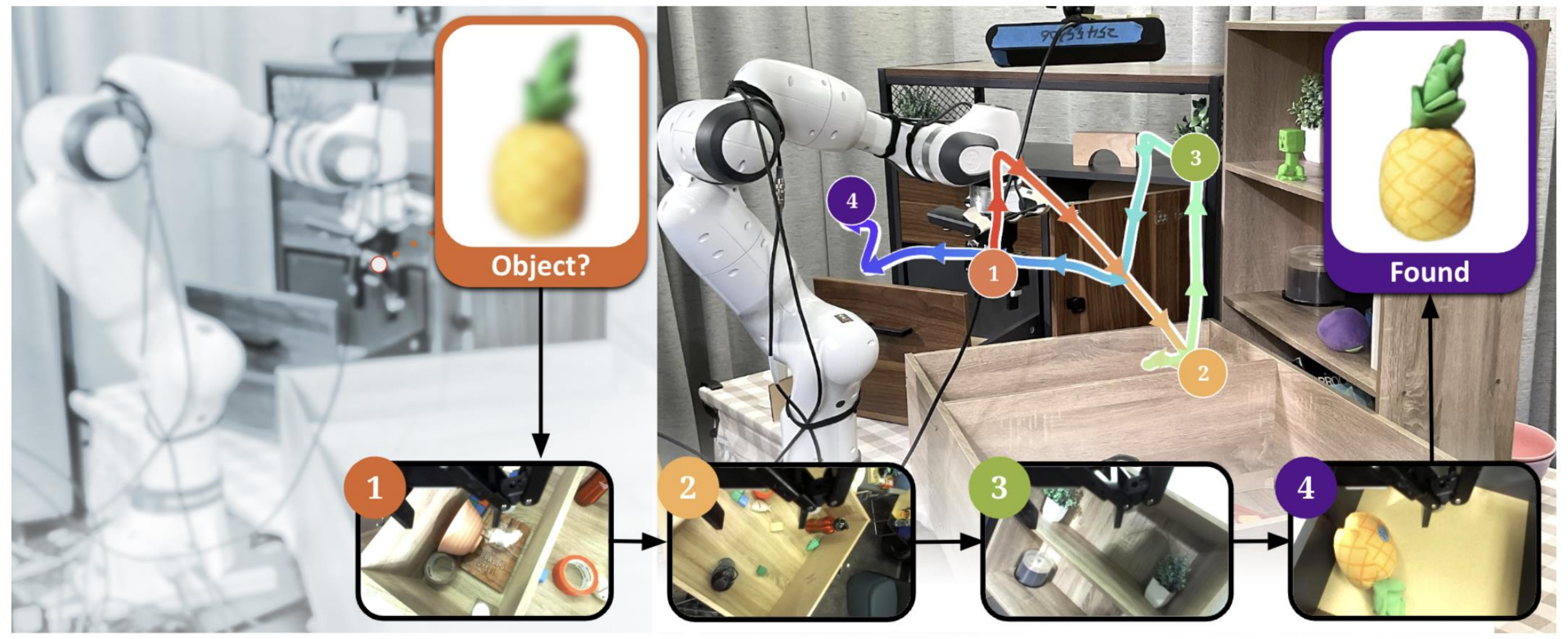
Real World Reinforcement Learning of Active Perception Behaviors

Edward S. Hu*, Jie Wang*, Xingfang Yuan*, Fiona Luo, Muyao Li, Gaspard Lambrechts, Oleh Rybkin, Dinesh Jayaraman

Website

VLAs Struggle to Search

How to act optimally from limited sensing?



Active Perception (AP): Move sensor around to improve perception.

