Autonomous Earthmoving: Classical Control Guidance Makes Long-Horizon Trenching *Reliably* Trainable

Evelin Fonseca Cruz · Joe Hewlett · Laszlo Kovacs · Sai Krishna Gottipati · Adam Bognat · Manhtriet Hoang · Clodéric Mars
CM Labs Simulations, Montréal, Canada | evelin.fonsecacruz@cm-labs.com

1. Problem & Task Decomposition

We examine excavator trenching for three reasons: [6] it is a representative earthmoving task; soil responds far less predictably than the arm's own geometry; and the horizon is long — what the bucket takes at the dig phase sets the load for every stage after it, and with it the number of cycles needed to finish the trench.

Neither approach survives on its own. Model-free RL sees reward long after the actions that earned it — *reward sparsity*, credit assignment failure, divergence. Classical control is untroubled by the horizon, but cannot anticipate the soil.

	Classical Control (IK)	Model-Free RL
Pros	Predictable, smooth trajectory tracking in free space.	Learns complex, non-linear dynamics without manual gain tuning or explicit system modeling.
Cons	Approximate unless tuned per machine, attachment and scenario.	Highly unstable when unguided; prone to cold-start exploration failures.

Core research question. Can Classical Control Guidance (via Curriculum Learning & Residual RL) accelerate Model-Free RL training and yield superior task execution on earthmoving tasks?

6-Stage Decomposition. Excavation cycles natively follow a sequential kinematic sequence (Start, Accommodate, Dig, Lift, Swing, Dump, Return). Segmenting into 6 distinct subtasks standardizes execution and prevents compounding geometric errors.

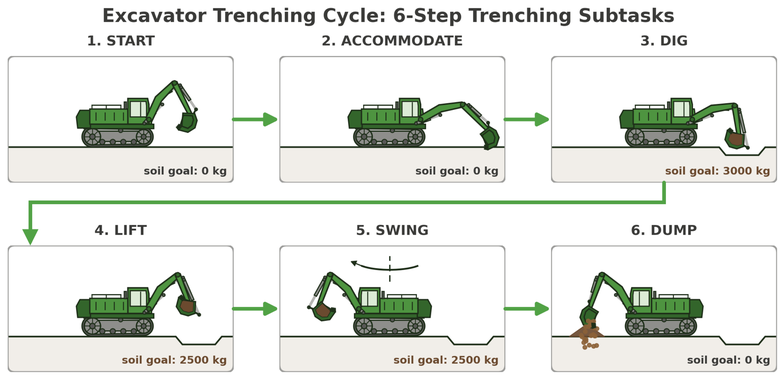

Fig. 1 The six-step trenching cycle. Each stage has one checkable goal — a bucket pose and a payload (§2) — and a run succeeds only when **all six complete in sequence**.

2. Learning Methodology

Segmenting the cycle into six short, goal-conditioned episodes keeps credit assignment tractable; within each, a classical IK command and a learned policy are blended into one action [1, 2]. The RL share of authority λ is either held **fixed** for a whole run, or **annealed** from IK toward RL as training proceeds — an **A**ction **B**lending **C**urriculum over **R**esidual **R**einforcement **L**earning (**ABC-RRL**).

The IK controller is untuned — a resolved-rate Jacobian controller on a trapezoidal velocity profile, with no per-task tuning. It is guidance for RL to improve on, not a baseline to beat.

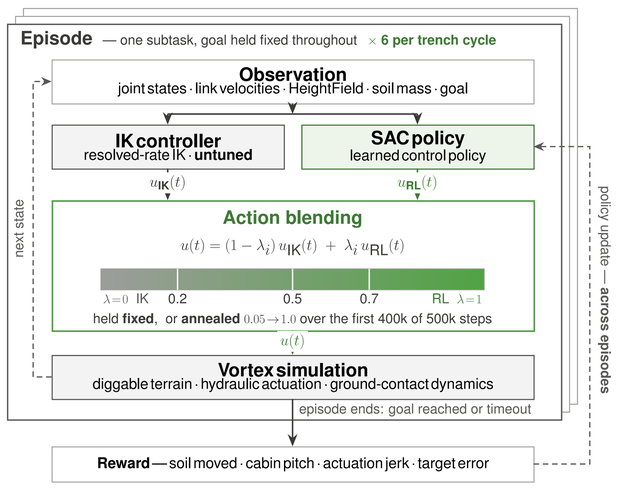

Fig. 2 One goal-conditioned episode. λ is the RL share of authority: at $\lambda = 0$ the action is the IK command alone, at $\lambda = 1$ the policy has the machine.

