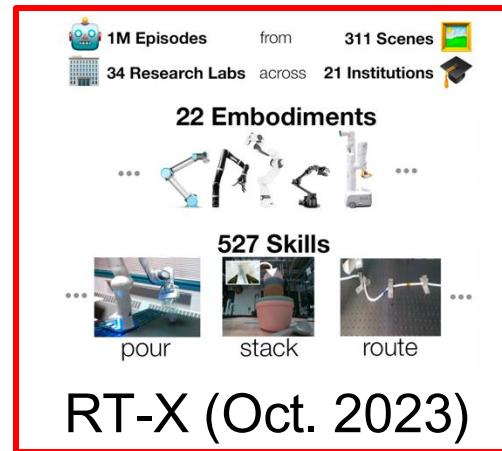
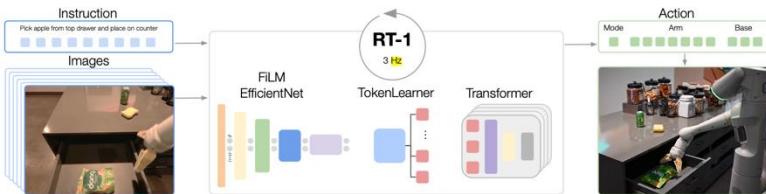
OpenVLA (Jun. 2024)



Robotic Foundation Models



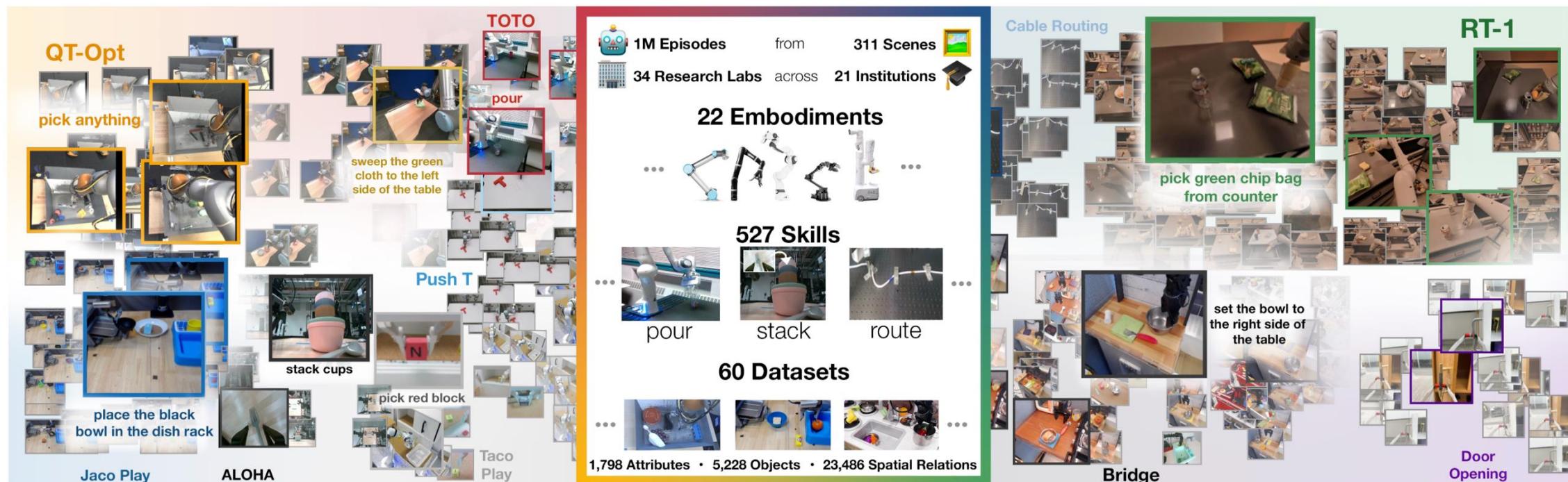
- What is a Robotic Foundation Model?
 - No explicit representation of states / transition functions
 - A policy that maps (observation/state, goal) to action



Robotic Transformer X (RT-X)



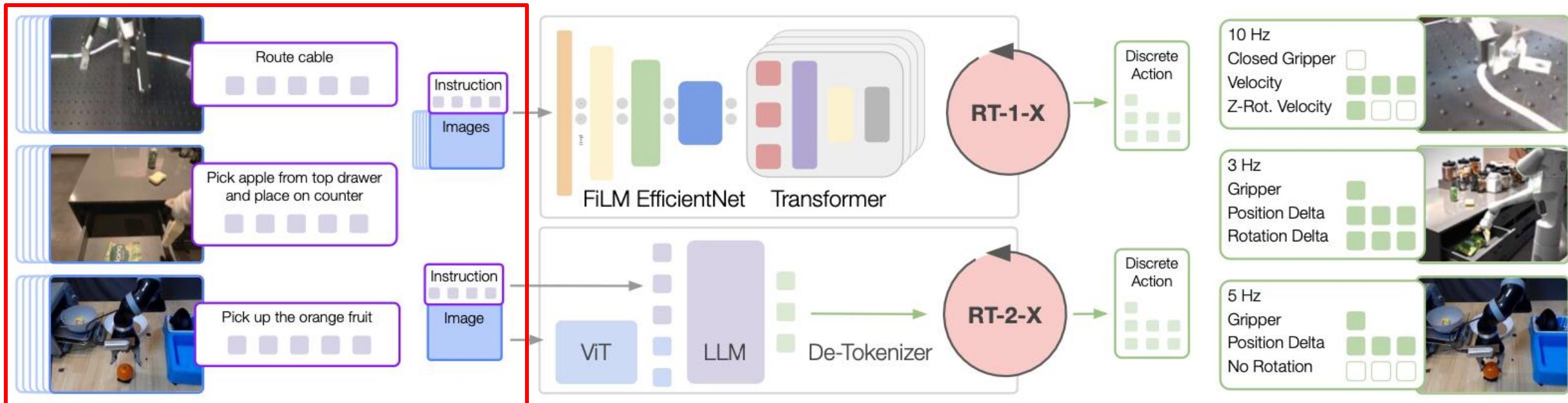
- ❑ First released in October 2023
- ❑ Instead of a single data source
 - ❑ 22 different robots collected through a collaboration between 21 institutions
 - ❑ demonstrating 527 skills (160,266 tasks)



Robotic Transformer X (RT-X)



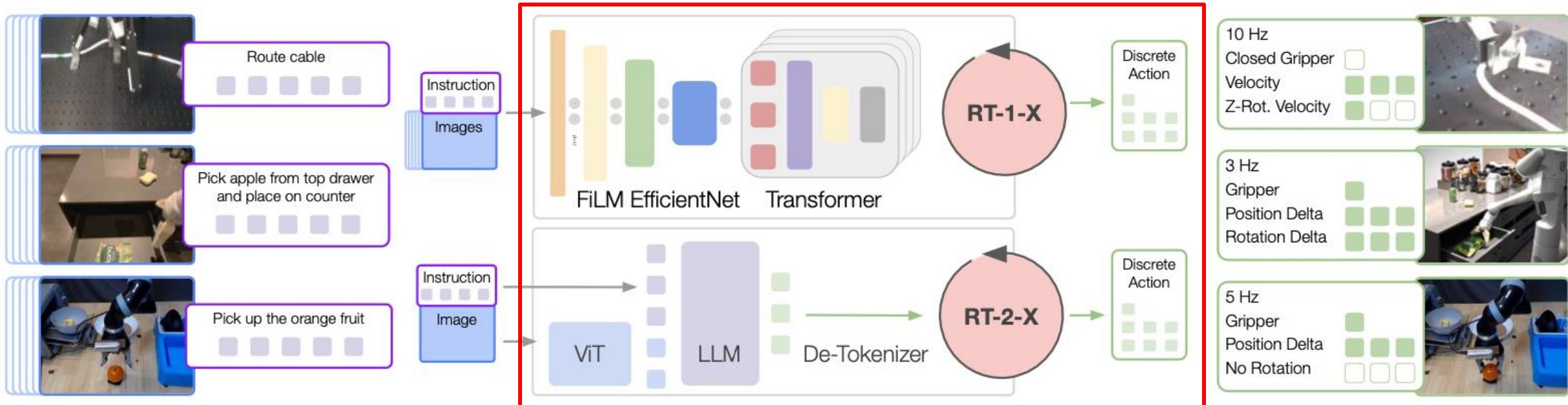
- Question: whether policies trained on data from many different robots and environments enjoy the benefits of positive transfer?



Robotic Transformer X (RT-X)



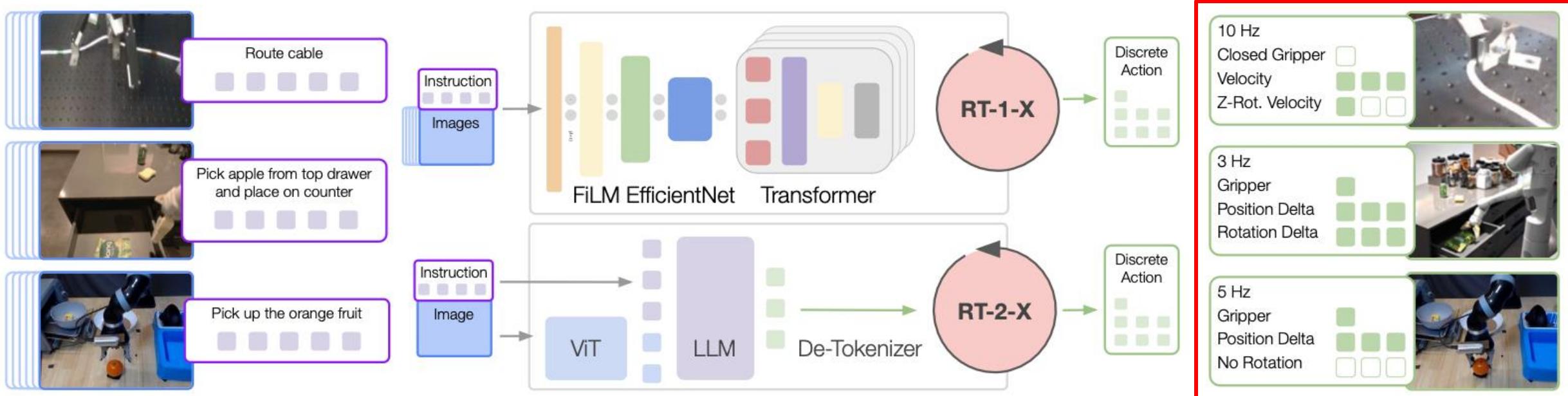
- Question: whether policies trained on data from many different robots and environments enjoy the benefits of positive transfer?



Robotic Transformer X (RT-X)



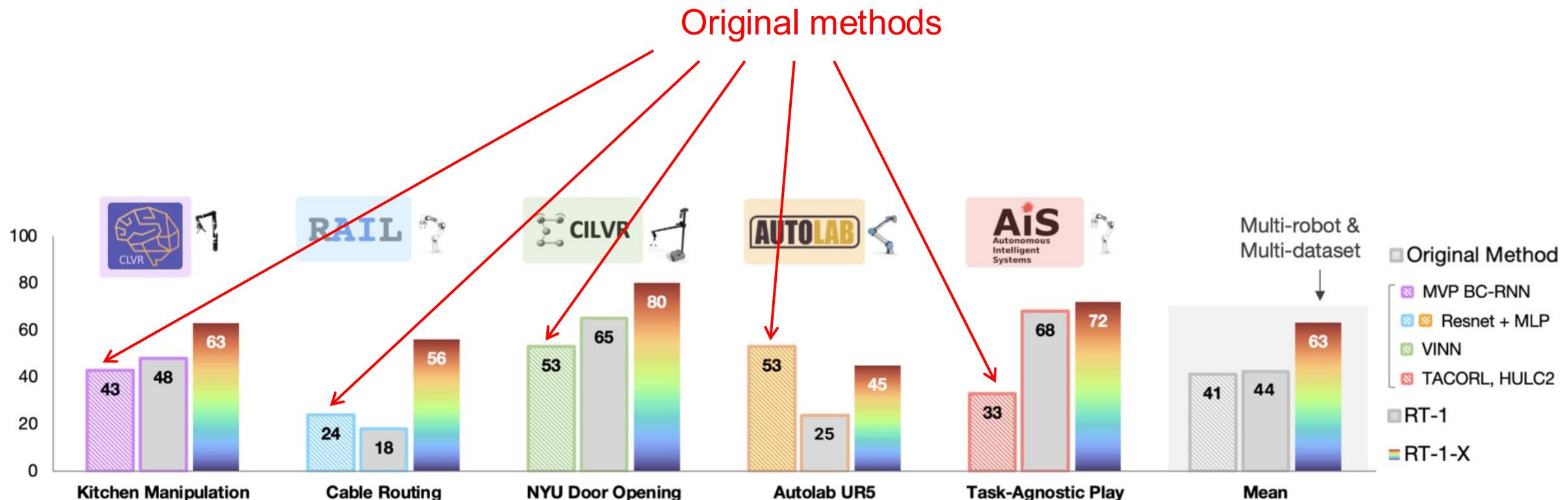
- Question: whether policies trained on data from many different robots and environments enjoy the benefits of positive transfer?



