Reset-pose curriculum · robotic manipulation

Wesley Wong

1 · PROBLEM

what and why

Reinforcement learning (RL) policies for multi-step robotic manipulation tasks train slowly and in- consistently: identical setups run with different random seeds can converge to very different final performance. Standard training also wastes time re-practicing sub-tasks an agent has already mas- tered instead of focusing on the one it is currently struggling with, and this inefficiency compounds with instability. Existing curriculum-learning methods generally reshape rewards, goals, or environ- ment parameters instead of using an agent's own measured, per-sub-task performance to decide where a training episode should begin.

2 · APPROACH

model / method, and why this one

I designed a reset-pose curriculum for episodic RL that tracks completion of each sub-task in a multi- step manipulation task, identifies the earliest sub-task the agent is currently struggling with (its bottleneck), and resets a fraction of training episodes directly to previously-successful states just before that bottleneck. A custom sampling algorithm blends a decisive selection with a more conservative, exploratory one, weighted by how clearly a single bottleneck currently stands out. This targets training experience at the edge of the agent's current capability without modifying the reward function or the task itself, unlike most existing curricula.

3 · DATASET

source, size, preprocessing

No external dataset is used. Training data is generated entirely online through simulation in NVIDIA IsaacLab (GPU-parallelized robot physics), across three benchmark manipulation tasks built on the Franka Emika robot arm: a cabinet-drawer-opening task, a nut-threading task, and a cube-lifting task. Each environment runs thousands of parallel simulated instances; sub-task completion and successful reset states are logged automatically during training, so no manual labeling or preprocessing is required.

4 · METRICS

how you'll know it worked

The primary metric is task success rate, compared between a curriculum-enabled condition and a fixed-reset baseline across 15–16 random seeds per configuration per environment, rather than relying on any single training run. Secondary diagnostics (policy entropy, reset-pose deviation, per-sub-task success) are used to explain why a result occurs, not just to report a headline number.

5 · COMPUTE

training time, memory, dependencies

All training was run on a single local workstation GPU using NVIDIA IsaacLab; no cluster or cloud allocation was required. Full training across all seeds and environments completes on the order of hours to low tens of hours in total. Dependencies: IsaacLab, PyTorch, and PPO implementations from RSL-RL and RL-Games. No compute is being requested – training is already complete.

6 · DELIVERABLES

model, paper, demo

A new type of RL curriculum: one that adapts an agent’s reset distribution using its own empirically measured, per-sub-task performance, rather than reshaping rewards, goals, or environments as existing curricula do. This is documented in a manuscript already drafted end-to-end (Related Work through Limitations), targeting an arXiv preprint by the end of September 2026 and submission to a peer- reviewed venue by the end of 2026. What I am seeking is mentorship on: (a) tightening the manuscript for a research audience, (b) selecting an appropriate venue, (c) the arXiv submission/endorsement process, and (d) feedback from someone with publication experience before I submit.

7 · RISK

what breaks first, and plan B

what breaks first, and plan B The main risk is training stability. RL is inherently fragile and prone to collapse when foreign modifica- tions are introduced into an environment, which is part of why this work builds on IsaacLab’s existing example environments rather than custom ones – they serve as a well-tested baseline and simplify generalization across tasks. Plan B: if the curriculum proves destabilizing in a way that outweighs its benefits, explore alternative formulations of the approach that are less disruptive to training.