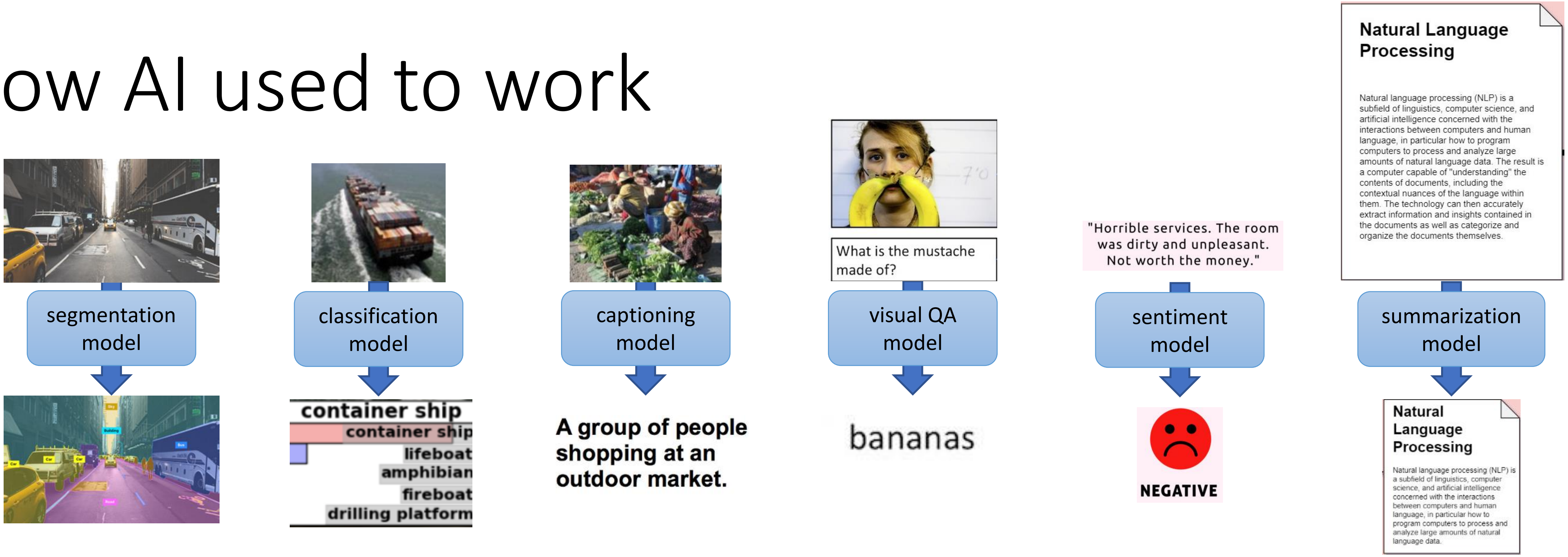
Robotic Foundation Models

Sergey Levine

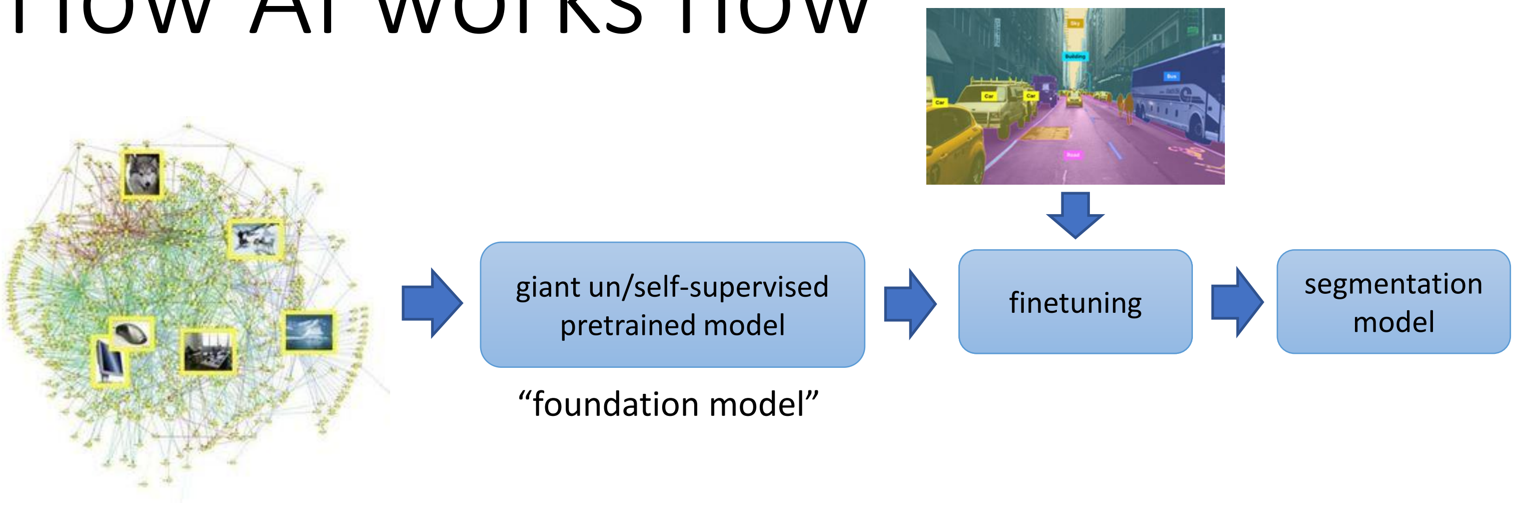
UC Berkeley

Physical Intelligence

How AI used to work



How AI works now



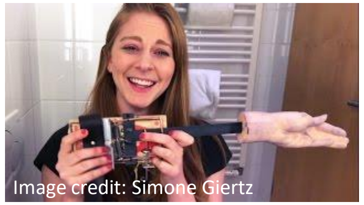
How robotic learning works now

PR2 pancake model

WAM pancake model

UR10 box picking model

lipstick robot model



mousetrap... hand... shake... model??

How robotic learning will work in the future

giant un/self-supervised pretrained **robot** model

"robot foundation model"

prompting or finetuning

trash sorting model

What do we need to figure out?

giant un/self-supervised pretrained **robot** model

"robot foundation model"

prompting or finetuning

trash sorting model

What kind of data?

what kind of model should we use?

what kind of algorithm or objective?

What type of representations do we hope to get?

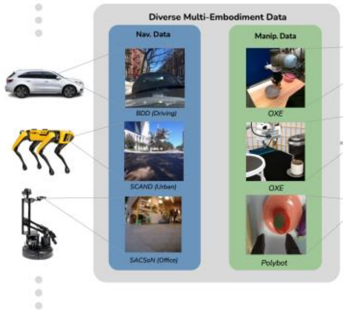
How do we adapt it to new tasks?

How different can the new tasks be from pretraining?

How do we build robotic foundation models?

Robotic foundation models for navigation

Manipulation, VLAs, and open-source models

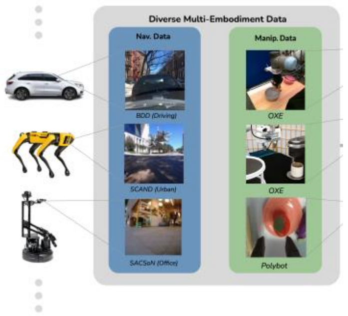


Taking cross-embodied learning to the limit

How do we build robotic foundation models?

Robotic foundation models for navigation

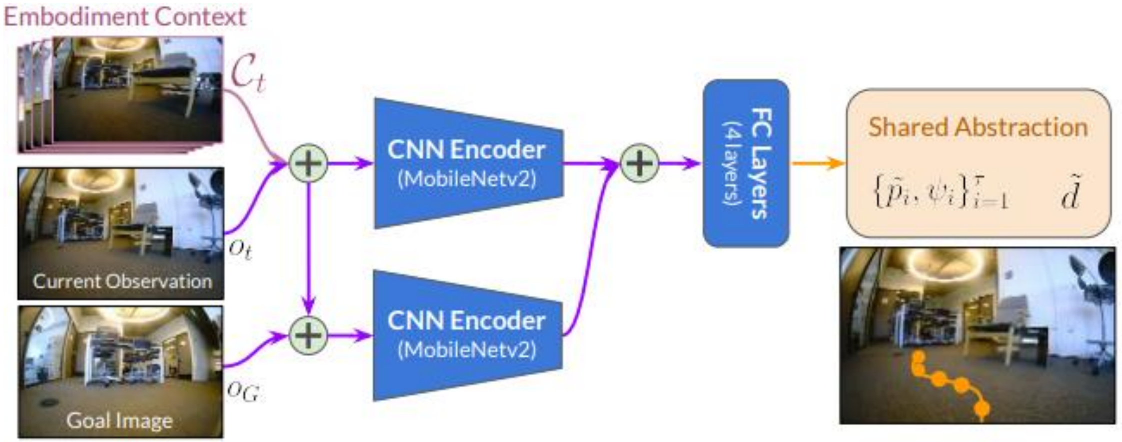
Manipulation, VLAs, and open-source models

