

AutoEval: Autonomous Evaluation of Generalist Robot Manipulation Policies in the Real World



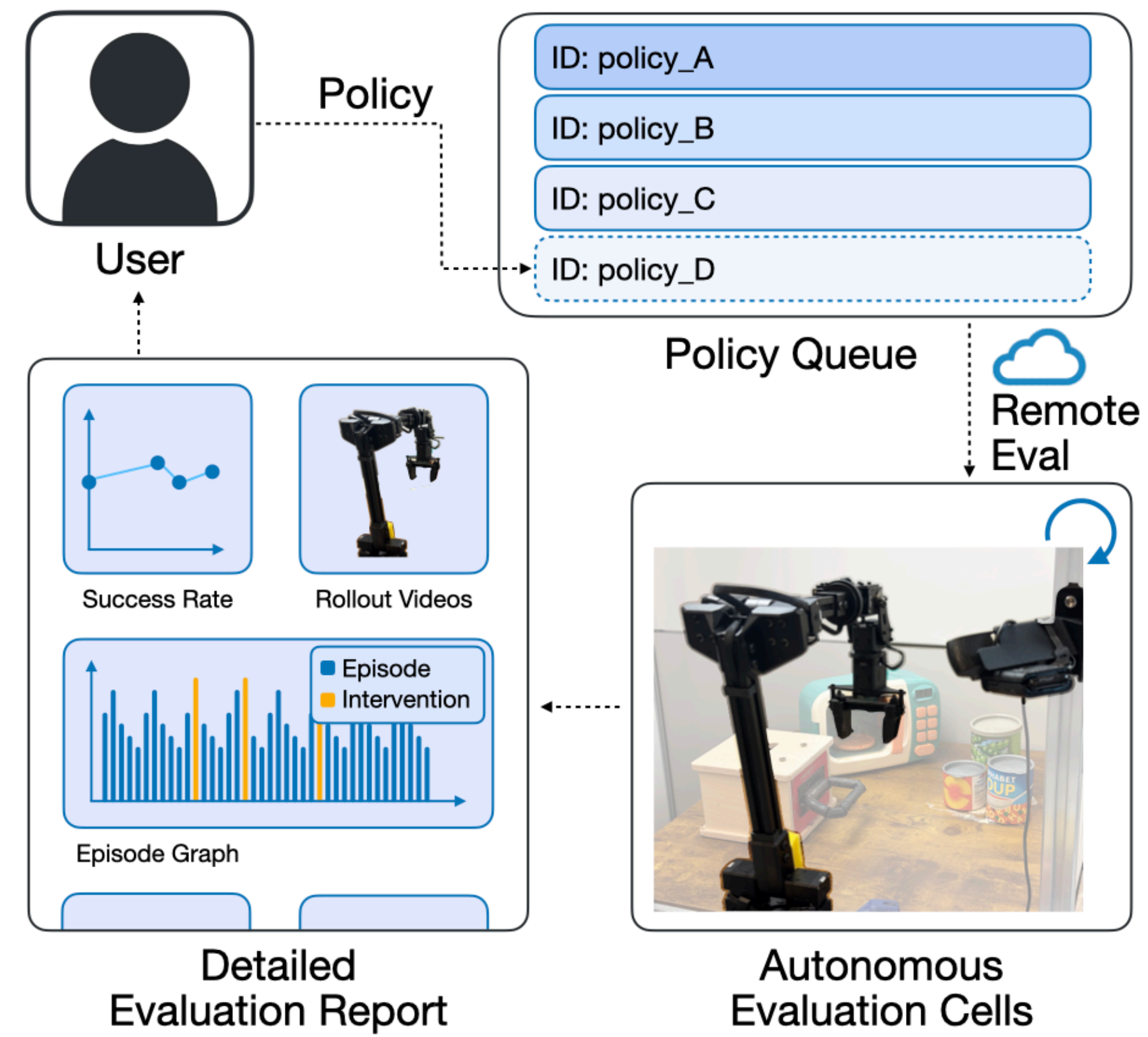
NVIDIA

<https://auto-eval.github.io>

Zhiyuan "Paul" Zhou, Pranav Atreya, You Liang Tan, Karl Pertsch, Sergey Levine

TL;DR

- ❖ AutoEval, a system to autonomously evaluate generalist robot policies in the real world 24/7, saving >99% human time
- ❖ Access our two public AutoEval stations online and submit your policy for eval!