Goal conditioning. One policy covers all six subtasks: no subtask IDs are passed, only the dual target vector $[e_t, m_{\text{target}}]$.

Spatial target error	$e_t = x_{\text{bucket}}(t) - g_i$ for active goal g_i
Payload target & tolerance	3000±1000 kg (*Dig*) · 2500±800 kg (*Lift*) · 2500±1000 kg (*Swing*) · 0 kg elsewhere
Stage complete when	bucket within **0.5** m and **0.1** rad of the goal *and* payload inside the stage tolerance
On subtask failure	Subtask timeout triggers early termination
Reward	R_t rewards step-wise progress in bucket pose (R_{bucket}) and soil payload (R_{soil}) relative to previous step values, penalizes duration to optimize cycle time, and awards a completion bonus on subtask success

3. Platform & Experimental Setup

The simulation platform is built on Vortex Studio. The excavator is a production model used in commercial operator training simulators, not a reduced stand-in built for RL. Hydraulic actuation, ground contact and terrain deformation [5] behave as they do for human trainees.

The Vortex Environment Interface (VEI) exposes a running Vortex simulation as a standard, synchronous RL environment: one gRPC stream, fields addressed by name, deterministic stepping.

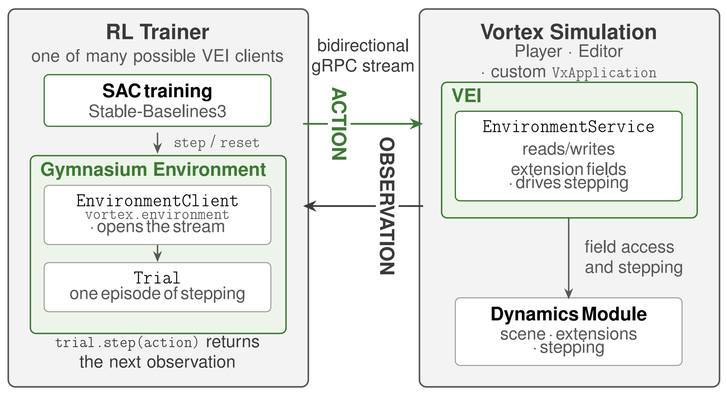

Fig. 3 The Vortex Environment Interface. One bidirectional gRPC stream; the RL trainer shown is one possible client, not the only shape.

Action	Box(4) — joint commands *Swing, Bucket, Stick, Boom*, continuous $[-1, 1]$
Observation	Box(159) — joint positions, linear & angular link velocities; world transforms for turret, arm, boom, chassis; HeightField array (Vortex Earthwork); and the stage's **target** payload — not the measured one, which feeds only the reward and the success test
Agent	SAC (Stable-Baselines3 [3]), chosen for stability on continuous action spaces; `MlpPolicy` 2×256, lr $3{\times}10^{-4}$, batch 256, buffer 10^6, $\gamma = 0.99$, $\tau = 0.005$, auto entropy; 500k steps per configuration, ≈24 h of wall clock each
Control rate	trained at **6 Hz** (one action per 10 simulation frames), evaluated at **60 Hz** — faster control than the training rate improved performance
Evaluation	**3 seeds** per configuration — 15 models, **150 episodes** each. Spread under ~5% is episode sampling, so no error bars [4]. Two of the 15 are earlier best checkpoints, not 500k

4. Results & Insights

3/3 seeds converge **with guidance** — against **1 in 3** under full RL authority with none

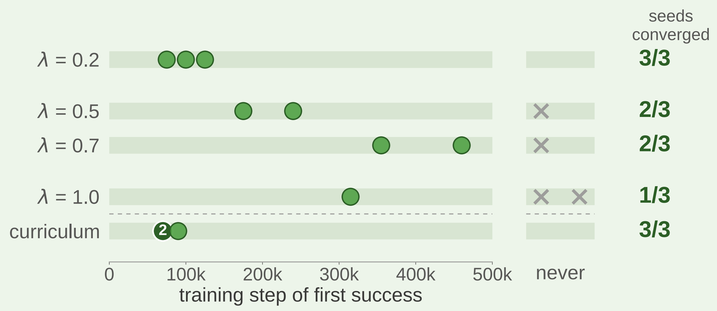

Fig. 4 Reliability falls as policy authority rises. One dot per seed, numbered where seeds coincide; λ to scale. × = never succeeded inside 500k.