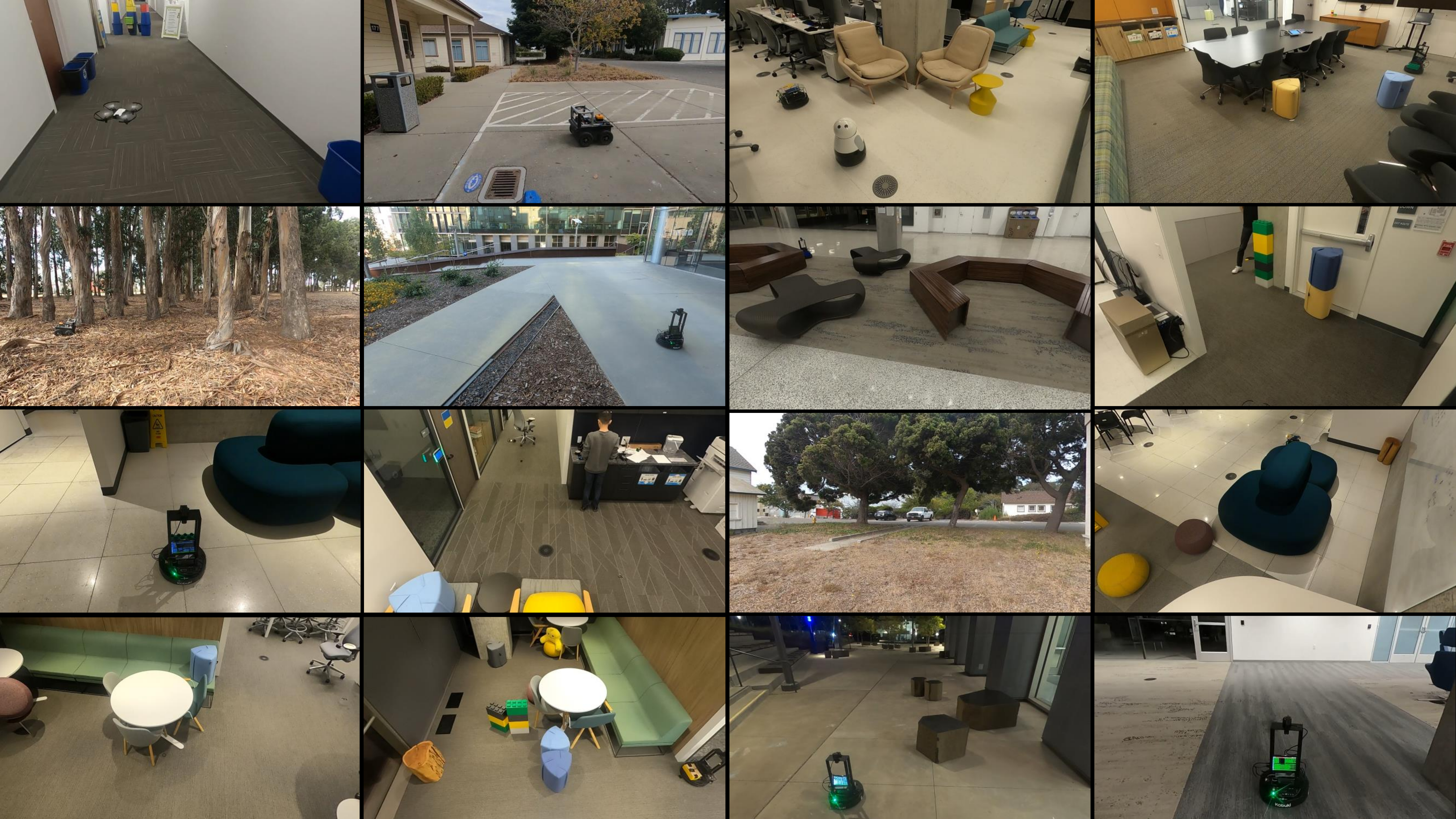


Taking cross-embodied learning to the limit

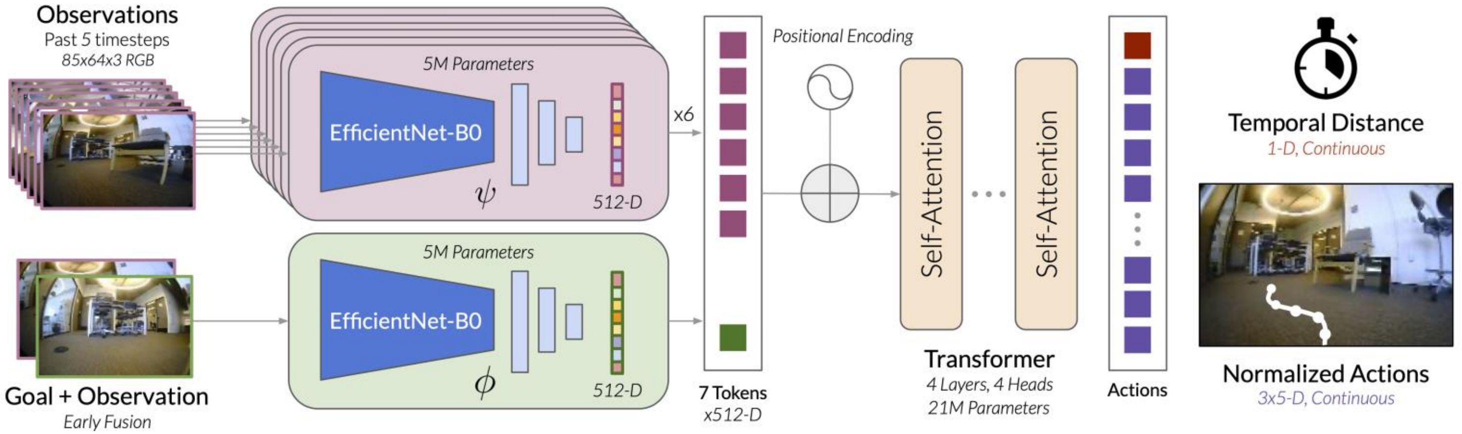
Robotic foundation models for navigation

	Dataset	Platform	Speed	Amt.	Environment
1	GoStanford [26]	TurtleBot2	0.5m/s	14h	office
2	RECON [32]	Jackal	1m/s	25h	off-road
3	CoryHall [35]	RC Car	1.2m/s	2h	hallways
4	Berkeley [33]	Jackal	2m/s	4h	suburban
5	SCAND-S [36]	Spot	1.5m/s	8h	sidewalks
6	SCAND-J [36]	Jackal	2m/s	1h	sidewalks
7	Seattle [37]	Warthog	5m/s	1h	off-road
8	TartanDrive [38]	ATV	10m/s	5h	off-road
	Ours			60h	



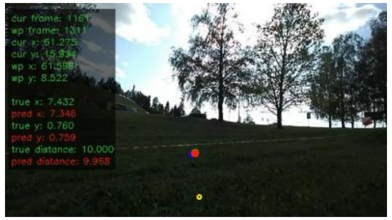


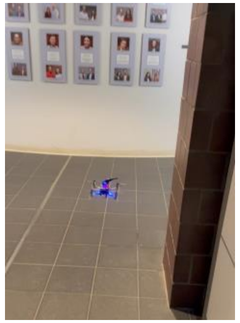
Scaling it up with Transformers

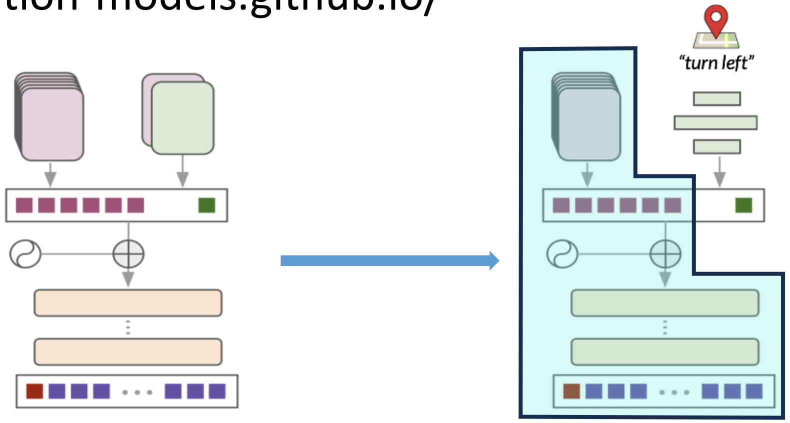


ViNT: Visual Navigation Transformer

https://general-navigation-models.github.io/





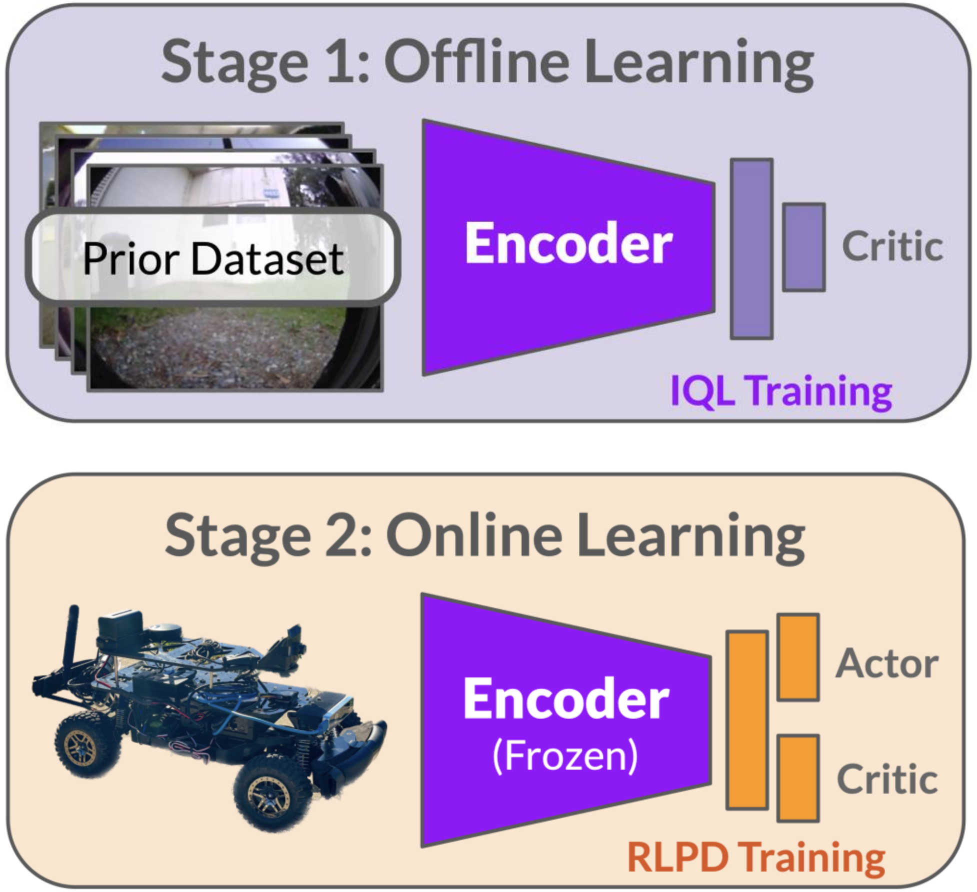


ViNT Foundation Model

Pre-trained ViNT

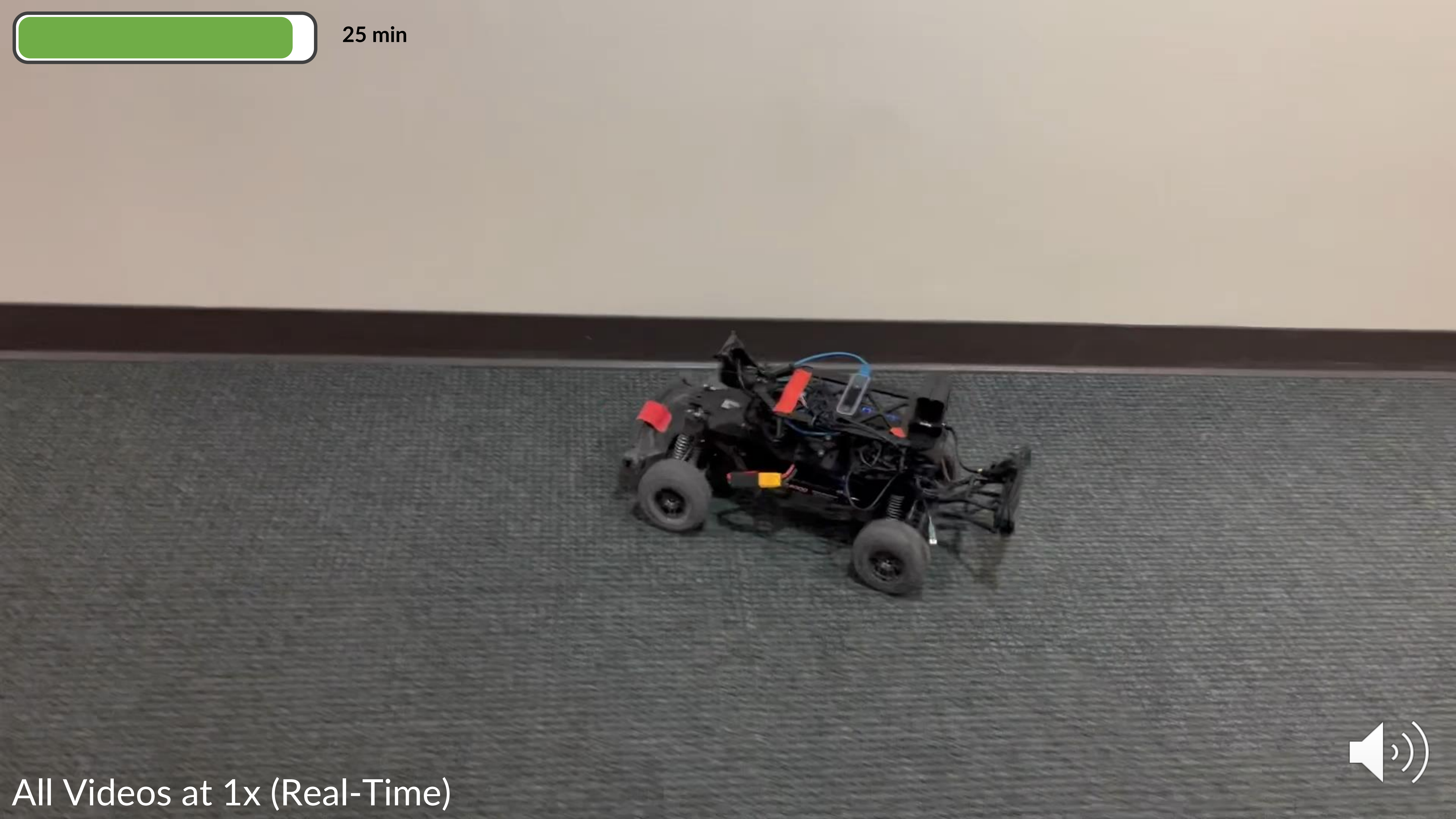
Shah, Sridhar, Dashora, Stachowicz, Black, Hirose, Levine. **ViNT: A Foundation Model for Visual Navigation.** 2023.

Now make it go fast!



Stachowicz, Shah, Bhorkar, Kostrikov, Levine. **FastRLAP: A System for Learning High-Speed Driving via Deep RL and Autonomous Practicing.** 2023.


