HORIZON: RECOVERABLE PHYSICAL-DOMAIN EXPANSION FOR POLICY GENERALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

In on-policy reinforcement learning, expanding physical randomization changes both the control problem and the trajectories that drive learning. Retracting a failed expansion cannot undo learner updates that erode acquired behavior. We introduce HORIZON, a recoverable curriculum that treats physical-domain expansion as a coupled update of the training distribution and learning state. The method accepts an expansion only when the updated policy satisfies candidate-range feasibility and retains performance under previously accepted conditions. After failure, it restores the accepted ranges and learning state while preserving the failed boundary for refinement. Failed exploration thus informs subsequent proposals without its learner updates persisting. In quadruped locomotion, HORIZON achieves 36.8% success under the strongest joint physical shift, compared with 19.8% for OpenAI ADR and 22.7% for GRAM. We further find that broader physical diversity need not improve joint generalization. Under the evaluated training allocations, a four-group curriculum achieves higher mean success on the tested joint shifts than full seven-group expansion, while the latter remains stronger on isolated shifts outside the four groups. The resulting fixed-weight policy transfers to hardware under payload changes, joint constraints, and leg extensions without online policy optimization. These findings identify recoverable expansion and domain composition as complementary considerations for learning generalizable motor control.

1 INTRODUCTION

Generalizable motor control requires policies to act under physical conditions beyond those encountered during training. Dynamics randomization broadens this experience by varying the simulator (Peng et al., 2018; Tan et al., 2018), but capability is not determined by range width alone. In on-policy reinforcement learning, the controller produces the trajectories from which it learns, so expanding the physical distribution changes both the control problem and the data used to update the learner. Harder conditions can reveal control errors and provide useful corrective data, but abrupt expansions can instead replace informative interaction with repeated early failures and erode behavior that was already acquired. Physical-domain generalization therefore raises two coupled questions: how can training expand into harder physics without allowing failed expansions to erode acquired behavior, and which physical variations should be learned together when their effects interact?

Performance-driven randomization adapts simulator distributions using boundary performance or success constraints (OpenAI et al., 2019; Tiboni et al., 2024). These methods make distribution expansion responsive to policy competence, but the state of the learner presents a distinct problem: retracting a failed range changes what the policy encounters next, yet does not undo updates already made under the rejected proposal. We call an expansion *training-time recoverable* when it sustains feasible on-policy rollouts under candidate conditions while retaining performance under previously accepted conditions. This view treats curriculum progress as a coupled change in the physical range and the learner state, rather than as a distribution update alone.

HORIZON implements this view through checkpointed physical-domain expansion. Each proposal is trained for a bounded phase and assessed both under candidate conditions and under accepted ranges relative to the last accepted checkpoint. When both criteria hold, HORIZON commits the expanded range and updated learner. Otherwise, it restores the accepted ranges and checkpointed

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

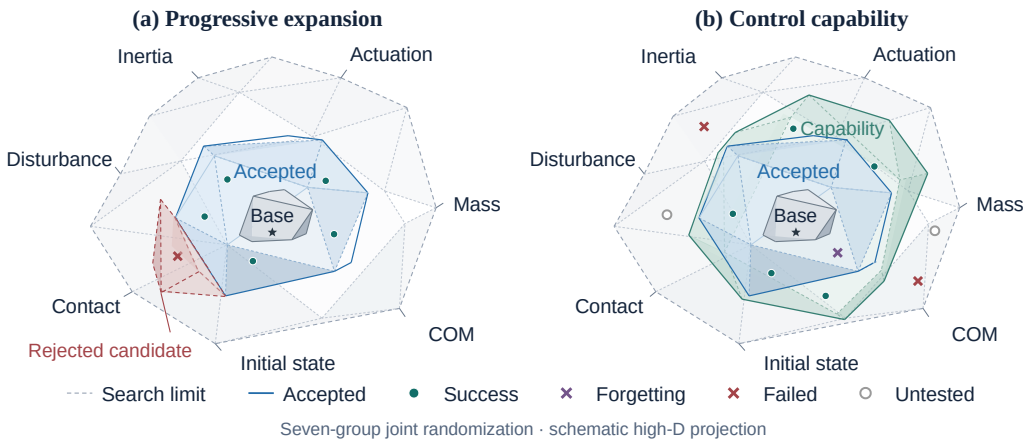


Figure 1: Progressive expansion and control capability in a schematic projection of seven physical groups. (a) The accepted domain expands from near-nominal Base within the search limit; the red extension marks a rejected candidate. (b) The capability region retains Base but differs from the accepted training domain.

learning state before refining the boundary. The policy updates produced by a rejected proposal therefore do not become the starting point for subsequent training, while the failed boundary still informs refinement. The deployed policy remains a fixed-weight, history-conditioned actor and receives no physical-parameter labels.

Training-time recoverability does not by itself determine which physical variations should be combined. In quadruped locomotion, seven physical-randomization groups act through the same closed-loop controller: actuation, mass, disturbance, contact, inertia, center-of-mass offset, and initial state. A payload shift, for example, may require corrective torque precisely when reduced actuation makes that torque unavailable. Robustness to each group separately therefore does not determine robustness to joint shifts, and broader expansion need not yield stronger joint control. Extending prior studies of which randomized properties are useful (Xie et al., 2021; Exarchos et al., 2021), we use single-group, pairwise, and three-group diagnostics to motivate a four-group Core comprising actuation, mass, disturbance, and center-of-mass offset. We then compare its fixed policy checkpoint with Full across isolated shifts and a broad set of joint shifts.

We therefore formulate physical-domain expansion as a coupled distribution-and-learning-state problem, with HORIZON governing candidate feasibility and retention of accepted behavior through checkpointed commits, learning-state restoration, and boundary refinement. Controlled mechanism studies isolate all three components: L_k requires relatively stable mastery of the current candidate domain, C_k limits excessive forgetting on accepted domains, and restoring the learner improves subsequent control beyond retracting physical ranges alone.

