AAAI 2025 Tutorial T04
Time: 2025-02-25 8:30-12:30
Location: Room 118A

Foundation Models Meet Embodied Agents

Manling Li
Northwestern

Yunzhu Li
Columbia

Jiayuan Mao
MIT

Wenlong Huang
Stanford

Northwestern University

COLUMBIA

MIT

Stanford University

AAAI 2025 Tutorial T04
Time: 2025-02-25 8:30-12:30
Location: Room 118A

Robotic Foundation Models

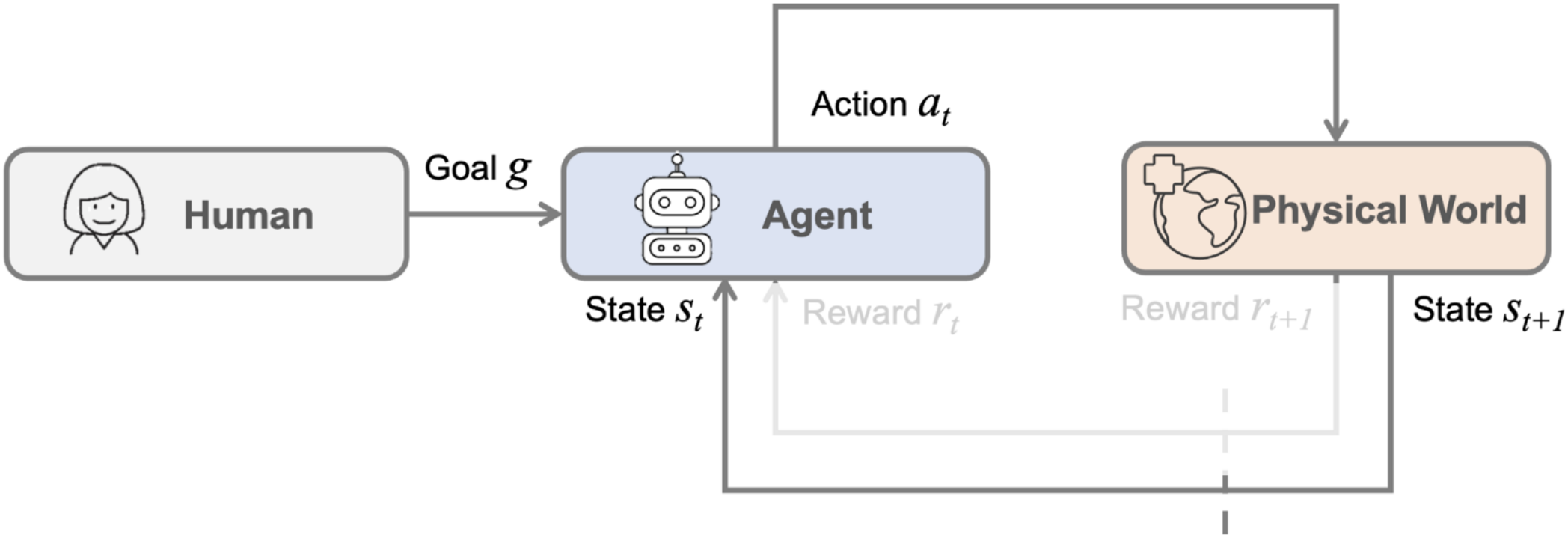
AAAI Tutorial: Foundation Models Meet Embodied Agents

Northwestern University

Stanford University

Robotic Foundation Models

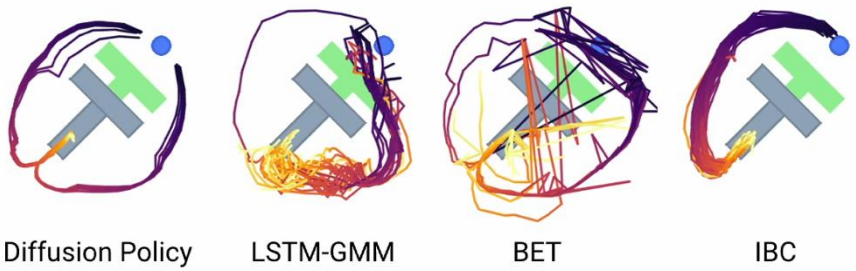
- What is a Robotic Foundation Model?
  - No explicit representation of states / transition functions
  - A policy that maps (observation/state, goal) to action



Robotic Foundation Models

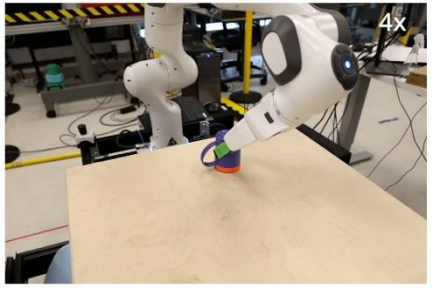
- What is a Robotic Foundation Model?
  - No explicit representation of states / transition functions
  - A policy that maps (observation/state, goal) to action

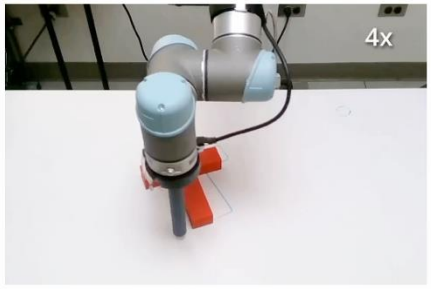
Imitation Learning
(Chi et al., Diffusion Policy)



Diffusion Policy 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.

Diffusion Policy predicts a sequence of action for receding-horizon control.





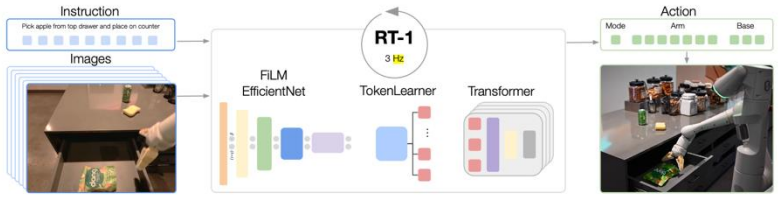
Reinforcement Learning
(OpenAI, Solving Rubik's Cube)

Robotic Foundation Models

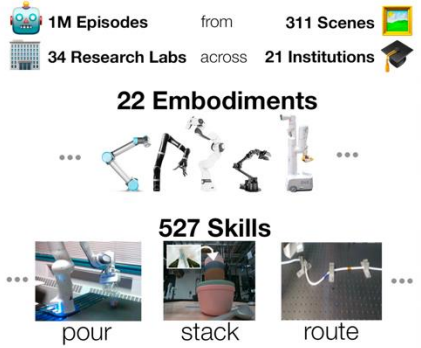
- What is a Robotic Foundation Model?
  - No explicit representation of states / transition functions
  - A policy that maps (observation/state, goal) to action
- Current Foundational Vision-and-Language Models
  - The output may **not** always be **perfect**.
  - It will always generate something **reasonable**.
- 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

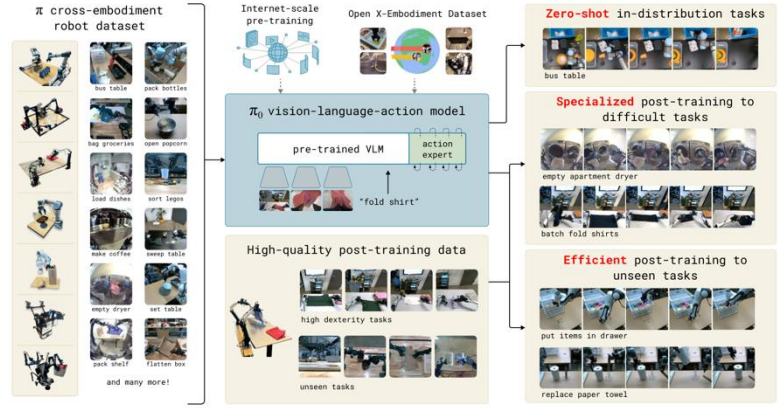
- What is a Robotic Foundation Model?
  - No explicit representation of states / transition functions
  - A policy that maps (observation/state, goal) to action



RT-1 (Dec. 2022)

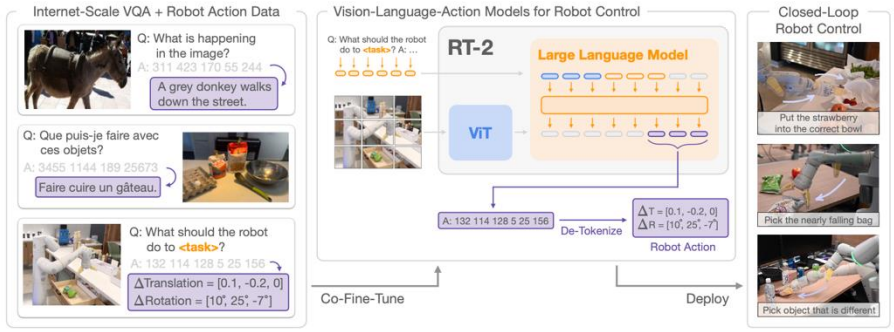


RT-X (Oct. 2023)

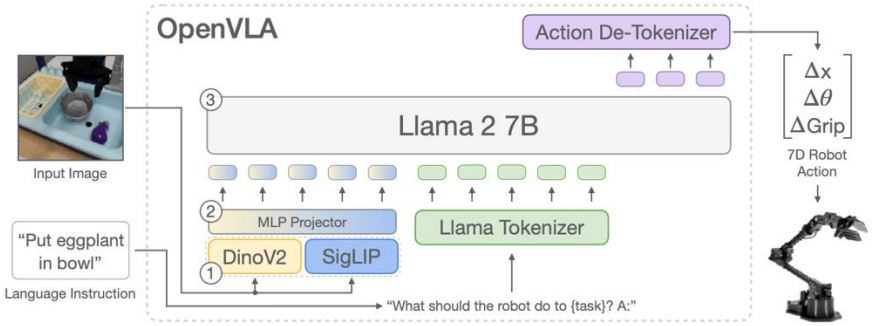


Pi-Zero (Oct. 2024)

RT-2 (Jul. 2023)

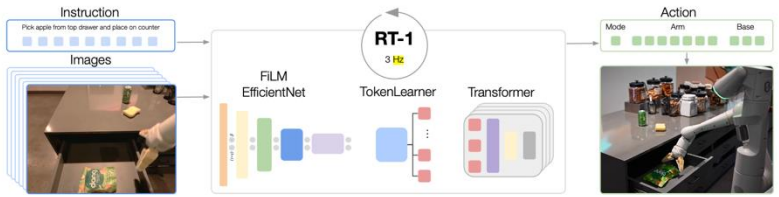


OpenVLA (Jun. 2024)

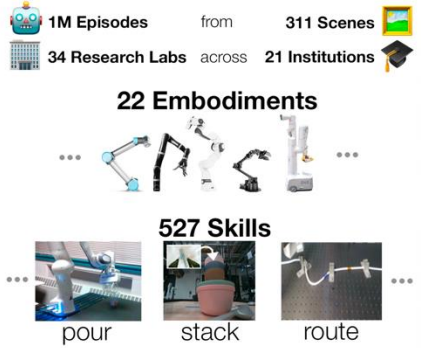


Robotic Foundation Models

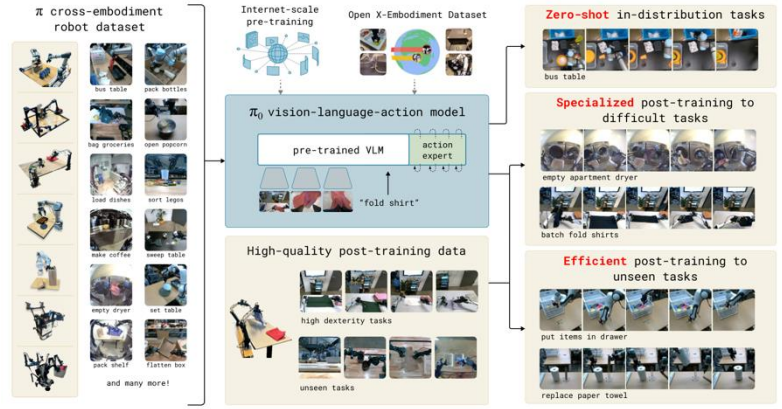
- What is a Robotic Foundation Model?
  - No explicit representation of states / transition functions
  - A policy that maps (observation/state, goal) to action



RT-1 (Dec. 2022)

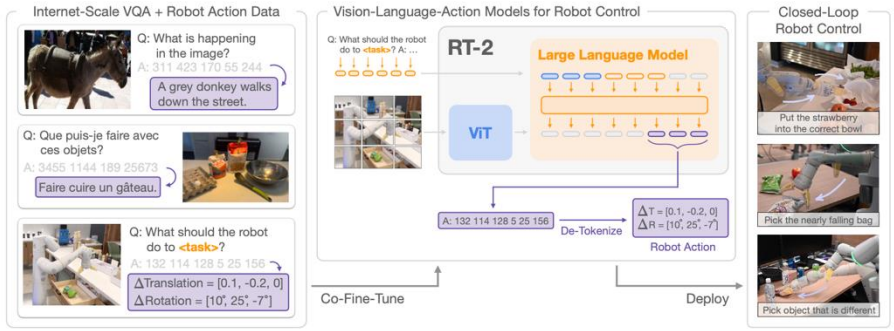


RT-X (Oct. 2023)

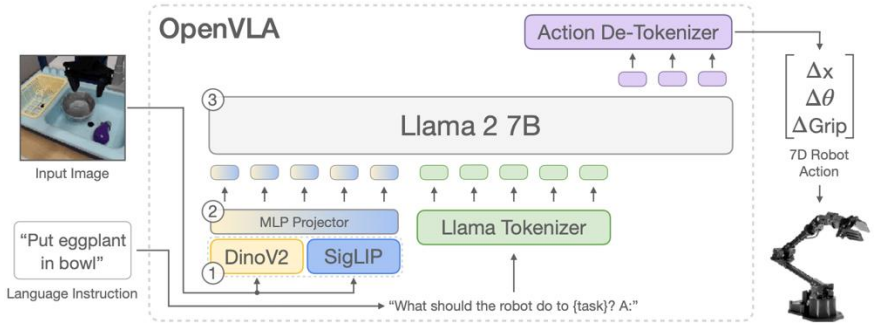


Pi-Zero (Oct. 2024)

RT-2 (Jul. 2023)



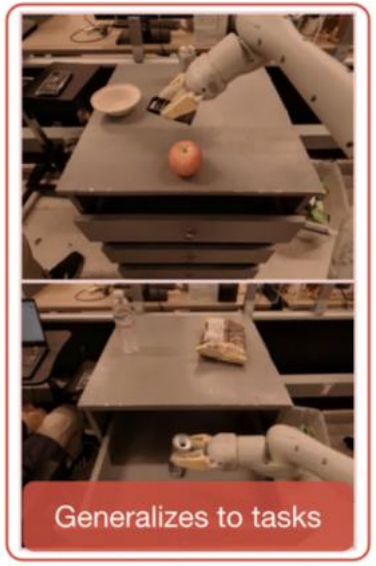
OpenVLA (Jun. 2024)

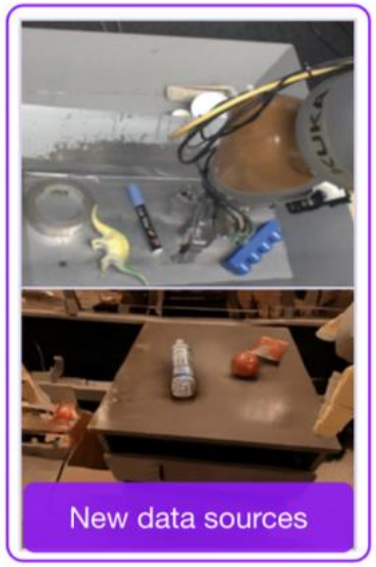


Robotic Transformer 1 (RT-1)

- 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

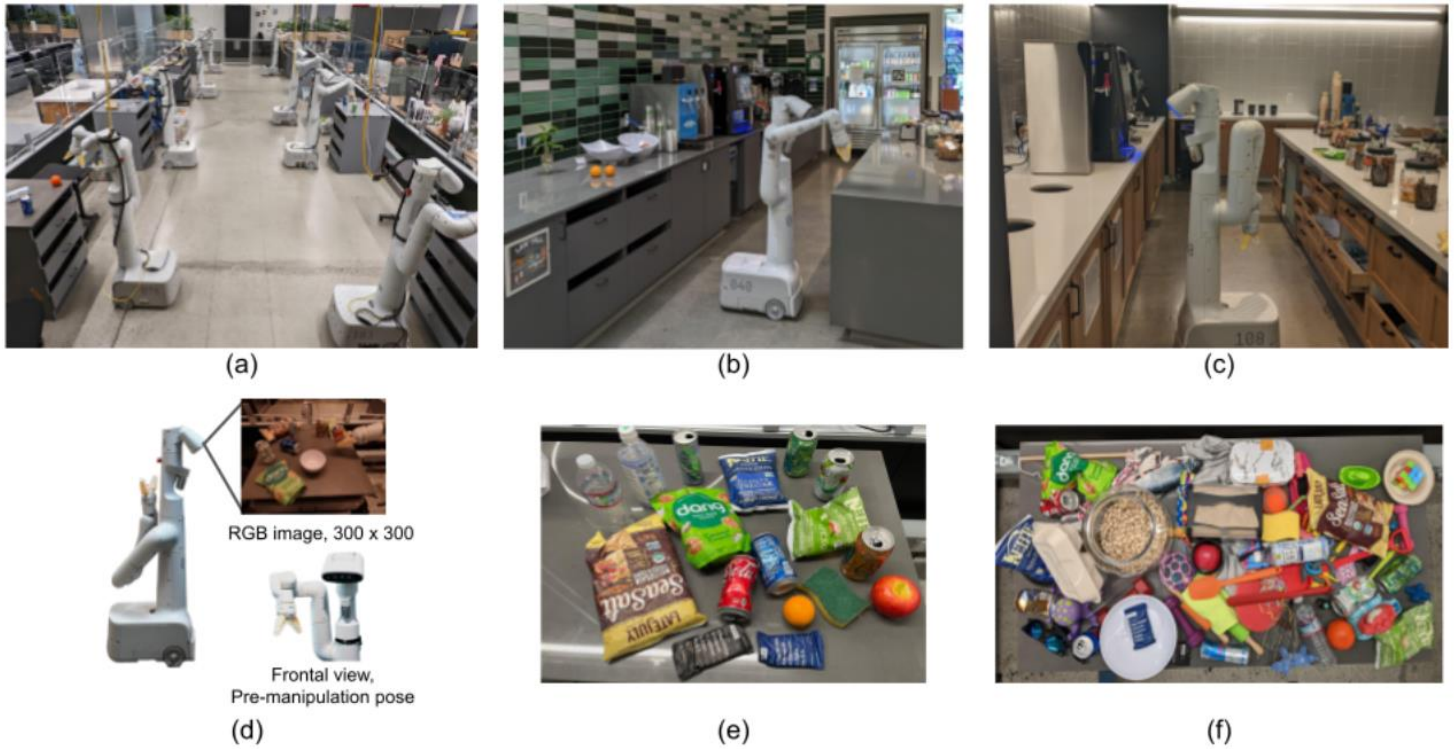
17 months with a fleet of 13 robots, containing ~130k episodes and over 700 tasks