15 min
All Videos at 1x (Real-Time)


25 min
All Videos at 1x (Real-Time)

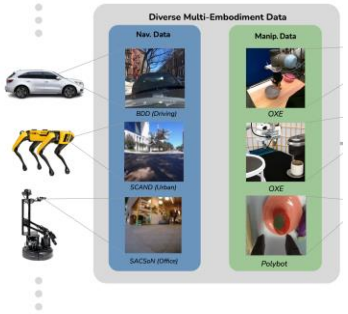

25 min
EXIT
All Videos at 1x (Real-Time)
HD Footage (not robot’s camera)


164

How do we build robotic foundation models?

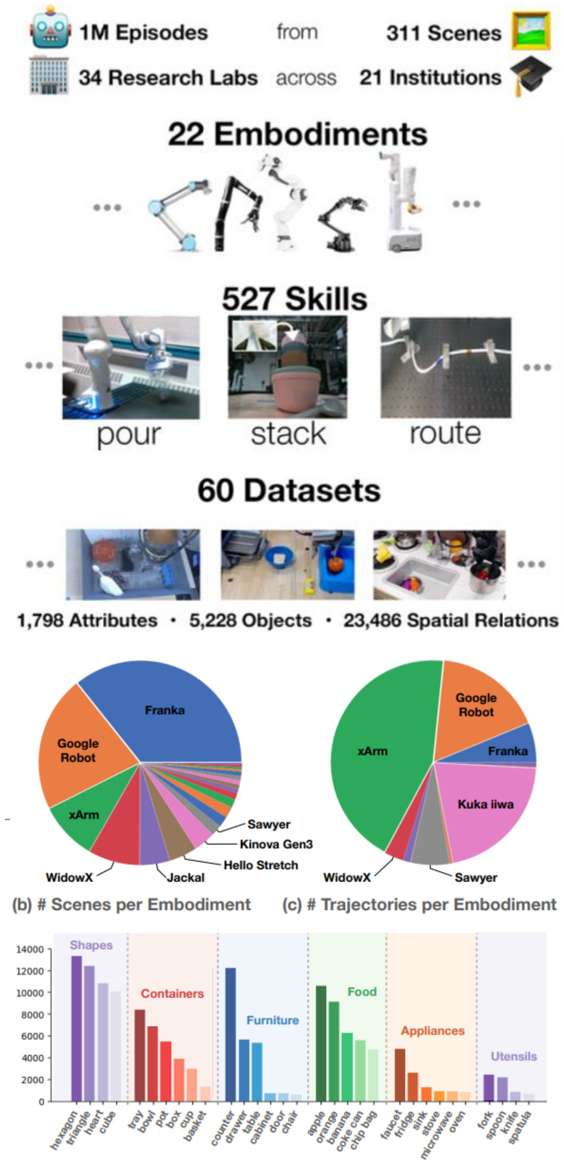
Robotic foundation models for navigation

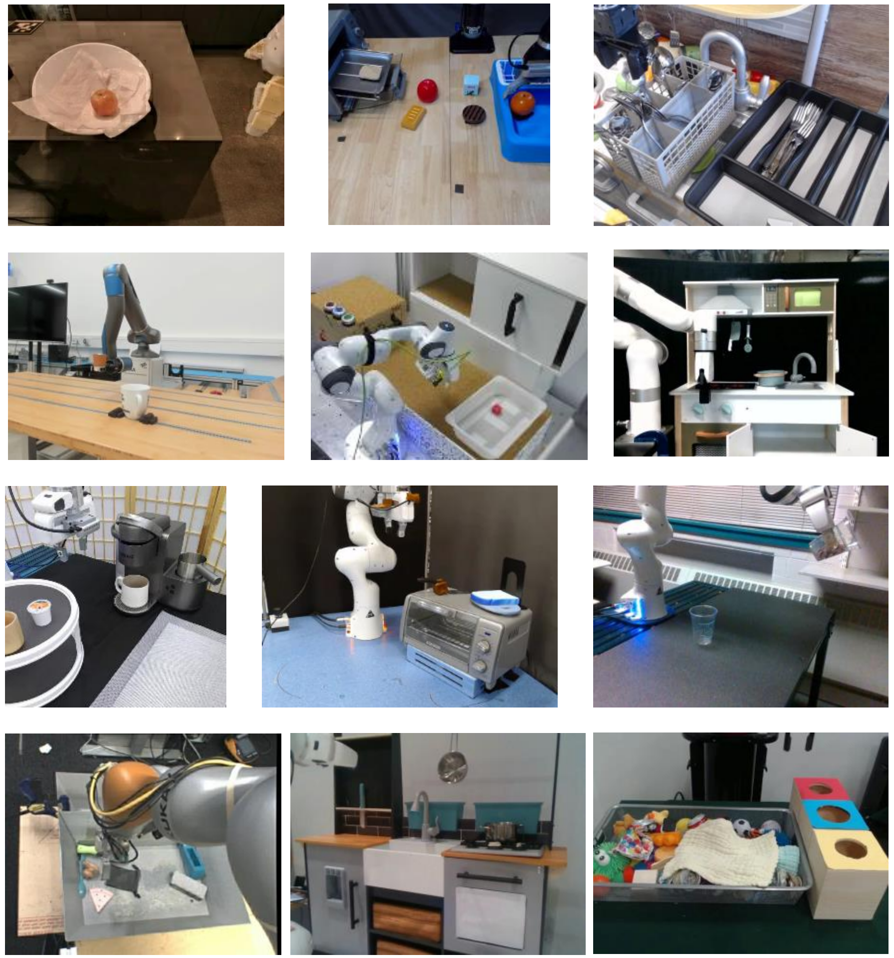
Manipulation, VLAs, and open-source models



Taking cross-embodied learning to the limit

RT-X: Combining many datasets for cross-embodiment manipulation

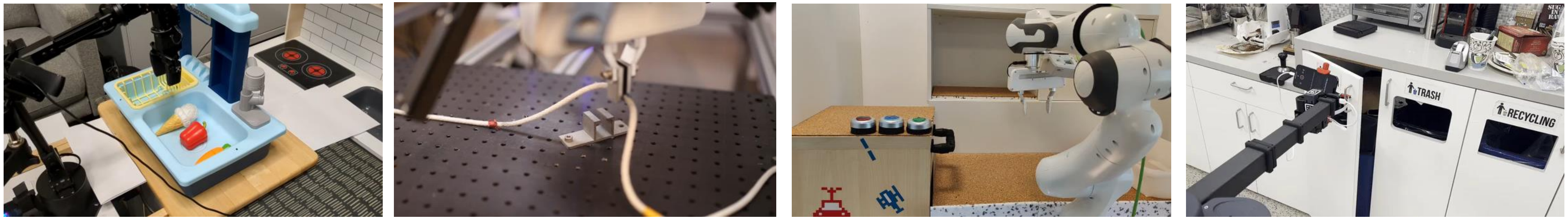


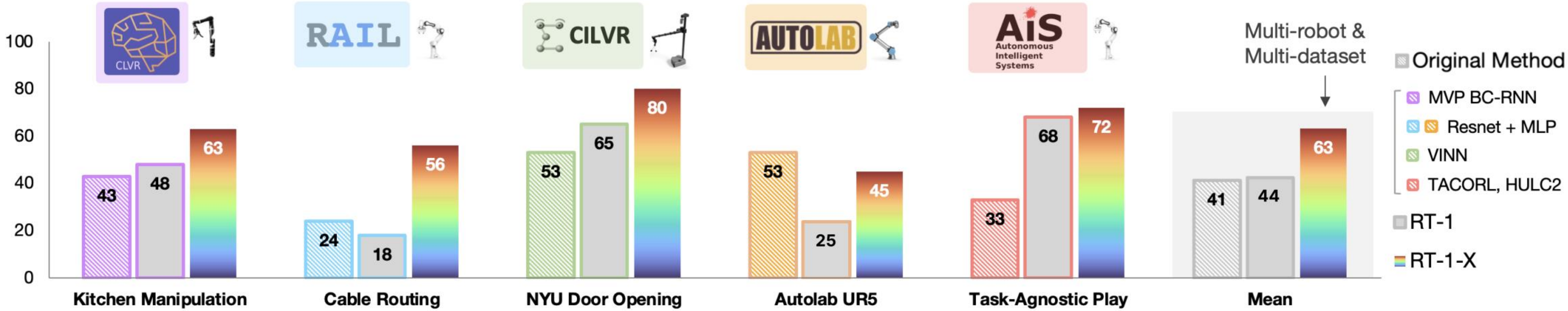


Visualizations: https://dibyaghosh.com/rtx_viz/
By Dibya Ghosh

RT-X: https://robotics-transformer-x.github.io/
Open X-Embodiment Collaboration

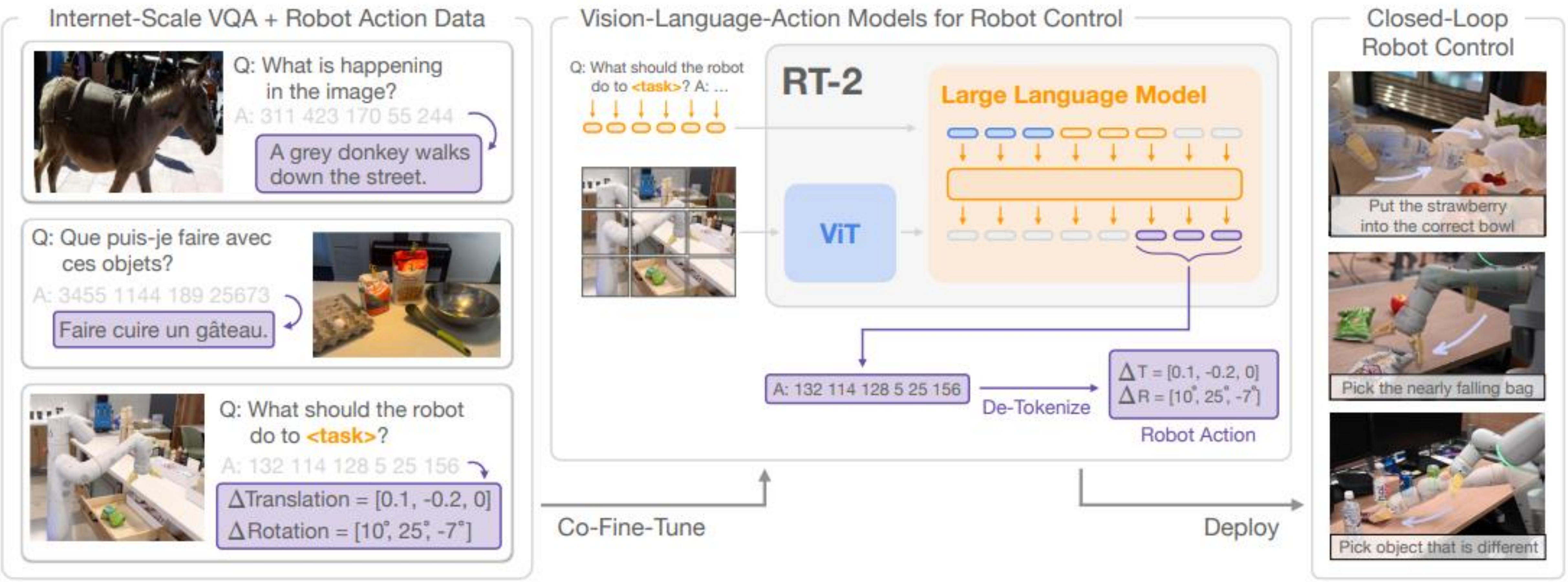
RT-X: Combining many datasets for cross-embodiment manipulation