Our experiments further show that broader physical coverage need not improve joint generalization under the evaluated training allocations: the four-group Core has higher success on 57 of 91 retested joint combinations and a 3.37-point higher mean success, whereas the broader curriculum remains stronger on isolated non-Core shifts. Comparisons with alternative curricula, OpenAI ADR, DORAEMON, GRAM, and expert distillation evaluate overall policy performance, while fixed-weight hardware trials test transfer under payload changes, joint constraints, and leg extensions.

2 Related Work

2.1 Adaptive Dynamics Randomization

Domain randomization broadens sim-to-real training by varying visual or physical properties (Tobin et al., 2017; Peng et al., 2018). Adaptive methods either fit simulator distributions to target-domain trajectories (Chebotar et al., 2019; Ramos et al., 2019; Muratore et al., 2022) or use simulation feedback. Automatic Domain Randomization moves uniform-distribution boundaries according to

boundary performance (OpenAI et al., 2019), while Active Domain Randomization selects informative conditions within predefined ranges (Mehta et al., 2020). DORAEMON maximizes distribution entropy subject to a success constraint and retracts the distribution when that constraint is violated (Tiboni et al., 2024). HORIZON couples group-wise distribution adaptation to retention under accepted conditions and checkpointed learning-state updates, using simulation feedback throughout.

2.2 Curriculum Expansion and Retention

Curriculum learning orders tasks by difficulty or learning progress (Bengio et al., 2009; Narvekar et al., 2020). Automatic variants generate tasks, expand initial-state sets, or revisit informative environments (Portelas et al., 2020; Florensa et al., 2017; Jiang et al., 2021). Continual reinforcement learning studies how new experience can alter acquired behavior (Khetarpal et al., 2022). Physical randomization links these issues through on-policy data collection: altering the dynamics changes the trajectories available for learning as well as the task distribution. This motivates assessing candidate feasibility and retention together when deciding whether to accept a physical-domain expansion.

2.3 Physical Variation and Motor-Control Generalization

Robust motor-control systems use history-based adaptation, terrain observations, fast parallel simulation, or expressive policy architectures (Kumar et al., 2021; Lee et al., 2020; Miki et al., 2022; Rudin et al., 2022; Liu et al., 2025). GRAM combines adaptation and robustness within the policy (Queeney et al., 2026). ADR and DORAEMON adjust the training distribution, while HORIZON also restores the accepted learning state after failed expansions, keeping the deployed actor interface fixed. Prior quadruped studies also show that the selected randomized properties affect transfer (Xie et al., 2021; Exarchos et al., 2021). We extend this question to group composition under a finite PPO-update budget, comparing group sets under the same joint evaluation. The resulting pairwise and composition effects remain conditional on the tested controller and physical ranges.

3 HORIZON

HORIZON expands physical parameter ranges during policy learning. In the quadruped benchmark, the actor is a two-layer decoder-only transformer with a 32-step key-value (KV) cache. At each control step, it receives the current observation and command together with the previous action. A privileged critic supports PPO training, while the actor receives only deployment-available inputs. Physical groups and their admissible bounds are inputs to the curriculum. Figure 2 summarizes the resulting loop from range proposal and policy updates to the joint feasibility and retention gate, followed by either commitment or restoration and boundary refinement.

3.1 Physical Ranges and Domain Groups

We group physical parameters into actuation, mass, disturbance, center-of-mass (COM) offset, initial state, contact, and inertia. A group may contain several intervals. For example, actuation includes motor effort and velocity limits together with stiffness and damping. A single expansion phase can therefore challenge several actuation parameters together.

For each group we distinguish Base ranges, allowed search limits, accepted ranges, and a current candidate. Base defines near-nominal randomization. Search limits bound the proposals available to the curriculum and are chosen to remain within the simulator's physically valid regime, preventing excessively broad distributions from producing invalid or unstable environments. Accepted ranges describe the intervals retained after successful phases, while candidate ranges describe the conditions currently being trained. Table 10 gives the Base and search intervals. These are training objects: success on a separate test distribution is required to establish control capability.

At each phase k, one eligible group g_k is selected for expansion. Its candidate replaces that group's accepted intervals. Other active groups continue to be randomized over their accepted intervals, whereas groups never designated for expansion during the run remain randomized at Base throughout training. Thus focused expansion changes one group's ranges while retaining joint variability in

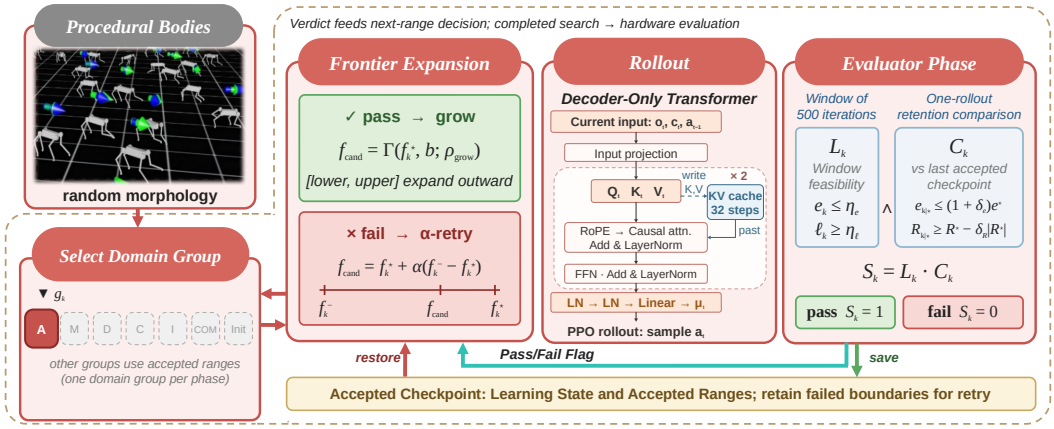


Figure 2: Checkpointed expansion with a cached decoder-only actor and privileged PPO critic. Each actor layer reads past K/V and writes current K/V (dashed arrow). LN denotes LayerNorm; the output LNs belong to the backbone and actor head. PPO samples actions; deployment uses μ_t. Both gates govern acceptance or restoration and refinement. Training uses Procedural Bodies, with group g_k from Table 10. Appendix A and Table 3 detail the algorithm and restored state.

the environment. Selection uses a seed-shuffled queue, with eligible groups revisited in later cycles. The method does not require a fixed sequence in which entire groups become enabled.