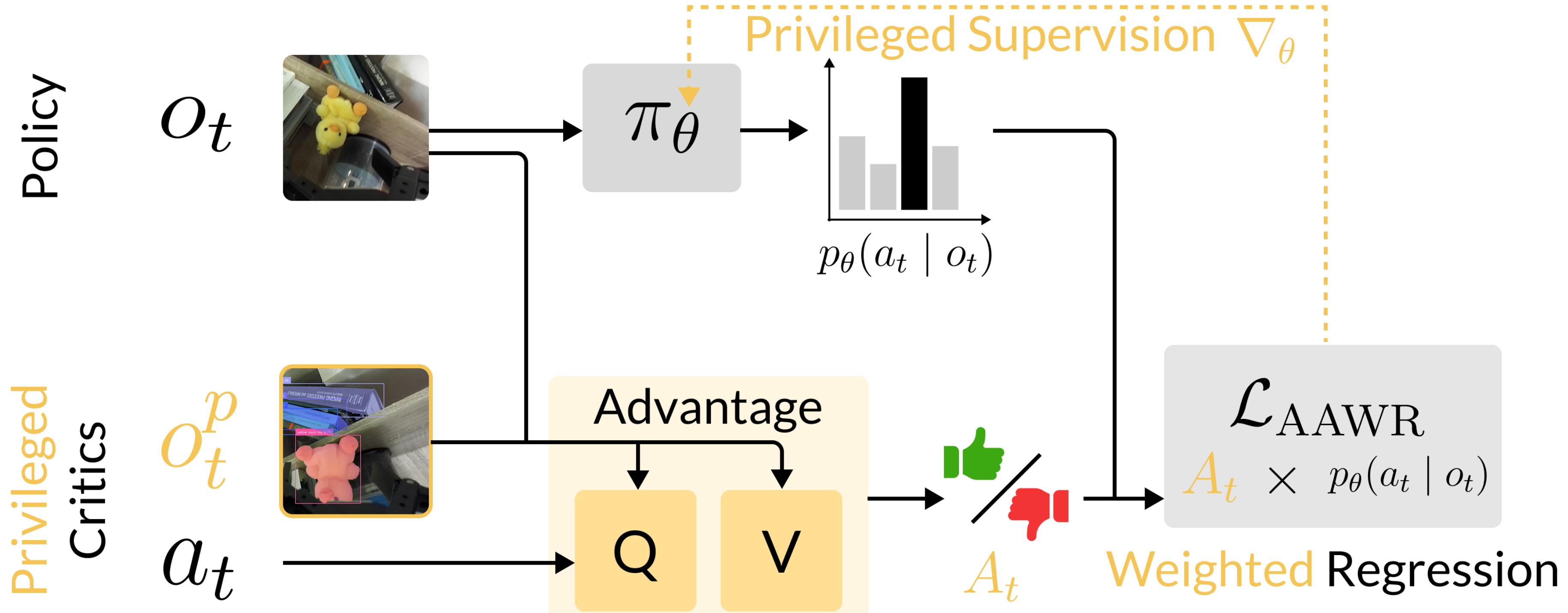
VLAs fail to find objects placed out of view.

Inefficient search, with meandering behavior.

Top half of scene never scanned!

How to teach Active Perception to VLAs?

Asymmetric Advantage Weighted Regression (AAWR)

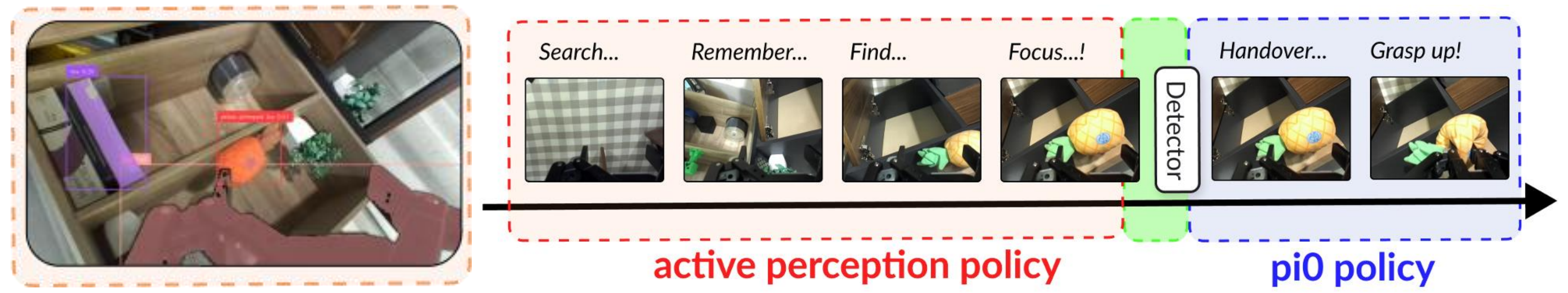


$$\mathcal{L}_{\text{AAWR}}(\pi) = \mathop{\mathbb{E}}_{(s,z)\sim d_\mu(s,z)} \mathop{\mathbb{E}}_{a\sim\mu(a|z)} \left[\exp\left(A^\mu(s,z,a)/\beta\right) \log \pi(a \mid z) \right]$$

$$A^\mu(s,z,a) = Q^\mu(s,z,a) - V^\mu(s,z)$$

Extend Advantage Weighted Regression[1,2] versatile weighted BC approach, to POMDPs

