Reset Dependence Is Embodied: Training Protocol Changes Which Morphologies Appear Optimal

Kevin Qiu[1] **Zhao Yang**[2] **Marek Cygan**[1]
[1]University of Warsaw [2]Vrije Universiteit Amsterdam
kevinxqiu@gmail.com

Abstract

Robot co-design benchmarks rank bodies after training their controllers under a fixed episode protocol, but that protocol can change the ranking itself. Consequently, two studies can select different bodies from the same corpus even when they use the same task and evaluator. We study one part of that protocol, the reset rule, which ends a training episode when the robot's torso falls below a height threshold. We train PPO controllers for 100 DERL-evolved UNIMAL bodies with this threshold at 20% of each body's starting torso height (Standard), at 50% (Strict), or with no threshold (Reset-free), and score every controller with the same evaluator. The threshold changes which bodies look best: the rankings after Strict and Reset-free training correlate at only 0.21, against about 0.89 if seed noise were the only difference. Removing fall resets costs some bodies far more than others. Bodies evolved on flat ground lose 45% of their per-step reward relative to Standard training, and those evolved on variable terrain only 16%. Permutation tests over bodies support this difference, and on 24 bodies, flat-ground bodies gain the most from Strict over Reset-free training under both PPO and SAC. Choosing bodies by their worst reward across the three protocols costs little: in sample, they keep at least 91% of the reward of each protocol's best bodies, and over 80% averaged across held-out seeds. Co-design benchmarks should state their reset rule, test whether their rankings depend on it, and report bodies that stay strong across protocols. Project page: reset-dependence.github.io.

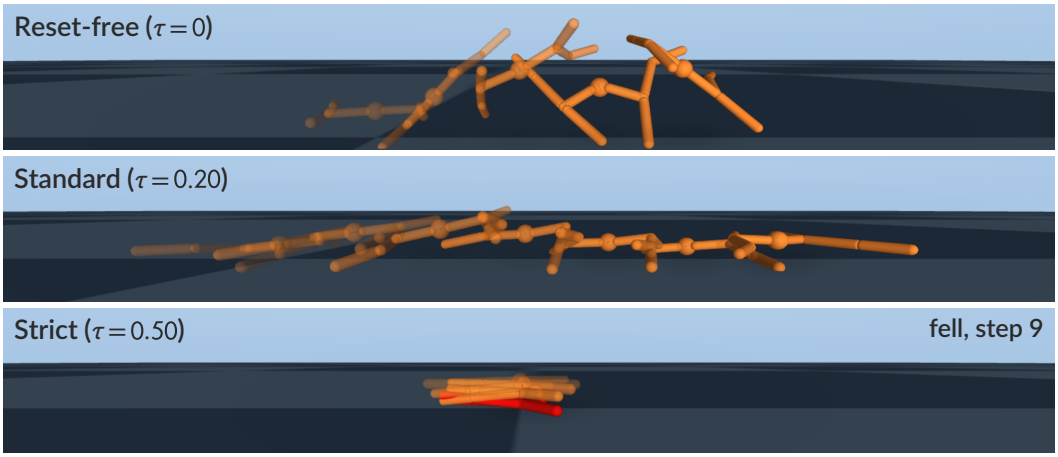


Figure 1: **Reset dependence is embodied.** In illustrative retrains, *Floor A* walks under Standard (τ=0.20), barely moves forward under Reset-free (τ=0), and its Strict (τ=0.50) episode ends at step 9. **Takeaway:** controllers of one body can behave very differently.

40th Conference on Neural Information Processing Systems (NeurIPS 2026). Track on Evaluations and Datasets.

1 Introduction

Robot co-design benchmarks rank bodies after training their controllers under a fixed episode protocol, but that protocol can change the ranking itself. Consequently, two studies can select different bodies from the same corpus even when they use the same task and evaluator. Co-design methods search over bodies, and recent studies also vary the task, but each commits to a single recipe for training the controllers that score the bodies [Gupta et al., 2021b, Lu et al., 2025, Ai et al., 2025]. Such training and evaluation protocols are known to shape the learned controller [Henderson et al., 2018, Pardo et al., 2018, Machado et al., 2018]. We ask whether they also change which *body* appears optimal.

We study one part of the protocol, the *reset rule*. During training, an episode ends when the robot's torso falls below a height threshold, and the robot starts again from its initial pose. Real robots often cannot rely on cheap fall resets [Sharma et al., 2022, Eysenbach et al., 2018], so a body selected under reset-rich training may be optimized for resets that are unavailable at deployment.

Marathon Runner vs. Sprinter

A sprinter and a marathon runner are both elite athletes, but they are not interchangeable. In a 100 m race, a single stumble often decides the outcome, whereas in a marathon, the ability to recover matters. The event, not anatomy alone, determines which body wins. Robot morphology co-design follows the same logic. A reset rule determines whether a fall ends an episode or becomes a state the controller must learn to handle. Changing this rule shifts which morphology appears optimal. Our experiments measure the resulting ranking shift.

We test this on 100 bodies from the UNIMAL design space, evolved by Deep Evolutionary Reinforcement Learning (DERL) [Gupta et al., 2021b] in three families: on flat ground (Floor), on variable terrain (VT), and for pushing a box across variable terrain (MVT). Here every controller is trained from scratch on the same task, walking forward on flat ground. We compare three training protocols that set the reset threshold to a fraction τ of the body's starting torso height: Standard (τ=0.20), Strict (τ=0.50), and Reset-free (τ=0), which never resets after a fall. Each protocol is run with three seeds, and every other episode setting is specified identically. We change the rule only during training and score every controller with the same evaluator. Our contributions are:

1. **The reset rule is part of the benchmark specification.** Changing the reset rule during training reorders the bodies. After Strict and Reset-free training, the rankings of all 100 bodies correlate at Spearman $\rho = 0.21$, against about 0.89 if seed noise were the only difference, and the reordering persists when runs are scored without a cutoff. As a control, retraining 60 bodies with a threshold below the torso's resting height leaves their Reset-free ranking almost unchanged. A morphology ranking is therefore underspecified without the reset rule that produced it.
2. **Reset dependence is embodied.** How much a body loses without fall resets differs across evolutionary families. Under PPO, Floor bodies lose 45% of their per-step reward under Reset-free training relative to Standard training, but VT bodies lose only 16%. The difference holds in permutation tests over bodies and when runs are scored without a cutoff, and on 24 bodies, Floor bodies gain the most from Strict over Reset-free training under both PPO and SAC (Table 1, Figure 3, Sections 4.5 and 4.6).
3. **Selecting for robustness costs little reward.** In sample, bodies chosen by their worst per-step reward across the three protocols keep at least 91% of the reward of each protocol's own best bodies, and over 80% averaged across held-out seeds (Section 4.3, Table 3). Section 5 turns this into a short audit for benchmark authors that also checks whether the reset rule acts on every body.

Every controller is trained from scratch, so the family difference does not come from controllers inherited from evolution. One explanation is what the bodies were selected for. Floor bodies were evolved for speed on predictable surfaces, so once a fall no longer ends the episode, they must act from states that were rare during their evolution. VT bodies were evolved on rough terrain that exposes them to such states. Our experiments are consistent with this account. The audit applies to any co-design benchmark that scores bodies under a reset rule. We release the code, the records of all 900 runs, and a script that recomputes 2,401 values from the released records on a CPU (Section 5).

2 Related Work

Morphology co-design. Methods that co-optimize morphology and control include evolutionary search [Sims, 1994, Gupta et al., 2021b, Bhatia et al., 2021], differentiable design [Chen et al., 2021], learned design policies [Yuan et al., 2022], efficient embodiment generators such as BodyGen [Lu et al., 2025], and LLM-guided morphology design such as RoboMorph [Qiu et al., 2026]. Recent LLM-based co-design also treats the reward itself as a design variable [Qiu and Cygan, 2026]. DERL [Gupta et al., 2021b] introduced the UNIMAL space used here, MetaMorph [Gupta et al., 2022] showed that universal controllers generalize across morphologies, recent work identifies embodiment scaling laws across about 1,000 bodies [Ai et al., 2025], and Yue et al. [2025] target humanoid morphology-control co-design for fall recovery. Strgar and Kriegman [2026] accelerate co-design through morphological pretraining.

These methods vary bodies, controllers, tasks, rewards, or design generators, but they generally fix one termination and evaluation protocol instead of auditing termination as a benchmark variable. We hold the reward fixed and show that the reset rule can change which body ranks best. Mertan and Cheney [2025] offer a complementary critique: simultaneous brain-body co-optimization can undervalue offspring with newly mutated bodies, so the search can miss strong bodies. We show instead that the training protocol used to score bodies can *redefine* which bodies are best.

Termination in RL. Prior work treats termination as part of the task. Pardo et al. [2018] showed that time limits introduce a non-Markov bias and proposed partial-episode bootstrapping. That correction changes how values are bootstrapped at episode boundaries, whereas our intervention changes which fall states the controller sees during training. CaT [Chane-Sane et al., 2024] recasts safety constraints as stochastic terminations for legged locomotion, on a fixed robot platform. Constraint- or bootstrapping-based alternatives may change how sensitive a controller is to termination, but they have not been used to audit morphology rankings under different reset rules. Li et al. [2023] report that offline RL agents develop a survival bias from the data distribution. These works study termination on a fixed task, robot or dataset, whereas in our setting the body matters, because the same threshold leads to qualitatively different outcomes for different bodies.

Reset-free RL and morphology. Sharma et al. [2022] formalized autonomous RL, and later work developed reset-free algorithms for fixed embodiments [Patil et al., 2024, Yang et al., 2025, Gupta et al., 2021a]. Powers et al. [2020] report that morphology shapes how hard a controller is to optimize, and Xiong et al. [2025] study reset-free learning across two swimmer topologies. None of these works trains a population of bodies under several reset rules, which is the question we study.

Evaluation sensitivity. RL results are sensitive to implementation details [Henderson et al., 2018, Agarwal et al., 2021, Engstrom et al., 2020, Andrychowicz et al., 2021, Patterson et al., 2024], and Karwowski et al. [2024] document Goodhart's law in RL reward optimization. Adkins et al. [2024] make hyperparameter sensitivity a first-class axis of evaluation. We carry this axis over to embodied design, where sensitivity to termination changes the ranking of *bodies*.

3 Method

3.1 Morphology Space

We use 100 articulated robots from the UNIMAL (UNIversal aniMAL) morphology space [Gupta et al., 2021b]. DERL evolved them under three selection pressures, which define our three families:

- **Floor** (n=41): evolved on flat terrain to move forward fast on predictable surfaces.
- **Variable terrain (VT)** (n=37): evolved on randomized terrain for robust locomotion under environmental perturbation.
- **Manipulation in variable terrain (MVT)** (n=22): evolved to push a box to a goal across variable terrain.

The bodies have 5 to 15 actuated hinge joints, 5 to 11 rigid parts and a kinematic tree 2 to 5 levels deep below the torso, and they differ in total mass. Every torso is a sphere of radius 0.1 m. Every

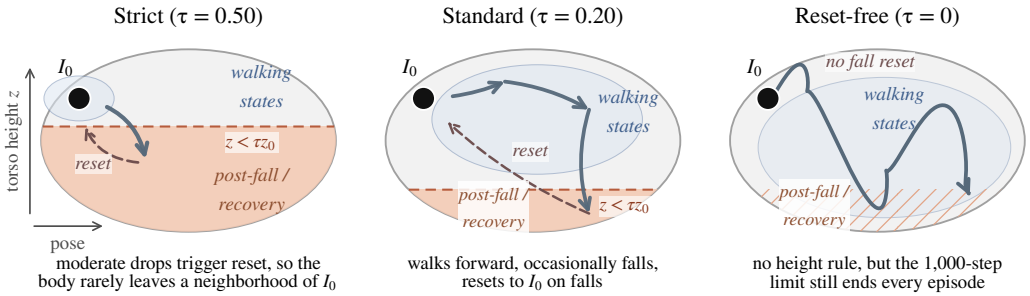


Figure 2: Each panel sketches the states one protocol trains on, with a generic pose dimension on the horizontal axis. I_0 is the start state, and reaching the region below the cutoff $z = \tau z_0$ ends an episode (Eq. 1), except under Reset-free. **Takeaway:** Strict training keeps rollouts near standing states, while Reset-free training exposes bodies to post-fall states the controller must handle.

controller, including those of the MVT bodies, is trained and evaluated on the same task, walking forward on flat ground. As in DERL, the reward is the forward velocity minus a small control cost, $r_t = v_x - 0.001 \cdot \|\mathbf{a}_t\|^2$, with no alive bonus, which makes the reward of a step close to the forward speed in m/s. The controller acts every 0.02 s. It observes the torso height and orientation, the joint angles and all velocities (21 to 41 dimensions, without the global x and y position or the time) and outputs hinge torques in $[-1, 1]$. Each episode starts from the body's spawn pose, with uniform noise in $[-0.01, 0.01]$ on the velocities and on every position except the global x and y. Table 5 (Appendix A.2) lists the full setup, and Figure 10 (Appendix K) shows three sample bodies from each family.

3.2 Training Protocol as an Independent Variable

The reset rule has one parameter, the threshold τ. During training, an episode ends early when the torso falls below the cutoff τz_0, where z_0 is the torso's starting height, and the body then restarts from its spawn pose. Because the reward has no alive bonus, the rule changes no step's reward, only whether an episode continues. Formally, with torso height z_t and ℓ_t steps elapsed in the episode, the episode ends as soon as one of three conditions holds:

$$\text{done}_t \;=\; \mathbf{1}[z_t < z_{\min}(\tau) \;\vee\; z_t > z_{\max} \;\vee\; \ell_t \geq T_{\max}]\,. \tag{1}$$

These are the fall cutoff, an upper height bound, and the time limit. The cutoff is $z_{\min}(\tau) = \tau z_0$ for $\tau > 0$, and Reset-free training has none, $z_{\min}(0) = -\infty$. The spawn pose holds the lowest point of the body 0.15 to 0.25 m above the floor, so every episode starts with a short drop. The upper bound $z_{\max} = \max(3z_0, 2\,\text{m})$ and the time limit $T_{\max} = 1000$ steps (20 s) are the same under every rule, and the time limit ends an episode as a terminal transition, without bootstrapping from the final state (Section 4.8 varies the time limit). Figure 2 sketches what each rule does to the states a controller trains on. The three training protocols are:

- **Standard** ($\tau = 0.20$): the reset rule of the locomotion environment our training code was adapted from (Appendix C). An episode ends only after the torso drops below 20% of its starting height.
- **Strict** ($\tau = 0.50$): DERL's ratio of 0.5 [Gupta et al., 2021b], applied to z_0 rather than to $z_0 - 0.2\,\text{m}$. An episode ends once the torso drops below half its starting height.
- **Reset-free** ($\tau = 0$): without a fall cutoff, the controller keeps acting from whatever state a failed gait leaves it in. Because episodes still end at the time limit and the upper height bound, *Reset-free* names the $\tau = 0$ setting only, not fully autonomous reset-free RL [Sharma et al., 2022].