3.2 Recoverability-Governed Expansion

HORIZON first trains under Base randomization and saves an accepted learning checkpoint. Each subsequent phase starts from that checkpoint and proposes wider ranges for the selected group. The phase has a maximum of 1,500 PPO updates, with candidate statistics aggregated over the most recent 500 iterations.

For a boundary with accepted value $f_k^\star$ and search limit b, a coarse candidate is $f_{\text{cand}} = \Gamma(f_k^\star, b; \rho_{\text{grow}})$, where $\Gamma(f, b; \rho) = b - (b - f)/\rho$ and $\rho_{\text{grow}} > 1$. This rule initiates expansion before a failed boundary is available. Parameter-specific minimum steps and caps determine the effective proposal, and upper and lower boundaries are tracked separately. Once an attempt fails, the accepted value $f_k^\star$ and failed value f_k^- define a bracket for a finer retry; the superscript $-$ labels the failed side rather than the sign of the physical value.

Candidate feasibility is measured over the recent training window. Retention is measured by the one-rollout retention comparison C_k, which evaluates the current policy under accepted ranges against references from the last accepted checkpoint. Let ℓ_k and e_k denote the averages over this window of the per-update mean normalized episode length and per-update mean tracking error under the candidate ranges, $e_{k|\star}$ and $R_{k|\star}$ the current policy's one-rollout statistics under accepted conditions, and $e^\star, R^\star$ the corresponding checkpoint references. We operationalize training-time recoverability through the conjunction of these checks:

$$\begin{aligned} L_k &: \quad e_k \le \eta_e, \quad \ell_k \ge \eta_\ell, \\ C_k &: \quad e_{k|\star} \le (1+\delta_e)e^\star, \quad R_{k|\star} \ge R^\star - \delta_R|R^\star|, \\ S_k &= L_k C_k, \end{aligned} \tag{1}$$

where Appendix A.3 specifies η_e, η_ℓ, and $\delta_e = \delta_R = 0.05$. L_k checks stable candidate control through separate window-average thresholds, unlike the joint episode-level test used for fixed evaluation. C_k constrains mean-performance degradation on accepted ranges relative to the last accepted checkpoint. Its 5% tolerances apply to this local comparison, not cumulative forgetting over the curriculum. The absolute value $|R^\star|$ handles either return sign. Passing S_k commits the candidate ranges and learning state; rejection triggers restoration and refinement (Section 3.3).

3.3 Learning-State Restoration and Boundary Refinement

After a rejected expansion, HORIZON restores the accepted physical ranges together with the actor, critic, optimizer, learning rate, and enabled observation normalizers. Evaluator windows, actor caches, and PPO rollout buffers are cleared before new collection. The failed boundary remains in the search record, allowing refinement to use the rejected attempt while policy learning resumes from the accepted checkpoint. Table 3 details the checkpoint and reset operations.

For the accepted boundary $f_k^\star$ and failed boundary f_k^-, the default retry is

$$f_{\text{cand}} = f_k^\star + \alpha\big(f_k^- - f_k^\star\big), \qquad \alpha = 0.4. \tag{2}$$

Acceptance advances a boundary; failure tightens its bracket or refines the proposal within configured bounds. The matched budget concerns PPO updates, including Base initialization, failed phases, and retries. Retention rollouts do not update the learner and are excluded from this budget; their simulator interactions and inference time constitute additional evaluation overhead (Appendix B.4). HORIZON finishes the current phase at the budget limit: success returns the new accepted checkpoint; failure restores and returns the previous one. Search also ends when no boundaries remain eligible (Appendix A).

4 Experiments

We first test training-time recoverability through complementary L_k-only and C_k-only gate ablations, phase diagnostics, and matched restoration continuations. We then examine range expansion and domain composition, establishing the Core and Full configurations before comparing training methods and ablations. Hardware trials evaluate transfer under mechanical interventions.

4.1 Evaluation Protocol

We evaluate Base randomization, isolated-group shifts, and joint physical shifts. In OOD-all, all seven groups use the predefined maximum search-limit intervals in Table 10. Direct-wide DR trains at the search limits of the four Core groups from its first update, while contact, inertia, and initial state remain at Base. OOD-all therefore combines these four trained ranges with shifts in the three non-Core groups. The common severity path interpolates each interval from Base to its search limit.

Success requires at least 950 of 1,000 control steps (20 s) and planar mean absolute velocity error below 0.4 m/s, excluding yaw error. Common-severity AUC (CS-AUC) integrates success percentage along the shared Base-to-limit path (0–100). RCS-AUC uses group-specific Base-DR-calibrated paths only in Figure 4a. We also report OOD-all success and Base tracking error. Appendices B.1 and B.3 define the task and metrics; Appendices B–C give evaluation and baseline details.

4.2 Training-Time Recoverability

The matched ablations in Table 1 test candidate feasibility, accepted-range retention, and learning-state restoration. Without C_k, a window-feasible candidate can be accepted despite reduced Base success, showing that L_k alone does not guarantee retention. The two L_k rows compare the latest accepted and budget-end checkpoints from the same continuation runs: the C_k-only branch deteriorates further by the budget endpoint, while the joint gate stays stable. Both branches face identical frozen four-group boundaries beyond their shared accepted ranges, ensuring a common evaluation distribution. Together with the consistent ordering at individual new boundaries, this shows that retention alone does not establish mastery of new ranges.

In the restoration comparisons, both branches retract to the same accepted ranges, and Restore additionally reloads the accepted learning state. Table 1 contrasts actuation collapse (P1) with disturbance without immediate collapse (P2), where Restore improves endpoint CS-AUC by 7.60 and 5.21 percentage points, respectively. With continuation ranges and update budgets matched within each pair, the gains isolate restoration's contribution. P2 extends this evidence to failure without immediate collapse.