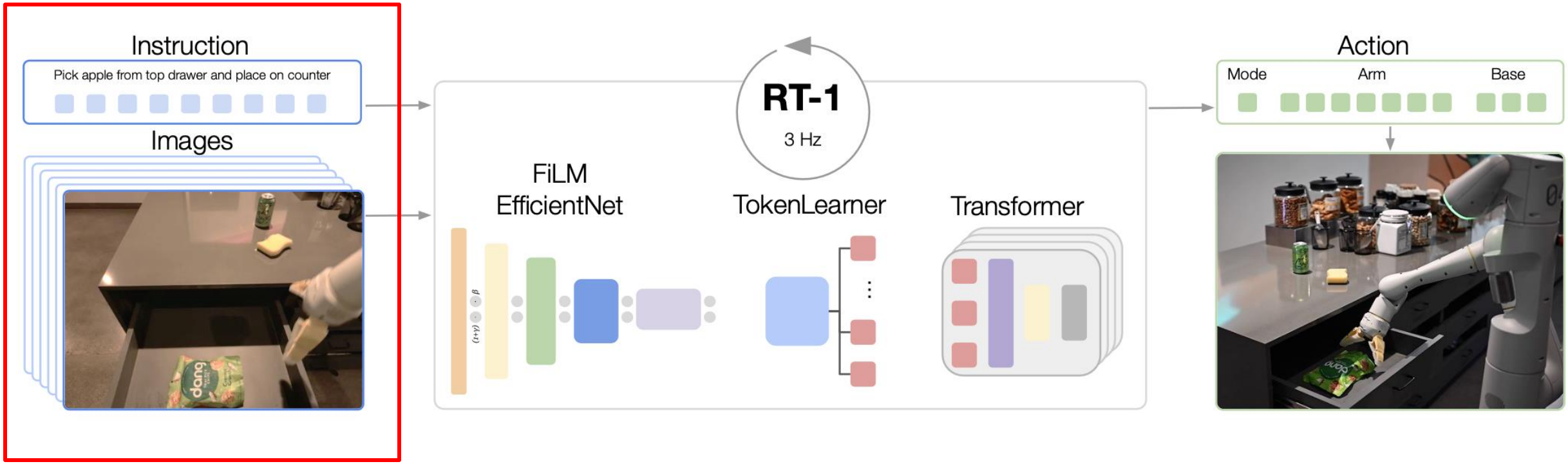
Robotic Transformer 1 (RT-1)

- 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



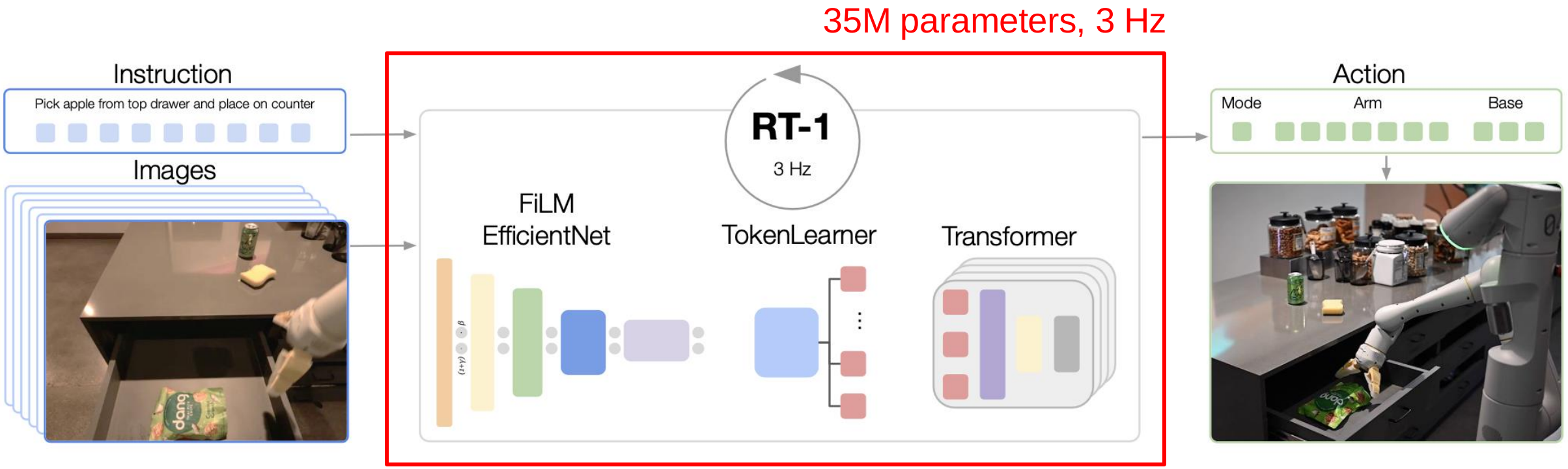
Robotic Transformer 1 (RT-1)

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



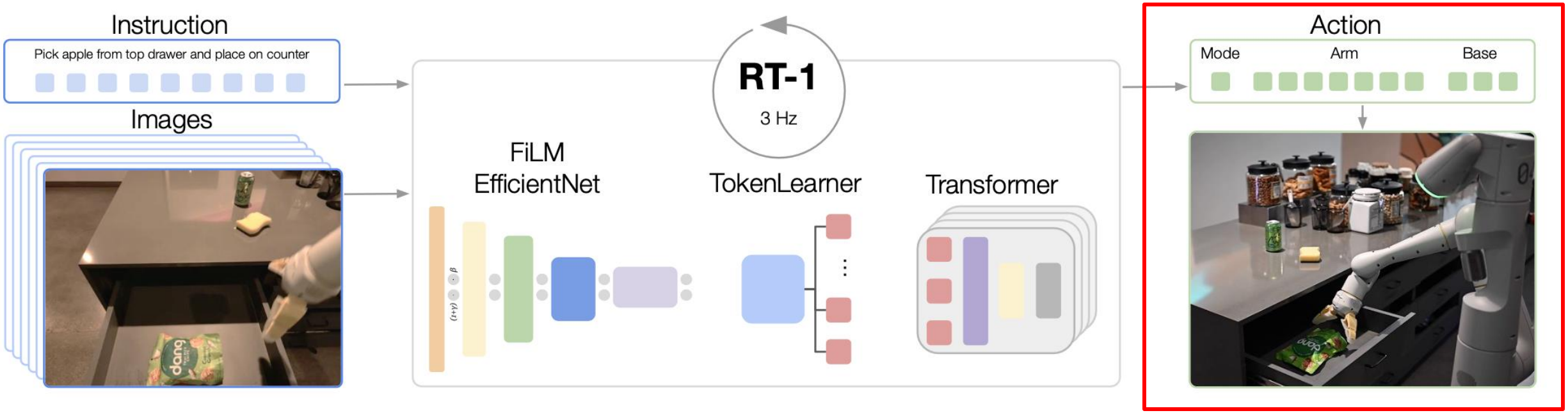
Robotic Transformer 1 (RT-1)

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



Robotic Transformer 1 (RT-1)

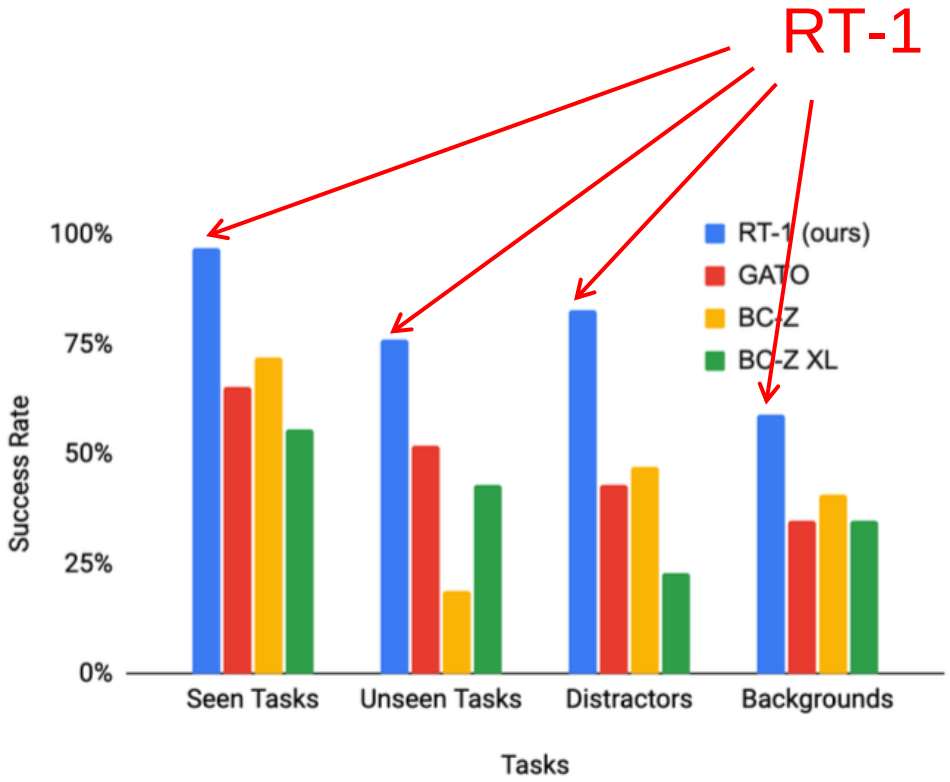
- Large-scale imitation learning
  - A policy that maps (observation/state, goal) to action



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**

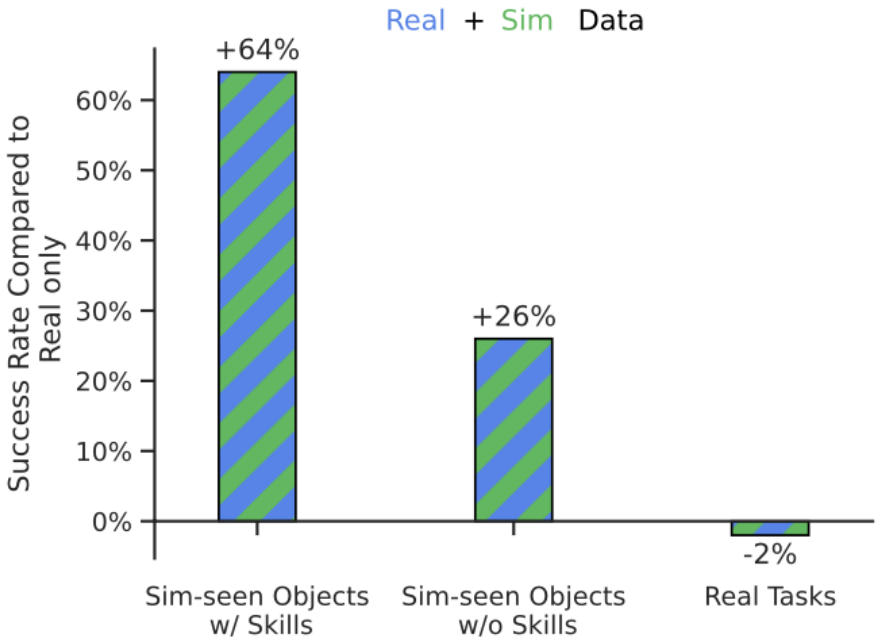


Robotic Transformer 1 (RT-1)

- Question #2: Does simulation data help with the performance?

Unseen object during real-world data collection

Models	Training Data	Real Objects: Seen Skill w/ Objects	Sim Objects (not seen in real): Seen Skill w/ Objects	Sim Objects (not seen in real): Unseen Skill w/ Objects
RT-1	Real Only	92	23	7
RT-1	Real + Sim	90(-2)	**87(+64)**	**33(+26)**

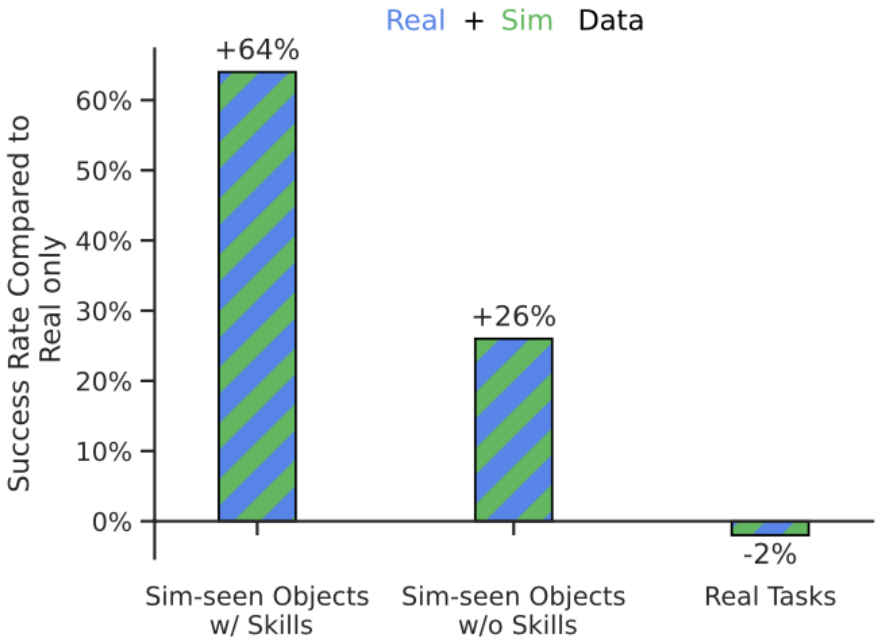


Robotic Transformer 1 (RT-1)

- Question #2: Does simulation data help with the performance?

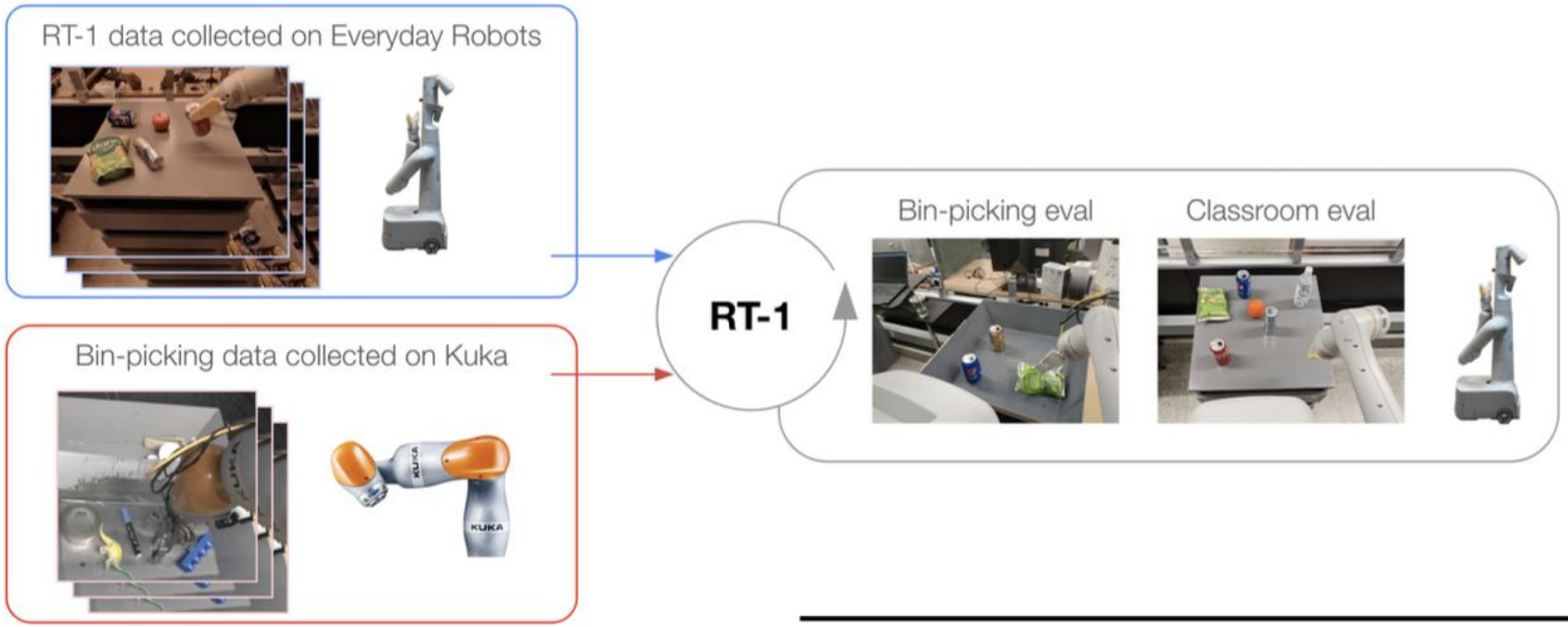
Models	Training Data	Real Objects: Seen Skill w/ Objects	Sim Objects (not seen in real): Seen Skill w/ Objects	Sim Objects (not seen in real): Unseen Skill w/ Objects
RT-1	Real Only	92	23	7
RT-1	Real + Sim	90(-2)	**87(+64)**	**33(+26)**

Training with sim data



Robotic Transformer 1 (RT-1)

- Question #3: Data from different robot?



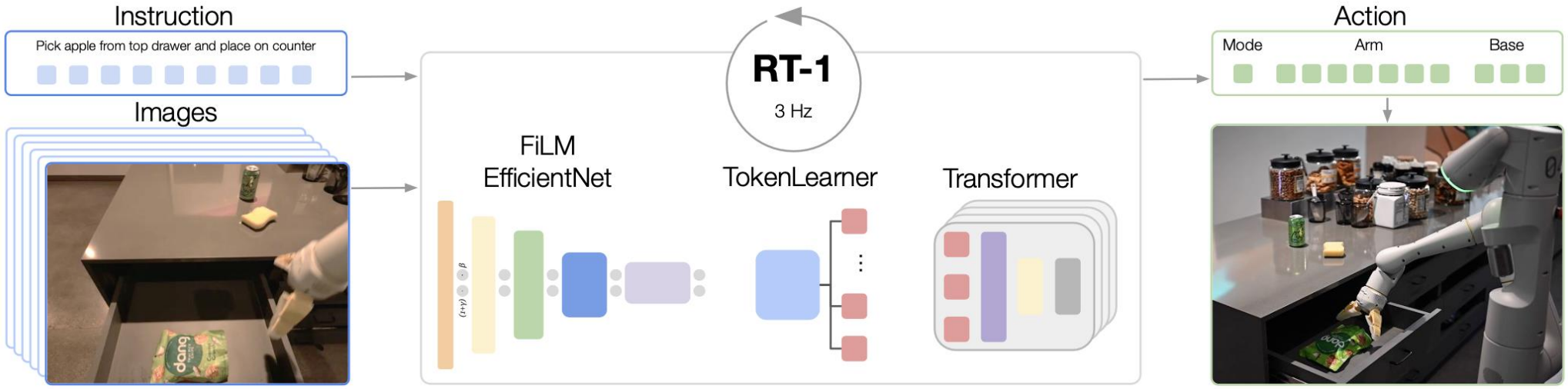
Bin-picking data from a different robot also helps

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)

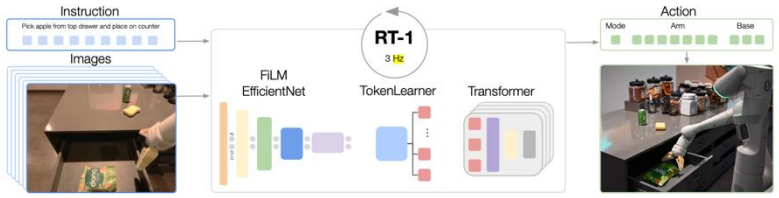
- Large-scale language-conditioned imitation learning.
- Significant data collection and engineering efforts.
- Among the initial investigations: (1) how to scale up and (2) what to expect.