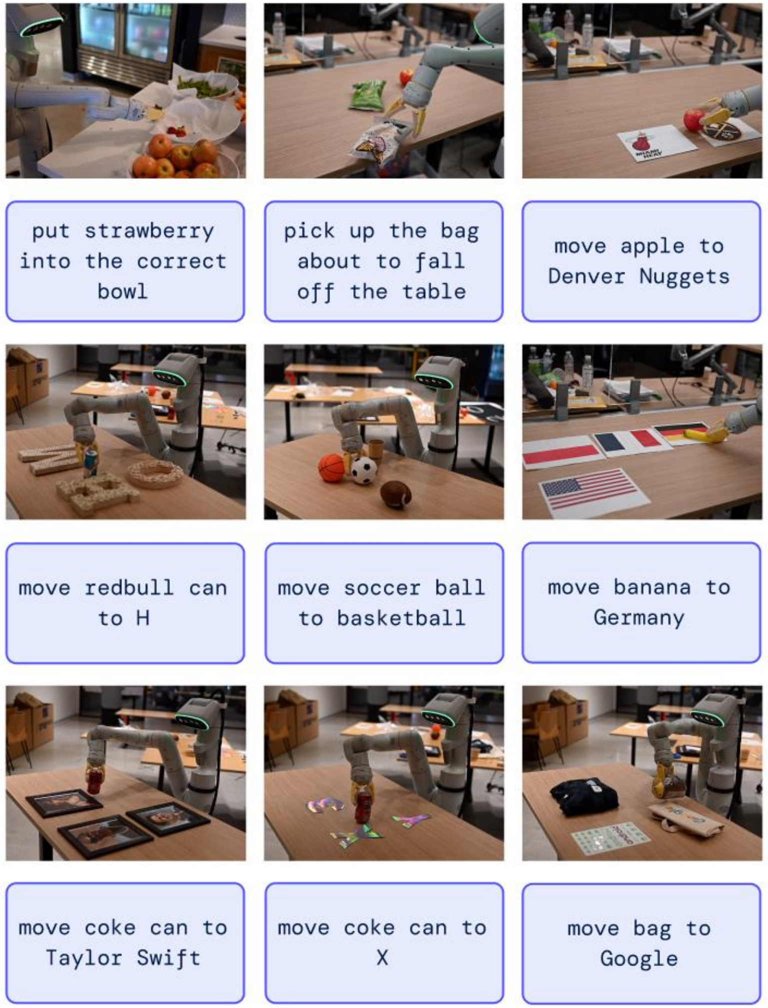
RT-X: https://robotics-transformer-x.github.io/
Open X-Embodiment Collaboration

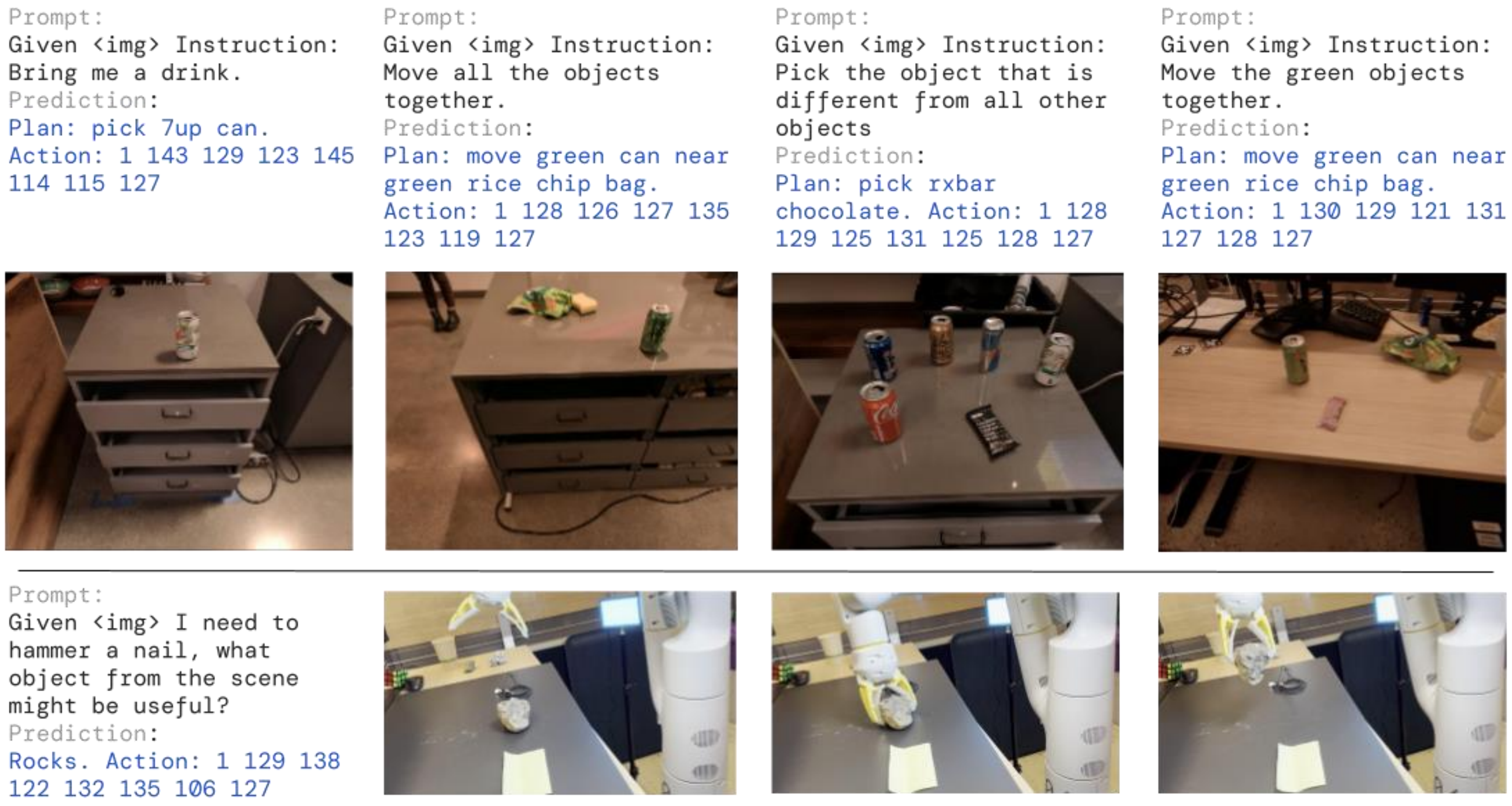
RT-2: A vision-language-action (VLA) model



RT-2: https://robotics-transformer2.github.io/
Fractal Team

RT-2: A vision-language-action (VLA) model





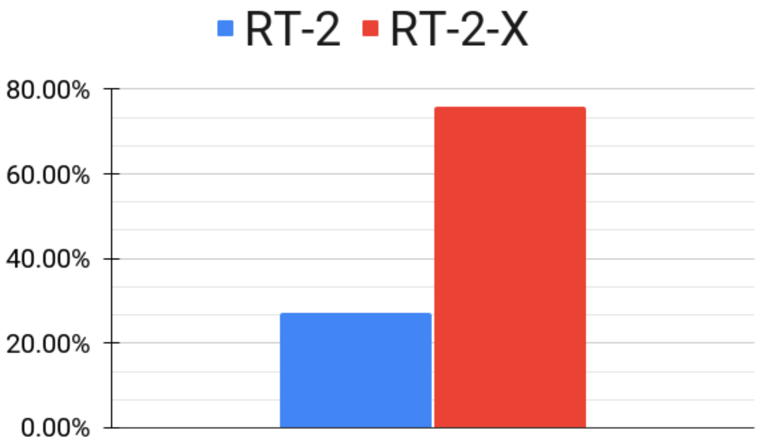
RT-2: https://robotics-transformer2.github.io/

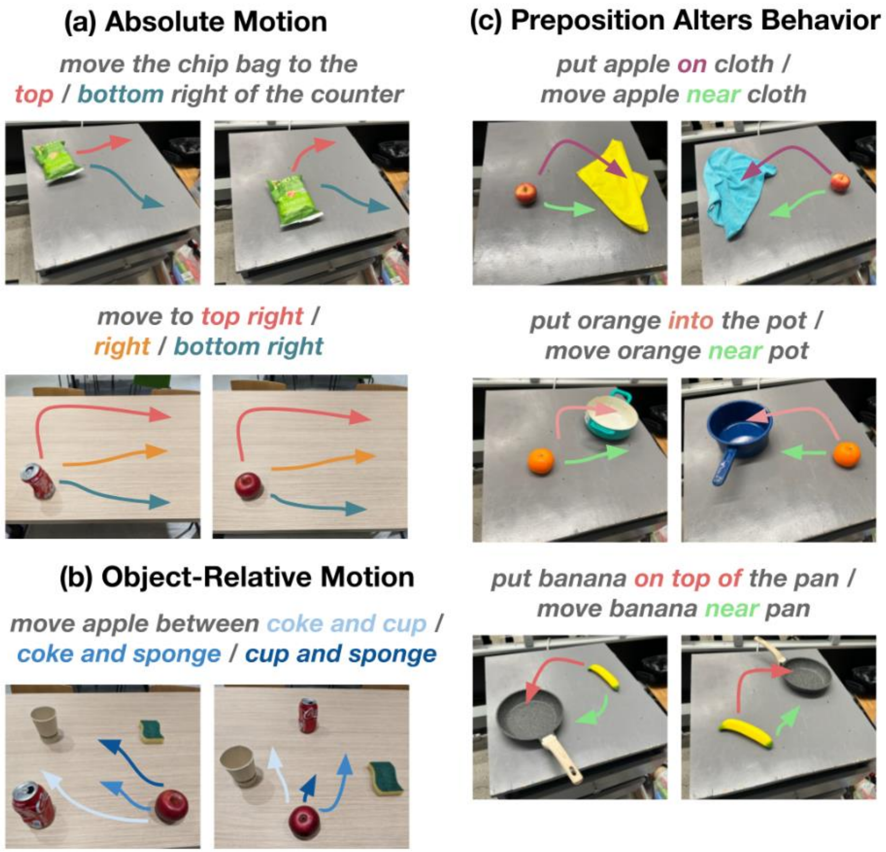
RT-2-X: Does cross embodiment training help VLAs?

“move apple between can & orange”

“move apple near cloth”

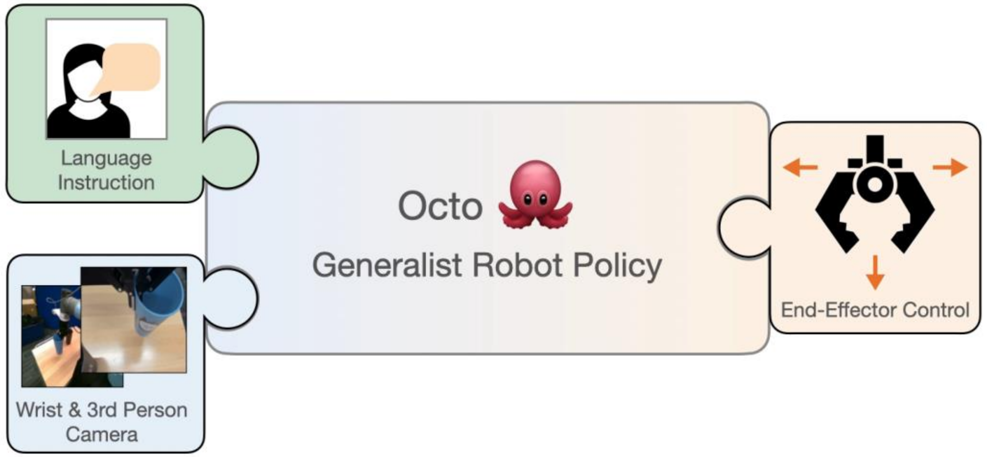
“move apple on cloth”