Existing setups use different thresholds and do not vary them when they rank bodies. DERL [Gupta et al., 2021b] ends an episode when the torso falls below $0.5\,(z_0 - 0.2\,\text{m})$, 0.1 m below our Strict cutoff, and MetaMorph [Gupta et al., 2022] uses the same UNIMAL environment. The default Gymnasium MuJoCo locomotion environments [Towers et al., 2025] use an *absolute* height range,

which corresponds to a fraction of about 0.27 for the Ant (starting near 0.75 m) and 0.56 to 0.71 for Hopper, Walker2d, and Humanoid, all above Standard. No value is privileged, so we vary it.

Per-body cutoff check. For some bodies, a rule does not act as its value suggests (Appendix C.1). A torso resting on the floor sits at about its radius, 0.1 m. For the 28 shortest bodies, the Standard cutoff lies below that, so a resting torso stays above it, and Standard training is in effect Reset-free training. The Strict cutoff has the opposite problem: it can end an episode during the drop from the spawn pose. With zero action, 45 of the 100 bodies are below it 1 s after leaving the spawn pose, so a Strict episode continues only if the controller catches the drop, which most controllers learn to do. For 11 bodies, however, the Strict-trained runs end within 20 steps (0.4 s) on average when evaluated with the Strict cutoff. We call these the cut-off bodies. Ten of them are among the 45, and 7 are among the 28 shortest. We therefore also report the main results on the 72 taller bodies and without the 11 cut-off bodies, as post hoc checks (Appendix L.1).

3.3 Experimental Design

Main sweep. We train 900 PPO [Schulman et al., 2017] controllers, one for each combination of body, protocol and seed ($100 \times 3 \times 3$), each from its own random initialization. Each controller trains for 5×10^6 environment steps in MuJoCo [Todorov et al., 2012], simulated on GPUs with MuJoCo MJX [Google DeepMind, 2024]. All runs are specified with the same PPO hyperparameters (Appendix A.1, Table 4), so no protocol receives its own tuning.

We evaluate each controller on 20 episodes. So that controllers trained under different protocols are compared on equal terms, the main results use one *common evaluator*, which ends episodes at the Standard cutoff ($\tau_{\text{eval}} = 0.20$). Because this evaluator shares its cutoff with Standard training, Section 4.6 checks the main result with a *neutral evaluator* that has no cutoff ($\tau_{\text{eval}} = 0$) but keeps the time limit. A Strict evaluator ($\tau_{\text{eval}} = 0.50$) identifies the cut-off runs of Section 3.2.

Cross-algorithm validation. To check that the contrast between Strict and Reset-free training does not depend on PPO, we also train SAC [Haarnoja et al., 2018], an off-policy algorithm, with Brax's implementation [Freeman et al., 2021] in the same MJX environment. We train it on 24 bodies (6 Floor, 13 VT, 5 MVT) under all three protocols with three seeds and 5×10^6 environment steps each (216 runs). We chose these bodies to cover the range of outcomes under Strict training, and Section 4.5 and Appendix J give the results.

3.4 Statistical Analysis

We score each controller by its *per-step reward* (PSR), the reward per step of each evaluation episode, averaged over its 20 episodes. A body's score under a protocol is the mean over its three seeds. Because this average covers only the steps before an episode ends, Appendix H.7 repeats the family comparison with the episode return divided by $T_{\max}$, which counts the steps after an early end as zero reward, and finds the same difference. A body's *retention* after Reset-free or Strict training is the share of its Standard reward that it keeps (RF/Std or Str/Std), and a family's *loss* is one minus the mean retention of its bodies. Total distance traveled is a secondary metric.

Our main statistical question is whether the protocol affects the three families differently. The three seeds of a body are repeated measurements of that body, so the body is the unit of analysis. The primary test is a permutation test that shuffles family labels across bodies and recomputes the protocol-by-family F statistic, checked by a Wald test with standard errors clustered by body. An ANOVA on all 900 runs, which treats seeds as independent, is only a supplement (Appendix H adds seed-averaged, rank-transformed, mixed-effects and GEE variants). We call a test planned if its decision rule was fixed before its outcome was known.

A rank correlation is hard to judge on its own, because even two runs of one protocol with different seeds rank the bodies differently. We therefore compare each correlation between two protocols with a *seed-noise reference*: the correlation we would expect if seed noise were the only difference between them. We estimate it from how well the single-seed rankings within each protocol agree, adjusted for averaging over three seeds with the Spearman–Brown formula (Appendix L.1). Confidence intervals for correlations and selection results are percentile bootstraps over $B = 10{,}000$ resamples of the bodies. The seed-noise references and the analyses of body subsets are post hoc.

Table 1: Family means of per-step reward over n bodies (seed-averaged, common evaluator). RF/Std and Str/Std average each body's share of its Standard reward, so they are not ratios of the family means. Strict values include the cut-off runs, which lower VT's Str/Std (Section 3.2). **Takeaway:** without fall resets, Floor bodies keep far less of their reward than VT bodies.

Family	n	Standard ↑	Strict ↑	Reset-free ↑	RF/Std ↑	Str/Std ↑
Floor	41	2.428	1.837	1.453	0.55	0.78
VT	37	2.304	1.467	1.999	0.84	0.66
MVT	22	2.651	2.317	2.088	0.72	0.91

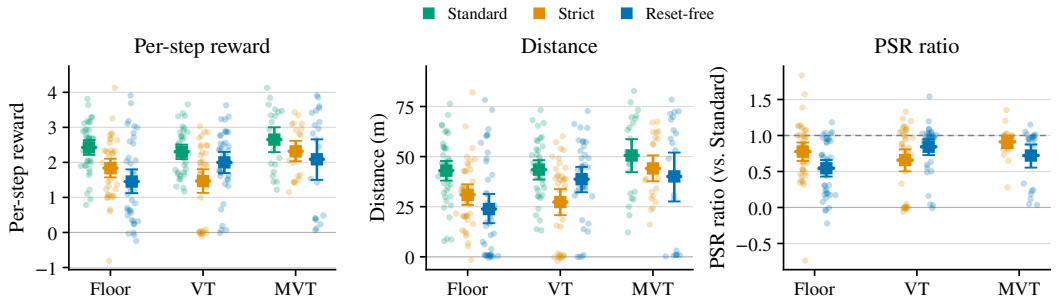


Figure 3: Per-step reward (left), distance (middle), and retention, the reward relative to Standard training (right) for each family. Points are bodies, and large markers with intervals show family means with 95% bootstrap CIs over bodies. Strict values near zero come from cut-off runs (Section 3.2). **Takeaway:** without fall resets, the whole Floor distribution shifts down relative to VT, not just a few bodies.

4 Results

Changing the reset rule during training changes how much each family loses (Section 4.1) and which bodies rank highest (Section 4.2), yet ranking bodies by their worst protocol still keeps most of the reward of each protocol's own best bodies (Section 4.3). Unless stated otherwise, results are per-step reward under the common evaluator ($\tau_{\text{eval}} = 0.20$), and the main results also hold without the bodies whose cutoffs misfire (Section 3.2, Appendix L.1).

4.1 Reset-Free Training Costs Floor Bodies Most

Removing fall resets costs Floor bodies far more than VT bodies (Table 1). After Reset-free training, Floor bodies lose on average 45% of the reward they reach after Standard training, VT bodies 16% and MVT bodies 28%. Every family loses more on the 72 taller bodies, where the Standard cutoff can act, and Floor still loses the most. This is consistent with the reset rule acting through the states a controller trains on: without a cutoff, a fall no longer ends the episode, so the states that follow it stay in training. Distance traveled shows the same pattern (Figure 3, middle, and Appendix M).

4.2 Protocol Choice Changes Which Bodies Appear Optimal

The protocol also changes which bodies rank highest (Figure 4). After Reset-free and Strict training, the rankings of all 100 bodies correlate at $\rho = 0.21$. If seed noise were the only difference between the two protocols, they would agree at about 0.89 (Table 2). The gap persists on the 72 taller bodies ($\rho = 0.26$ against a reference of 0.89) and without the 11 cut-off bodies (0.25 against 0.85). The top-20 lists of Reset-free and Strict training share only 10 bodies, against about 16 in simulated three-seed reruns of one protocol (Appendix L.1). Every pair of protocols agrees less than its seed-noise reference. Measured on single-seed rankings, the excess disagreement, which is zero when changing the rule moves a ranking no more than a new seed does, is 0.53 [0.36, 0.70] for Reset-free versus Strict, 0.29 [0.17, 0.43] for Standard versus Strict and 0.17 [0.08, 0.28] for Reset-free versus Standard (post hoc, Appendix L.1). Standard and Reset-free agree the most ($\rho = 0.72$), partly because their rules coincide for the 28 short bodies. Scored without an evaluation cutoff, the Reset-free and Strict

Table 2: Spearman correlations between the protocols' rankings of the 100 bodies by per-step reward (mean over three seeds, common evaluator), with 95% percentile bootstrap CIs over bodies ($B = 10{,}000$). The seed-noise reference is the correlation expected if seed noise were the only difference between the two protocols (Section 3.4). **Takeaway:** every pair of protocols ranks the bodies less alike than seed noise alone would, most of all Reset-free and Strict.

Comparison	Spearman ρ [95% CI]	Seed-noise reference [95% CI]
Standard vs. Reset-free	0.72 [0.58, 0.83]	0.91 [0.87, 0.94]
Standard vs. Strict	0.46 [0.26, 0.62]	0.87 [0.82, 0.91]
Reset-free vs. Strict	0.21 [−0.02, 0.40]	0.89 [0.84, 0.92]

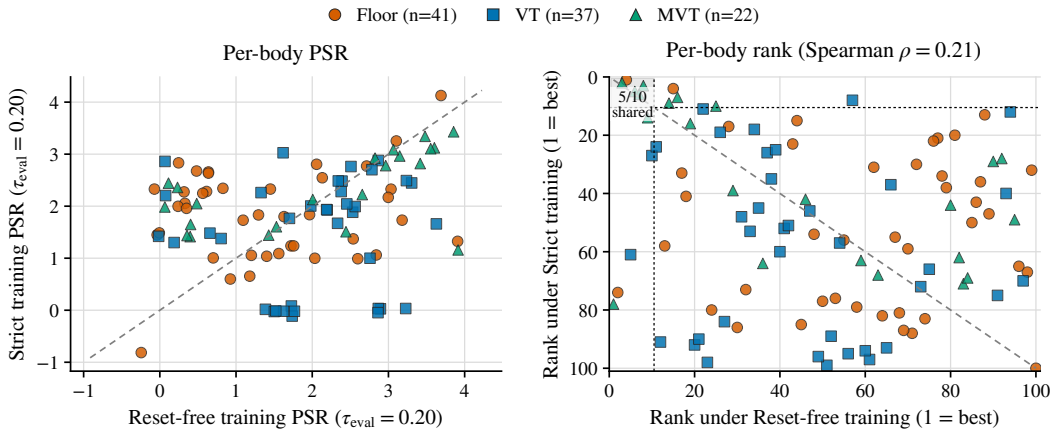


Figure 4: Reset-free vs. Strict training per body (common evaluator). Left: per-step reward. Floor bodies lie mostly above the diagonal, where Strict training does better, and VT bodies mostly below it. Strict values near zero come from cut-off runs (Section 3.2). Right: ranks, with dotted lines at the top-10 cutoff. **Takeaway:** Reset-free and Strict training rank the bodies very differently, even within the Floor and VT families ($\rho = 0.02$ and 0.13).

rankings still correlate at only 0.31. A control agrees far more: on 60 bodies that never fell below a tenth of their starting height in our probes, retraining with a 0.10 threshold ranks them nearly as Reset-free training does ($\rho = 0.92$), while Reset-free and Strict training of the same bodies give $\rho = 0.44$ (post hoc, Appendix L.1). Figure 5 and Appendix Q show single bodies under each protocol.

4.3 Protocol-Robust Selection

When the reset rule is not known in advance, the useful bodies are those that stay strong under every plausible protocol. A simple way to find them is to score each body by its worst reward across the three protocols and keep the top k. In sample, these bodies keep at least 91% of the mean reward of each protocol's own top k, for $k = 5$, 10 and 20 (Table 3), and 83 to 85% averaged across held-out seeds, the most of the five criteria (post hoc, Appendix E). Choosing by one protocol is riskier: the five best bodies after Standard training, once trained under Strict, earn only 49.0% of the reward of the five best Strict-trained bodies (75.7% on the 72 taller bodies). The selector describes performance across the tested protocols, and a planned test with one protocol held out did not support it as a predictor for an untested protocol (Appendix E).

4.4 Statistical Tests of the Family Difference

The permutation test finds that the three families respond differently to the three protocols ($p = 0.004$), and the Wald test with body-clustered standard errors ($p = 0.0014$) and the supplementary ANOVA agree (Appendix H.5). A mixed-effects model with a random intercept per body gives a VT-by-Reset-free interaction of 0.67 [0.39, 0.95] (Table 11). On the Floor and VT bodies alone, the interaction also

Table 3: Worst-case retention by criterion. Each row picks k bodies by one criterion. For each protocol, we divide the mean reward of the picked bodies by that of the protocol's own top k, and each cell reports the smallest of the three ratios (in sample, common evaluator). † marks the selector we use, and Appendix E reports each protocol separately with bootstrap CIs. **Takeaway:** choosing bodies with all three protocols in view avoids the large losses of choosing by one protocol.

Selection criterion	k=5 ↑	k=10 ↑	k=20 ↑
Standard-only	49.0%	74.9%	85.1%
Strict-only	91.1%	87.9%	77.4%
Reset-free-only	67.7%	81.9%	78.6%
Mean over protocols	92.3%	93.9%	89.9%
Min over protocols† (robust)	92.3%	92.8%	91.4%

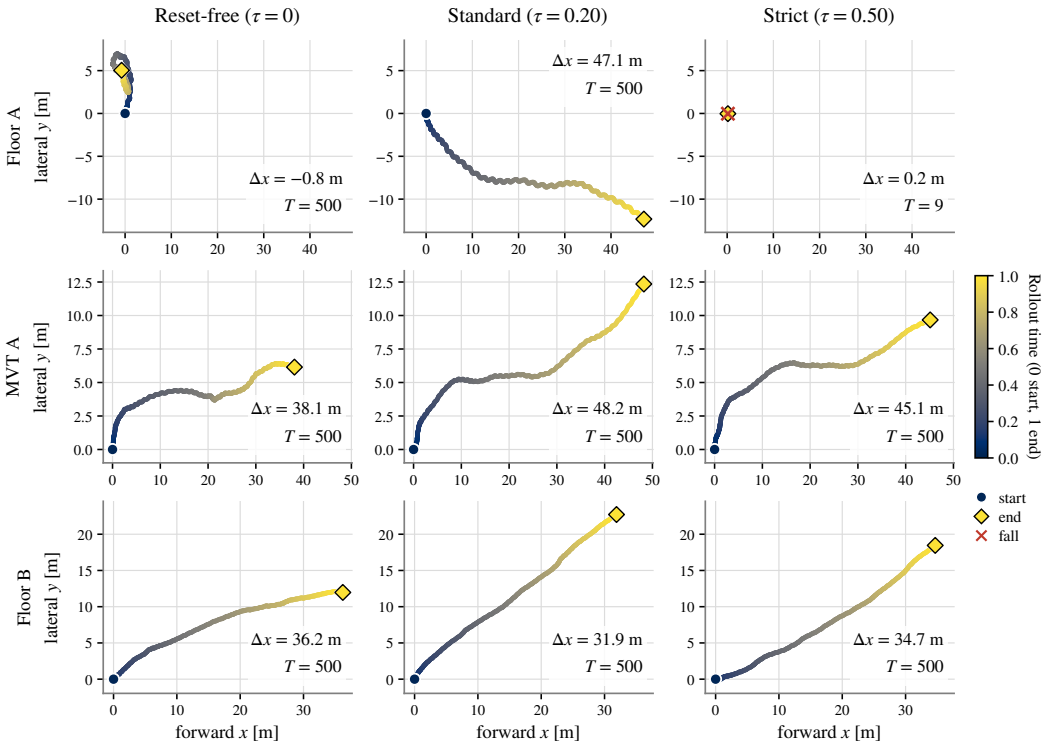


Figure 5: Top-down torso (x, y) trajectories from illustrative retrains of three bodies under three training protocols. Panels without early termination show the first 500 control steps (10 s). *Floor A* walks under Standard and barely moves forward under Reset-free, and its Strict episode ends at step 9. *MVT A* and *Floor B* walk under all three. Full IDs are in Appendix B and chronophotographs in Figures 1 and 13. **Takeaway:** controllers of the same body can differ in their whole path, not only in the distance they cover.

holds in a body-level permutation test ($p < 0.01$, Appendix H.3). Under Strict training, the family difference depends on the 11 cut-off bodies (Table 17). The central contrast, that Floor bodies lose more than VT bodies without fall resets, holds on every body set: the share of Standard reward that Floor bodies keep after Reset-free training lies 0.30 below VT's on all 100 bodies, 0.32 below on the 72 taller bodies and 0.25 below without the 11 cut-off bodies (post hoc permutation $p < 0.01$, Table 17). It also holds when both problems are removed at once, on the 65 of the 72 taller bodies with no cut-off Strict run (Appendix L.1), in four further robustness tests (Table 11) and in all 1,000 subsamples with 22 bodies per family (Appendix H.2).