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

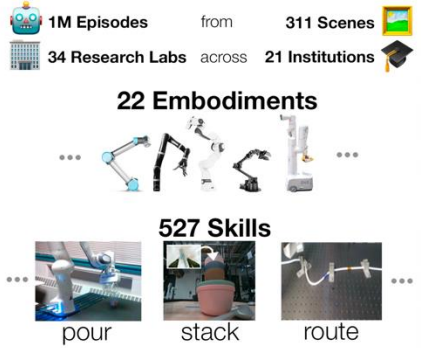


Robotic Foundation Models

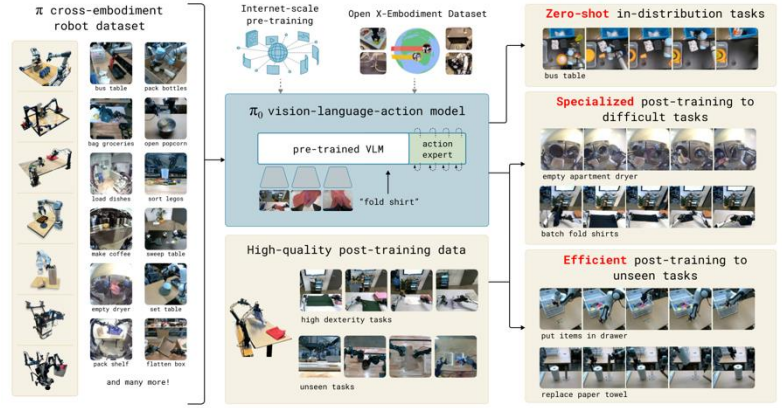
- What is a Robotic Foundation Model?
  - No explicit representation of states / transition functions
  - A policy that maps (observation/state, goal) to action



RT-1 (Dec. 2022)

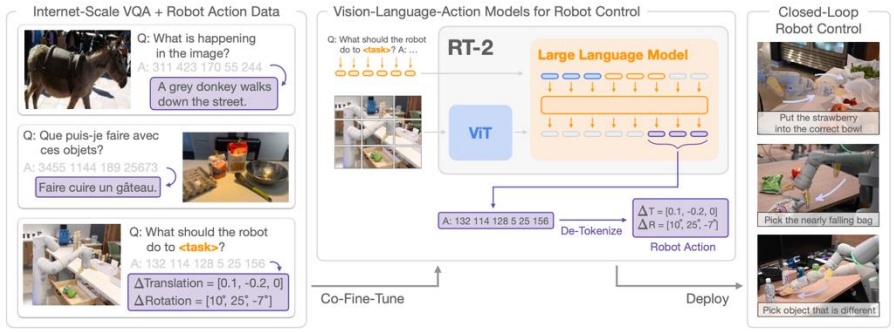


RT-X (Oct. 2023)

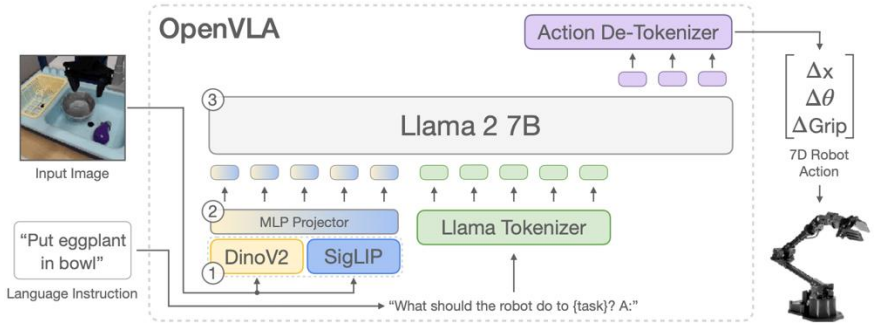


Pi-Zero (Oct. 2024)

RT-2 (Jul. 2023)

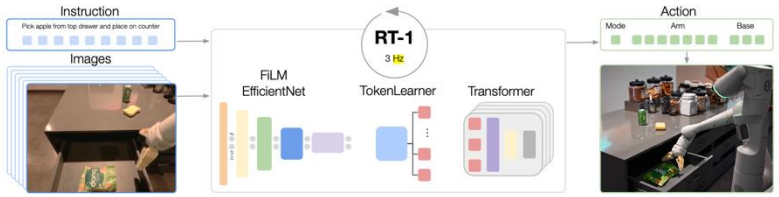


OpenVLA (Jun. 2024)

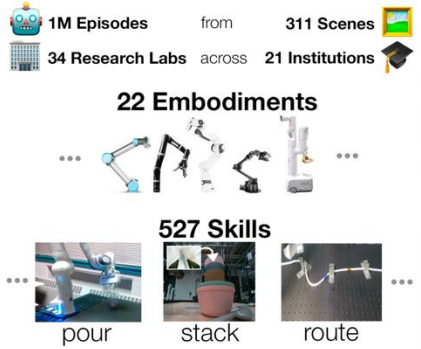


Robotic Foundation Models

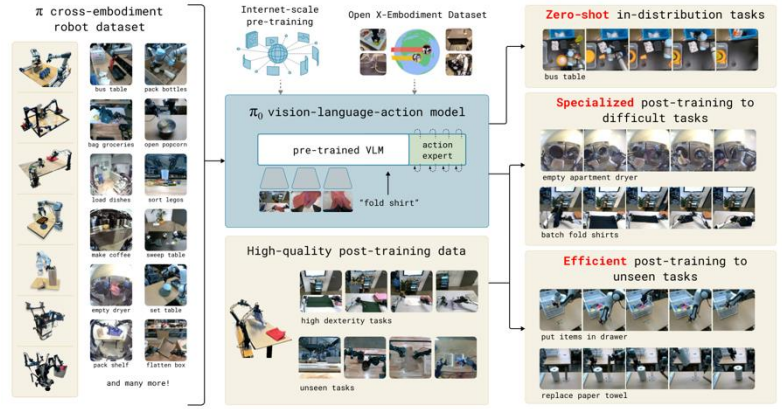
- What is a Robotic Foundation Model?
  - No explicit representation of states / transition functions
  - A policy that maps (observation/state, goal) to action



RT-1 (Dec. 2022)

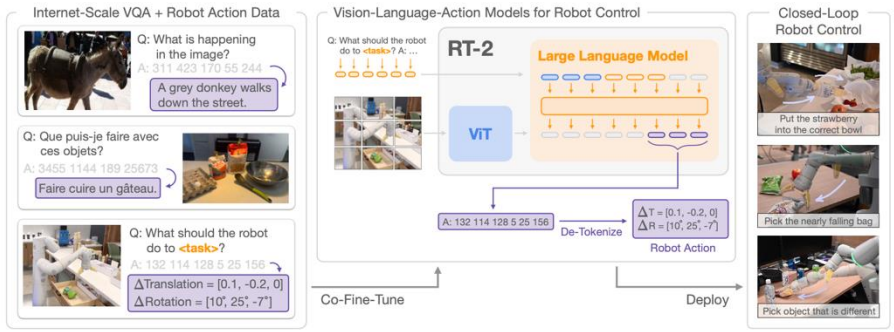


RT-X (Oct. 2023)

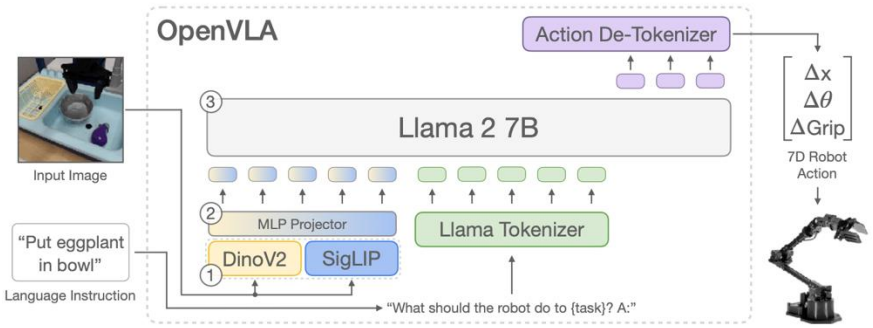


Pi-Zero (Oct. 2024)

RT-2 (Jul. 2023)



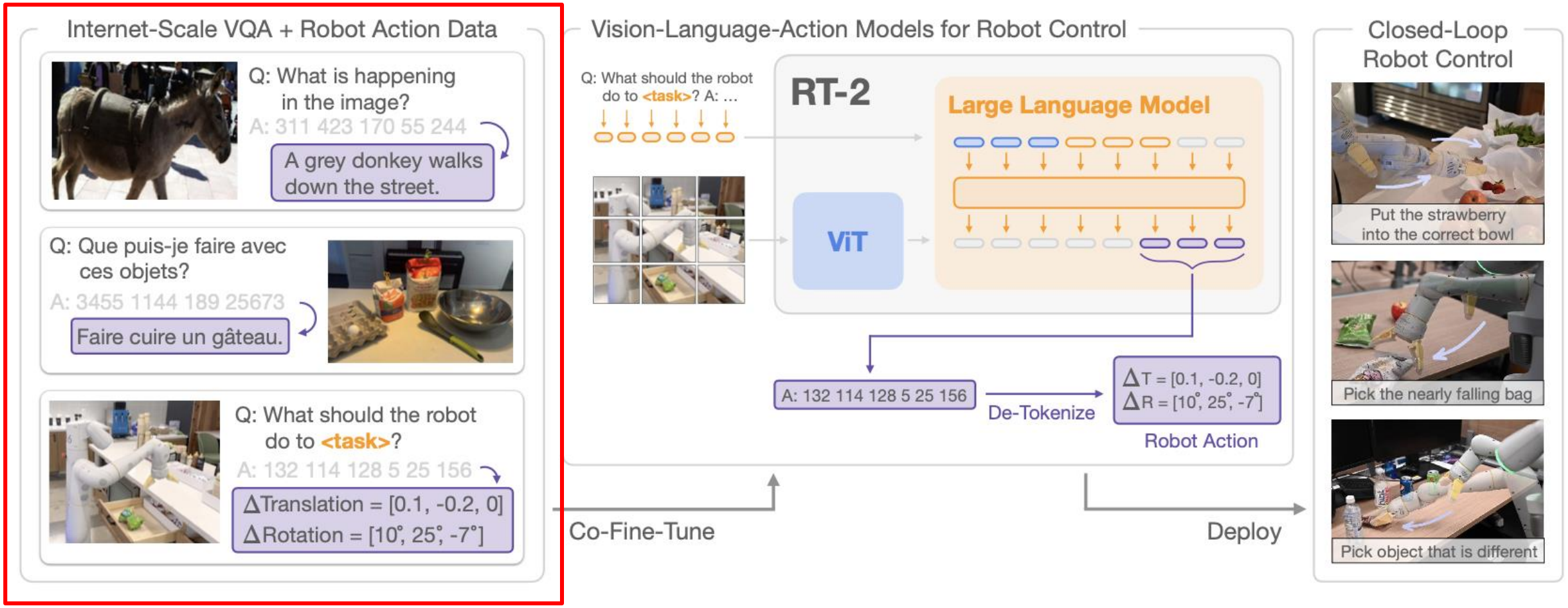
OpenVLA (Jun. 2024)



Robotic Transformer 2 (RT-2)

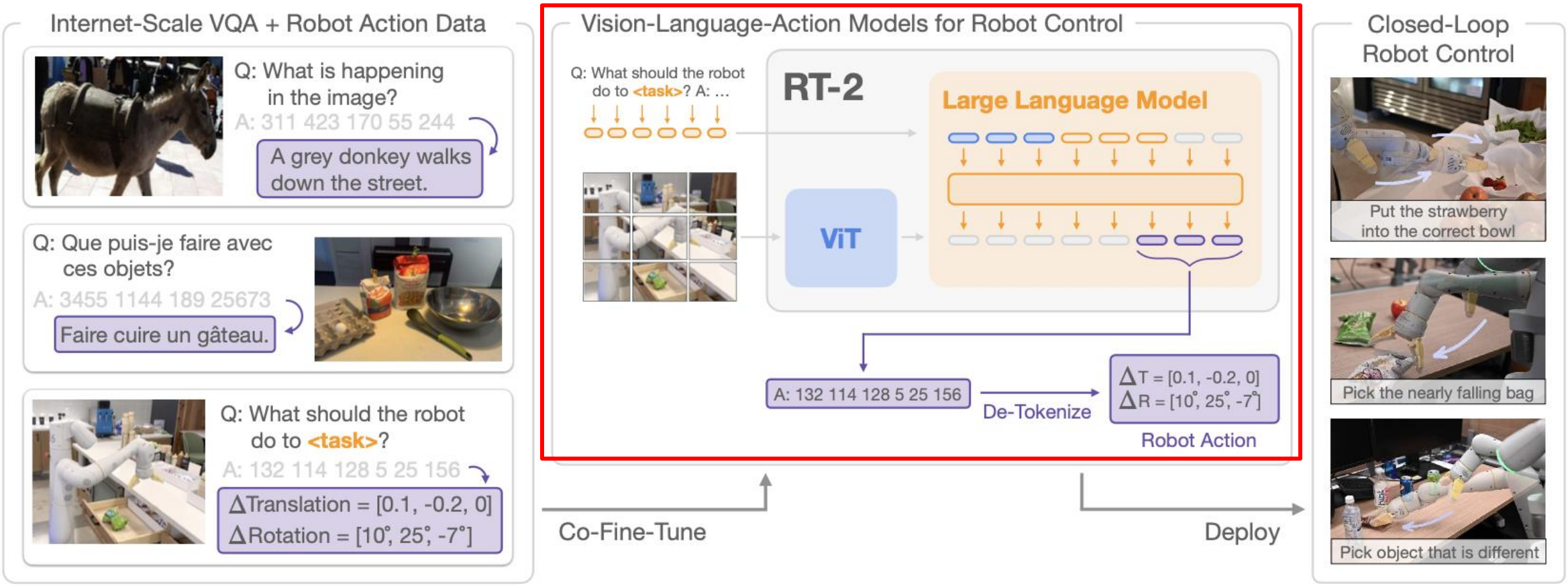
- First released in July 2023
- How VLMs can be incorporated into Robotic Foundation Models?
- Key idea: co-fine-tune VLMs on both
  - (1) robot data
  - (2) Internet-scale vision-language tasks (e.g., VQA)
- Introduced the name: Vision-Language-Action Models (VLA)

Robotic Transformer 2 (RT-2)



Tokenize the robot actions

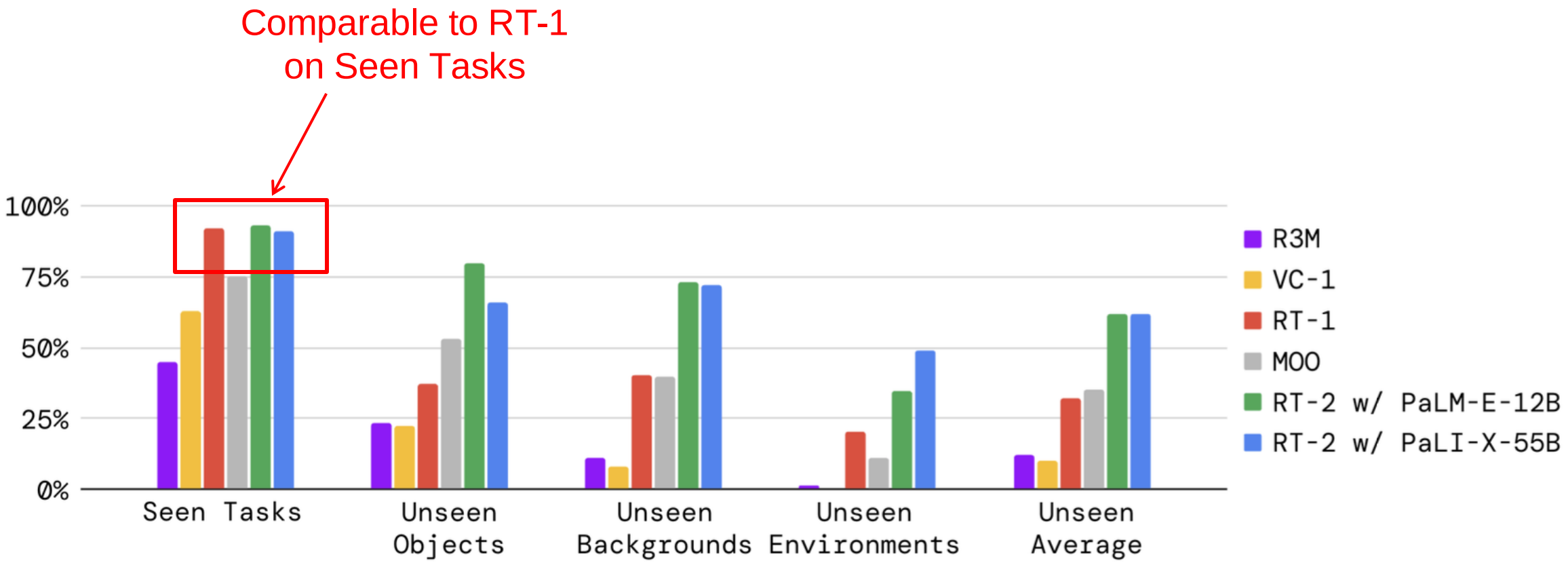
Robotic Transformer 2 (RT-2)



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

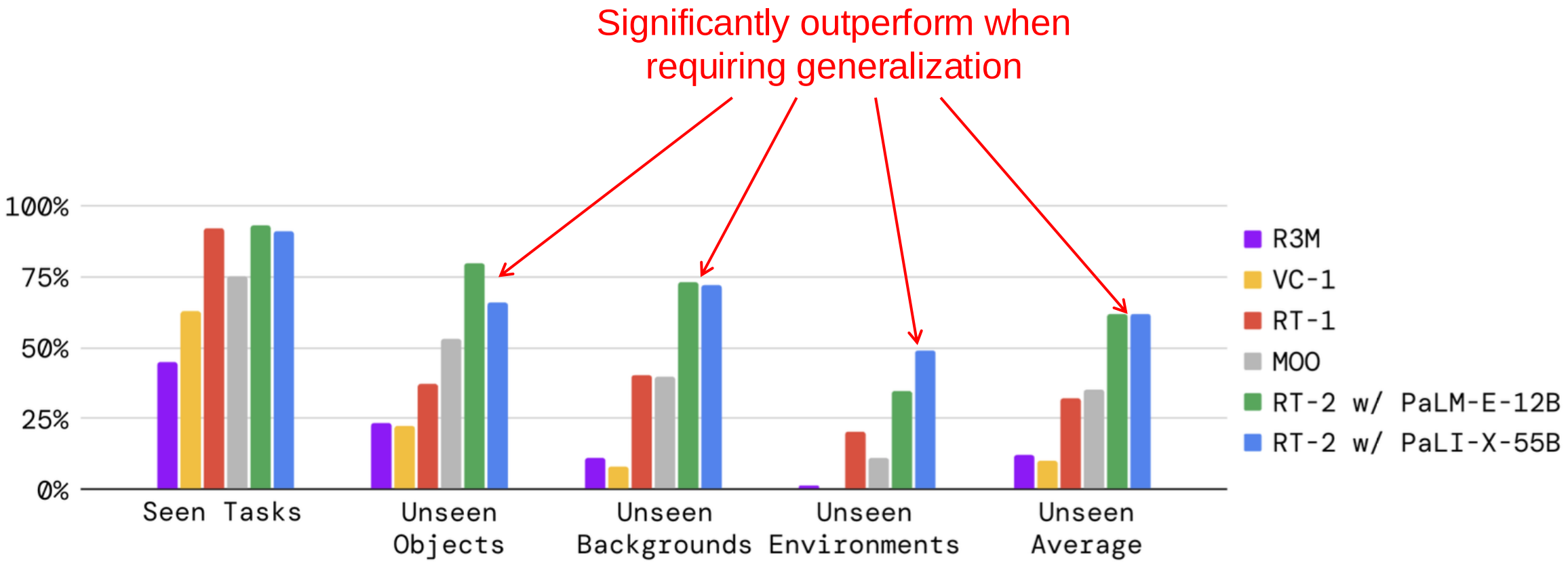
Robotic Transformer 2 (RT-2)

- 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

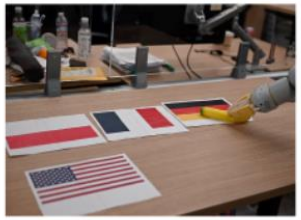
move apple to Denver Nuggets

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 @bhea

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.

joining The Verge, she covered the economy.