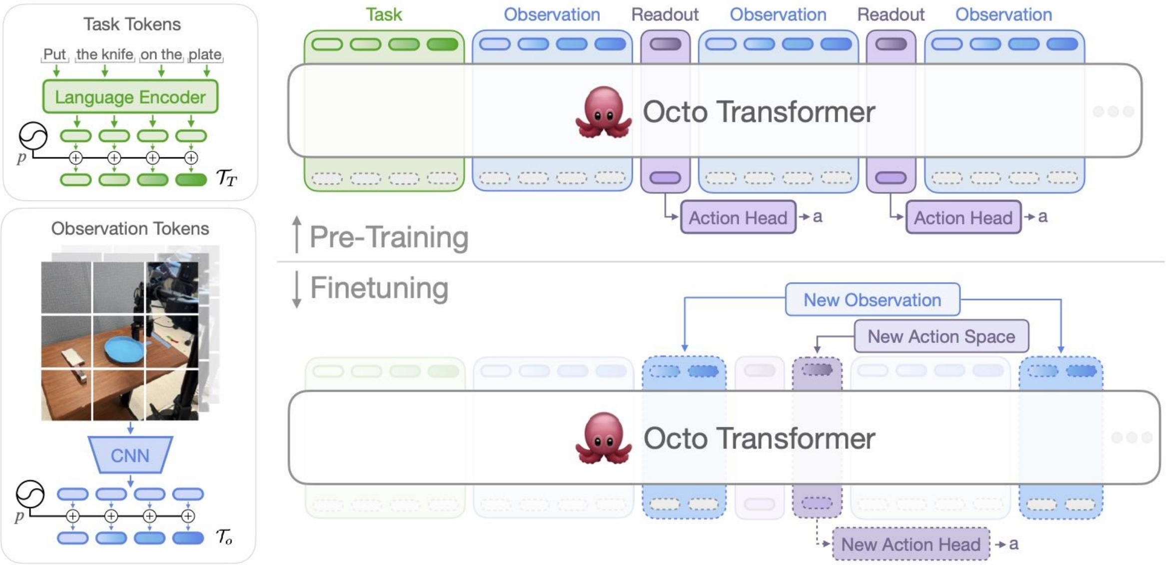




RT-X: https://robotics-transformer-x.github.io/
Open X-Embodiment Collaboration

Octo: an open-source robotic foundation model

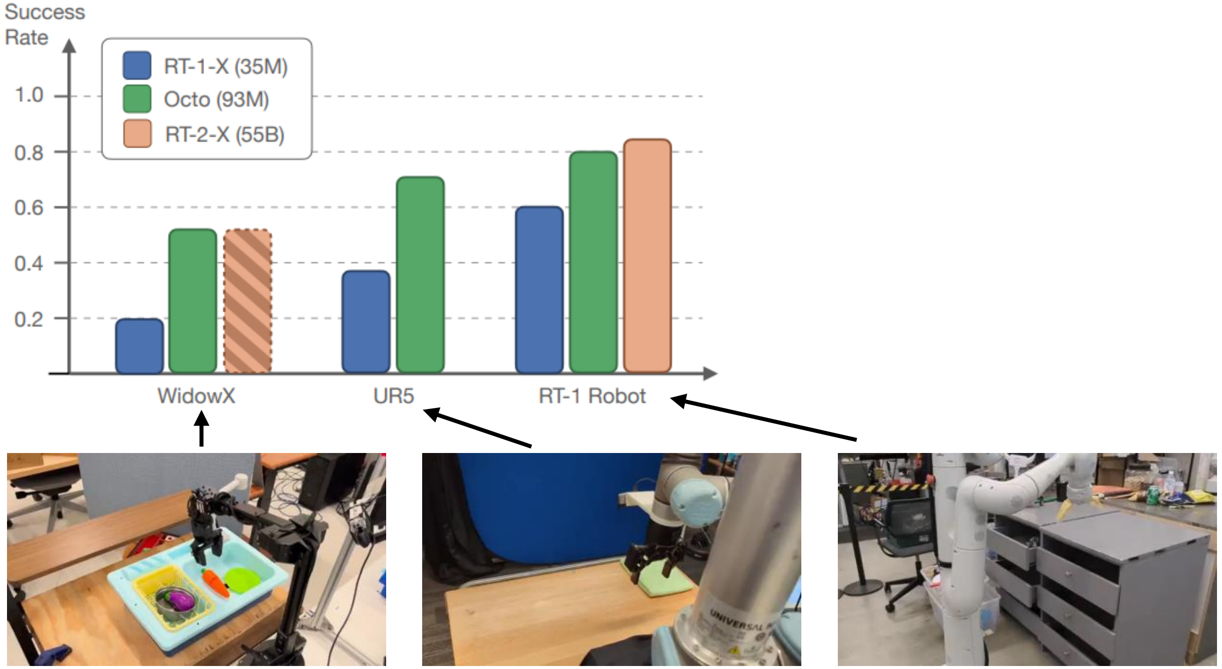




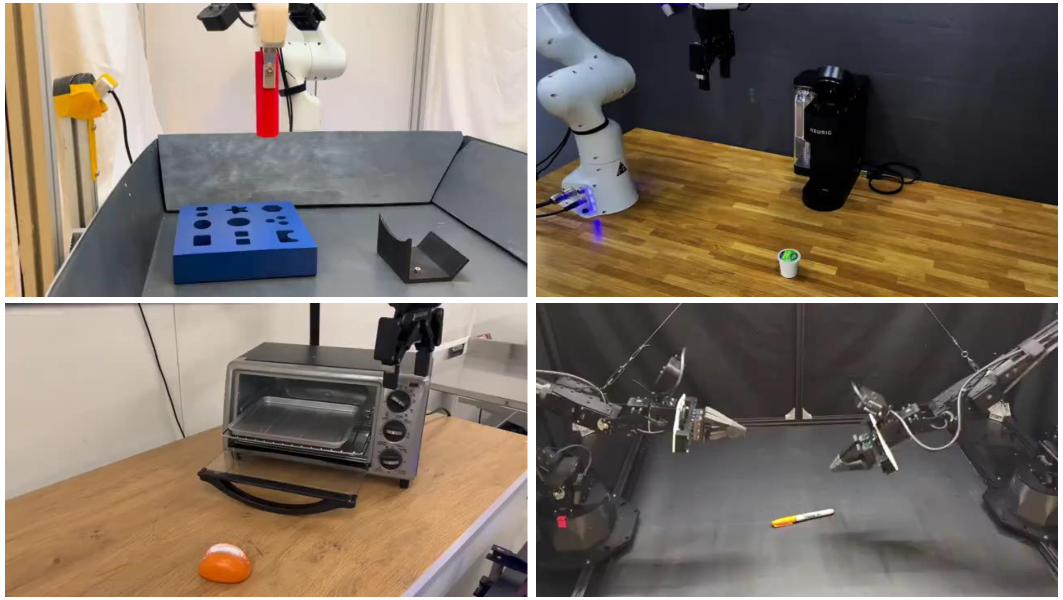
Octo: https://octo-models.github.io/
Ghosh*, Walke*, Pertsch*, Black*, Mees*, et al.

Octo: an open-source robotic foundation model

zero-shot evaluation



Berkeley Insertion
Stanford Coffee



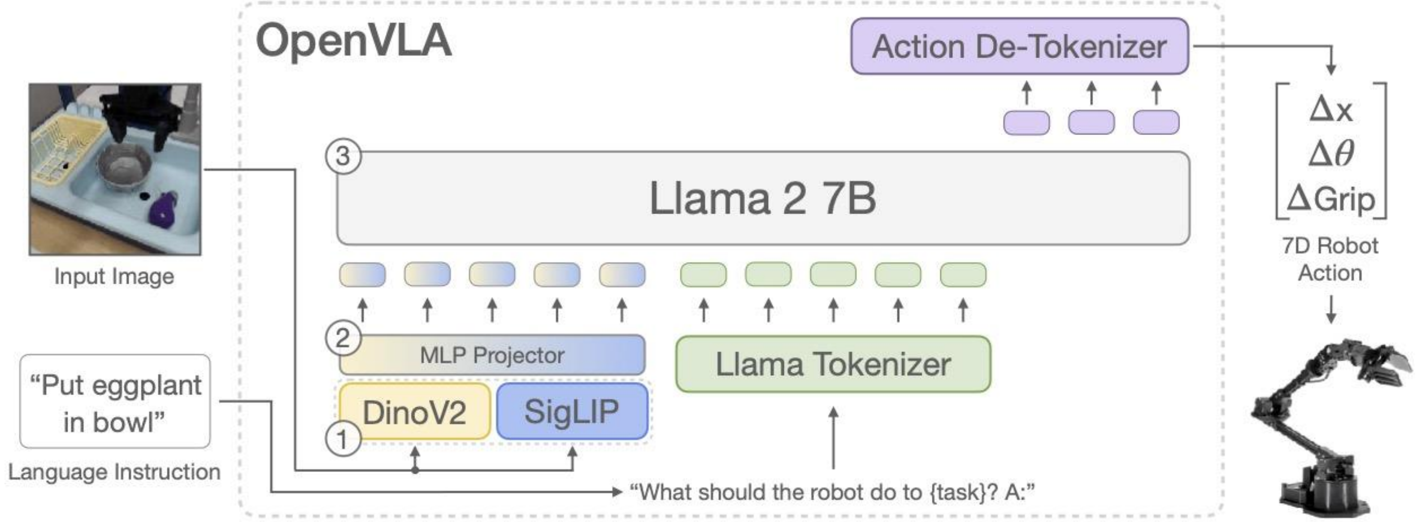
CMU Baking
Berkeley Bimanual

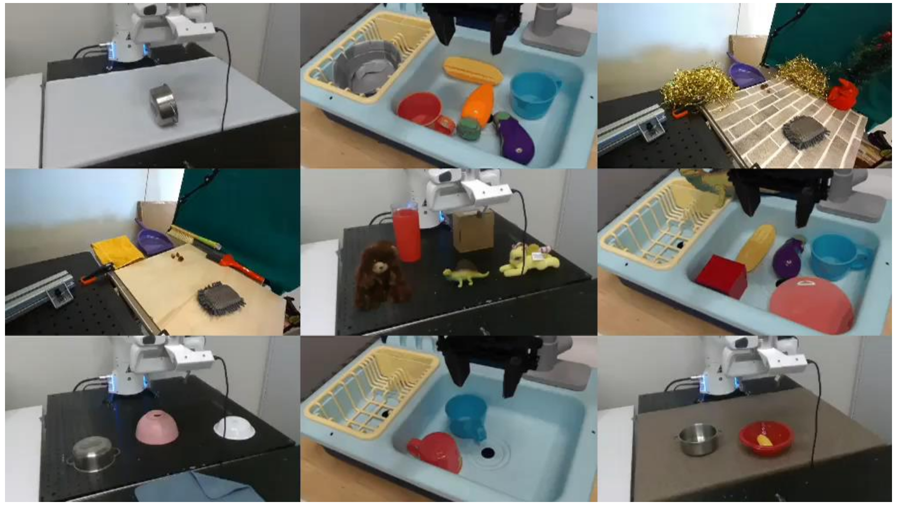
finetuning

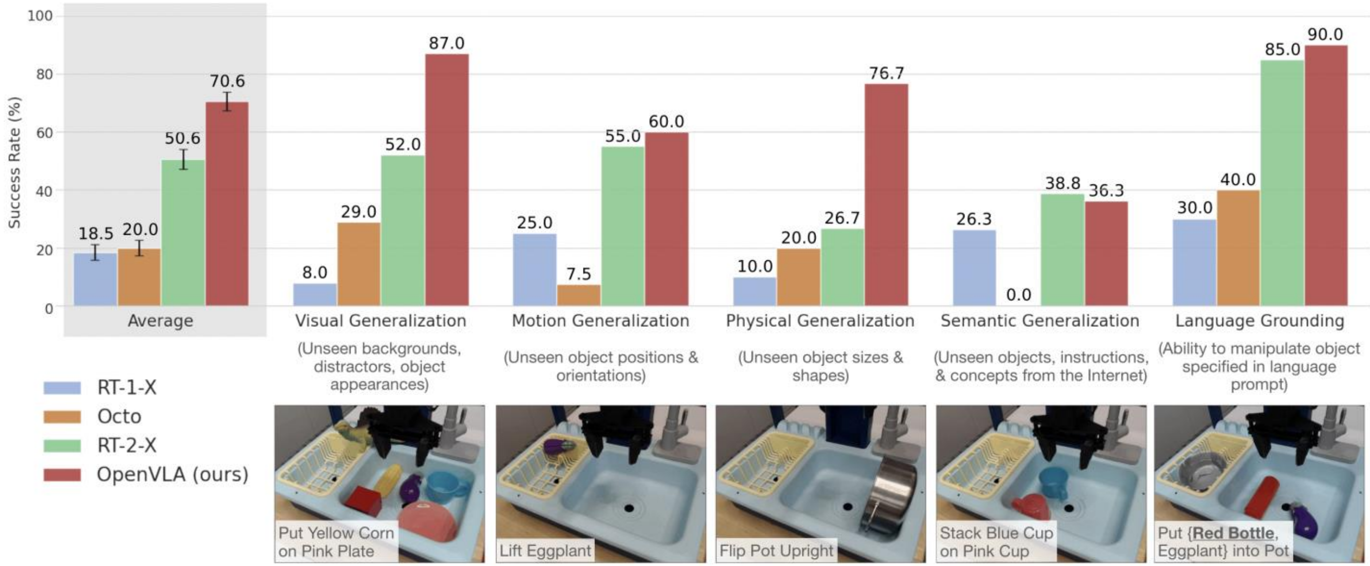
	Berkeley Insertion*	Stanford Coffee	CMU Baking	Berkeley Pick-Up†	Berkeley Coke	Berkeley Bimanual†	Average
ResNet+Transformer Scratch	10%	45%	25%	0%	20%	20%	20%
VC-1 [57]	5%	0%	30%	0%	10%	50%	15%
Octo (Ours)	**70%**	**75%**	**50%**	**60%**	**100%**	**80%**	**72%**

Octo: https://octo-models.github.io/
Ghosh*, Walke*, Pertsch*, Black*, Mees*, et al.