Robotic Transformer X (RT-X)



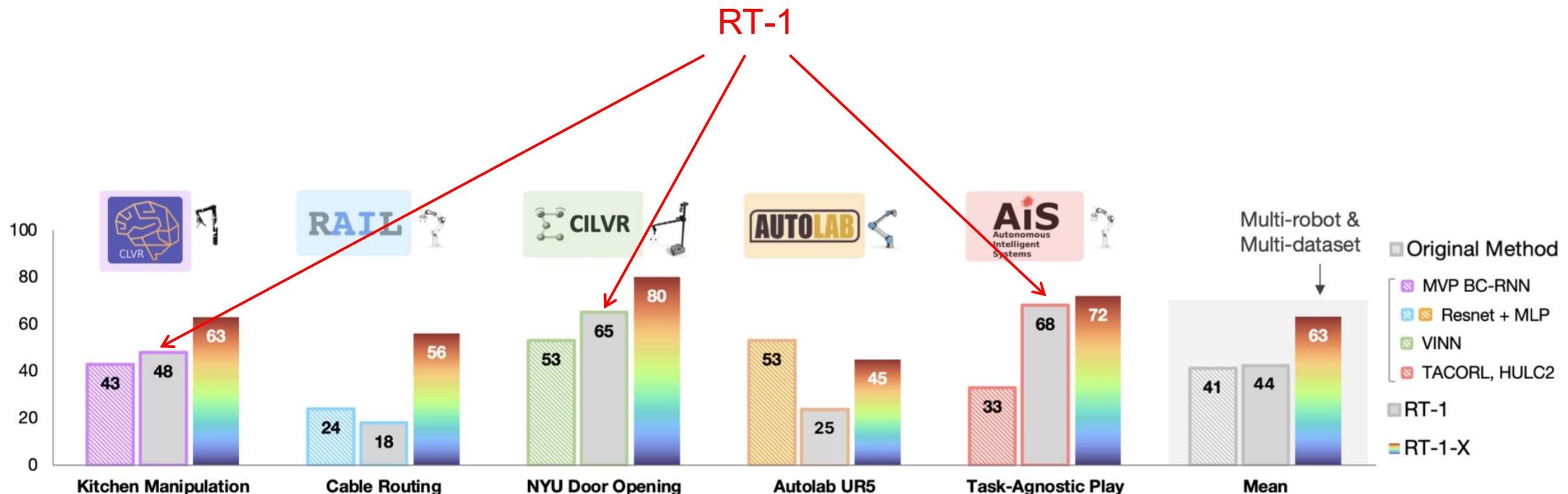
- Question: whether policies trained on data from many different robots and environments enjoy the benefits of positive transfer?



Robotic Transformer X (RT-X)



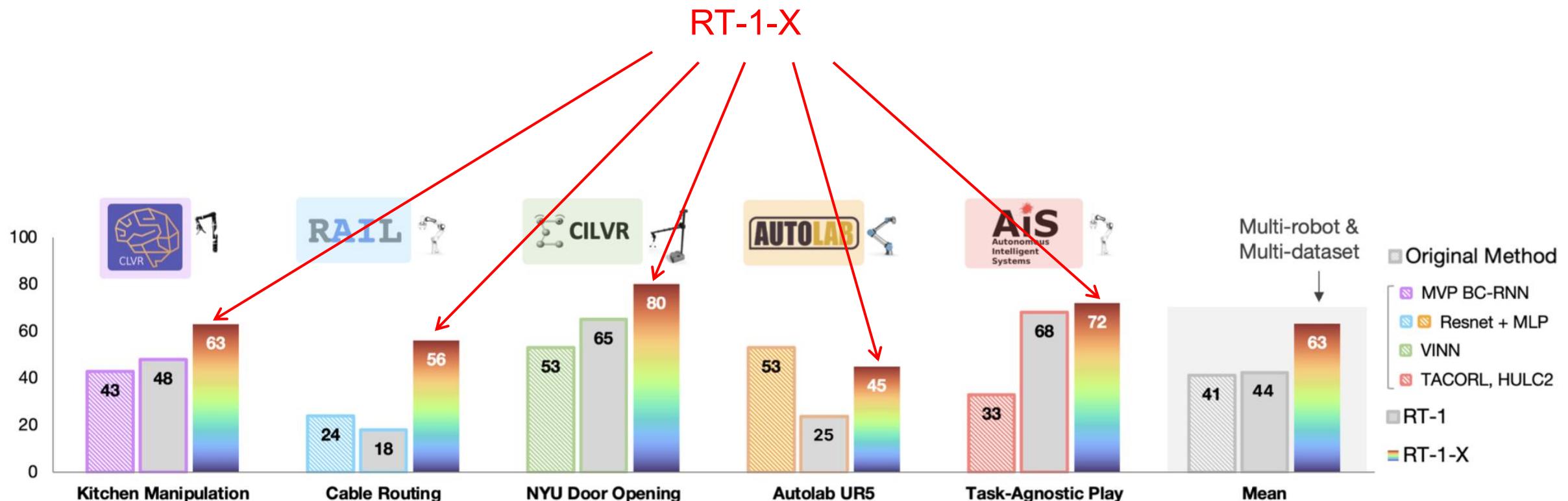
- Question: whether policies trained on data from many different robots and environments enjoy the benefits of positive transfer?



Robotic Transformer X (RT-X)



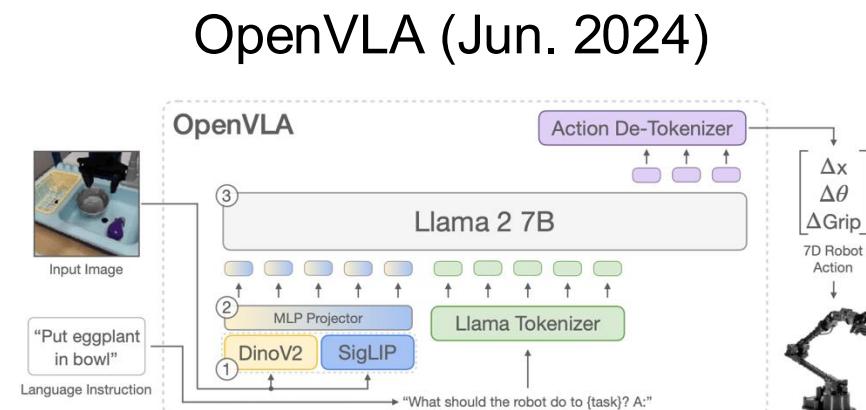
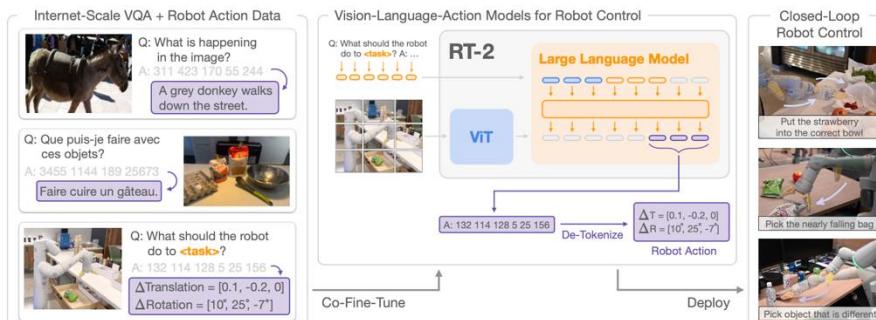
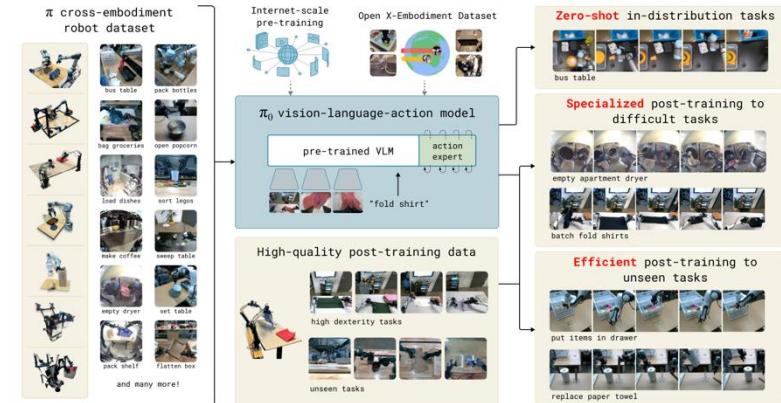
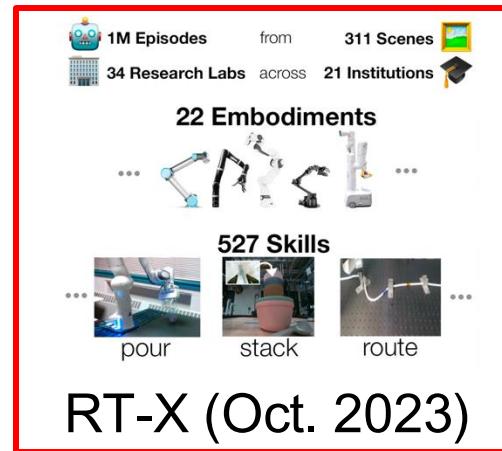
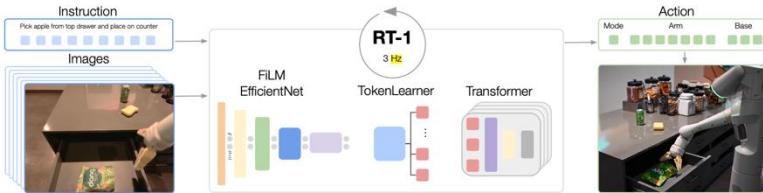
- Question: whether policies trained on data from many different robots and environments enjoy the benefits of positive transfer?



Robotic Foundation Models



- What is a Robotic Foundation Model?
 - No explicit representation of states / transition functions
 - A policy that maps (observation/state, goal) to action

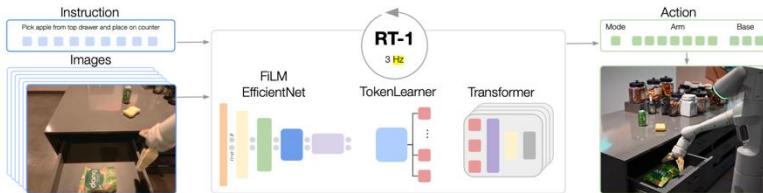


Robotic Foundation Models

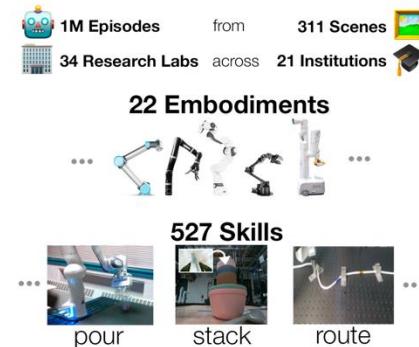


□ What is a Robotic Foundation Model?

