WHAT YOU MEASURE, ABLATE, AND POOL DECIDES THE CONCLUSION: AUDITING SENSORIMOTOR PRETRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

Statistical rigour does not protect against instrument artefacts. We report a preregistered multi-body audit of a developmental sensorimotor curriculum — four MuJoCo MJX embodiments (DM humanoid, Unitree G1, Go2 quadruped, Franka Panda), a 318-run grid plus follow-ups, Holm families, bootstrap CIs, TOST — in which three routine choices each independently changed what we concluded: which *metric* you read, how you *ablate* a component, and whether you *pool* across bodies.

The first indicts our own instrument. The curriculum's headline training-time advantage (Hedges g up to 4.4, raw ratios up to $195\times$) survived preregistration, four embodiments and Holm correction, yet a gain-matched diagnostic shows it is largely manufactured by the curriculum's own action-gain schedule. A fixed evaluation gain amplifies the curriculum arm's measured motion by 15–$26\times$ and the control arm's not at all; removing that asymmetry removes 74–93% of the raw ratio (29–63% of the log-ratio). Measured at each policy's own training gain on a corrected trainer (six instrument fixes), the curriculum improves the proxy on **0/4** bodies. Nor does it transfer, at final performance: on **none of four** downstream routes does the curriculum improve on its own ablation, and on the frozen routes the audited intrinsic-motivation objectives (curriculum, empowerment, prediction-error; ICM/RND) sit at or near *random init*, while a self-supervised dynamics control, itself reward-free, transfers near-ceiling and none overtakes random init at a $10\times$ pretraining budget (P27; ICM health caveat in Appendix K).

Two further lessons reach beyond this system, on less direct evidence. First, ablating prediction error by *reward weight* understates it: removing the *component* shows it suppresses motion ($g = -3.0/-1.2$); under one matched evaluation the gating gap on H1_dm is larger by 21 $[13.5, 28.0]$ $|q_{\text{vel}}|$ units, and on Go2 the raw difference is borderline; the representation effect reproduces on shared-encoder ICM and, dose-dependently, RND. Second, a cross-body diagnostic that looks predictive pooled ($\rho \approx 0.5$) is a leave-one-body-out Simpson's paradox (every diagnostic's mean $R^2 < 0$, best -0.19). That pooling failure reproduces on *released benchmark data*: naive pooling erases 85% of a real within-task Atari advantage, and under URLB's own seed noise no "best" unsupervised-RL method can be crowned. Running that benchmark's standard pipeline ourselves (144 fine-tunings), no method tops all three of its domains at either seed budget we ran; the sharper claim — that two poolings of the same runs disagree about the winner — held at three seeds and *failed* at five, and we print that failure rather than the budget that suited us. We derive both failures as properties of the estimators, and release the protocol, per-seed data, and a checklist whose first row is the one we needed: check that a headline effect is not produced by the instrument that measures it.

1 INTRODUCTION

The Active Sensorimotor Curriculum (ASC) framework proposes that an embodied agent can acquire motor competence *without extrinsic reward*: a developmental body-range-of-motion (ROM) curriculum is combined with intrinsic-motivation signals, typically a multi-step empowerment lower

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

bound (Klyubin et al., 2005; Mohamed & Rezende, 2015; Belghazi et al., 2018) and a prediction-error / curiosity term (Pathak et al., 2017; Sekar et al., 2020). The framework is attractive. It echoes developmental robotics (Cangelosi & Schlesinger, 2015) and curriculum learning (Bengio et al., 2009), and it composes well with modern unsupervised RL pretraining (Laskin et al., 2021).

But ASC bundles several mechanisms, and whether any of them "works" hinges on three choices the community routinely makes by default, any one of which can flip the conclusion. **Which *metric* you read:** the training-time motion proxy $|q_{\text{vel}}|$, or zero-shot downstream task success? **How you *ablate* a component:** zero its reward weight, the usual shortcut, or actually remove it from representation learning? **Whether you *pool* across bodies:** does a diagnostic validated on pooled cross-body data survive a within-body, leave-one-body-out decomposition?

Reporting a large ablation gap on *one* body and *one* metric, *pooled* where convenient, is the single-cell trap the RL-reproducibility literature warns about (Henderson et al., 2018; Agarwal et al., 2021; Bean et al., 2025), sharpened here to a multi-body setting. We adopt that literature's inferential discipline: effect sizes, bootstrap CIs, corrected families. We part from it on *what the estimate is for*. A stratified bootstrap (rliable-style) robustly estimates an *aggregate*, but an aggregate is not a cross-body *prediction*. When bodies are structurally non-exchangeable (different action dimensions, dynamics, reward scales), the pooled correlation is dominated by its between-body component. It can be large with *zero* held-out-body predictability, and stratifying the bootstrap only tightens the CI around that misleading number; we verify this in simulation (P23). A cross-body *predictive* claim must instead be certified by *leave-one-body-out* generalization (C3; we use a stratified IQM only as a within-body robustness check, §3.1). Leave-one-*domain*-out evaluation is standard in domain generalization (Gulrajani & Lopez-Paz, 2021; Koh et al., 2021). Our delta is that in embodied RL the routinely reported cross-body *aggregates* answer a different question than LOBO does, and that implementation-level choices decide conclusions here as they do in Engstrom et al. (2020). To our knowledge this is the first *measurement-validity audit across embodiments*: reliable-RL standardizes aggregation *within* a domain (Agarwal et al., 2021), and URLB (Laskin et al., 2021) benchmarks transfer, but neither asks whether a measurement survives a change of body.

Our deliverable is a *validated, executable protocol*: downstream transfer, within-body decomposition and component gating, each diagnostic a released implementation, validated by a self-supervised dynamics positive control (§3.1) and distilled into a one-table checklist (§5.1). ASC is the demonstration vehicle. Applied to it on a preregistered multi-body grid, each choice independently changes what one would conclude (Figure 1). The *ablation* and *pooling* failures *reproduce off this system*, on shared-encoder ICM/RND (§3.2) and on released benchmark data (§4). The *metric* artefact reproduces on a six-fix corrected re-run (§3(e)) and, as a generic gain-mismatch effect, in a trainer-independent simulation with no ASC code (§3(d)): a saturated policy's measured motion scales as $1/g_{\text{train}}$, while a genuinely competent controller is gain-invariant.

The ROM curriculum has a precise antecedent. Bernstein described skill acquisition as the *freezing* of biomechanical degrees of freedom and their progressive *freeing* (Bernstein, 1967; Vereijken et al., 1992), which a monotonically widening actuator control range approximates computationally; morphology-aware curricula are an active line (Cangelosi & Schlesinger, 2015; Badie et al., 2025). Our result cautions against reading that lineage into a proxy effect: the effect we measured under a multi-body, multiple-comparison-corrected discipline was an artefact of how the curriculum is implemented and evaluated (§3(d)), a failure mode single-body studies could not have surfaced.

Three questions were locked in a public preregistration before any training run: whether curriculum dominance generalizes across bodies (H1), where the curriculum's effect lives (H2), and whether a single training-time statistic of the representation predicts competence out-of-body (H3, with a preregistered leave-one-body-out accept rule $R^2 \geq 0.5$). H1 is met on all four bodies on the proxy, H2 yields a body-graded mechanism, and H3 is rejected. The audit below shows why the first two of those answers should not be trusted.

Our contributions are a **reusable measurement protocol** for multi-body developmental RL, demonstrated by the four ASC case studies that motivate each of its rules. Relative to the reliability literature, two findings are new: an instrument artefact produced by a curriculum's own action-gain schedule (C1), and a body-identity Simpson's paradox that a stratified bootstrap cannot detect (C3). The evidence is not uniform: C1 rests on a direct intervention, C2 on ablations compared across evaluation regimes, C3's external half on descriptive benchmark analyses. Across all 4 bodies (H1

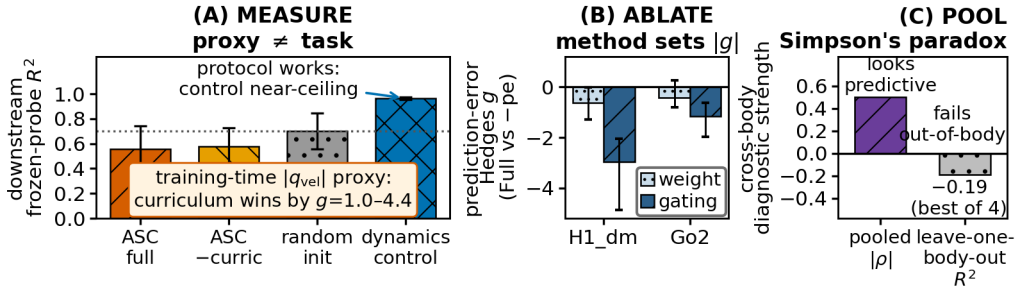

Figure 1: **One method (ASC), three measurement choices, three changed conclusions. (A) Measure ($C1$):** downstream frozen-probe R^2 (mean over the three locomotor bodies). On the training-time $|q_{\text{vel}}|$ proxy the curriculum wins by $g = 1.0$–4.4, yet downstream ASC (0.56) is no better than *random* init (0.70) — while a self-supervised dynamics *control* representation is near-ceiling (0.97). Two of the probe's three targets are functions of its input, so we read it as information preservation rather than task relevance (Appendix K); the "not the yardstick" argument rests on the RL-transfer battery, where the same control also wins. **(B) Ablate ($C2$):** prediction-error Hedges g (Full vs $-$pe, 95% CIs). Reward-weight ablation gives $-0.64/-0.43$, not significant after Holm; true component gating gives $-2.96/-1.16$, significant — same direction, a different verdict. Re-evaluated under one matched regime, gating's Full$-(-$pe) gap is larger by $21.0\,[13.5, 28.0]$ $|q_{\text{vel}}|$ on H1_dm; on Go2 by $2.05\,[-0.01, 4.05]$ (raw CI touching 0; $\Delta g = -0.73\,[-1.74, -0.08]$). **(C) Pool ($C3$):** a cross-body diagnostic looks predictive when pooled ($|\rho| \approx 0.5$) but fails leave-one-body-out (mean $R^2 < 0$ for every diagnostic) — a body-identity Simpson's paradox.

is met on every morphology under the realized m=12 family, Panda at n=20) and a 318-run grid plus follow-ups:

1. **(C1) The headline effect was manufactured by our own instrument.** The ROM curriculum produced a large, body-robust, apparently compute-irreplaceable effect on the training-time $|q_{\text{vel}}|$ proxy ($g = 1.0$–4.4, the full preregistered H1). A gain-matched diagnostic attributes most of it to the curriculum's own action-gain schedule: removing the eval-gain asymmetry removes 74%/93% of the ratio and 29%/63% of the log-ratio. A corrected re-run makes it null: the curriculum improves the proxy on **0/4** bodies (negative point estimates on three, which we do not over-interpret; §3(d,e)). Independently, whatever the proxy measures **does not transfer under four downstream measurements**, each comparing the curriculum with its own ablation: zero-shot Tier-1 (0/8 cells), a frozen probe (0/3 bodies), and RL fine-tuning (0/2 modes each) on a reach and a locomotion task. A self-supervised dynamics control transfers near-ceiling on the identical protocol (§3.1). A preregistered, corrected, multi-body result can still be an instrument artefact.
2. **(C2) Ablation-method dissociation: the method sets what is detected.** Ablating intrinsic motivation by *reward weight*, the usual shortcut, understates what ablating the *component* reveals. Empowerment shows no detectable effect, but *removing* prediction-error makes the policy move substantially *more* ($g = -2.96$ H1_dm, -1.16 Go2; G1 ns; original instrument): prediction-error *suppresses* motion. Weight ablation points the same way; under matched evaluation the gating gap is larger by 21 $|q_{\text{vel}}|$ units on H1_dm (CI $[13.5, 28.0]$) and borderline on Go2: a change of size, not sign. The representation effect reproduces on shared-encoder ICM and RND with a monotonic dose-response wherever the shared encoder carries task information (§3.2).
3. **(C3) Pooling dissociation: Simpson's paradox** (Yule, 1903; Simpson, 1951)**.** A training-time diagnostic looks moderately predictive of competence when pooled across bodies ($|\rho| \approx 0.5$) but fails completely under leave-one-body-out regression (every diagnostic's mean $R^2 < 0$, best -0.19). The pooled signal is a body-identity artefact (§4). H3 is rejected against the *downstream Tier-1* target too (max LOBO $R^2 = -3.5$, all four diagnostics; §4), so the failure is not a proxy artefact: no training-time diagnostic predicts out-of-body competence. Two further ways pooling misleads appear *outside* our system: on released Atari data it attenuates a real within-task advantage by 85%, and in our own

rerun of the URLB pipeline the mean and rank aggregates answer different questions — no method tops all three of its domains at either seed budget (§4).

4. **(C4) A reusable measurement guide.** We package the dissociations into a concrete guide: check that an effect is not produced by the measuring instrument, measure task not proxy, ablate components not reward weights, decompose pooled claims. It is backed by git-locked decision rules, Holm-locked families, a seed-power calculator, and full per-seed data release (§6). Its rows are recommendations with the evidence of one audit behind them, not a validated standard.

2 SETUP

Bodies, trainer and ROM mask. Four MuJoCo MJX (Todorov et al., 2012) bodies (DM humanoid, Unitree G1, Unitree Go2, Franka Panda; menagerie pinned by SHA; Appendix Table 7) span two humanoids, a quadruped and a manipulator. ASC as audited here is our own composition (M5.2 pilot design, Appendix G), not a reproduction of a published implementation; the components are standard, the combination is ours. The trainer is the zero-extrinsic-reward ASC loop: a shared state encoder, a stochastic actor, a critic regressed on the intrinsic reward, a one-step empowerment surrogate and a forward predictor, over 512 parallel envs for 1000 steps. The ROM curriculum multiplies the actor's output by a schedule-dependent mask; for the humanoids and the manipulator that mask is a *uniform scalar gain* rising to 0.4375 by the end of training, a detail that turns out to matter (§3(d)). Appendix G gives the full specification, including per-body actuator counts, hyperparameters and the two ablation gating modes.

Evaluation. The primary H1 outcome is a *training-time proxy*: mean $|q_{\text{vel}}|$ across the 512 training envs at full ROM ($= 1$) over the final 100 sim-steps. Two eval regimes exist and the distinction matters (§3(c)): *terminal-state* (continuing from where training ended) vs *fresh-reset*; unless stated, §3 reports terminal-state and §3.2 fresh-reset. We additionally report the preregistered *downstream Tier-1* evaluation — three body-matched zero-shot tasks per body against n=20 RANDOM baselines (§3.1) — against which H3 is tested, as preregistered.

Statistical protocol and a deviation we state up front. The confirmatory comparisons use Welch's t-test, Mann–Whitney U, Hedges' g (Hedges, 1981) with 20,000-resample bootstrap (Efron, 1979) 95% CIs, and Holm-Bonferroni FWER correction with $\alpha = 0.05$ applied to the Welch p. The benchmark audits resample less (B=10k, 2k) and the LOBO fits report no interval; Appendix K.1 lists the exceptions. The preregistration (`d67f51d`, before any P1+ run) defined a *primary family* of $|q_{\text{vel}}|$, Tier-1, H2a and H2b contrasts. **As realized, our Holm-corrected primary family is the $|q_{\text{vel}}|$ contrasts only (the m=12 family of 3 ablations $\times$ 4 bodies); the Tier-1 and H2 contrasts were not run as part of a single confirmatory family.** We label H2/H2c and the scaling/transfer follow-ups exploratory or post-amendment, in place. Seed counts: n=20/arm for the primary $|q_{\text{vel}}|$ cells, n=12 for mechanism cells, n=4–8 for the exploratory empowerment-estimator sweep. This family narrowing affects borderline calls (notably G1, §3). Codes of the form Pn name released run logs and their locked protocols; Appendix K maps each to its experiment.

3 CURRICULUM DOMINANCE (H1) — AND ITS RETRACTION

Read this section as the seemingly-rigorous positive we set out to establish and then retract: (a)–(c) are what a single-body study would report and believe; (d)–(e) show it is an instrument artefact (its mechanism, dissected in Appendix A, is the artefact's structure).

The primary headline. For each completed body we run the Full configuration (curriculum on, α=5, β=0.5) and the $-$curric ablation (ROM $= 1$ from t=0), 20 seeds each, and report the contrast on training-time $|q_{\text{vel}}|$. The full-sample-size Holm-corrected primary-family results are in Table 1.

Two observations matter.

(a) The preregistered four-body H1 is met. Every body shows $g \gg 0$ with bootstrap 95% CI excluding 0 and Holm-corrected $p < 0.05$, meeting the per-body sub-criterion ($|g| > 0.8$ *and* CI

Body	Full	−curric	Hedges g	g 95% CI	Holm p	Ratio (95% CI)	Decision
H1_dm (humanoid)	10.39 ± 8.91	0.053 ± 0.033	**+1.607**	[+1.29, +2.21]	**0.0006**	**195× [120, 303]**	✓
Go2 (quadruped)	7.21 ± 1.03	0.896 ± 1.71	**+4.389**	[+2.99, +8.90]	**<0.0001**	**8× [4.2, 28.6]**	✓
G1 (humanoid)	23.30 ± 14.63	12.38 ± 3.98	**+0.998**	[+0.72, +1.47]	**0.0357**	**2× [1.4, 2.5]**	✓
Panda (manip.)	4.41 ± 1.21	1.99 ± 2.75	**+1.121**	[+0.45, +2.35]	**0.0126**	**2× [1.3, 5.0]**	✓

Table 1: H1 primary-family results: Full vs. −curric on training-time $|q_{\text{vel}}|$.

excludes 0 *and* Holm-significant) — including the manipulator. (At an earlier, partial Panda cell of $n \leq 9$ this contrast was underpowered, g=0.38, Holm p=1.0; completing Panda to n=20 resolves it: g=1.12, Holm p=0.013.) At n=20/arm the a-priori MDE is $g \approx 0.91$ uncorrected, ≈ 1.23 at the family's worst-case $\alpha/12$ (power 0.8); we report MDE, not observed power (a monotone transform of p). **All four morphologies — two humanoids, a quadruped, and a manipulator — satisfy the locked H1 criterion**, so the curriculum's effect on the $|q_{\text{vel}}|$ proxy generalizes across the full preregistered body set. We flag that **G1 is the borderline pass** ($g = 0.998$, Holm $p = 0.036$ — below the family-corrected MDE, implied power ≈ 0.55), whereas Go2/H1_dm/Panda pass with comfortable margins (heterogeneity $I^2 = 92\%$). This 4-body result is on the *proxy*; §3.1 shows it does not transfer.

(b) Effect magnitude is body-dependent. The primary magnitude measure is the effect size (Hedges $g = 1.0$–4.4; per-body forest plot in Appendix Figure 3) and its log-ratio; raw Full/−curric ratios (2×–195×) are reported for completeness but should be read with care — the 195× on H1_dm has a denominator (0.053) at the noise floor, which inflates the ratio, and the preregistration retired raw-ratio headlines for exactly this reason. On the primary g/log-ratio scale the effect still varies markedly across bodies (Figure 4) — a spread any single-body study would miss. Larger-DoF G1 (29 actuators) shows a *smaller* relative gap than Go2 (12) or H1_dm (21); we discuss why in §5.

(c) The proxy magnitude is itself eval-protocol-dependent. Under a fresh-reset eval the same-checkpoint fold-changes compress sharply and the curriculum motion effect becomes null-to-small, so the raw terminal-state fold-changes are not competence gaps — the dissociation is unchanged either way (Appendix K, Figure 7).

(d) Most of the proxy contrast is produced by the schedule's action gain, not by acquired motion. The ROM mask is a *uniform scalar gain*: Full trains with actions scaled by 0.4375 while −curric trains at 1.0, yet both are evaluated at 1.0. A trainer-free gain diagnostic on the released checkpoints (P16; n=5/arm, exploratory — per-seed ranges in Appendix K) finds every arm saturated and parked at joint limits, and that removing this eval-gain asymmetry removes 74% (H1_dm) and 93% (Go2) of the training-terminal Full-vs-−curric ratio (29%/63% on the log scale our effect sizes use; H1_dm's share falls with a shorter roll-in, Go2's does not) (protocol and per-cell amplification in Appendix K). So the proxy effect is largely a measurement of who receives a torque boost at evaluation, not acquired motor competence. The mechanism is not specific to our trainer: in a minimal trainer-independent damped system (no ASC code; P22), a saturated policy's measured motion inflates by exactly $1/g_{\text{train}}$ at a mismatched eval gain while a genuinely competent controller stays gain-invariant — so the artefact needs no ASC code. Evaluating both arms at both gains from common, policy-independent states removes the advantage at either matched gain (Go2 1.01×, H1_dm 1.05×, CIs spanning 1; Appendix K).