Across four additional actuation parent trajectories (P3–P6), Restore improves endpoint CS-AUC in three cases, with a mean paired gain of 2.13 ± 1.86 percentage points. These results support a

Table 1: Matched ablations (%; higher is better). Gate rows report mean ± sample SD; restoration rows report endpoint estimates for the P1–P2 cross-type contrast. Training L_k uses window-average thresholds; evaluation uses the episode-level criterion. Restoration CS-AUC sweeps only the failed group from Base to its search limits, keeping all other groups at Base.

Component / failure case	Evaluation	With	Without
C_k retention	Base success	$\mathbf{96.93 \pm 1.23}$	87.56 ± 1.91
L_k feasibility	Four-group success, latest accepted	$\mathbf{77.11 \pm 0.79}$	63.09 ± 1.95
	Four-group success, budget endpoint	$\mathbf{77.38 \pm 1.11}$	58.11 ± 1.76
Learning-state restoration		Restore	Range-only
Actuation P1 (collapse)	Endpoint CS-AUC	**70.62**	63.02
Disturbance P2 (no collapse)	Endpoint CS-AUC	**80.62**	75.41

recovery benefit beyond the initial actuation case, with variation across failure events; Appendix D.2 reports the complete results, including the -0.38-point difference in P5. The gates assess expansion acceptance, while restoration determines the learner state for retries.

4.3 Learning over Broader Physical Ranges

Having separated the roles of L_k, C_k, and learning-state restoration, we next ask whether recoverable search broadens the training distribution over a complete curriculum. Figure 3a traces all seven physical groups from a single joint transformer curriculum run, while Figure 3b compares its final group-wise coverage with the corresponding Base and MLP configurations. This separates whether a failed group resumes expansion from how far each group ultimately progresses. We quantify the latter using *range reach*, which summarizes accepted training-range coverage rather than control capability. Success is evaluated independently below. Appendix B.2 defines the normalization for two- and one-sided parameters, within-group aggregation, and the treatment of the nonzero Base range.

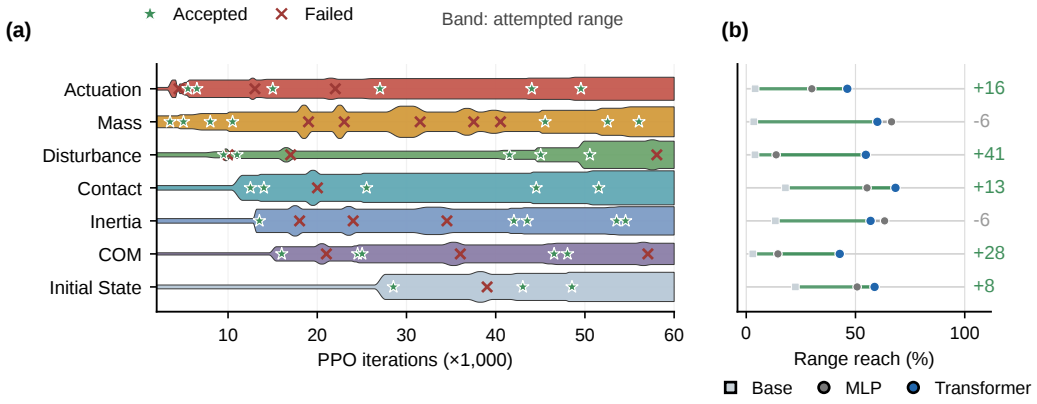


Figure 3: Physical-domain expansion during training. (a) Attempted ranges for all seven groups in a single joint transformer curriculum run. Crosses mark unsuccessful proposals and stars mark accepted boundaries. A later star after a cross shows that an eligible group was revisited and progressed. (b) Group-wise range reach for Base, MLP, and transformer. Right-margin values report transformer-minus-MLP differences in percentage points, whereas Base values reflect initial randomization. Appendix B.2 defines configuration-specific normalization and group aggregation. Range reach measures accepted training coverage, not control success.

An unsuccessful proposal need not end expansion of its group. The actuation trajectory contains failed proposals followed by acceptance of wider ranges, with related patterns in disturbance and COM. The reported range reach rises from 4.1%, 4.0%, and 3.0% under Base to 46.3%, 54.7%, and 42.8% for the transformer curriculum on these three groups. The transformer exceeds the cor-

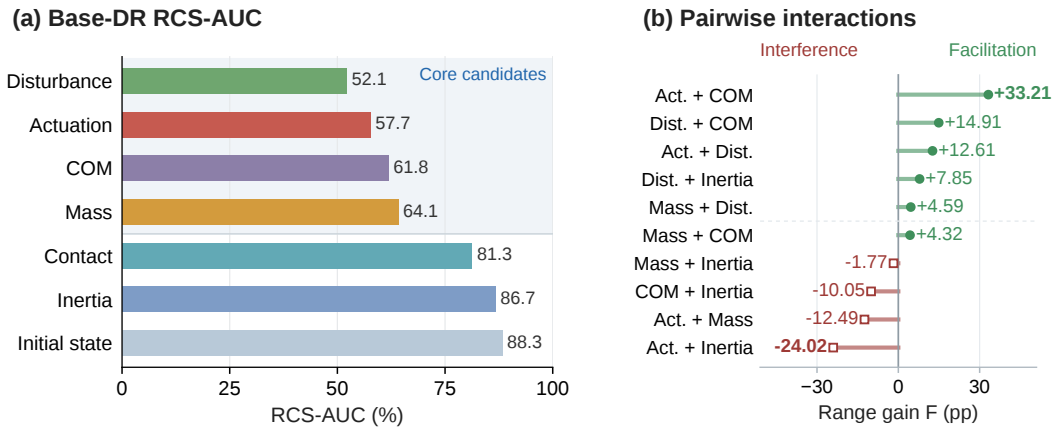


Figure 4: Domain sensitivity and pairwise expansion evidence used to form the Core curriculum. (a) Base-DR RCS-AUC. (b) Pairwise range gain relative to the two single-group references (F, percentage points). The fixed-checkpoint Core–Full sweep over joint combinations is reported in Appendix E.

responding MLP range summaries by approximately 16, 41, and 28 percentage points. Thus the curriculum's progress depends on the physical group and the controller learning under it.