OpenVLA: an open-source **vision-language-action** model





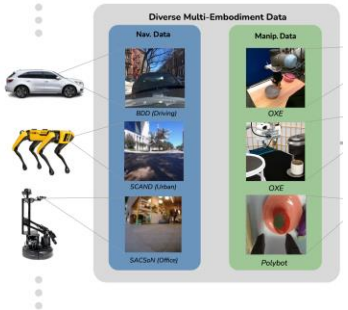


OpenVLA: https://openvla.github.io/
Kim*, Pertsch*, Karamcheti*, et al.

How do we build robotic foundation models?

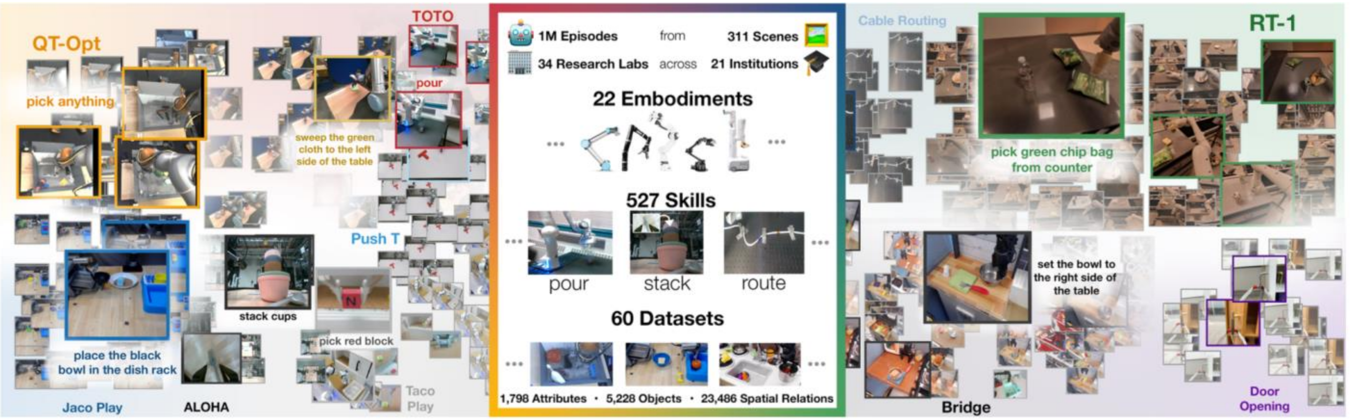
Robotic foundation models for navigation

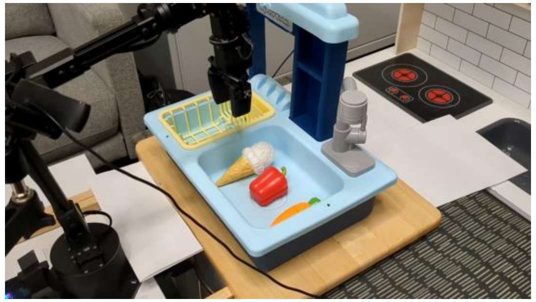
Manipulation, VLAs, and open-source models

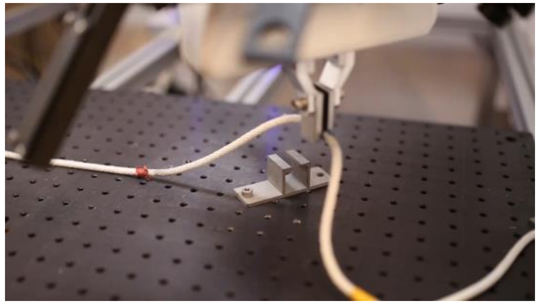


Taking cross-embodied learning to the limit

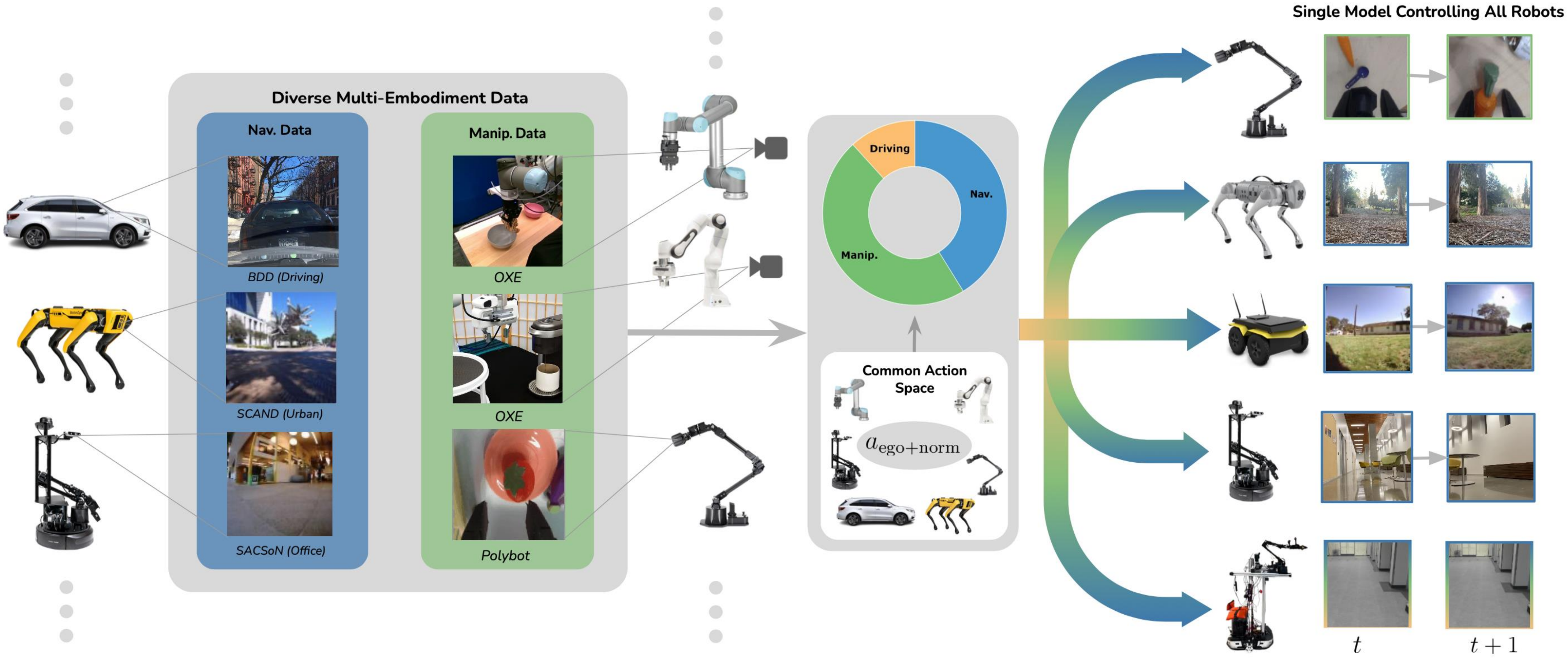
How diverse can the data be?