Prove AWR needs privileged state info to converge to optimal POMDP policy



Train AP policy to find good viewpoints, then switch to VLA policy

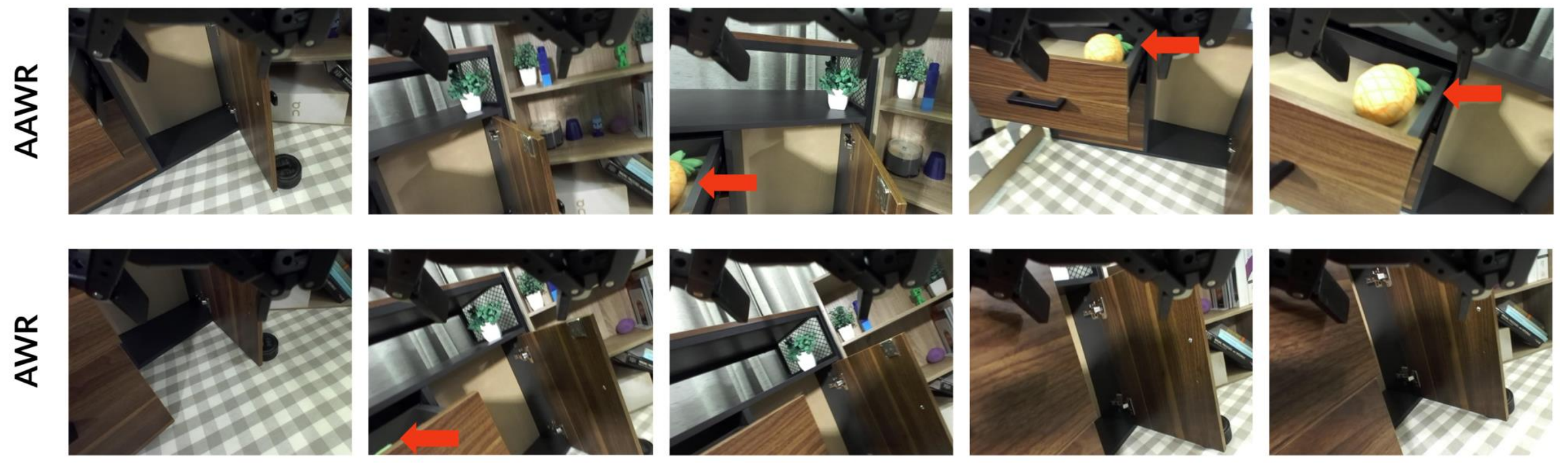
Privileged info / reward: object detector outputs

[1] Neumann, Gerhard, and Jan Peters. "Fitted q-iteration by advantage weighted regression." *NeurIPS* (2008).
[2] Peng, Xue Bin, et al. "Advantage-weighted regression: Simple and scalable off-policy reinforcement learning." *arXiv* (2019).