Range expansion alone does not determine success under a new joint condition. We therefore evaluate isolated robustness separately and use the resulting group sensitivity to guide the composition study.

4.4 Domain Composition and Joint Generalization

Figure 4a measures Base-DR's vulnerability within the tested ranges using RCS-AUC. Disturbance, actuation, COM, and mass are more challenging than contact, inertia, and initial state, marking them as candidate groups for curriculum expansion.

In the pairwise development diagnostics (Figure 4b), actuation–COM gains 33.21 percentage points in mean range reach and 13.35 points in mean isolated-group success over single-group training. Effects depend on the partner: actuation–inertia reduces mean range reach, whereas disturbance–inertia increases it. Because actuation remains limiting in the intermediate actuation–mass–disturbance curriculum, the favorable actuation–COM and COM–disturbance development results motivate adding COM. This progression yields the four-group *Core*: actuation, mass, disturbance, and COM. The other three groups remain at Base. Appendix E details the selection progression and full pairwise results.

Full expands all seven groups. Under targeted Core extensions with contact, inertia, or initial-state shifts, Core leads Full by 10.0–12.6 percentage points (Table 22). A fixed-checkpoint sweep over 91 joint OOD combinations of three to five active groups favors Core on 57 combinations, with mean success of 48.94% versus 45.57% (Table 23). Within this fixed-checkpoint sweep, the mean advantage increases from 0.88 to 3.84 and 6.75 percentage points as the number of active groups rises from three to four and five. The comparison across five training seeds also favors Core by 8.7 points in seven-group OOD-all success and 3.4 points in CS-AUC (Table 2).

Core improves isolated success on its four expanded groups relative to Full and retains 91.0–94.5% success on the other three, where Full scores higher (Table 12). Full's stronger isolated performance on the additional groups does not translate into better joint control, supporting evaluation of domain composition through combined shifts. Core uses approximately 45,000 PPO updates, and Full approximately 60,000 (Table 11). Thus broader coverage need not improve joint generalization under the tested controller and training allocations.

Table 2: Method comparison and ablations. Entries report mean ± sample SD across five independent training seeds, except Linear, whose SD is across five evaluation seeds of one training run. OOD-all success and CS-AUC are percentages, and lower Base tracking error is better. † Linear uses its last successful pre-expansion checkpoint. Appendix Table 11 specifies cost accounting and returned-policy selection.

Method	OOD-all success ↑	Base tracking ↓	CS-AUC ↑
HORIZON (Core)	$\mathbf{36.8 \pm 2.5}$	0.106 ± 0.002	$\mathbf{62.5 \pm 1.2}$
Base-DR	17.4 ± 1.1	$\mathbf{0.076 \pm 0.001}$	41.4 ± 0.7
Direct-wide DR	10.3 ± 0.8	0.079 ± 0.001	44.9 ± 0.7
OpenAI ADR (OpenAI et al., 2019)	19.8 ± 1.9	0.206 ± 0.002	61.5 ± 0.6
HORIZON + Active-DR (Mehta et al., 2020)	14.7 ± 0.7	0.362 ± 0.003	28.7 ± 0.4
Linear†	22.0 ± 1.9	0.196 ± 0.004	51.4 ± 0.8
DORAEMON (Tiboni et al., 2024)	18.9 ± 0.9	0.182 ± 0.001	54.8 ± 1.0
GRAM (Queeney et al., 2026)	22.7 ± 1.1	0.407 ± 0.004	34.3 ± 0.9
Ablations			
HORIZON (Full)	28.1 ± 1.8	0.127 ± 0.002	59.1 ± 0.4
Multi-expert distillation	16.5 ± 1.2	0.233 ± 0.004	45.3 ± 1.5
w/o rollback	26.3 ± 2.6	0.215 ± 0.007	60.4 ± 1.1

4.5 Method Comparison and Ablation

Table 2 compares HORIZON (Core) with fixed and scheduled randomization, adaptive distribution methods, and history adaptation, alongside Full, distillation, and rollback ablations. Except for Full, methods randomizing beyond Base use the same four Core groups; the other three remain at Base. The *w/o rollback* variant keeps HORIZON's gates, range retraction, and retry refinement but retains the failed phase's learning state. Appendix C gives the detailed baseline and ablation designs, training budgets, and checkpoint-selection rules.

Core has the highest mean OOD-all success and CS-AUC. Its CS-AUC is close to OpenAI ADR (62.5% versus 61.5%), but OOD-all success is 17.0 points higher (36.8% versus 19.8%) and Base tracking error is lower (0.106 versus 0.206). Similar integrated success therefore need not imply comparable performance at the strongest joint shift. Direct-wide trains at the four Core groups' search limits from the outset, yet its lower OOD-all success than Core shows that broad exposure alone does not ensure strong control at the joint test endpoint. Base-DR and Direct-wide achieve lower Base tracking error, indicating a nominal-tracking trade-off.

Without rollback, OOD-all success is $26.3 \pm 2.6\%$, Base tracking error 0.215 ± 0.007, and CS-AUC $60.4 \pm 1.1\%$. Table 2 evaluates the complete training-and-return procedure: Core returns its latest accepted checkpoint, whereas the ablation returns the terminal learner. Matched-range, matched-update continuations in Appendix D.2 isolate restoration's effect on subsequent learning. Distillation reaches $16.5 \pm 1.2\%$ joint success versus Core's $36.8 \pm 2.5\%$, favoring joint curriculum training over this procedure without isolating state-distribution coverage as the cause (Appendix C.7).

