ROBOTICS CENTER

TECHNICAL WHITEPAPER / V0.1

RL²

Reinforcement Learning in Real Life

The deployment-and-learning loop that takes robot policies from “works in the lab” to “works — and pays — in production.” In Physical AI, model capability is commoditizing. The durable moat is the real-world learning loop.

Robotics Center of Silicon Valley
YC S25 · San Francisco · roboticscenter.ai

PHYSICAL AI · INFRASTRUCTURE

ABSTRACT

Physical AI is bottlenecked not by model capability but by the loop between a policy that works in the lab and one that survives — and pays — in production. This paper defines that loop, **RL² (Reinforcement Learning in Real Life)**: deploy → capture failures → evaluate against real acceptance criteria → collect targeted data → redeploy. We argue that models commoditize while the deployment-and-learning loop is the durable moat, describe the one-stack architecture (hardware · data · evaluation) that runs it, and specify how we measure production readiness — stating plainly what is built today versus on the roadmap.

01

THE PROBLEM

Why Physical AI Stalls

After two years at the center of the Bay Area robotics ecosystem — supplying hardware, data, and integration to frontier labs, funded startups, and stealth teams — one pattern is undeniable: the blocker is rarely model capability. It is the loop between a policy that works in the lab and one that survives, and pays, in a real deployment. Two gaps stall teams.

1.1 The sim-to-real gap: three failure modes

- **Simulation ≠ hardware reality.** A VLA trained in Isaac Sim or MuJoCo meets real actuator limits — thermal throttling, joint backlash, sensor drift. Even two units of the same robot model have different dynamics.
- **Teleoperation data is off-policy.** Human demonstrations bootstrap a model, but the model is not the human. That distribution gap closes through on-policy rollouts and targeted collection — not more of the same data.
- **“Working” is undefined until deployment.** Teams optimize lab success rates, then meet the long tail — the wrinkled shirt, 4pm glare, a pallet 3cm off spec. Without a real-world eval harness, readiness is a guess.

1.2 The pilot-to-production economics wall

Even a policy that passes every eval is a cost center: CapEx, maintenance, operators, downtime. Most fleets die at the pilot stage — not in the lab — when unit economics never close. Physical AI does not only have a model problem; it has a business-model problem. Any real-world loop must be cheap enough to run continuously, or it stalls regardless of technical merit.

02

THESIS

The Loop Is the Moat

Models commoditize. The loop does not. Value in Physical AI accrues to whoever closes the real-world learning loop fastest and cheapest.

Within roughly 18 months there will be several “good-enough” open VLAs, and frontier labs will give the rest away to sell compute. Value accrues to the layer that does **not** commoditize: the deployment-and-learning loop that turns a raw policy into a deployed, reliable, economically viable system.

Wall	State in 2026	Where it is solved
Model capability	Commoditizing	Not our layer
Sim-to-real	Technical wall	The RL² loop
Pilot-to-production economics	The wall that kills companies	Loop + unit economics

The same mechanism clears all three: a real-world learning loop cheap enough to run.

03

FRAMEWORK

The RL² Framework

We call the loop **RL² — Reinforcement Learning in Real Life**: a cyclic process that improves a deployed policy using real-world outcomes, not simulator reward alone.