3 New

COURTESY OF GOOGLE

FORBES > INNOVATION > CLOUD

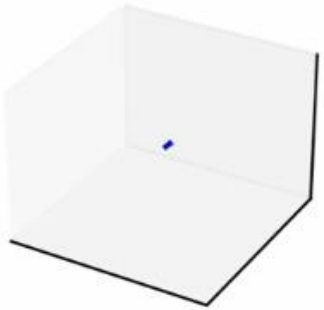
EDITORS' PICK

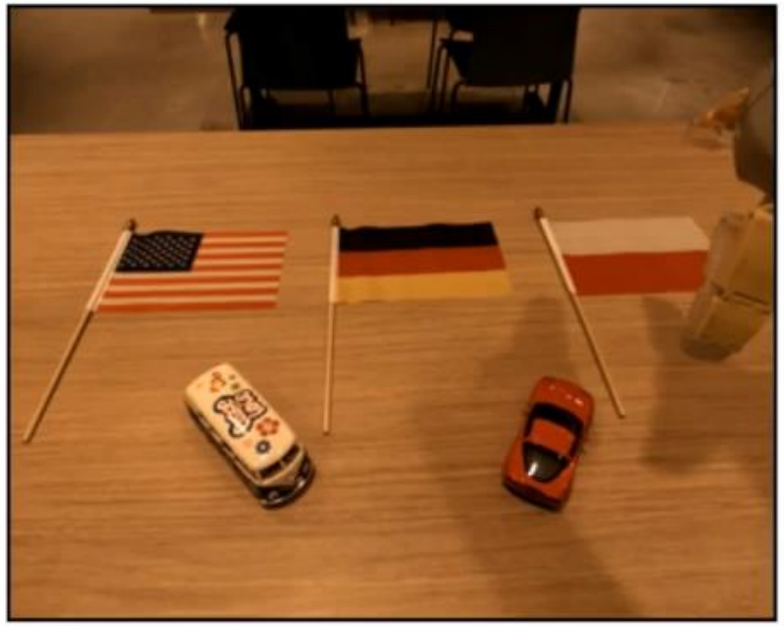
Google's RT-2 AI Model: A Step Closer To Robots That Can Learn Like Humans

Janakiram MSV Senior Contributor

I cover emerging technologies with a focus on infrastructure and AI

Follow

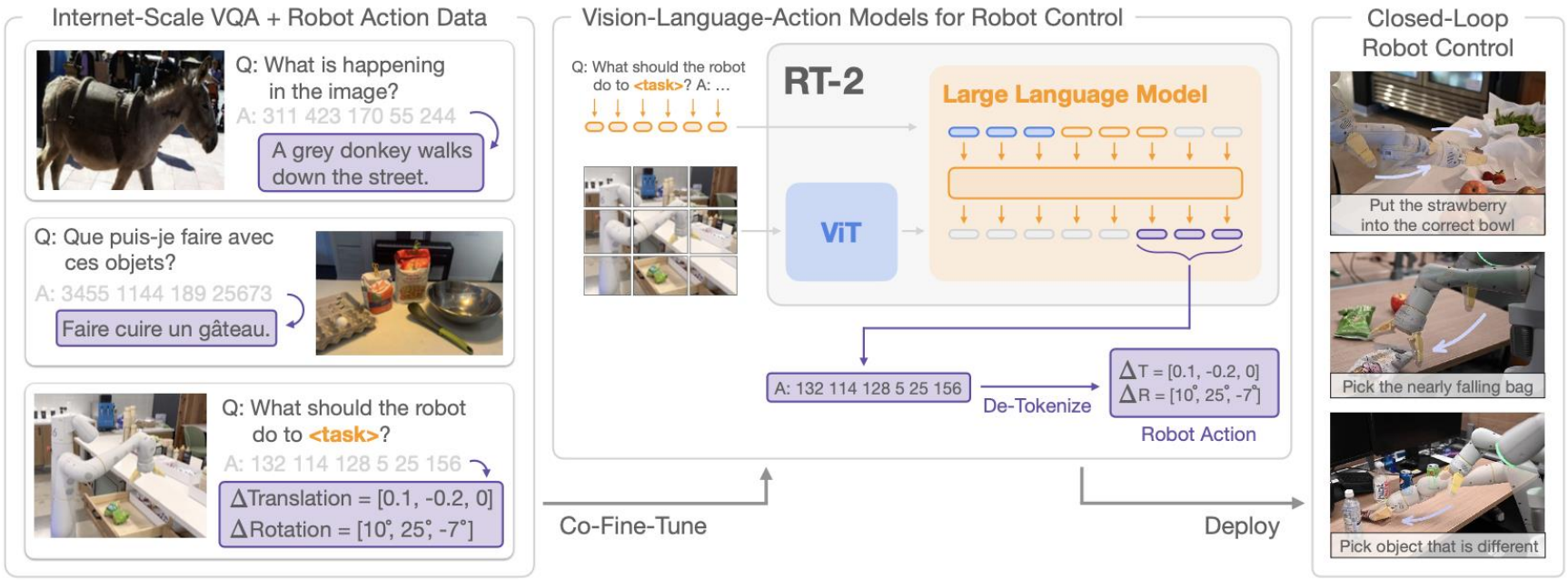




Robotic Transformer 2 (RT-2)

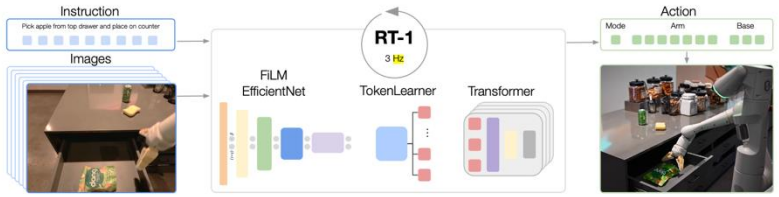
- Co-fine-tuning boosts generalization over semantic and visual concepts

- Limited to seen skills but can deploy them in new ways
- Efficiency is still an issue
- The absolute performance is still not ideal

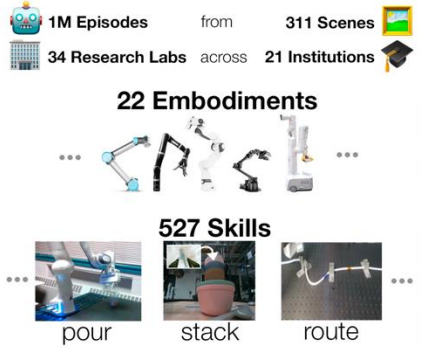


Robotic Foundation Models

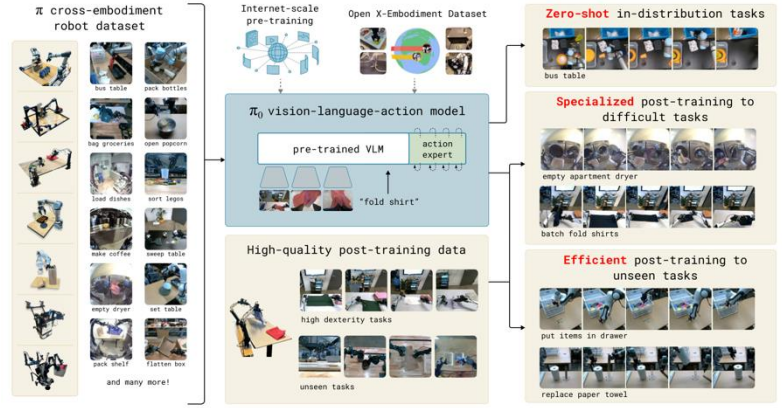
- What is a Robotic Foundation Model?
  - No explicit representation of states / transition functions
  - A policy that maps (observation/state, goal) to action



RT-1 (Dec. 2022)

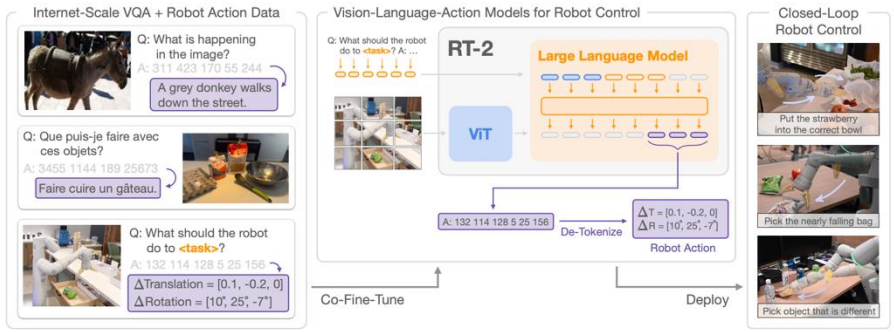


RT-X (Oct. 2023)

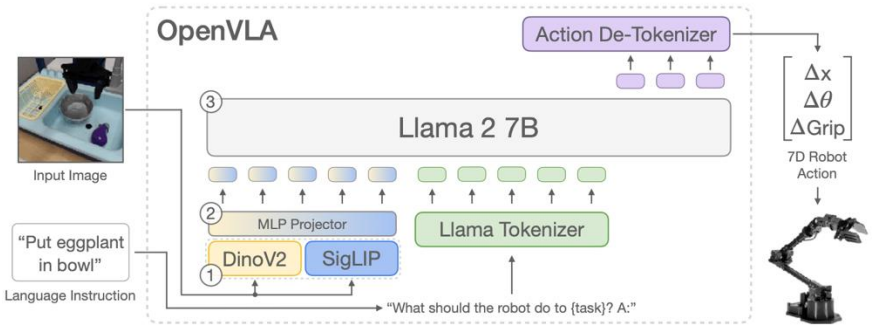


Pi-Zero (Oct. 2024)

RT-2 (Jul. 2023)

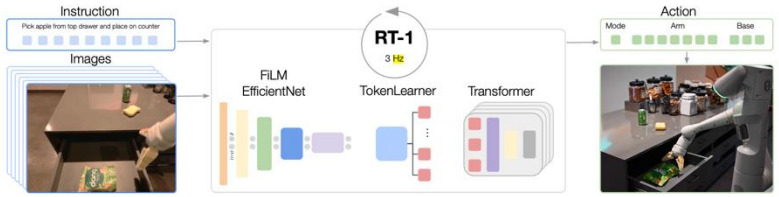


OpenVLA (Jun. 2024)

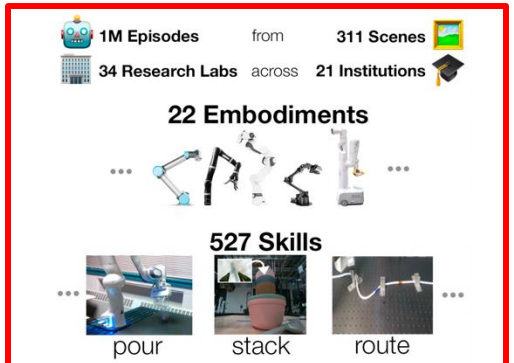


Robotic Foundation Models

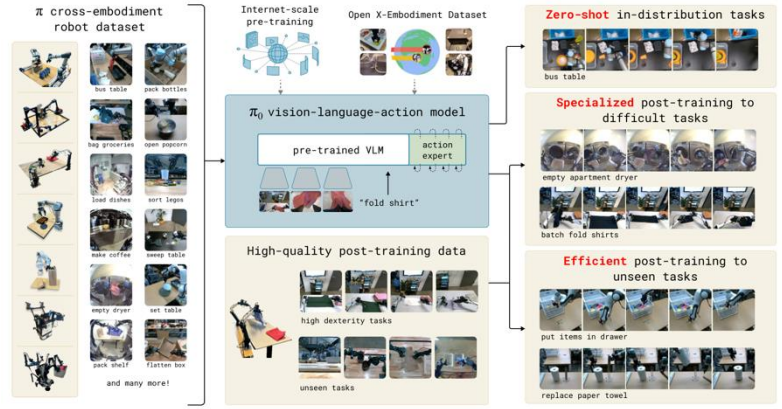
- What is a Robotic Foundation Model?
  - No explicit representation of states / transition functions
  - A policy that maps (observation/state, goal) to action



RT-1 (Dec. 2022)

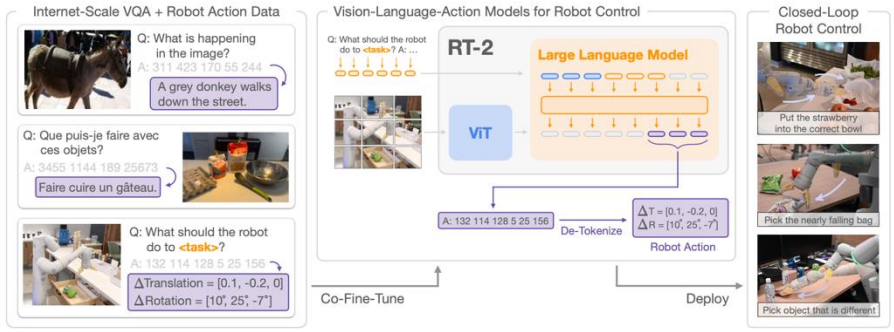


RT-X (Oct. 2023)

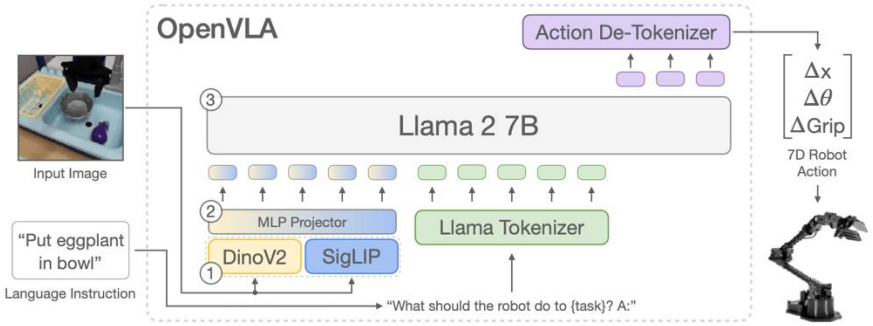


Pi-Zero (Oct. 2024)

RT-2 (Jul. 2023)

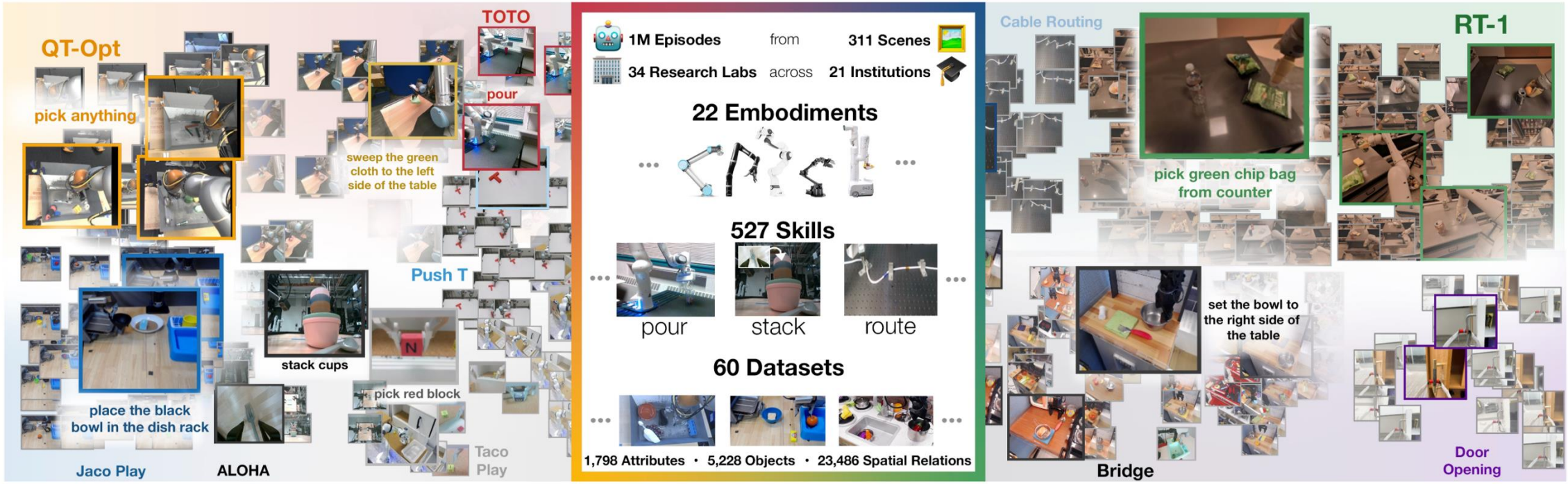


OpenVLA (Jun. 2024)



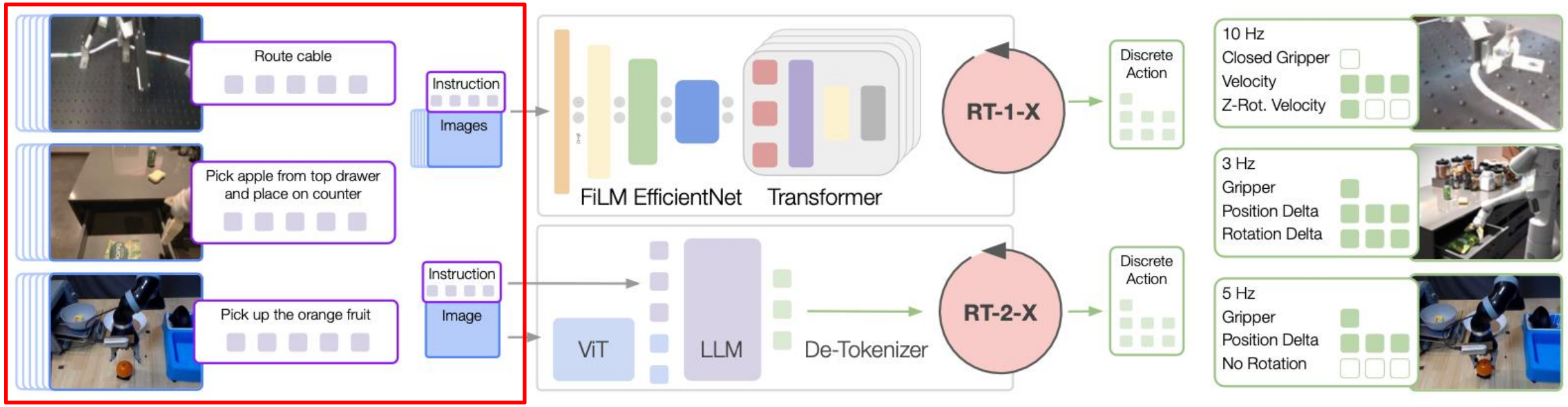
Robotic Transformer X (RT-X)

- 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)



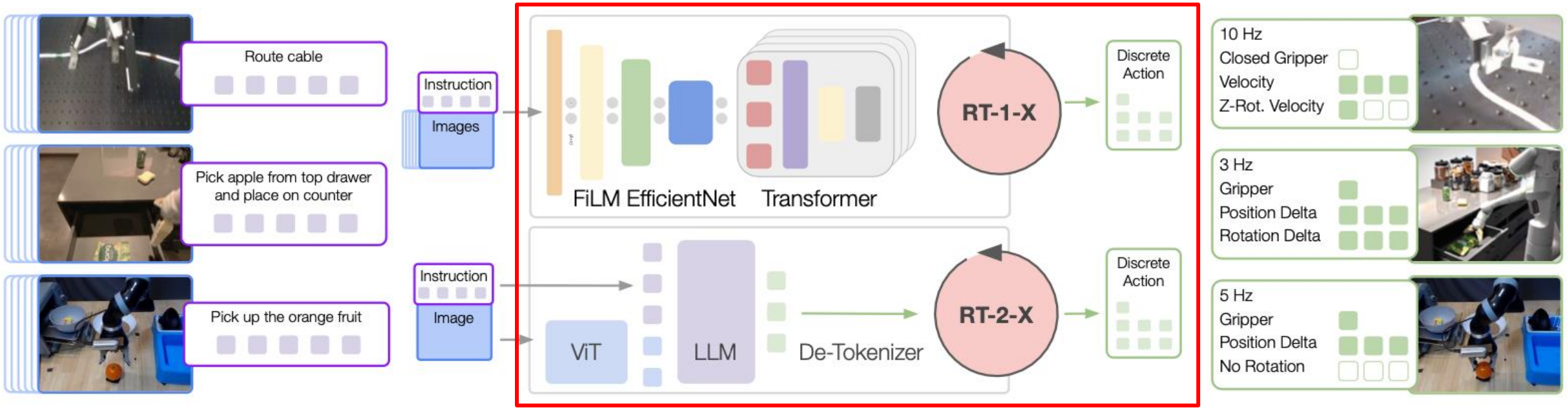
Robotic Transformer X (RT-X)

