Prompt Driven Exploration

Sunshine Jiang[1,3], John Marangola[1,3], David Zhang[1,3], Raghuram Kowdeed[3], Ruiyang Luo[1,3], Nitish Dashora[1,3], Richard Li[1,3], Pulkit Agrawal[1,3], Zhang-Wei Hong[1,2,3]

Massachusetts Institute of Technology[1], MIT-IBM Computing Research Lab[2], Improbable AI Lab[3]

Why is reinforcement learning (RL) sample inefficient?

Standard RL explores through action noise (e.g., Gaussian noise, random actions, entropy bonuses), which is inherently local; discovering new strategies requires global exploration.

Local exploration
Jittering, local

Global exploration
Consistent, global

How to perform global exploration?

Intuition: If a policy fails under one prompt, try describing the task differently.

Task: *"Pick up the green container on the top rack"*

Original prompt

"put the green container on the top rack"

Picked up the green container, but failed to put it on the rack

Optimized prompt

"put the green container completely on the top rack"

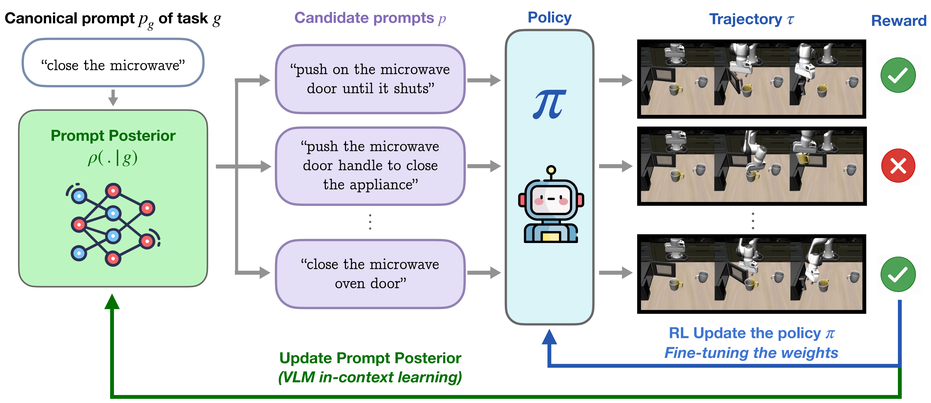
Picked up the green container and put it on the rack successfully

Takeaway: One prompt fails, but another prompt can unlock successful behavior.

Question: How do we discover those prompts?

Method: Prompt Driven Exploration (PDE)

Intuition: Instead of perturbing actions, sample prompts. Each prompt induces a coherent behavior for the whole rollout.



PDE mirrors Posterior Sampling RL (PSRL): Sample one hypothesis, execute it for the whole episode, then update from feedback. Instead of sampling a policy, PDE samples a prompt from a VLM posterior. Each prompt induces a coherent policy $\pi_p(a \mid o) = \pi_\theta(a \mid o, p)$, shifting exploration from noisy actions to high-level language behaviors.

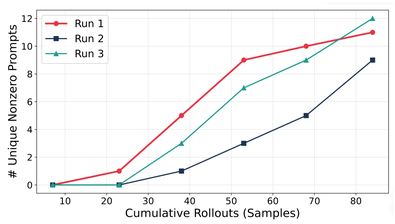
Why and how PDE improves exploration?

Why does the initial policy fail?

- Reuses a wrong motor program from a visually similar training task
- Action noise only perturbs locally, so rollouts repeat the same failure

Can PDE discover successful prompts?

- Discovery consistently scales linearly with rollout number
- 3 types of successful rollouts
  - Explicit contact/action
    "push the microwave door closed"
  - Spatial reference
    "close the appliance door on the left"
  - Subgoal ordering
    "first move to the microwave, then close the door"

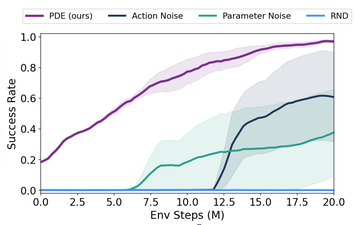


Original prompt

"close the microwave"

Optimized prompt

"push on the microwave door until it shuts"



PDE bootstraps RL from zero success!

PDE bootstraps RL efficiently from low initial success

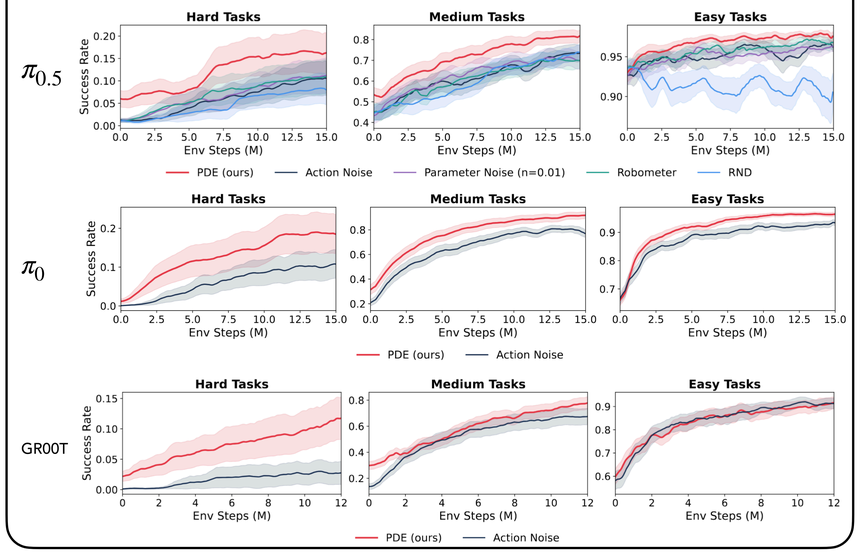
Tasks: LIBERO-PRO

3 Perturbation
1. Task
2. Swap
3. Object

4 Suites
1. Goal
2. Object
3. Spatial
4. Long

[1] Plappert, Matthias, et al. "Parameter space noise for exploration."
[2] Liang, Anthony, et al. "Robometer: Scaling general-purpose robotic reward models via trajectory comparisons."

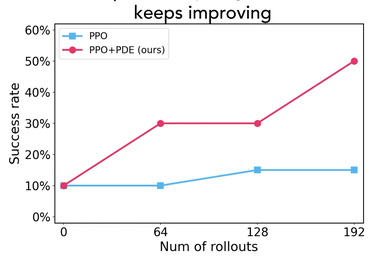
Baselines:
- Action noise
- Parameter noise [1]
- Robometer [2]
- RND



PDE improves sample efficiency for realworld RL

Setup

PPO plateaus quickly while PDE keeps improving



Original prompt

"Put the green container in the bowl"

Wrong object

"Open the top drawer and put the tape inside"

Collide with drawer

Optimized prompt

"put the green container in the green bowl"

Right object

"Open the top drawer and put the tape above"

Move higher