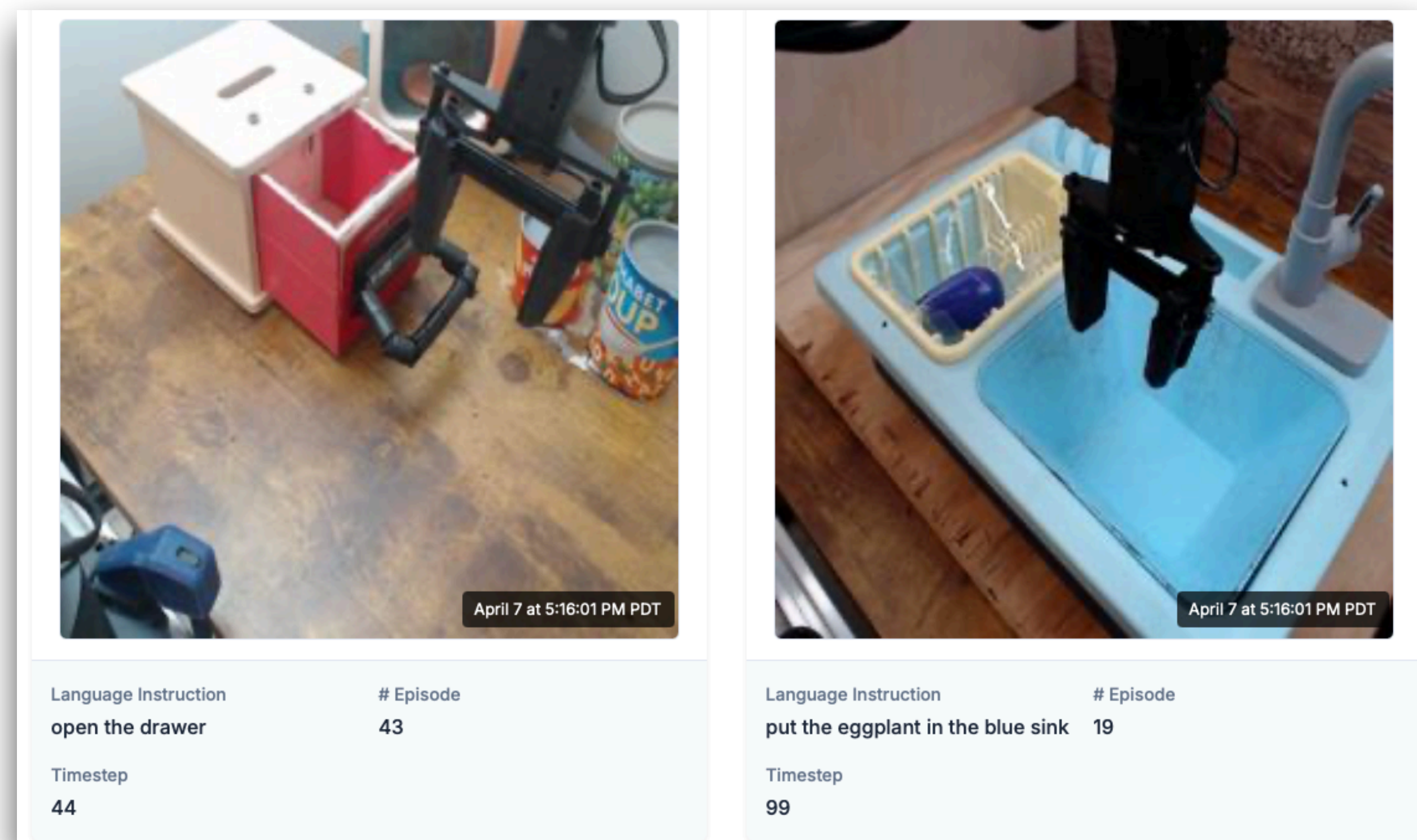


1. **Submit** Your Eval Now
2. Watch Robot **Live Stream**



Publicly Available Tasks

We **open access** to two AutoEval stations with four BridgeDataV2-style tasks.



