World Modeling Challenge

CVPR Nashville JUNE 11-15, 2025

OpenDriveLab

About us

Humanoid robot company, founded in 2015

Headquartered in Palo Alto, CA

~300 employees (AI team is 34 people)

Robotics Scaling Issues

More params is not always better

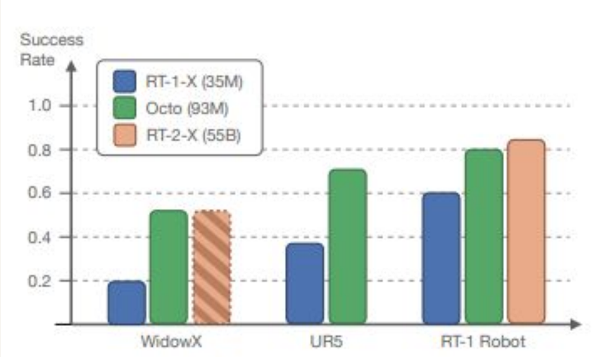


Fig. 5: **Zero-Shot Evaluation.** Out-of-the-box, Octo can control multiple robots in environments from the pretraining data. When using natural language to specify tasks, Octo outperforms RT-1-X [67], the current best openly available generalist robot policy across three different robot embodiments and setups. Octo also performs similarly to RT-2-X [103] on the tested WidowX and RT-1 Robot tasks.[1]

Octo 93M vs. 55B
(Octo Team et al. 2024)

More data is not always better

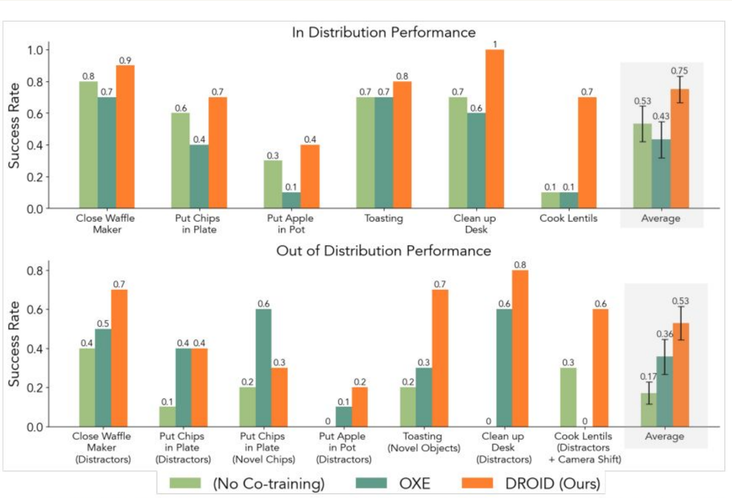


Fig. 8: **Does DROID Improve Policy Performance and Robustness?** We find that across all our evaluation tasks, co-training with DROID significantly improves both in distribution and OOD performance over both no co-training and co-training with the Open-X dataset. We compare success rate averaged across all tasks with standard error, and find DROID outperforms the next best method **by 22%** absolute success rate in-distribution and **by 17%** out of distribution.

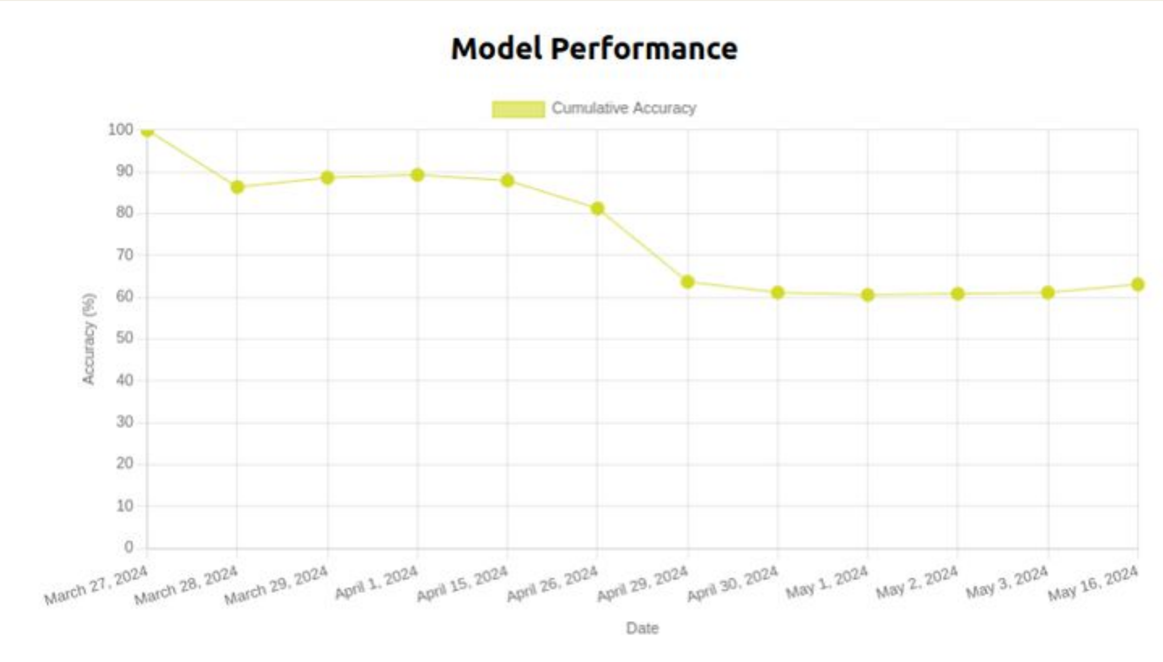
DROID dataset
(Khazatsky et al. 2024)

Why predictable scaling is important?

- **Generative Models lead the rest of ML:** Robotics models lags behind generative modeling tech (LLMs, Images, Video) by ~3-5 years. Let's catch up!
- **ML Scaling is Expensive**. If we are to make a big investment in data collection & compute, we need predictable ways to match $ spend with capability increase. ChatGPT for Robotics can only happen once the Scaling Laws for Robotics happens.

Evaluation

- In order to establish "scaling laws", reliable and consistent evaluation is critical
- In constantly evolving environments such as homes, previously valid experimental results quickly become outdated due to shifts in conditions



1X World Model Challenge

compression

sampling

Level 1: Compression Challenge

- Scaling up learning in NLP + Vision has been achieved by optimizing a simple token-prediction loss to model all the tokens jointly, let’s do the same for robotics
    - Lower loss indicates a better understanding of the data
- Given previous frames, predict next frame’s logits, scored on cross entropy loss on a held-out test set

Goal: Minimize loss

Level 2: Sampling Challenge

- Given previous frames, sample next frame, scored against PSNR
- Future predictions should be coherent and plausible continuations of the video
- Admits broader set of solutions than the compression challenge (e.g. latent diffusion)

Goal: Generate realistic future frames

Overall Winner: Team Duke

Compression Challenge

rank	id	ce
1	Duke	7.4976
2	a27sridh	7.9869
3	WaterlooVipLab	7.9869
4	Shortnapse	8.2723
5	Be off soon	11.0881
6	jmonas	11.5883
7	USTC	11.5948
8	lzzzzzm	11.5952
9	jhonQ	11.7778

Sampling Challenge

rank	id	psnr
1	Duke	21.5578
2	Micheal	18.5083
3	vjango	18.4823
4	WaterlooVipLab	18.0394
5	Jason	17.1983
6	jmy	17.1051
7	Shortnapse	17.0652
8	JJanGGoo	15.5284
9	plumwine	15.2858
10	a27sridh	11.954