(e) A corrected trainer — six fixes bundled — finds no effect. We rebuilt the trainer (`m15_asc_corrected.py`) with six instrument fixes (enumerated in Appendix K) and re-ran the full primary grid (4 bodies × 20 seeds, G1's −curric arm completing 10/20; P19). Evaluated at the gain each policy was actually trained at, **the curriculum improves the $|q_{\text{vel}}|$ proxy on 0/4 bodies** (point estimates negative on three — Hedges $g = -1.0\,[-1.8, -0.4]$ H1_dm, $-5.8\,[-8.9, -5.1]$ G1, $-1.9\,[-2.9, -1.3]$ Panda; Go2 not detected, $-0.43\,[-1.18, +0.17]$; bootstrap CIs, Holm/TOST in Appendix K — but we claim only the *absence of improvement*), versus the $g = 1.0$–4.4 the broken trainer reported on all four. A per-body positive-control battery (P24/P24b/P24c; Appendix K) grades the instrument itself. Scripted movers and three independently-trained Go2 walkers validate the proxy against stillness with monotone dose-responses on H1_dm, G1 and Go2, so on the three

bodies whose instrument passes its own control the null is 0/3; saturated controllers expose the same gain artefact. On the fixed-base Panda the proxy *fails its control outright*, so Panda's reading, and per-body competence generally, rests on the eval-independent transfer battery. Re-evaluating at full ROM flips two bodies positive, the discrepancy living *entirely* in the curriculum arm (Appendix K): the headline effect came from evaluating a range-restricted policy at an unrestricted gain.

3.1 METRIC: THE PROXY DOES NOT TRANSFER TO DOWNSTREAM TASK (C1)

The §3 result above establishes a large curriculum effect on the training-time $|q_{\text{vel}}|$ proxy. But the proxy is not the thing the field ultimately cares about. We preregistered a *downstream Tier-1* evaluation — three body-matched zero-shot tasks per body (stand / walk / reach), run on the zero-reward-pretrained policy with no fine-tuning — and report it here (P10b; Full / −curric / RANDOM, n=20 per arm on H1_dm, G1, Go2; Panda partial n=10, no claim).

The curriculum's proxy advantage does not transfer. Of the 8 discriminative (body, task) cells (Go2 `stand_z` is excluded as non-discriminative), Full vs −curric is Holm-significant on **0/8** (per-cell values in Appendix Table 8): every cell has Holm $p = 1.0$, and the effect sizes are small and frequently negative ($|g| < 0.4$ throughout; e.g. H1_dm `stand_z` $g = +0.40$, H1_dm `walk_dx` $g = -0.16$, Go2 `walk_dx` $g = -0.34$, G1 `reach_d` $g = +0.05$).

The null is not a broken-metric artefact. The same Tier-1 metric *does* discriminate trained-vs-RANDOM on several cells (H1_dm `walk` $g = +0.91$, Go2 `walk` $g = +4.11$; point estimates, not part of the tested family), so it can detect competence differences — it finds Full ≈ −curric. (Full vs RANDOM is itself body/task-specific: some cells score *below* random — e.g. G1 `stand`/`reach` — so the intrinsically-pretrained policies are not uniformly competent zero-shot.)

An eval-regime nuance, stated plainly. Zero-shot Tier-1 is evaluated at full ROM — exactly the −curric training regime — so it may *favor* −curric's distribution (Appendix K). A skeptic could still object that zero-shot is the wrong test, since the field's claim is that pretraining helps *after* downstream RL; we close that objection with three further measurements on the same checkpoints.

Confirmations 2–4 — nor under RL fine-tuning or frozen probing. Three further measurements agree, each reported in full in Appendix C: fine-tuning each pretrained encoder on a learnable Panda reach (frozen and end-to-end, n=10; P11) improves over −curric on **0/2** modes; a frozen-encoder probe across the three locomotor bodies (P11-C) is **0/3** (RANDOM marginally ahead); and a locomotion transfer on H1_dm, after making locomotion learnable (P14), is **0/2** with the curriculum the *worst* frozen source (g=−1.94; Go2 missed the learnability gate and is excluded). Across all four routes — zero-shot Tier-1 ($0/8$), frozen probing ($0/3$), and Panda-reach ($0/2$) and H1_dm-locomotion ($0/2$) RL fine-tuning — **the curriculum's apparent $|q_{\text{vel}}|$ advantage buys no downstream task competence**, and on the frozen probe it scores below random init with an MLP head (a linear head puts it marginally above). This is the metric dissociation (C1) — the headline reason to report task-level outcomes, not just the training-time proxy.

The protocol is not vacuous — it discriminates good representations (positive control). One might still object that *nothing* transfers in this setup, making the dissociation a property of our yardstick rather than of ASC. It is not. We ran the *identical* transfer and probe protocols on a control representation pretrained by self-supervised forward-dynamics regression on uniform-random rollouts — predicting the next-observation delta from the latent and action, with no reward, label or task signal, so it is as reward-free as the ASC arms (P12). This dynamics representation transfers *near-ceiling*: it is the *best* source on Panda-reach RL transfer (dyn-vs-full $g = +1.25$ frozen) and it tops the frozen probe. We rest the argument on the transfer result, because the probe is weaker than we first treated it as — two of its three targets are functions of the observation it decodes from, an untrained encoder already scores well on it, and a next-observation objective optimizes that preservation by construction (Appendix K). On the transfer routes the protocol still ranks dynamics ≫ random > ASC. The ASC proxy↔task dissociation is therefore a property of ASC, not of an insensitive evaluation — strikingly, ASC's representation sits *below* random init on the probe. This validates the protocol as a discriminator; read as a whole, the battery is a controlled comparison in which *only the direct-regression objective transfers*: the dividing line is not reward-freeness but

whether the representation is shaped by regression on transitions or by RL through an intrinsic reward, and no intrinsic-motivation source overtakes random init even under a $10\times$-budget control (P27; Appendix K).

Aggregated with rliable-style estimators. A stratified-IQM and probability-of-improvement recomputation (Agarwal et al., 2021) gives the same ordering with non-overlapping CIs; $P(\text{ASC} > \text{random})$ is at floor *within* every body while naive across-body pooling inflates it (C3 in miniature; Appendix K, Figure 5).

3.2 Intrinsic motivation: no detectable empowerment effect, prediction-error suppresses motion (C2)

Ablation method changes the conclusion. Under the original *reward-weight* ablation (zeroing α/β only), the intrinsic-motivation contributions are Holm-non-significant on every body — but this ablation is confounded, because the empowerment and prediction-error *auxiliary losses* still update the encoder/predictor when their reward weight is zero. We reran the ablation with **true component gating** (the auxiliary loss *and* its parameters dropped, `gated_components` recorded; fresh-reset eval; P10a, `p10a_gated_results.md` / `p10a_gated_aggregate.json`; H1_dm/G1 n=20, Go2 n=15/14 after a few seeds NaN'd). The true-gating result *splits* the original "intrinsic motivation contributes nothing" headline:

- **Empowerment: no detectable effect, not a certified null.** Full vs no_emp is inconclusive on all three bodies (G1 $g = -0.08$, Go2 $g = -0.00$, H1_dm $g = +0.65$ ns) under both ablations, and TOST certifies equivalence on no cell (all $p_{\text{TOST}} \geq 0.21$): the contribution of our one-step empowerment surrogate is *undetected*, not null.
- **Prediction-error is NOT null — it *suppresses* motion.** *Removing* the prediction-error component *increases* fresh $|q_{\text{vel}}|$: H1_dm $26.0 \to 53.3$ ($g = -2.96$, Holm $p < 10^{-4}$), Go2 $19.0 \to 21.8$ ($g = -1.16$, Holm $p = 0.04$); G1 is ns. So prediction-error has a real, large, *motion-suppressing* effect on 2/3 bodies (original trainer). Weight ablation ($g = -0.64$, -0.43, n.s. after Holm, other eval regime) leaves the pe loss training the encoder and, via the undetached action, the actor; gating removes both, so the carrying path is unidentified (Table 10; matched-regime contrast in Appendix K.1).

This is the **ablation-method dissociation (C2)**: the shortcut ablation (reward weight) and the true ablation (component) yield opposite conclusions about prediction-error. The methodological lesson is to ablate *components*, not just reward weights, before declaring an intrinsic signal inert.

C2 is not specific to ASC. The mechanism — an intrinsic module's auxiliary loss reshaping the *shared* encoder even when its reward weight is zero, so that reward-weight ablation misreads the component — is a property of intrinsic-motivation methods with a shared representation, not of our trainer. We reproduce it on shared-encoder variants of two public methods, ICM (Pathak et al., 2017) and RND (Burda et al., 2019), routed through a shared encoder on a public task (P20): holding the reward coefficient at zero and toggling only whether the auxiliary loss trains the encoder, ICM shows a robust frozen-probe gap and RND a clean *dose-response* on high-dimensional MiniGrid (Chevalier-Boisvert et al., 2023), while the effect *vanishes* on low-dimensional MountainCar where a near-random encoder already saturates the probe (magnitudes and dose curve in Appendix K). So the dissociation is real and method-general *wherever the shared encoder carries task information* — exactly ASC's regime — rather than universal. (Honest prereg deviation: the MiniGrid $A \neq B$ gap is a probe *degradation*, opposite in sign to our preregistered directional guess; the *dissociation*, not its sign, is what C2 asserts, and that holds — see Appendix K.)

Robustness of the intrinsic results. We report prediction-error's suppression as a direct effect and explicitly *reject* our own "predictability attractor" explanation, which dies once the residual is normalized by motion scale (Appendix I; exploratory). Two $50\times$-compute controls hold: a scaled-intrinsic run detects no "train-longer" rescue of the empowerment null (Appendix D; n=5, wide CI), and a $50\times$-steps no-curriculum baseline does not catch up — on Go2/H1_dm it regresses *below* its $1\times$ counterpart into near-stillness (g=+7.43/+1.24, Holm $p \leq 5 \cdot 10^{-5}$; Appendix D), which under (d) reads as the most compute-intensive check measuring the same gain artefact.

4 H3 REJECTED: NO DIAGNOSTIC PREDICTS OUT-OF-BODY COMPETENCE

We preregistered four training-time representation-space diagnostics and a leave-one-body-out (LOBO) ridge protocol testing whether any predicts *downstream Tier-1 competence* out-of-body, with the locked rule: max LOBO $R^2 \geq 0.5 \Rightarrow$ H3 supported; 0.3–0.5 partial; < 0.3 rejected.

H3 against the preregistered downstream target. All four diagnostics D1–D4 are tested with the LOBO target set to **downstream Tier-1 competence** (the §3.1 body-matched composite), not the $|q_{\text{vel}}|$ proxy (P10c; 135 checkpoints, 4 bodies). Three notes: (i) **D2** is checkpoint-constant within a body, so it has no within-body power and its LOBO score measures only whether the held-out body's *level* is predictable; (ii) a tag-parsing bug that dropped every H1_dm row was fixed (Appendix K) and all 4 bodies are included; (iii) D2 has no Panda fold (no holdout bank).

LOBO ridge result. Against the preregistered downstream Tier-1 target, all four training-time diagnostics fail to predict held-out-body competence: with leakage-free train-fold standardization the best leave-one-body-out R^2 is -3.5 (D2, 3 folds) and the others run -7.8 to -17.2 (Table 9). Most of that error (60–80% of SS_{res}) is the held-out body's level, but the rejection does not rest on it: given each held-out body's true level, within-body R^2 is at most $+0.003$. Against the $|q_{\text{vel}}|$ proxy the best is -0.19 (Appendix H: every fold, the decomposition, and an earlier reporting error).

Why pooled Spearman is misleading: a Simpson's paradox. The naive pooled Spearman rank correlation across all (body $\times$ cfg $\times$ seed) checkpoints *looks* moderately strong (D3 $\rho = -0.547$, D4 $+0.522$, both pooled $p < 10^{-10}$; Table 11 in Appendix K). But per-body Spearman correlations (within-body, $n \geq 29$ per body) tell a different story: of the 11 (diagnostic $\times$ body) cells with a defined within-body ρ, only **2 reach** $p < 0.05$ (G1 / D4, $\rho = +0.231$; H1_dm / D1, $\rho = -0.253$), and the latter has the *opposite* sign to its pooled correlation ($+0.258$). Our LOBO regression script auto-detects this pattern (the "Simpson warning" fires when pooled $|\rho| > 0.3$ but every within-body $|\rho| < 0.3$) on both D3 and D4 (after a detector fix, Appendix K). The same paradox holds against the **downstream Tier-1 target** (P10c): D3's pooled $\rho = +0.33$ likewise dissolves under within-body decomposition (the Simpson warning fires), so the pooled-vs-LOBO gap is a property of the diagnostics across *both* targets, not an artefact of the proxy.

The pooled correlations are dominated by body identity, not within-body signal (Go2 is low on both diagnostic and qvel, H1_dm high on both; Figure 8, Appendix K). **In multi-body ML, pooled rank correlation does not validate a predictive metric**; LOBO ridge with train-fold standardization is the minimum bar for a cross-body diagnostic claim. Proposition 1 (Appendix H) makes this precise: two processes can share the same unlabelled pooled distribution — hence every pooled correlation, however bootstrapped — while their LOBO R^2 differ in sign; only holding a body out tells them apart. Our result is about these four diagnostics on these four bodies, not all training-time diagnostics.

This sharpens the reproducibility literature's "single-metric reads can mislead" caveat (Henderson et al., 2018; Agarwal et al., 2021; Spoor et al., 2026): reads *pooled across bodies* mislead while looking more rigorous than they are.

Three distinct ways a pooled number misleads, and where each reproduces. The Simpson's paradox above is one of three phenomena we group under C3, and they should not be conflated. *(a) Ecological reversal of a correlation*: a pooled association that the within-stratum associations do not carry — the diagnostics above, Proposition 1, and a 2000-trial simulation in which pooling lends a correlation a sign no stratum exhibits (Appendix K). *(b) Attenuation of a within-stratum comparison*: when strata differ in level, a pooled probability-of-improvement is pulled toward $\frac{1}{2}$. This reproduces on *released, third-party* data: on the openrlbenchmark Atari comparison of CleanRL PPO (Huang et al., 2022) vs OpenAI-baselines PPO2 (26 games as strata, human-normalized with rliable's constants), CleanRL leads in $18/26$ games with a stratified $P(\text{improvement}) = 0.67$ against a naive pooled 0.53 — 85% **of the within-task signal erased**. *(c) Dependence of "the winner" on what the aggregate is for*: a mean over tasks and a mean rank answer different questions (average utility over *this* task mixture vs consistency across domains), and neither is wrong; what is wrong is reading either as "best". Bootstrapping URLB's published appendix numbers (Laskin et al., 2021) (8 methods, 12 tasks, mean$\pm$SEM over 10 seeds) under their own seed noise, **the table cannot**

statistically crown a best URL method: the raw-mean winner is a coin flip ($P(\text{RND}) = 0.54$), the two estimators disagree in about a third of draws, no method tops all three domains in 94–95%, and LODO is undecidable at $n = 8$ (Appendix K). Reading a table is not running a pipeline, so we ran theirs (27 pretrainings and 144 finetunings of the official URLB code, extended to 5 seeds). In *both* the published panel and our runs, at either seed budget, no method is top-1 on all three domains — $P(\text{sweep}) = 0.0006$ at n=3 and 0.0409 at n=5 (bootstrap frequencies, descriptive; not a test). Domain, not method, decides the ranking here. Estimator-dependence is weaker, and a seed extension *withdrew* it: the two poolings disagreed at n=3 but agree at n=5, both crowning RND at the point estimate. How confidently depends on the estimand: our preferred one, a checkpoint-cluster bootstrap counting a tie as no winner, gives RND a sole-winner frequency of 0.84 on rank and 0.999 on raw mean (the locked grader's 0.946 resampled tasks independently; Appendix K); we therefore claim it from the published tables and our n=3 runs, not from n=5 (P28/P28b). None of the three is downstream of our instrument: (a) reproduces in simulation, (b) on third-party data, (c) inside the field's own pipeline run by us end-to-end. Pooling is not wrong; it answers the average-over-these-bodies question, and only that one.

5 Discussion

5.1 The measurement guide: what you measure, ablate, and pool decides the conclusion

The through-line is methodological: three default choices each flip the ASC conclusion, and we distill the audit into a five-row checklist a multi-body embodied-RL study can apply directly (Table 12, Appendix K): (0) check that the effect survives the instrument — re-evaluate every arm at matched gain and reset; (i) do not read a training-time proxy as competence — evaluate a *downstream task*; (ii) do not ablate by zeroing a *reward weight* — use *true component gating*; (iii) do not validate a diagnostic on *pooled* cross-body data — decompose *within-body / LOBO*; (iv) do not report one body/seed/metric — run a Holm-locked multi-body grid with CIs and TOST.

The meta-lesson. Our headline effect passed preregistration, four morphologies, Holm correction *and* a 50× compute control, and was still an artefact only a gain-matched re-measurement caught (§3(d,e)): rigour operates *within* a measurement and cannot validate the instrument.

6 Conclusion

Our central finding is methodological: **what you measure, ablate, and pool decides the conclusion** about ASC. *Check the instrument before the effect*: the curriculum's large, preregistered, Holm-corrected proxy effect is largely an artefact of its own action-gain schedule (null on 0/4 bodies corrected, §3(d,e)) and transfers on none of four routes while a dynamics control does (§3.1). *Ablate components, not reward weights*: true gating shows prediction-error *suppresses* motion ($g = -2.96$, H1_dm) where the reward-weight shortcut stayed non-significant (-0.64) (§3.2). *Decompose pooled claims*: a diagnostic that looks predictive pooled ($|\rho| \approx 0.5$) fails leave-one-body-out ($R^2 \leq 0.0$) on both the proxy and the downstream Tier-1 target — a body-identity Simpson's paradox (§4).

Beneath all three, the larger-effect bodies' proxy gain rides a body-specific monotone unlock that order perturbations degrade (§A.2): the artefact's structure, not evidence for development. The rules are the contribution; the ASC dissociations are the evidence.

Limitations

Scope. 4 bodies, one trainer family; no intrinsic-motivation source overtakes random init at 10×, and 100× is open (P27). **One preregistered prediction failed:** estimator-dependence (P2) held at n=3 but not n=5; that claim now rests on the published tables and n=3 (§4). **The external-baseline rerun is a consistency check**, not a superiority claim. **Threats to each claim.** C1's gain attribution rests on P16 and P22 (P19 bundles six fixes) and on H1_dm depends on roll-in length; C2's gating removes the encoder path, actor path and predictor at once, and is strong on H1_dm only. All fourteen caveats: Appendix K.1.