4.5 Cross-Algorithm Validation with SAC

SAC orders the families as PPO does. The SAC runs were set up to check the contrast between Strict and Reset-free training, so here we use the ratio of a body's Strict to its Reset-free reward, not RF/Std as in Section 4.1. On these 24 bodies, chosen to cover the range of outcomes under Strict training (Section 3.3), both algorithms give Floor bodies the highest mean ratio and MVT bodies the lowest (2.18, 1.27 and 0.91 for Floor, VT and MVT under PPO, and 2.41, 1.52 and 0.96 under SAC, each over its bodies whose Reset-free reward exceeds 0.05). The two algorithms also agree body by body: their ratios correlate at Spearman $\rho = 0.75$ on the 23 bodies where this holds under both, and at 0.72 to 0.83 with any one body left out. SAC thus supports the direction of the family effect (Appendix J).

4.6 Robustness to the Evaluator and a Distance Audit

The family difference also holds when runs are scored without a cutoff. Under the neutral evaluator, Floor bodies still keep the least of their Standard reward after Reset-free training: 64%, against 95% for VT and 85% for MVT (Appendix D). Distance points the same way. Without a cutoff at evaluation, Reset-free-trained Floor controllers cover 31.2 m per episode on average, against 44.5 m after Standard training (Appendix H.8). Appendix M shows that the reward gap reflects slower gaits rather than controllers that stop moving, and Appendix N that control effort is similar across protocols.

4.7 Replication Across Seeds and Body Structure

The family difference replicates in each of the three seeds, with the same ordering of Reset-free retention: Floor bodies keep 53 to 59% of their Standard reward, against 83 to 88% for VT bodies (Appendix H). Here a body's sensitivity is the gap, not the ratio, between its Strict and Reset-free reward. In our corpus, five simple descriptors of body structure, such as starting height and limb count, explain 46% of the variation in this gap, while a body's overall skill, its Standard reward, explains 11%. Taller bodies and bodies with deeper limb trees also keep less of their reward after Reset-free training (Spearman $\rho = -0.53$ and -0.48 with RF/Std), and bodies with more first-level limbs keep more ($\rho = 0.50$). With starting height and limb count controlled, the families still differ in this gap (Appendix I), so these two descriptors do not account for the family difference.

4.8 Time Limit and Other Protocol Variables

Our Reset-free training removes the fall cutoff but keeps the 1,000-step time limit, so our results describe training without fall resets, not fully autonomous training. A planned experiment retrained 30 bodies under Reset-free with no time limit at all. Their ranking correlates with the ordinary Reset-free ranking at $\rho = 0.85$ (95% CI $[0.67, 0.93]$). This meets the plan's criterion for a changed ranking (interval below 0.95), and the Floor and MVT family means drop (Figure 12). These runs, however, used a later version of our training code, which by itself lowers Reset-free Floor scores by a similar amount (post hoc), so the experiment does not isolate the time limit. On the same bodies, a planned factorial test finds that the reset rule and the time limit interact ($p = 0.025$), and Reset-free rankings with limits of 3,001 and 10,001 steps correlate with the ordinary ones at 0.92 and 0.85 (Appendix P). Appendix P.1 applies the same ranking test to other protocol variables in Ant, Evolution Gym and Humanoid bodies.

5 Discussion

Morphology modulates sensitivity to the reset rule. For co-design, a body's score under one reset rule mixes how good the body is with how well it suits that rule, and optimizing it selects for both. Across many bodies, the cost of removing fall resets is a property of the body. Floor bodies lose about three times as much of their reward as VT bodies, which fits the evolutionary account of Section 1. Reset-free RL usually studies one fixed embodiment [Sharma et al., 2022, Yang et al., 2025, Gupta et al., 2021a, Eysenbach et al., 2018], and our results suggest testing reset-free RL methods on several body families, with Floor bodies, which lose the most, as stress tests for recovery from falls (Appendix O). Universal controllers trained across many bodies, such as MetaMorph [Gupta et al., 2022] and URMA [Bohlinger et al., 2025], are a natural next test of whether one policy can absorb this interaction.

A pragmatic audit. A ranking from a single training protocol can be specific to that protocol, so a co-design benchmark should test protocol sensitivity before it names a winning body. The test need not be a full sweep. Retrain the top k bodies, or a subset stratified by family, under one looser and one stricter threshold (120 runs for 20 bodies and three seeds), score every controller with a common evaluator, and report:

Field	Report	This paper
Training reset	Predicate, thresholds, other mechanics	$\tau_{\text{train}} \in \{0, 0.20, 0.50\}$; all else specified identically
Cutoff check	Each body's cutoff vs. torso and reset drop	28 Standard cutoffs below the torso, 11 cut-off bodies
Evaluator	Common reset and metric	$\tau_{\text{eval}} = 0.20$; per-step reward
Audit set	Bodies, seeds, settings	100 bodies $\times$ 3 protocols $\times$ 3 seeds; $k = 5, 10, 20$
Statistics	Rank agreement against seeds, retention, CIs	Spearman ρ vs. seeds; top-10 overlap; top-k retention
Selector	Rule for declaring a best body	Minimum reward across protocols (in sample)

If the top bodies change by more than they do between seeds, the benchmark has found a protocol-specific ranking. The seed-noise reference needs no extra runs, because it comes from the agreement between single-seed rankings within each protocol, and retraining with a threshold below the torso's resting height, as in Section 4.2, gives a control for differences between separate training runs. The per-body cutoff check is worth running on standard setups too. DERL's own rule puts the cutoff of our 25 bodies that start at 0.3 m at half the torso radius, and the lower height bound of Gymnasium's Ant, 0.2 m, lies below its own 0.25 m torso radius, so in both cases a torso resting on the floor stays above the cutoff (Appendix C).

Released artifacts. Everything is in the public repository at `https://github.com/kevinxqiu/reset-dependence`: the training and analysis code, the run summaries of the PPO and SAC UNIMAL experiments and of 5,860 runs on Humanoid, Ant and Evolution Gym bodies, and a seeded CPU script that recomputes 2,401 values from them in a few minutes without GPU training. The code is under the MIT license and the run records under CC BY 4.0, with a data card and Croissant metadata (Appendix B.1, Table 7).

Scope and limitations. This study establishes that changing the fall-termination threshold reorders DERL-evolved UNIMAL morphologies and produces a family-specific interaction in this corpus. The named family ordering is not claimed for other design spaces. More generally, which morphology ranks best can depend on the controller-training protocol used to score it, so that protocol should be reported and reasonable alternatives audited. Whether a particular benchmark is sensitive remains an empirical question. Our experiments use height-based resets, a fixed compute budget, and simulated UNIMAL bodies trained on a single locomotion task. SAC supports the direction of the family ordering (Section 4.5), but loss percentages are PPO estimates, and tuning the learner for each protocol or moving to hardware or other morphology spaces could change them. The family difference also need not hold for other termination rules. In a planned test with a rule ending episodes when the torso touched the ground, a body's sensitivity to that rule correlated only weakly with its sensitivity to the height cutoff ($\rho = 0.19$, below the planned 0.60).

6 Conclusion

Which morphology ranks best can depend on the controller-training protocol used to score it. In the UNIMAL corpus, changing the fall threshold, with the other settings specified identically, changes rank agreement beyond seed variation, the top-ranked bodies, and the response of the three evolutionary families to Reset-free training. The specific family ordering remains scoped to this corpus. Reporting the reset rule, checking that it acts on every body, and using a common evaluator are low-cost specification safeguards that do not assume every benchmark will be sensitive. Benchmark authors should then test whether selected bodies survive reasonable protocol alternatives and distinguish setting-specific from protocol-robust selections. Because real robots often cannot rely on cheap fall resets, co-design methods that select bodies for hardware should treat the training protocol as part of the design problem.

Acknowledgments and Disclosure of Funding

This work was partially supported by the National Science Centre, Poland, under PRELUDIUM Grant No. 2024/53/N/ST6/03370 and Sonata Bis Grant No. 2024/54/E/ST6/00388. We gratefully acknowledge Polish high-performance computing infrastructure PLGrid (HPC Centers: ACK Cyfronet AGH) for providing computer facilities and support within computational grant no. PLG/2026/019509. We thank the Faculty of Mathematics, Informatics and Mechanics of the University of Warsaw for access to its Entropy GPU cluster.

References

Jacob Adkins, Michael Bowling, and Adam White. A method for evaluating hyperparameter sensitivity in reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 37, pages 124820–124842, 2024. doi: 10.52202/079017-3964.

Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron C. Courville, and Marc G. Bellemare. Deep reinforcement learning at the edge of the statistical precipice. In *Advances in Neural Information Processing Systems*, volume 34, pages 29304–29320, 2021.

Bo Ai, Liu Dai, Nico Bohlinger, Dichen Li, Tongzhou Mu, Zhanxin Wu, K. Fay, Henrik I. Christensen, Jan Peters, and Hao Su. Towards embodiment scaling laws in robot locomotion. In *Proceedings of the 9th Conference on Robot Learning*, volume 305 of *Proceedings of Machine Learning Research*, pages 3483–3515, 2025.

Marcin Andrychowicz, Anton Raichuk, Piotr Stańczyk, Manu Orsini, Sertan Girgin, Raphaël Marinier, Léonard Hussenot, Matthieu Geist, Olivier Pietquin, Marcin Michalski, Sylvain Gelly, and Olivier Bachem. What matters for on-policy deep actor-critic methods? A large-scale study. In *International Conference on Learning Representations*, 2021.

Jagdeep Bhatia, Holly Jackson, Yunsheng Tian, Jie Xu, and Wojciech Matusik. Evolution Gym: A large-scale benchmark for evolving soft robots. In *Advances in Neural Information Processing Systems*, volume 34, pages 2201–2214, 2021.

Nico Bohlinger, Grzegorz Czechmanowski, Maciej Piotr Krupka, Piotr Kicki, Krzysztof Walas, Jan Peters, and Davide Tateo. One policy to run them all: an end-to-end learning approach to multi-embodiment locomotion. In *Proceedings of The 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pages 3356–3378, 2025.

Elliot Chane-Sane, Pierre-Alexandre Leziart, Thomas Flayols, Olivier Stasse, Philippe Souères, and Nicolas Mansard. CaT: Constraints as terminations for legged locomotion reinforcement learning. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 13303–13310, 2024. doi: 10.1109/IROS58592.2024.10802334.

Tianjian Chen, Zhanpeng He, and Matei Ciocarlie. Hardware as policy: Mechanical and computational co-optimization using deep reinforcement learning. In *Proceedings of the 2020 Conference on Robot Learning*, volume 155 of *Proceedings of Machine Learning Research*, pages 1158–1173, 2021.

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*, 2020.

Benjamin Eysenbach, Shixiang Gu, Julian Ibarz, and Sergey Levine. Leave no trace: Learning to reset for safe and autonomous reinforcement learning. In *International Conference on Learning Representations*, 2018.

C. Daniel Freeman, Erik Frey, Anton Raichuk, Sertan Girgin, Igor Mordatch, and Olivier Bachem. Brax: A differentiable physics engine for large scale rigid body simulation. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/d1f491a404d6854880943e5c3cd9ca25-Abstract-round1.html.

Google DeepMind. MuJoCo XLA (MJX). https://mujoco.readthedocs.io/en/stable/mjx.html, 2024.

Abhishek Gupta, Justin Yu, Tony Z. Zhao, Vikash Kumar, Aaron Rovinsky, Kelvin Xu, Thomas Devlin, and Sergey Levine. Reset-free reinforcement learning via multi-task learning: Learning dexterous manipulation behaviors without human intervention. In *International Conference on Robotics and Automation*, pages 6664–6671, 2021a. doi: 10.1109/ICRA48506.2021.9561384.

Agrim Gupta, Silvio Savarese, Surya Ganguli, and Li Fei-Fei. Embodied intelligence via learning and evolution. *Nature Communications*, 12(1):5721, 2021b. doi: 10.1038/s41467-021-25874-z.

Agrim Gupta, Linxi Fan, Surya Ganguli, and Li Fei-Fei. MetaMorph: Learning universal controllers with transformers. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=Opmqtk_GvYL.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1861–1870, 2018.

Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. Deep reinforcement learning that matters. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, pages 3207–3214, 2018. doi: 10.1609/aaai.v32i1.11694.

Jacek Karwowski, Oliver Hayman, Xingjian Bai, Klaus Kiendlhofer, Charlie Griffin, and Joar Max Viktor Skalse. Goodhart's law in reinforcement learning. In *International Conference on Learning Representations*, 2024.

Anqi Li, Dipendra Misra, Andrey Kolobov, and Ching-An Cheng. Survival instinct in offline reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36, pages 62062–62120, 2023. doi: 10.52202/075280-2713.

Haofei Lu, Zhe Wu, Junliang Xing, Jianshu Li, Ruoyu Li, Zhe Li, and Yuanchun Shi. BodyGen: Advancing towards efficient embodiment co-design. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=cTR17xl89h.

Marlos C. Machado, Marc G. Bellemare, Erik Talvitie, Joel Veness, Matthew Hausknecht, and Michael Bowling. Revisiting the Arcade Learning Environment: Evaluation protocols and open problems for general agents. *Journal of Artificial Intelligence Research*, 61:523–562, 2018. doi: 10.1613/jair.5699.

