Fast Reinforcement Learning without Rewards or Demonstrations via Auxiliary Task Examples

Trevor Ablett[1], Bryan Chan[2], Jayce Haoran Wang[1], Jonathan Kelly[1]

[1]University of Toronto, [2]University of Alberta

MRM-D Workshop

Website
papers.starslab.ca/vpace

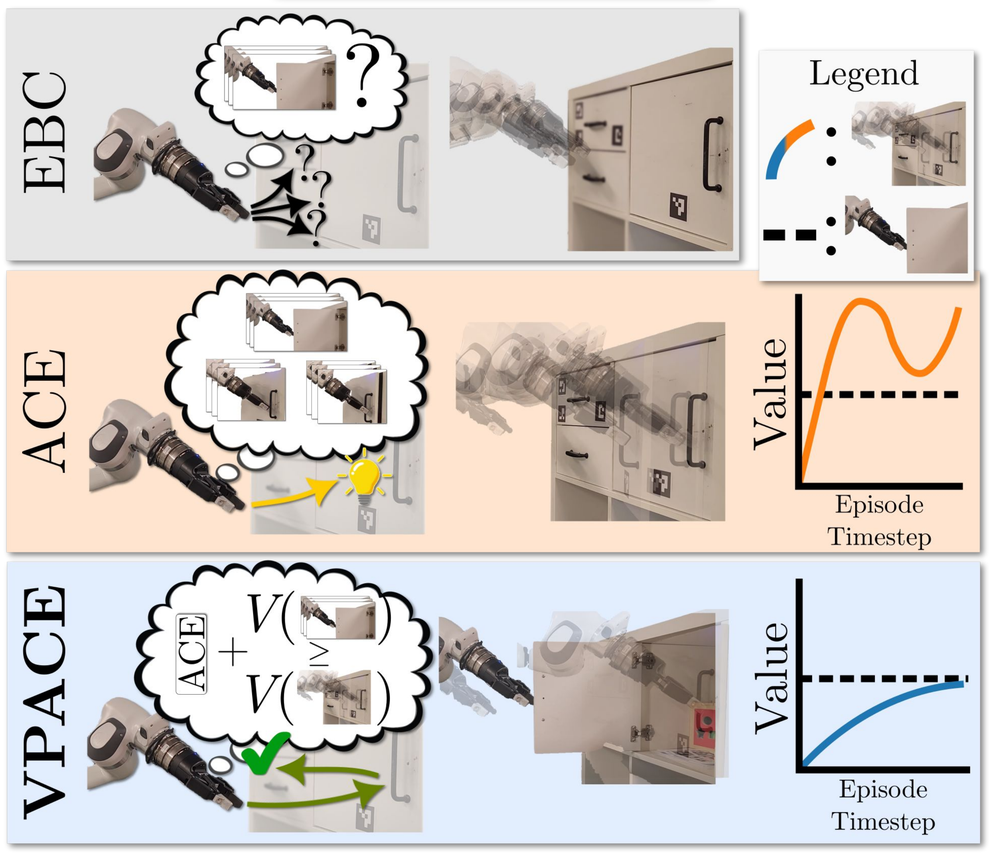
Code
github.com/utiasSTARS/vpace

Motivation

Problem: Example-based control (**EBC**) (RL from examples) is very inefficient for learning even moderately complex tasks.

Question: Can we use *auxiliary* task examples in a hierarchical model to improve exploration?

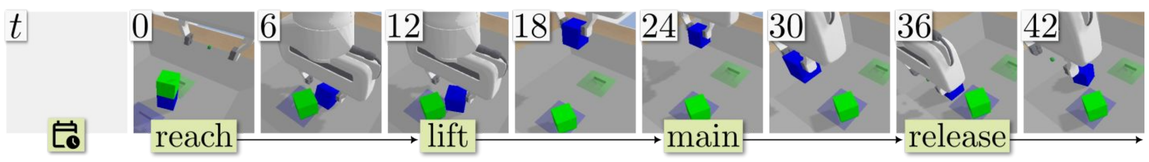
Summary



VPACE

*Auxiliary Control from Examples (**ACE**)*:

- Off-policy learning with **multitask policy** and **multitask Q-function** (discriminator optional).
- Implementation of SAC-X[1,2] framework for IRL, where scheduler selects between individual policies during training.