REPRODUCIBILITY

All training scripts, ablation runners, the statistical-tests library (`src/phase2/stats_tests_v2.py`: Welch + Mann–Whitney + Hedges g + 20K bootstrap + Holm + BH + TOST + power calculator), every preregistration and protocol amendment, and the per-seed raw JSON for all experiments are released as the anonymized supplementary material accompanying this submission. There are two reproduction paths. Every statistical claim is recomputable on a CPU from the released per-seed summaries; `scripts/verify_bundle_stats.py` re-runs every such analysis from the extracted bundle, compares each output with the shipped JSON field by field, and exits non-zero on any difference. Model checkpoints are omitted except the ten H1_dm checkpoints behind the C1 intervention, which `scripts/reproduce_c1_minimal.sh` re-measures on one GPU from inside the extracted bundle (checked: $99\times \rightarrow 22\times$ against the released $101\times \rightarrow 26\times$). The bundle's README lists what is and is not included and marks every result as preregistered, extension or post hoc. The preregistration is git-locked at SHA `d67f51d`; each follow-up (P5 scaled baseline, P6 statistics, the shuffled_tail disambiguator, P7 scaled-intrinsic) was locked before its run, and the lock commits precede the corresponding result commits in the released history. Total compute: 4 $\times$ A100 80 GB. Summing the recorded wall time over the 1,492 timed summaries among 1,767 released gives **32.5 GPU-days**; the largest contributors are the downstream transfer grids (P11 9.8, P15 6.2), the primary P2 grid (2.9) and P21 (2.1). The URLB rerun is *not* in that total — the upstream pipeline records no per-run time — and occupied two GPUs for about 44 h. The downstream measurements are released with the same discipline: the replay-buffer SAC (Haarnoja et al., 2018) fine-tuner (`m8_sac_transfer_rb.py`) and its locomotion variant with reset-on-fall (`m8b_sac_locomotion_reset.py`), the frozen-representation probe (`m9_repr_probe.py`), the self-supervised dynamics positive control (`m10_dyn_pretrain.py`), and the rliable-style aggregator (`rliable_analysis.py`), each with a lock-before-run protocol committed ahead of its results. One exception we state explicitly: the prediction-error mechanism analysis (`m11_pe_mechanism.py`) and its results were committed together with no preceding protocol, so those tests are exploratory (`pre_registered: false` in the released JSON) and are reported as such. The external-baseline study (P15) follows the same discipline: the ICM/RND pretrainer built on the identical backbone (`m12_url_pretrain.py`), the fresh-reset proxy re-measurement (`m13_fresh_eval_backfill.py`, which writes non-destructive sidecars beside the original summaries), and the preregistered rank analysis (`p15_rank_analysis.py`), with the protocol — including the method set, the hyperparameter budget, the validation/confirmatory seed split, and both success criteria — locked in `p15_url_baseline_protocol.md` and committed before any P15 run produced a result. Every headline number in Sections 3.1–5.1 recomputes from the released per-seed JSON via the named aggregators.

AI USE STATEMENT

In this work, we used LLM-based agents (as research assistants under author direction) for research execution and paper writing: implementing and executing the experiment grids, drafting preregistered protocols and analysis code, retrieving and verifying citations against primary sources, proposing and refining candidate hypotheses and their preregistered predictions, drafting the proof of Proposition 1, drafting interpretations of results, and drafting and revising the manuscript text. We did *not* use generative AI to create synthetic datasets or to generate any experimental measurement: every number reported here comes from the released simulation runs. The authors set the research questions and audit framing, chose which hypotheses to test, reviewed and approved every preregistered protocol before it was executed, verified all reported numbers against the released artifacts, accepted or rejected each proposed interpretation, and take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work is a methodological audit of simulated sensorimotor pretraining (MuJoCo MJX); it involves no human subjects, no personal data, and no deployment. We see no direct ethical risks.

Indirectly, the paper argues *against* over-claiming from training-time proxies, which we believe supports more honest reporting in embodied-AI research.

REFERENCES

Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron Courville, and Marc G. Bellemare. Deep reinforcement learning at the edge of the statistical precipice. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Nadine Badie, Firas Al-Hafez, Pierre Schumacher, Daniel F. B. Haeufle, Jan Peters, and Syn Schmitt. Bioinspired morphology and task curricula for learning locomotion in bipedal muscle-actuated systems. *Communications Engineering*, 2025. doi: 10.1038/s44172-025-00443-0.

Andrew M. Bean, Ryan Othniel Kearns, Angelika Romanou, Franziska Sofia Hafner, Harry Mayne, et al. Measuring what matters: Construct validity in large language model benchmarks. In *Advances in Neural Information Processing Systems 38 (NeurIPS 2025) Datasets and Benchmarks Track*, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/1967e0fc3aa6cbbace562f5cb8e3954e-Paper-Datasets_and_Benchmarks_Track.pdf.

Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and R. Devon Hjelm. Mutual information neural estimation. In *International Conference on Machine Learning (ICML)*, 2018.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *International Conference on Machine Learning (ICML)*, 2009.

Nikolai A. Bernstein. *The Co-ordination and Regulation of Movements*. Pergamon Press, Oxford, 1967.

Yuri Burda, Harrison Edwards, Amos Storkey, and Oleg Klimov. Exploration by random network distillation. In *International Conference on Learning Representations (ICLR)*, 2019. arXiv:1810.12894.

Angelo Cangelosi and Matthew Schlesinger. *Developmental Robotics: From Babies to Robots*. MIT Press, 2015.

Maxime Chevalier-Boisvert, Bolun Dai, Mark Towers, Rodrigo de Lazcano, Lucas Willems, Salem Lahlou, Suman Pal, Pablo Samuel Castro, and Jordan Terry. Minigrid & miniworld: Modular & customizable reinforcement learning environments for goal-oriented tasks, 2023.

Bradley Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1): 1–26, 1979.

Logan Engstrom, Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Firdaus Janoos, Larry Rudolph, and Aleksander Madry. Implementation matters in deep RL: A case study on PPO and TRPO. In *International Conference on Learning Representations (ICLR)*, 2020.

Andrew Gelman and Jennifer Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, 2006. doi: 10.1017/CBO9780511790942.

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *International Conference on Learning Representations*, 2021.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International Conference on Machine Learning (ICML)*, pp. 1861–1870, 2018.

Larry V. Hedges. Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, 6(2):107–128, 1981.

Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. Deep reinforcement learning that matters. In *AAAI Conference on Artificial Intelligence*, 2018.

Shengyi Huang, Rousslan Fernand Julien Dossa, Chang Ye, Jeff Braga, Dipam Chakraborty, Kinal Mehta, and João G. M. Araújo. CleanRL: High-quality single-file implementations of deep reinforcement learning algorithms. *Journal of Machine Learning Research*, 23(274):1–18, 2022.

Alexander S. Klyubin, Daniel Polani, and Chrystopher L. Nehaniv. Empowerment: A universal agent-centric measure of control. In *IEEE Congress on Evolutionary Computation*, pp. 128–135, 2005.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton Earnshaw, Imran Haque, Sara M. Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. WILDS: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning (ICML)*, pp. 5637–5664, 2021.

Daniël Lakens. Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, 8(4):355–362, 2017.

Michael Laskin, Denis Yarats, Hao Liu, Kimin Lee, Albert Zhan, Kevin Lu, Catherine Cang, Lerrel Pinto, and Pieter Abbeel. URLB: Unsupervised reinforcement learning benchmark. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2021.

Michael Laskin, Hao Liu, Xue Bin Peng, Denis Yarats, Aravind Rajeswaran, and Pieter Abbeel. Unsupervised reinforcement learning with contrastive intrinsic control. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022a.

Michael Laskin, Hao Liu, Xue Bin Peng, Denis Yarats, Aravind Rajeswaran, and Pieter Abbeel. CIC: Contrastive intrinsic control for unsupervised skill discovery, 2022b.

Timothy P. Lillicrap, Jonathan J. Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. In *International Conference on Learning Representations (ICLR)*, 2016.

Shakir Mohamed and Danilo Jimenez Rezende. Variational information maximisation for intrinsically motivated reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2015.

Deepak Pathak, Pulkit Agrawal, Alexei A. Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *International Conference on Machine Learning (ICML)*, 2017.

Deepak Pathak, Dhiraj Gandhi, and Abhinav Gupta. Self-supervised exploration via disagreement. In *International Conference on Machine Learning (ICML)*, pp. 5062–5071, 2019.

W. S. Robinson. Ecological correlations and the behavior of individuals. *American Sociological Review*, 15(3):351–357, 1950. doi: 10.2307/2087176.

Ramanan Sekar, Oleh Rybkin, Kostas Daniilidis, Pieter Abbeel, Danijar Hafner, and Deepak Pathak. Planning to explore via self-supervised world models. In *International Conference on Machine Learning (ICML)*, 2020.

E. H. Simpson. The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 13(2):238–241, 1951. doi: 10.1111/j.2517-6161.1951.tb00088.x.

Lindsay Spoor, Aske Plaat, and Thomas Moerland. Evaluation metrics for safe reinforcement learning. *arXiv preprint arXiv:2609.15315*, 2026. URL `https://arxiv.org/abs/2609.15315`. v1.

Richard S. Sutton. The bitter lesson. `http://www.incompleteideas.net/IncIdeas/BitterLesson.html`, 2019.

Emanuel Todorov, Tom Erez, and Yuval Tassa. MuJoCo: A physics engine for model-based control. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 5026–5033, 2012.

Beatrix Vereijken, Richard E. A. van Emmerik, H. T. A. Whiting, and Karl M. Newell. Free(z)ing degrees of freedom in skill acquisition. *Journal of Motor Behavior*, 24(1):133–142, 1992. doi: 10.1080/00222895.1992.9941608.

G. Udny Yule. Notes on the theory of association of attributes in statistics. *Biometrika*, 2(2):121–134, 1903. doi: 10.1093/biomet/2.2.121.

A CURRICULUM MECHANISM: SCHEDULE PERTURBATIONS IN DETAIL

This appendix reports the three schedule perturbations summarised in §3 (Table 5) in full.

A.1 H2A — EARLY-STEP CURRICULUM: BODY-SPECIFIC RECOVERY

H2a tests whether the curriculum benefit is carried by the *early* training steps (ROM ramps to 1 over the first N=200 of 1000 steps, then holds ROM=1). For each of the 3 fully-completed bodies we run 12 seeds of the `early_only` schedule and compare to the body's Full and −curric cells (each n=20) via Hedges' g.

We report two retention metrics. The preregistered *literal* formula was `ratio = g(Full, early_only) / g(Full, -curric)` $\geq$ 0.70 — but as recorded in `experiments/phase2-iclr/p3_h2_issues.md` Issue #1, this formula is inverted relative to the natural-language hypothesis (high $g(F, \text{early})$ means early_only *fails* to recover Full). The text-faithful interpretation is `retention = 1 - g(F, early)/g(F, -curric)` ≥ 0.70. We report both:

Body	g(F, early)	g(F, −curric)	literal ratio	text-faithful retention	text PASS?
G1 (29-DoF humanoid)	−0.07	+1.00	−0.07	**+1.07**	✓
Go2 (12-DoF quadruped)	+1.40	+4.39	+0.32	+0.68	× (borderline)
H1_dm (21-DoF humanoid)	+1.12	+1.61	+0.70	**+0.30**	×

Table 2: H2a retention metrics by body.

The text-faithful result is the substantive H2a verdict: **curriculum-effect retention by early-only is body-specific** — G1 fully recovers the curriculum benefit from the first 200 steps alone; Go2 retains 68% (just under threshold); H1_dm retains only 30% (Figure 2; per-body values in Table 2).

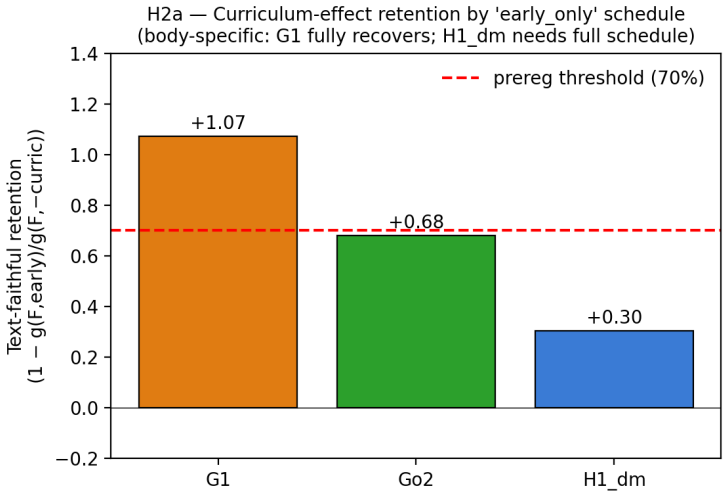

Figure 2: H2a — text-faithful retention by body. G1 fully recovers Full; Go2 just under the 70% threshold; H1_dm clearly fails. Body-specific curriculum mechanism.

A first-pass interpretation: the bodies whose curriculum benefit is *smaller* (G1 g=1.0) are also the ones whose effect lives entirely in early-training. Bodies with *larger* curriculum benefits (Go2 g=4.4, H1_dm g=1.6) need the full schedule. Cleanly stated: **the size of the late-schedule contribution scales with the size of the total curriculum benefit, not inversely.** A multi-body finding inaccessible to single-body studies; we discuss possible causes (state-space coverage during full ROM, exploration-saturation, etc.) in §5.

A.2 H2B — SCHEDULE ORDER MATTERS

We test whether the curriculum's information content lies in its *cumulative* ROM-time exposure or in its *temporal ordering* (Table 3). The preregistered original H2b ran `rom_schedule=reverse` (ROM= $1 - s$), but that schedule ends training at ROM= 0 — the actor learns to output zero actions in its final, most-fit phase, and eval at ROM= 1 reads the now-unmasked near-zero outputs as low qvel by *training artefact*, not by schedule mechanism. We documented this implementation confound (`p3_h2_issues.md` Issue #2) and ran a corrected test with a locked amendment (`p3_h2b_correction.md`, committed before any vshape run).

Corrected schedule (vshape). $\text{ROM}_t(s) = |1-2s|$ — starts at 1, dips to 0 at the midpoint, returns to 1 at the end. Two properties: (i) $\int_0^1 \text{ROM}_t(s)\,ds = 0.5$, exactly matching Full's $\int_0^1 s\,ds = 0.5$ — same cumulative ROM-time; (ii) ROM= 1 at end-of-training, so eval inherits no zero-action artefact. vshape vs Full therefore isolates *temporal ordering* with *cumulative ROM-time controlled*.

Locked decision rule (per body, n=12 vshape vs n=20 Full, Holm m=3): Holm $p > 0.05$ AND $|g| < 0.30 \Rightarrow$ order doesn't matter; Holm $p < 0.05$ AND $|g| > 0.50 \Rightarrow$ order matters; else inconclusive.

Body	g(vshape, Full)	g 95% CI	p_{Holm}	Verdict
Go2 (quadruped)	−2.52	[−7.71, −1.32]	**0.0002**	**order DOES matter**
H1_dm (humanoid)	−1.34	[−1.85, −1.05]	**0.0002**	**order DOES matter**
G1 (humanoid)	−0.54	[−0.95, −0.06]	0.0748	inconclusive (marginal)

Table 3: H2b' (vshape vs Full) results by body.

Result — H2b' MIXED, 2/3 bodies "order matters": The two bodies with the largest curriculum effects (Go2 P2 g=4.4, H1_dm P2 g=1.6) show vshape strongly degrading the curriculum benefit — the effect cannot be reproduced by re-arranging the same cumulative ROM-time. G1 (small overall curriculum effect g=1.0) sits right at the decision boundary, with the same sign but missing Holm significance (p=0.075, just above the per-test threshold).

Disambiguating ordering from the midpoint dead zone (shuffled_tail). vshape has a confound of its own: at s=0.5 the actor passes through ROM=0 — a mid-training dead zone — so its deficit could reflect *discontinuity* rather than *ordering*. A second preregistered perturbation (`p3_h2b_future_shuffled.md`, locked before any run) removes the dead zone: `shuffled_tail` randomly permutes Full's per-step ROM values over the first 90% of training (per-seed RNG; Table 4) and forces ROM=1 over the final 10% — same dose, scrambled order, no ROM=0 dwell. Same locked rule (n=12 vs Full n=20, Holm m=3):

The refined mechanism claim is therefore graded. On Go2 — the body with the largest curriculum effect — temporal ordering matters per se: two independent perturbations with identical cumulative dose (deterministic reversal, random scramble) both degrade the benefit under Holm correction, and vshape's larger deficit (−2.52 vs. −1.30) is consistent with the dead zone *adding* to the ordering cost. On H1_dm the pure scramble is directionally negative but not Holm-significant, so part of vshape's large deficit there is attributable to the midpoint discontinuity; we claim ordering on H1_dm only as partial evidence. On G1 neither perturbation matters, consistent with its curriculum effect living in the early phase (§A.1). **Where the curriculum effect is large, its benefit depends on the temporal structure of the unlock — not merely on cumulative ROM-time exposure.**

Body	g(shufftail, Full)	g 95% CI	p_{Holm}	Verdict	Joint reading with vshape
Go2	-1.30	$[-3.47, -0.40]$	**0.032**	order DOES matter	both perturbations significant → **ordering confirmed**
H1_dm	-0.70	$[-1.45, -0.03]$	0.098	inconclusive	scramble hurts less than vshape (-0.70 vs -1.34) → part of vshape's deficit is the dead zone; ordering evidence partial
G1	-0.22	$[-0.63, +0.57]$	0.450	order DOESN'T matter	consistent with §A.1 early-phase sufficiency

Table 4: shuffled_tail (dead-zone-free scramble vs Full) results by body, with joint reading.

B CURRICULUM MECHANISM: WHAT WE MOVED OUT OF THE MAIN TEXT

Read this section with C1 in hand. The analysis below asks which property of the ROM schedule carries the proxy effect, and was run before we established (§3(d)) that most of that effect is generated by the schedule's action gain rather than by acquired motion. We keep it because the perturbation results are real measurements of the trained policies and because the exercise is instructive — a coherent, body-graded mechanism story fitted comfortably around an artefact — but its conclusions describe the artefact's structure, not a learned competence, and should not be read as a mechanism of sensorimotor development.

Three preregistered perturbations ask *which property* of the ROM schedule carries its effect: `early_only` (ramp to full ROM over the first 200 of 1000 steps, then hold) removes the schedule's late phase; `vshape` (ROM$_t = |1 - 2s|$) inverts its *ordering* at matched cumulative ROM-dose; and `shuffled_tail` permutes the first 90% of the trajectory while forcing ROM=1 at eval, removing monotone continuity without creating a dead zone. Appendix A reports each in full (12 seeds per perturbation cell against the body's n=20 Full and −curric cells, Hedges g with Holm correction). The result is graded rather than uniform: on the small-effect body (G1, g=1.0) the early phase alone recovers the whole benefit and neither ordering perturbation is significant, whereas on both large-effect bodies (Go2 g=4.4, H1_dm g=1.6) early-only fails to recover and reordering at fixed dose significantly degrades the benefit ($g = -2.5$ and $g = -1.3$). On the bodies where the curriculum matters most, what it contributes is therefore not the *amount* of ROM exposure but its *gradual, monotone ordering* (C4; §B.1).

Table 5 collects the three mechanism perturbations on one line per body, so the graded claim is legible at a glance: each column removes one property of the Full schedule (its early phase, its ordering at fixed dose, or — for shuffled_tail — its monotone continuity without a dead zone), and a × marks where that removal significantly degrades the curriculum benefit.

Body (P2 curric g)	early-only	vshape	shuffled_tail	reading
G1 (g=1.0, small)	✓ recovers	– ns	– ns	early phase suffices; structure dispensable
Go2 (g=4.4, large)	× partial	× -2.5	× -1.3	ordering matters (both perturbations agree)
H1_dm (g=1.6, large)	× fails	× -1.3	∼ -0.7 ns	ordering + continuity; partial

Table 5: Mechanism summary. Columns remove one property of the Full schedule: *early-only* its late phase (H2a); *vshape* its ordering at matched cumulative ROM-dose (H2b′); *shuffled_tail* its monotone continuity without creating a dead zone. Entries under the perturbation columns are Hedges g versus Full. ✓ = curriculum benefit preserved; × = significantly degraded (Holm-corrected); ∼/– = directional/ns. The large-effect bodies depend on the schedule's temporal structure; the small-effect body does not.

B.1 THE CURRICULUM'S MECHANISM IS GRADUAL MONOTONE UNLOCK

Two mechanism results (Sections A.1 and A.2) point to the same interpretation. H2a finds early-only sufficient on the smallest-effect body (G1, retention $+1.07$) but losing 32% (Go2) to 70% (H1_dm) of the curriculum effect on bodies with larger overall gaps. H2b' (the corrected vshape test that controls cumulative ROM-time) finds the same two bodies showing strong "order matters" verdicts ($g = -2.52$ Go2, $g = -1.34$ H1_dm, both Holm p=0.0002) while G1 sits at the decision boundary ($g = -0.54$, p_{Holm}=0.075).

The shuffled_tail disambiguation (§A.2) sharpens this picture per body. On Go2, *both* dose-controlled perturbations — deterministic reversal and random scramble, which share no structure except violating monotonicity — degrade the benefit under Holm correction: the ordering claim there is confirmed, and vshape's larger deficit suggests the midpoint dead zone adds a separate cost on top. On H1_dm, the dead-zone-free scramble is directionally negative but not significant, so vshape's large deficit is partly a *continuity* effect: training through a ROM=0 dwell costs H1_dm more than order scrambling alone does. On G1 neither perturbation matters.