1. Open the drawer
2. Close the drawer
3. Put eggplant in yellow basket
4. Put eggplant in blue sink

Submitting Evals & Getting Results

Submit New Job

Job Description

Enter job description

Robot

Select a robot

Task

Select a robot first

Policy Name

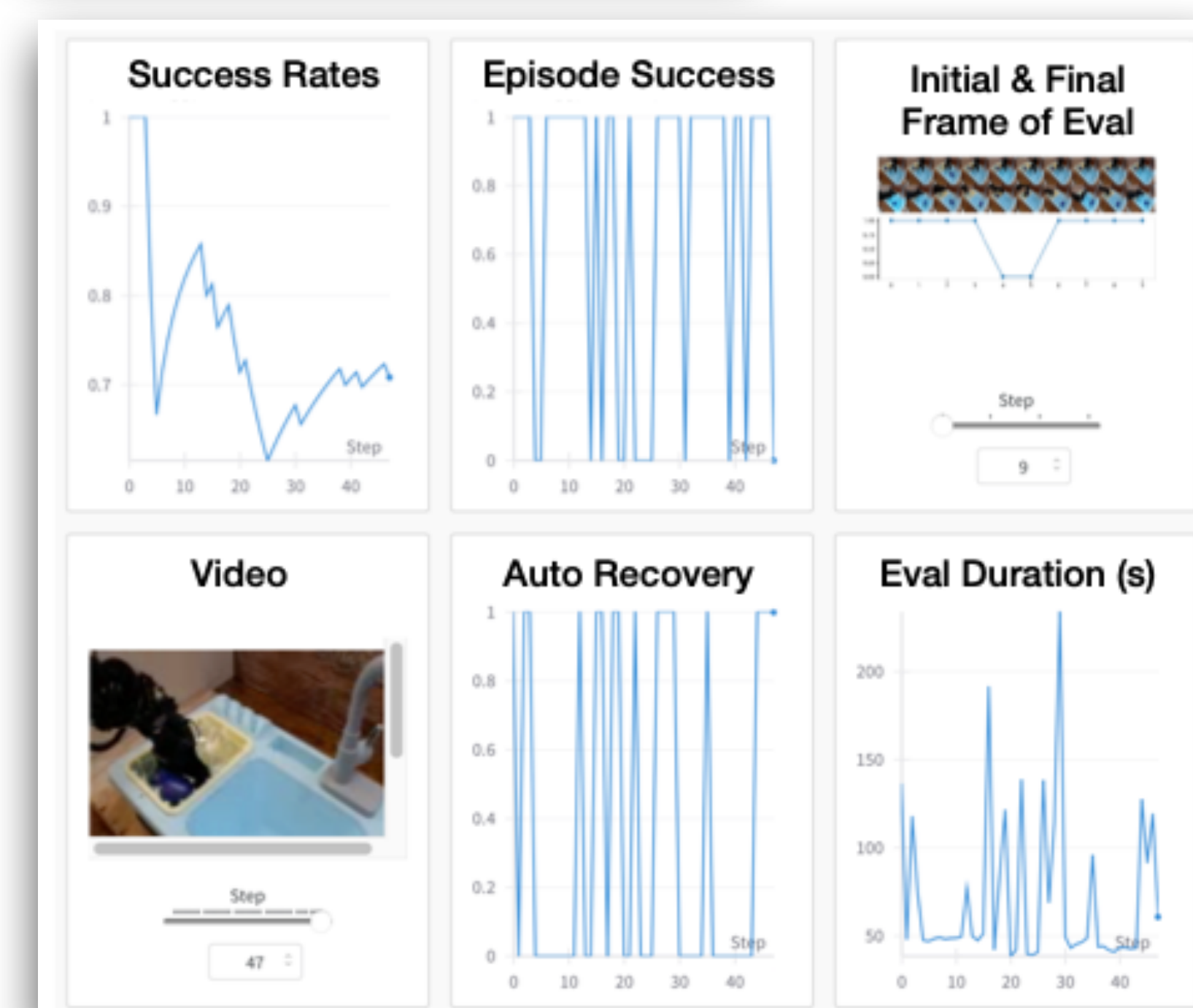
Enter policy name

Policy Server IP

Enter policy server IP

Submit Job

Simply host your policy as a *server* (see code examples), and evaluate it by submitting an "*evaluation job*" to the AutoEval queue.



After AutoEval finishes, a *report* with success rates, videos, etc is provided.