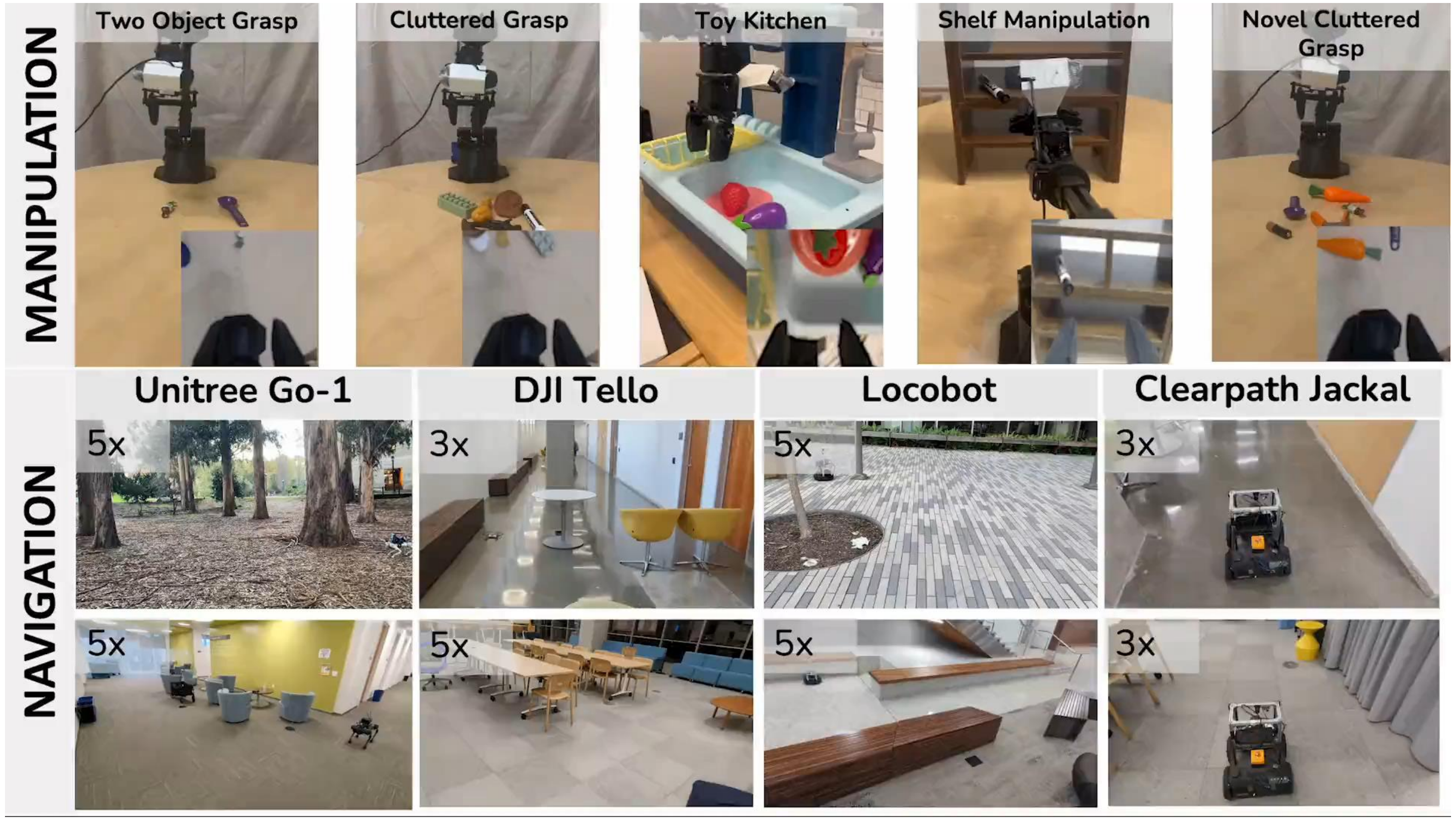


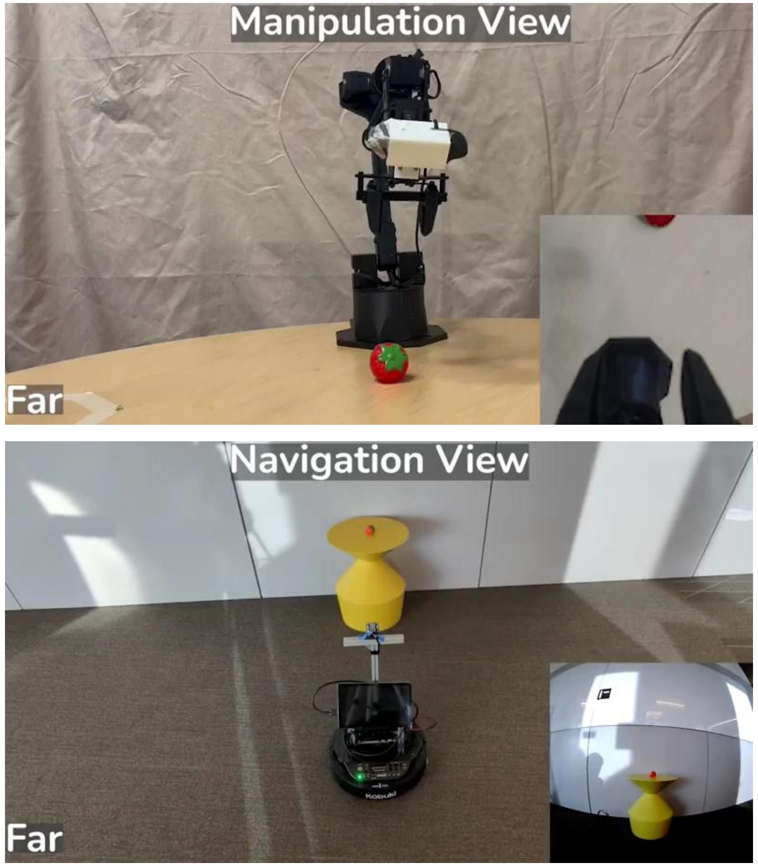
An "extreme" cross-embodiment recipe



Yang, Glossop, Bhorkar, Shah, Vuong, Finn, Sadigh, Levine. **Extreme Cross-Embodiment Learning for Manipulation and Navigation.** 2024.

Why might this work?

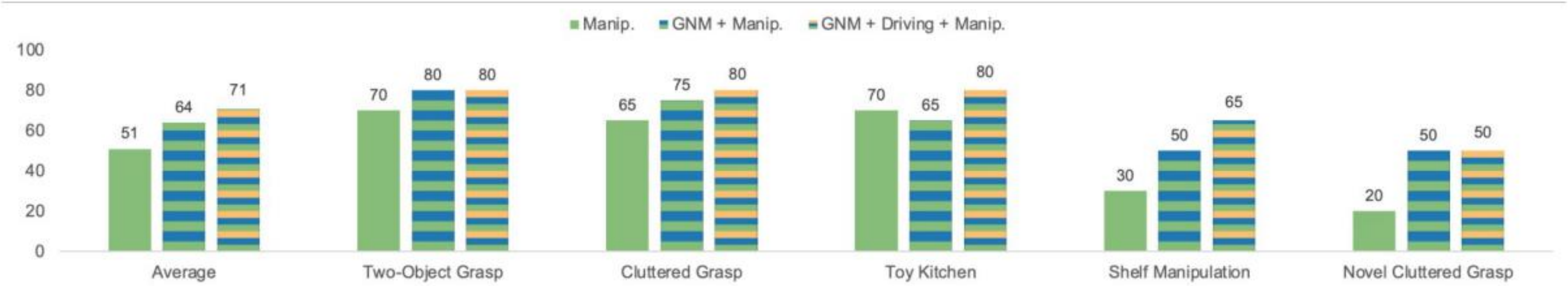




Yang, Glossop, Bhorkar, Shah, Vuong, Finn, Sadigh, Levine. **Extreme Cross-Embodiment Learning for Manipulation and Navigation.** 2024.

Some results

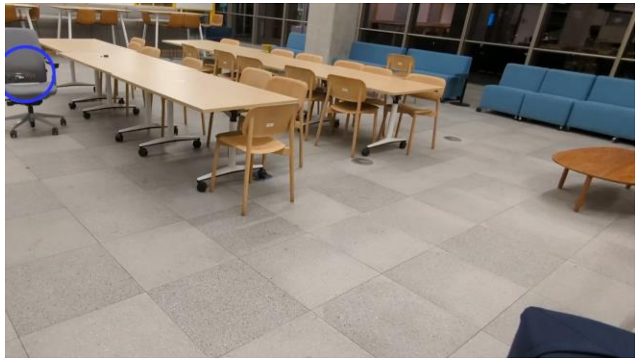
Does navigation help with manipulation?

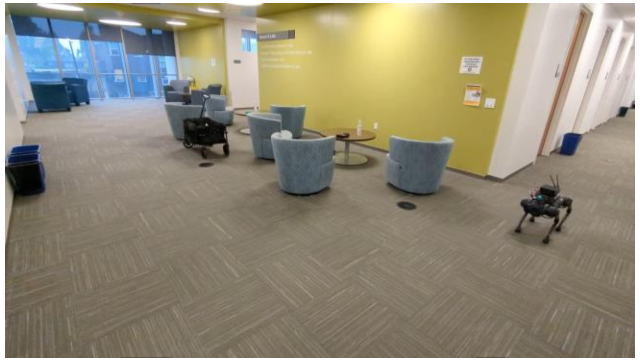


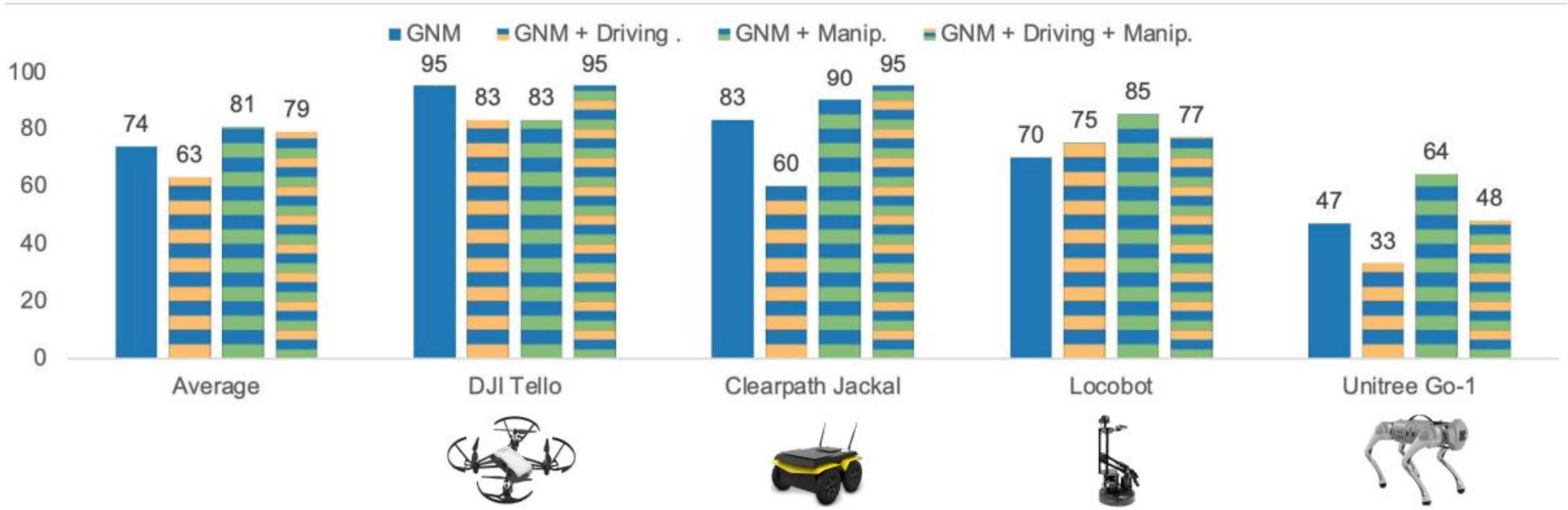
Yang, Glossop, Bhorkar, Shah, Vuong, Finn, Sadigh, Levine. **Extreme Cross-Embodiment Learning for Manipulation and Navigation.** 2024.

Some results

Does manipulation help with navigation?