The joint interpretation: **the curriculum benefit on bodies with a large curriculum effect is carried by the *gradual monotone unlock* structure — not by early-step exposure alone (H2a refutes that), not by cumulative ROM-time alone (H2b' refutes that), and on Go2 specifically not by mere schedule continuity either (shuffled_tail refutes that)**. On bodies with a small curriculum effect (G1), simpler explanations (early initialization, cumulative ROM-time) are sufficient — there is less "effect" to carry, so the schedule's structural details matter less.

A natural follow-up hypothesis: the late-stage monotone unlock allows the actor to incrementally extend its support over states that only become reachable at higher ROM, while a non-monotone schedule forces the actor to spend its later training in a regime its earlier training prepared it poorly for. Measuring per-step state-occupancy distributions under each schedule would test this directly; we leave it to follow-up.

This pattern resembles Bernstein's freezing/freeing of degrees of freedom (Bernstein, 1967; Vereijken et al., 1992): the learner first masters a low-dimensional, range-restricted body and only then releases the remaining degrees of freedom. We no longer read it as evidence for that account. A scalar action-gain schedule is not a biomechanical freezing of degrees of freedom, and C1 shows the effect whose temporal structure is dissected here is largely produced by that schedule's interaction with the evaluation gain. What this section establishes is the *structure of the artefact*: its value is not universal but scales with the body. On the bodies where the curriculum effect is large (Go2, H1_dm) the freeze-then-free ordering is doing real work — re-arranging the same cumulative ROM exposure (vshape) or scrambling its order (shuffled_tail) destroys much of the benefit — whereas on the small-effect body (G1) an early initialization suffices and the ordering is dispensable. A testable prediction follows for both robotics and motor-development research: the benefit of a freezing/freeing curriculum should grow with a morphology's effective instability or contact complexity, an axis our four-body slice only begins to sample (we return to this in the scaling discussion of the limitations).

The original H2b (`reverse` schedule) was invalidated by an implementation confound — it ended training at ROM$= 0$, so the actor learned a near-no-op policy that gets unmasked at eval. The corrected vshape test (Section A.2) controls this by ending at ROM$= 1$ and matching Full's cumulative ROM-time.

C DOWNSTREAM TRANSFER MEASUREMENTS IN DETAIL

Confirmation 2 — the advantage survives no RL fine-tuning either. We fine-tuned each pretrained encoder on a *learnable* downstream task (Panda end-effector reach) with a stabilized replay-buffer SAC, under both a frozen-encoder protocol (representation quality) and an end-to-end fine-tuning protocol (warm-start quality); Full / −curric / RANDOM, n=10 seeds each (P11, `p11_transfer_results.md`). The curriculum improves transfer over −curric on **0/2** encoder modes (Holm); Full is never the best source (g(Full − −curric) $= +0.02$ frozen / -0.08 finetune, both p_{Holm}=1.0 — no detected benefit, though not certified equivalent: a Welch-TOST at ± 0.5 SD gives $p_{\text{TOST}} = 0.27$ on the frozen cell, so we report "no detected benefit", not "equivalent"; frozen

RANDOM beats Full at $g{=}-1.12$). The ROM curriculum does not give a better fine-tuning starting point.

Confirmation 3 — nor a better frozen representation. A frozen-encoder linear/MLP probe predicting task-relevant quantities (root height, forward velocity, end-effector-to-target distance) from the encoder latent, across the three locomotor bodies $\times$ {Full, $-$curric, RANDOM} $\times$ 8 seeds (P11-C, `p11_probe_results.md`), finds the curriculum improves probe R^2 on **0/3** bodies (Holm), under both heads and under a trajectory-separated holdout. Where the two heads differ is Full versus RANDOM: the MLP head puts RANDOM ahead on every body (still so with whole trajectories held out, $\Delta R^2 = -0.09$, CIs excluding 0), while the linear head puts Full marginally ahead ($+0.01$ to $+0.04$). The proxy advantage is not a downstream-task-relevant representation signal; whether ASC is *worse* than an untrained encoder depends on the probe's capacity.

Confirmation 4 — nor on locomotion, on the body where the proxy effect is largest. The obvious rejoinder to Confirmation 2 is that Panda reach is manipulation, not the curriculum's home turf. So we made *locomotion* learnable — adding per-env reset-on-fall and an alive bonus to the fine-tuner (P14, `p14_locomotion_results.md`), under which a random-init SAC learns a genuine forward gait on H1_dm (0.51 m/s). H1_dm is the body with the **largest** proxy curriculum effect (up to $195\times$). If the proxy were competence it should transfer *here* most of all — yet the curriculum improves forward-velocity transfer over $-$curric on **0/2** modes: under a frozen encoder it is the *worst* source (Full 0.35 vs RANDOM 0.55, $g{=}-1.94$), and under fine-tuning all sources converge to the same gait (0.52–0.55, g(Full $-$ $-$curric)$=-0.07$, $p{=}0.91$). (Go2 did not clear the pre-registered learnability gate and is excluded; the deterministic eval understates steady-state speed equally across sources.)

D SCALED-COMPUTE BASELINE (P5) IN DETAIL

P7 — the intrinsic null is stable from $1\times$ to $50\times$ compute. Empowerment might only begin to pay off with far more compute, once the curriculum has done its work. A preregistered scaled-intrinsic experiment (P7, `p7_scaled_intrinsic.md`, locked before any run) tests this on Go2 — the largest-curriculum body — with the curriculum ON and $n_{\text{steps}} = 50{,}000$ ($50\times$): removing empowerment changes nothing measurable (scaled_full 7.94 ± 2.30 vs. scaled_noemp 8.26 ± 3.16, $g = -0.10$, $[-1.23, +1.21]$, Welch $p = 0.86$, $n{=}5$), nowhere near the preregistered "contributes at scale" bar ($|g| > 0.8$); $50\times$ compute does not resurrect an empowerment effect (per-body values in Table 6). At $n{=}5$ the equivalence test does not certify a null (TOST at ± 0.5 SD, $p = 0.28$), so we report "no detectable contribution at scale," not certified equivalence; and since the interval spans ± 1.2, P7 does not *exclude* a "train-longer" rescue of the intrinsic arm either — it fails to detect one.

Before turning to the compute question we note what we did *not* keep in the main text. Three preregistered schedule perturbations (`early_only`, `vshape`, `shuffled_tail`) asked which property of the ROM schedule carries its proxy effect, and produced a coherent, body-graded answer: on the large-effect bodies the benefit depends on gradual monotone ordering, on the small-effect body it does not. Appendix A reports them in full. We have moved them out of the main narrative because §3(d) shows the effect they dissect is largely produced by the schedule's action gain — the mechanism story fitted comfortably around an artefact, which is itself the cautionary point of this paper.

The strongest objection to "the curriculum dominates the $|q_{\text{vel}}|$ proxy" is the Bitter-Lesson contender (Sutton, 2019): maybe the no-curriculum variant simply needs more compute to match the curriculum *on the proxy*. We test this directly with a preregistered scaled baseline (protocol locked at `p5_scaled_baseline.md` before any production run): `scaled_nc` = identical trainer with the ROM curriculum removed (ROM $= 1$ from $t{=}0$) and intrinsic motivation left ON at P2-Full operating points, trained for $50\times$ **the steps** ($n_{\text{steps}} = 50{,}000$ vs. 1,000), $n{=}5$ seeds on the two bodies with the largest H1 effects (Go2, H1_dm), compared against P2 Full ($n{=}20$) at standard compute, Holm $m{=}2$.

Both bodies meet the locked "curriculum scaffold persists at scale" criterion ($g > 0.8$, CI excludes 0, Holm $p < 0.05$). The scaled contender does not arrive. H1_dm's contrast, while clearly significant, is run at $n{=}5$, where the a-priori minimum detectable effect is $g \approx 2.02$ ($\alpha{=}0.05$, power 0.8, t-

Body	Full (1×)	scaled_nc (50×)	1× −curric	Hedges g	g 95% CI	Holm p	Verdict
Go2	7.21 ± 1.03	0.0004 ± 0.0003	0.896	**+7.43**	[+5.53, +12.08]	$< 10^{-5}$	scaffold persists
H1_dm	10.39 ± 8.91	0.0064 ± 0.0009	0.053	**+1.24**	[+1.00, +1.70]	$5 \cdot 10^{-5}$	scaffold persists

Table 6: P5 scaled no-curriculum baseline ($50\times$ steps) vs P2 Full.

based) — only very large effects are detectable at this n, so we treat the H1_dm cell as supporting, not load-bearing; Go2's observed effect far exceeds that MDE.

A stillness fixed point, not a slow climb (exploratory). Outside the locked rule, the $50\times$ no-curriculum runs land *below* even the $1\times$ no-curriculum baseline (Go2: 0.0004 vs. 0.896; H1_dm: 0.0064 vs. 0.053), with near-zero seed variance. The per-step training traces (n=5 each) show *why*, and the two bodies reach the same place by different routes: on **Go2** the policy first moves and then collapses — training $|q_{\text{vel}}|$ falls $0.50 \rightarrow 0.0004$ while prediction error falls $0.64 \rightarrow 0.02$, i.e. the optimizer buys lower prediction error by bringing the body to rest; on **H1_dm** the policy never leaves the low-motion basin ($|q_{\text{vel}}|$ $0.008 \rightarrow 0.007$, prediction error flat at ≈ 0.12). In both, the empowerment estimate stays at the noise floor throughout ($|\hat{\mathcal{E}}| \lesssim 5{\times}10^{-1}$ transient on Go2, $\sim 10^{-4}$ on H1_dm), so it supplies essentially no gradient — consistent with the §3 intrinsic null. The shared endpoint is a *stillness fixed point* of the unscaffolded intrinsic objective: without the ROM schedule to keep early motion in-distribution, more optimization either drives the body to rest (Go2) or fails to start it (H1_dm). Within this trainer family compute is thus not an inefficient substitute for the developmental scaffold but, at least on these two bodies, *counterproductive without it*. We flag this as an observation, not a proven universal: n=5, two bodies, one trainer.

E Discussion material moved from the main text

E.1 Cross-body diagnostics: pooled correlations as Simpson's paradoxes

The H3 rejection (Section 4) was the surprise of the project. We had expected at least one of three reasonable representation-space diagnostics — effective rank, linear-probe transfer, or prediction-error residuals — to support an out-of-body downstream prediction. The LOBO ridge R^2 values are all < 0 (best -0.19), while the pooled Spearman rank correlations look moderate-and-significant ($|\rho| \approx 0.5$, $p < 10^{-10}$ for D3 and D4).

Per-body decomposition reveals these pooled correlations as Simpson's-paradox artefacts: only 2 of the 11 (diagnostic × body) cells reach within-body $p < 0.05$ (G1 / D4, $\rho = +0.231$; H1_dm / D1, $\rho = -0.253$), and the latter has the *opposite* sign to its pooled correlation ($+0.258$). The bodies cluster — Go2 has uniformly low diagnostic values *and* low qvel; H1_dm has uniformly high in both — creating a between-body trend that masquerades as within-body signal.

The methodological implication is sharp: **for any predictive- diagnostic claim in multi-body RL/embodied-AI, within-body decomposition is required.** Pooled Spearman or pooled correlation in the presence of body-identity confounding is precisely the kind of metric that a single-body community has not been forced to question. Our LOBO ridge with train-fold standardization (statistics fit on the training bodies pooled, applied to the held-out body; not per-body z-scoring) is one minimum bar; we recommend at minimum a Simpson's-paradox audit on any pooled cross-body correlation.

We ran the LOBO ridge against *both* targets: the training-time $|q_{\text{vel}}|$ proxy and the preregistered *downstream Tier-1* competence (§4). H3 is rejected on both (max LOBO $R^2 = -0.19$ proxy / -3.5 Tier-1), and the Simpson's pattern holds against both targets — so the diagnostics' pooled-vs-LOBO gap is a property of the diagnostics, not of the proxy. This *strengthens* the Simpson's-paradox point and is consistent with §3.1's separate finding that the proxy and downstream competence dissociate.

We do not claim that *no* training-time diagnostic could predict downstream qvel across bodies. The H3 rejection is locally for the specific D1/D3/D4 family we preregistered. Richer diagnostics — e.g. sparse-coding-of-latent dynamics, mutual-information bounds between encoder and downstream control policies, or learned linear weights across body-aware features — remain promising

candidates. Our contribution is the demonstration that the *simplest* preregistered family of diagnostics fails, and that pooled correlation does not substitute for cross-body decomposition.

F SUPPORTING TABLES AND FIGURES

ID	Source	Morph	$n_q/n_v/n_u$
H1_dm	DM `model/humanoid` (Apache-2.0)	humanoid	28 / 27 / 21
G1	menagerie `unitree_g1/scene_mjx.xml` (BSD-3)	humanoid	36 / 35 / 29
Go2	menagerie `unitree_go2/scene_mjx.xml` (BSD-3)	quadruped	19 / 18 / 12
Panda	menagerie `franka_emika_panda/mjx_scene.xml` (Apache-2.0)	manipulator	9 / 9 / 8

Table 7: Bodies used in the multi-body grid.

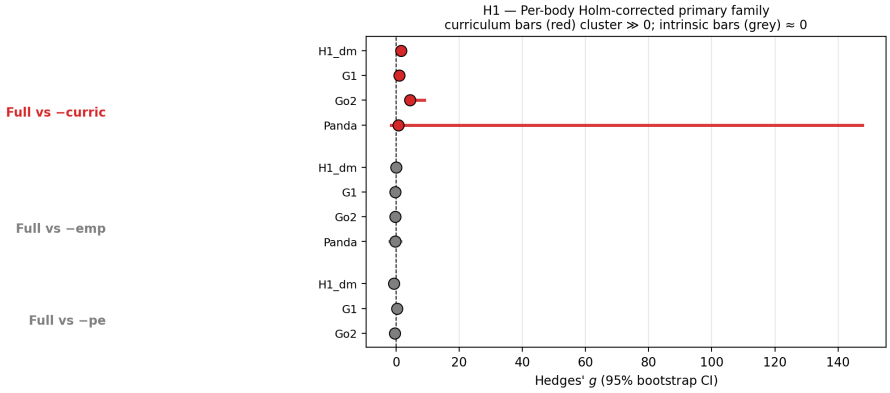

Figure 3: Per-body forest plot (Hedges g with bootstrap 95% CI) for the three preregistered primary-family contrasts. Curriculum bars (red) cluster strongly above 0 on every body; intrinsic-motivation bars (grey) cluster around 0.

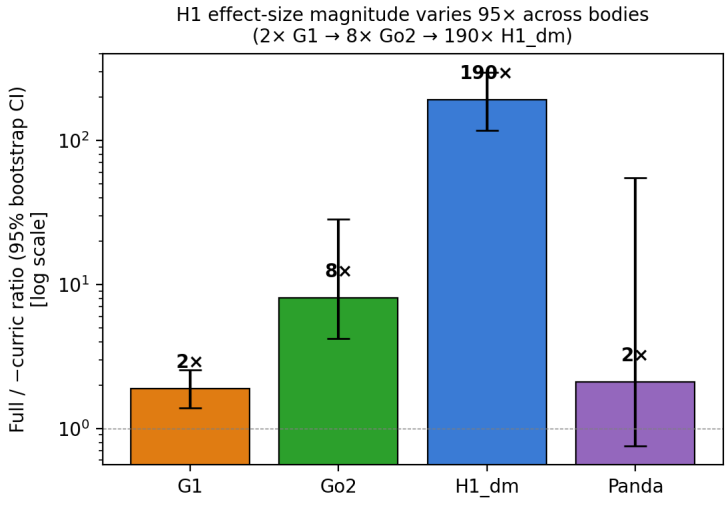

Figure 4: Full/−curric ratio per body on log scale (point estimate + bootstrap 95% CI). 95× spread across bodies; G1 2×, Go2 8×, H1_dm 195×. Panda is shown at its completed n=20 (g=1.12, Holm p=0.013); at an earlier partial $n \leq 9$ cell this contrast was underpowered (see §3(b)).

Body	stand (g)	walk (g)	reach (g)	Holm p (all cells)
H1_dm	$+0.40$	-0.16	-0.38	1.0
G1	$+0.22$	-0.21	$+0.05$	1.0
Go2	— (excl.)	-0.34	-0.26	1.0

Table 8: Zero-shot downstream Tier-1: Full vs $-$curric by (body, task).

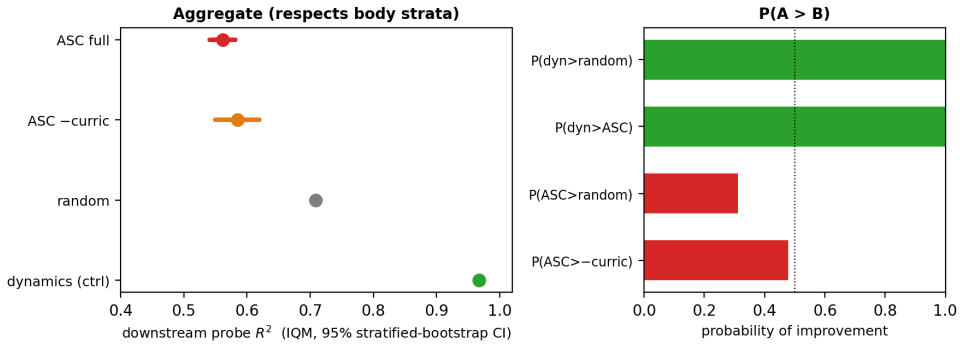

Figure 5: **rliable-style aggregate of the downstream probe** (Agarwal et al., 2021), using a *stratified* bootstrap that resamples within each body (respecting C3, not pooling naively). **Left:** interquartile-mean R^2 with 95% CIs — ASC-full $0.56\,[0.54, 0.58]$ and $-$curric 0.59 sit below random $0.71\,[0.70, 0.72]$, while the dynamics control is near-ceiling $0.97\,[0.966, 0.968]$ (CIs non-overlapping). **Right:** probability of improvement — $P(\text{dyn}{>}\text{ASC}){=}P(\text{dyn}{>}\text{random}){=}1.00$, whereas $P(\text{ASC}{>}\text{random}){=}0.00$ within every body: the curriculum representation is below random init with this (MLP) head.

G Full experimental setup

Bodies and renderer. We use four MJX-compatible bodies from `mujoco_menagerie` (pinned to SHA `b846dd12bc459d776cccb3dee0b1d02acbf7a9c7`) plus the DeepMind humanoid (Apache-2.0):

We tested the Unitree H1 humanoid in menagerie, but it fails MJX with a `(CYLINDER, BOX)` collision-pair not-implemented error on both its `scene.xml` and `h1.xml`; menagerie publishes no `_mjx` variant. Per the preregistration's substitution rule we use the DM humanoid in its place.

Trainer. Trainer architecture follows the M5.2 design used in our pilot single-body study: a state encoder $(q_{\text{pos}}, q_{\text{vel}}) \to z \in \mathbb{R}^{128}$, a 2-layer MLP Gaussian actor with tanh squash, a 5-MLP critic ensemble regressed (smooth-L_1) to a detached intrinsic reward, a MINE-style empowerment *surrogate* (Belghazi et al., 2018) (defined precisely below), and a 4-layer MLP predictor with L_2 prediction error. Two separate optimizers (critic lr 10^{-4}, actor lr $3{\times}10^{-4}$); intrinsic reward target-clip at ± 5 to prevent the critic-loss blow-ups we encountered in early development. The intrinsic objective is $r = \alpha\,\hat{\mathcal{E}} + \beta\,\text{pe}$ with $\alpha{=}5$, $\beta{=}0.5$ as the primary operating point.

Body-aware ROM mask. A per-body unlock policy: uniform for humanoids, legs-first for the quadruped, distal-last for the manipulator (implementation in `env_wrapper_v2.py`).