4.6 Hardware Transfer

We deploy the policy on a Unitree Go2 quadruped under joint constraints, leg extensions, payload shifts, and coupled mechanical interventions. Figure 5 shows knee locking, four-leg extensions, hind-leg extensions with payload, moving payloads, combined loading and extensions, and hind-knee locking. These trials use a fixed-weight history-conditioned controller, without privileged physical parameters or online policy optimization. The repeated-trial comparison uses fifteen trials per policy in each of four conditions: 3 kg payload, 5 kg payload, a locked front-left leg, and 15 cm extensions on all four legs. Each trial begins with 5 s of balance and continues for 15 s under $(v_x, v_y, \omega_z) = (0.3, 0, 0)$; completing both stages without intervention counts as success. Failures during initial balance also enter the denominator.

HORIZON maintains initial balance in 57 of 60 trials and completes commanded locomotion in 49 of 60, compared with 31 of 60 and 0 of 60 for Base-DR. Thus, eight HORIZON trials pass the balance stage but do not complete locomotion; none of the 31 Base-DR trials that pass bal-



Figure 5: Motor-control transfer under mechanical interventions. The annotated 2 s transient in (a) reflects history-conditioned inference with fixed network weights, not online policy optimization or explicit system identification. Repeated-trial success is evaluated separately using the 20 s protocol described in the text.

ance complete it. HORIZON succeeds in 15/15 trials for each payload, 8/15 with the front-left leg locked, and 11/15 with four-leg extensions. Of its eleven failures, locking accounts for two delayed falls and five uncontrolled-motion outcomes; extensions account for one immediate imbalance and three uncontrolled-motion outcomes. This distinguishes complete payload success from the remaining failures under joint and leg-geometry changes. Appendix F gives the condition definitions and Base-DR failure outcomes. The additional moving-payload and coupled-intervention demonstrations remain outside the 120-trial denominator, while cross-robot simulation rollouts provide a separate transfer illustration in Appendix F.2.

5 Conclusion

Physical-domain generalization in on-policy RL depends on both the reach of the training distribution and the learner's ability to keep producing useful experience as that distribution expands. HORIZON turns this observation into checkpointed physical-domain expansion by coupling every range proposal to retention under accepted conditions and restoring the accepted learning state after a failed proposal. The no-rollback ablation's CS-AUC remains close in magnitude to HORIZON's, but it has lower OOD-all success and higher Base tracking error, while the matched failure continuations show that retracting the sampling range does not undo learner updates accumulated during a rejected phase. The evidence therefore supports treating training-time recoverability as a property of the joint evolution of the physical distribution and learner state, rather than as a criterion on range width alone.

The development diagnostics motivate the four-group Core. Under the evaluated training allocations, Core achieves higher success on the tested joint shifts than Full, showing that broader physical coverage need not improve joint generalization. The tested distillation procedure also falls short of joint curriculum training, while fixed-weight hardware transfer connects the resulting robustness to concrete mechanical interventions. Together, these findings connect two decisions: accepting expansions that retain acquired performance within the specified tolerances, and choosing variations to learn jointly. Both inform physical-domain generalization in the tested quadruped setting.

AI USE STATEMENT

Generative AI tools assisted research planning, code preparation, manuscript drafting and revision, and figure and table preparation. This includes organizing the experimental questions, editing mathematical and procedural descriptions, and preparing LaTeX and plotting code. The authors are responsible for the methods, evidence, references, and final content.

REPRODUCIBILITY STATEMENT

Appendix A, including Algorithm 1, describes range selection, candidate evaluation, state restoration, and refinement. Appendix B specifies physical ranges, episode success, cost accounting, and statistical aggregation. Baseline interfaces, mechanism diagnostics, composition protocols, and hardware counts are given in the remaining appendices. Upon acceptance, we will publicly release the training and evaluation code, configuration files, and reproduction instructions.

REFERENCES

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 41–48. ACM, 2009. doi: 10.1145/1553374.1553380. URL https://doi.org/10.1145/1553374.1553380.

Yevgen Chebotar, Ankur Handa, Viktor Makoviychuk, Miles Macklin, Jan Issac, Nathan Ratliff, and Dieter Fox. Closing the sim-to-real loop: Adapting simulation randomization with real world experience. In *2019 International Conference on Robotics and Automation (ICRA)*, pp. 8973–8979, 2019. doi: 10.1109/ICRA.2019.8793789.

Ioannis Exarchos, Yifeng Jiang, Wenhao Yu, and C. Karen Liu. Policy transfer via kinematic domain randomization and adaptation. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 45–51, 2021. doi: 10.1109/ICRA48506.2021.9561982. URL https://ieeexplore.ieee.org/document/9561982.

Carlos Florensa, David Held, Markus Wulfmeier, Michael Zhang, and Pieter Abbeel. Reverse curriculum generation for reinforcement learning. In *Proceedings of the 1st Annual Conference on Robot Learning*, volume 78 of *Proceedings of Machine Learning Research*, pp. 482–495. PMLR, 2017. URL https://proceedings.mlr.press/v78/florensa17a.html.

Minqi Jiang, Michael Dennis, Jack Parker-Holder, Jakob Foerster, Edward Grefenstette, and Tim Rocktäschel. Replay-guided adversarial environment design. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/hash/0e915db6326b6fb6a3c56546980a8c93-Abstract.html.

Khimya Khetarpal, Matthew Riemer, Irina Rish, and Doina Precup. Towards continual reinforcement learning: A review and perspectives. *Journal of Artificial Intelligence Research*, 75:1401–1476, 2022. doi: 10.1613/jair.1.13673. URL https://www.jair.org/index.php/jair/article/view/13673.

Ashish Kumar, Zipeng Fu, Deepak Pathak, and Jitendra Malik. RMA: Rapid motor adaptation for legged robots. In *Proceedings of Robotics: Science and Systems*, Virtual, July 2021. doi: 10.15607/RSS.2021.XVII.011. URL https://www.roboticsproceedings.org/rss17/p011.html.