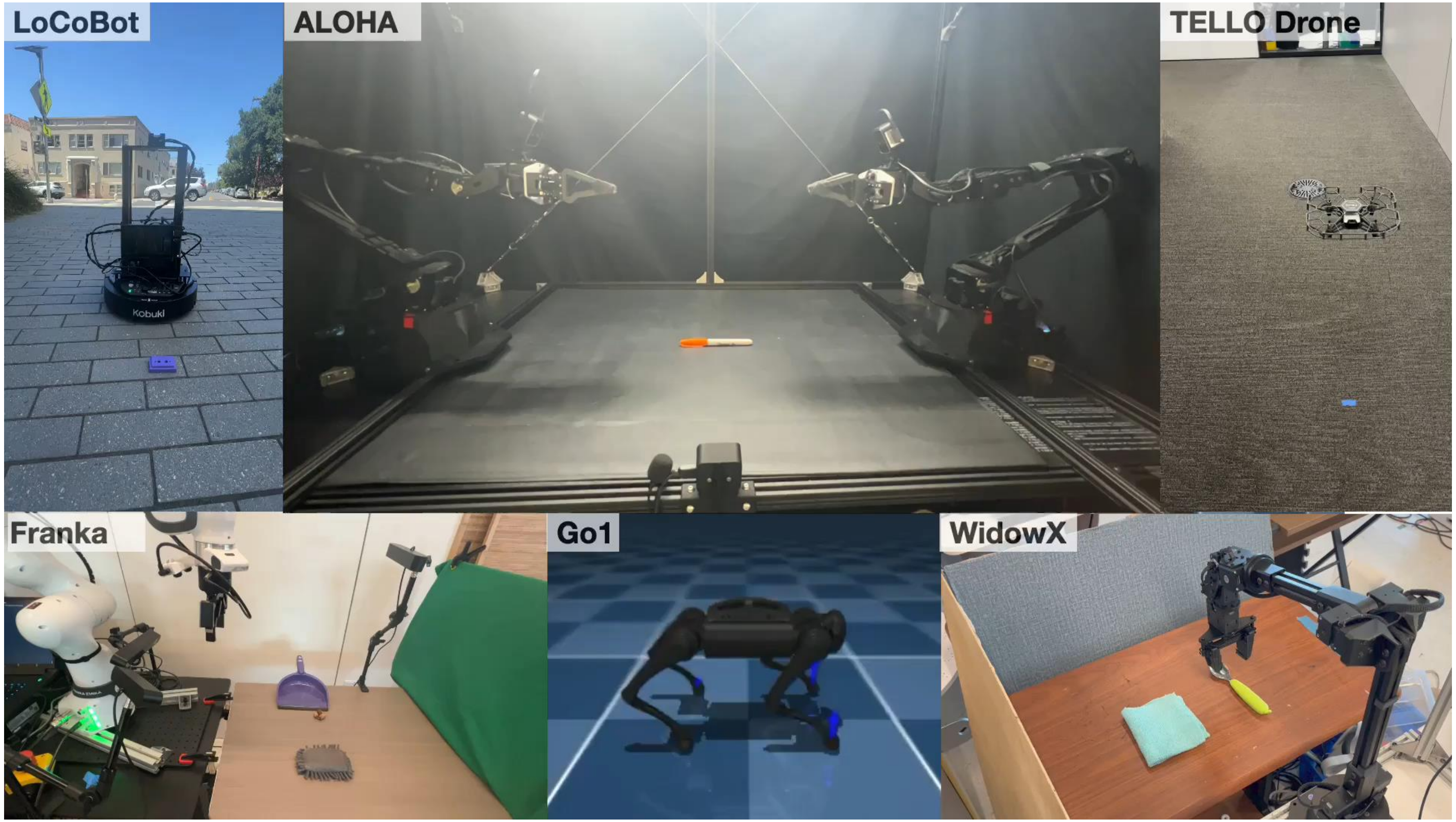






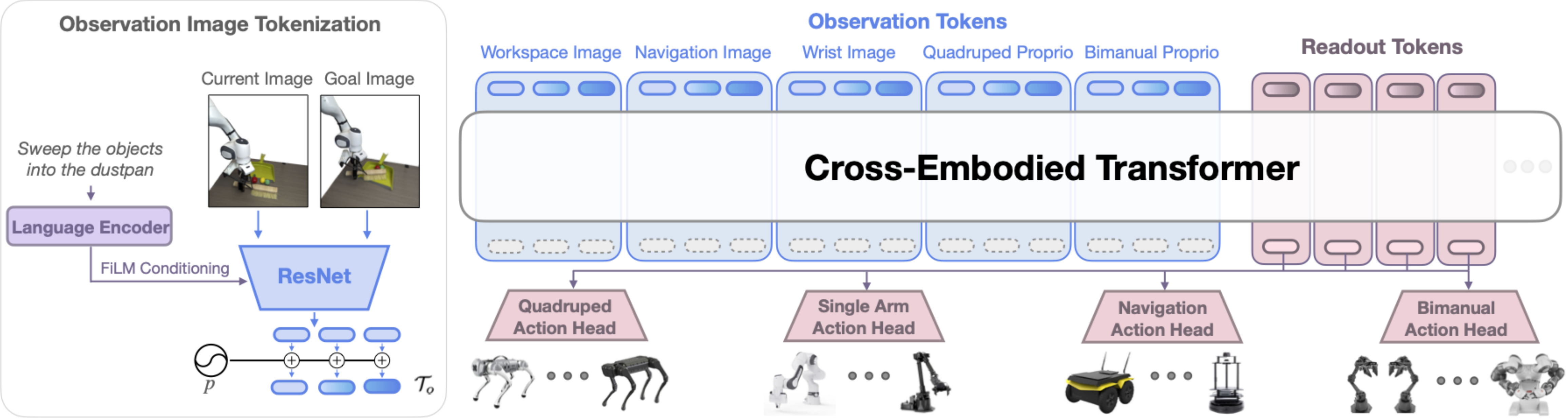
Yang, Glossop, Bhorkar, Shah, Vuong, Finn, Sadigh, Levine. **Extreme Cross-Embodiment Learning for Manipulation and Navigation.** 2024.

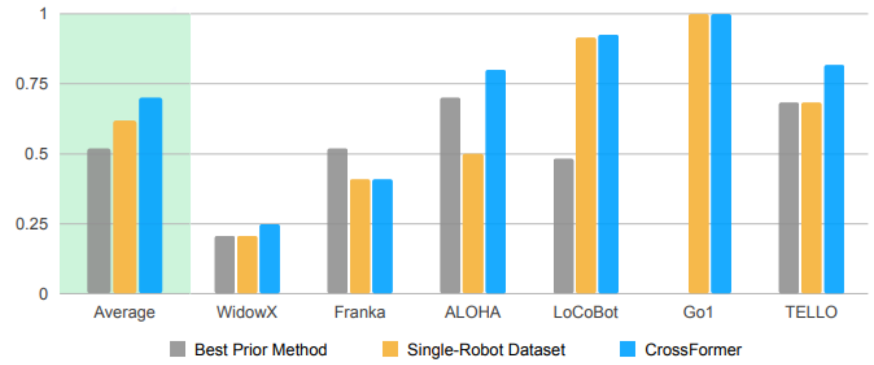
Can we make it even more “extreme”?

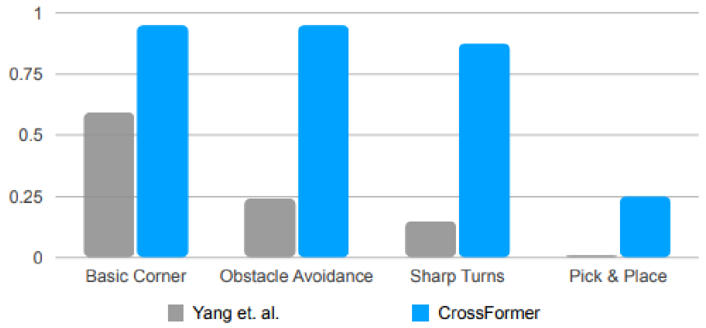


Doshi, Walke, Mees, Dasari, Levine. **Scaling Cross-Embodied Learning: One Policy for Manipulation, Navigation, Locomotion and Aviation.** 2024.

The CrossFormer architecture

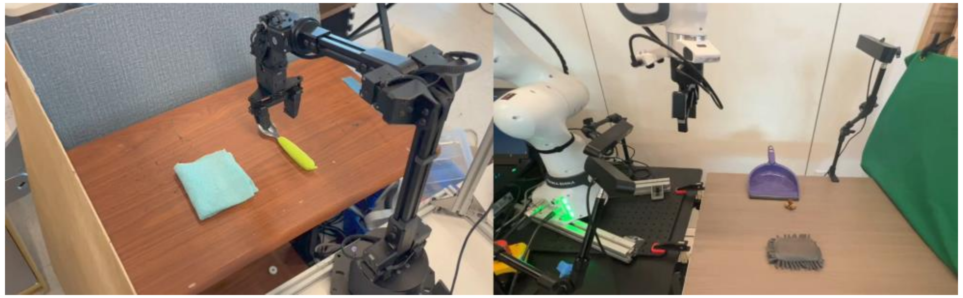




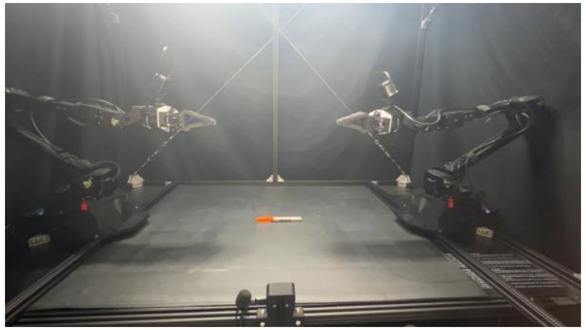


Doshi, Walke, Mees, Dasari, Levine. **Scaling Cross-Embodied Learning: One Policy for Manipulation, Navigation, Locomotion and Aviation.** 2024.

How diverse do the embodiments get?

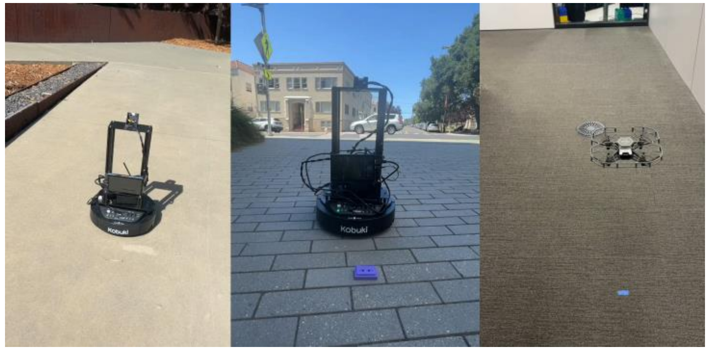


- Robotic manipulation matches prior robotic foundation models (e.g., Octo)
- Can use **either** third person or wrist-mounted cameras



High-frequency bimanual manipulation (50 Hz) matches dedicated bimanual models

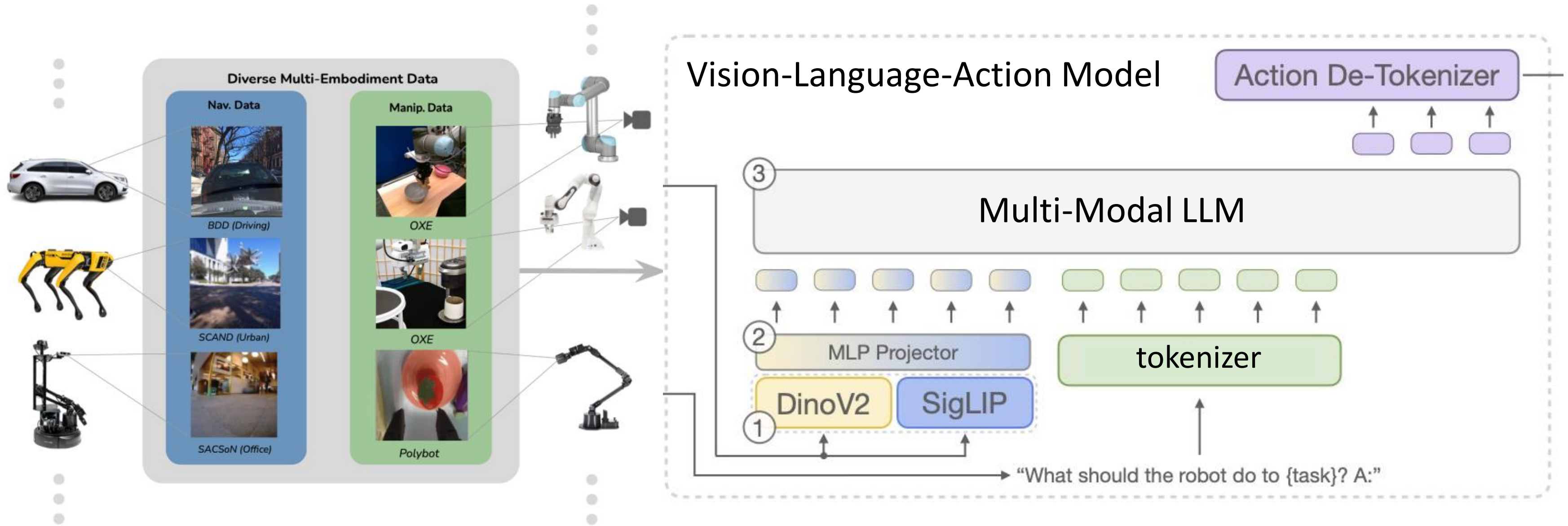
Same model performs low-level joint control for a quadruped



Integrates with topological graphs for long-horizon navigation (ground robots & quadcopters)

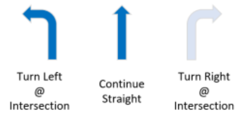
Doshi, Walke, Mees, Dasari, Levine. **Scaling Cross-Embodied Learning: One Policy for Manipulation, Navigation, Locomotion and Aviation.** 2024.

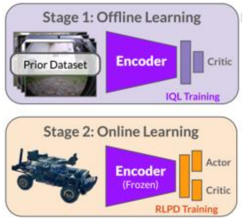
Summary



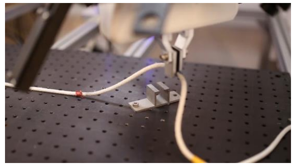
zero shot to new robots

few shot to new tasks



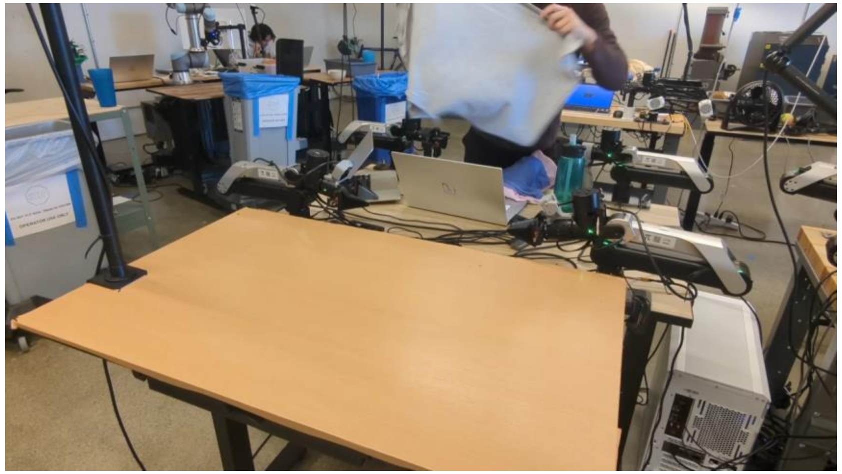


pretrain for super fast online RL



utilize for downstream instruction following

π Physical Intelligence



Can we scale up robotic foundation models to tackle the breadth of real-world tasks and robotic platforms?

http://physicalintelligence.company

website: http://rail.eecs.berkeley.edu

RAIL
Robotic AI & Learning Lab