H LOBO ridge specification and per-fold results

LOBO ridge result (vs. downstream Tier-1). With leakage-free standardization (scaler fit on the training bodies only, then applied to the held-out body; § code note), mean LOBO R^2 is **strongly negative for every diagnostic**:

max LOBO $R^2 = -3.53 \ll 0.5$ (and $\ll$ the 0.3 partial bar; every fold predicts a held-out body worse than that body's own mean). *A reporting error in earlier drafts, and what it hid.* Until a

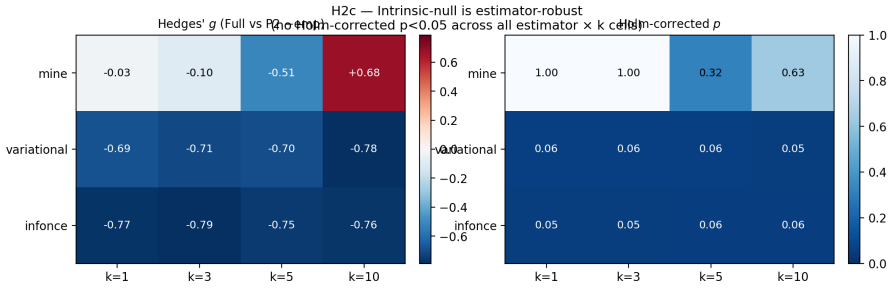

Figure 6: H2c — estimator × window-size k heatmap (k copies of the *current* action; not a horizon, see Appendix J). Left: Hedges g vs P2 −emp baseline (cool = Full lower than −emp; warm = higher). Right: Holm-corrected p. Every cell has $p_{\text{Holm}} \geq 0.05$; the intrinsic null is not rescued by any estimator in this exploratory, underpowered (n=4) sweep.

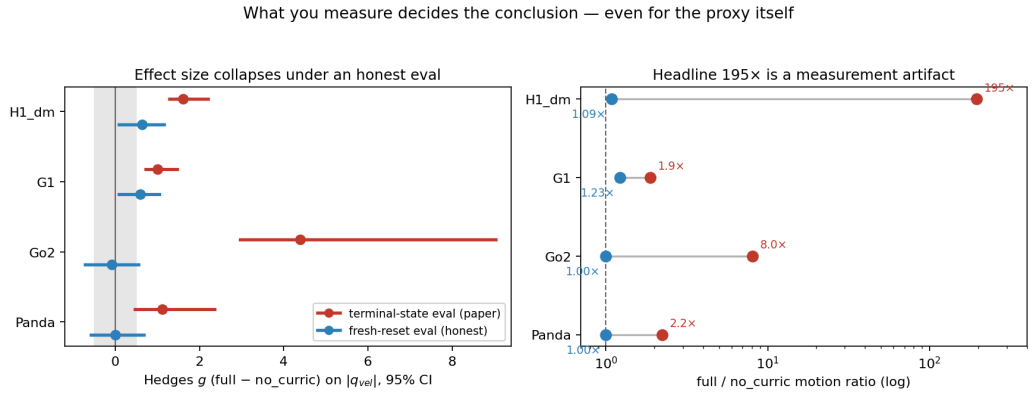

Figure 7: **The curriculum's motion-proxy advantage is largely a terminal-eval artefact.** Hedges g (Full − −curric) on $|q_{\text{vel}}|$ with 95% bootstrap CIs (left) and fold-change on a log axis (right), under the terminal-state eval (the paper's primary) vs. a fresh-reset eval of the same checkpoints. The dramatic terminal-eval effects ($g = 1.0$–4.4; ratios 2–195×) collapse under fresh-reset to null on Go2/Panda (CI spans 0) and a small surviving effect on H1_dm/G1 ($g \approx 0.6$); the H1_dm headline 195× becomes 1.09×.

pre-submission audit caught it, this table — and every LOBO figure derived from it — reported the mean of fold values *floored at* -1 and called it R^2. R^2 has no lower bound; the floor entered the analysis script in the P4 results commit, was never part of the preregistered rule (max LOBO $R^2 \geq 0.5$), and produced the $-0.88/-0.78/-1.00/-0.96$ row. We now report the raw statistic. Two consequences. First, the verdict is unchanged and the correction is in the conservative direction: flooring can only raise a value, so the raw maximum (-3.53) is further from the 0.5 bar than the floored one was. Second, and more useful, the raw folds show *where* the failure lives: the Panda and G1 folds reach -23 to -40 because a held-out body's target *offset* is unpredictable from the other bodies (the argument in the next paragraph), whereas the H1_dm folds sit between -0.3 and -1.1. The floor had compressed exactly that information. The proxy-target numbers quoted in the main text (D3 -0.19, D4 -0.46) never touched the floor and are unchanged; the proxy D1/D2 means move from $-0.87/-0.81$ to $-3.22/-1.87$ and the proxy verdict is likewise unchanged. **H3 is rejected against the downstream target** — no training-time diagnostic predicts out-of-body Tier-1 competence. The rejection is threshold-insensitive (identical at 0.05/0.10/0.30/0.50) and now uses the *right* target, so this is no longer the weaker proxy-only result — it is the preregistered downstream-competence test, and it fails cleanly. *How much of the negativity is the held-out body's level, measured.* Each fold's residual splits exactly into a level term, $n\,(\bar{y} - \bar{\hat{y}})^2$, and a within-body term (`p10c_lobo_offset_decomp.py`, which reproduces every fold R^2 above to 10^{-9}). The level term is 60%, 62%, 80% and 64% of SS_{res} on average for D1–D4, and 92–98% on the folds below -10 (Go2 and Panda for D1; G1 and Panda for D3; Panda for D4). Supplying each held-out

Diagnostic	D1 eff-rank	D2 occ-entropy	D3 probe→body z	D4 pe-residual
Mean LOBO R^2 (raw)	-8.82	-3.53 (body-const.)	-17.21	-7.78
G1 held out	-0.92	-7.77	-39.69	-2.20
Go2 held out	-10.65	-2.48	-2.75	-3.69
H1_dm held out	-0.60	-0.35	-1.11	-0.84
Panda held out	-23.10	—	-25.30	-24.40
Floored at -1 (earlier drafts)	-0.88	-0.78	-1.00	-0.96

Table 9: LOBO ridge R^2 for the H3 diagnostic family, vs. downstream Tier-1 competence (leakage-free train-fold standardization). The primary rows are the raw fold R^2 and their mean; the last row is the statistic earlier drafts reported *as* R^2, shown for continuity and discussed below. D2 has no Panda fold (no holdout bank). Per-fold $\mathrm{SS}_{\mathrm{res}}$, $\mathrm{SS}_{\mathrm{tot}}$ and n are in the released JSON.

body's true level — the most favourable reading possible — leaves within-body R^2 of -0.14, 0.00, $+0.003$ and -0.30: H3's rejection does not depend on the level problem. D2 is the clean case: it is constant within a body, so its within-body R^2 is exactly 0 and its entire LOBO score (-3.53) is the level term. Two limits of scope. The result is about these four diagnostics, these four bodies and this ridge procedure, not an impossibility claim for training-time diagnostics. And the Tier-1 composite averages body-specific tasks on body-specific scales, so cross-body levels are only partly commensurable — one more reason the level term should not be read as the diagnostic's failure and was separated out.

Proposition 1 (the pooled correlation does not identify the sign of leave-one-body-out predictability). The decomposition is the classical between/within argument (Yule, 1903; Simpson, 1951; Robinson, 1950; Gelman & Hill, 2006); we claim no new statistics. What we state is a *non-identifiability* result with a fixed estimand, a fixed learner and a stated asymptotic regime, which is narrower than the general reading an earlier draft invited ("no aggregate is a consistent estimator of LOBO R^2") and is the version the proof actually supports.

Setting. Bodies $k = 1, \dots, K$ are exchangeable draws from a body-level law: body k has means (μ_k^D, μ_k^Y) drawn from a bivariate normal law with variances (τ_D^2, τ_Y^2) and correlation r_b; conditional on them, body k contributes n_k i.i.d. within-body pairs (D, Y) from a bivariate normal law with variances (σ_D^2, σ_Y^2) and correlation r_w, the same for every body. *Estimand:* the leave-one-body-out predictive R^2 of ordinary least squares of Y on D, fit to the pooled sample of the other bodies and scored on the held-out body against that body's own mean, $R^2_{\mathrm{LOBO}} = 1 - \mathbb{E}[\mathrm{SS}_{\mathrm{res}}]/\mathbb{E}[\mathrm{SS}_{\mathrm{tot}}]$. *Regime:* $K \to \infty$ and $n_k \to \infty$, so the fit converges to the population regression of the pooled law. Write Σ for the covariance of a single pooled observation with its body label discarded; under the setting, $\Sigma_{DD} = \sigma_D^2 + \tau_D^2$, $\Sigma_{YY} = \sigma_Y^2 + \tau_Y^2$, $\Sigma_{DY} = r_w \sigma_D \sigma_Y + r_b \tau_D \tau_Y$.

Claim. There exist two processes $\mathcal{P}$ and $\mathcal{P}'$ under which a single pooled observation (D, Y), body label discarded, has the *same* distribution, while $R^2_{\mathrm{LOBO}}(\mathcal{P}) > 0 > R^2_{\mathrm{LOBO}}(\mathcal{P}')$. Hence no population quantity of the unlabelled pooled law — the pooled Pearson correlation, the pooled Spearman correlation, and the limit to which a bootstrap or IQM of either converges among them — determines the sign of cross-body predictability. We do not claim more: statistics that use the body labels (a within-body correlation, a LOBO fit itself) *can* separate the two processes, and that is exactly the recommendation.

Proof. Take $\mathcal{P}'$: $r_w = 0$ (within a body D and Y are independent, since the within-body law is normal), $0 < r_b < 1$, $\tau_D, \tau_Y > 0$. Take $\mathcal{P}$: no body offsets ($\tau_D = \tau_Y = 0$) and a within-body normal law whose covariance is the Σ of $\mathcal{P}'$. A single unlabelled observation is $\mathcal{N}(0, \Sigma)$ under both, so every functional of that law coincides, including $\rho^\star = r_b \tau_D \tau_Y / \sqrt{\Sigma_{DD}\Sigma_{YY}}$. Both pooled fits converge to the same line, $\hat{Y} = \beta D$ with $\beta = \Sigma_{DY}/\Sigma_{DD} = r_b \tau_D \tau_Y / (\sigma_D^2 + \tau_D^2)$. Under $\mathcal{P}$ every body has the pooled law, so the held-out residual variance is $\Sigma_{YY}(1 - \rho^{\star 2})$ against a within-body variance Σ_{YY}: $R^2_{\mathrm{LOBO}}(\mathcal{P}) = \rho^{\star 2} > 0$. Under $\mathcal{P}'$, on a held-out body k the residual is

$(\mu_k^Y - \beta\mu_k^D) + (\varepsilon_Y - \beta\varepsilon_D)$ with independent within-body noises, so

$$\frac{\mathbb{E}[\mathrm{SS}_{\mathrm{res}}]}{n_k} = \sigma_Y^2 + \beta^2\sigma_D^2 + \underbrace{(1-r_b^2)\tau_Y^2 + (\beta - r_b\tau_Y/\tau_D)^2\tau_D^2}_{\mathbb{E}[(\mu_k^Y - \beta\mu_k^D)^2]}, \qquad \frac{\mathbb{E}[\mathrm{SS}_{\mathrm{tot}}]}{n_k} = \sigma_Y^2,$$

giving $R^2_{\mathrm{LOBO}}(\mathcal{P}') = -\big[\beta^2\sigma_D^2 + (1-r_b^2)\tau_Y^2 + (\beta - r_b\tau_Y/\tau_D)^2\tau_D^2\big]/\sigma_Y^2 < 0$, strictly because $r_b < 1$ and $\tau_Y > 0$. □

The learned predictor is literally the same function under both processes; what differs is only whether a new body's level is predictable from its diagnostic level. (Our real LOBO folds have exactly this structure: Appendix H, the offset decomposition.) Any $\rho^\star \in (0, r_b)$ is attainable by scaling τ_D/σ_D and τ_Y/σ_Y together.

What the proposition does and does not license. It licenses one sentence: a pooled cross-body correlation, however tightly estimated, does not tell you whether a diagnostic will predict a *new* body. It does not say pooling is wrong for the question pooling answers (which method is best on average over *these* bodies), and it is silent on the rank-vs-mean disagreement and the no-domain-sweep observations in §4, which are different phenomena and are argued there on their own evidence.

Checks, and what each one is. Two numerical checks accompany the proof, and they are different kinds of object. *(i) The construction itself*, simulated (`prop1_mc_check.py`; $\sigma = 1$, $\tau = 2$, $r_b = 0.8$, 400 bodies of 200): pooled Pearson 0.64 vs 0.64 and pooled Spearman 0.62 vs 0.63 for $\mathcal{P}$ vs $\mathcal{P}'$, OLS LOBO $R^2 = +0.41$ vs -2.05 against the proof's $+0.41$ vs -1.95. *(ii) A finite-sample illustration* in P23's simulator, which differs from the proposition in three stated ways: 4 bodies of 40, Spearman rather than Pearson correlation, and ridge (α=1) rather than OLS. Its two locked configurations have unequal pooled ρ (0.61 confounded, LOBO R^2=-1.15; 0.77 genuine, $+0.55$; P23's locked file records the first as -0.77, the value floored at -1 disclosed for Table 9), so post hoc we scaled the confounded configuration's offset spread until the pooled correlations matched (`p23_equal_rho_witness.py`): at pooled $\rho = 0.77$, LOBO $R^2 = -1.37$ confounded vs $+0.55$ genuine; at $\rho = 0.61$, -1.15 vs $+0.35$; over 200 independent draws the matched confounded process averages -2.34 and is negative in every draw. The matching is on the correlation only, not on the whole pooled law, so (ii) illustrates the proposition in the regime our data occupy; (i) is the check of the proposition as stated. Stratifying does not rescue the confounded case — its bootstrap IQM is 0.61 with CI $[0.53, 0.69]$, a tight interval around a number that says nothing about new bodies.

I PREDICTION-ERROR MECHANISM INVESTIGATION

Why does prediction-error suppress motion? — a mechanism, and a measurement caveat inside it. The natural explanation is a *predictability attractor*: by rewarding a well-modeled world, the pe term could push the policy toward intrinsically predictable, low-motion states. We tested this directly (P11 mechanism, `p11_mech_results.md`): on the two suppressing bodies (H1_dm, Go2, n=20/15) we rolled out the true-gated `full` and `no_pe` policies and measured, along each trajectory, (i) mean $|q_{\mathrm{vel}}|$, (ii) planar-root occupancy entropy, and (iii) a *post-hoc* 1-step forward-model residual (post-hoc because under true gating the `no_pe` predictor is untrained, hence not comparable across arms). Motion suppression re-confirms strongly ($g = -3.28/-1.61$). The *raw* predictor residual is also much lower for `full` ($g = -3.32/-1.01$), seemingly supporting the attractor — **but this is a motion-scale artefact**: a body that barely moves produces small state deltas that are trivially predictable in absolute terms. After normalizing by state variance ($1 - R^2$), `full`'s predictability advantage is negligible on H1_dm ($g = -0.87$) and *reverses* on Go2 ($g = +0.41$); occupancy entropy is null on both ($g \approx +0.11$). So the honest mechanism is **direct motion suppression** — the pe objective, computed on latent transitions, penalizes the large state changes vigorous movement produces — *not* a preference for predictable states. That the "obvious" mechanistic story survives the raw metric but dies under the motion-normalized one is itself an instance of this paper's thesis: even a mechanism explanation is a measurement choice.

For completeness we retain the original reward-weight analysis (the documented P2 baseline), which we describe next as the confounded contrast it is.

How large each path is (post hoc, `p10a_gradient_paths_replay.py`). Released checkpoints are 90–100% saturated with the actor's log-std at its clamp, so gradients measured there are

Ablation of prediction error	intrinsic reward	aux. gradient → encoder	aux. gradient → actor	predictor in optimizer
Full (both trainers)	on	on	on (action not detached)	yes
Reward weight, P2 (β=0)	**off**	on	on	yes
Component gating, P10a	**off**	**off**	**off**	**no**

Table 10: What each prediction-error ablation actually removes, read from `m5_v4_run.py` (the P2 arms predate its gating flag and set only β=0). Gating removes three paths at once, so the C2 difference is attributable to the auxiliary loss as a whole, not to its encoder path or its actor path separately. The same holds for empowerment. Separating them needs a reward × encoder-gradient × actor-gradient design, which we have not run; the paragraph below measures how large each path is.

mostly zero on the actor. We therefore replayed the P2-era trainer (10 seeds × {Full, β=0} × 3 bodies; every logged value matches the released runs) and read each loss's gradient on the encoder and actor parameter groups every 10 steps without changing the update. As a share of the summed gradient's squared norm (Full arm, seed-bootstrap CIs): prediction error supplies 83–94% of the *encoder* gradient on H1_dm and Go2 in the first 100 steps (56% on G1) and 98–99% on every body after step 500; on the *actor* it supplies 34–45% in the first 100 steps on every body, falling to 13% (H1_dm) and 4% (Go2) late as the policy saturates but rising to 84% on G1. Its direction is roughly orthogonal to the policy gradient ($|\cos| \leq 0.25$ throughout). The β=0 arm looks the same, confirming that weight ablation leaves both paths open; the gated arm has exactly zero prediction-error gradient and no predictor in its optimizer state (30/30 checkpoints). So gating removes a real actor-side term as well as the encoder-side one, and the reading "C2 is an encoder-representation effect" is not supported. Which path produces the C2 difference is still unknown: a gradient share is not an effect, AdamW rescales per coordinate, and the global clip at 1.0 is active in $\geq$ 99% of steps. Detaching the action inside the prediction-error loss (encoder path only) or detaching z_t (actor path only) would separate them.

Reward-weight ablation (the confounded baseline). An important caveat on what `-emp`/`-pe` remove in the P2 baseline: the flags zero the empowerment/prediction-error *reward weight* (α or β), but in the current trainer the empowerment and prediction-error *auxiliary losses* still update the encoder/predictor/estimator (they are summed into the actor-side objective regardless of the flags). So these are **reward-weight ablations, not removals of the intrinsic components from representation learning** — we do not claim the latter; the true component-gated rerun is the main-text result (§3.2, P10a). With that scope: on the three bodies, all six Full-vs-{−emp, −pe} contrasts on $|q_{\text{vel}}|$ are Holm-non-significant after the primary-family correction (the closest near-miss is H1_dm Full-vs-−pe: $g = -0.64$ $[-1.30, -0.03]$, raw $p = 0.054$, $p_{\text{Holm}} = 0.48$). A Welch-TOST equivalence analysis at margin $\pm 0.3 \cdot$ SD (Lakens, 2017) (analysis plan locked post-hoc in `p6_stats_upgrade.md`) shows that **no cell reaches statistical equivalence either** (all $p_{\text{TOST}} \geq 0.24$): across the six completed-body cells, two are inconclusive ($|g| < 0.3$ but underpowered for the equivalence bound) and four carry point estimates $|g| \geq 0.3$ (the two underpowered Panda cells add one of each). Read on its own this contrast would suggest the intrinsic-motivation half of ASC contributes nothing *detectable* at n=20/cell — but, as the true-gating rerun above shows, that reading is an artefact of the ablation method: it conflates a genuine empowerment null with a masked, large prediction-error *suppression*. Per the preregistration §3.4 underpower-flag rule we do not assert a *true* null for empowerment, and we defer to §3.2's component-gated split for the substantive intrinsic verdict.

J Estimator robustness of the intrinsic result (H2c)

Estimator-robustness of the intrinsic result (exploratory, H2c). An exploratory sweep on H1_dm — 3 empowerment estimators (MINE, the variational lower bound, InfoNCE/contrastive) × 4 window sizes $k \in \{1, 3, 5, 10\}$ × 4 seeds, 48 runs — finds **no estimator/window cell** whose Full $|q_{\text{vel}}|$ differs from the P2 −emp baseline after Holm m=12 correction (all $p_{\text{Holm}} \geq 0.052$; Figure 6). At n=4/cell this is a robustness check, not a confirmatory family: it indicates the §3

intrinsic result is not an artefact of the MINE estimator family. It says nothing about *horizon*: in both trainers the k-step "action window" is the current action repeated k times and the future state is the next step's latent (`m5_v4_run.py`, `m15_asc_corrected.py`), so k changes the estimator's input width, not the control horizon, and repeated copies add no action-sequence information. *What the surrogate estimates.* Negatives are actions shuffled across environments, i.e. drawn from the marginal $p(a)$ rather than $p(a \mid s)$; with an ideal critic the Donsker–Varadhan objective then targets $I(A; S, S')$, which exceeds the one-step empowerment $I(A; S' \mid S)$ by $I(A; S)$, and a finite-sample estimate is not guaranteed to be a bound. Every empowerment statement in this paper is therefore about this one-step surrogate; we make no claim about multi-step or conditional-negative empowerment, which would need a new estimator validated on a task with known mutual information. The consistent negative direction of the variational/InfoNCE cells ($g \approx -0.7$, raw $p \approx 0.005$–0.010) is logged as a follow-up hypothesis requiring ≥ 20 seeds per cell.