Alican Mertan and Nick Cheney. Evolutionary brain-body co-optimization consistently fails to select for morphological potential. In *Proceedings of the Artificial Life Conference 2025 (ALIFE 2025)*, volume 37, page 63. MIT Press, 2025. doi: 10.1162/ISAL.a.891.

Fabio Pardo, Arash Tavakoli, Vitaly Levdik, and Petar Kormushev. Time limits in reinforcement learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4045–4054, 2018.

Darshan Patil, Janarthanan Rajendran, Glen Berseth, and Sarath Chandar. Intelligent switching for reset-free RL. In *International Conference on Learning Representations*, 2024.

Andrew Patterson, Samuel Neumann, Martha White, and Adam White. Empirical design in reinforcement learning. *Journal of Machine Learning Research*, 25(318):1–63, 2024.

Joshua P. Powers, Ryan Grindle, Sam Kriegman, Lapo Frati, Nick Cheney, and Josh C. Bongard. Morphology dictates learnability in neural controllers. In *Artificial Life Conference Proceedings*, volume 32, pages 52–59, 2020. doi: 10.1162/isal_a_00243.

Kevin Qiu and Marek Cygan. Debate2Create: Robot co-design via multi-agent LLM debate. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, pages 101918–101945, 2026. URL https://proceedings.mlr.press/v306/qiu26c.html.

Kevin Qiu, Władysław Pałucki, Krzysztof Ciebiera, Paweł Fijałkowski, Marek Cygan, and Łukasz Kuciński. RoboMorph: Evolving robot morphology using large language models. In *2026 IEEE International Conference on Robotics and Automation (ICRA)*, pages 17611–17618, 2026. doi: 10.1109/ICRA57385.2026.11696020.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347, 2017.

Archit Sharma, Kelvin Xu, Nikhil Sardana, Abhishek Gupta, Karol Hausman, Sergey Levine, and Chelsea Finn. Autonomous reinforcement learning: Formalism and benchmarking. In *International Conference on Learning Representations*, 2022.

Karl Sims. Evolving virtual creatures. In *Proceedings of the 21st Annual Conference on Computer Graphics and Interactive Techniques*, pages 15–22. ACM, 1994. doi: 10.1145/192161.192167.

Luke Strgar and Sam Kriegman. Accelerated co-design of robots through morphological pretraining. In *International Conference on Learning Representations*, 2026.

Emanuel Todorov, Tom Erez, and Yuval Tassa. MuJoCo: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 5026–5033, 2012. doi: 10.1109/IROS.2012.6386109.

Mark Towers, Ariel Kwiatkowski, John U. Balis, Gianluca De Cola, Tristan Deleu, Manuel Goulão, Andreas Kallinteris, Markus Krimmel, Arjun KG, Rodrigo Perez-Vicente, Jordan K. Terry, Andrea Pierré, Sander Schulhoff, Jun Jet Tai, Hannah Tan, and Omar G. Younis. Gymnasium: A standard interface for reinforcement learning environments. In *Advances in Neural Information Processing Systems (Datasets and Benchmarks Track)*, volume 38, pages 163114–163129, 2025. doi: 10.52202/085713-4916.

Tongzhao Xiong, Zhaorong Liu, Yufei Wang, Chong Jin Ong, and Lailai Zhu. Chemotactic navigation in robotic swimmers via reset-free hierarchical reinforcement learning. *Nature Communications*, 16:5441, 2025. doi: 10.1038/s41467-025-60646-z.

Zhao Yang, Thomas M. Moerland, Mike Preuss, Aske Plaat, and Edward S. Hu. Reset-free reinforcement learning with world models. *Transactions on Machine Learning Research*, 2025. URL https://openreview.net/forum?id=ZdMIXltJzK.

Ye Yuan, Yuda Song, Zhengyi Luo, Wen Sun, and Kris M. Kitani. Transform2Act: Learning a transform-and-control policy for efficient agent design. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=UcDUxjPYWSr.

Bo Yue, Sheng Xu, Kui Jia, and Guiliang Liu. Toward humanoid brain-body co-design: Joint optimization of control and morphology for fall recovery. arXiv preprint arXiv:2510.22336, 2025.

Appendix Contents

A Experimental Details **15**
A.1 Training Hyperparameters 15
A.2 Environment Details 15
A.3 Morphology Space 16
A.4 Compute 17
A.5 Data Validation 17
B Detailed Reproducibility Analysis **17**
B.1 Released Artifacts 17
C Protocol Glossary and Per-Body Cutoff Checks **18**
C.1 Per-Body Cutoff Checks 18
D Evaluation-Protocol Robustness **19**
E Protocol-Robust Selection: Full Retention Matrix **20**
F Continuous-Threshold Sweep **21**
G Within-Family Ratio Distributions **21**
H Per-Seed Replication and Alternative Blocked Models **22**
H.1 Robustness of the Protocol-by-Family Interaction 22
H.2 Balanced-Family Robustness 22
H.3 Leave-One-Family-Out Robustness 22
H.4 Per-Seed RF/Std ANOVA 23
H.5 Alternative Blocked Models 23
H.6 Morphology-Level Permutation Test 24
H.7 Metric Sensitivity 24
H.8 Compute-Budget Audit 24
I Morphological Features Track Sensitivity **24**
J Cross-Algorithm Validation with SAC **26**
K Morphology Gallery **27**
L Rank Agreement Across Protocols and Seeds **27**
L.1 Main Results on Each Body Set and the Seed-Noise Reference 27
M Locomotion-Quality Audit **29**
N Control-Effort Confound Audit **30**
O Extended Discussion **30**
P Time Limit and Other Protocol Variables **30**
P.1 Other Design Spaces and Protocol Variables 31
Q Companion Chronophotograph **32**

A Experimental Details

A.1 Training Hyperparameters

All PPO training is specified with identical hyperparameters across thresholds, morphologies, and seeds. The run records store only the rule, seed, and step budget, so the other entries of Table 4 are the values set by the launch code.

Table 4: PPO training hyperparameters for the MJX main sweep.

Parameter	Value
Algorithm	PPO [Schulman et al., 2017]
Implementation	MJX-CustomPPO
Total environment steps	5,000,000
Parallel environments	128
Steps per rollout	16
Minibatch size	256
PPO epochs per update	10
Learning rate	3×10^{-4}
Discount (γ)	0.99
GAE (λ)	0.95
Clip range (ϵ)	0.2
Value loss coefficient	0.5
Entropy coefficient	0.01, applied to half the policy entropy
Maximum gradient norm	0.5
Optimizer	Adam
Advantage normalization	Per minibatch
Policy network	MLP [64, 64]
Value network	MLP [64, 64]
Activation	Tanh
Max episode length ($T_{\max}$)	1,000 steps
Physics timestep	0.005s
Frame skip	4
Effective control dt	0.020s

A.2 Environment Details

Simulator. MuJoCo 3.x [Todorov et al., 2012] with MJX acceleration for training.

Reward function. $r_t = v_x - 0.001 \cdot \|\mathbf{a}_t\|^2$, where $v_x = (x_{t+1} - x_t)/\Delta t$ is the forward velocity and $\mathbf{a}_t \in [-1, 1]^{n_{\text{act}}}$ is the normalized action vector.

Termination. As formalized in Equation 1, the episode ends on a fall ($z_t < \tau z_0$ when $\tau > 0$), on the upper height bound ($z_t > z_{\max} = \max(3z_0, 2\,\text{m})$), or at the shared horizon $T_{\max} = 1000$, which is a terminal transition (not bootstrapped) under every rule. Here z_0 is the torso height in the reset (spawn) pose, which places the lowest point of the body 0.15 to 0.25 m above the floor, so every episode begins with a short drop. At $\tau = 0$ the fall clause is inactive; the shared horizon and upper height bound still apply.

Evaluation. Each trained policy is evaluated for 20 episodes under three evaluation settings: $\tau_{\text{eval}} = 0.20$, $\tau_{\text{eval}} = 0.50$, and neutral evaluation ($\tau_{\text{eval}} = 0$). The main text uses $\tau_{\text{eval}} = 0.20$ unless noted otherwise. For PPO, actions are sampled stochastically (from the Gaussian policy distribution) at evaluation to match the training-time action distribution across all conditions. Rank comparisons use PSR averaged over evaluation episodes and training seeds (Table 5).

Table 5: Training and evaluation setup. Every setting is specified identically for the three fall rules except the threshold τ, and the last column notes a public setup with the same value where one exists.

Setting	Value	Public reference
Bodies	100 UNIMAL bodies (41 Floor, 37 VT, 22 MVT, 97 distinct files), torso a sphere of radius 0.1 m	UNIMAL-100 benchmark (DERL-evolved)
Task, reward	Forward locomotion on flat ground, $r_t = v_x - 0.001 \, \|a_t\|^2$, no alive bonus	DERL also has no alive bonus
Observations, actions	Torso height and orientation, joint angles and all velocities (21 to 41 dimensions, no global x, y or time). Hinge torques in $[-1, 1]$ (5 to 15 dimensions)	
Reset	Spawn pose (torso at reset height z_0, lowest point 0.15 to 0.25 m above the floor), uniform noise in $[-0.01, 0.01]$ on velocities and on positions except global x, y	Spawn heights from DERL files
Episode end	Fall ($z_t < \tau z_0$, none at $\tau = 0$), upper bound $z_t > \max(3z_0, 2\,\text{m})$, or 1,000 steps, a terminal transition (not bootstrapped) under every rule	DERL also uses 1,000 steps but bootstraps
Fall rule	$\tau \in \{0, 0.20, 0.50\}$, cutoffs 0.06 to 0.21 m (Standard) and 0.15 to 0.52 m (Strict), a later sweep adds 0.10 and 0.35	Section 3.2
Training	PPO, one preset for every rule (Appendix A), 5M environment steps, seeds 42 to 44, MuJoCo MJX	5M steps, DERL's per-body budget
Evaluation	20 evaluation episodes under the neutral (no fall clause), common ($\tau_{\text{eval}} = 0.20$) and Strict evaluators, scored by per-step reward (PSR) and fixed-horizon return (episode return divided by 1,000)	Section 3.4

A.3 Morphology Space

We use the 100 UNIMAL morphologies of MetaMorph's UNIMAL-100 training set [Gupta et al., 2022], evolved by DERL [Gupta et al., 2021b], comprising three evolutionary families:

- **Floor** ($n = 41$): Evolved for flat-terrain locomotion.
- **VT** ($n = 37$): Evolved for variable-terrain locomotion (slopes, steps, rugged ground).
- **MVT** ($n = 22$): Evolved for manipulation in variable terrain (pushing a box to a goal).

Morphologies vary in number of rigid bodies (5–11), actuated hinge joints (5–15), total mass, kinematic tree depth (2–5 levels below the torso), and limb geometry. All share a central torso with tree-structured limbs.

Three pairs of the 100 body files are byte-identical, and dropping one body of each pair changes no three-seed rank correlation between rules by more than 0.04.

A.4 Compute

All UNIMAL training ran on a shared GPU cluster, with up to four PPO processes packed per GPU. The corpus records do not store the GPU model, and the panels that recorded it ran on NVIDIA TITAN V (12 GB), GeForce RTX 2080 Ti, TITAN X (Pascal), and GeForce GTX 1080 Ti GPUs. Typical wall time is 1.09 hours per 5M-step PPO run of the corpus, which ran one process per GPU. The table below sums the recorded run time of each released panel's runs (for the UNIMAL panels, training plus evaluation). It equals GPU-hours where each run had its own GPU, and where several runs shared a GPU, GPU-hours are lower: the 108 runs of the retraining with later code took 228 hours of wall time at four runs per GPU, about 57 GPU-hours. The review-period panels ran on NVIDIA GH200 and H100 GPUs and, for Evolution Gym, on CPUs.

Panel	Hardware	Runs	Run-hours
Main corpus (Section 3)	GPU, 1 run per GPU	900	980
Thresholds 0.10 and 0.35 (Appendix F)	GPU, 4 runs per GPU	600	1,006
SAC panel (Appendix J)	GPU, 1 run per GPU	216	105
Retraining with later code	GPU, 4 runs per GPU	108	228
Structural transfer to non-evolved bodies	GPU, up to 4 per GPU	540	1,025
Existing robots	GPU, up to 4 per GPU	60	53
Longer budget	GPU, up to 4 per GPU	162	183
Contact-based termination	GPU, up to 4 per GPU	270	552
No time limit	GPU, up to 4 per GPU	90	180
Rule by time limit	GPU, up to 4 per GPU	180	370
Time-limit durations	GPU, up to 4 per GPU	180	357
Humanoid, fall rules (Appendix P.1)	NVIDIA GH200	1,800	1,427
Humanoid, time limit	NVIDIA GH200	900	725
Ant, reset initialization	NVIDIA H100	160	85
Ant, three other variables	NVIDIA GH200	1,800	not recorded
Evolution Gym, time limit	CPU	1,200	2,155

A.5 Data Validation

Before statistical analysis, we ran an automated validation pass over the PPO run corpus covering file integrity, numerical sanity, seed independence, reset-free semantics, family labels, recorded configuration labels, and cross-checks against main-text summary statistics. The pass checked recorded labels and values, not the settings that were executed, the code version, the hardware, or whether each threshold can act on each body (Appendix C.1).

B Detailed Reproducibility Analysis

Body-alias lookup. Table 6 maps the body aliases of Figures 1, 5 and 13 to their full UNIMAL identifiers.

Table 6: Aliases used in figure captions, and the full UNIMAL morphology identifiers.

Alias	Family	UNIMAL morphology identifier
Floor A	Floor	`floor-5506-10-6-01-15-48-35`
MVT A	MVT	`mvt-5506-13-10-17-12-25-45`
Floor B	Floor	`floor-1409-0-7-01-15-15-26`

B.1 Released Artifacts

The release is one code repository (`https://github.com/kevinxqiu/reset-dependence`), with code under the MIT license and run records under CC BY 4.0 (Table 7). One seeded CPU script recomputes 2,401 values from the run summaries without GPU training, among them nearly every main-text statistic of the UNIMAL panels, and its output lists the values it does not recompute. The code retrains the controllers of the 100 evolved bodies from scratch.

Table 7: Contents of the release.

Item	Status
Run summaries	Every reported UNIMAL panel except the A2C panel, and a panel on four existing legged robots, 3,306 runs in all
Additional panels	Per-run summaries of the 5,860 Humanoid, Ant and Evolution Gym runs with each panel's specification and frozen report. Their raw run records and training code are not released
Probe records	Zero-action rollouts and settle heights of all 100 bodies and the random-action searches of Appendix C.1, with the scripts that made all but one of them
Planned-test records	Specification, report and analysis output of each planned test except the A2C panel
Code	Training and launch code (the corpus-era trainer and the later revision) and CPU scripts that recompute the statistics and figures from the run summaries and probe records
Bodies	Morphology features and a script that downloads the UNIMAL body files from the upstream release and checks their hashes. The files are not redistributed, because the upstream repositories declare no license
Rollouts and logs	The traces behind Figures 1, 5 and 13, and training logs where kept (none for the corpus and SAC runs)
Not available	Policy checkpoints and the per-run code versions of the corpus and SAC runs, which were not saved, and the analyzed A2C runs (Appendix J)

C Protocol Glossary and Per-Body Cutoff Checks

The three protocols in the main text are specified to differ only in the height-based early-termination threshold τ applied during training; all other episode mechanics are specified identically (Table 8).