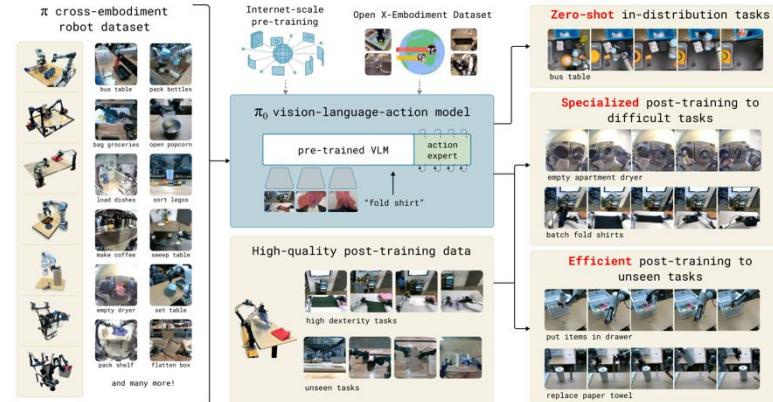
- No explicit representation of states / transition functions
- A policy that maps (observation/state, goal) to action



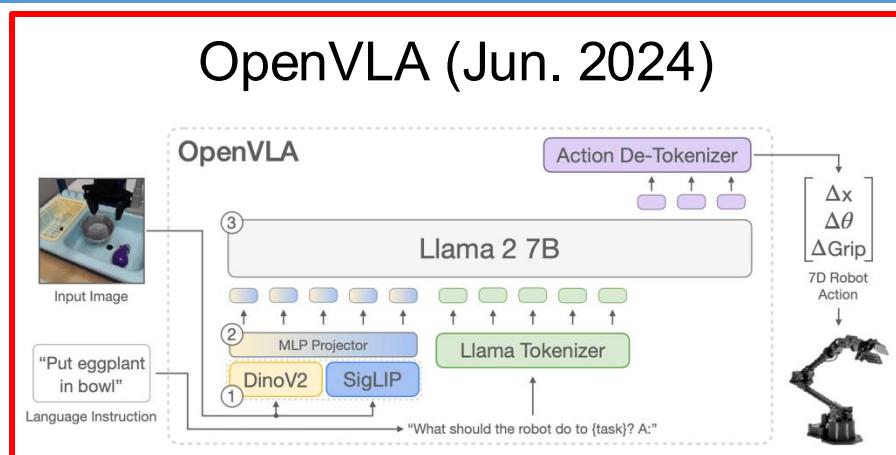
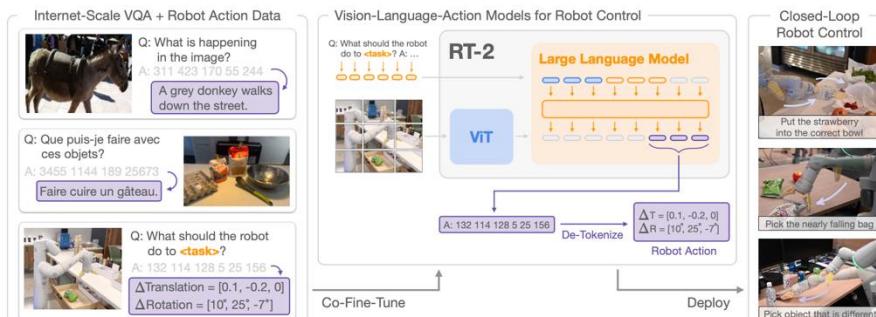
RT-1 (Dec. 2022)



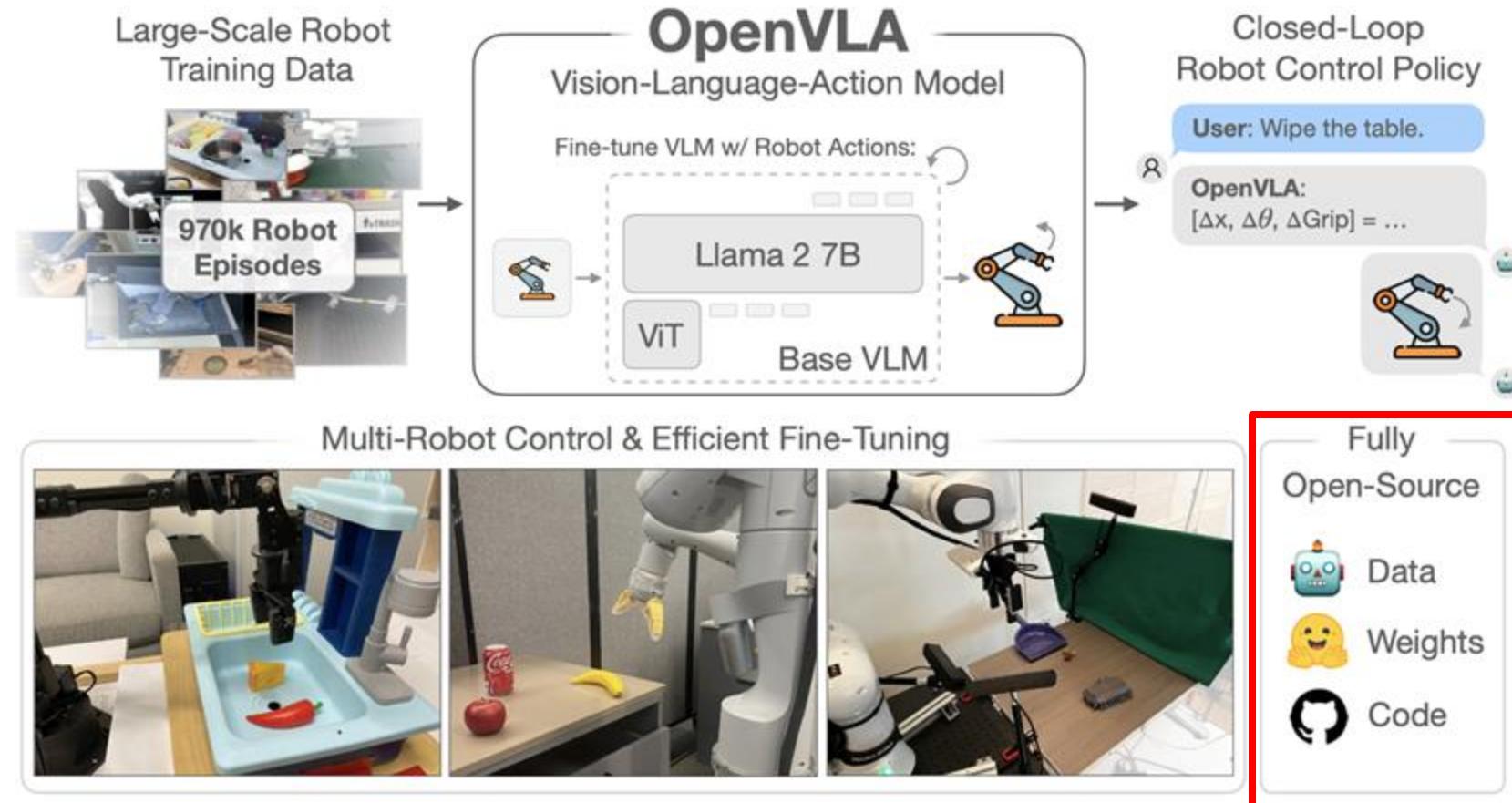
RT-X (Oct. 2023)

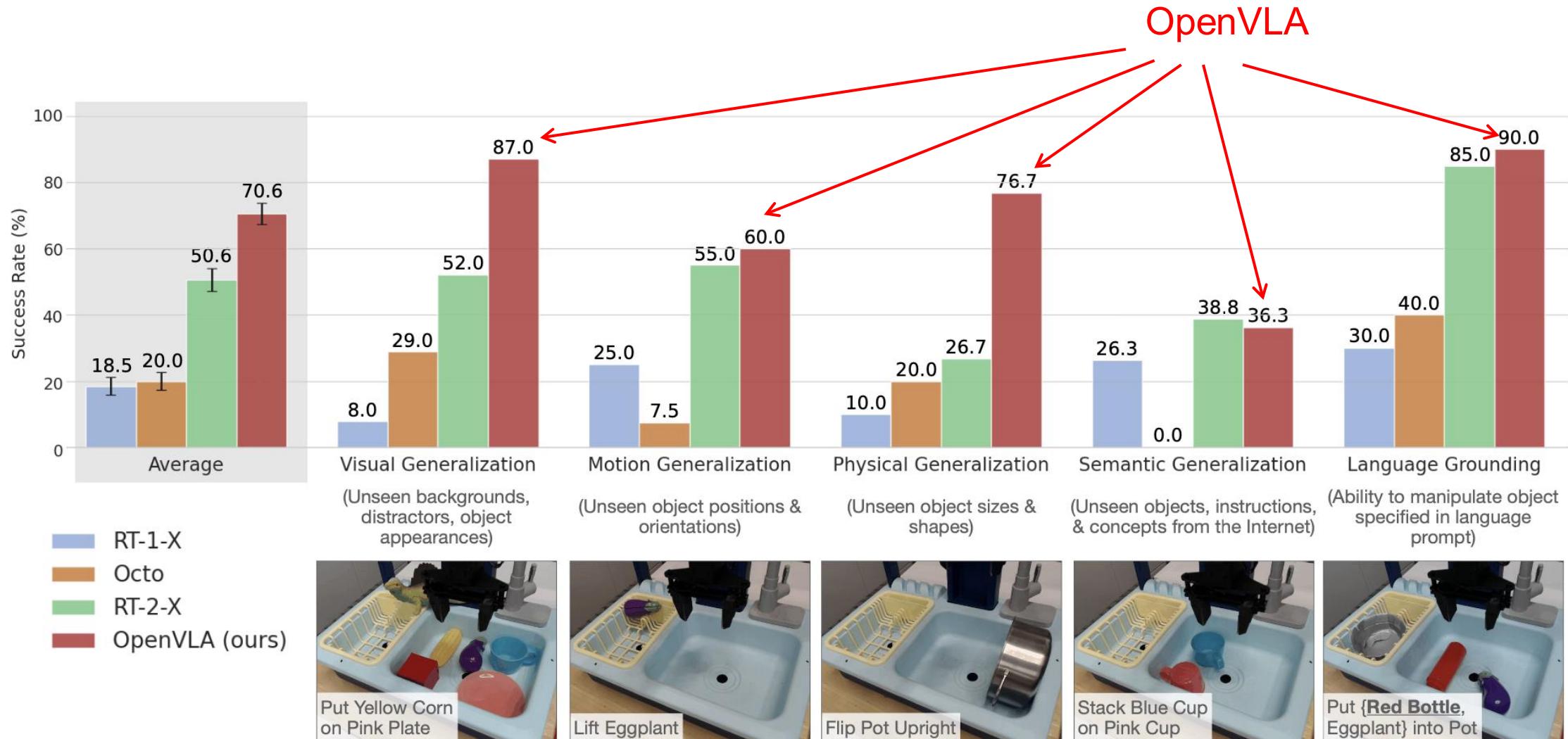


Pi-Zero (Oct. 2024)

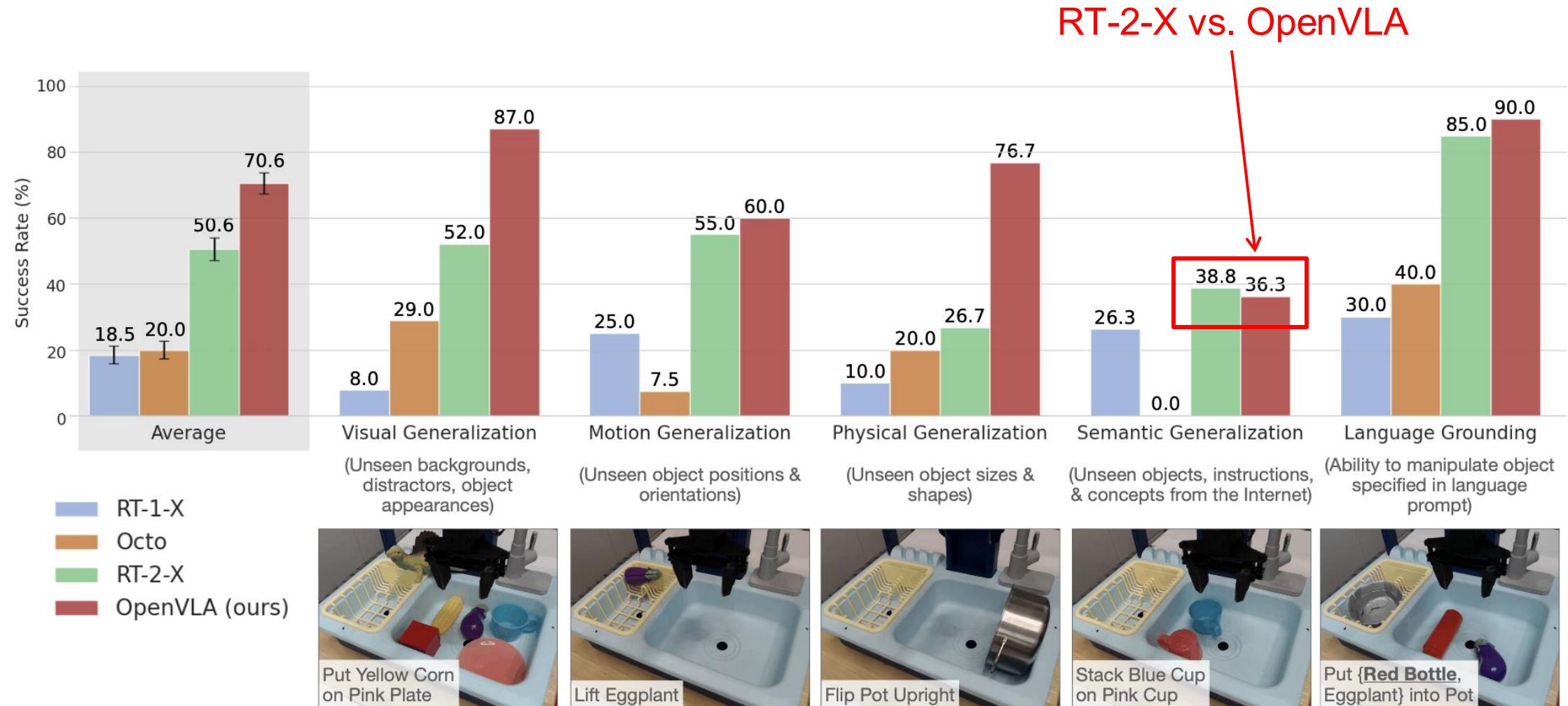


- First released in June 2024
- RT-2 / RT-2-X (55B params) were closed-source
- OpenVLA (7B params)