K MAIN-TEXT DETAIL RELOCATED FOR THE ICLR PAGE LIMIT

The following passages were moved here verbatim from the main text to meet the 9-page submission limit; each is pointed to from its original location. No result is changed.

What the frozen probe measures, and a consistency check we do not build on. The probe decodes three targets from a frozen encoder's latent, with the encoder's input being $(\text{qpos}, \text{qvel})$. Two of those targets — root height and forward velocity — are functions of that input (forward velocity *is* one of its columns), so a high probe R^2 says mainly that the encoder preserved its input, not that it learned task-relevant structure. Two consequences we state rather than leave for a reader to find: an *untrained* encoder already scores well on this probe (random init beats ASC on every locomotor body: 0.714 vs 0.563 on H1_dm, 0.519 vs 0.334 on G1), and the self-supervised dynamics control's objective — predict the next-observation delta — optimizes exactly that preservation, so its near-ceiling probe score is close to definitional. We therefore rest the positive-control argument on the RL-transfer routes (§3.1), which carry no such circularity, and report the probe as an information-preservation measure. A related minor observation, which we do *not* claim as a contribution for the same reason: the cheap probe ranks the reward-free methods consistently with frozen-transfer across all four embodiments (mean Spearman $+0.77$, top-1 in 3/4, combined $p = 0.033$ respecting C3); the agreement is partly by construction, fragile, and fails under fine-tuning. A further caveat on the probe's split: frames from 64 parallel environments are pooled and split $80/20$ at random, so a test frame's temporal neighbours sit in train; a contiguous split was tried first and gave negative R^2 for every source, and we switched after seeing that — an analysis decision made on the data, which inflates the absolute levels. *Trajectory-separated re-evaluation (post hoc, `p11_probe_trajsplit.py`).* Re-scoring the same rollouts and frozen features with whole environments held out (5-fold over the 64 envs; the random-frame split reproduces all 96 committed cells exactly) leaves the source ordering *identical* within each head on every body, for all 96 cells. Absolute R^2 falls by 0.04–0.17 (MLP) and 0.03–0.08 (linear) — that much was leakage. The MLP gap Full$-$RANDOM shrinks by 40–50% on H1_dm and G1 but stays negative on all three bodies ($\Delta R^2 = -0.089, -0.096, -0.094$; seed-bootstrap CIs excluding 0; exact paired sign-flip $p = 0.008$, the floor at n=8); Full$--$curric stays null (CIs covering 0); the dynamics control stays near-ceiling (0.89–0.93). The original contiguous split's negative R^2 turns out not to be leakage but non-stationarity: under the random policy the humanoid falls (H1_dm root height $1.28 \rightarrow 0.20$ by step 240), so the last 20% of each rollout is out of support and only the dynamics encoder extrapolates to it.

K.1 COLLECTED CAVEATS

Each of these is stated where it arises; we collect them here so a reader does not have to assemble the list themselves.

1. **The proxy fails its own positive control on Panda** (§3(e)): the fixed-base arm's instrument is not validated, so Panda's competence reading rests on the transfer battery alone.
2. **G1's corrected $-$curric arm completed $10/20$ seeds** (§3(e)): the highest-DoF body diverges under full-ROM control. The divergence rate is itself an outcome and we report it as one ($10/20$ at $1\times$, $9/10$ at $10\times$). Reported as two parts (`p19c_g1_two_part.py`): completion is $20/20$ for

Full against $10/20$ for $-$curric (Fisher exact $p = 0.0004$); conditional on completion, Full moves less (44.7 vs 65.0 fresh-reset $|q_{\text{vel}}|$ at training gain, $g = -5.78\,[-8.88, -5.11]$). We impute nothing for the diverged runs, which have no fixed-time measurement. Instead we compute the value all ten would have to share for the conditional contrast to vanish: 24.4, below every completed $-$curric seed (59.8–74.5) and below the Full mean. The contrast is conditional on survival, and survival itself differs by arm.

3. **The corrected-trainer null in (e) is not attributable to the gain fix alone** (§3(e)): P19 bundles six changes (per-env intrinsic reward among them; enumerated in Appendix K), so its $0/4$ is a property of the corrected trainer as a whole. The attribution of the *original* effect to the evaluation gain rests on (d) — the same policies re-measured with only the eval gain changed (P16) — and on the trainer-independent simulation (P22), not on (e). A factorial crossing of gain handling against the other fixes would separate them; we have not run it. Own-gain evaluation also gives each arm a different control authority, so (e) removes the evaluation artefact but is not a same-authority comparison of learned competence.
4. **"Empowerment" is a one-step surrogate, and k is not a horizon** (Appendix J): the estimator repeats the current action k times, looks one step ahead, and draws negatives from the marginal action distribution, so it targets $I(A; S, S')$ rather than $I(A; S' \mid S)$. The "no detectable empowerment effect" finding and the H2c window sweep are statements about this surrogate only.
5. **H1_dm's gain attribution depends on the roll-in length** (Appendix K): matching the gain removes 74% of the training-terminal ratio but only 21% of the much smaller ratio after a 200-step roll-in; on Go2 it removes the advantage at both lengths. From a common, policy-independent state bank with matched gain and common noise, the Full advantage is absent on Go2 and non-significant on H1_dm (1.05–$1.14\times$); all of this is $n{=}5$ and post hoc. The damped system (P22) shows the artefact needs no ASC code; how often it alters published conclusions elsewhere we have not measured.
6. **The realized Holm family is narrower than preregistered** (§2): only the $|q_{\text{vel}}|$ contrasts ($m{=}12$) were run as one confirmatory family; Tier-1 and H2 are labelled exploratory or post-amendment in place.
7. **The frozen probe's train/test split was chosen after seeing the data** (this appendix, above): a contiguous split gave negative R^2 for every source, and we switched to a random frame split. It inflated absolute levels by 0.03–0.17; a post-hoc trajectory-separated holdout leaves the source ordering unchanged and the MLP Full$-$RANDOM gap negative (this appendix, above). "ASC below random" holds for the MLP head only; a linear head puts ASC marginally above random. The probe's targets are also largely functions of its input, which is why the positive-control argument rests on the transfer routes.
8. **P20 has no lock-before-run commit** (this appendix, above), unlike every other study reported here.
9. **The P28 grader resolves an exact rank tie by dictionary order** (this appendix, above); we report the locked verdict and a post-hoc tie-aware re-grade side by side rather than editing the locked instrument.
10. **The intrinsic-reward weights were selected on the primary outcome metric, and only the Full arm was tuned** (§3(a)): α, β come from a 3×3 single-seed pilot grid (`run_m5_grid.sh`; 400 steps, 256 envs, one body) scored on `eval_qvel_mean` — the same $|q_{\text{vel}}|$ that becomes the H1 outcome. $\alpha{=}5$ is the argmax by roughly $6\times$ (79.4 vs 12.8 for the best $\alpha \in \{0, 1\}$ cell), and $\beta{=}0.5$ is *not* the argmax ($\beta{=}0$ scores 79.4, $\beta{=}0.5$ scores 47.0); we kept $\beta{=}0.5$ to retain a prediction-error term. No cell in the grid ran `--no_curriculum`, so the $-$curric arm inherits $\alpha{=}5$ untuned and the primary contrast is tuned-vs-untuned. The selection was never re-validated at the production scale (1000 steps, 512 envs, four bodies). The direction is the reason we state it plainly rather than bury it: asymmetric tuning inflates the *Full* arm, so it biases the primary contrast *toward* the curriculum advantage we report and then retract — it strengthens C1 rather than threatening it. Disclosed, this is one more instance of the paper's own thesis; left undisclosed, it would be a finding against it.
11. **Transfer is compared at final performance, and the tallies are cells of one battery** (§3.1): the $0/8$, $0/3$ and $0/2$ counts are cells of a single downstream battery on the same checkpoints, not independent replications, and every route compares end-of-fine-tuning performance. A pretraining benefit that shows up as fewer downstream interactions to a fixed success rate — the benefit

practitioners usually want — is not excluded by a tie in final performance; we did not compare learning curves or time-to-threshold.

12. **C2's direct contrast was added post hoc, and it is a re-evaluation, not a factorial training design** (§3.2): the original reward-weight and gating analyses used different evaluation regimes. We re-evaluated all four arms' checkpoints under *one* regime (`p10a_matched_contrast.py`; the re-evaluation reproduces the in-training values at $r > 0.9998$; the two Full arms are separate runs of two trainer versions whose fresh-reset means agree within 0.5). Fresh-reset, in raw $|q_{\text{vel}}|$ (the primary scale, since standardized differences also move with each arm's SD): the gating gap exceeds the weight-ablation gap by $21.0\,[13.5, 28.0]$ on H1_dm, by $2.05\,[-0.01, 4.05]$ on Go2 (bootstrap $p = 0.051$), and by $0.8\,[-10.3, 12.2]$ on G1. Standardized: weight ablation $g = -0.80\,[-1.36, -0.31]$ vs gating -2.97 on H1_dm ($\Delta g = -2.16\,[-4.17, -1.09]$), Go2 $\Delta g = -0.73\,[-1.74, -0.08]$, G1 $-0.13\,[-0.97, +0.80]$. So the direct evidence is strong on H1_dm only; Go2 is borderline in raw units. Terminal-state gives the same pattern (H1_dm -2.17, Go2 -1.25, CIs excluding 0). On H1_dm weight ablation does detect the effect under fresh-reset, so "null under weight ablation" was regime-specific; the robust statement is the difference in size. The ablated arms come from two training grids (P2 and P10a); a reward-gate $\times$ auxiliary-gate design trained together would be cleaner and has not been run.
13. **The statistical protocol is not uniform across studies** (§2): the confirmatory $|q_{\text{vel}}|$ family uses Welch, Mann–Whitney, Hedges' g with B=20,000 CIs and Holm; the benchmark audits use B=10,000 (P28) and B=2,000 stratified (P23); P25 is point-estimate only, superseded by P26; and the LOBO regressions report R^2 with no interval. Holm is applied to the Welch p-values only — the Mann–Whitney p is reported raw and is never family-corrected, so it should not be read as a second corrected test.
14. **All three of our own ranking instruments break ties toward the first-listed method** (this appendix, above): P26 by `argmin`, P28 by `min(dict, key=dict.get)`, P25 by a stable `sorted()`. Each is disclosed with the numbers the convention moves; P25's affects one graded verdict on a panel this paper does not read.

Our own rank estimator's tie handling moves these numbers. The P26 bootstrap picks a rank winner with `argmin`, which under an exact tie silently credits the method earliest in the method list. Mean ranks over 4 integer task-ranks tie often enough to matter: 6.7% of draws for the pooled rank, and far more per domain. Splitting tied credit evenly instead moves the pooled rank-winner split from Disagreement 0.685 / RND 0.315 to 0.651 / 0.349, the estimator-disagreement rate from 34.2% to 33.0%, the sweep probability from 4.6% to 5.5%, and — the largest single shift — Quadruped's $P(\text{Disagreement top-1})$ from 0.79 to 0.69, with RND taking the difference. None of these reverse the claim we make from this panel (no method is crownable; the two estimators part company on a third of draws), which is why the main text now quotes ranges rather than the single locked digits. We report it because it is the paper's own thesis applied to us: a tie-breaking convention inside an estimator moved a published probability by ten points.

P20 was not locked before running, and we say so. Every other study we report has a protocol commit that precedes its results commit with no result files in the tree. P20 — the C2 externalization on shared-encoder ICM/RND — does not: its two result commits (MountainCar, then MiniGrid 22 minutes later) have no lock ahead of them. The order is the part a reader should weigh, because the first environment produced the *negative* result and the second the positive one. What we can show is that the negative result was committed and kept rather than discarded, with its own commit message recording that it "weakens C2 generality"; the boundary condition it established (the effect vanishes where a near-random encoder already saturates the probe) is reported in §3.2 as a limit on C2, not omitted. But we cannot claim preregistration for P20, and we do not.

Effect-size heterogeneity, and which body "wins" depends on the measure. The two arms' variances differ by orders of magnitude on some bodies — on H1_dm the Full arm's variance is over 7×10^4 times the $-$curric arm's (8.91 vs 0.033 SD), because the $-$curric arm sits at the noise floor. Hedges' g standardizes by a pooled SD, which under that much heterogeneity is effectively the larger arm's SD, so g should be read as a standardized contrast and not as a scale-free statement about which body gained most. It matters here: ranked by g the order is Go2 (4.39) > H1_dm (1.61) > Panda (1.12) > G1 (1.00), but ranked by raw ratio it is H1_dm $(195\times)$ > Go2 $(8.0\times)$ > Panda $(2.2\times)$ > G1 $(1.9\times)$ — the top two swap. This is a small instance of the paper's own thesis,

and we report it rather than picking the measure that reads best. What both measures agree on, and what §3(b) actually claims, is that the largest-DoF body (G1) shows the *smallest* relative gap. All significance testing uses Welch's t, which does not assume equal variances.

Our seed-sizing calculator uses the equal-variance degrees of freedom, and that is the unsafe direction. The power routine behind the preregistered sizing table takes $\nu = n_1 + n_2 - 2$ rather than the Welch–Satterthwaite ν. Its docstring says so, but the consequence was not stated: since the pooled ν is the larger one, the routine *overstates* power and therefore *under*-sizes seed budgets — it errs toward telling us we had enough seeds, which is exactly the error a preregistration is supposed to prevent, and this project has twice been short of seeds. We quantified it rather than leaving it qualitative. At the variance ratios we observe (up to $\approx 16\times$ between arms) the overstatement is at most 0.05 in absolute power, and the required n per arm at $\alpha/12$ and 80% power moves from 11 to 13 at g=1.8 and from 6 to 7 at g=2.9; at the underpowered Panda cell (g=0.38) it moves from 193 to 194, which changes nothing about a cell we already report as hopeless. The approximation is also the conventional one for *planning*, where the variance ratio is not yet known. We leave the calculator as it was locked and record the correction here.

The Simpson detector, audited. We release the "Simpson warning" as a checklist item, so its own failure modes matter. A pre-submission audit found the detector compared undefined within-body correlations with $<$, which is always false in floating point, so a single body with no defined ρ (a body-constant diagnostic, or one with no holdout bank) silently prevented the warning from ever firing. Two sentences in §4 consequently disagreed with our own released JSON. The detector now ignores undefined bodies and requires at least two defined ones; the fix changed no LOBO R^2 and no verdict, only the flag (D3 now fires against both the proxy and the Tier-1 target, matching the pattern the text describes). One limit remains by construction and we state it rather than hide it: on a diagnostic that is constant within every body — D2 here, whose within-body ρ is undefined on all three locomotor bodies — there is nothing to decompose, so the warning cannot fire and the user must notice the structural nullity themselves, as we report for D2 in §4.

Map of the Pn run codes used in the main text. Each code names a released run log and its locked protocol in the supplementary material (`experiments/` for results and per-seed data, `src/phase2/` for the runner and aggregator that produced them).

Code	What it is
P2	The primary 318-run H1 grid (four bodies, four arms, Holm family)
P5	Scaled vanilla baseline: the compute-matched no-curriculum contender
P6	Statistical upgrade: bootstrap CIs, TOST, power/MDE calculator
P7	Scale-robustness of the empowerment null
P10a/P10b/P10c	Gated re-ablation; Tier-1 downstream target; LOBO on that target
P11/P11-C	Off-policy task transfer; frozen-encoder representation probe
P12	Generality positive control (does the transfer protocol discriminate?)
P14	Locomotion transfer on the largest-proxy body
P15	External reward-free baselines (ICM/RND) under a locked validity gate
P16	Gain/parking diagnostic: the instrument artefact isolated
P19/P19b	Corrected-trainer re-run; quantification of the G1 divergent seeds
P20	C2 externalized on a public intrinsic-motivation setup
P22	Trainer-independent damped-point-mass simulation of the gain artefact
P23	Stratified-bootstrap vs leave-one-body-out demonstration
P24/P24b/P24c	Per-body positive-control battery for the corrected proxy
P25/P26	URLB published-table pooling audit; its inferential upgrade
P27/P27b	$10\times$ budget-scaled control; its optimization-health diagnostics
P28	End-to-end rerun of the URLB standard pipeline (144 runs)

External-baseline rerun: scope of the claim (from Limitations). P15 failed two validity checks (collapsed ICM/RND configs, r_{int} decaying 300–1000$\times$; an underpowered permutation test, $p_{\min} = 0.083$); the corrected-backbone rerun fixes both ($p = 0.035/0.017$) and leaves the corrected ICM/RND transfer *below random* — so the comparison supports a consistency check, not an absolute-superiority claim in either direction. Downstream transfer is measured across four routes (§3.1); all four agree that the curriculum does not transfer (Go2 gate miss $0.12 < 0.15$).

(c) The proxy magnitude is itself eval-protocol-dependent. The fold-changes in §3(a) use the terminal-state eval (a 100-step eval continuing from the training terminal state). Re-measuring the *same* checkpoints under a fresh-reset eval (fresh init, full ROM, stochastic policy — a more faithful stand-in for competence) compresses the advantage sharply: the H1_dm fold-change falls from $195\times$ to $1.09\times$, and by Hedges g the curriculum motion effect is not detected on Go2 and Panda (95% CI spans zero) and only small on H1_dm/G1 ($g \approx 0.6$; Figure 7). Even the *size* of the proxy effect depends on how the proxy is measured; the dissociation is unchanged (the curriculum fails to transfer under either eval, §3.1), but the raw fold-changes are terminal-state-eval quantities, not competence gaps.

An eval-regime nuance, stated plainly (from §3.1). Zero-shot Tier-1 is evaluated at *full ROM*, which is exactly the $-$curric *training* regime (ROM $= 1$ from $t{=}0$), whereas Full trained mostly under restricted ROM. The full-ROM eval may therefore *favor* $-$curric's training distribution. A skeptic could still object that zero-shot is the wrong test — the field's claim is that pretraining helps *after* downstream RL. We close that objection with three further measurements on the same checkpoints.

Aggregated with rliable-style estimators (from §3.1). To confirm the rankings are not an artefact of mean-based summaries, we recompute the downstream probe with rliable's robust estimators (Agarwal et al., 2021) — the interquartile mean (IQM) with a *stratified* bootstrap that resamples within each body (respecting C3, not pooling naively) and the probability of improvement (Figure 5). The ordering is sharp with non-overlapping 95% CIs (ASC-full 0.56, random 0.71, dynamics 0.97), and $P(\text{ASC} > \text{random}) = 0.00$ on *each* of the three bodies — the curriculum representation scores below random init on this probe, not merely no better (with the MLP head; the gap shrinks by 40–50% but survives a trajectory-separated holdout, Appendix K). (Computing that same probability by pooling seeds across bodies instead returns 0.31, because pooling counts "ASC on one body beats random on another" as an improvement. We report the within-body value; the gap between the two is C3 in miniature.)

(d) Gain-diagnostic mechanics (full detail from §3(d)). In a trainer-free diagnostic (P16, `p16_gain_diagnostic_results.md`; released checkpoints, $n{=}5$ per cell) we rolled each policy in for 1000 steps at *its own* training gain and then, *from the identical state*, evaluated twice changing only the gain. Every cell is saturated (99.6–100% of actions at the bound) and parks (joint-limit occupancy 0.25–0.79, roll-in $|q_{\text{vel}}|$ falling to 0.05–0.93); evaluating at 1.0 multiplies Full's measured motion by $15.6\times$ (H1_dm; per-seed 2.4–$33.7\times$) and $26.3\times$ (Go2; per-seed 5.7–$49.3\times$) while leaving $-$curric's unchanged ($1.0\times$, its training gain already being 1.0), and removing that asymmetry removes 74% (H1_dm) and 93% (Go2) of the Full-vs-$-$curric ratio. P16 is exploratory ($n{=}5$, no significance testing), and it is the only analysis that changes the evaluation gain *alone*; the $n{=}20$ corrected re-run (e) changes six things at once and is evidence about the corrected pipeline, not an independent confirmation of P16's attribution. So the proxy effect is not acquired motor competence but largely a measurement of who receives a torque boost at evaluation and how deeply each policy parked — the mechanism we diagnosed for the invalidated `reverse` schedule (§B.1) and failed to generalize to `linear`. P16 is trainer-free and cannot reproduce the terminal state exactly, and the corrected re-run (e) does not remove that caveat for the reason just given.