Table 8: Scope of Standard, Strict, and Reset-free as used in this paper. The column "Early-termination on falls?" refers to the height-based predicate during training; non-threshold episode mechanics are specified identically across protocols.

Name in paper	τ	Early-termination on falls?	Literature anchor
Standard	0.20	Yes, at 20% of reset height	The default Gymnasium MuJoCo [Towers et al., 2025] height range roughly maps to an equivalent fraction of ≈ 0.27 on an Ant with $z_0 \approx 0.75$ m (0.56, 0.64, and 0.71 of the initial torso height on Hopper, Walker2d, and Humanoid), all above 0.20. Our value is the fall rule of the locomotion environment our training code was adapted from.
Strict	0.50	Yes, at 50% of reset height	DERL default [Gupta et al., 2021b] sets a per-body ratio of 0.5, but applies it to its spawn height minus a 0.2 m clearance and bootstraps at its time limit, so our cutoff is 0.1 m higher; MetaMorph [Gupta et al., 2022] uses the same configurable UNIMAL substrate.
Reset-free	0	No early-termination on falls	Fall-based resets are disabled during training, matching the endpoint of our height-threshold sweep.

Throughout the paper, *reset-free* denotes this height-threshold endpoint: the fall predicate is disabled, while the rest of the training setup is specified identically across protocols. Main-text body-level claims are scoped to this height-based termination mechanism. The geometry of Appendix C.1 also appears in standard setups: DERL's own rule puts the cutoff of our 25 shortest bodies at half the torso radius, and Gymnasium Ant's lower bound of 0.2 m lies below its own 0.25 m torso radius, so the per-body check of Section 3.2 is worth running on standard setups too.

C.1 Per-Body Cutoff Checks

The Standard cutoff lies below the torso radius for 28 bodies. Every torso in the corpus is a sphere of radius 0.1 m. For the 28 bodies with $z_0 \leq 0.45$ m (8 Floor, 13 VT, 7 MVT), the Standard

cutoff $0.20\,z_0$ is 0.06 or 0.09 m and lies below the torso radius. In simulation probes with random actions, flipped landings and drops, the torso of the 25 bodies with $z_0 = 0.30$ m never reached the cutoff, and the torso of the three MVT bodies with $z_0 = 0.45$ m went at most 5 mm below it, so Standard training may have ended some of their early episodes, which the corpus did not record. All 15 policies of each of the 28 bodies produce the same record under the common evaluator as under the neutral one, so Standard training is in effect Reset-free training for these bodies, whose Reset-free to Standard ratio averages 0.98. The same geometry limits the threshold sweep: the $\tau = 0.10$ cutoff lies below the torso radius for 94 bodies, and the $\tau = 0.35$ cutoff is reached whenever the torso of one of the 25 shortest bodies rests on the floor.

The Strict cutoff can be crossed during the drop from the reset pose. Because z_0 is a spawn height, the Strict cutoff $0.5\,z_0$ can lie above the height at which a body comes to rest, and with zero action 45 of the 100 bodies are below it after 1 s. A Strict episode on such a body continues only if the controller catches the drop, and most controllers learn to. We call a Strict-trained run cut off if its episodes under the Strict evaluator average at most 20 steps. Cut-off runs are 17 of 123 Floor, 31 of 111 VT and 1 of 66 MVT Strict runs, and they account for most of the Strict loss of Floor and VT bodies (about 15 of 22 and 28 of 34 points, post hoc). For ten VT bodies, every PPO controller we trained averages at most 26 steps under the Strict evaluator, while SAC keeps one of them up. An eleventh body, a Floor body with $z_0 = 0.75$ m, has cut-off Strict-trained runs in every seed, which this geometry does not explain. We call these 11 bodies, whose Strict-trained runs are cut off on average, the cut-off bodies, and the 23 bodies with at least one cut-off seed the cut-off-seed bodies. Both sets are defined by outcomes, so results without them are post hoc sensitivity checks. The numbers record of the release (`records/camera_ready_numbers.json`) lists the affected body sets by identifier, and its probe records hold the probe details (Table 7).

D Evaluation-Protocol Robustness

Selection-retention robustness. The main-text selection analysis (Section 4.3) scores every policy at $\tau_{\text{eval}} = 0.20$, and Table 9 reports the same retention at $\tau_{\text{eval}} \in \{0.00, 0.20, 0.50\}$. The common and neutral rows agree closely, partly because the two evaluators coincide for the 28 bodies whose Standard cutoff lies below the torso radius: top-5 Standard-selected retention against the Strict oracle is 49.0% and 49.3%, with overlapping confidence intervals. The $\tau_{\text{eval}} = 0.50$ rows are diagnostic only, because Standard rewards approach zero for many bodies under that evaluator.

Table 9: Selection retention (target-training oracle retained by a Standard-training selection) across evaluation thresholds. Morphology bootstrap with $B = 10{,}000$ resamples. The $\tau_{\text{eval}} = 0.20$ rows match Section 4.3; the $\tau_{\text{eval}} = 0$ rows are the threshold-free robustness check. The $\tau_{\text{eval}} = 0.50$ rows are diagnostic because Standard denominators approach zero.

τ_{eval}	k	vs. Strict oracle ↑	vs. RF oracle ↑
0.20	5	49.0% [22.8, 93.5]	95.1% [90.9, 100.0]
0.20	10	74.9% [50.1, 93.3]	96.4% [92.3, 100.0]
0.20	20	85.1% [66.9, 93.0]	94.4% [83.9, 100.0]
0.00	5	49.3% [23.0, 93.5]	96.3% [92.5, 100.0]
0.00	10	75.1% [50.0, 92.6]	96.9% [94.3, 100.0]
0.00	20	83.7% [65.6, 92.7]	94.8% [85.8, 99.9]
0.50	5	89.6% [78.3, 99.0]	100.0% [56.2, 100.0]
0.50	10	93.9% [85.8, 98.6]	75.8% [51.1, 100.0]
0.50	20	90.7% [84.1, 96.1]	73.3% [62.7, 89.8]

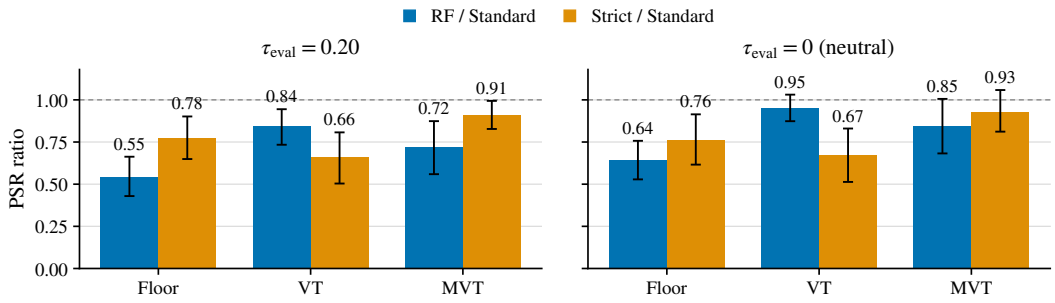


Figure 6: Family-level RF/Std and Str/Std PSR ratios at $\tau_{\text{eval}} = 0.20$ and neutral $\tau_{\text{eval}} = 0$. Bars show family means over per-body ratios, with 95% morphology-bootstrap CIs. The Floor and VT difference in Str/Std comes mostly from runs cut off during the drop from the reset pose (Appendix C.1). **Takeaway:** the Reset-free family ordering persists under neutral evaluation, so the evaluation cutoff does not produce it.

E Protocol-Robust Selection: Full Retention Matrix

Table 10 gives the full criterion-by-oracle-by-k matrix with bootstrap 95% CIs ($B = 10{,}000$ body resamples). Min-over-protocols is the only criterion with point estimates above 90% against every oracle at every k ($92.3/92.8/91.4\%$), and mean-over-protocols is near 90% throughout ($92.3/93.9/89.9\%$). At $k = 5$ the two select the same top-5 set, so their point retentions coincide. All values are in sample, because the same seed-averaged scores choose the bodies and define the oracles. The Standard-only 49.0% against the Strict oracle at $k = 5$ comes mostly from the bodies whose cutoffs do not act as intended: it is 75.7% on the 72 bodies whose Standard cutoff lies above the torso radius.

Table 10: In-sample retention of each single-protocol oracle's top-k PSR under five selection criteria, at common $\tau_{\text{eval}} = 0.20$. Point estimates with bootstrap 95% CIs. Oracle diagonals are 100% by construction. † marks our conservative selector.

Criterion	k	vs. Standard oracle ↑	vs. Strict oracle ↑	vs. Reset-free oracle ↑
Standard-only	5	n/a	49.0% [22.8, 93.5]	95.1% [90.9, 100.0]
	10	n/a	74.9% [50.1, 93.3]	96.4% [92.3, 100.0]
	20	n/a	85.1% [66.9, 93.0]	94.4% [83.9, 100.0]
Strict-only	5	91.1% [81.6, 97.2]	n/a	93.3% [74.6, 99.2]
	10	90.4% [81.3, 97.0]	n/a	87.9% [61.7, 97.7]
	20	90.1% [84.6, 95.5]	n/a	77.4% [60.2, 90.6]
Reset-free-only	5	95.6% [92.5, 100.0]	67.7% [36.5, 97.9]	n/a
	10	98.9% [95.6, 100.0]	81.9% [60.1, 94.2]	n/a
	20	98.9% [96.7, 100.0]	78.6% [60.9, 90.5]	n/a
Mean over protocols	5	92.3% [86.9, 99.0]	99.1% [88.5, 100.0]	95.7% [88.1, 100.0]
	10	96.7% [91.9, 99.9]	93.9% [80.9, 100.0]	95.8% [91.0, 99.5]
	20	97.9% [94.0, 99.7]	89.9% [82.7, 96.9]	98.4% [93.5, 99.8]
Min over protocols†	5	92.3% [84.3, 97.9]	99.1% [98.2, 100.0]	95.7% [85.2, 99.6]
	10	93.5% [85.2, 97.9]	99.4% [96.6, 100.0]	92.8% [85.2, 98.3]
	20	91.4% [85.4, 96.3]	96.7% [92.1, 99.7]	92.1% [84.7, 96.9]

Held-out seed. To gauge how optimistic the in-sample values are, we also choose bodies on the mean of two seeds and score them against the third seed's own top k, averaging over the three held-out seeds (post hoc). Min-over-protocols keeps 82.8%, 84.3% and 85.2% at $k = 5$, 10 and 20, the highest worst-case retention of the five criteria at every k (the next best, the mean over protocols, keeps 82.2%, 78.9% and 81.6%). Of the paired body-bootstrap intervals for its differences from Standard-only, Strict-only and mean-over-protocols, only those against Standard-only at $k = 5$ and 20 and against Strict-only at $k = 20$ exclude zero.

Held-out protocol. A planned test held out each protocol in turn. It required that, at $k = 5$, the best criterion built from the other two protocols beat the better of those two single protocols on the held-out one by at least 0.05 PSR in two of three splits, with an interval above zero, and that some combination keep at least 85% of the held-out top-10 reward in every split. Neither part passed: the gains are $+0.02$, -0.13 and $+0.23$ PSR (mean $+0.04$, $[-0.42, 0.23]$), and with Strict held out the best combination keeps 81.9% of the held-out top-10 reward. The test therefore does not support the selector as a predictor for an untested protocol.

Family composition under each criterion. The top-10 picks under each selection criterion have the following family split (Floor / VT / MVT): Standard-only $2/2/6$; Strict-only $2/1/7$; Reset-free-only $2/2/6$; mean $2/0/8$; min $2/0/8$. The two protocol-robust criteria shift the MVT count from 6-7 to 8 without requiring family labels as input. Five of the eight MVT bodies in the min-over-protocols top 10 are among the 28 bodies whose Standard cutoff lies below the torso radius, and on the other 72 bodies the mean and min criteria each pick 4 MVT bodies.

F Continuous-Threshold Sweep

Figure 7 shows the mean per-step reward (seed-averaged, evaluated at $\tau_{\text{eval}} = 0.20$) as a function of the training termination threshold τ at five measured points ($\tau \in \{0.00, 0.10, 0.20, 0.35, 0.50\}$), per evolutionary family. The discrete main-text comparison at $\tau \in \{0, 0.20, 0.50\}$ samples the same threshold response: Floor peaks sharply at $\tau = 0.20$ (2.43) and drops steeply toward reset-free ($\tau = 0$, 1.45), VT drops steeply toward strict ($\tau = 0.50$, 1.47, or 2.01 without the ten VT bodies whose Strict runs are cut off during the drop from the reset pose), and MVT's curve is the flattest (range $[2.09, 2.65]$ across all five thresholds). The intermediate thresholds do not act as their values suggest, because the $\tau = 0.10$ cutoff lies below the torso radius of 94 bodies and the $\tau = 0.35$ cutoff is reached whenever the torso of one of the 25 shortest bodies rests on the floor (Appendix C.1). We therefore read the sweep for the shape of reward across τ only, not as showing the same family response at every threshold.

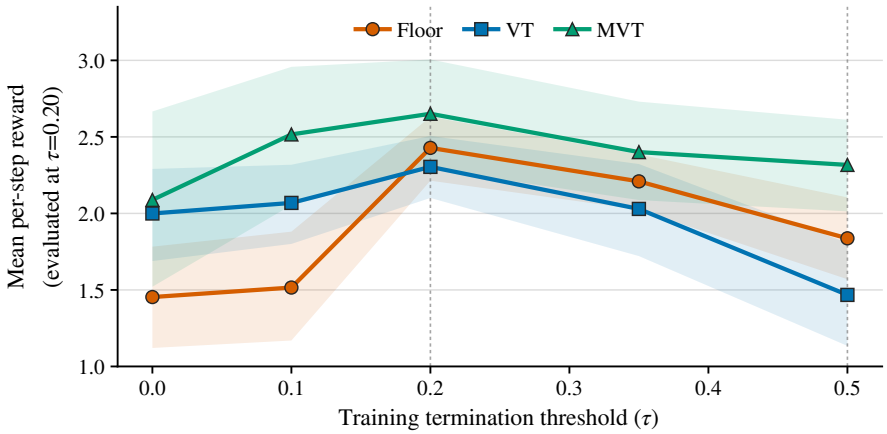


Figure 7: Mean per-step reward by training threshold, per family. Shaded ribbons are 95% morphology-bootstrap CIs ($B = 5{,}000$) computed on seed-averaged morphologies, with each policy evaluated at $\tau_{\text{eval}} = 0.20$. VT's drop at strict termination comes mostly from the ten VT bodies whose Strict runs are cut off during the drop from the reset pose. The $\tau = 0.10$ and 0.35 points show the shape of reward across τ only, because those cutoffs do not act on every body as their values suggest. **Takeaway:** Floor bodies lose the most without fall termination.

G Within-Family Ratio Distributions

The family-level means reported in Table 1 average over heterogeneous individual morphologies. Figure 8 shows the full within-family distribution of RF/Std and Str/Std ratios. Every Standard denominator is safely positive (minimum seed-averaged Standard PSR $= 0.789$), so the ratios are not artifacts of near-zero baselines.

