OpenVLA

An Open-Source Vision-Language-Action Model

Kim, Pertsch, Karamcheti et al. · CoRL 2024 · arXiv:2406.09246

The problem

- Strong VLAs (RT-2, RT-2-X) are closed
- No recipe to fine-tune a VLA
- No best practice for consumer GPUs

so

The model

Image → DINOv2 + SigLIP

↓

MLP projector → Llama 2 7B

↓

+ instruction → 7 tokens

↓

256 bins → Δpos, Δrot, grip

7B

parameters, 7x fewer than RT-2-X (55B)

970k

robot episodes from Open X (RT-2-X: 350k)

Recipe lessons

- Fine-tune the vision encoder
- 224 px: as good as 384 px
- 27 epochs, past 95% token accuracy
- Bridge only, no Open X: −30 points

fine-tune

+20.4

points vs Diffusion Policy, fine-tuned

+16.5

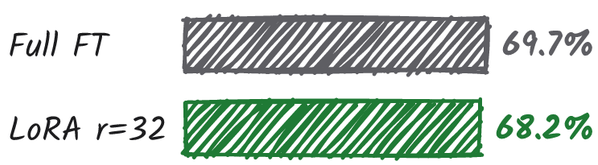
points vs RT-2-X, 29 tasks

Out of the box: WidowX



17 tasks, partial credit · Google: 85.0% vs 78.3%

Adapt: 2 Franka tasks



33 trials each · LoRA: 1.4% of params · SigLIP-only

Serve: 8 Bridge tasks



80 trials each · SigLIP-only variant, smaller mix

Takeaway

- Open 7B VLA beats closed 55B RT-2-X
- Broad robot data boosts generalization
- Limit: one image, one action, ~6 Hz on a 4090

bushuyeu.com/read-along-openvla · paper CC BY 4.0 · sketchnote by Pavel Bushuyeu

Active listening · OpenVLA

Write while you listen. One line per box is enough.

Kim, Pertsch, Karamcheti et al. · CoRL 2024 · arXiv:2406.09246

Date _____ Presenter _____ Role _____

Their main claim, in my words

The evidence they show

What I would push on

My question for the room

How it connects to my work

Listen for

- ☐ Borrowed vs new: Prismatic, RT-2 bins, Octo's mix
- ☐ How one robot action becomes 7 tokens
- ☐ Test robots: are they in the training data?
- ☐ Where RT-2-X and Diffusion Policy still win
- ☐ Which model and GPU the LoRA and int4 claims use

Ask the room

1. One image and action, ~6 Hz (4090): which fast tasks suffer?
2. Is co-training with web data worth it for semantics?
3. Data cleaning and API quirks: how to compare VLAs fairly?

Page, video and FAQ:
bushuyeu.com/read-along-openvla