How long the roll-in is, and how much of the ratio the gain explains (post hoc). P16 rolls each policy in for 1000 steps — the training horizon, so evaluation starts from the training-terminal state that the headline proxy measures. Rerunning the released script at a 200-step roll-in (`p16_rollin200_{H1_dm,Go2}.json`) shows how much depends on that choice. The ratio itself is far smaller early: $3.8\times$ on H1_dm and $2.3\times$ on Go2, against $101\times$ and $68\times$ at 1000 steps, because most of the terminal-state ratio comes from both arms *parking* over the roll-in, $-$curric most deeply. What gain-matching removes then diverges by body. On Go2 it removes the whole advantage at 200 steps ($2.3\times \rightarrow 0.78\times$, the curriculum arm now moving less) as at 1000 (93%). On H1_dm it removes only 21% at 200 steps ($3.8\times \rightarrow 3.0\times$) against 74% at 1000. So the gain attribution is robust to roll-in length on Go2 and depends on it on H1_dm, where a modest early advantage survives gain-matching. None of this revives a transfer benefit (§3.1 measures competence independently of the proxy), but it does narrow C1: on H1_dm the headline's *size* is mostly parking and gain, not its existence. Two reruns of the released 1000-step configura-

tion, one from inside the extracted supplementary bundle, gave $105\times \rightarrow 26\times$ and $99\times \rightarrow 22\times$ (75% and 78% removed), within GPU non-determinism of the released $101\times \rightarrow 26\times$ (74%) (`scripts/reproduce_c1_minimal.sh`, which ships the ten H1_dm checkpoints needed).

Both policies at both gains, from common states, with common noise (post hoc, `p16_gain_surface.py`). P16 compares each arm only at its own training gain, after its own roll-in, and its two branches do not share the actor's sampling noise. We re-ran it as a full gain $\times$ policy surface: both arms at both gains (0.4375, 1.0), every branch drawing the same pre-generated action noise (the explicit-noise sampler matches the actor's own to machine precision), same 5 seeds and 100-step evaluation. *(A) Own roll-in*, as in P16, reproduces it ($96\times \rightarrow 23\times$ on H1_dm, $80\times \rightarrow 5.4\times$ on Go2) and shows it is an interaction: gain 1.0 raises Full's motion and *lowers* $-$curric's, because each policy parks only at the gain it was trained with (interaction $J = 3.5\,[1.9, 5.2]$ on H1_dm, $13.0\,[11.7, 14.7]$ on Go2, $5/5$ seeds). At the matched gain 0.4375 there is no Full edge on H1_dm ($0.21\,[-0.70, 1.47]$) and a reversal on Go2 ($-$curric moves $16\times$ more). *(B) A common state bank* that does not depend on the evaluated policy — our primary one mixes states parked by held-out Full and $-$curric policies; fresh resets and a zero-control roll-in are sensitivity banks — removes the Full advantage at either matched gain: Go2 $1.01\times$ at gain 1.0 and $1.00\times$ at 0.4375 (all three banks 0.97–$1.01\times$); H1_dm $1.05\times$ $[0.96, 1.14]$ and $0.99\times$ on the primary bank, with a small edge at gain 1.0 only under fresh and zero-control starts (1.11–$1.14\times$, $4/5$ seeds, sign-flip $p \geq 0.125$). Actions are saturated in every cell (0.93–1.00) and mean $|\text{ctrl}|$ is set by the gain alone, identical across arms. With $n{=}5$ the smallest attainable sign-flip p is 0.0625 and percentile CIs are anti-conservative. This is the cleanest form of C1 we have: the headline proxy advantage is a gain $\times$ own-roll-in state interaction, and from policy-independent states at matched control authority it is absent on Go2 and at most a non-significant 5–14% on H1_dm.

Corrected-trainer per-body uncertainty (P19). Train-gain (own-gain) Full-vs-$-$curric, bootstrap g CI95 and Holm p: H1_dm $g = -1.01$ $[-1.83, -0.40]$, $p_{\text{Holm}} = 0.005$; G1 -5.78 $[-8.88, -5.11]$, $p_{\text{Holm}} = 1.3{\times}10^{-6}$; Panda -1.89 $[-2.93, -1.26]$, $p_{\text{Holm}} < 10^{-3}$; Go2 -0.43 $[-1.18, +0.17]$, $p_{\text{Holm}} = 0.17$. (These intervals changed in the second decimal between drafts: the aggregator iterated over files in filesystem order, which the seeded bootstrap then depended on; it now sorts its inputs, and the released bundle's verifier reproduces the values above exactly.) For the Go2 null a Welch-TOST at the paper's ± 0.3 SD margin does *not* certify equivalence ($p_{\text{TOST}} = 0.67$ — the deficit side cannot be ruled out, the point estimate being negative), but the one-sided upper test *does* bound any curriculum improvement below 0.3 SD ($p = 0.012$) — exactly the "absence of improvement" scope the main text claims, and nothing more.

(e) Corrected-trainer detail (full detail from §3(e)). We rebuilt the trainer (`m15_asc_corrected.py`) with six fixes — per-env intrinsic reward (the base broadcasts one scalar to all 512 envs), corrected SAC entropy sign, TD bootstrap, a per-joint `ctrlrange` *clamp* that replaces the uniform action gain (a command inside the clamped interval is unchanged, so re-evaluating at a different ROM cannot silently amplify a policy — the property the fix needed), dual-gain evaluation, and a collapse-proof smoke gate. We then re-ran the full primary grid (4 bodies $\times$ 20 seeds, the p2 seeds; P19, `p19_corrected_results.md`). Re-evaluating the same policies at full ROM — the old eval — flips two bodies back to positive ($g = +1.8/+1.9$); the discrepancy lives *entirely* in the full-curriculum arm ($-$curric, trained at full range throughout, is gain-invariant to within 0.1%), whose motion is amplified $+71\%$ (H1_dm) by the full-ROM eval. (G1 $-$curric completed 10/20 seeds — the highest-DoF body diverged under full-ROM control from step 0. Per seed (P19b), 9/10 diverged seeds' last *finite* $|q_{\text{vel}}|$ met or exceeded the completed-seed mean of 65.8 (the remaining one at 62.8), but a last-finite reading is not a fixed-time measurement and a standardized g also depends on the variance, so we do not claim a direction for the excluded runs: the G1 contrast is conditional on survival, and the divergence rate ($10/20$) is reported as an outcome in its own right.)

H3 pooled-Spearman table and Simpson scatter (from §4).

Positive-control transfer magnitudes (full detail from §3.1). The self-supervised dynamics control transfers *near-ceiling*: frozen-probe $R^2 = 0.95$–0.98 (vs ASC's 0.33–0.78 and random's 0.52–

Diagnostic	Pooled Spearman ρ	Pooled p
D1	$+0.258$	0.0000
D3	-0.547	$< 10^{-10}$
D4	$+0.522$	$< 10^{-10}$

Table 11: Pooled cross-body Spearman correlations for H3 diagnostics.

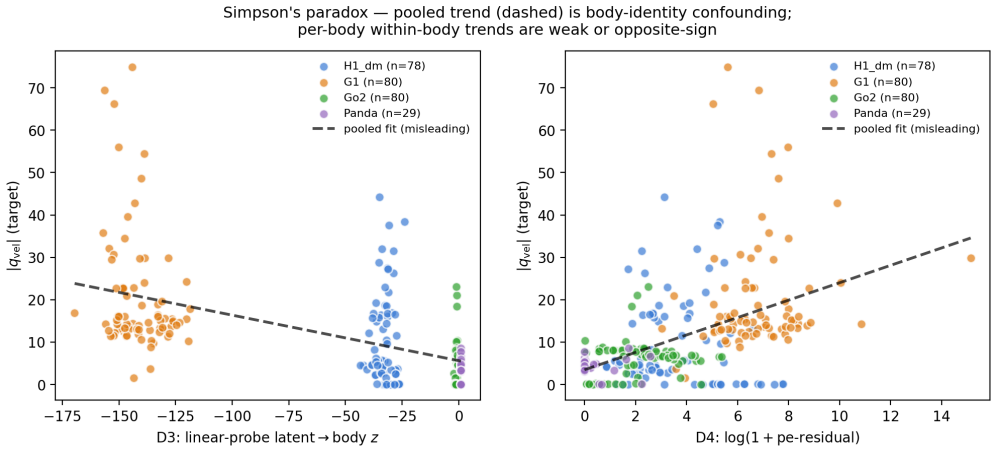

Figure 8: Simpson's paradox in the H3 diagnostics. Diagnostic values (D3 left, D4 right) vs target $|q_{\text{vel}}|$, colored per body. The pooled linear fit (dashed black) looks moderately strong, but per-body trends are weak or opposite-sign. The pooled signal is body-identity confounding.

0.87; dyn-vs-random Hedges $g = +8$ to $+23$) and it is the *best* source on Panda-reach RL transfer (dyn-vs-full $g = +1.25$ frozen / $+0.72$ finetune, n=10).

C2 external MiniGrid dose-response (full detail from §3.2). *Architecture note, stated first.* Both public modules are attached to a shared learned encoder and read its feature h. For RND this is a departure from the canonical method: the target network is frozen, but it takes h rather than the raw observation, so its mapping on observations drifts as h trains, whereas canonical RND fixes random features of the observation itself. The predictor loss back-propagates into the encoder (`p20_c2_external.py`, `RNDModule.compute`). What P20 tests is therefore the shared-encoder variant of each method — the configuration in which reward-weight ablation can mislead — not the methods as released; we do not claim the canonical fixed-target RND would show the same gap. Holding the reward coefficient at zero and toggling only whether the auxiliary loss trains the shared encoder changes a frozen-encoder probe of task-relevant state: on high-dimensional MiniGrid (147-d observations, where the encoder has representational headroom) ICM shows a robust gap (Hedges $g = -1.78$, $p = 0.001$, 8 seeds) and RND a clean *dose-response* (the gap grows monotonically with auxiliary weight: $p =$ ns $\to 0.012 \to 0.0004$ at $1\times$ / $5\times$ / $20\times$). Two honest bounds: the effect is modest ($\sim$5–10% of probe R^2) and it *vanishes* on low-dimensional MountainCar (2-d observations), where a near-random encoder already saturates the probe and the auxiliary loss has no room to matter.

C3 estimator simulation (full detail from §4). On synthetic strata that differ in level, the pooled probability of improvement attenuates toward $\frac{1}{2}$ (median attenuation 0.67 at 8 SD of between-stratum separation) and a pooled correlation acquires a sign no stratum exhibits (up to $+0.75$ from purely between-stratum spread), with sign reversal of the probability of improvement essentially never occurring (`p17_pooling_simulation_results.md`, 2000-trial simulation). The Atari attenuation (85%) is even larger than the 0.63 in our own probe.

Measurement checklist (full table from §5.1).

Default choice (pitfall)	Uncorrected, you conclude	Diagnostic that catches it	Cost
Read an effect at one fixed evaluation setting (gain, reset, episode length)	"the arm learned more"	*instrument audit*: re-evaluate every arm at its own training gain **and** at the others'; the effect should survive the cross, and a saturated (bang-bang) controller and a competent one pushed through the same evaluation should not reproduce it. Warning sign (heuristic, scale-dependent): most of the effect disappears under gain-matching — ours lost 74–93% of the raw ratio but 29–63% of the log-ratio. Gain-invariance is not required of a competent policy when control authority genuinely differs; the check is that the *comparison* does not change with it	re-evaluate existing checkpoints; no training
Read a training-time proxy ($\lvert q_{\text{vel}}\rvert$, coverage, rank) as competence	"the method works"	evaluate a *downstream task* — zero-shot **and** frozen-representation probe **and** RL fine-tuning	zero-shot and probe: 1 eval per checkpoint; fine-tuning: a short downstream training run per checkpoint
Ablate a component by zeroing its *reward weight*	"the component is inert"	*true component gating*: drop the auxiliary loss **and** its parameters	re-run the ablation arm
Validate a diagnostic on *pooled* cross-body data	"the diagnostic predicts competence"	*within-body / LOBO* decomposition + a Simpson's-paradox audit	re-fit per body
Report one body / one seed / one metric	"a large, robust effect"	multi-body grid, Holm-locked family, bootstrap CIs, TOST for nulls	$N_{\text{body}} \times N_{\text{seed}}$ runs

Table 12: Measurement checklist for multi-body developmental RL: recommendations, not a validated standard. Each row is a dissociation observed in one audit (§5.1); the first row rests on a direct intervention (C1), the third and fourth on weaker designs (C2 across evaluation regimes; C3 descriptively on external benchmarks), and the rules state warning signs rather than calibrated thresholds. The first row is the check that would have caught C1 before the others were run.

Internal-consistency observation on the diagnostic (relocated from §3.1). Pairing, per method, the frozen-probe R^2 (cheap, seconds) with the frozen-transfer outcome (costly, hours), and extending the four in-house sources with ICM (Pathak et al., 2017) and RND (Burda et al., 2019) on the *corrected* backbone, the two agree across all four embodiments. On the two bodies with the full method set the Spearman is $\rho = +0.89$ on Panda-reach over six methods ($p = 0.035$) and $+1.00$ on H1_dm-locomotion over five ($p = 0.017$); the two remaining bodies, run as four-method locomotion cells ($\{$full, $-$curric, random, dynamics-control$\}$), give $+0.80$ on G1 (correct top-1, the dynamics control) and $+0.40$ on Go2 (a miss — the collapse-prone quadruped's locomotion transfer sits near floor for every method, leaving little signal to predict). The mean frozen Spearman is $+0.77$ (positive on all four bodies) with correct top-1 selection in 3/4; a combined test that convolves the independent per-cell random-guess nulls — respecting non-exchangeability and never pooling across bodies (C3) — gives $p = 0.033$. We still do not claim this as a contribution: the agreement is partly by construction (probe and frozen-transfer both read the frozen representation), the non-trivial probe→*fine-tune* version fails (H1_dm $\rho = -0.5$), the corrected ICM/RND fall below random ($R^2 = -0.15/-0.50$) so the six-method significance leans on them, the per-cell permutation p has a $2/4! = 0.083$ floor for the four-method cells, and the combined p is fragile — it is carried by the two individually-significant cells and would lose significance if any one correct cell flipped. We record it as evidence the cheap diagnostic is coherent with its costly target across embodiments, nothing more (`p21_selection_results.md`).

Corrected-proxy positive-control battery (P24/P24b/P24c), per body. On H1_dm a scripted mover registers $2.7\times$ a motionless and $2.1\times$ a parked policy; on G1 a mover registers $2.6\times$ a motionless one; on Go2 three independently-trained SAC walkers register 1.4–$1.6\times$ a motionless controller, each with a monotone amplitude dose-response and the proxy strictly monotone in walker competence (deterministic fwd-vel $0.073/0.139/0.166 \rightarrow$ proxy $26.7/28.2/31.4$). The saturated "parked" foil tops the G1 mover; on Go2 the two stronger walkers clear it (walker$-$parked $+4.0$ and $+0.8$, rollout-level CI95 $[3.9, 4.1]$ / $[0.5, 1.1]$; P24d, n=64 env rollouts, measurement noise not training seeds) while the weakest sits slightly below it (-0.7, $[-0.9, -0.5]$) — the same gain/saturation artefact §3(d) diagnoses. All mover-vs-still and inversion contrasts above likewise carry rollout CIs excluding zero (P24d). On the fixed-base Panda the battery *fails*: a saturated arm registers $2.8\times$ a genuine mover and the dose-response inverts, so the $|q_{\text{vel}}|$ proxy does not track motion content there at all — a per-body instrument-validity fact that pooling across bodies would erase, and that only a per-body control reveals (checklist row 1). Full tables: `p24{,b,c}_*_results.md`.

$10\times$ **budget-scaled control (P27).** Scaling every audited objective's pretraining by a common $10\times$ factor (curriculum stretched proportionally; dyn/ICM/RND likewise — a uniform scaling, so inter-objective compute *ratios* are unchanged, not equalized) and re-running the original frozen transfer battery: H1_dm loco dyn 0.574, random 0.566, full 0.211, $-$curric 0.074, RND, ICM; G1 loco dyn $0.752 >$ random $0.529 >$ full $0.309 >$ $-$curric $0.265 >$ RND $>$ ICM; Panda reach (n=3, reduced arm) random -0.491, dyn -0.514, full -0.584. The preregistered P1, as locked (a conjunction: dyn top *and* intrinsic-motivation $\leq$ random on every route), **fails as a conjunction**: dyn-vs-random is a tie on H1_dm ($g = +0.18$ $[-1.06, +1.77]$) and on Panda ($\Delta = 0.023$, n=3), holding inferentially only on G1 ($g = +2.08$ $[+0.96, +7.80]$). Its second clause — **no intrinsic-motivation source overtakes random init** — holds on all 9 reward-free-vs-random comparisons across the three routes ($4 + 4 + 1$), and the pre-committed reversal clause (withdraw the abstract's claim if any reward-free source overtook random/dyn; "reward-free source" in the locked text means the intrinsic-motivation arms, the dynamics control being itself reward-free) did not fire. P2: no curriculum benefit detectable at the control's MDE ≈ 1.8 ($g_{\text{full}-\text{nc}} = +0.63$ $[-0.44, +2.65]$ on H1_dm at 5-vs-5 seeds, $+0.13$ $[-0.73, +3.43]$ on G1 at 5-vs-3 — not detected, and the intervals still admit large positive effects, so this is absence of evidence at this n, not evidence of a small effect; `p27_effect_cis.json`). Two disclosures about the grader behind that verdict: its second disjunct is implemented as $g < 0$ rather than the docstring's "CI spans 0" and computes no interval — it never fired, since both routes pass on $|g| < 0.8$, and the intervals quoted here are computed separately and released — and a missing arm would be graded as a failure rather than as missing data, which does not arise in our runs but would on a partial checkout. Honest scope: (i) G1 $-$curric at $10\times$ diverged on first attempt on 4/5 seeds (NaN by step 250 under full-ROM control; two survived a retry, two deterministically $\rightarrow$ n=3) — a survivorship bias whose direction favors the P2 null, disclosed as such, and evidence that budget *worsens* the no-curriculum arm's instability; (ii) optimization-health diagnostics (Table 13) show degraded late-training health at $10\times$ — critic losses one to two orders higher (H1_dm RND $3.1 \rightarrow 85$, G1 RND $\rightarrow 421$), ICM's r_{int} clip-saturation rising ($45\% \rightarrow 82\%$ H1_dm) — and re-applying the preregistered P15 validity checks at $10\times$: **both RND arms pass all three checks** (predictor MSE falls, e.g. $0.0015 \rightarrow 0.0006$ H1_dm), while **both ICM arms fail the module-learning check** (forward loss rises late, $0.17 \rightarrow 53$ H1_dm; finite and policy-motion checks pass on all four; no old-style $300\times$ r_{int} collapse; the ICM policy-motion pass is a change-magnitude pass, at $-96\%/-93\%$ motion). The same late ICM fwd-loss rise is present at $1\times$ — a non-stationarity property under an evolving policy, not a $10\times$-specific breakage — and its direction is disclosed: degraded ICM health makes the negative result *easier*, so the ICM $10\times$ cell is not evidence that budget cannot help ICM. The budget-robustness claim is therefore firmest for the ASC arms, random/dyn *and RND*, with ICM $10\times$ as supporting evidence only; (iii) n=5/3 per cell (MDE $\approx$ 1.8/2.9), one trainer family (`p27_scaled_control_results.md`).