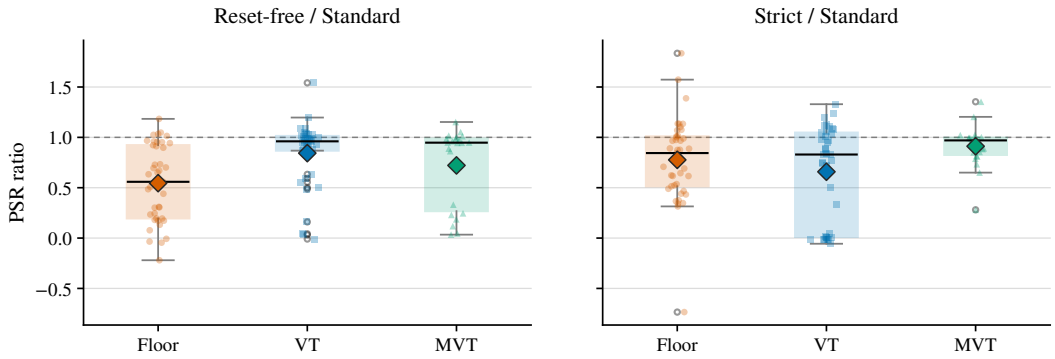


Figure 8: Per-morphology RF/Std (left) and Str/Std (right) ratios by family, using seed-averaged PSR at $\tau_{\text{eval}} = 0.20$. Boxes show medians and IQRs, and diamonds family means. The low VT Str/Std values come mostly from runs cut off during the drop from the reset pose (Appendix C.1). **Takeaway:** the family difference in RF/Std is a shift of the whole distribution, not an outlier effect.

H Per-Seed Replication and Alternative Blocked Models

H.1 Robustness of the Protocol-by-Family Interaction

Table 11 adds five robustness checks to the per-seed ANOVA of the protocol-by-family interaction. The ANOVAs on per-seed rows treat a body's three seeds as independent, so they are supplementary, and the tests that treat the body as the unit are primary. All five confirm the interaction direction: four return $p < 0.01$, and the mixed-effects model returns a sign-consistent VT-by-reset-free coefficient with a 95% CI excluding zero. Without the 11 bodies whose Strict runs are cut off on average, the permutation p is 0.14 (Appendix L.1) and the clustered Wald p is 0.007.

Table 11: Protocol-by-family interaction: supplementary per-seed ANOVA plus five robustness checks (Section 4). The mixed-effects row reports the VT-by-reset-free interaction coefficient with 95% CI in place of a p-value.

Test	Statistic	p-value
Per-seed two-way ANOVA	$F(4, 794) = 12.54$	6.8×10^{-10}
Morphology-level permutation	$F_{\text{obs}} = 12.54$ vs. null 95th $= 7.39$	5.2×10^{-3} (B=5,000); 4.5×10^{-3} (B=10,000)
Seed-averaged two-way ANOVA	$F(4, 194) = 4.79$	1.1×10^{-3}
Rank-transformed per-seed ANOVA	$F(4, 794) = 11.43$	5.1×10^{-9}
Cluster-robust Wald + GEE backup	$F(4, 99) = 4.79$; $\chi^2_4 = 21.91$	1.4×10^{-3}; 2.1×10^{-4}
Mixed-effects (random morph intercept)	VT×RF $= +0.67$, CI $[0.39, 0.95]$	n/a

H.2 Balanced-Family Robustness

Family sizes are imbalanced (Floor $n = 41$, VT $n = 37$, MVT $n = 22$), so we repeated the PSR interaction analysis on 1,000 balanced subsamples with 22 morphologies per family (deterministic seed). The directional family ordering is stable under this reanalysis: Floor RF/Std remains below VT RF/Std in all 1,000 balanced draws.

H.3 Leave-One-Family-Out Robustness

To identify which families drive the interaction we refit the per-seed blocked ANOVA on every 2-family subset. The two anchored contrasts are stable in this per-seed test, which treats seeds as independent: dropping MVT preserves the Floor-vs-VT contrast ($F(2, 620) = 21.01$, $p = 1.5 \times 10^{-9}$, body-level permutation $p < 0.01$), and dropping Floor preserves the VT/MVT strict-termination contrast ($F(2, 468) = 11.09$, $p = 2.0 \times 10^{-5}$, body-level permutation $p = 0.043$). The Floor-vs-VT

contrast on per-body RF/Std (Floor 0.55 vs. VT 0.84) has $\Delta = -0.298$, 95% CI $[-0.454, -0.137]$ ($B = 10{,}000$, bootstrap within families), and the VT-vs-MVT contrast on Str/Std (VT 0.66 vs. MVT 0.91) has $\Delta = -0.254$, $[-0.430, -0.082]$. Of these two contrasts, only the Floor-vs-VT contrast in RF/Std is a family result, because the VT-vs-MVT Strict contrast comes from the ten VT bodies whose Strict runs are cut off during the drop from the reset pose: without them it is -0.01, $[-0.15, 0.12]$ (Appendix C.1).

H.4 Per-Seed RF/Std ANOVA

The family interaction replicates across all three random seeds (Figure 9 and Table 12).

Table 12: Per-seed replication of the family interaction. The last row averages rewards over seeds before forming ratios, as in Table 1.

Seed	Floor RF/Std ↑	VT RF/Std ↑	MVT RF/Std ↑	F	p
42	0.528	0.875	0.747	5.02	0.008
43	0.545	0.846	0.773	4.90	0.009
44	0.594	0.833	0.679	3.24	0.043
Seed-averaged	0.545	0.843	0.722	6.46	0.002

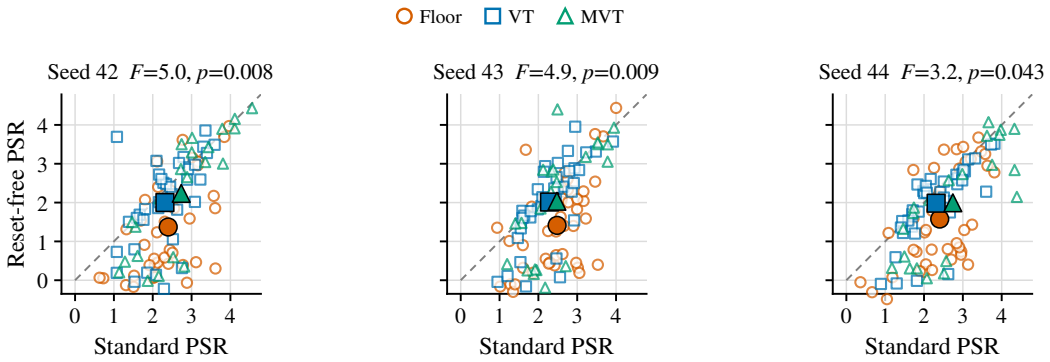


Figure 9: Per-seed scatter: Standard vs. Reset-free per-step reward for each random seed. Larger filled markers show family means. **Takeaway:** the family separation is consistent across seeds.

The RF/Std ordering (VT > MVT > Floor) holds in all three seeds; per-seed ANOVA statistics range from $F = 3.24$ to 5.02 ($p \in [0.008, 0.043]$). The seed-averaged ANOVA on RF/Std ratios ($F = 6.46$, $p = 0.002$) and the bootstrap confidence interval for the VT-Floor difference ($[0.14, 0.45]$) confirm the robustness of the effect.

H.5 Alternative Blocked Models

The main-text per-seed interaction test, which is supplementary because it treats a body's three seeds as independent, uses a per-seed blocked ANOVA ($n = 900$) with morphology fixed effects, protocol, and the protocol-by-family interaction; morphology is nested in family, so the family main effect is absorbed and not interpreted. The interaction gives $F(4, 794) = 12.54$, $p = 6.8 \times 10^{-10}$, partial $\eta^2 = 0.059$. Two alternative specifications preserve the interaction:

- **Seed-averaged blocking** (one observation per cell after averaging across the three seeds per (morphology, protocol); $n = 300$): interaction $F(4, 194) = 4.79$, $p = 0.00106$, partial $\eta^2 = 0.090$; protocol main $F(2, 194) = 25.31$, $p = 1.7 \times 10^{-10}$. Each body belongs to one family, so once each body has its own effect a family main effect is not identifiable, and we do not report one.
- **Mixed-effects random-intercept model** with protocol, family, their interaction, and a random intercept per morphology on per-seed data recovers the main reset-free contrast:

reset-free main $= -0.975$ (95% CI $[-1.17, -0.78]$), which is the effect in the Floor reference family (the average over all bodies is -0.64), and VT-by-reset-free interaction $= +0.670$ (95% CI $[0.39, 0.95]$). We draw no strict-side family conclusion from these models, because the strict-side family contrast comes from runs cut off during the drop from the reset pose (Appendix C.1).

H.6 Morphology-Level Permutation Test

The parametric F-distribution on the blocked model could in principle be unreliable because morphology is nested in family. We therefore add a non-parametric permutation complement: shuffle family labels across the 100 morphologies while preserving family sizes (41, 37, 22), refit the same blocked two-way ANOVA, and record the Type II interaction F. Under the null that protocol response is family-independent, the observed $F = 12.54$ should fall within the permutation F distribution; instead it exceeds the 95th percentile (7.4) of the null and yields $p_{\text{perm}} = 0.005$ at $B = 5{,}000$ (null mean $= 2.75$, null max $= 19.02$), and the released recomputation gives $p_{\text{perm}} = 0.004$ at $B = 10{,}000$. Without the 11 bodies whose Strict runs are cut off on average, the permutation p is 0.14 (Appendix L.1). A simpler one-way ANOVA on per-morphology RF/Std ratios (no morph block, so no nesting concern) gives $F = 6.46$ with $p_{\text{perm}} = 0.003$ at $B = 10{,}000$ (null 95th percentile $= 3.06$). Both permutation tests confirm the interaction is not an artifact of the parametric F distribution or the nested design.

H.7 Metric Sensitivity: Per-Step Reward vs. Total-Reward Per Horizon

Per-step reward (PSR) estimates a reward rate; total reward over the fixed horizon $T_{\max}$=1000 (fitness$/T_{\max}$) treats post-termination reward as zero and is the natural conservative complement when episodes end early. PSR can over-credit a fast-then-fall policy by averaging only over surviving steps; fitness$/T_{\max}$ cannot. Recomputing every Table 1 ratio under fitness$/T_{\max}$ gives RF/Std Floor 0.473, VT 0.847, MVT 0.678 (vs. PSR ratios 0.545, 0.843, 0.722), and the one-way family ANOVA on fitness RF/Std yields $F(2, 97) = 8.37$, $p = 4.4 \times 10^{-4}$, larger than the PSR-based $F = 6.46$. The per-body PSR-vs-fitness rank correlation exceeds 0.95 at every protocol. The family difference in RF/Std does not depend on the PSR normalization.

H.8 Compute-Budget Audit: Where the 5M Steps Go

A natural concern is that under reset-free training (no fall termination), Floor bodies might consume the 5M-step budget in unrecoverable post-fall states, so the Floor RF underperformance would reflect an effective compute deficit rather than a body-by-protocol learning effect. The trained policies argue against this. Under neutral evaluation ($\tau_{\text{eval}} = 0$, no z-fall termination), every (family, protocol) cell averages at least 999 of 1000 steps (Floor RF $= 1000.0$; Floor Strict $= 999.0$; all other cells $= 1000.0$), because without a fall clause only the upper height bound can end an episode early, and it did so for one controller only. Eval-time fall behavior at the paper's headline threshold ($\tau_{\text{eval}} = 0.20$) shows Floor RF drops below $0.20\, z_0$ in 61% of evaluation episodes (mean episode length 512 vs. 828 for Floor Standard), but neutral-eval distance for Floor RF is 31.2 m, comparable to Floor Strict (31.2) and below Floor Standard (44.5). Floor RF is therefore not simply a collapsed policy: under neutral evaluation it keeps moving, but its torso often drops below the standard height threshold.

I Morphological Features Track Sensitivity

The families differ in evolutionary origin, so we also ask whether structural features alone track protocol sensitivity in this corpus. We correlate simple graph-derived morphology descriptors (tree depth, joint count, rigid-body count, first-level limb count, and reset height) with RF/Std and Str/Std PSR ratios (Table 13).

Initial height and number of first-level limbs are the clearest opposite-sign predictors. Taller bodies are less robust to reset-free training ($\rho = -0.533$) but more robust to strict termination ($\rho = 0.382$). More first-level limbs show the reverse pattern ($\rho = 0.496$ for RF/Std, $\rho = -0.370$ for Str/Std). Both survive Bonferroni correction across the feature-ratio correlations in Table 13 on all 100 bodies. The Str/Std correlations depend on the 28 shortest bodies, and none survives the correction on the 72 bodies whose Standard cutoff lies above the torso radius (Appendix C.1), where height has

Table 13: Spearman correlations between representative morphological features and PSR sensitivity ratios. All correlations are computed on per-morphology RF/Std and Str/Std ratios.

	RF/Std		Str/Std	
Feature	ρ	p	ρ	p
Initial height	-0.533	<0.001	0.382	<0.001
Num. first-level limbs	0.496	<0.001	-0.370	<0.001
Tree depth	-0.483	<0.001	0.403	<0.001
Num. joints	-0.117	0.246	0.480	<0.001
Num. bodies	-0.004	0.965	0.375	<0.001

$\rho = 0.03$ ($p = 0.81$). These feature-wise correlations are partially confounded with family labels (because families differ systematically on these features); the linear model and leave-one-family-out cross-validation below probe whether a family-controlled structural predictor remains informative.

Linear feature model for protocol sensitivity. We fit an ordinary least squares regression of the paired delta $\Delta_{\text{str-rf}}(m) = \text{PSR}_{\text{strict}}(m) - \text{PSR}_{\text{RF}}(m)$ on the same five structural features across all 100 corpus morphologies. In-sample $R^2 = 0.46$; leave-one-out cross-validated Spearman $\rho = 0.653$ ($p = 1.9 \times 10^{-13}$). Standardized coefficients (per standard-deviation feature increment) are: initial height $+0.460$, tree depth $+0.352$, number of joints $+0.245$, number of first-level limbs -0.167, number of bodies -0.137. Four of the five coefficients agree in sign with the Spearman correlations (the rigid-body coefficient does not, because the number of rigid bodies is collinear with the number of joints, $r = 0.84$), and they show that taller, deeper bodies tend to be worse under reset-free training (positive $\Delta_{\text{str-rf}}$), while wider, more-limbed bodies tend the other way.

Are structural features doing work beyond baseline skill? A simpler explanation is that the feature model recapitulates overall learnability: a high-skill body wins under any protocol, so $\Delta_{\text{str-rf}}$ is just a function of PSR_{Std}. To check, we fit three nested models on the 100-body corpus (Table 14): M1 uses only baseline Standard skill, M2 only structural features, and M3 both. If the effect were purely a skill artifact, M1 would dominate and M3 would equal M1. Instead, M2 alone explains four times as much variance as M1 ($R^2 = 0.46$ vs. 0.11) and achieves Spearman $\rho = 0.71$ (vs. 0.28); M3 adds only small lift over M2 ($R^2 = 0.55$, $\rho = 0.74$). In this corpus, structural features carry information about protocol sensitivity that baseline skill does not, and the two signals are largely complementary.