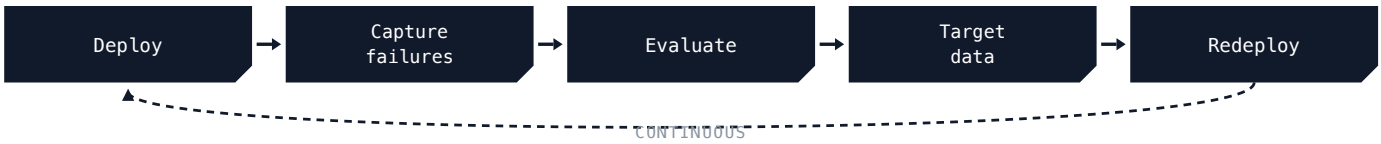

Fig 1 — The RL² loop: continuous learning on real hardware

3.1 What makes it learning, not data-hoarding

- **On-policy, not more human demos.** Roll out the current policy on real hardware and capture where it fails — not where a human differs.
- **Diagnose failures by cause.** Cluster by perception vs. contact/dynamics vs. planning vs. novel object, then collect data that targets the cause.
- **Uncertainty-aware collection.** The policy flags low-confidence / out-of-distribution states and hands off; those become the targeted data.
- **Human recovery is supervision.** When the robot hits novel uncertainty, a human recovers it — and that recovery is the correction signal.
- **Eval is the ground truth.** Held-out real-world variation, not training loss, decides whether the policy generalized.

3.2 Handling new uncertainty

You cannot pre-train the unknown. The honest answer: **detect it, fail safe (defer to a human), capture it, and close the loop fast.** The metric that matters is not “never surprised” — it is **time-to-close a new failure mode.**

04

ARCHITECTURE

System Architecture: One Stack

RL² runs on a single integrated stack. Most teams stitch these layers together from vendors; owning them as one loop is the differentiator.

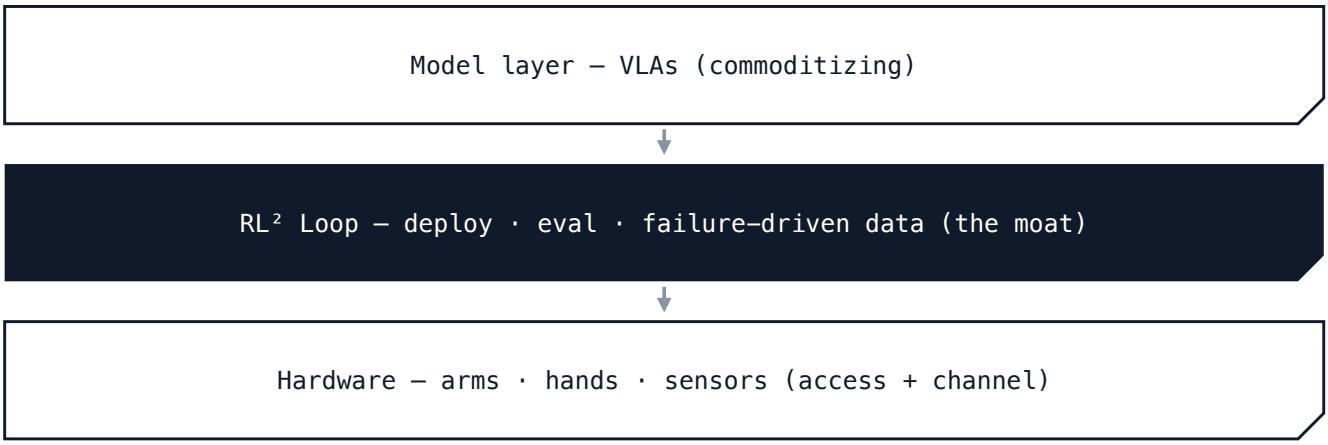

Fig 2 — One stack; the loop is the layer that does not commoditize

4.1 Hardware layer

Access and channel. Official US distribution of leading platforms (AgileX, Unitree, AGI Bot, Booster) plus our own OpenArm and Mobile ALOHA. Robots already in customers' hands are the deployment surface the loop runs on.

4.2 Data layer

Managed teleoperation and egocentric collection; a refinement pipeline (time-sync, QC, subtask segmentation, packaging); delivery as raw ros2 rosbags (MCAP) with optional LeRobot v2 / HDF5 / RLDS conversion.

4.3 Evaluation layer

The readiness gate and observability: encode acceptance criteria, run policy versions against them, produce a readiness report, and route failures back into targeted collection.

05

EVALUATION

Evaluation Methodology

Lab success rate is not readiness. We encode the customer's real acceptance criteria — the specific tasks, environments, and edge cases they are judged on in production — into an eval that reruns on every policy version.