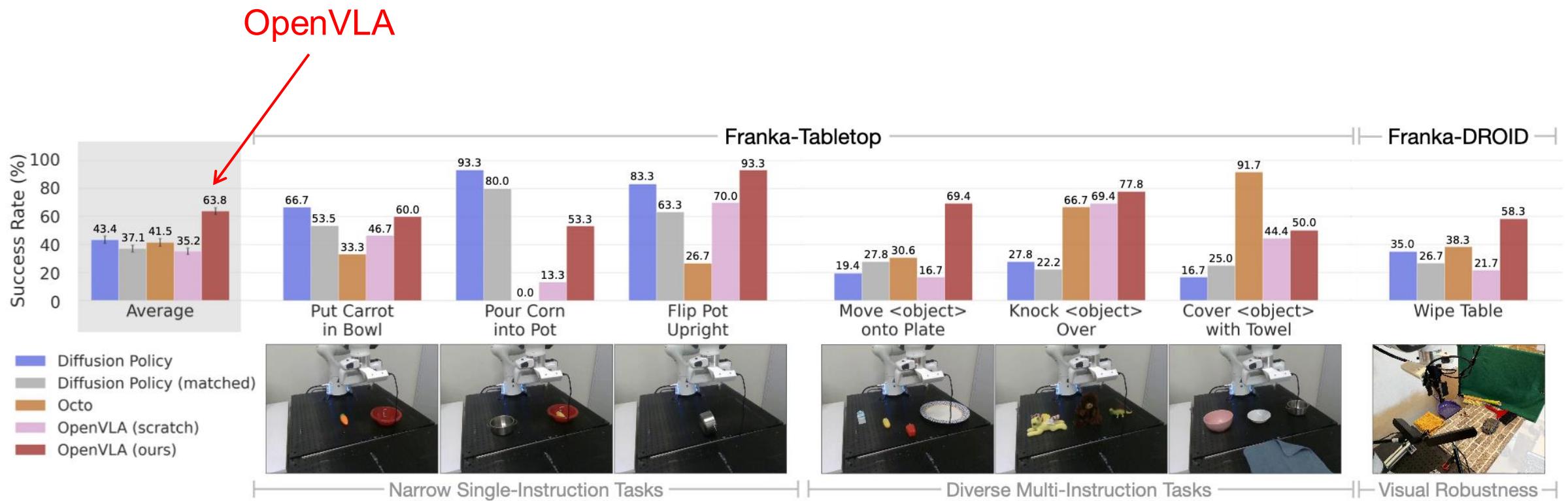


Question #1: How well does it work out of the box?

□ Question #1: How well does it work out of the box?

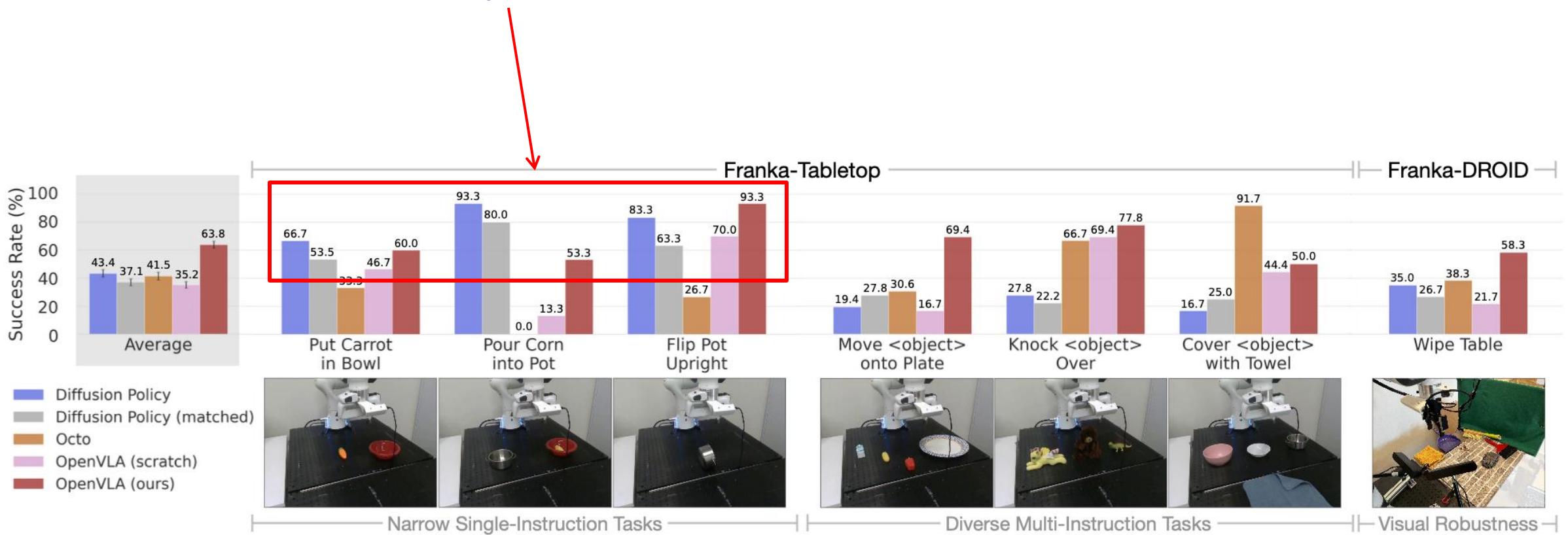


□ Question #2: Fine-tuning to adapt to new robot setups



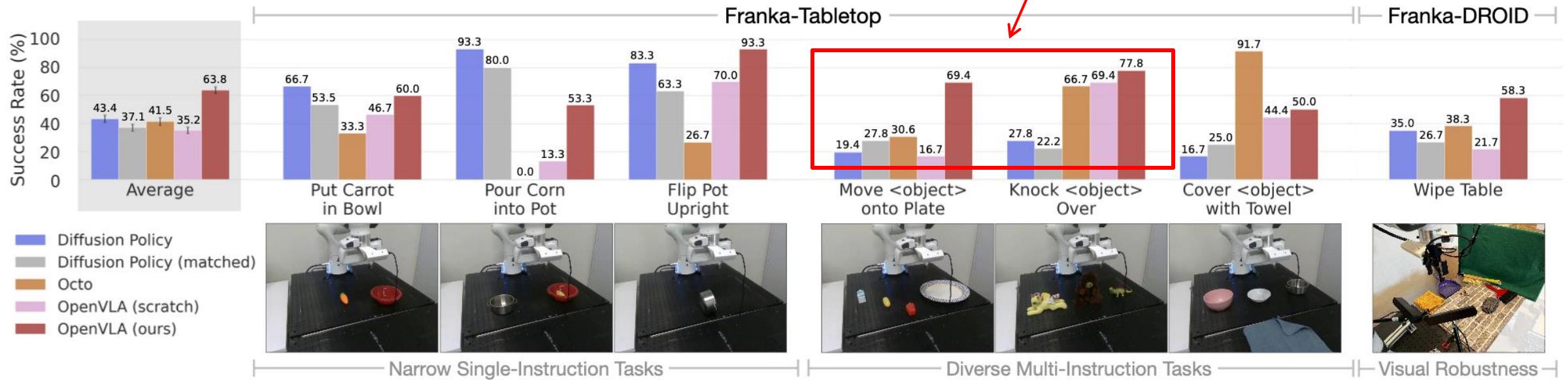
□ Question #2: Fine-tuning to adapt to new robot setups

Diffusion Policy vs. OpenVLA



Question #2: Fine-tuning to adapt to new robot setups

Diffusion Policy vs. OpenVLA



❑ Fully open sourced

openvla / openvla

Issues 18 Pull requests 6 Actions Projects Security Insights

 **openvla** Public
forked from [TRI-ML/prismatic-vlms](#)

 main  2 Branches  0 Tags Add file 

This branch is [46 commits ahead of TRI-ML/prismatic-vlms:main](#). #212

File	Description	Time
moojink Update README: "50 episodes" per task in LIBERO	1b024f2 · 2 months ago	 61 Commits
experiments/robot	Pin robosuite==1.4.1 in libero_requirements.txt	2 months ago
prismatic	Add check for empty token at end of prompt in openvla.p...	5 months ago
scripts	OpenVLA Release	8 months ago
vla-scripts	Update default LR (set to 5e-4)	4 months ago
.gitignore	Add BridgeData V2 eval script and instructions	6 months ago
.pre-commit-config.yaml	Lint, add 224px optimized Prism models	10 months ago
LICENSE	OpenVLA Release	8 months ago
Makefile	Initial commit	last year
README.md	Update README: "50 episodes" per task in LIBERO	2 months ago
pyproject.toml	Pin torchvision, torchaudio versions in pyproject.toml	6 months ago

 **About**
OpenVLA: An open-source vision-language-action model for robotic manipulation.

 Readme
 MIT license
 Activity
 Custom properties
 2k stars
 21 watching
 265 forks
[Report repository](#)

 **Releases**
No releases published

 **Packages**
No packages published

 **Hugging Face**

 **OpenVLA Collaboration** University
<https://openvla.github.io/>

 **AI & ML interests**
Robot Learning

 **Recent Activity**
 KarlP authored a paper about 1 month ago [FAST: Efficient Action Tokenization for Vision-Language-Actio...](#)
 moojink updated a model 5 months ago [openvla/openvla-7b-finetuned-libero-10](#)
 moojink updated a model 5 months ago [openvla/openvla-7b-finetuned-libero-goal](#)
[View all activity](#)

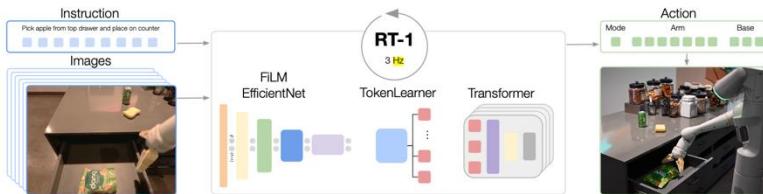
 **Team members** 3


Robotic Foundation Models

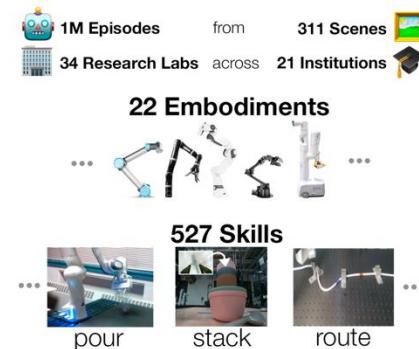


□ What is a Robotic Foundation Model?