A smaller two-feature control using only initial height and first-level limb count explains $R^2 = 0.42$ of $\Delta_{\text{str-rf}}$. Its family residual means remain opposite-signed (Floor $+0.177$, VT -0.342, MVT $+0.244$; one-way ANOVA $F = 3.32$, $p = 0.04$, against $F = 5.12$, $p = 0.008$ for the families before the control), so these two structural features reduce but do not erase the family pattern.

Table 14: In-sample fit for three nested predictors of $\Delta_{\text{str-rf}} = \text{PSR}_{\text{strict}} - \text{PSR}_{\text{RF}}$ on the 100 UNIMAL corpus bodies. M2 features: initial height, num. first-level limbs, tree depth, num. joints, num. bodies. Spearman 95% CIs are bootstrap (B=10,000) over morphologies.

Model	Predictors	$R^2 \uparrow$	AIC $\downarrow$	Spearman $\rho \uparrow$ [95% CI]
M1	PSR_{Std}	0.11	54.3	0.28 [0.08, 0.45]
M2	structural features (5)	0.46	11.1	0.71 [0.59, 0.79]
M3	PSR_{Std} + structural (6)	0.55	-3.4	0.74 [0.63, 0.82]

Leave-one-family-out cross-validation. To test whether the structural-feature signal transfers across DERL lineages within UNIMAL, we hold out one family, fit M1/M2/M3 on the other two, and evaluate the held-out family (Table 15). M2 achieves Spearman point-estimate $\rho \geq 0.48$ on all three held-out families with 95% CIs excluding zero. The transfer is strongest on the smaller MVT fold and remains positive on Floor and VT, but reset height alone ranks each held-out family at least as well (ρ of 0.62, 0.49, and 0.79), so we treat the structural model as a description of this corpus rather than a predictor.

Table 15: Leave-one-family-out cross-validation. For each held-out family, M1/M2/M3 are fit on the other two families' bodies, and Spearman ρ between predicted and observed $\Delta_{\text{str-rf}}$ is reported with bootstrap 95% CIs (B=10,000, resampling held-out bodies). Reset height alone ranks each held-out family at least as well as M2.

Held-out family	n_{test}	M1		M2		M3	
		$\rho \uparrow$	95% CI	$\rho \uparrow$	95% CI	$\rho \uparrow$	95% CI
Floor	41	0.19	$[-0.13, 0.48]$	0.51	$[0.21, 0.73]$	0.53	$[0.26, 0.73]$
VT	37	0.36	$[-0.00, 0.64]$	0.48	$[0.17, 0.69]$	0.60	$[0.32, 0.78]$
MVT	22	0.69	$[0.36, 0.89]$	0.76	$[0.42, 0.93]$	0.80	$[0.49, 0.94]$

J Cross-Algorithm Validation with SAC

As an optimizer-robustness check, we trained SAC [Haarnoja et al., 2018] (the Brax implementation [Freeman et al., 2021]) on a 24-morphology subset (6 Floor, 13 VT, 5 MVT) at $\tau \in \{0.00, 0.20, 0.50\}$ across three seeds for 5×10^6 environment steps each, giving a 3×3 cell per body and $n = 216$ total runs. The bodies were fixed before any SAC analysis to cover three outcomes under Strict training: 8 that fell often, 8 that stayed upright but barely moved, and 8 that were unaffected. The panel therefore covers the range of outcomes rather than estimating family averages. Its evaluation differs from PPO's in three details: SAC is scored with its deterministic action, its reset noise perturbs positions but not velocities, and its reset draws vary by seed.

The panel met its validation gate, a per-run protocol-by-family interaction with $p < 0.01$ and the Floor $>$ VT $>$ MVT ordering on the Strict to Reset-free ratio, which was set after a 96-run pilot of Reset-free and Strict training had shown that ordering.

Table 16: SAC per-step reward by family (seed-averaged across three SAC seeds, eval at $\tau_{\text{eval}} = 0.20$) and the Str/RF validation ratio. Str/RF averages $\text{PSR}_{\text{strict}}/\text{PSR}_{\text{RF}}$ across bodies in each family. Values above 1 favor strict termination.

Family	n	Standard $\uparrow$	Strict $\uparrow$	Reset-free $\uparrow$	Str/RF
Floor	6	3.92	4.34	2.80	2.41
VT	13	3.43	3.10	2.97	1.52
MVT	5	4.07	3.92	4.10	0.96

The planned family-ordering check, Floor $>$ VT $>$ MVT on per-body Str/RF, holds under SAC $(2.41, 1.52, 0.96)$ for the mean of per-body ratios on these 24 bodies (Table 16), and SAC gives VT bodies a smaller Strict penalty than PPO (VT Strict minus Reset-free PSR $+0.12$ under SAC against -0.51 under PPO). SAC changes some absolute magnitudes, so we interpret the panel as a Str/RF ordering check rather than cross-algorithm replication of all PPO ratios. Cross-algorithm rank agreement is strong on the 23 bodies with denominator-valid PPO and SAC Str/RF ratios: Spearman $\rho = 0.75$ ($p = 3.1 \times 10^{-5}$). The omitted body has nonpositive PPO reset-free PSR, so its PPO Str/RF denominator fails the pre-specified denominator threshold (PSR > 0.05); all SAC cells are valid. Leave-one-body-out sensitivity gives min $\rho = 0.720$, median 0.761, max 0.833.

A blocked ANOVA on per-seed SAC PSR ($n = 216$) with body fixed effects, protocol, and the protocol-by-family interaction gives $F(4, 186) = 6.71$ ($p = 4.6 \times 10^{-5}$), matching the main-text per-seed specification, which treats a body's seeds as independent. Tests that treat the body as the unit give $p = 0.016$ (clustered Wald), 0.070 (seed-averaged), and 0.09 (permutation of family labels). Body is nested in family, so the family main effect is absorbed rather than interpreted. We use this as cross-algorithm support for the Str/RF ordering; the PPO corpus remains the source for RF/Std magnitudes and percentage-loss claims.

QC. 216/216 cells are valid (no NaN). Median seed-level standard deviations are $\sigma_{\text{RF}} = 0.148$, $\sigma_{\text{Std}} = 0.181$, $\sigma_{\text{strict}} = 0.257$, all below the planned reliability threshold (0.514).

A2C. During the review period, we also set up synchronous A2C on the same design (24 bodies, three rules and three seeds, 216 runs) under the same gate. The analyzed runs are no longer available, and the surviving record of the analysis reports an interaction $p = 0.073$, above the gate's 0.01.

K Morphology Gallery

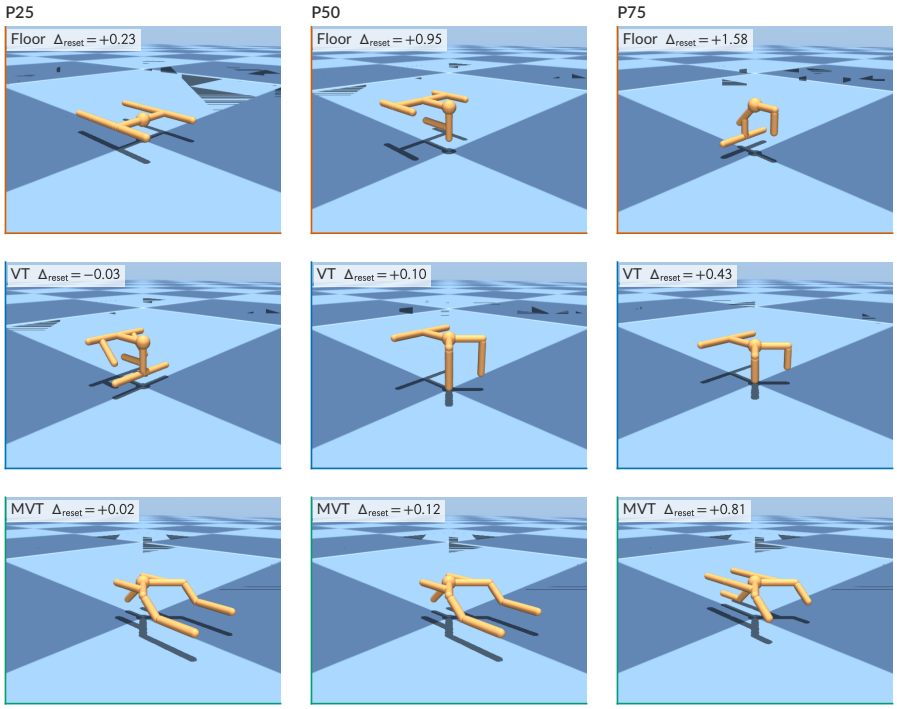


Figure 10: Nine representative UNIMAL morphologies rendered at their reset (spawn) pose. Rows show Floor, VT, and MVT families. Columns show the 25th, 50th, and 75th percentile of $\Delta_{\text{reset}} = \text{PSR}_{\text{Std}} - \text{PSR}_{\text{RF}}$ within each family. For the Floor 25th-percentile body the Standard cutoff lies below the torso radius, and for the MVT 25th- and 50th-percentile bodies it lies 4 mm above the radius and never ended a recorded evaluation episode, so their Δ_{reset} should not be read as an effect of the Standard rule. **Takeaway:** protocol sensitivity appears across visibly different bodies within each evolutionary family.

L Rank Agreement Across Protocols and Seeds

Leave-one-seed-out reselection for Figure 11. Re-running the most-protocol-specialized selection within VT preserves the qualitative pattern: the Standard and Strict specialists reappear across folds, and reset-free alternatives remain VT bodies with the same protocol-specialization profile.

L.1 Main Results on Each Body Set and the Seed-Noise Reference

Table 17 gives the main results on each body set, including a fourth, without the 23 cut-off-seed bodies (Appendix C.1). The reordering and the larger Reset-free cost of Floor bodies hold on every set, and also when both problems are removed at once, on the 65 of the 72 bodies without a cut-off seed (Reset-free vs. Strict $\rho = 0.37$; Floor minus VT on RF/Std -0.30, permutation $p = 0.010$, post hoc). The Strict-side family difference does not: the gap between the Floor and VT losses under Strict training is not significant even on all 100 bodies (post hoc permutation $p = 0.25$) and vanishes on the 72, and the low share kept by Standard-only selection comes mostly from the bodies whose cutoffs do not act as intended.

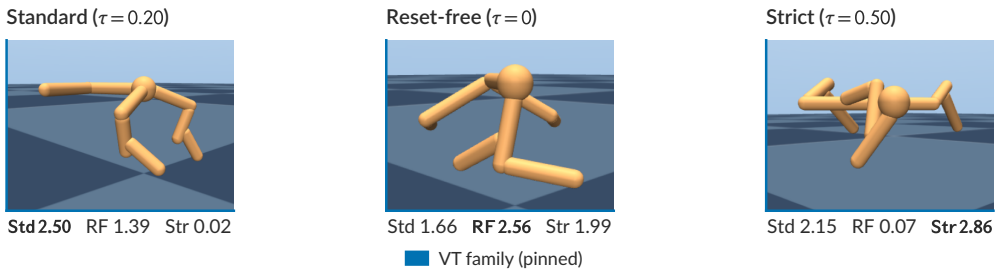


Figure 11: Protocol-specialization exemplars within VT. Family is held fixed, and specialization $=$ PSR$_p$ $-$ max PSR over the other two protocols. Triplet: cross-protocol PSR (bold $=$ training protocol). The Standard and Strict exemplars show sharp cross-protocol collapse (PSR ≤ 0.07), and the Standard exemplar's Strict score of 0.02 comes from Strict runs cut off during the drop from the reset pose (Appendix C.1). **Takeaway:** protocol specialization remains visible even within one evolutionary family.

Table 17: Main results on all 100 bodies, on the 72 whose Standard cutoff lies above the torso radius, and without the 11 cut-off or the 23 cut-off-seed bodies (Section 3.2). The seed-noise reference is the agreement that seed noise alone would leave between the two three-seed rankings, and p-values permute family labels across bodies. Strict loss is one minus Strict retention, and the last row is the in-sample share of the Strict top-5 reward kept by the five best bodies after Standard training. The three right-hand columns are post hoc, as are the seed-noise, D, Strict-loss and Standard top-5 rows and the p-values in the Floor minus VT and Reset-free vs. Standard rows.

Result	All 100	The 72	Without the 11 (89)	Without the 23 (77)
Reset-free vs. Strict ρ	.21	.26	.25	.45
Seed-noise reference for that ρ	.89	.89	.85	.90
Reset-free vs. Strict D	.53	.52	.42	.38
Reset-free vs. Standard D	.17	.20	.16	.20
RF/Std, Floor/VT/MVT	.55/.84/.72	.44/.76/.60	.56/.81/.72	.47/.80/.71
Floor minus VT on RF/Std (p)	$-.30$ ($<.001$)	$-.32$ (.002)	$-.25$ (.008)	$-.33$ (.001)
Reset-free vs. Standard $\times$ family (p)	$<.001$	.003	.007	$<.001$
Three-rule $\times$ family (p)	.004	.01	.14	.008
Strict loss, Floor/VT/MVT (%)	22/34/9	16/17/1	19/10/9	8/4/6
Standard top 5, % of Strict top-5 PSR	49.0	75.7	72.7	95.6

Seed-noise reference. Averaging three seeds makes a ranking more reliable. If W is the mean Spearman correlation between single-seed rankings of one rule, the Spearman-Brown formula $3W/(1+2W)$ gives the reliability of a three-seed ranking as 0.93 for Reset-free, 0.90 for Standard and 0.85 for Strict. The formula matches the data: it predicts the agreement between one seed and the mean of the other two at 0.86 for Reset-free and 0.72 for Strict, against 0.86 and 0.73 observed. If seed noise were the only difference between the rules, the three-seed Reset-free and Strict rankings would therefore agree at about $\sqrt{0.85 \times 0.93} = 0.89$ $[0.84, 0.92]$, against the observed 0.21.

Controls for the reordering. Scored by the neutral evaluator, which has no cutoff, the Reset-free and Strict rankings of the 100 bodies correlate at 0.31 $[0.11, 0.49]$. On the 60 bodies whose $0.10\,z_0$ cutoff lies below the torso radius and that no probe rollout pushed below it, training with $\tau = 0.10$ is in effect Reset-free training, run as a separate launch with the same seeds. The two three-seed rankings correlate at 0.92 $[0.84, 0.96]$, close to the 0.89 expected with independent seeds, while Reset-free and Strict training of the same bodies correlate at 0.44 $[0.20, 0.64]$ (difference 0.48 $[0.27, 0.72]$). Both checks are post hoc.

Excess disagreement. For rules a and b with within-rule agreements W_a and W_b and a mean cross-rule correlation C between single-seed rankings, the excess disagreement $D = (W_a + W_b)/2 - C$ is zero in expectation when changing the rule moves a ranking no more than drawing a new seed does. On all 100 bodies, D is 0.17 $[0.08, 0.28]$ for Reset-free versus Standard, 0.29 $[0.17, 0.43]$ for

Standard versus Strict and 0.53 $[0.36, 0.70]$ for Reset-free versus Strict, with percentile intervals that resample bodies. A null that shuffles each body's Reset-free and Strict runs between two pseudo-rules centers at 0.64, with 95% of draws between 0.55 and 0.73, again far above the observed 0.21.

