

RL²

真实世界强化学习 · Reinforcement Learning in Real Life

把机器人策略从“实验室能跑”带到“真实产线能用、且能赚钱”的部署-学习闭环。在物理 AI 里，模型能力正在商品化；持久的护城河是真实世界的学习闭环。

摘要

物理 AI 的瓶颈不在模型能力，而在于"实验室能跑"到"真实产线能用、且能赚钱"之间的那个闭环。本白皮书定义了这个闭环——**RL²（真实世界强化学习，Reinforcement Learning in Real Life）**：部署 → 捕获失败 → 对照真实验收标准评测 → 定向采集数据 → 再部署。我们论证：模型会商品化，而部署-学习闭环才是持久护城河；并给出运行这一闭环的一体化技术栈（硬件·数据·评测）以及量化生产就绪度的方法——文末诚实标注"已上线"与"路线图"。

01

问题

物理 AI 为什么卡住

在湾区机器人生态的中枢运营两年——为前沿实验室、拿到融资的创业公司和隐身团队提供硬件、数据与集成——一个规律无法忽视：卡点几乎从来不是模型能力，而是"实验室能跑的策略"到"真实部署里能活下来、且能赚钱"之间的那个闭环。两道坎让团队停滞。

1.1 sim-to-real 鸿沟：三个失败模式

- **仿真 ≠ 硬件现实**。在 Isaac Sim / MuJoCo 里训练的 VLA 一撞上真实执行器极限就露馅——热漂移、关节回差、传感器漂移。同型号的两台机器，动力学都不一样。
- **遥操作数据本质是 off-policy**。人类示范能给模型开个头，但模型不是人。这个分布差距靠"更多同样的数据"补不上，只能靠 on-policy rollout + 失败捕获 + 定向采集来收敛。
- **"能用"在部署前是未定义的**。团队盯着实验室成功率优化，然后在长尾里翻车——起皱的衬衫、下午四点的反光、偏了 3cm 的托盘。没有真实世界的 eval harness，"就绪"只是猜。

1.2 从试点到量产的经济墙

就算一个策略过了所有 eval，它仍是成本中心：CapEx、维护、操作员、停机。大多数机队不是死在实验室，是死在试点阶段——单位经济性始终算不平。物理 AI 不只有模型问题，还有商业模式问题。任何真实世界闭环都必须便宜到能持续跑，否则技术再好也会停滞。

闭环才是护城河

模型会商品化，闭环不会。物理 AI 的价值，归于能最快、最便宜关闭真实世界学习闭环的那一个。

大约 18 个月内会出现几个"够用"的开源 VLA，前沿实验室也会把其余的免费送出去以卖算力。价值沉淀在不会商品化的那一层：把原始策略变成已部署、可靠、经济可行系统的部署-学习闭环。

墙	2026 年状态	在哪一层被解决
模型能力	正在商品化	不是我们这层
Sim-to-real	技术墙	RL ² 闭环
试点到量产的经济性	真正杀死公司的墙	闭环 + 单位经济性

同一个机制能同时清掉这三堵墙：一个便宜到能持续跑的真实世界学习闭环。

我们把这个闭环称为 **RL² — 真实世界强化学习 (Reinforcement Learning in Real Life)**：用真实世界的结果、而非仅靠仿真奖励，来持续改进已部署策略的循环过程。



图 1 - RL² 闭环：在真实硬件上持续学习

3.1 什么让它是"学习"而不是"囤数据"

- **On-policy**，而不是更多人类示范。在真实硬件上跑当前策略，捕获它在哪失败——而不是人在哪不一样。
- **按成因诊断失败**。按感知 / 接触-动力学 / 规划 / 新物体聚类，再采集针对成因的数据。
- **不确定性感知采集**。策略标记低置信 / 分布外状态并交接；这些状态正是要采的定向数据。
- **人类接管即监督信号**。机器人撞上新的不确定性时，人接管——这个接管就是纠正信号。
- **Eval 才是 ground truth**。由留出的真实世界变化、而非训练 loss，判断策略是否真的泛化。

3.2 如何处理"新的不确定性"

你无法为未知预训练。诚实的答案是：**检测它 → 安全降级（交给人）→ 捕获它 → 快速闭环**。真正重要的指标不是"永不意外"，而是"关闭一个新失败模式所需的时间"（time-to-close）。

系统架构：一体化技术栈

RL² 跑在一个一体化的技术栈上。多数团队从各家供应商拼凑这几层；把它们作为一个闭环整体拥有，才是差异化所在。



图 2 - 一体化技术栈；闭环是唯一不会被商品化的一层

4.1 硬件层

渠道与获客。领先平台（AgileX、Unitree、AGI Bot、Booster）的美国官方分销，加上我们自研的 OpenArm 与 Mobile ALOHA。已经在客户手里的机器人，就是闭环运行的部署面。