- **Automated regression gates** — metric thresholds tracked across versions; a regression is caught before it ships (a CI gate for policies).
- **Operator-in-the-loop evaluation** — physical tasks are hard to score automatically; our operator network runs real rollouts and judges outcomes against the criteria.
- **Failure-driven data targeting** — every failed rollout points at the exact scene, object, or condition to collect next.

Example readiness report (illustrative):

Field	Value
Model	example-vla v0.3
Status	Scored — 3 finalized sessions, 52 decisive actions
Success rate	88.5% (46 pass / 6 fail)
Avg action latency	161 ms
Top failure modes	operator_rejected ×4, timeout ×2
Acceptance gate	PASS (success ≥ 85%, latency ≤ 200 ms)

Illustrative sample only — not customer data. Full automated policy scoring across arbitrary tasks is on the roadmap (§9); today readiness combines automated gates with operator-in-the-loop review.

06

DATA ENGINE

The Data Engine

Data is the fuel of the loop — but generic data is a commodity. What makes our data engine defensible is that **the eval decides what to collect.**

- **Failure-driven targeting.** Eval failures define the next collection batch — data aimed at the actual failure mode, not more of the same.
- **Managed operators.** Trained, NDA-bound specialists, certified against gold-standard episodes, with inter-operator consistency checks across people, shifts, and sites.
- **Automated QA suite.** Sync/integrity, image quality, task completion (against provided criteria), kinematic sanity, label/metadata, and distribution coverage — automated first pass, then human review, with rejected episodes re-collected so delivered usable-hours hold.
- **Formats.** Raw ros2 rosbags (MCAP) with optional LeRobot v2 / HDF5 / RLDS training-ready conversion.

07

WORKED EXAMPLE

An Illustrative Loop

A representative manipulation task — pick a bin from a stack, reorient it, place it on a conveyor — shows the loop end to end:

1. **Deploy** the current policy on the real cell.
2. **Capture** per-subtask rollouts; flag failures (missed grasp, failed reorient, timeout).
3. **Evaluate** against acceptance criteria; produce a readiness report.
4. **Target**: failures cluster on the reorient under glare — collect targeted demonstrations of exactly that condition.
5. **Redeploy**: retrain, gate on the eval, run again; track time-to-close the reorient failure mode.

Values illustrative; the point is the mechanism, not the numbers.

08

WHY NOW

The Window

- Physical AI is the hottest capital cycle in tech — tens of billions of dollars flowing in — concentrated in the model / “brain” layer.
- Models are commoditizing; the deployment, evaluation, and learning-loop layer is under-built and under-funded.
- The winners will close the real-world loop the fastest and cheapest — precisely the layer we operate, from a seat (physical footprint, hardware channel, data operations) that software-only and model-only players cannot reach.

09

STATUS

What Is Built vs. Roadmap

BUILT — LIVE TODAY

· Hardware distribution and integration; managed teleoperation and egocentric collection; a refinement pipeline (QC, subtask segmentation, packaging).

· VLA model registry and inference sessions (shadow / execute) that record actions and outcomes.

· An evaluation-runs API that scores policies from real rollout telemetry, with pass-criteria gates and readiness reports; operator-in-the-loop review; CLI and API access.

ROADMAP

· Fully automated policy scoring across arbitrary tasks; automatic failure-mode clustering; eval leaderboards.

· Edge deployment and optimization (compile / quantize / serve VLAs on-robot) — motivated by the latency reality of real-time control.

· Uncertainty-aware automated active collection.

10

ABOUT

Robotics Center of Silicon Valley

Robotics Center of Silicon Valley (YC S25). One stack for Physical AI: hardware, data, and evaluation. We host robotics builders at 90 Welsh St, San Francisco, every Friday.

Contact: **contact@roboticscenter.ai** · roboticscenter.ai · 90 Welsh St, San Francisco, CA 94107

The winners in Physical AI will not have the biggest models. They will run the fastest, cheapest real-world learning loop. RL² is the infrastructure that gets them there.