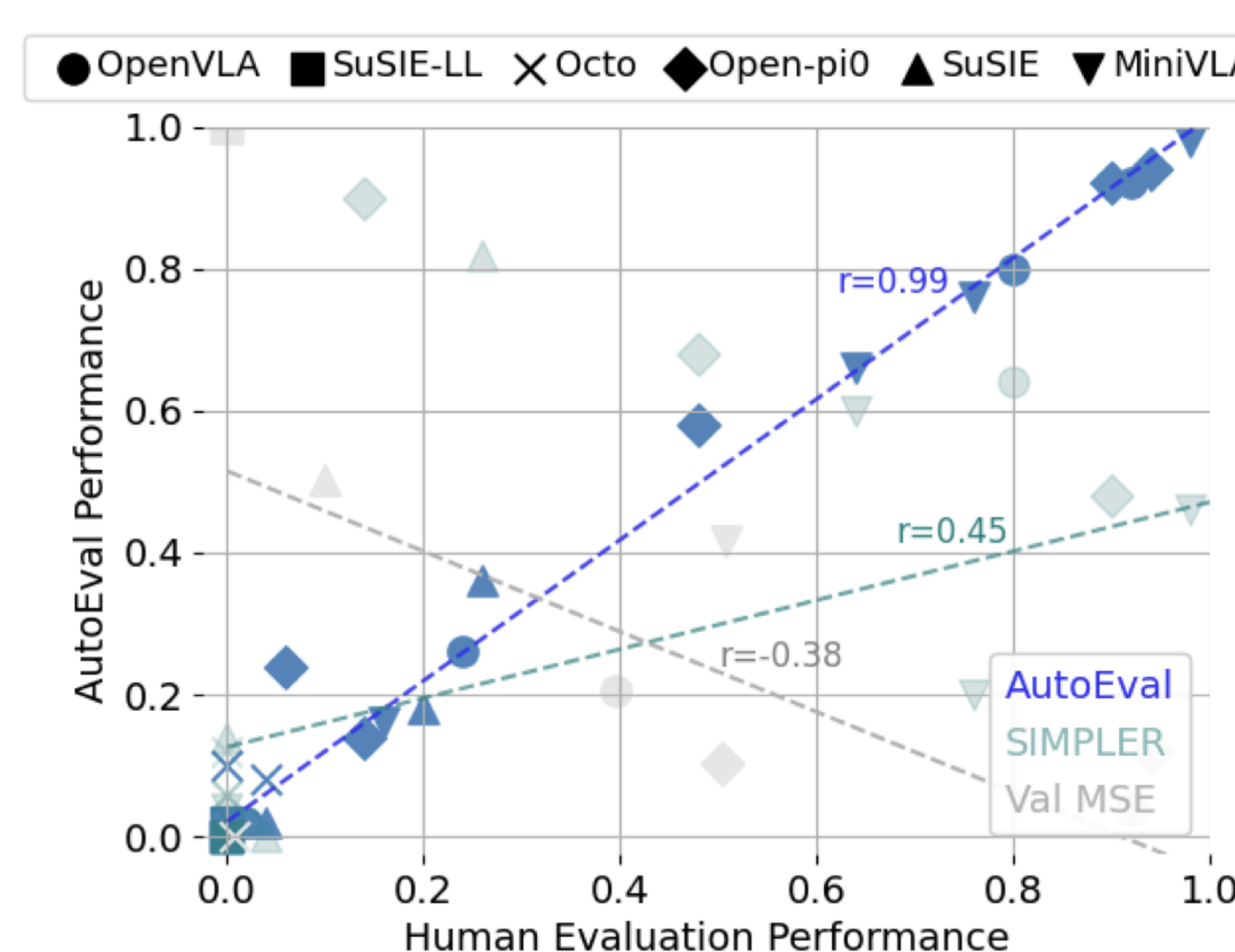
How does it work?

Replace human success detection & scene reset with **learned components**, by fine-tuning **foundation models** (VLMs, VLAs).

Algorithm 1 Autonomous Policy Evaluation Loop

- 1: **Input:** Task T , policy π to be evaluated, initial state distribution $\rho(s)$, success classifier C_T , reset policy π_T , reset classifier $C_{\rho(s)}$
- 2: **Output:** Estimated prob. of success for task T
- 3: **for** each trial **do**
- 4: **Start State:** Start from initial state $s_0 \sim \rho(s)$
- 5: **Policy Rollout:** Rollout π for K steps
- 6: **Success Check:** Label success using $C_T(s_K)$
- 7: **Reset Scene:** Rollout reset policy π_T to return initial state to $\rho(s)$
- 8: **Failure:** If unable to reset or robot unhealthy, notify human operator to help
- 9: **end for**

Results



1. **Correlates well** with human-run evals, better than simulated evals and offline metrics
2. Saves **>99%** human time
3. Runs **robustly over long time** periods & results are **reproducible**