0% of cycles completed by the **untuned IK controller alone** — it stalls at *dig* on 849 kg of a ~2000 kg target

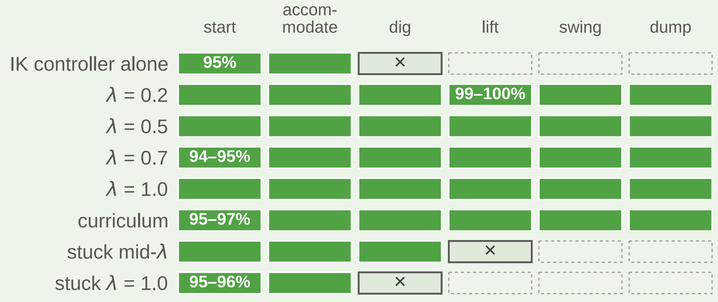

Fig. 5 Each method stops at a different link. Episodes completing each stage, pooled over the row's seeds; a range where it is not 100%. × = stops here, dashed = never reached.

6× sooner to a working policy: the curriculum's median run passes 50% success at **75k** steps, $\lambda = 0.7$'s not until **467k**

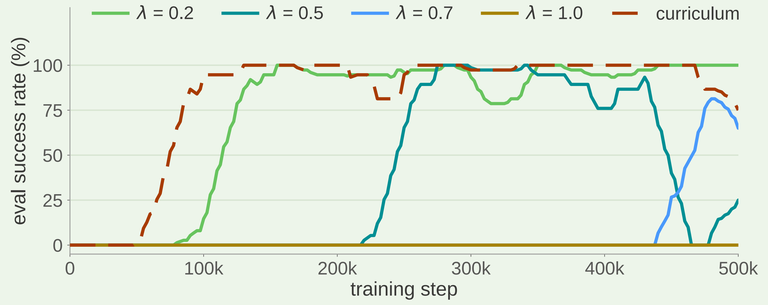

Fig. 6 Guided runs rise earlier. Median over all three seeds, 15-point rolling mean. At $\lambda = 1$ it never leaves zero.

Conclusions

Neither extreme is reliable. The untuned IK controller never completes a cycle — it stalls at *dig* on 849 kg. Unguided RL is not impossible, but only 1 seed in 3 converged; the other two dug 24–76 kg, worse than no policy at all.

Guidance removes the downside, not the ceiling. The best single model here is unguided ($\lambda = 1$, 100%). What guidance buys is that *every* seed gets there — 3/3 at $\lambda = 0.2$, and first success 3–4× earlier.

The curriculum keeps that reliability at full autonomy. Annealing λ from 0 to 1 gave 3/3 deployable policies at 94.7–97.3% — within sampling noise of the fixed blends, with the IK controller switched out by the end.

Future Work

Richer reward shaping; other earthmoving tasks, longer horizons and other machines; then sim-to-real transfer. Everything here is simulation, on one soil model and one trench.

References [1] Dragan & Srinivasa, *A Policy-Blending Formalism for Shared Control*, IJRR 2013. · [2] Schaff & Walter, *Residual Policy Learning for Shared Autonomy*, RSS 2020. · [3] Raffin et al., *Stable-Baselines3*, JMLR 22(268) 2021. · [4] Agarwal et al., *Deep RL at the Edge of the Statistical Precipice*, NeurIPS 2021. · [5] Holz, Beer & Kuhlen, *Soil Deformation Models for Real-Time Simulation*, VRIPHYS 2009. · [6] Egli et al., *Soil-Adaptive Execution of Excavation Tasks*, IEEE RAL 2022.