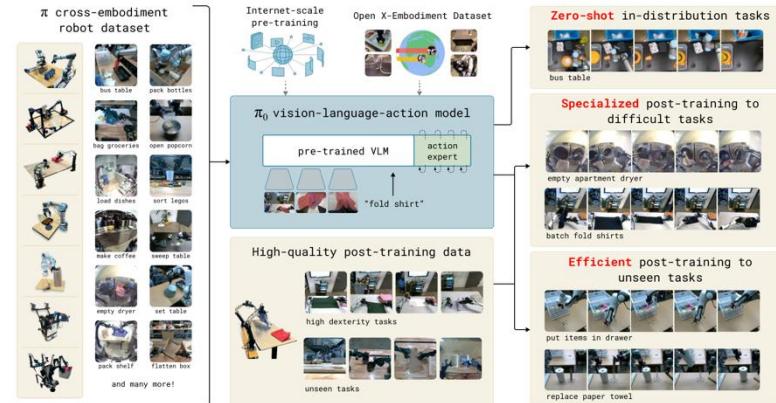
- No explicit representation of states / transition functions
- A policy that maps (observation/state, goal) to action



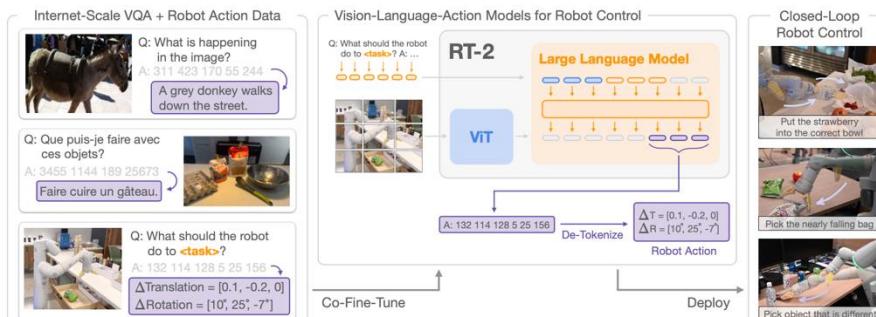
RT-1 (Dec. 2022)



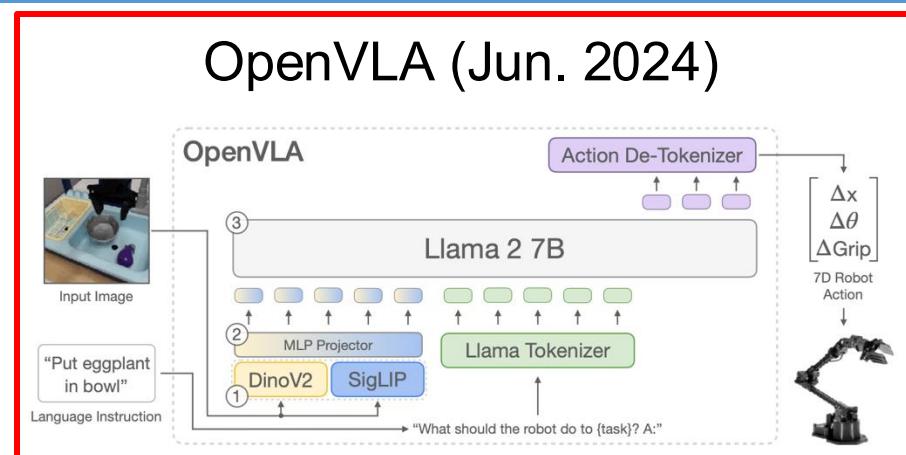
RT-X (Oct. 2023)



Pi-Zero (Oct. 2024)



RT-2 (Jul. 2023)

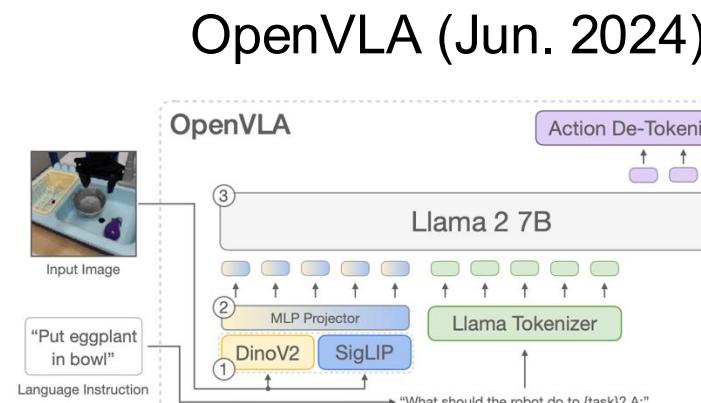
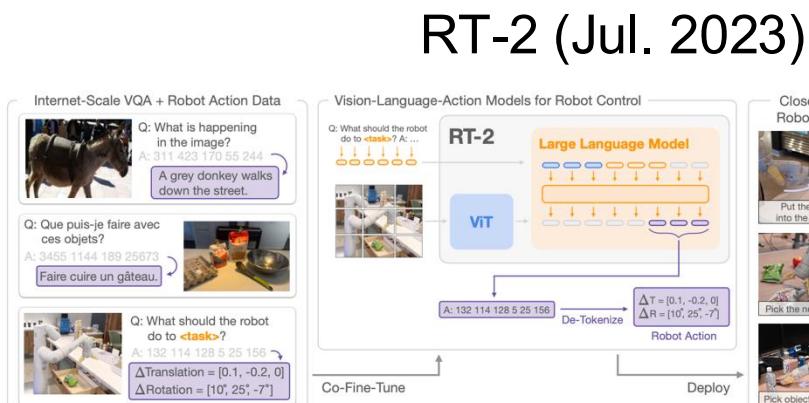
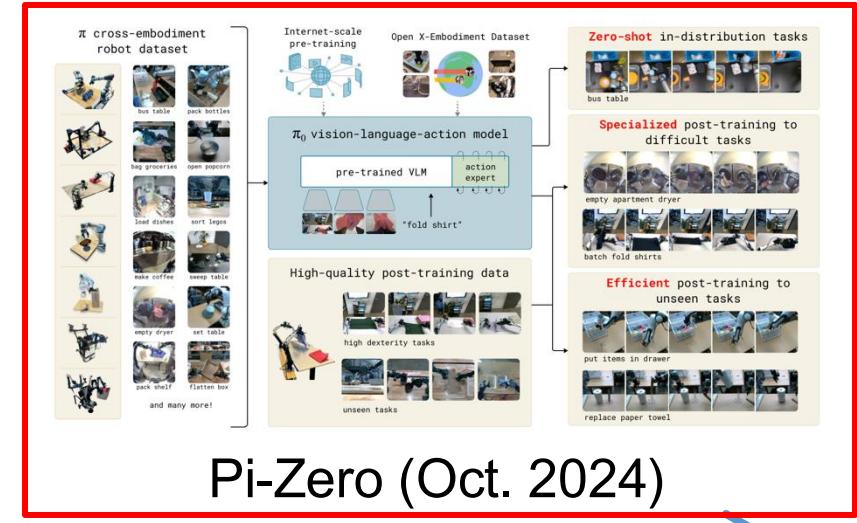
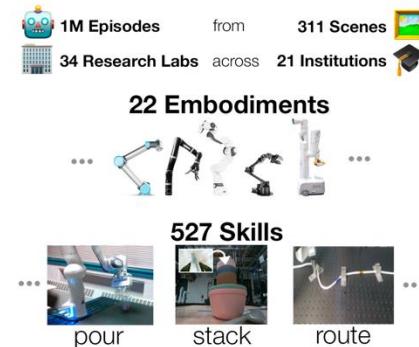
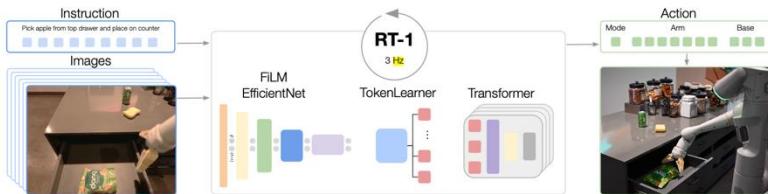


Robotic Foundation Models



□ What is a Robotic Foundation Model?

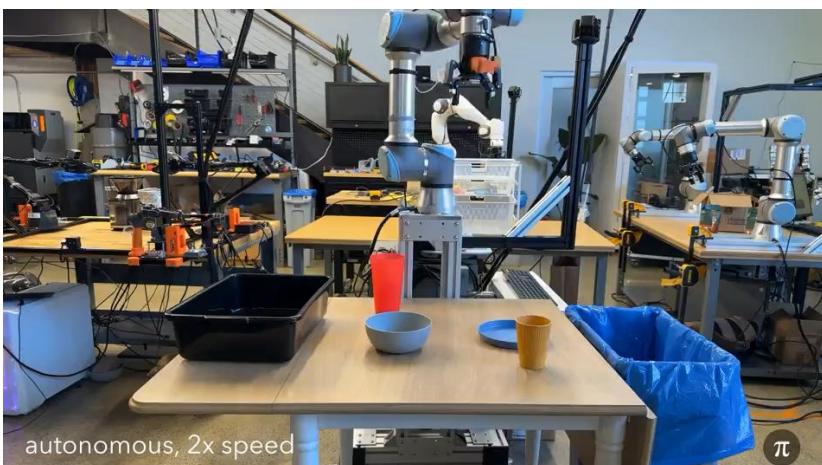
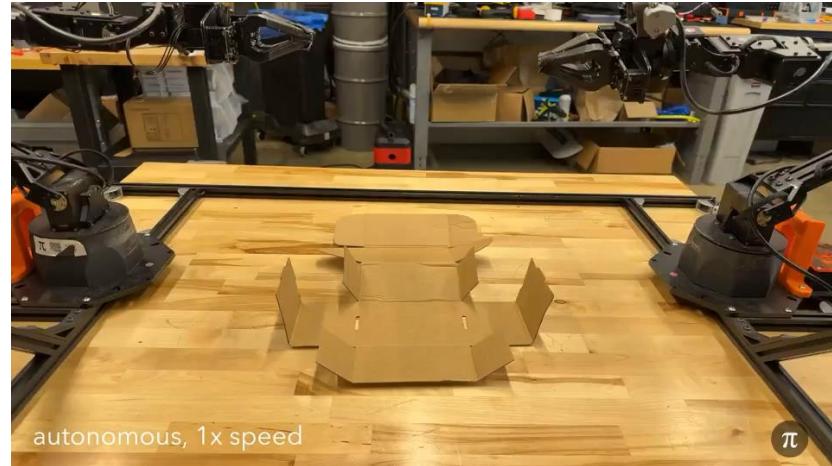
- No explicit representation of states / transition functions
- A policy that maps (observation/state, goal) to action