Joonho Lee, Jemin Hwangbo, Lorenz Wellhausen, Vladlen Koltun, and Marco Hutter. Learning quadrupedal locomotion over challenging terrain. *Science Robotics*, 5(47):eabc5986, 2020. doi: 10.1126/scirobotics.abc5986.

Min Liu, Deepak Pathak, and Ananye Agarwal. Locoformer: Generalist locomotion via long-context adaptation. In *Proceedings of The 9th Conference on Robot Learning*, volume 305 of *Proceedings of Machine Learning Research*, pp. 532–546. PMLR, 2025. URL https://proceedings.mlr.press/v305/liu25a.html.

Bhairav Mehta, Manfred Diaz, Florian Golemo, Christopher J. Pal, and Liam Paull. Active domain randomization. In *Proceedings of the Conference on Robot Learning*, volume 100 of *Proceedings of Machine Learning Research*, pp. 1162–1176. PMLR, 2020. URL https://proceedings.mlr.press/v100/mehta20a.html.

Takahiro Miki, Joonho Lee, Jemin Hwangbo, Lorenz Wellhausen, Vladlen Koltun, and Marco Hutter. Learning robust perceptive locomotion for quadrupedal robots in the wild. *Science Robotics*, 7(62):eabk2822, 2022. doi: 10.1126/scirobotics.abk2822.

Mayank Mittal, Pascal Roth, James Tigue, Antoine Richard, Octi Zhang, Peter Du, Antonio Serrano-Muñoz, Xinjie Yao, René Zurbrügg, Nikita Rudin, Lukasz Wawrzyniak, Milad Rakhsha, Alain Denzler, Eric Heiden, Ales Borovicka, Ossama Ahmed, Iretiayo Akinola, Abrar Anwar, Mark T. Carlson, Ji Yuan Feng, Animesh Garg, Renato Gasoto, Lionel Gulich, Yijie Guo, M. Gussert, Alex Hansen, Mihir Kulkarni, Chenran Li, Wei Liu, Viktor Makoviychuk, Grzegorz Malczyk, Hammad Mazhar, Masoud Moghani, Adithyavairavan Murali, Michael Noseworthy, Alexander Poddubny, Nathan Ratliff, Welf Rehberg, Clemens Schwarke, Ritvik Singh, James Latham Smith, Bingjie Tang, Ruchik Thaker, Matthew Trepte, Karl Van Wyk, Fangzhou Yu, Alex Millane, Vikram Ramasamy, Remo Steiner, Sangeeta Subramanian, Clemens Volk, CY Chen, Neel Jawale, Ashwin Varghese Kuruttukulam, Michael A. Lin, Ajay Mandlekar, Karsten Patzwaldt, John Welsh, Huihua Zhao, Fatima Anes, Jean-Francois Lafleche, Nicolas Moënne-Loccoz, Soowan Park, Rob Stepinski, Dirk Van Gelder, Chris Amevor, Jan Carius, Jumyung Chang, Anka He Chen, Pablo de Heras Ciechomski, Gilles Daviet, Mohammad Mohajerani, Julia von Muralt, Viktor Reutskyy, Michael Sauter, Simon Schirm, Eric L. Shi, Pierre Terdiman, Kenny Vilella, Tobias Widmer, Gordon Yeoman, Tiffany Chen, Sergey Grizan, Cathy Li, Lotus Li, Connor Smith, Rafael Wiltz, Kostas Alexis, Yan Chang, David Chu, Linxi "Jim" Fan, Farbod Farshidian, Ankur Handa, Spencer Huang, Marco Hutter, Yashraj Narang, Soha Pouya, Shiwei Sheng, Yuke Zhu, Miles Macklin, Adam Moravanszky, Philipp Reist, Yunrong Guo, David Hoeller, and Gavriel State. Isaac lab: A gpu-accelerated simulation framework for multi-modal robot learning. *arXiv preprint arXiv:2511.04831*, 2025. doi: 10.48550/arXiv.2511.04831. URL https://arxiv.org/abs/2511.04831.

Fabio Muratore, Theo Gruner, Florian Wiese, Boris Belousov, Michael Gienger, and Jan Peters. Neural posterior domain randomization. In *Proceedings of the 5th Conference on Robot Learning*, volume 164 of *Proceedings of Machine Learning Research*, pp. 1532–1542. PMLR, 2022. URL https://proceedings.mlr.press/v164/muratore22a.html.

Sanmit Narvekar, Bei Peng, Matteo Leonetti, Jivko Sinapov, Matthew E. Taylor, and Peter Stone. Curriculum learning for reinforcement learning domains: A framework and survey. *Journal of Machine Learning Research*, 21(181):1–50, 2020. URL https://www.jmlr.org/papers/v21/20-212.html.

OpenAI, Ilge Akkaya, Marcin Andrychowicz, Maciek Chociej, Mateusz Litwin, Bob McGrew, Arthur Petron, Alex Paino, Matthias Plappert, Glenn Powell, Raphael Ribas, Jonas Schneider, Nikolas Tezak, Jerry Tworek, Peter Welinder, Lilian Weng, Qiming Yuan, Wojciech Zaremba, and Lei Zhang. Solving rubik's cube with a robot hand, 2019. URL https://arxiv.org/abs/1910.07113.

Xue Bin Peng, Marcin Andrychowicz, Wojciech Zaremba, and Pieter Abbeel. Sim-to-real transfer of robotic control with dynamics randomization. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3803–3810, 2018. doi: 10.1109/ICRA.2018.8460528.

Rémy Portelas, Cédric Colas, Katja Hofmann, and Pierre-Yves Oudeyer. Teacher algorithms for curriculum learning of deep RL in continuously parameterized environments. In *Proceedings of the Conference on Robot Learning*, volume 100 of *Proceedings of Machine Learning Research*, pp. 835–853. PMLR, 2020. URL https://proceedings.mlr.press/v100/portelas20a.html.