URLB pooling audit, inferential detail (P25/P26). Propagating URLB's published seed noise (mean$\pm$SEM, 10 seeds; parametric bootstrap, B=20k; each cell drawn independently as $\mathcal{N}(\text{mean}, \text{SEM}^2)$, because the covariance that shared seeds induce across a domain's tasks cannot be recovered from published mean$\pm$SEM — every winner frequency below is therefore conditional on that independence, and is a bootstrap frequency, not a posterior probability of being best) through the 2×10^6-frame states table: the raw-mean pooled winner is RND in 54% of draws vs Disagreement 46%; the within-task mean-rank winner is Disagreement in 69%; the two estimators disagree

Table 13: P27b optimization-health diagnostics: median-over-seeds of first-10% → last-10% critic loss, r_{int} clip-saturation fraction, and intrinsic-module loss (ICM forward loss / RND predictor MSE), at $1\times$ vs $10\times$ pretraining (n=20/8 seeds at $1\times$, 5/3 at $10\times$; smoke runs excluded). P15 checks at $10\times$ apply to the four ICM/RND arms: finite ✓ and policy-motion ✓ on all four (the ICM passes are change-magnitude passes at -96%/-93% motion); module-learning ✓ on both RND arms, × on both ICM arms (also not met at $1\times$ over the full run). G1 −curric is non-finite at both budgets (divergence, §(e); P19b).

cell	budget	critic (first→last)	r_{int} clip%	module loss (first→last)	P15@10×
H1_dm full	1× / 10×	1.5 → 1.5 / 0.2 → 7.8	30 / 39	—	n/a
H1_dm −curric	1× / 10×	1.2 → 0.7 / 0.5 → 3.6	93 / 73	—	n/a
H1_dm ICM	1× / 10×	2.9 → 3.2 / 1.0 → 24.8	45 / 82	0.06 → 15 / 0.17 → 53	fail (c2)
H1_dm RND	1× / 10×	0.0 → 3.1 / 0.0 → 85	0 / 0	.013 → .001 / .0015 → .0006	**pass**
G1 full	1× / 10×	0.3 → 2.9 / 0.7 → 14.0	70 / 77	—	n/a
G1 −curric	1× / 10×	non-finite (diverges)	100 / 100	—	n/a
G1 ICM	10× only	1.8 → 108.8	96	3.4 → 128	fail (c2)
G1 RND	10× only	0.3 → 421.0	0	.0013 → .0009	**pass**
Panda full	1× / 10×	2.1 → 0.9 / 0.7 → 1.4	0 / 1	—	n/a

in 34% of draws; a single method tops all three domains in only 4.6%. Held-out-domain (LODO) Spearman: Jaco $+0.08$ (exact 8!-permutation $p = 0.85$; bootstrap CI $[-0.22, 0.62]$), Quadruped $+0.27$ ($p = 0.52$; $[-0.07, 0.41]$), Walker $+0.40$ ($p = 0.33$; $[-0.04, 0.62]$) — undecidable at $n = 8$ methods, while the pixels panel shows significantly *positive* LODO (Jaco $+0.86$, Walker $+0.85$, both $p = 0.011$; our per-task ranks are ordinal, and recomputing them as average ranks moves exactly one of the six LODO values — pixels Jaco to $+0.88$, $p = 0.007$ — so the figure we quote is the conservative one, and the states panel is unchanged to every digit): the cross-domain question is panel-dependent and unanswerable from the headline table, which is precisely why a single printed ordering should not be read as crowning a method. An earlier point-estimate version of this audit (P25) reported the estimator flip and low LODO values without this uncertainty; P26 supersedes it. A late audit of our own instruments found that P25 carries the same tie-break defect we disclose for P26 and P28: it selects winners with a stable `sorted()`, which under an exact tie returns whichever method is listed first. Re-running the locked P25 analysis with the method list reversed — a probe that changes nothing but which member of a tie is favoured — leaves seven of its eight panels untouched and flips one graded verdict, Q2 (pooling-estimator sensitivity) on the states-5×10^5 panel, where Disagreement and RND tie exactly at mean rank 2.083; the states-2×10^6 panel's Quadruped top-1 likewise turns on a tie at 1.75, without moving any verdict. Neither panel is the one this paper reads and P25's verdicts are superseded, so no claim of ours rests on them — but all three of our own instruments carried the defect, which is the sharpest illustration we can offer of why we treat it as a property of the estimator rather than as a lapse a careful author would have caught. The locked script is unedited; the enumeration is in `p25_pooling_tie_audit.json`. The crowning inference is not hypothetical: e.g., CIC's arXiv abstract ranks itself against "the next leading overall exploration algorithm" by a single cross-domain multiplier ($1.18\times$) aggregated over the same benchmark (Laskin et al., 2022b); the NeurIPS version (Laskin et al., 2022a) states the comparison without that multiplier.

Running the URLB pipeline ourselves (P28). P25/P26 read published numbers; P28 runs the benchmark. We executed the official URLB code (`rll-research/url_benchmark`, commit `bb98f0c`) through its own `pretrain.py`/`finetune.py` entry points with hydra CLI overrides only: states observations, methods {ICM, RND, Disagreement (Pathak et al., 2019)}, all three domains, 3 seeds, 10^6-frame reward-free pretraining, then the full 12-task finetune battery (10^5 frames, 10 eval episodes) from the 10^6 snapshot, plus the repository's DDPG (Lillicrap et al., 2016) from-scratch baseline on all 12 tasks — 27 pretrains and 144 finetunes at the preregistered budget, 0 failures. Including the seed extension below, 45 pretrains of which 44 completed, and 236 finetunes; the single failed pretrain (RND on jaco at seed 4) is why its four downstream tasks are absent, as detailed there. Three compatibility patches were required and are released with the run (`p28-urlb/compat.patch`), none algorithmic: `np.int`→`int` (removed in numpy ≥ 1.24), re-applying the repository's own float32 action-spec cast after `action_scale.Wrapper` (dm_control ≥ 1.0 rebuilds it as float64), and accepting the `update_encoder` keyword in `DDPGAgent.__init__` (an upstream bug that makes the repository's own scratch baseline uninstantiable as released; we note this of the released snapshot only and make no inference about how

the benchmark's published DDPG results were produced). *Reproduction check, and the limits of that check.* Our 12-task profile correlates with the published states-10^6 table at Spearman $+0.916$ (ICM), $+0.930$ (RND), $+0.902$ (Disagreement), $+0.965$ (DDPG). That statistic is weaker than it looks, and the failure is our own C3 failure mode: it is a within-method correlation *across tasks*, so it is dominated by task difficulty, which every method shares. The cross-method matrix (ours in rows, published in columns) shows the diagonal is not distinctive: ICM $0.916/0.615/0.622$, RND $0.755/0.930/0.944$, Disagreement $0.713/0.909/0.902$ — our RND matches the published *Disagreement* (0.944) marginally better than the published RND (0.930). A pooled correlation driven by between-stratum variance cannot certify a within-stratum claim, which is exactly what this paper says elsewhere; we therefore report the matrix rather than the diagonal, and treat the rerun as reproducing the benchmark's *task-difficulty profile*, not as certifying its between-method ordering. On the axis the section's claim lives on, the honest comparison is coarser and less flattering: per-domain winners agree with the published panel on $1/3$ domains (ours walker ICM / quadruped RND / jaco Disagreement; published walker RND / quadruped Disagreement / jaco Disagreement), and the exact three-method ordering reproduces on $6/12$ tasks against a chance rate of $2/12$. Decomposed within domain — the level at which Proposition 1 says such a correlation is interpretable — the fidelity is uneven and lowest exactly where it matters: ICM $1.00/1.00/0.80$, RND $1.00/0.80/0.00$, Disagreement $1.00/0.40/-0.40$ and DDPG $1.00/0.80/0.80$ on walker/quadruped/jaco. Disagreement's within-domain fidelity is -0.40 on jaco, the domain in which our runs crown it. What survives in *both* datasets is P1: no method sweeps all three domains. The "three different winners" pattern is a property of our n=3 runs — the published n=10 panel gives two distinct winners — and we flag it as such rather than presenting it as a property of the benchmark. *Preregistered result* (predictions and an honest-reversal clause locked before any output existed): per-domain seed-means give walker ICM $646.9 >$ RND $639.1 >$ Disagreement 556.6, quadruped RND $664.6 >$ Disagreement $395.5 >$ ICM 133.5, jaco Disagreement $126.2 >$ ICM $115.6 >$ RND 97.5 — each method top-1 exactly once and last exactly once, with a nonparametric bootstrap (B=10k, resampling seeds within task) putting $P(\text{any method tops all three domains}) = 0.0006$. Pooling those same runs: raw-mean crowns RND (467.1 vs Disagreement 359.4, ICM 298.6) with $P(\text{win}) = 1.000$, while within-task mean-rank returns 2.000 for all three methods. The three mean ranks always sum to 6 (per-task ranks are $\{1, 2, 3\}$ and the strata are balanced), so the tie is a single point on a plane the data already lies in, and we should not oversell its exactness. What is structural is the *near*-degeneracy: under cyclic dominance — each method top-1 in one domain, second in another, last in the third — every per-domain rank vector is a permutation of $\{1, 2, 3\}$ and every mean rank is pinned at 2.000. Our data sits essentially on that configuration: RND is exactly cyclic ($2.00/1.00/3.00$), while ICM ($1.25/3.00/1.75$) and Disagreement ($2.75/2.00/1.25$) deviate by ± 0.25 in equal and opposite directions. All three lie within 0.25 of the centroid, which is the informative statement; landing exactly on it is the part that is fragile. Consistent with that, an exact tie recurs in 10.8% of bootstrap draws. We report its uncertainty rather than the point estimate alone: under the same bootstrap the mean-rank winner is RND in 80.5% of draws, ICM in 13.3% and Disagreement in 6.2%, so the honest statement is not "the estimators crown different methods" but that one estimator crowns with certainty (1.000) where the other cannot separate the methods at the point estimate and stays well short of certainty under resampling; the two disagree about the winner in 19.5% of draws. The reversal clause (a domain-uniform winner crowned by both estimators at $P \geq 0.9$, which would have required weakening this claim in the main text) did not fire. *A graded verdict that turns on a tie-break, and a post-hoc re-grade that does not.* The locked grader resolves the rank estimator's winner with `min(mr, key=mr.get)`, which under an exact tie returns the first method in a Python dict rather than reporting no winner — and the observed point estimate *is* an exact tie, so P2's graded first arm ("the winners differ") fired only because `rnd` $\neq$ `icm`. We report this rather than repair it in place: the grader is the preregistered instrument and `git diff` against its lock commit is empty, which we would rather keep than trade for a cosmetic improvement. All graded numbers above are the locked grader's. A clearly post-hoc re-grade (`src/phase2/p28_posthoc_tie.py`, separate output) treats an exact tie as no winner and leaves every preregistered verdict unchanged (P1 true, P2 true, reversal unfired), with rank-estimator $P(\text{win})$ RND 0.763 / ICM 0.067 / Disagreement 0.062 and 23.7% disagreement. The size of the artifact is worth naming: of the locked grader's 13.3% for ICM, 6.6 points are tied draws awarded to it by dictionary order, so its genuine untied share is 6.7%. Two things follow that matter more than the arithmetic. First, an exact three-way tie recurs in 10.8% of bootstrap draws, so it is a recurrent feature of this design, not a measure-zero coincidence. Second, under the re-grade P2 also passes the *strict* reading of its second arm — requiring *both* estimators to crown at $P \geq 0.9$, which the rank estimator's 0.763 fails — so the verdict no longer depends on the tie-break at all. Its

fragility is still worth stating: flipping the closest between-method margin in the grid (0.65%, Disagreement 431.9 vs RND 429.1 on `walker_flip`) removes the tie, after which P2 holds strictly and fails permissively. A preregistered verdict movable by one task at a 0.65% margin is not robust, which is why the main text rests on P1. Against the scratch baseline (pooled 313.8), reward-free pretraining wins 10/12 tasks for RND but 7/12 for ICM and 7/12 for Disagreement — again a per-domain statement. *The seed extension (P28b), and the prediction it cost us.* n=3 was this study's most obvious soft spot, so after seeing the n=3 result we extended to n=5 under a protocol locked first: seeds 4 and 5 and no further extension, the locked grader untouched, n=3 remaining the preregistered primary, both outcomes reported. **P1 survived and P2 failed.** At n=5 no method still tops all three domains (P(sweep) = 0.0409), but the two poolings now agree: both crown RND, the rank estimator at P(win) = 0.946 under the locked grader — above the 0.9 bar the protocol set for "statistically crownable"; 0.844 under the post-hoc cluster, tie-aware estimand below — the exact tie is gone (1.583 RND / 2.083 / 2.333), and estimator disagreement falls from 19.5% to 5.4%. Walker's winner flipped ICM → RND on a 1.4% margin (627 vs 636), which the amendment had named in advance as the vulnerable number (63.0% bootstrap stability at n=3). Per the locked rule we therefore claim estimator-dependence from the published tables (8 methods, 10 seeds, 34% disagreement) and from our n=3 runs, and *not* from our n=5 runs; the "three different winners" phrasing is withdrawn. We also decline the convenient reading that agreement-on-more-seeds itself demonstrates crowning instability: one n=3 → n=5 comparison on three methods cannot separate that from a small-sample artefact, and the artefact reading is the honest one for our own runs. *Divergences:* RND on jaco at seed 4 hit an invalid physics state (`mjWARN_BADQACC`) on both its attempt and its retry. It is the *pretraining* run that failed, so all four jaco tasks are missing for that method-seed and each of those four task cells is n=4 rather than n=5; no other cell is affected. ICM on jaco at the same seed diverged once and completed on retry; seeds 1–3 never diverged. Dropping a diverging RND seed on jaco could only help RND there, and RND does not win jaco at either budget. *A dependency the locked bootstrap ignores, and what restoring it moves (post hoc).* Each (method, domain, seed) pretraining checkpoint is fine-tuned on all four of its domain's tasks, so the four task scores share a checkpoint; the locked grader resamples seeds independently per (method, task) cell and discards that covariance. It is real but modest: the mean pairwise Pearson r among a domain's four task scores across seeds is +0.15 (RND) to +0.47 (Disagreement), strongest for ICM on quadruped (+0.86) and Disagreement on jaco (+0.77), though the from-scratch DDPG control, which shares no checkpoint, shows +0.63 on quadruped, so part of this is task similarity rather than the shared checkpoint. A cluster bootstrap that resamples the (method, domain, seed) block as a unit (`src/phase2/p28_posthoc_cluster.py`, the locked grader imported unmodified, separate output) widens every interval and moves nothing that was graded. At n=3: rank-estimator P(win) for RND 0.805 → 0.751 (tie-aware 0.688, exact ties in 13.2% of draws), estimator disagreement 19.5% → 24.9%, P(sweep) 0.0006 → 0.0028; P1 and P2 both still hold. At n=5: rank P(win) 0.946 → 0.896, disagreement 5.4% → 10.4%, P(sweep) 0.0409 → 0.0771 (P1, a point-estimate predicate, holds); the raw-mean estimator stays at 0.999, so the locked P2 rule (crownable if *either* estimator reaches 0.9) still fails and the withdrawal above stands. The one reading that moves is the strict one — both estimators at 0.9 — which the rank estimator now misses by 0.004 and straddles across RNG seeds (mean 0.898, range 0.892–0.906 over ten seeds, below 0.9 in eight). *Which winner frequency?* The number depends on the tie rule, and the locked grader never defined one: breaking an exact rank tie toward the first-listed method (the locked behaviour) gives 0.896; splitting tied credit would give a value in between; counting a tie as no winner gives RND a sole-winner frequency of 0.844, with ties in 6.4% of draws. We regard the sole-winner frequency under the cluster bootstrap as the correct post-hoc estimand and report it as such, alongside the locked verdicts, which stand as graded. On that estimand the rank estimator does *not* reach 0.9 at n=5, while the raw-mean estimator does (0.999); these are bootstrap frequencies, not probabilities that a method is best. *Whole learning curves, not just the final evaluation (post hoc, `p28_learning_curves.py`).* Each fine-tune logs its evaluation every 10k frames, so the curves can be summarized without new training: area under the reward curve over 0–100k frames, compared within task. Against the from-scratch DDPG baseline, pretraining raises this area with a seed-bootstrap CI excluding 0 on 9/12 tasks for RND, 8/12 for Disagreement and 3/12 for ICM — the adaptation benefit a final-performance comparison can hide, here for the URLB methods rather than for ASC. Domain winners by area are the same as by final return (RND, RND, Disagreement), so no method tops all three domains at the point estimate, but a single method does so in 22% of seed-block draws (against 4–8% on final return): on this summary the domain non-uniformity is weaker, and we report it as such. *Reproducing either verdict.* The released results directory holds all five seeds, so running the locked grader as-is

now reproduces the n=5 verdict, *not* the n=3 primary this section reports — a reader would get numbers that disagree with the paper and reasonably suspect an error. We did not edit the grader to fix that. `src/phase2/p28_reproduce.py` imports it unmodified and applies it to a chosen seed budget (`--seeds 1,2,3` for the preregistered primary), and both verdict files are released for comparison (`p28_urlb_rerun_results_n3_locked.json`, `_n5.json`).

Scope, and the direction small n pushes. 3 seeds as preregistered (5 after the extension above) against URLB's 10 (a control-scale audit, not a leaderboard rerun), states only, the 10^6 snapshot, 3 of 8 methods; LODO was preregistered as ungraded (degenerate at 3 methods). Our locked caveat claimed small n makes a *spurious* crowning easier and so biased the reversal clause toward firing. That is wrong, and wrong in our favour: the clause is a conjunction whose binding term is domain-uniformity, and seed noise makes a uniform sweep *less* likely, so small n biases both graded predictions toward passing. The comparison sometimes offered — a sweep would occur with probability $3(1/3)^3 = 1/9$ if winners were exchangeable noise, against our bootstrap sweep frequency of 0.0006 at n=3 (0.0028 under the cluster bootstrap) and 0.0409 (0.0771) at n=5 — is *not* a test of that null: the bootstrap resamples our observed data rather than simulating the null, and its sweep frequency moves with the seed budget. We therefore state the non-uniformity descriptively — the point estimate has three different domain winners at n=3 and two at n=5 — and do not claim to have excluded a small-sample explanation. The identities are a different matter: per-domain top-1 is stable only on quadruped (RND, $P = 1.000$) and is closer to a coin flip on walker (ICM 0.630 vs RND 0.370) and jaco (Disagreement 0.767), with all three holding jointly in 0.49 of draws. We therefore headline the no-sweep finding, not the winners' names. The raw-mean certainty carries its own n caveat: $P(\text{win}) = 1.000$ rests on RND's worst seed (464.0 pooled) clearing Disagreement's best (459.4) by 4.6 points, so it is certainty about three seeds, not about the benchmark.

What the ROM clamp restricts, per body. Read off the loaded models, the same clamp line means different physical quantities. On G1 (position servos whose `ctrlrange` equals the joint range), Go2 (general actuators configured as PD servos) and Panda (position servos on the seven arm joints; the eighth actuator sets gripper width in metres) it bounds the commanded joint target — an approximate range-of-motion limit. On the DM humanoid (21 torque motors, gear 20–120, all `ctrlrange` $[-1, 1]$) it bounds commanded *torque*, not joint angle, so "range of motion" is a misnomer there; "control-range curriculum" is the accurate name for what the four bodies share. The clamp is also centred on the `ctrlrange` midpoint, not the home pose (e.g. G1 knee $+1.40$ rad, Go2 calf -1.74 rad), so a low mask holds the body near a mid-range posture rather than freezing it at its initial configuration. None of this affects the property the fix was for — a command inside the clamped interval passes unchanged, so a different evaluation ROM cannot amplify it — but it is one more reason the developmental reading is not available. We did not log realized joint-angle ranges per body.

Corrected-trainer objective: a symmetric quirk we disclose. In `m15_asc_corrected.py` the actor, forward predictor and empowerment estimator share one optimizer, and the prediction-error term added to that objective is not detached from the action, so it sends a small gradient to the actor toward *predictable* actions — opposing the prediction-error *reward*. This m5_v4-inherited quirk is applied identically to Full, −curric and every method. Symmetry is not a guarantee of no confounding: the same gradient acts on arm-specific state–action distributions, so the Full-vs-−curric contrast in (e) should be read as conditional on this term being present, not as the contrast under a trainer without it. We disclose it rather than change it, since re-running the full corrected grid to remove the term is not warranted by the contrast-based claim.