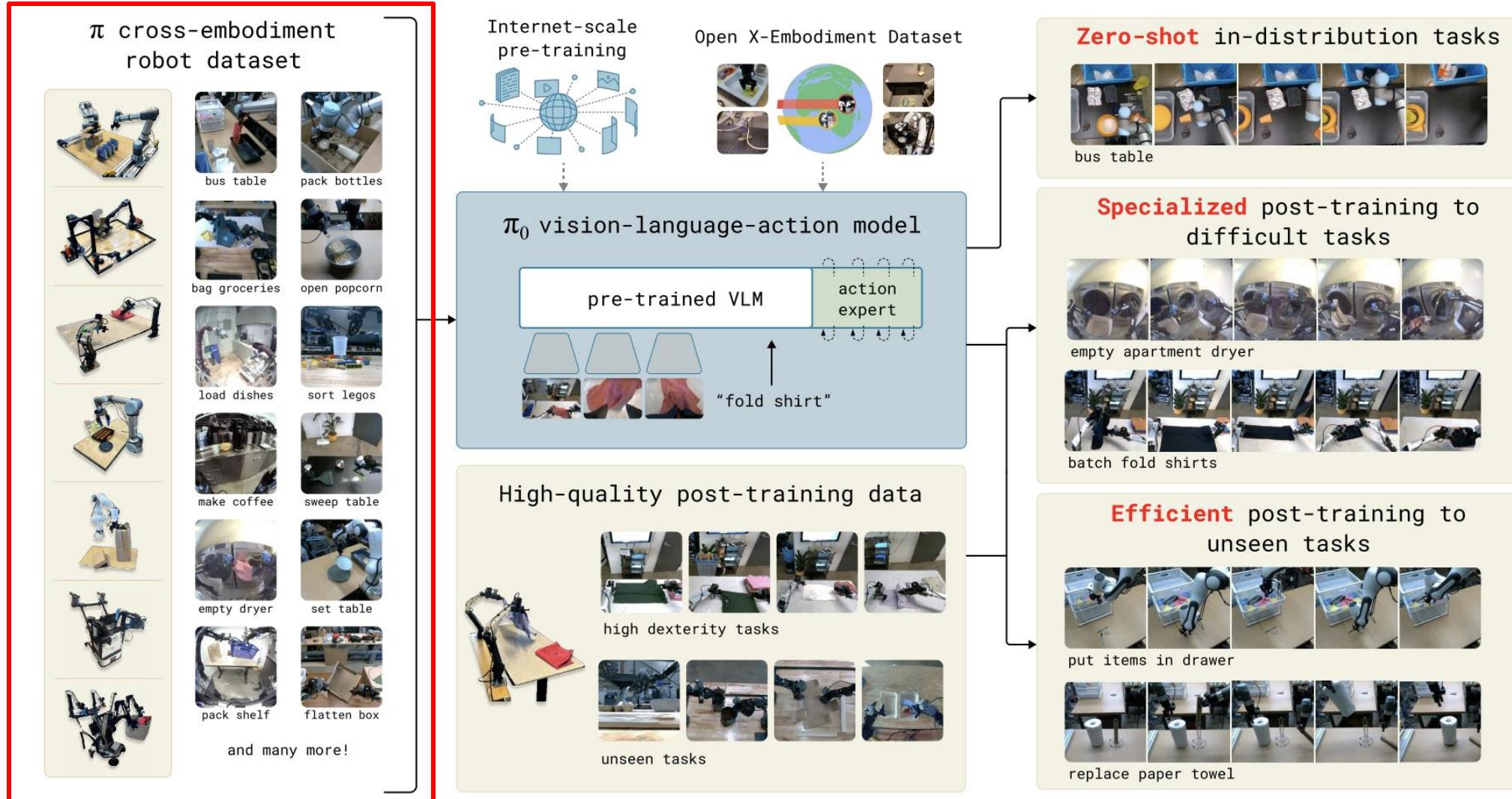
Pi-Zero by Physical Intelligence



- First released in October 2024



Pi-Zero by Physical Intelligence



Cross-Embodiment Dataset

Pi-Zero by Physical Intelligence



Pre-Training

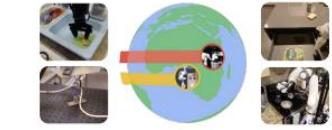
π cross-embodiment robot dataset



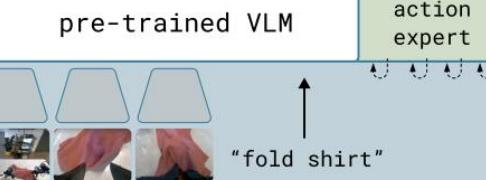
Internet-scale pre-training



Open X-Embodiment Dataset



π_0 vision-language-action model



High-quality post-training data



high dexterity tasks



unseen tasks

Zero-shot in-distribution tasks



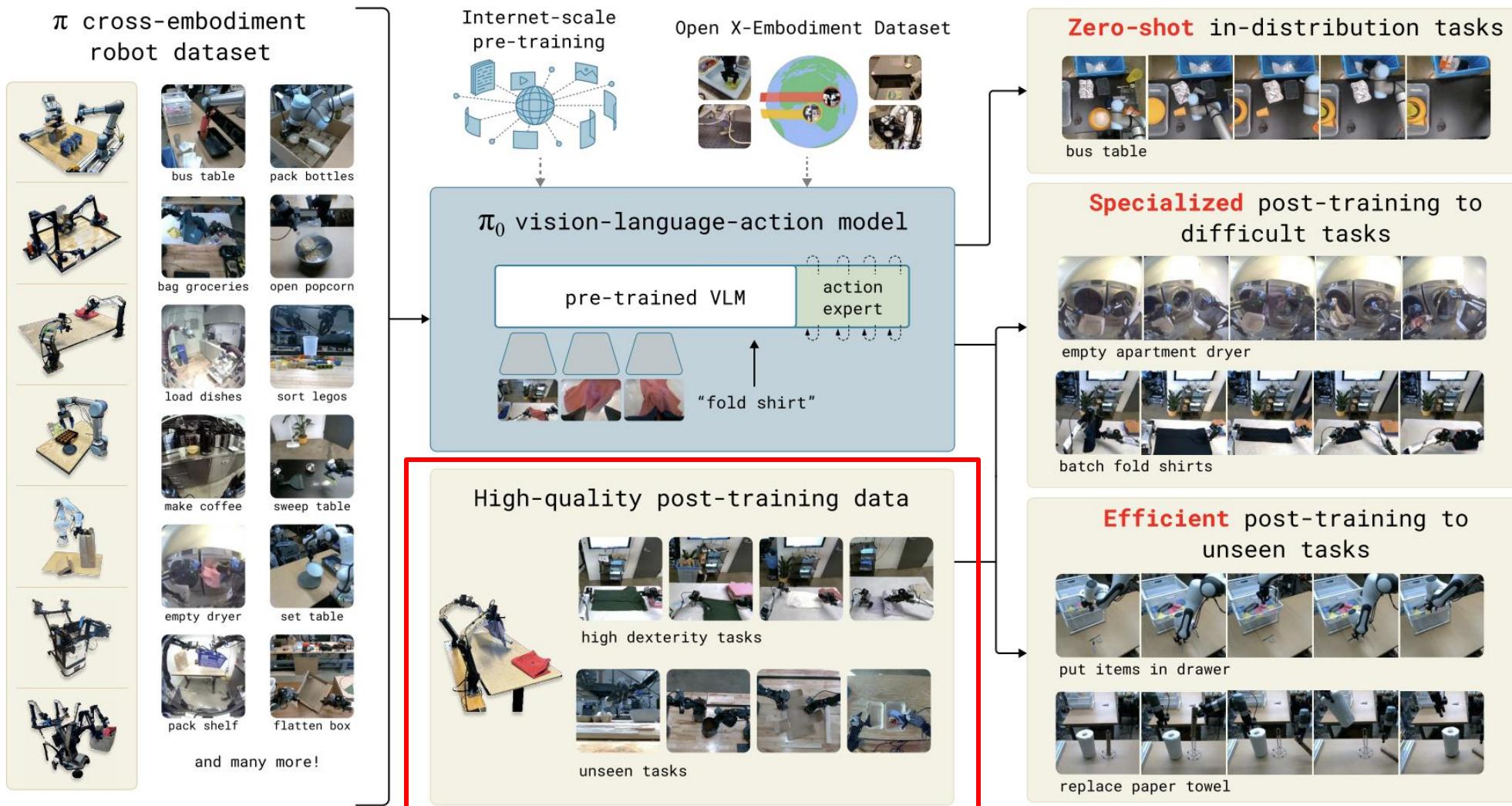
Specialized post-training to difficult tasks



Efficient post-training to unseen tasks



Pi-Zero by Physical Intelligence



Post-Training

Pi-Zero by Physical Intelligence



Simple in-distribution tasks



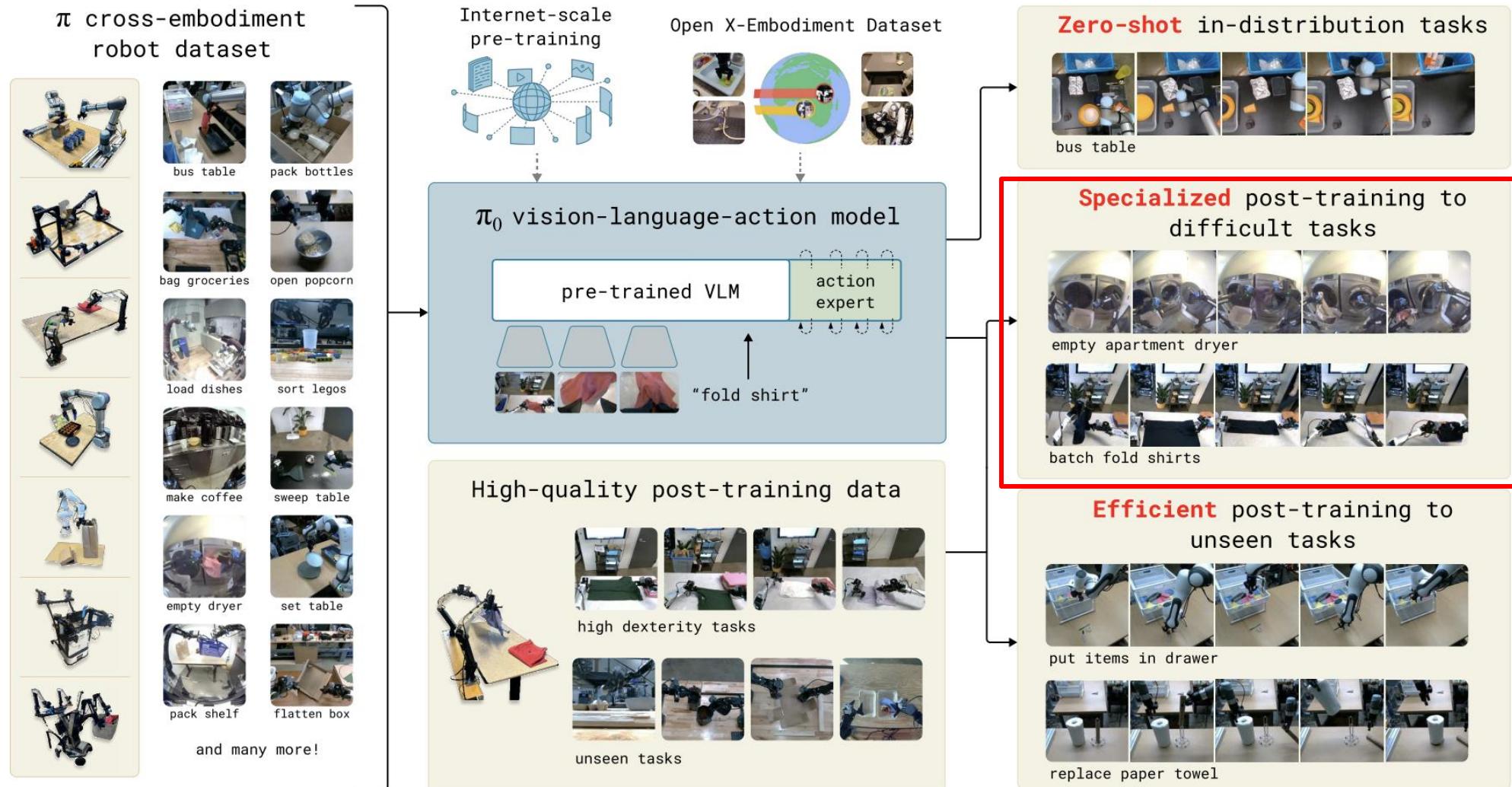
Specialized post-training to difficult tasks