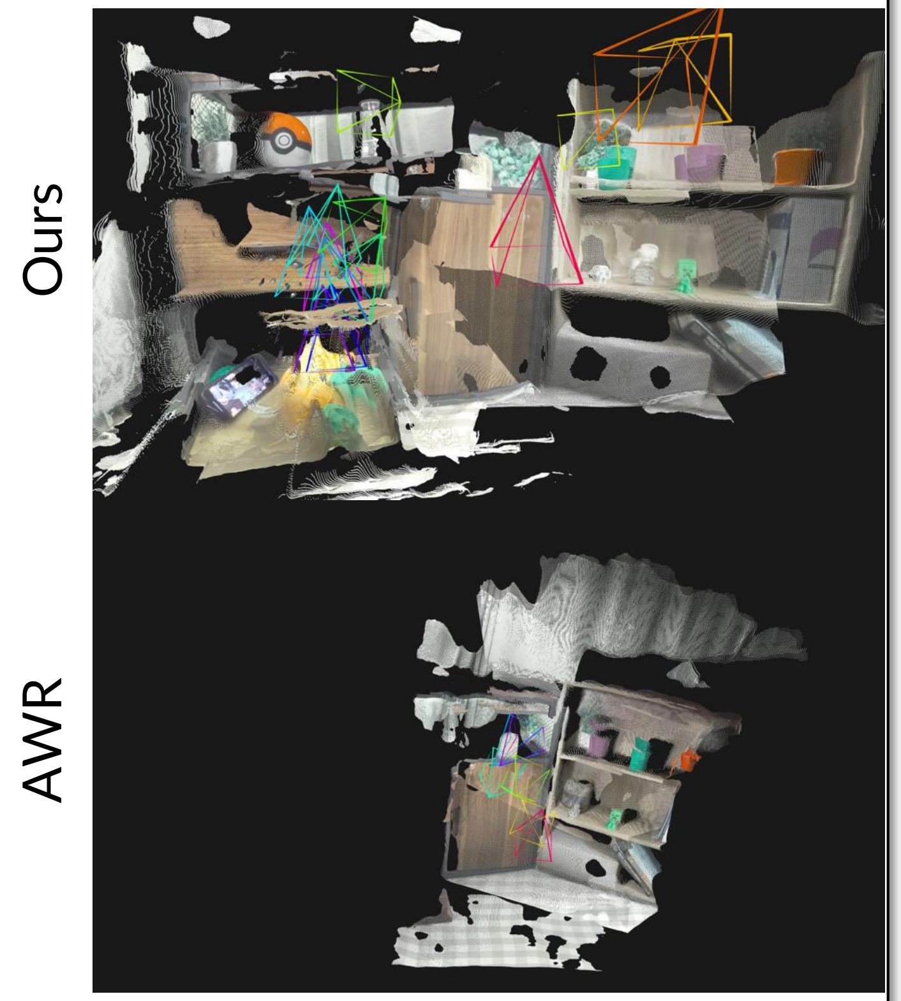
Experiments

Method	Bookshelf-P			Bookshelf-D			Shelf-Cabinet			Complex		
	Search ↑	Completion ↑	Steps ↓	Search ↑	Completion ↑	Steps ↓	Search ↑	Completion ↑	Steps ↓	Search ↑	Completion ↑	Steps ↓
AAWR	**92.4**	**44.4**	36.6	81.3	**44.4**	26.9	**78.2**	40.0	46.3	73.2	**50.0**	**43.0**
AWR	79.6	0.0	**34.0**	62.6	16.7	30.2	52.3	10.0	**38.0**	33.2	40.0	67.0
BC	29.9	20.0	84.0	47.7	16.7	**22.5**	28.1	15.0	125.0	31.5	15.0	77.0
π_0	11.0	16.7	263.3	66.7	33.3	229.7	10.0	10.0	280.0	29.6	20.0	252.5
Exhaustive	64.2	44.0	105.4	**96.0**	22.2	106.7	52.8	**45.0**	183.0	**78.2**	30.0	297.0
VLM+π_0	31.4	27.8	322.3	33.2	16.7	281.8	28.2	15.0	382.0	14.8	10.0	374.7

With ~50 demos, AAWR outperforms RL/IL/TAMP approaches in search quality, VLA success, and time.

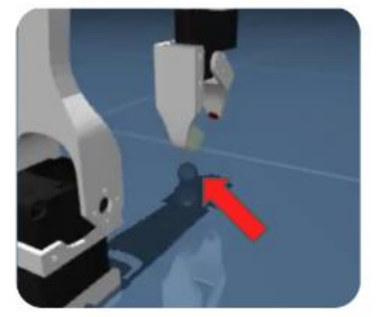


Wrist Cam 3D reconstruction

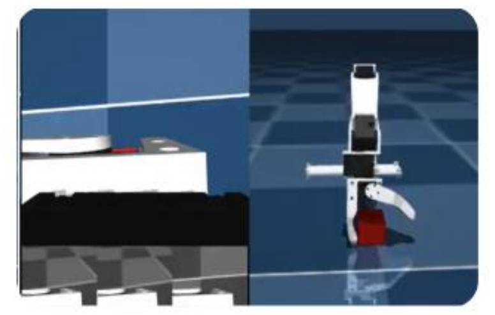


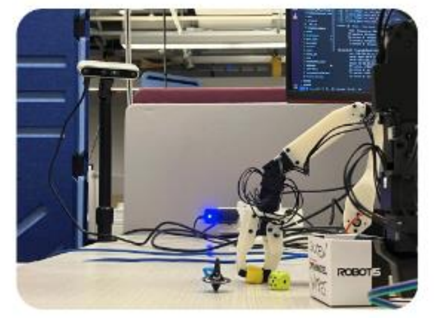
AAWR locates and fixates on the target object, but baselines fail.

See paper for more experiments; offline-to-online RL in real world!


Camouflage Pick

Fully Ob. Pick


Active Perception Koch


Blind Pick

Method	Grasp %	Pick %
BC	47	41
Off. AWR	65	62
On. AWR	71	55
Off. AAWR (ours)	88	71
On. AAWR (ours)	**94**	**89**

Takeaway: Efficiently train AP policies in real world by using privileged value networks to supervise the policy.