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



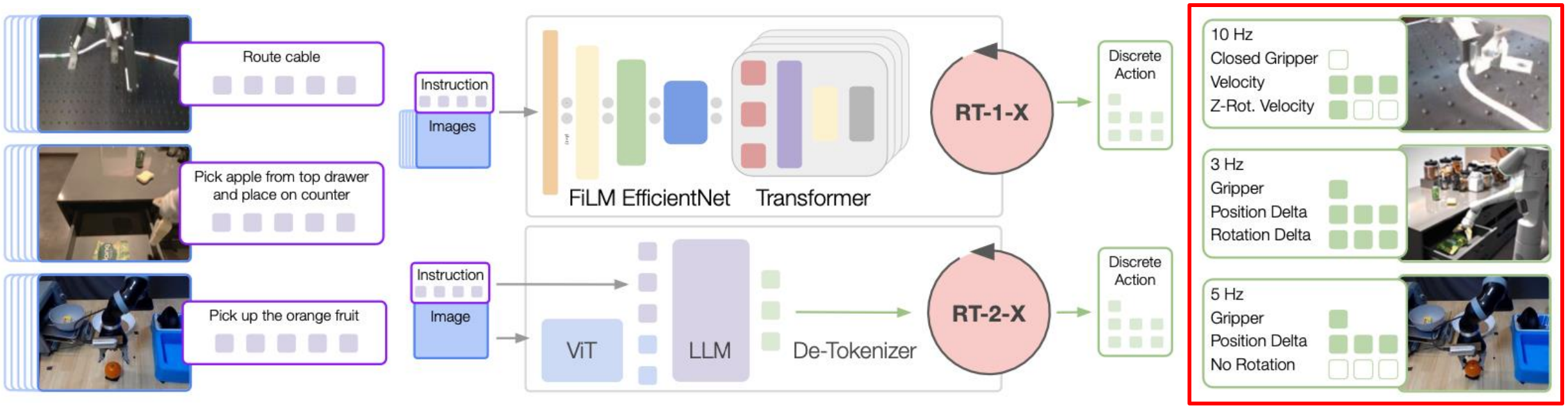
Robotic Transformer X (RT-X)

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



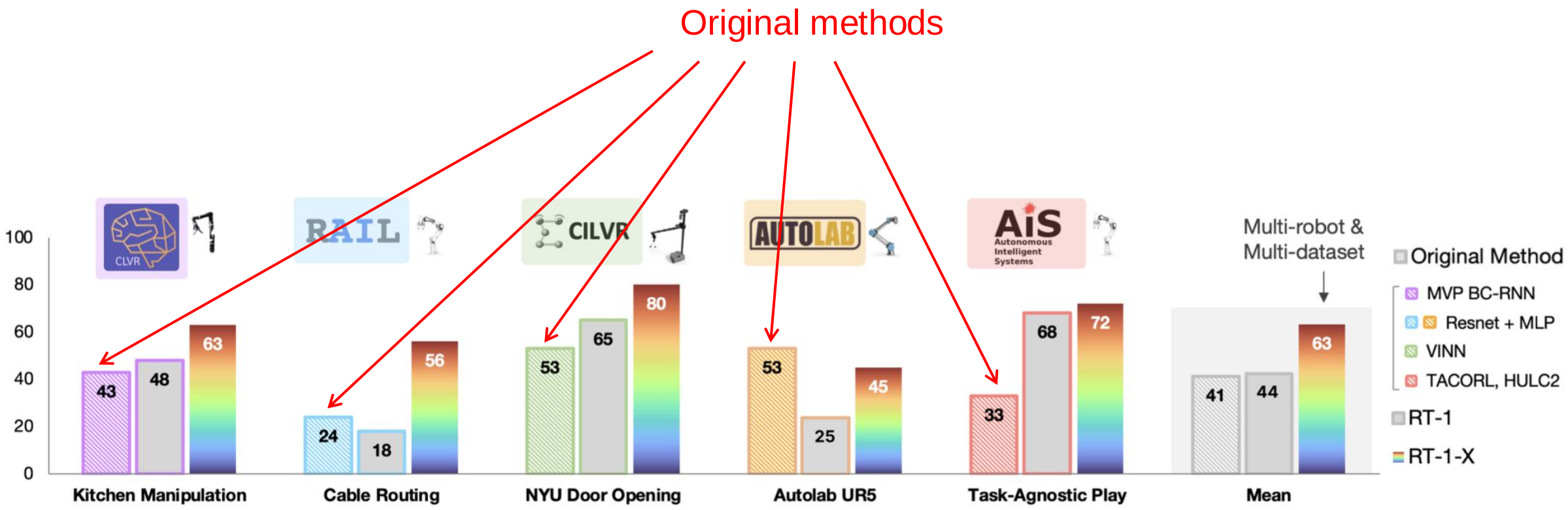
Robotic Transformer X (RT-X)

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



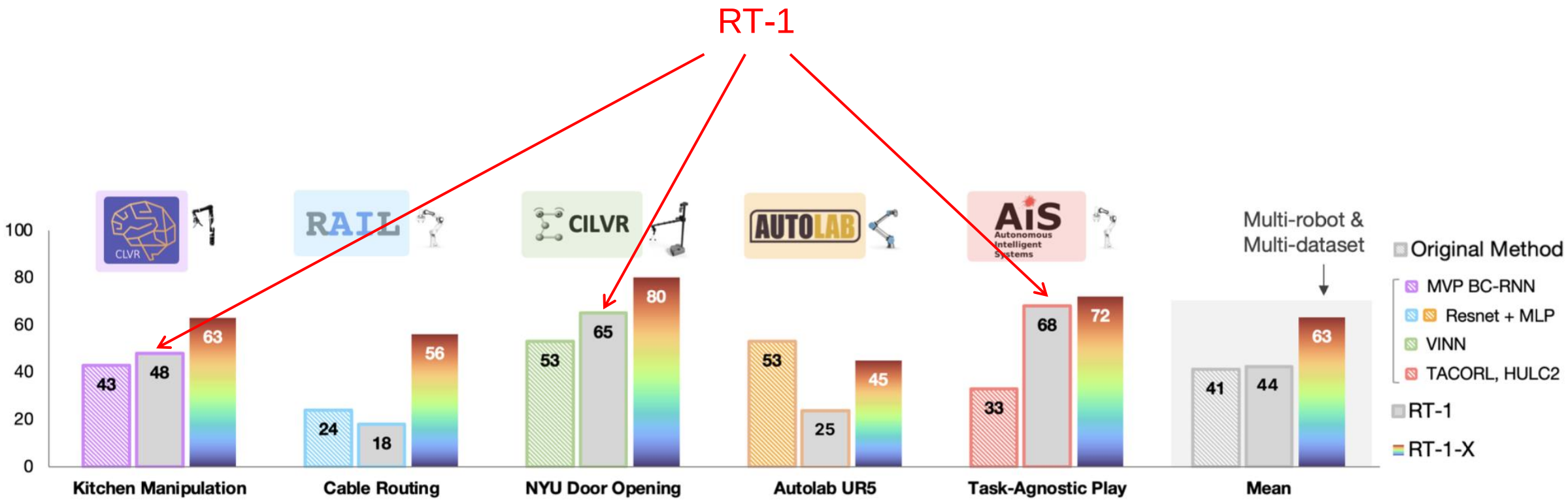
Robotic Transformer X (RT-X)

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



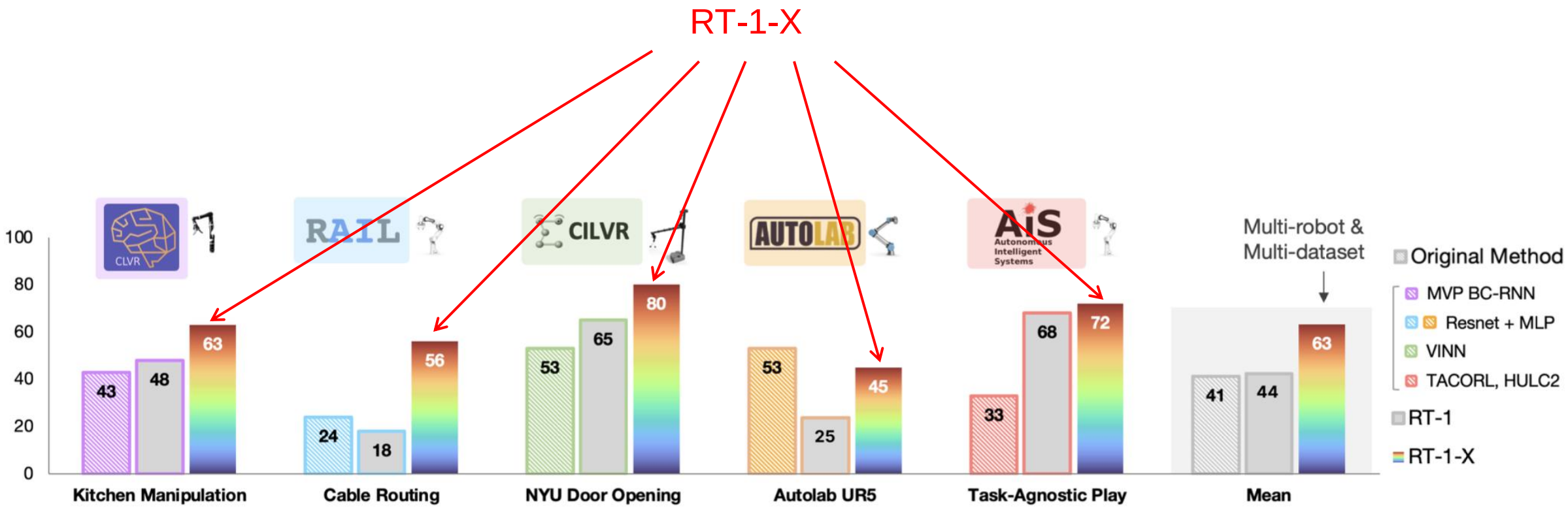
Robotic Transformer X (RT-X)

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



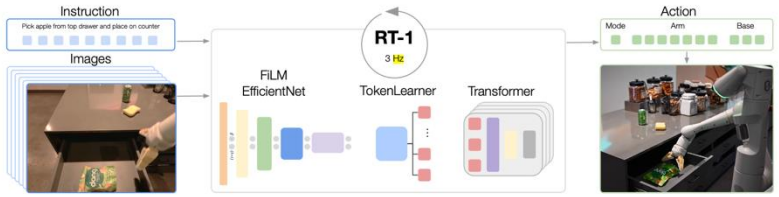
Robotic Transformer X (RT-X)

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

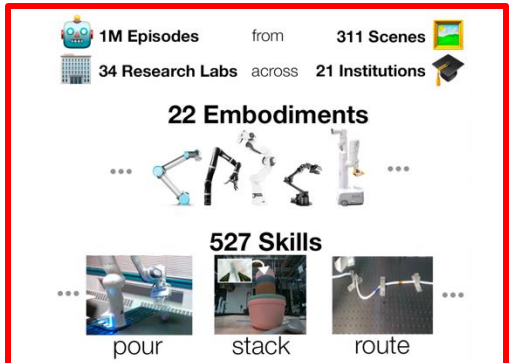


Robotic Foundation Models

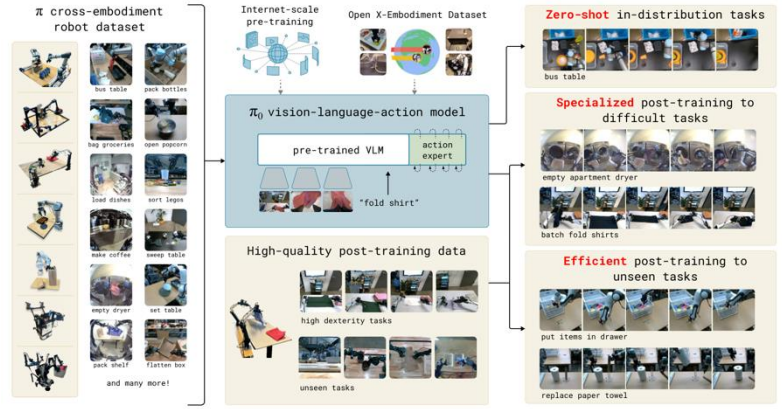
- What is a Robotic Foundation Model?
  - No explicit representation of states / transition functions
  - A policy that maps (observation/state, goal) to action



RT-1 (Dec. 2022)

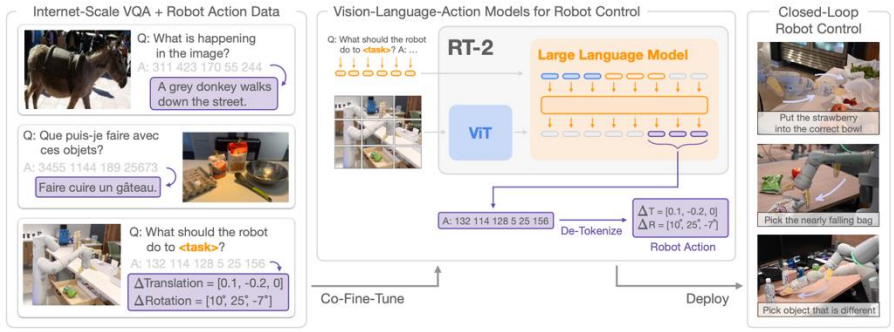


RT-X (Oct. 2023)

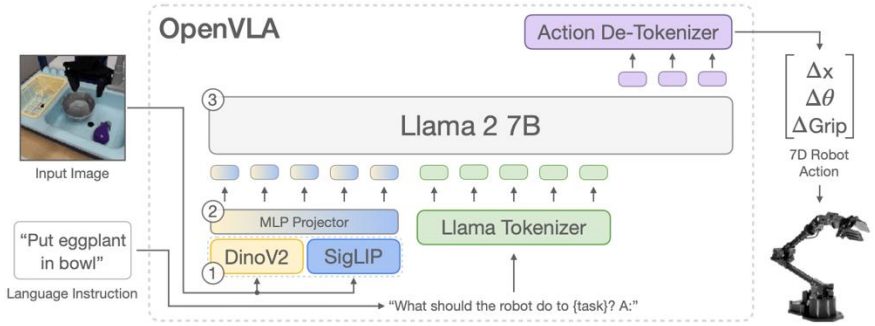


Pi-Zero (Oct. 2024)

RT-2 (Jul. 2023)

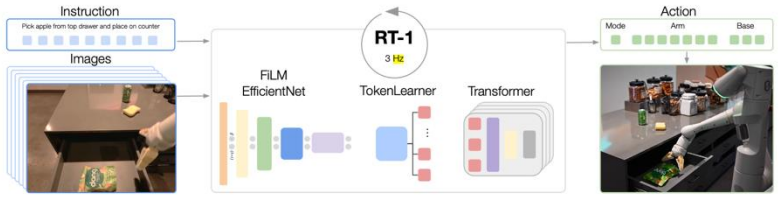


OpenVLA (Jun. 2024)

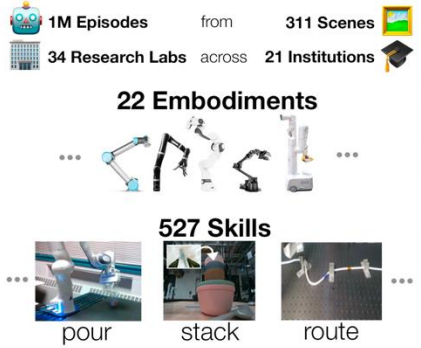


Robotic Foundation Models

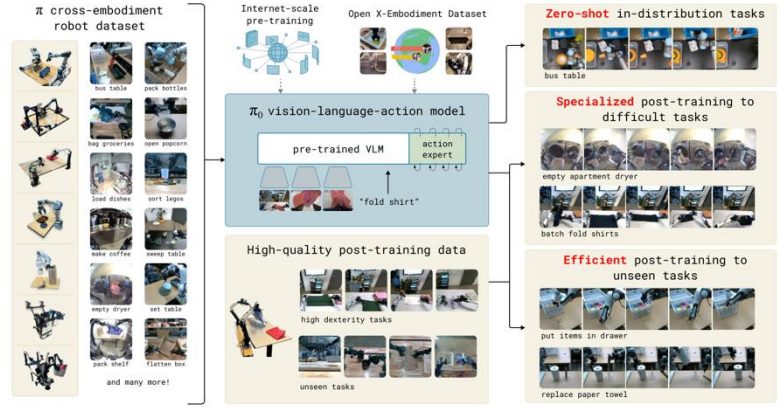
- What is a Robotic Foundation Model?
  - No explicit representation of states / transition functions
  - A policy that maps (observation/state, goal) to action



RT-1 (Dec. 2022)

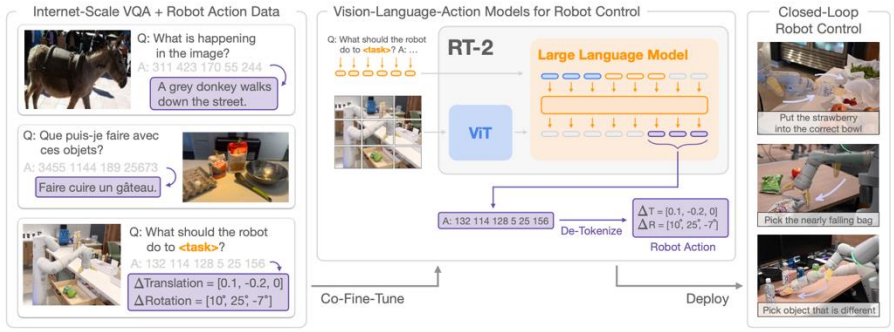


RT-X (Oct. 2023)

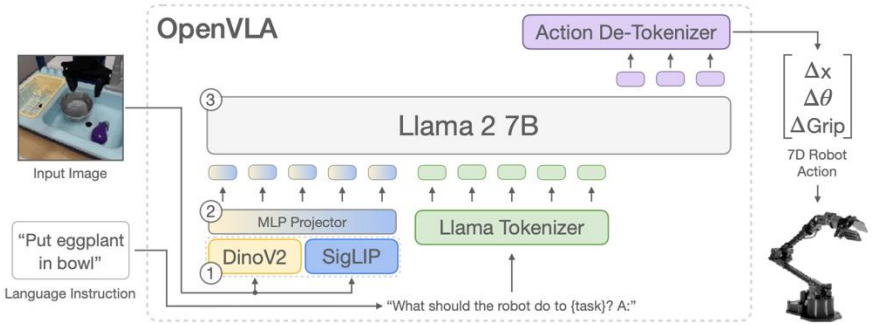


Pi-Zero (Oct. 2024)

RT-2 (Jul. 2023)