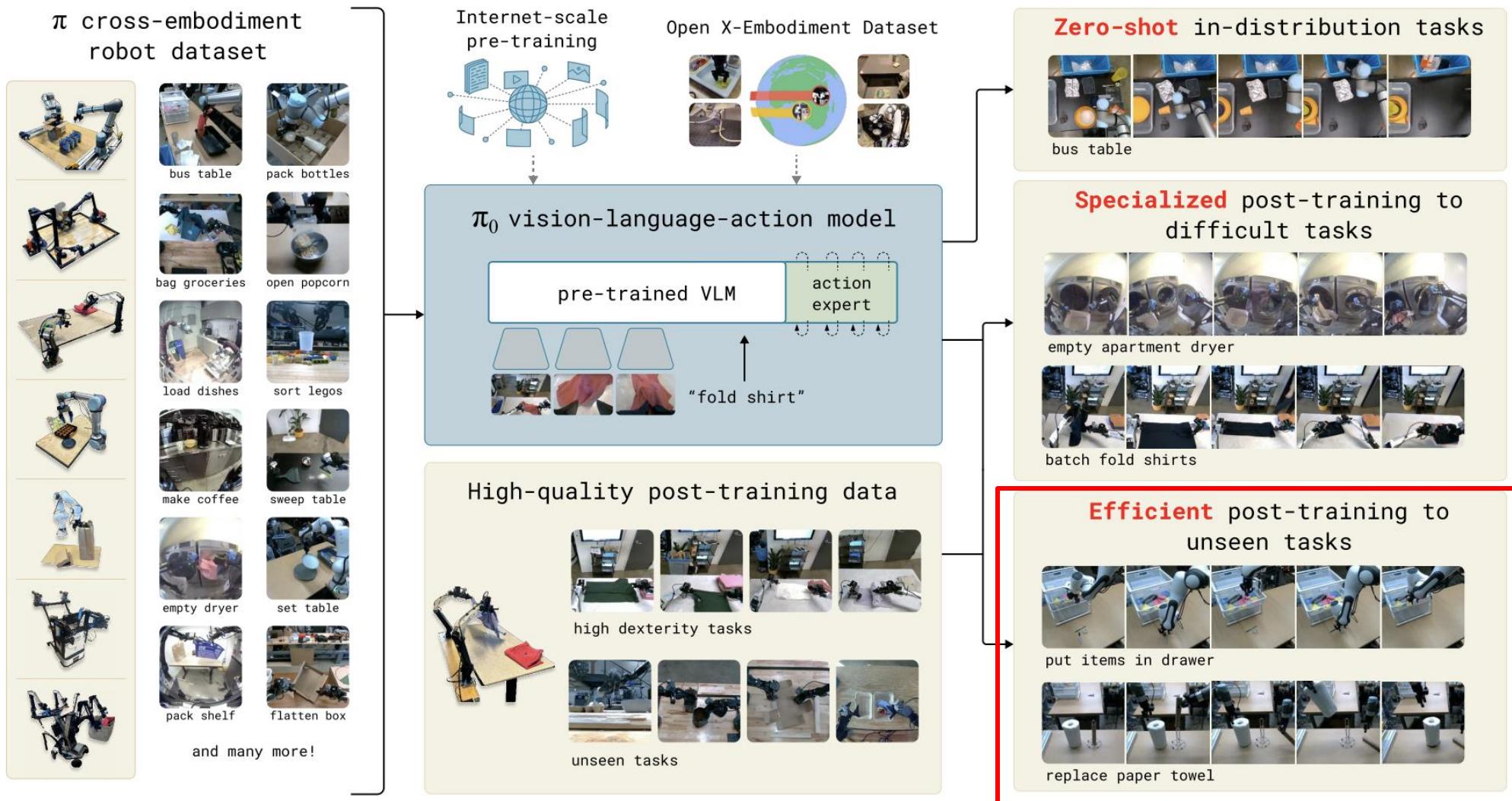
Efficient post-training to unseen tasks



Pi-Zero by Physical Intelligence



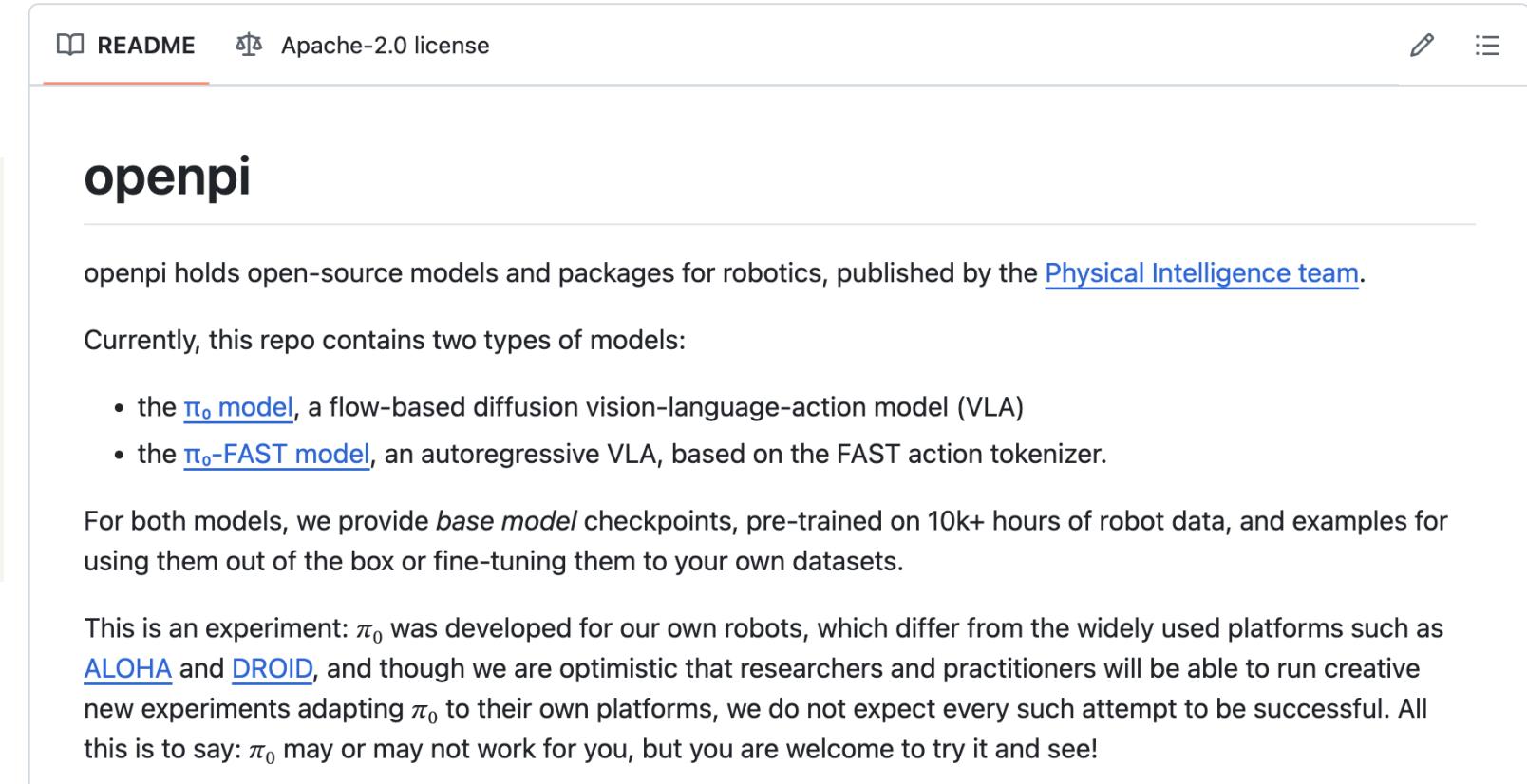
Pi-Zero by Physical Intelligence



Physical Intelligence (π)

Open Sourcing π_0

Published February 4, 2025
Email research@physicalintelligence.company
Repo [Physical-Intelligence/openpi](#)



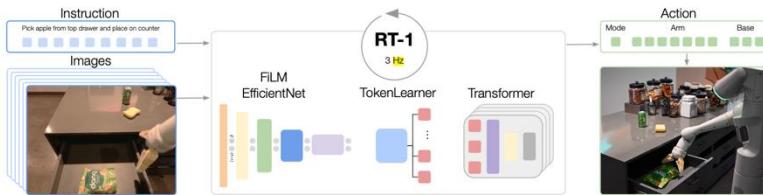
The screenshot shows a GitHub repository page for 'openpi'. At the top, there are links for 'README' and 'Apache-2.0 license'. Below the title 'openpi', it says: 'openpi holds open-source models and packages for robotics, published by the [Physical Intelligence team](#)'. It then states: 'Currently, this repo contains two types of models:' followed by a bulleted list: '• the [\$\pi_0\$ model](#), a flow-based diffusion vision-language-action model (VLA)' and '• the [\$\pi_0\$ -FAST model](#), an autoregressive VLA, based on the FAST action tokenizer.' Further down, it says: 'For both models, we provide *base model* checkpoints, pre-trained on 10k+ hours of robot data, and examples for using them out of the box or fine-tuning them to your own datasets.' At the bottom, it concludes: 'This is an experiment: π_0 was developed for our own robots, which differ from the widely used platforms such as [ALOHA](#) and [DROID](#), and though we are optimistic that researchers and practitioners will be able to run creative new experiments adapting π_0 to their own platforms, we do not expect every such attempt to be successful. All this is to say: π_0 may or may not work for you, but you are welcome to try it and see!'

Robotic Foundation Models

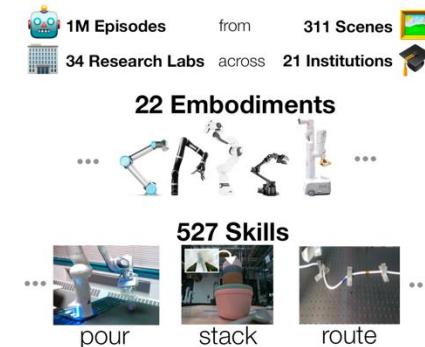


□ What is a Robotic Foundation Model?

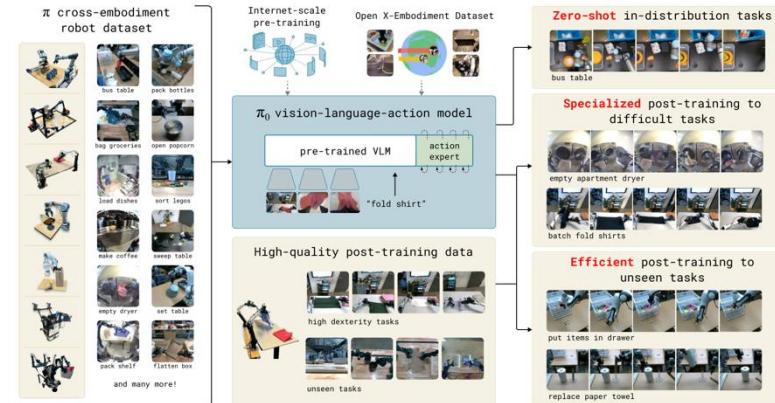
- No explicit representation of states / transition functions
- A policy that maps (observation/state, goal) to action



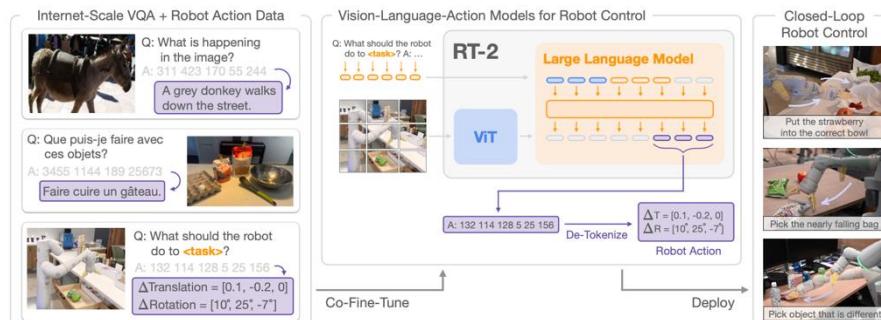
RT-1 (Dec. 2022)



RT-X (Oct. 2023)

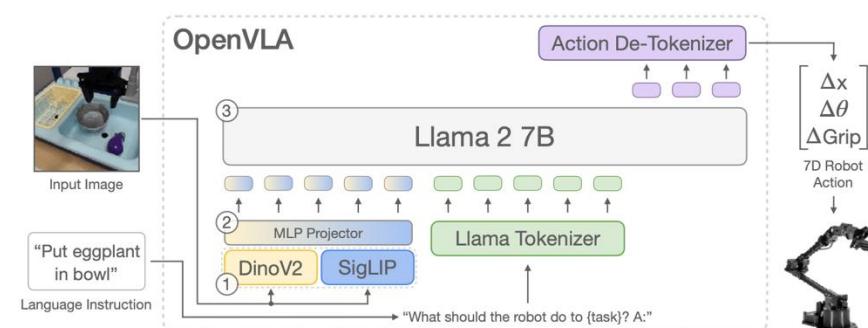


Pi-Zero (Oct. 2024)



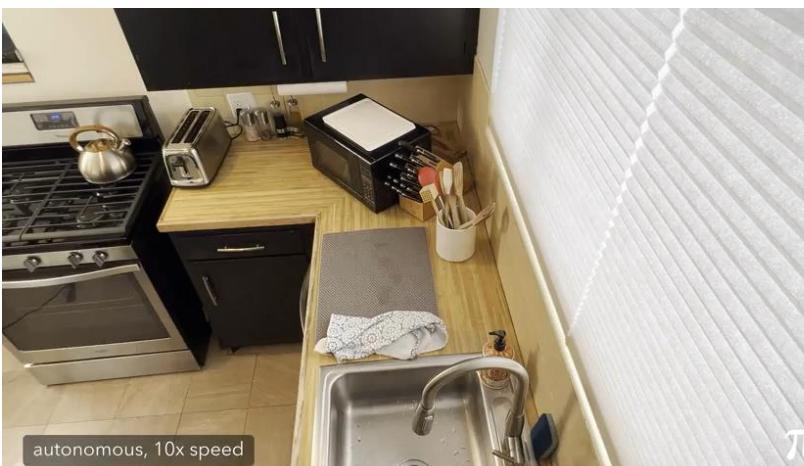
RT-2 (Jul. 2023)

OpenVLA (Jun. 2024)