Top-k overlap. Seeds alone turn over the top of a ranking, so we use top-k overlap as a description rather than a test. Two single-seed runs of the same protocol share 3 to 8 of their top-10 bodies, and two simulated three-seed reruns of one rule share about 7 of their top-10 bodies, so the observed 5 of 10 has a probability of 0.03 to 0.15 under seed noise alone. The top-20 lists of Reset-free and Strict training share 10 bodies, against 15.5 to 16.1 in simulated seed reruns and 13.7 under the exchangeability null ($p = 0.001$). For top-k overlap intervals, the bootstrap holds one protocol's observed top-k fixed.

M Locomotion-Quality Audit

Low per-step reward under Reset-free training could reflect a body that *learned to walk but was penalized by the protocol*, or a body that *never learned to walk at all*. We compare the distance traveled after each training protocol under the common and the neutral evaluator.

We classify a (morphology, training protocol) cell as *walks* if the mean evaluation distance is at least 5 m and as *stationary* if below 1 m. We report cross-protocol transitions at both $\tau_{\text{eval}} = 0.20$ (the main-text evaluation) and $\tau_{\text{eval}} = 0$ (no early termination, so distance is integrated over the full evaluation horizon), in Tables 18 and 19.

Table 18: Per-family cross-protocol locomotion transitions at $\tau_{\text{eval}} = 0.20$. A body that walks under Standard but is stationary under the target protocol either learned locomotion and was disrupted by the protocol or, for the Floor-reset-free cell, learned a policy whose evaluation episodes the fall cutoff often ends early.

Family	n	Std→RF halt	Std→Str halt	Std halt→RF walks	Std halt→Str walks
Floor	41	12	1	0	0
VT	37	4	10	0	0
MVT	22	5	0	0	0

Table 19: Per-family cross-protocol locomotion transitions at $\tau_{\text{eval}} = 0$ (no early termination; distance integrated over the full evaluation horizon). The VT-Strict and MVT-Strict counts hold unchanged, but all ten VT-Strict stops are bodies with cut-off Strict runs, so they are not a family effect (Appendix C.1).

Family	n	Std→RF halt	Std→Str halt	Std halt→RF walks	Std halt→Str walks
Floor	41	0	1	0	0
VT	37	0	10	0	0
MVT	22	0	0	0	0

Twelve of 41 Floor bodies walk under Standard but become stationary under reset-free at $\tau_{\text{eval}} = 0.20$, compared to only one that becomes stationary under strict; four VT bodies become stationary under reset-free, ten under strict; MVT shows five vs. zero. Under $\tau_{\text{eval}} = 0$ with no early termination, none of the 12 Floor bodies stays under 1 m after reset-free training, and 5 of the 12 walk at least 5 m, and the VT-Strict and MVT-Strict counts are unchanged (10 and 0). All ten VT-Strict stops are bodies with cut-off Strict runs, so they are not a family effect (Appendix C.1). This audit supports attributing the family ordering to the training reset rule: under neutral evaluation, RF/Std ratios remain ordered Floor 0.643 < MVT 0.847 < VT 0.953 (Figure 6), and fixed-horizon fitness ratios similarly keep Floor lowest (Appendix H.7).

N Control-Effort Confound Audit

We derive $\widehat{\sum_t \|a_t\|^2}_{\text{per episode}} = (\text{distance}/dt - \text{fitness})/10^{-3}$ with $dt = 0.02\,\text{s}$ (this recovers each run's mean per-episode sum from its recorded mean distance and return, given the reward in Appendix A.2). Per-step $\|a_t\|^2$ is family-protocol-flat at 8.8–9.8, so the family difference in Section 4 is not explained by one family pushing actuators harder. Among runs whose controllers travel at least 5 m per episode, the squared action norm per meter (not torque, work, or physical energy) is 178–266 across all nine family×protocol cells; differences do not mirror the PSR family-by-protocol interaction.

O Extended Discussion

Implications for autonomous reset-free RL. Reset-free RL [Sharma et al., 2022, Yang et al., 2025, Patil et al., 2024, Gupta et al., 2021a] removes episodic resets; our τ=0 condition removes one component of the reset schedule, the height-triggered early termination. The body-dependent sensitivity we observe suggests a concrete test for broader reset-free settings: if morphology changes rankings under matched finite-horizon training, algorithms developed for settings with fewer resets [Sharma et al., 2022, Yang et al., 2025, Patil et al., 2024] should be evaluated across body families rather than on a single fixed morphology. Floor-family bodies are candidate stress tests for recovery from difficult states. This motivates co-designing bodies and algorithms under the reset schedule of interest.

Why sweep the termination threshold. Co-design and locomotion implementations configure the height-termination threshold in different ways (Appendix C), so the body rankings they report can depend on a choice that the audited sources do not sweep as a benchmark variable. Removing the height predicate during training, our τ=0 condition, is also a finite-horizon analogue of what an autonomous-RL pipeline would face if it inherited a co-design-selected body and had to learn without hand-tuned fall detectors. The τ=0 condition is a co-design-relevant corner of the protocol space; the paper's claims are scoped to the simulator (cf. Section 5).

P Time Limit and Other Protocol Variables

Section 4.8 reports that, by its planned criterion, the no-boundary variant changes the ranking ($\rho = 0.85$, 95% CI $[0.67, 0.93]$, whose upper end lies below the planned bound of 0.95) and family magnitudes relative to corpus τ=0. Its runs used a later code revision than the corpus. A panel of 12 bodies retrained with that revision under all three rules (108 runs) ranks the bodies as the corpus does at $\rho = 0.97$, 0.88 and 0.73 for Reset-free, Standard and Strict, and in a post hoc comparison the revision lowers Reset-free Floor scores by a similar amount as the no-time-limit panel (a median of 0.42 PSR on that panel's 4 Floor bodies and 0.48 on 7 Floor bodies of a longer-budget panel, both with the ordinary 1,000-step limit), so the panel does not isolate the time limit.

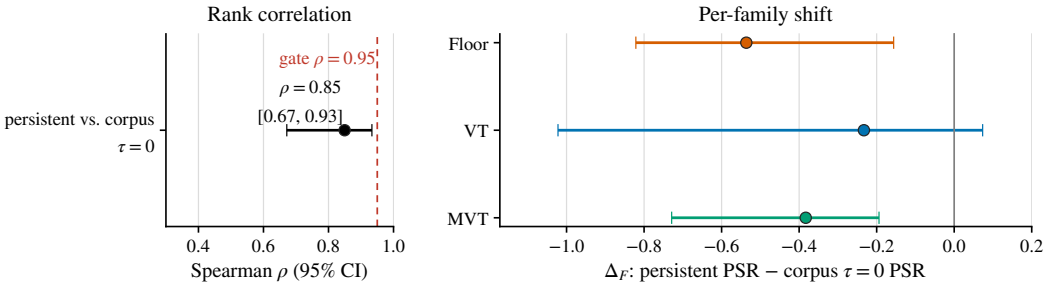


Figure 12: Persistent no-boundary PPO vs. corpus τ=0 on a 30-body × 3-seed panel, trained with a later code revision than the corpus. Left: per-body rank correlation with corpus τ=0 is $\rho = 0.85$, and the upper end of its interval lies below the planned gate of 0.95. Right: family-level shifts (persistent minus corpus), median Δ_F with 95% bootstrap CIs. **Takeaway:** the no-time-limit runs, trained with later code, rank the bodies differently from the corpus and score Floor and MVT bodies lower.

A planned factorial test on the same 30 bodies, which also trained Standard and Strict without a time limit (180 runs), finds that the reset rule and the time limit interact ($F(2, 505) = 3.72$, $p = 0.025$), which meets the plan's criterion for the appendix. Additional boundary-duration panels give the same qualitative reading. Spearman correlations against the corpus B=1000 setting are $\rho_{3001} = 0.916$ (95% CI $[0.784, 0.971]$), $\rho_{10001} = 0.852$ ($[0.680, 0.945]$), and $\rho_{5\text{M}} = 0.850$ ($[0.672, 0.934]$). The point estimates decrease as the boundary lengthens, while confidence intervals overlap. A per-body regression of $\Delta(B) = \text{PSR}(1000) - \text{PSR}(B)$ on $\log B$ has slope 0.099 (95% CI $[0.053, 0.150]$), consistent with larger magnitude shifts under longer boundaries. These runs used the same later code. The no-boundary panel is therefore a boundary-setting check for the corpus reset-free condition, not an isolated test of the time limit.

P.1 Other Design Spaces and Protocol Variables

We also trained panels in other design spaces and on other protocol variables, each with a report-only decision rule. In three pools of 30 parametric Humanoid bodies trained with Brax PPO under four fall rules (none, the native Gymnasium height range, and cutoffs at 0.2 and 0.5 of the root height) and scored by one evaluator without fall termination, rankings under a fall rule agree with rankings without one at 0.58 to 0.73 in two pools and 0.11 to 0.19 in the third. Seed halves of one rule also agree poorly (0.26 to 0.69), and although D is above zero in the third pool (0.28 $[0.01, 0.55]$ for the native rule), no D stays above zero after adjustment for multiple comparisons, so this panel cannot resolve an effect of the fall rule beyond seed variation. Changing how time limits are handled (Humanoid and Evolution Gym), how resets are initialized, or adding observation noise (Ant) left rankings within the agreement between seeds ($\rho = 0.65$ to 0.99). The terminal Humanoid arm used later code than its bootstrapped arm, and much of the Ant agreement comes from a wide spread of skill. In these checks, reordering beyond seed variation appears only for fall termination on UNIMAL bodies, so whether another benchmark is sensitive has to be checked rather than assumed.

Q Companion Chronophotograph: Bodies that Walk Under All Three Protocols

Figure 13 shows *MVT A* and *Floor B* under the three protocols: both walk across all three, as in all 9 recorded corpus runs of each.



Figure 13: 3D-rendered chronophotograph of *MVT A* and *Floor B* from illustrative retrains across three training protocols. Each panel overlays poses from one rollout seen by a fixed camera, so their spread shows the distance covered. **Takeaway:** protocol dependence is body-specific, as *MVT A* and *Floor B* walk under all three protocols.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state three specific claims (a family-specific difference in Reset-free cost, ranking instability across protocols beyond seed noise, and an in-sample protocol-robust selection). Each is supported by quantitative results in Section 4 (Table 1, Table 2, and Table 3). Limitations are discussed in Section 5.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The Scope and limitations paragraph of Section 5 and the appendix itemize the scope: (1) family and ranking claims scoped to UNIMAL forward locomotion in simulation; (2) height-based termination with non-threshold episode mechanics specified identically across protocols (Appendix C), and a per-body cutoff check, with the main results repeated without the affected bodies (Section 3.2, Appendices C.1 and L.1); (3) balanced-family reanalysis preserves the directional Floor/VT ordering (Appendix H); (4) a fixed 5M-step training budget (Appendix H.8); (5) per-morphology reset-free benefit correlates with baseline Standard skill, controlled via structural features in Section 4.7.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.

- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper does not contain formal theoretical results.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: PPO hyperparameters are listed in Appendix Table 4, and the SAC settings are in the released SAC trainer (Appendix B.1). The morphology space (UNIMAL) is publicly available. Training uses a custom PPO implementation on MuJoCo MJX with three fixed seeds; the cross-algorithm validation uses Brax SAC. Statistical analyses (one-way and blocked two-way ANOVA, Spearman correlations with percentile bootstrap CIs, top-k overlap, the M1/M2/M3 linear feature regressions) are standard and specified in Section 3 and Appendices E, H, I and L. The public code repository (`https://github.com/kevinxqiu/reset-dependence`) and its release include run summaries, morphology features, validation reports, planned-test records, CPU analysis scripts that recompute the reported statistics from completed runs, and training, analysis, rendering and launch code, with a script that downloads the UNIMAL body files and checks their hashes. No checkpoints of the UNIMAL runs were saved (Appendix B.1). Absolute per-step-reward magnitudes are sensitive to training-time implementation details (Appendix P), so our quantitative claims target rankings, family orderings, and the protocol-by-family interaction.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed

instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.

- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The public code repository (`https://github.com/kevinxqiu/reset-dependence`) provides run summaries and CPU analysis scripts sufficient to recompute the main UNIMAL statistics, validation checks, and quantitative figures from the completed experiments. It documents the artifact layout and exact commands, and includes training, analysis, rendering and launch code, with a script that downloads the UNIMAL body files and checks their hashes. These are reproducibility artifacts for an empirical evaluation audit, not a new dataset or benchmark-suite contribution. No checkpoints of the UNIMAL runs were saved, so no trained policies or rollout videos of these runs are released (Appendix B.1). The UNIMAL morphology space is publicly available from Gupta et al. [2021b].

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: Section 3 describes the training protocol, cross-evaluation procedure, and analysis methods. Appendix A provides the PPO hyperparameters, environment details and compute, and Appendix J gives the SAC details (main-text summary in Section 4.5).

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Percentile 95% bootstrap confidence intervals are reported for ranking correlations using $B = 10{,}000$ morphology resamples. Per-seed replications, a morphology-level permutation test ($p_{\text{perm}} = 0.004$, $B = 10{,}000$) and a cluster-robust Wald test are reported, with a two-way blocked ANOVA on runs as a supplementary test. A balanced-family reanalysis ($B = 1{,}000$ subsamples) is reported as a robustness check. A threshold-free distance audit (Section 4.6) reproduces the Floor/VT difference without ratios.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: Appendix A.4 lists the hardware and, where recorded, the run time of every released panel. Corpus runs used one GPU each, at about 1.09 GPU-hours per 5M-step PPO run, so the 900-run PPO corpus took about 980 GPU-hours. The SAC panel's 216 runs include its 96-run pilot; runs outside the released records, such as those of the A2C panel, are not counted.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: This research uses simulated robots only and does not involve human subjects, personal data, or dual-use concerns. It conforms with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: This work studies evaluation methodology for simulated robot co-design and does not introduce deployable robot capabilities, human-subject data, or dual-use models. Its primary positive impact is methodological: reducing the chance that a benchmark selects a morphology because of an implicit simulator reset rule rather than robust task performance. The main negative risk is indirect: if protocol sensitivity is ignored, benchmark results could steer expensive morphology fabrication toward bodies that are brittle under a different reset regime. The paper mitigates this by recommending explicit reset-rule reporting, protocol-sensitivity audits, and protocol-robust selection when choosing bodies.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.

- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The release contains simulated-locomotion run summaries and analysis scripts, not pretrained models or datasets with misuse risk.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We use the UNIMAL morphology space [Gupta et al., 2021b] and the public UNIMAL-100 benchmark associated with MetaMorph [Gupta et al., 2022]; the code repository does not redistribute the UNIMAL body files, whose upstream repositories declare no license, but downloads them and checks their hashes; our code is under the MIT license and the run records under CC BY 4.0, with a third-party notice. MuJoCo / MuJoCo MJX [Todorov et al., 2012] and the Brax SAC implementation [Freeman et al., 2021] are Apache 2.0 licensed. All are properly cited.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The repository README and the data card of its release document the released run summaries, morphology features, cached statistical reports and CPU analysis scripts (Appendix B.1). These are run records rather than a new dataset resource. No checkpoints of the UNIMAL runs were saved.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: This research does not involve crowdsourcing or human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: This research does not involve human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLMs are not used as a component of the core methodology. The research involves training RL policies in simulation.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.