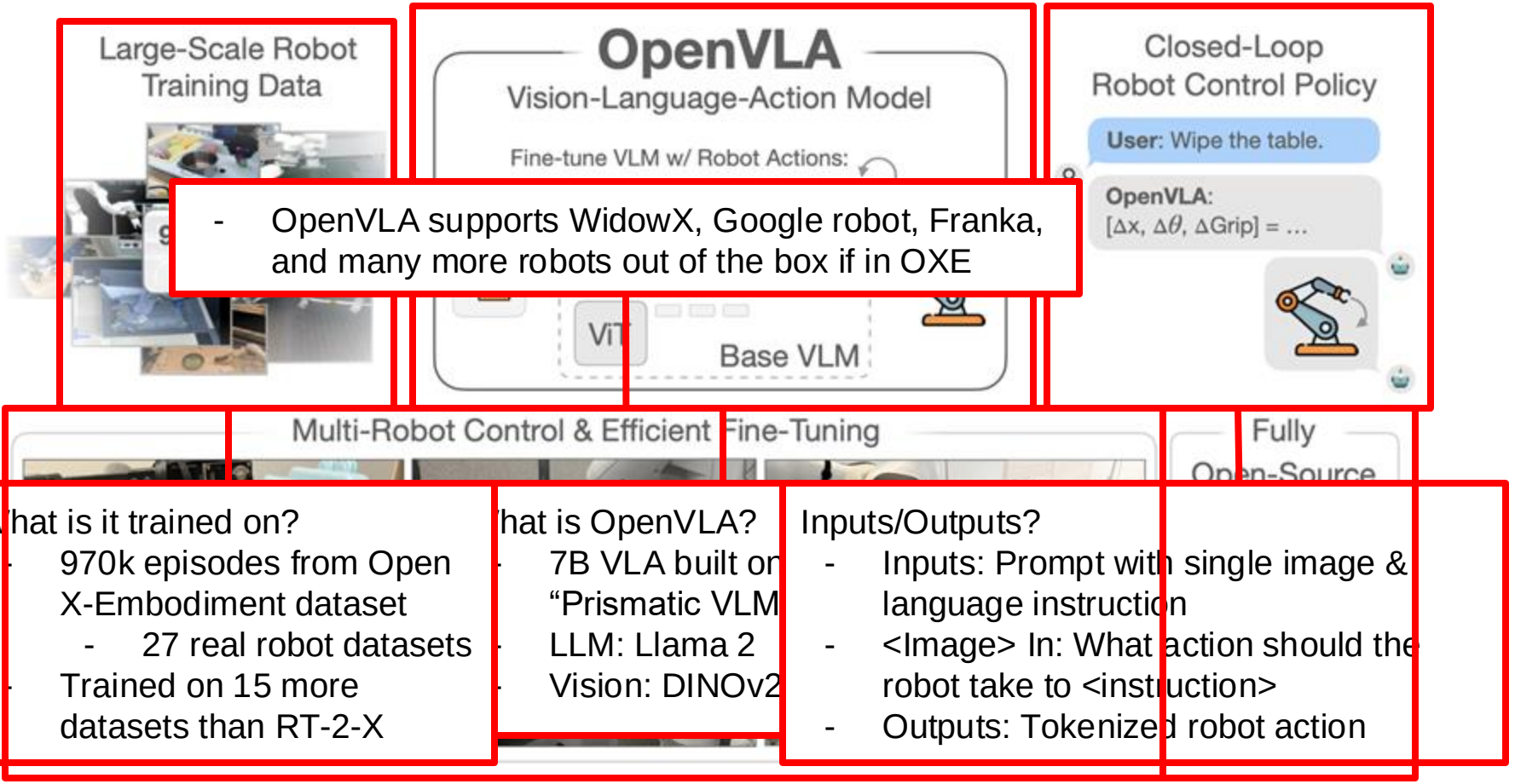
OpenVLA (Jun. 2024)



OpenVLA

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

OpenVLA



- OpenVLA supports WidowX, Google robot, Franka, and many more robots out of the box if in OXE

What is it trained on?

- 970k episodes from Open X-Embodiment dataset
  - 27 real robot datasets
- Trained on 15 more datasets than RT-2-X

hat is OpenVLA?

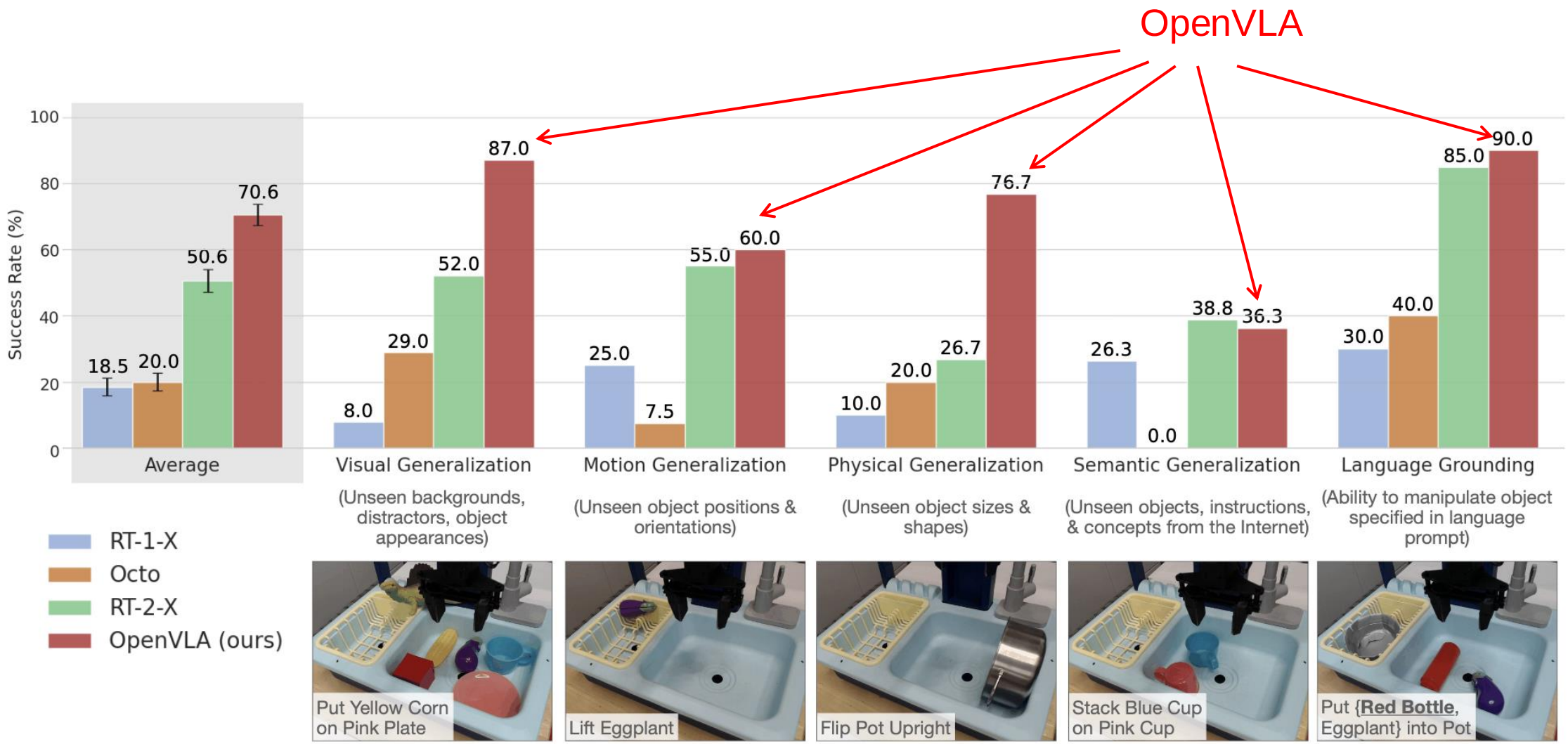
- 7B VLA built on
- “Prismatic VLM
- LLM: Llama 2
- Vision: DINOv2

Inputs/Outputs?

- Inputs: Prompt with single image & language instruction
- <Image> In: What action should the robot take to <instruction>
- Outputs: Tokenized robot action

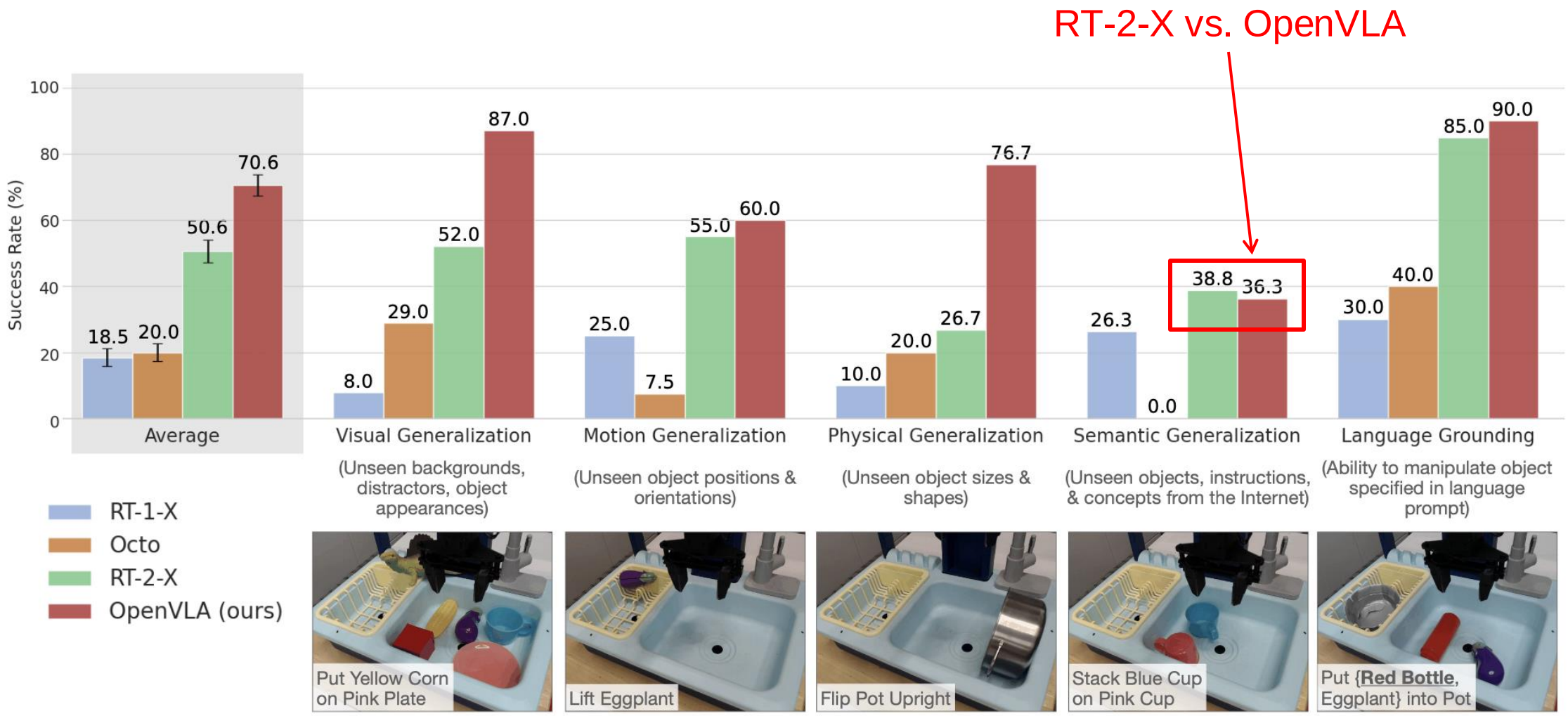
OpenVLA

- Question #1: How well does it work <u>out of the box</u>?



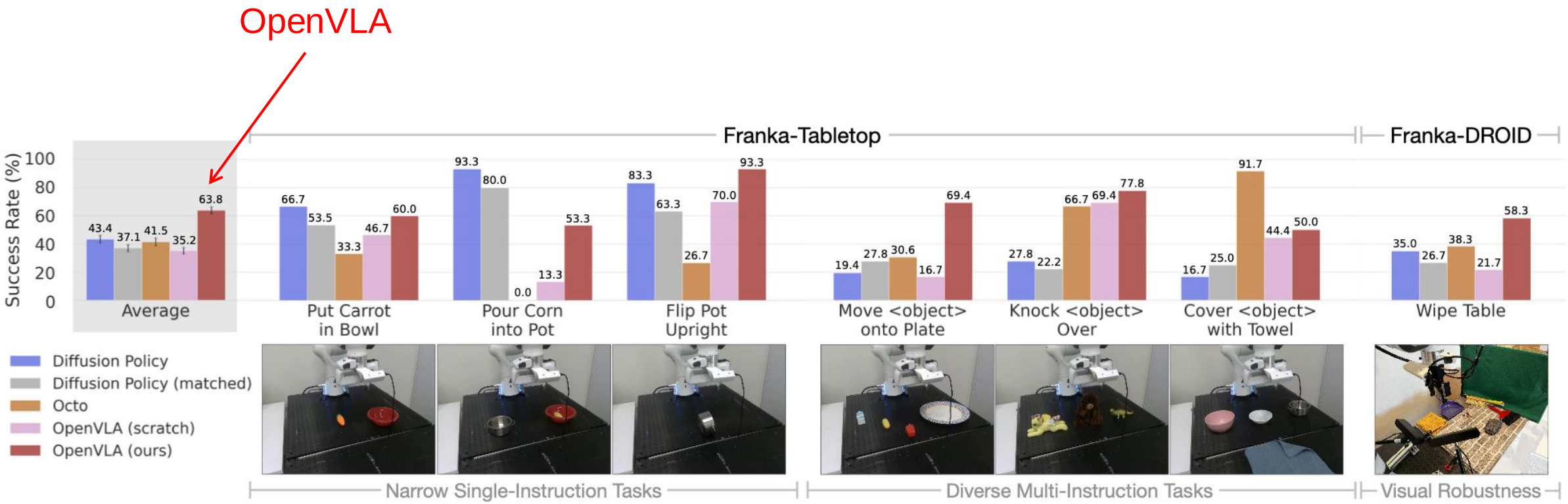
OpenVLA

- Question #1: How well does it work <u>out of the box</u>?



OpenVLA

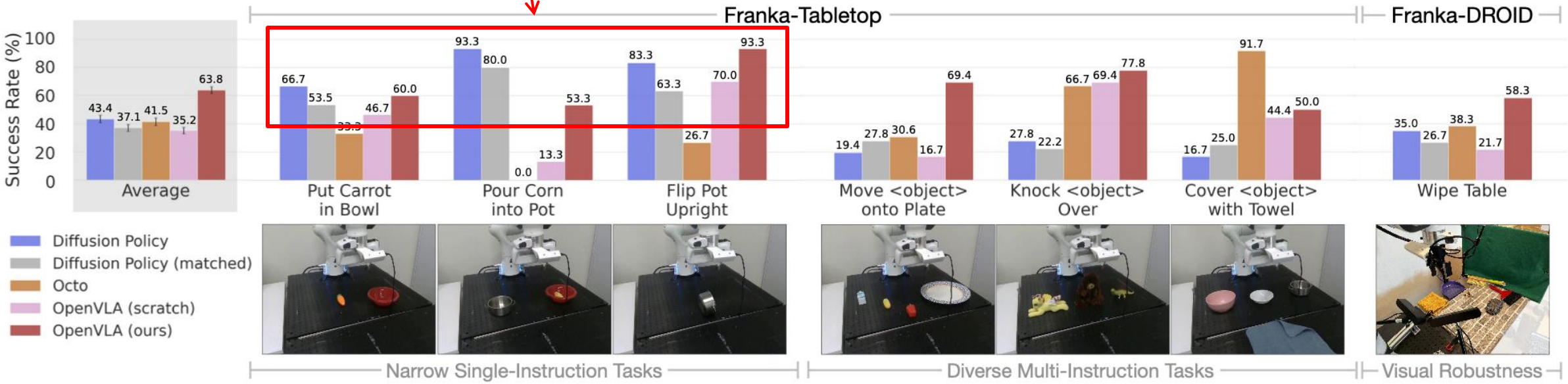
- Question #2: Fine-tuning to adapt to new robot setups



OpenVLA

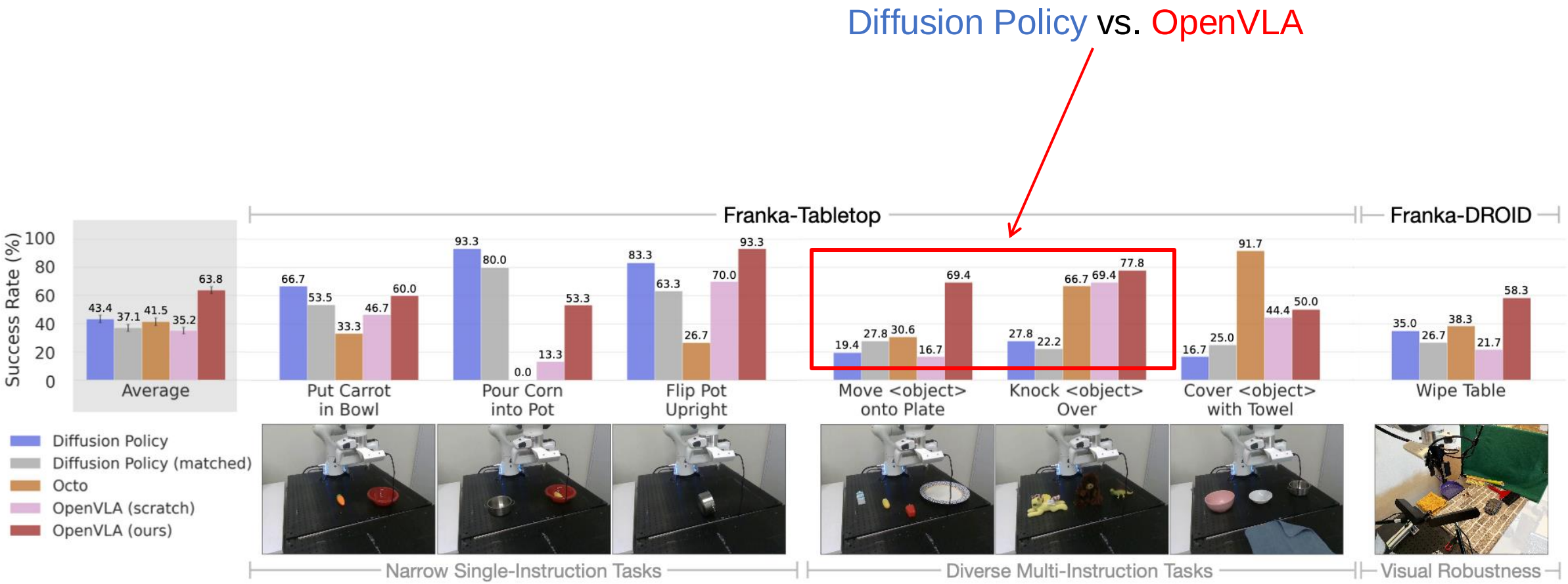
- Question #2: Fine-tuning to adapt to new robot setups

Diffusion Policy vs. OpenVLA



OpenVLA

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



OpenVLA

- Fully open sourced

openvla / openvla

Type / to search

Issues 18 | Pull requests 6 | Actions | Projects | Security | Insights

openvla Public

forked from TRI-ML/prismatic-vlms

Watch 21 | Fork 265 | Star 2k

main | 2 Branches | 0 Tags | Go to file t | Add file | Code

This branch is 46 commits ahead of `TRI-ML/prismatic-vlms:main`. #212

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 — Search models, datasets, users...