- All policies and Q-functions learn from all data.

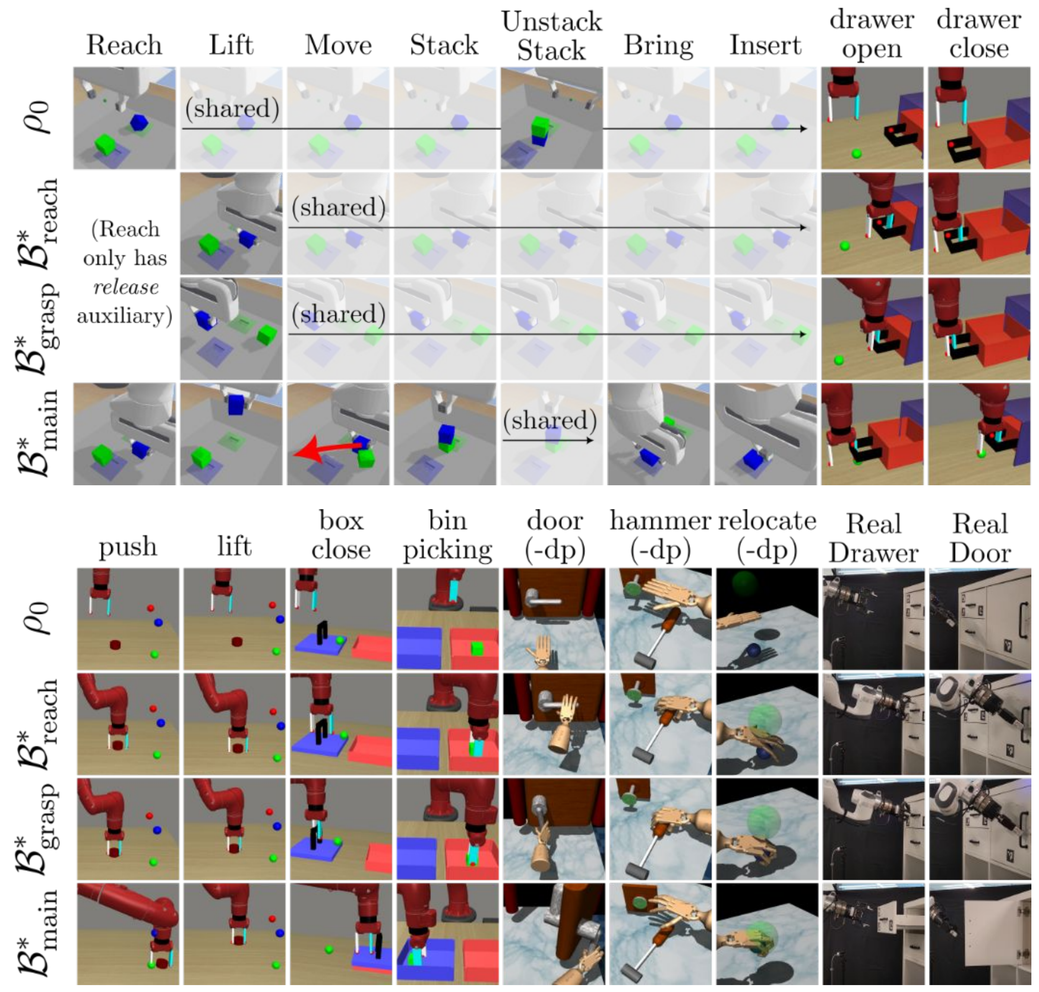
Highly exploratory policies have unstable Q values due to bootstrapping. We use **over-success-level** value penalization with ACE (***VPACE***):

$$Q^{\pi}_{\min} = \hat{R}_{\min}/(1-\gamma) \qquad Q^{\pi}_{\max} = \mathbb{E}_{\mathcal{B}^*}\left[V^{\pi}(s^*)\right]$$