4.2 数据层

托管式遥操作与第一人称（egocentric）采集；精炼管线（时间同步、QC、子任务分割、打包）；以原始 ros2 rosbag（MCAP）交付，可选转 LeRobot v2 / HDF5 / RLDS。

4.3 评测层

就绪门禁与可观测性：把验收标准编码，让各策略版本跑一遍，产出 readiness report，并把失败反哺定向采集。

实验室成功率不等于就绪。我们把客户真实的验收标准——他们在生产中被考核的具体任务、环境和边缘情形——编码成一个可在每个策略版本上重跑的 eval。

- **自动回归门禁**——跨版本追踪指标阈值，回归在上线前就被拦下（相当于给策略的 CI gate）。
- **Operator-in-the-loop 评测**——物理任务难以全自动打分；我们的操作员网络跑真实 rollout，按标准判定结果。
- **失败驱动的数据定向**——每一次失败 rollout 都指向下一步该采哪个场景、物体或条件。

示例 readiness report（示意）：

字段	值
模型	example-vla v0.3
状态	已评分 — 3 个已完成会话，52 个决定性动作
成功率	88.5%（46 通过 / 6 失败）
平均动作延迟	161 ms
Top 失败模式	operator_rejected ×4, timeout ×2
验收门禁	PASS（成功率 ≥ 85%，延迟 ≤ 200 ms）

仅为示意样例，非客户数据。跨任意任务的全自动策略打分在路线图上（见第 9 节）；当前就绪度由自动门禁 + operator-in-the-loop 复核共同得出。

数据是闭环的燃料——但通用数据是商品。我们的数据引擎之所以有护城河，是因为由 **eval** 决定采什么。

- **失败驱动的定向采集。** eval 的失败定义下一批采集——瞄准真实失败模式，而不是"更多同样的"。
- **托管操作员。** 经培训、签 NDA 的专员，通过 gold-standard 参考 episode 认证，跨人、跨班次、跨站点做一致性校验。
- **自动 QA 套件。** 时间同步/完整性、图像质量、任务完成（对照给定标准）、运动学合理性、标签/元数据、分布覆盖——自动初筛 → 人工复核，被拒 episode 会重采，保证交付可用小时数达标。
- **格式。** 原始 ros2 rosbag (MCAP)，可选转 LeRobot v2 / HDF5 / RLDS 训练就绪格式。

一个贯穿性示例

一个有代表性的操作任务——从堆栈取箱、翻转、放上传送带——把闭环从头到尾走一遍：

1. **部署**：在真实工位跑当前策略。
2. **捕获**：按子任务记录 rollout；标记失败（抓取失败、翻转失败、超时）。
3. **评测**：对照验收标准打分，产出 readiness report。
4. **定向**：失败集中在“反光下的翻转”——就采这个条件的定向示范。
5. **再部署**：重训、过 eval 门禁、再跑；追踪该失败模式的 time-to-close。

数值为示意；重点是机制，不是数字。

为什么是现在

- 物理 AI 是科技领域最热的资本周期——数百亿美元涌入——且高度集中在模型/"大脑"层。
- 模型正在商品化；而部署、评测、学习闭环这一层却建得不足、投得不足。
- 赢家会最快、最便宜地关闭真实世界闭环——正是我们运营的这一层，而且是从一个软件-only、模型-only 玩家够不到的位置（实体落脚点、硬件渠道、数据运营）。

已上线（今天可用）

- 硬件分销与集成；托管遥操作与第一人称采集；精炼管线（QC、子任务分割、打包）。
- VLA 模型注册 + 推理会话（shadow / execute），记录动作与结果。
- evaluation-runs API：从真实 rollout 遥测给策略打分，带 pass-criteria 门禁与 readiness report；operator-in-the-loop 复核；CLI 与 API 接入。

路线图

- 跨任意任务的全自动策略打分；失败模式自动聚类；eval 排行榜。
- Edge 部署与优化（在机器人上编译 / 量化 / 服务 VLA）——由实时控制的延迟现实驱动。
- 不确定性感知的自动主动采集。

硅谷机器人中心（Robotics Center of Silicon Valley, YC S25）。物理 AI 的一体化技术栈：硬件、数据、评测。我们每周五在旧金山 90 Welsh St 接待机器人 builder。

联系方式： contact@roboticscenter.ai · roboticscenter.ai · 90 Welsh St, San Francisco, CA 94107

物理 AI 的赢家不会是模型最大的那个，而是能最快、最便宜地跑通真实世界学习闭环的那个。RL² 就是把他们送到那里的基础设施。