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

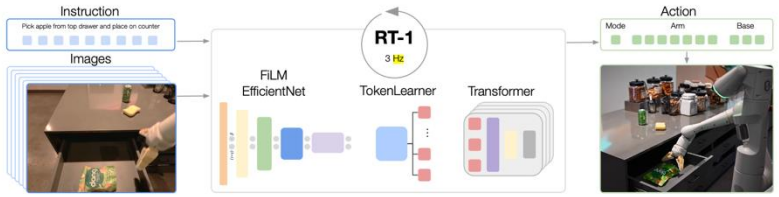
moojink updated a model 5 months ago
openvla/openvla-7b-finetuned-libero-goal

View all activity

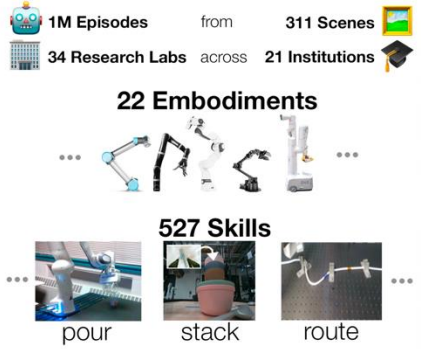
Team members 3

Robotic Foundation Models

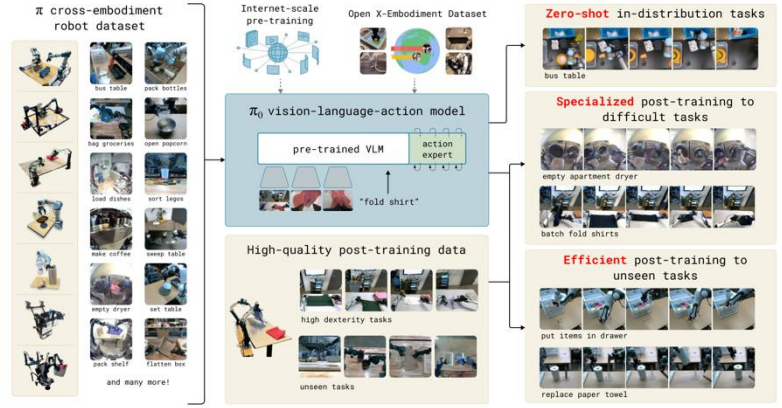
- What is a Robotic Foundation Model?
  - No explicit representation of states / transition functions
  - A policy that maps (observation/state, goal) to action



RT-1 (Dec. 2022)

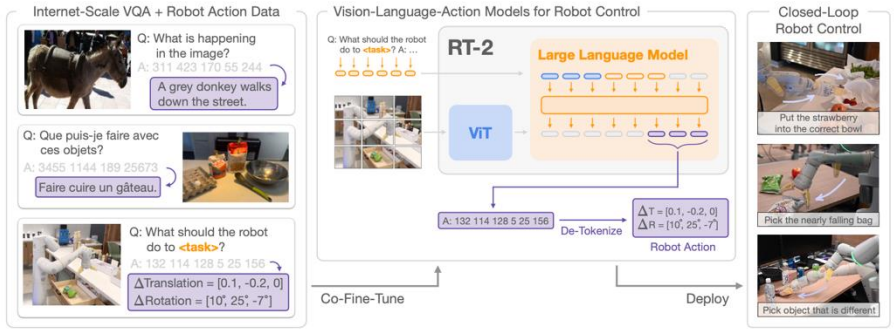


RT-X (Oct. 2023)

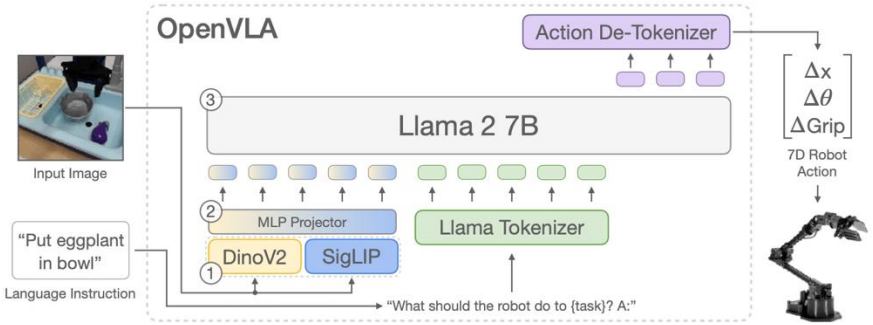


Pi-Zero (Oct. 2024)

RT-2 (Jul. 2023)

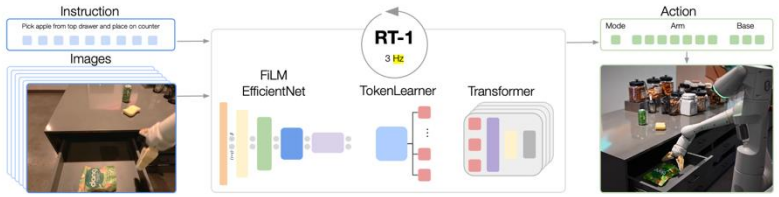


OpenVLA (Jun. 2024)

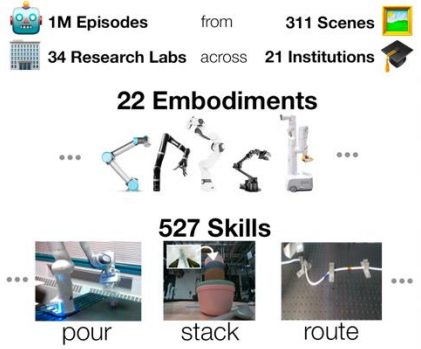


Robotic Foundation Models

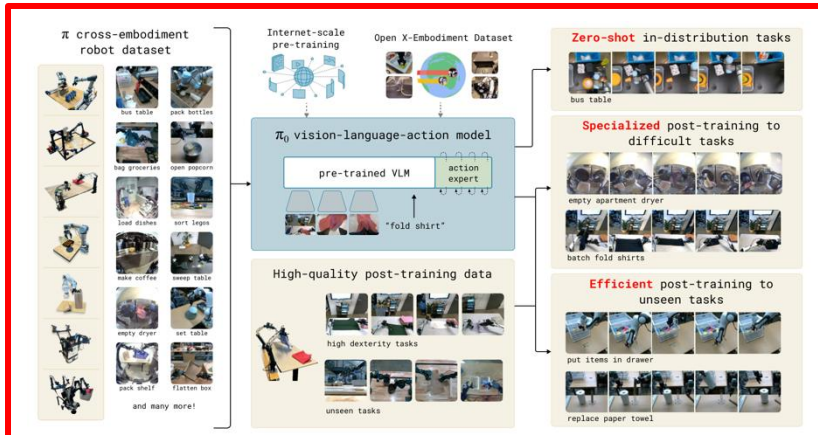
- What is a Robotic Foundation Model?
  - No explicit representation of states / transition functions
  - A policy that maps (observation/state, goal) to action



RT-1 (Dec. 2022)

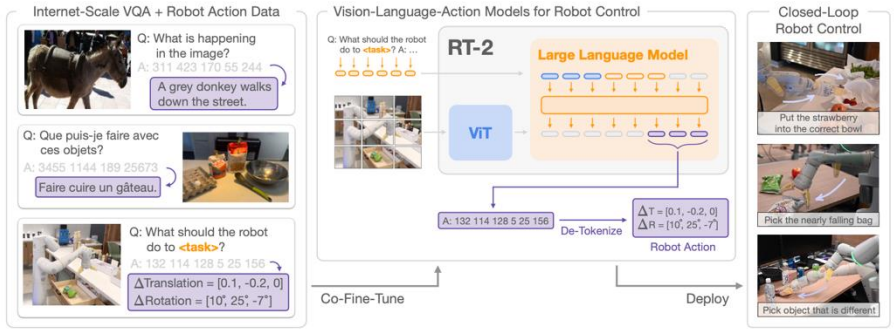


RT-X (Oct. 2023)

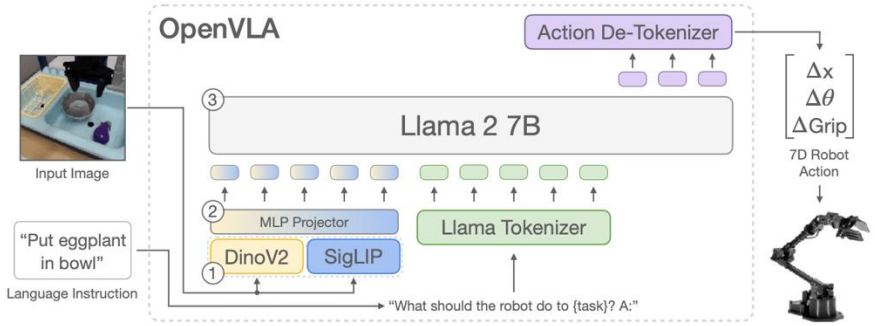


Pi-Zero (Oct. 2024)

RT-2 (Jul. 2023)

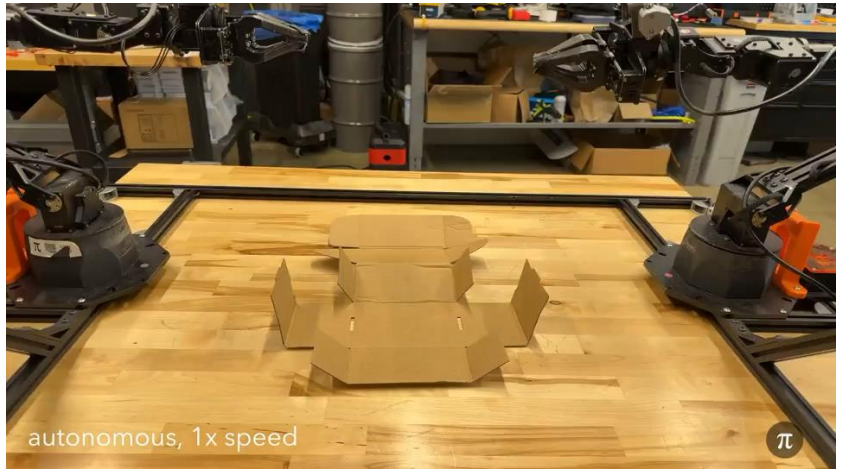


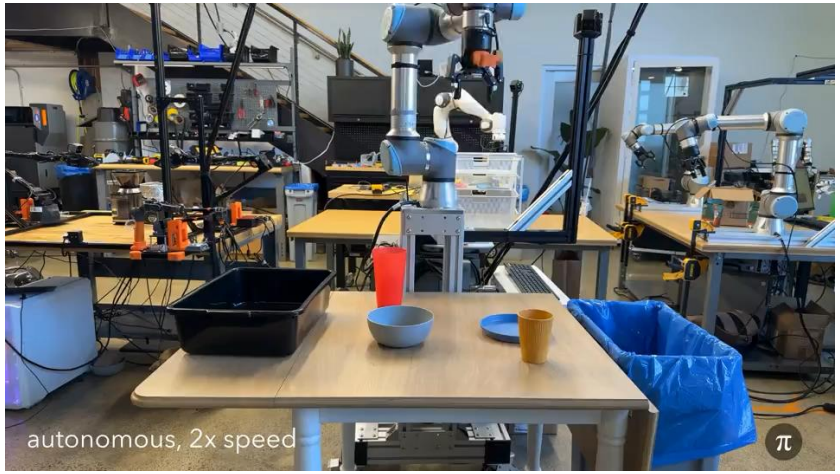
OpenVLA (Jun. 2024)