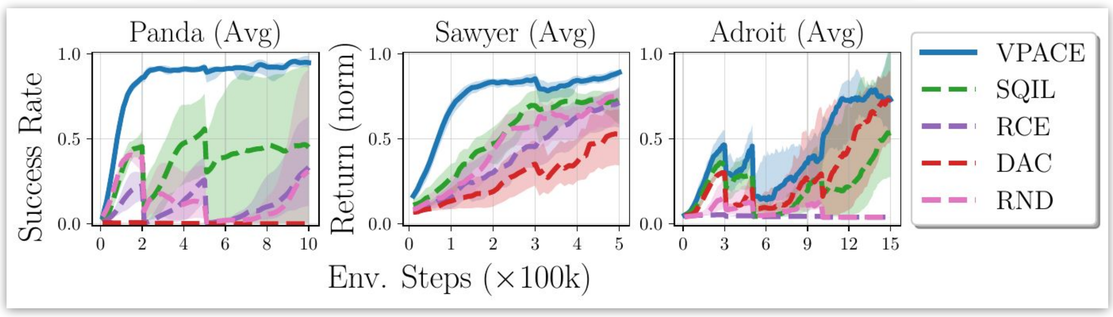
$$\mathcal{L}^{\pi}_{\text{pen}}(Q) = \lambda \mathbb{E}_{\mathcal{B}}[\,(\max(Q(s,a) - Q^{\pi}_{\max}, 0))^2 + (\max(Q^{\pi}_{\min} - Q(s,a), 0))^2]$$

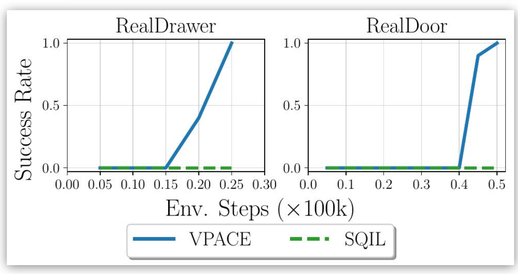
[1] M. Riedmiller et al., "Learning by Playing Solving Sparse Reward Tasks from Scratch," ICML'18
[2] T. Ablett, B. Chan, and J. Kelly, "Learning From Guided Play: Improving Exploration for Adversarial Imitation Learning With Simple Auxiliary Tasks," RAL'23

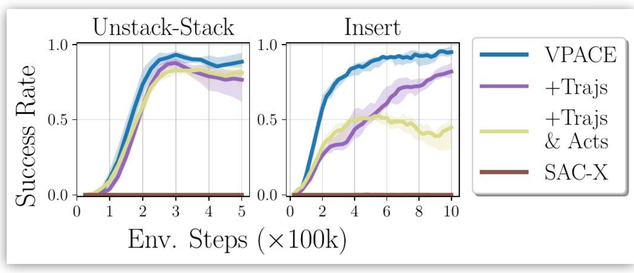
Experimental Environments & Data



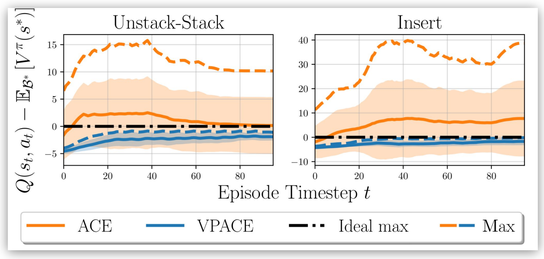
Results and Analysis

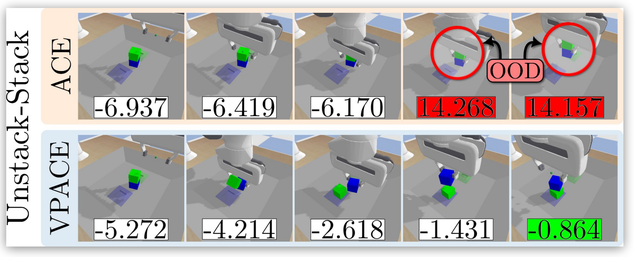






VPACE strongly outperforms baselines, and learns tasks in 1-3 hours from scratch on a real robot. Preliminary results show that learning from examples may outperform learning from full trajectories.





Values are highly overestimated without penalization, especially for OOD states.

Conclusion

✅ VPACE enables fast RL from examples.
❌ Requires manual task selection.
🔮 Future work: apply to offline RL.