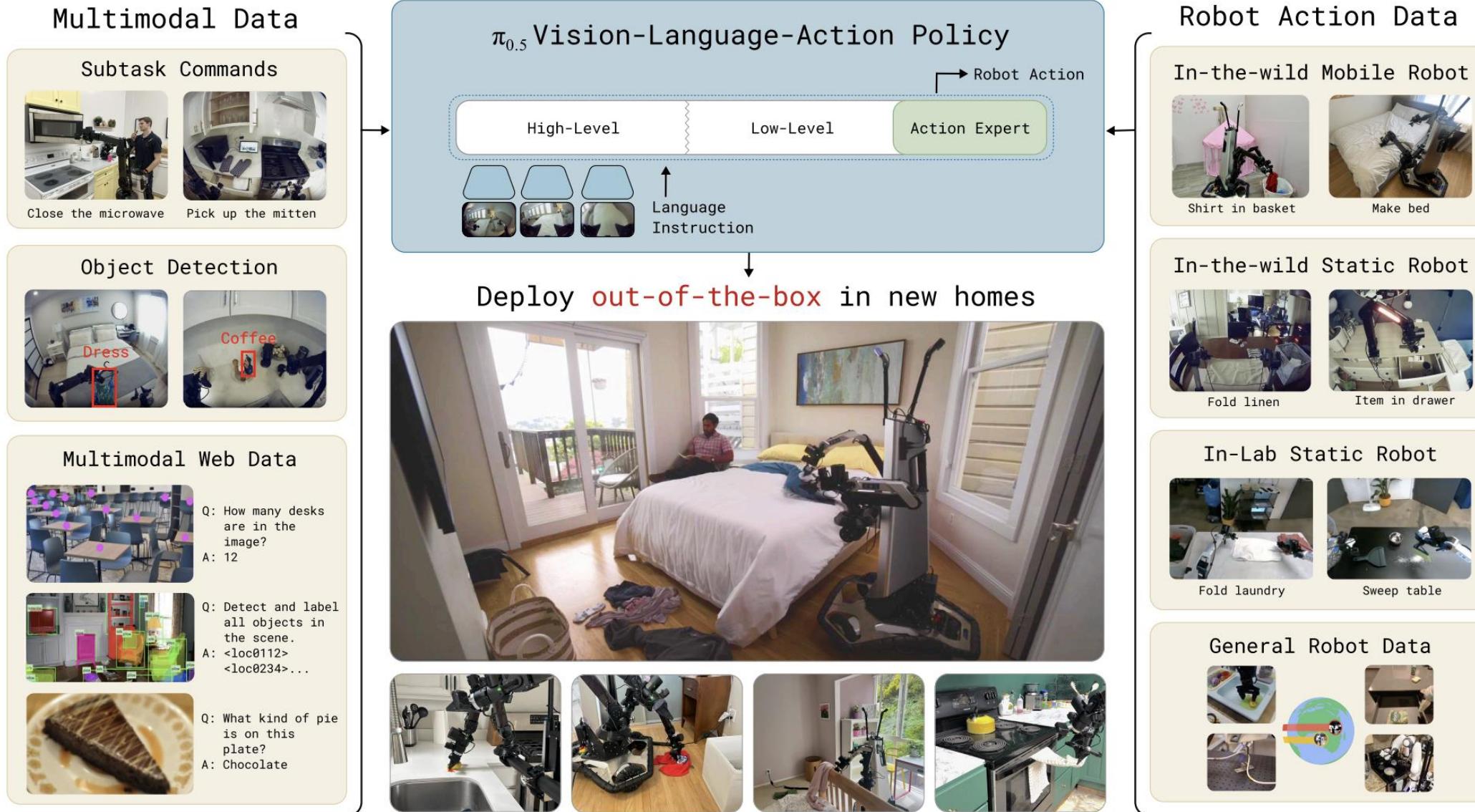
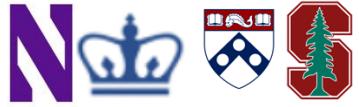


Helix (Figure)
Hi-Robot (PI)
Gemini Robotics
Pi-0.5 (PI)
GR00T (Nvidia)
DYNA-1
LBM (TRI)

- ❑ Technical report released in April 2025
- ❑ Code released in September 2025



π_{0.5}: a Vision-Language-Action Model with Open-World Generalization



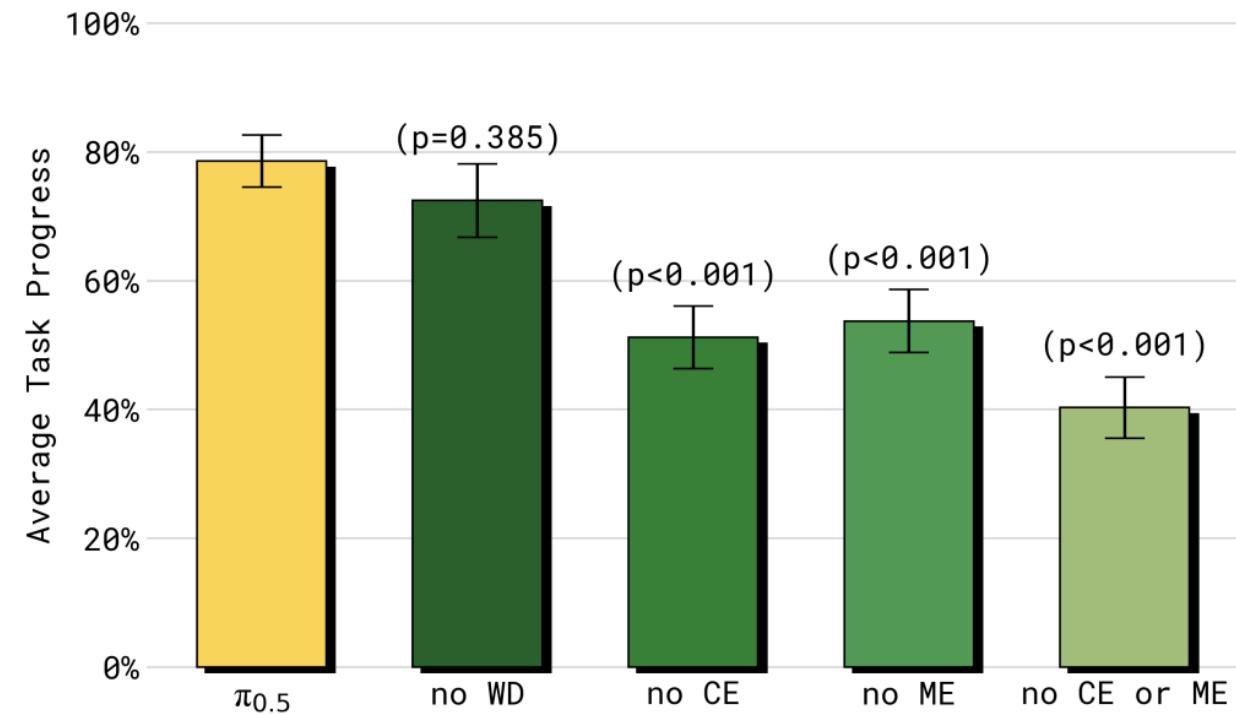
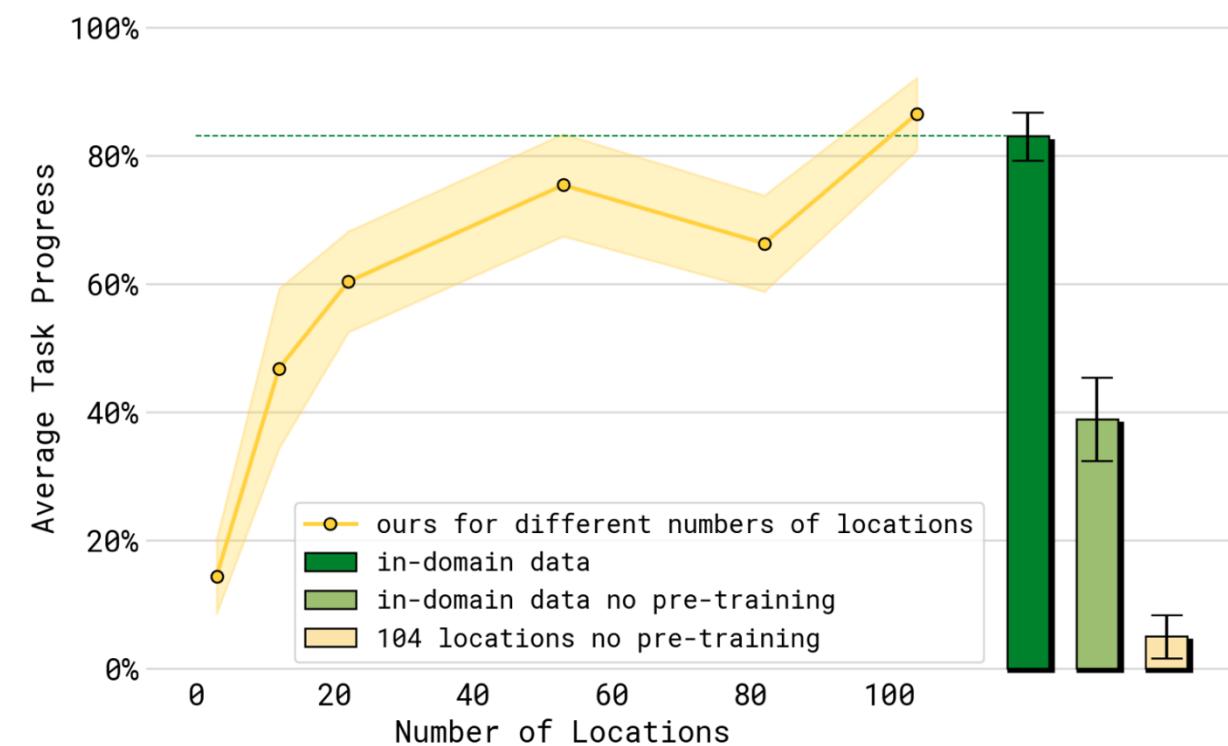
Pre-training

<p>Laboratory cross-embodiment</p>  <p>Sort drawer</p>  <p>Pack bottles</p>  <p>Sweep table</p>  <p>Fold laundry</p>  <p>Open X-Embodiment</p>	<p>Diverse mobile manipulator</p>          <p>Shirt in basket Spatula in holder Wipe plate</p> <p>Hang dress Tissue on stand Dish in sink</p> <p>Make bed</p>	<p>High-level subtask</p>  <p>How would you clean the bedroom?</p> <p>Bounding boxes: <loc0405><loc0011><loc0911><loc0197>closet Subtask: move to closet</p>  <p>How would you clean the kitchen?</p> <p>Bounding boxes: <loc0571><loc0376><loc0815><loc0484>mitten <loc0787><loc0346><loc1003><loc0490>drawer Subtask: move left arm forward and pick up mitten</p>	<p>Verbal instruction</p>  <p>Put cup in sink</p>  <p>Place pillow on bed</p>  <p>Policy: put plate in sink Relabeled: put plate on rack</p>
<p>Diverse non-mobile manipulator</p>       <p>Item in drawer Fold linen Tidy table</p> <p>Cabinet putaway Kettle on base Towel on oven handle</p>	<p>Multi-modal web data</p>  <p>Describe this region: <loc0470><loc0390><loc0605><loc0484> Front legs of elephant</p>  <p>What kind of pie is this? This is a delicious-looking pecan pie. The image shows a classic pecan pie with its characteristic dark brown filling studded with pecans.</p>		

Post-training

□ Take-home messages:

- Performance improves with more training environments.
- Training recipe: web data (WD), cross embodiment (CE), multiple environments (ME)



Chris Paxton reposted



Here's an unedited 10-minute demo of pi-0.5 straight out of the box. We gave it just one instruction, "clean the table by putting items into the basket", and let it run. It worked!

No fine-tuning required: it handled multiple tabletop scenes robustly, with smoother and faster motions than pi-0.

Full evaluations of pi0.5 and comparisons with pi-0 are on the way, stay tuned!

Video credit: [@KC_Q1015](#)



I did a bit of light red-teaming on the pi-0.5 model from [@physical_int](#) on our DROID setup at [@Princeton](#).

I was curious: does it inherit biases/unsafe behavior from its co-training with internet data?



Spoiler ⚠: the model thinks I'm a criminal!

"Put the tomato on the criminal"



Evaluating π_0 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

*Corresponding author

Gemini Robotics by Google DeepMind



MODELS

Gemini Robotics 1.5 brings AI agents into the physical world

25 SEPTEMBER 2025

Carolina Parada

Share



賴嘉滿 @ FIRSTLIGHT Capital ✅
@chiamin_lai

Auto-translated from Japanese by Grok Show original ⚙

The day after the Gemini Robotics announcement, I experienced a real-machine demo at CoRL 2025. It understood and acted on a person's instruction: "Put the AirPods with the pink label into the box."

The dexterity of the hand is still lacking, but the comprehension accuracy is surprisingly high.

Rate this translation: ⬆️ ⬇️

A photograph of a woman with dark curly hair, wearing a black top and jeans, standing next to a blue robotic arm on a table. The table has various objects, including a yellow container, a white container, a green cup, and some small colorful blocks. A timer in the bottom left corner of the photo says '0:25'. In the background, other people are visible at the event.