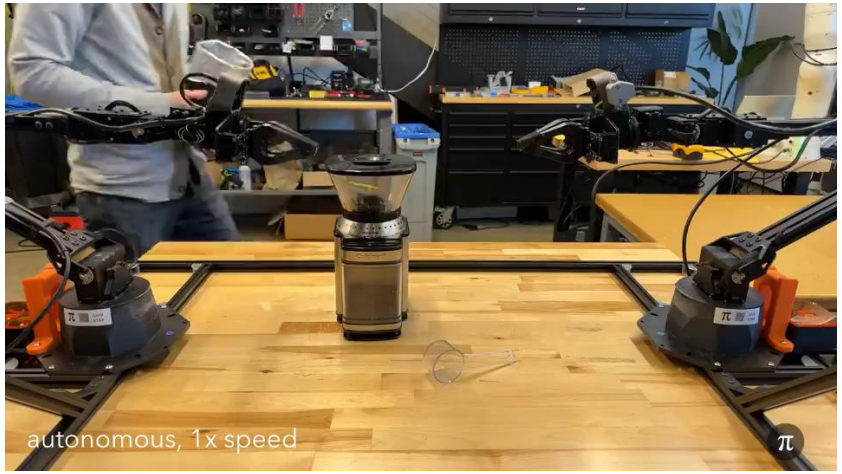


Pi-Zero by Physical Intelligence

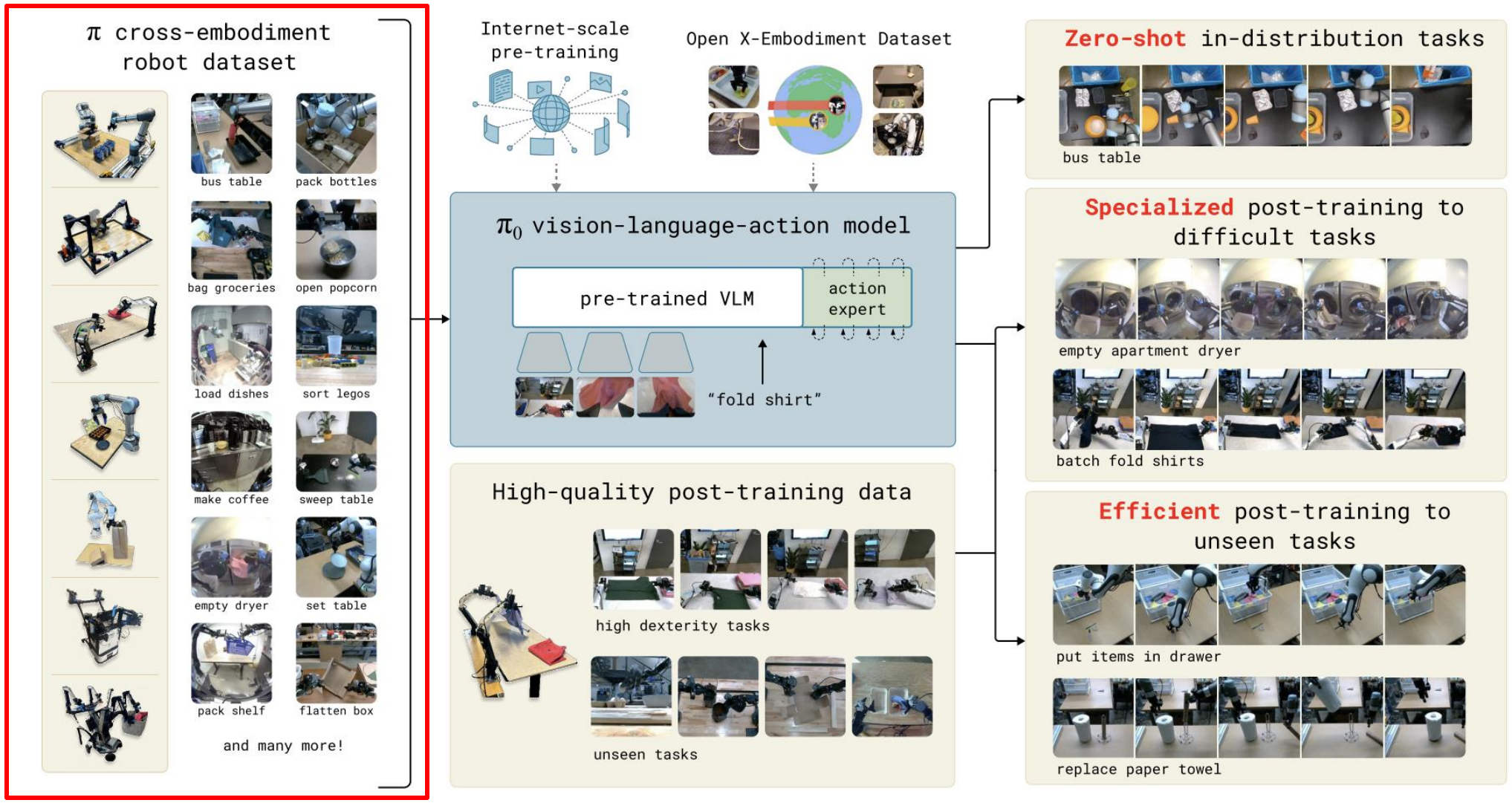
- First released in October 2024





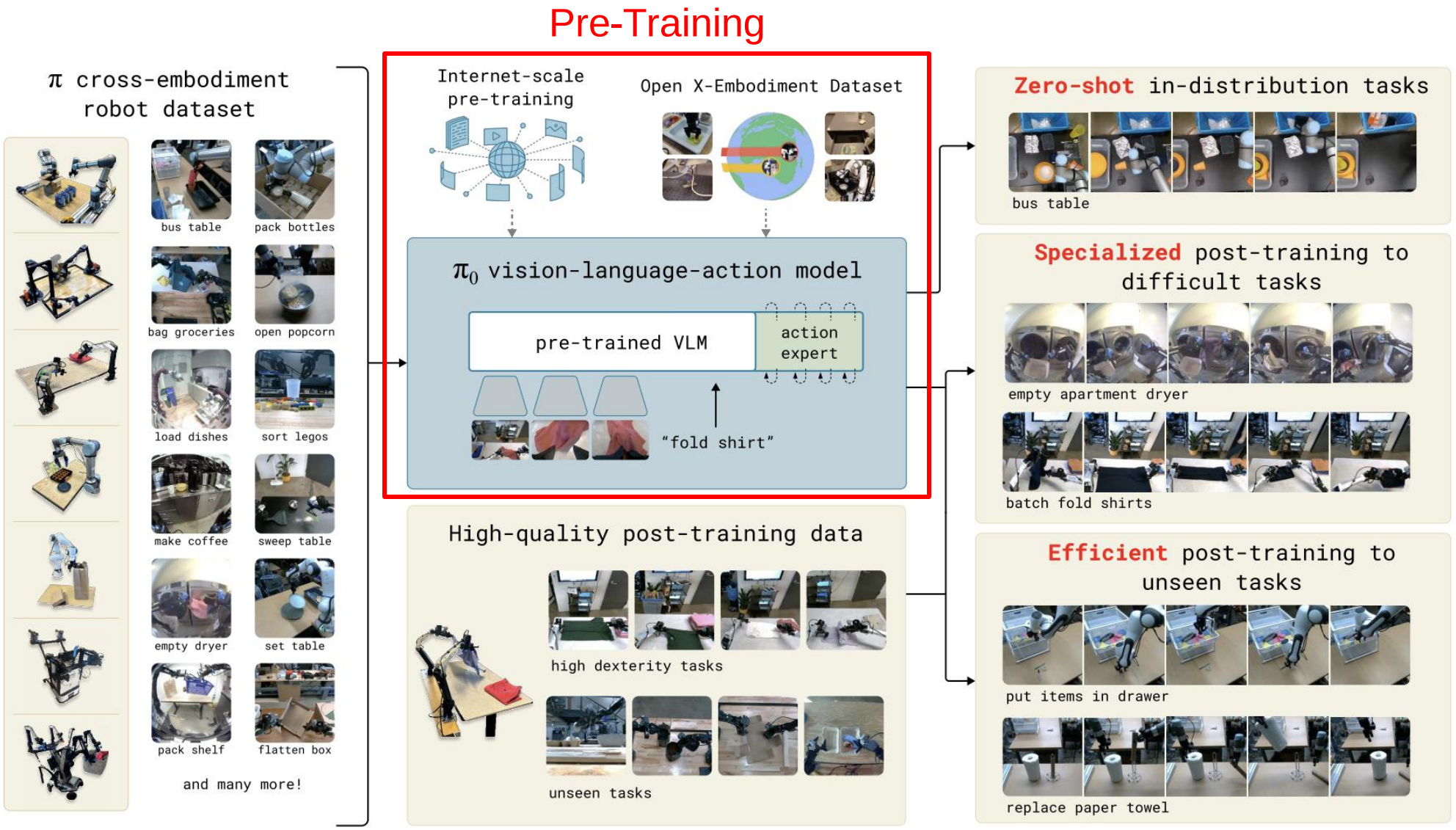


Pi-Zero by Physical Intelligence

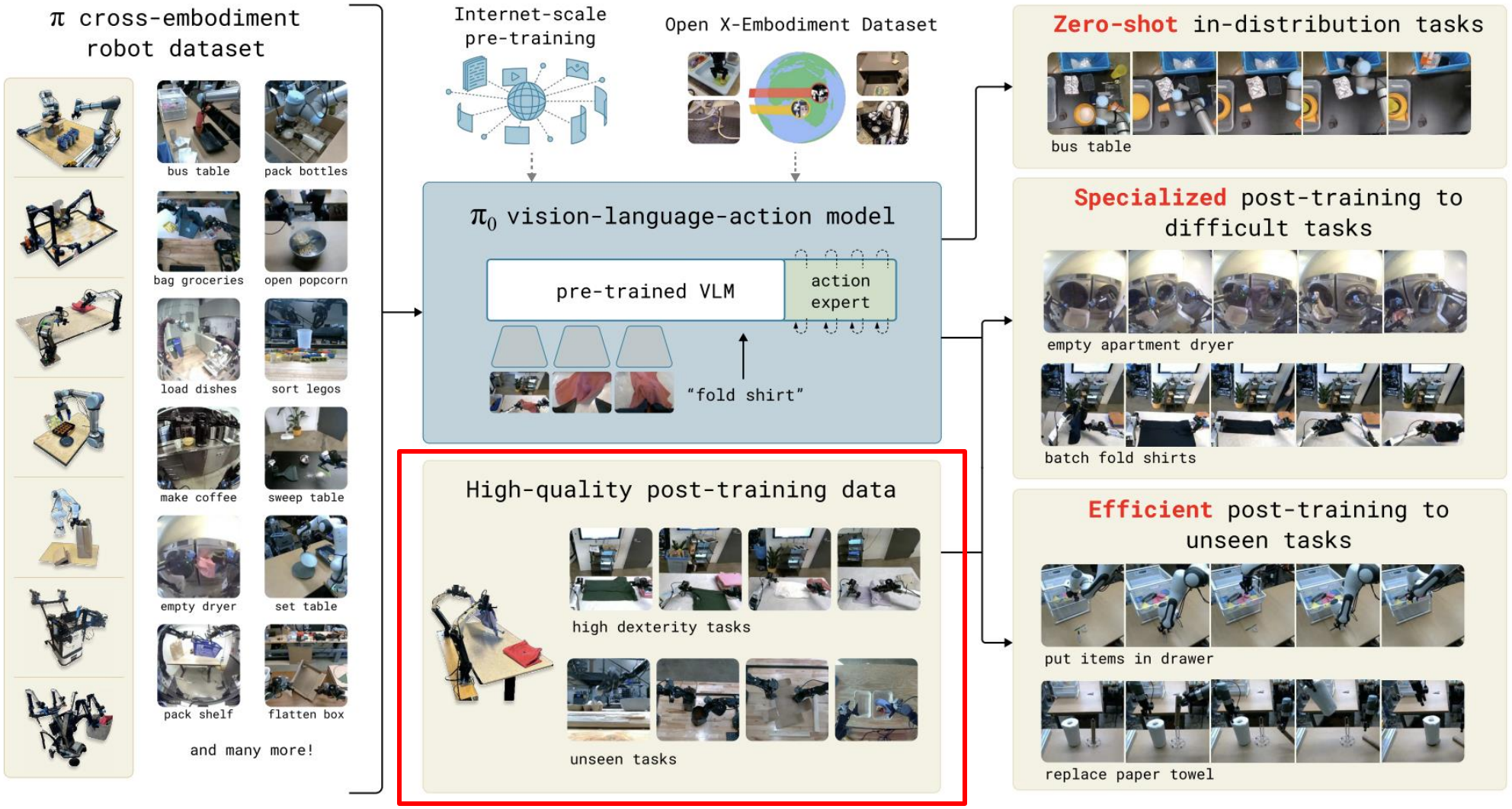


Cross-Embodiment Dataset

Pi-Zero by Physical Intelligence

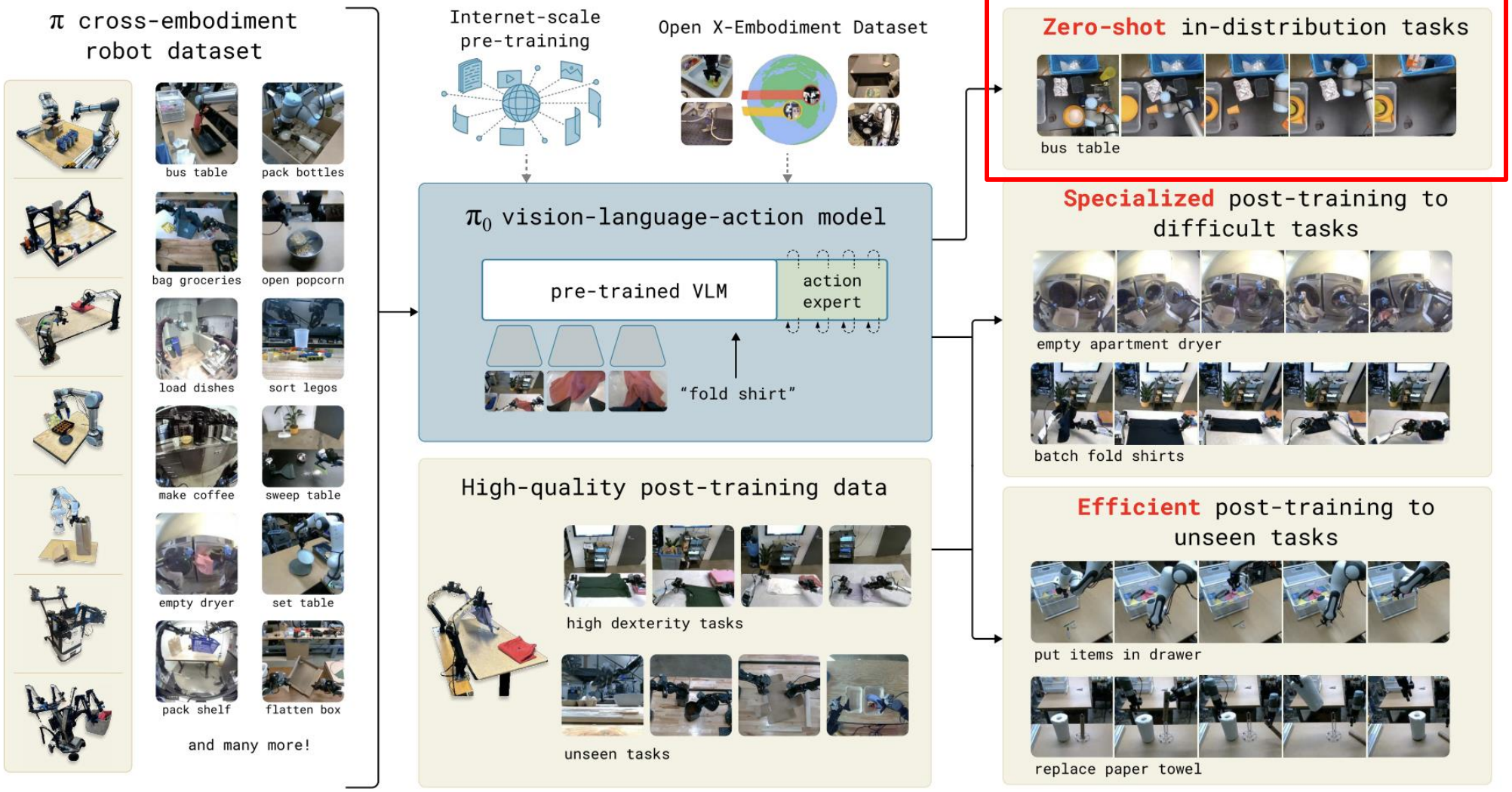


Pi-Zero by Physical Intelligence



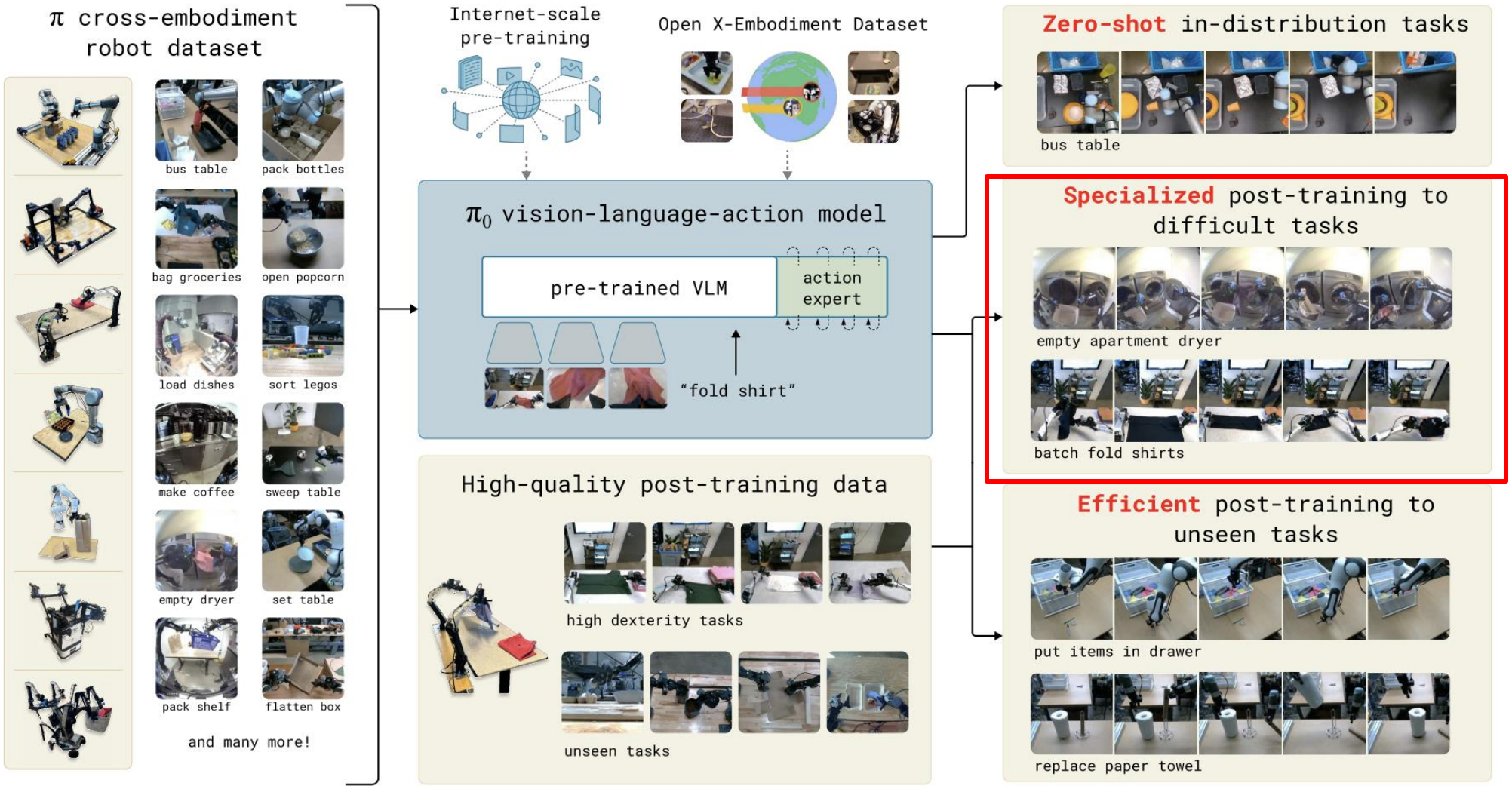
Pi-Zero by Physical Intelligence

Simple in-distribution tasks

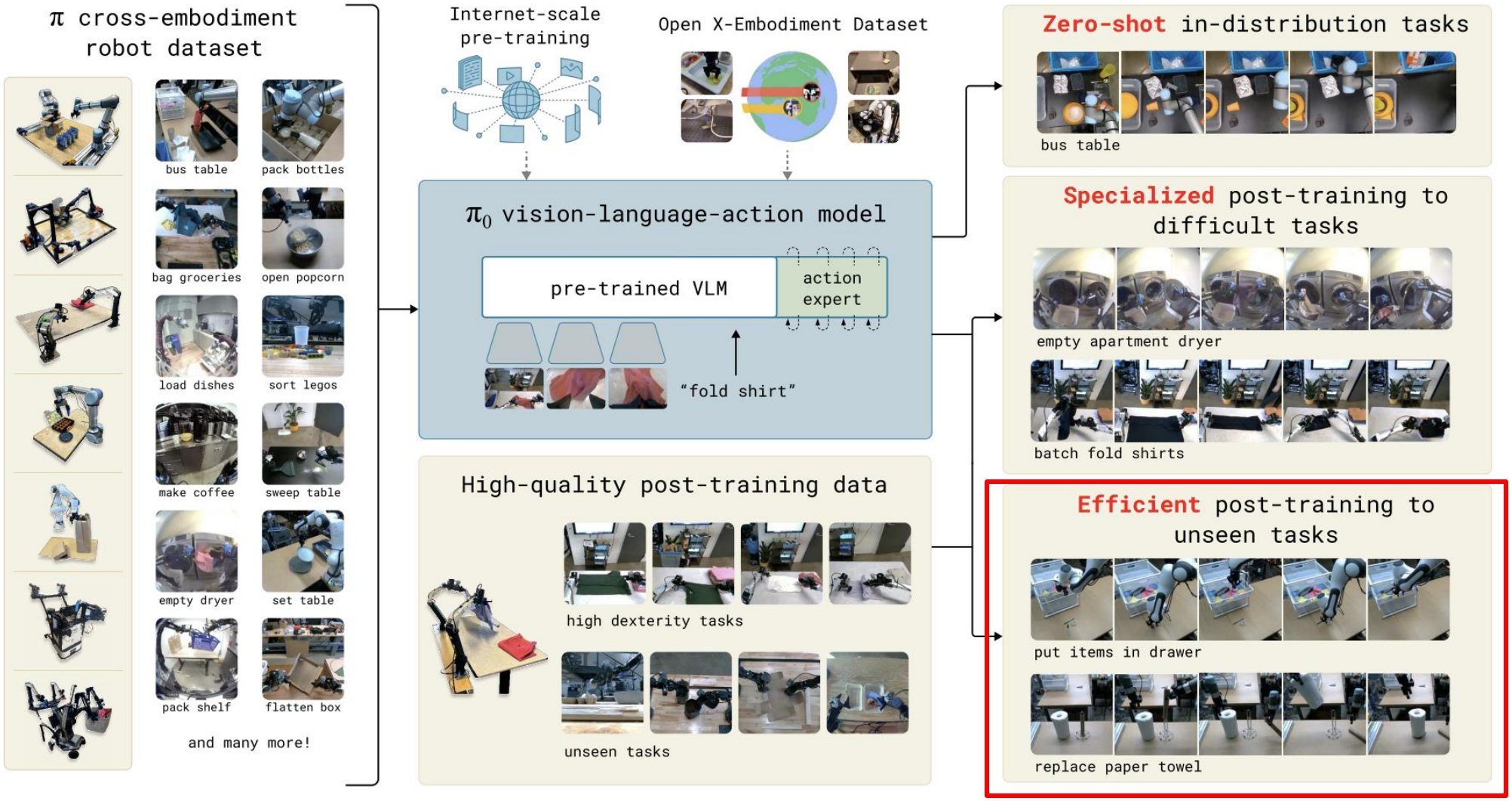


Pi-Zero by Physical Intelligence

Complicated in-distribution tasks



Pi-Zero by Physical Intelligence



Unseen tasks

Pi-Zero by Physical Intelligence

Physical Intelligence (π)

Open Sourcing π_0

Published	February 4, 2025
Email	research@physicalintelligence.company
Repo	Physical-Intelligence/openpi

README | Apache-2.0 license

openpi

openpi holds open-source models and packages for robotics, published by the Physical Intelligence team.

Currently, this repo contains two types of models:

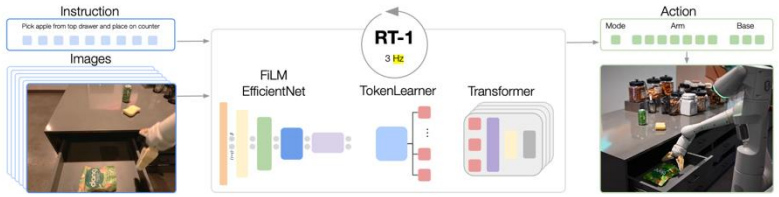
- the π_0 model, a flow-based diffusion vision-language-action model (VLA)
- the π_0-FAST model, an autoregressive VLA, based on the FAST action tokenizer.

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.

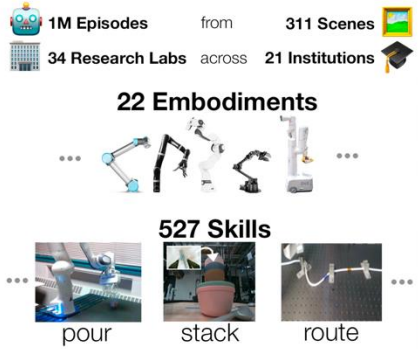
This is an experiment: π_0 was developed for our own robots, which differ from the widely used platforms such as ALOHA and DROID, and though we are optimistic that researchers and practitioners will be able to run creative new experiments adapting π_0 to their own platforms, we do not expect every such 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

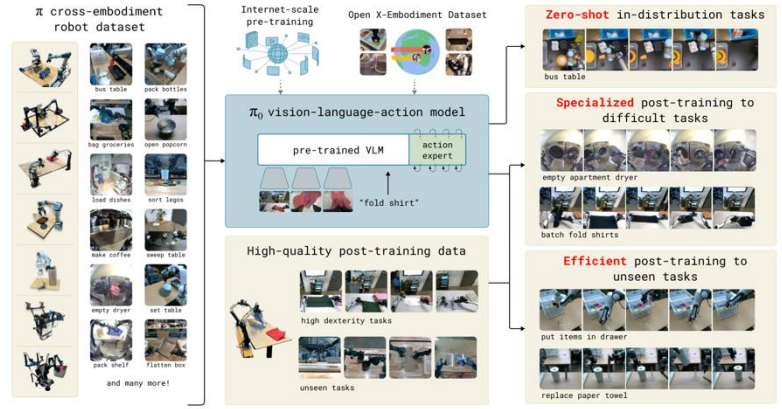
- What is a Robotic Foundation Model?
  - No explicit representation of states / transition functions
  - A policy that maps (observation/state, goal) to action



RT-1 (Dec. 2022)

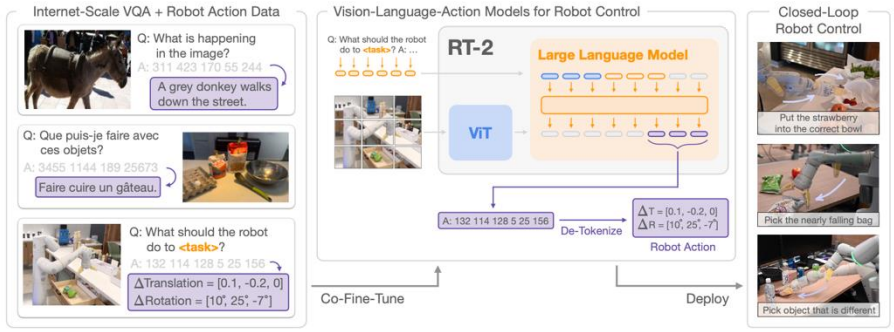


RT-X (Oct. 2023)



Pi-Zero (Oct. 2024)

RT-2 (Jul. 2023)



OpenVLA (Jun. 2024)