James Queeney, Xiaoyi Cai, Alexander Schperberg, Radu Corcodel, Mouhacine Benosman, and Jonathan P. How. GRAM: Generalization in deep RL with a robust adaptation module. *IEEE Robotics and Automation Letters*, 11(2):2218–2225, 2026. doi: 10.1109/LRA.2025.3641155. URL https://www.merl.com/publications/TR2026-057.

Fabio Ramos, Rafael Carvalhaes Possas, and Dieter Fox. BayesSim: Adaptive domain randomization via probabilistic inference for robotics simulators. In *Robotics: Science and Systems*, 2019. doi: 10.15607/RSS.2019.XV.029.

Nikita Rudin, David Hoeller, Philipp Reist, and Marco Hutter. Learning to walk in minutes using massively parallel deep reinforcement learning. In *Proceedings of the 5th Conference on Robot Learning*, volume 164 of *Proceedings of Machine Learning Research*, pp. 91–100. PMLR, 2022. URL https://proceedings.mlr.press/v164/rudin22a.html.

Andrei A. Rusu, Sergio Gomez Colmenarejo, Caglar Gulcehre, Guillaume Desjardins, James Kirkpatrick, Razvan Pascanu, Volodymyr Mnih, Koray Kavukcuoglu, and Raia Hadsell. Policy distillation. In *International Conference on Learning Representations*, 2016. URL https://mlanthology.org/iclr/2016/rusu2016iclr-policy/.

Clemens Schwarke, Mayank Mittal, Nikita Rudin, David Hoeller, and Marco Hutter. Rsl-rl: A learning library for robotics research. *arXiv preprint arXiv:2509.10771*, 2025. doi: 10.48550/arXiv.2509.10771. URL https://arxiv.org/abs/2509.10771.

Jie Tan, Tingnan Zhang, Erwin Coumans, Atil Iscen, Yunfei Bai, Danijar Hafner, Steven Bohez, and Vincent Vanhoucke. Sim-to-real: Learning agile locomotion for quadruped robots. In *Robotics: Science and Systems*, 2018. doi: 10.15607/RSS.2018.XIV.010.

Gabriele Tiboni, Pascal Klink, Jan Peters, Tatiana Tommasi, Carlo D'Eramo, and Georgia Chalvatzaki. Domain randomization via entropy maximization. In *International Conference on Learning Representations*, volume 2024, pp. 19841–19863, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/56adf9cb91aedfa41ce24398782a012f-Paper-Conference.pdf.

Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 23–30, 2017. doi: 10.1109/IROS.2017.8202133.

Zhaoming Xie, Xingye Da, Michiel van de Panne, Buck Babich, and Animesh Garg. Dynamics randomization revisited: A case study for quadrupedal locomotion. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 4955–4961, 2021. doi: 10.1109/ICRA48506.2021.9560837.

A Algorithm and Training Details

Algorithm 1 HORIZON: checkpointed physical-domain expansion

Require: Active groups, Base range $\mathcal{D}^0$, search limits, PPO-update budget
1: $\mathcal{D}^{\mathrm{acc}} \leftarrow \mathcal{D}^0$; train on Base; save a checkpoint; initialize failure brackets and a seed-shuffled group queue
2: **while** an eligible group and remaining budget exist **do**
3: Select a group; propose $\mathcal{D}^{\mathrm{cand}}$ from $\mathcal{D}^{\mathrm{acc}}$ and its failed boundary
4: Reset phase buffers; complete one bounded PPO phase, even if the budget is reached within it
5: Evaluate window feasibility L_k and the one-rollout retention comparison C_k
6: **if** $S_k = L_k C_k = 1$ **then**
7: $\mathcal{D}^{\mathrm{acc}} \leftarrow \mathcal{D}^{\mathrm{cand}}$; save the new checkpoint
8: **else**
9: Record failed boundaries and restore the checkpoint
10: **end if**
11: Reset transition buffers; requeue groups with unfinished boundaries
12: **end while**
13: **return** the latest accepted policy after the final phase decision

A.1 Learning State, Search State, and Rollout Resets

The checkpointed cycle maintains two kinds of information. Learning state contains the actor and critic parameters and the optimizer state used to update them. Search state contains accepted group intervals, parameter-side stopping flags, failed boundaries, pending retries, and group-selection state. Accepting a candidate records the updated learned parameters together with the accepted ranges. Failure restores the saved learning state and accepted ranges, while preserving the newly observed failed boundary for the next proposal. Table 3 distinguishes restoration from resetting rollout state and from accumulating cost.

Table 3: State handling in checkpointed expansion. Restoration replaces learning state with the saved state; reset clears transient rollout information. Cost and the latest failed proposal are retained across restoration.

State	On acceptance	After failure
Actor and critic parameters	Save	Restore saved parameters
Optimizer moments and learning rate	Save	Restore saved optimizer state
Accepted parameter ranges	Save accepted candidate	Restore accepted intervals
Failed boundary and retry information	Update search record	Keep latest failure and form retry
Group queue and refinement counters	Record	Restore the queue, then merge failure counters and the pending retry
Observation normalization, if enabled	Save	Restore actor and critic normalizer states
Evaluator windows	Store reference statistics	Clear candidate-window statistics before the next phase
Actor temporal caches	Transient rollout state	Clear for the new rollout phase
PPO rollout storage	Consumed by updates	Clear before new collection
Simulator state and commands	Save a snapshot	Restore the baseline snapshot, then reset and resample
Random-number generator states	Save	Restore checkpoint state before reset and resampling
Cumulative interactions / elapsed time	Accumulate	Continue accumulating
Saved checkpoint iteration	Record index	Keep separate from cumulative cost

The training budget is checked between phases. If it is reached within a phase, HORIZON completes that phase and evaluates S_k before stopping. A successful final phase returns its newly accepted

checkpoint; a failed final phase restores and returns the preceding accepted checkpoint. Thus the returned HORIZON policy has passed the joint gate. The nominal update budget may be exceeded by the remainder of one phase, whose cap is 1,500 updates; all these updates, including unsuccessful candidates, count toward the realized training cost. On failure, HORIZON restores the checkpointed simulator, command, and random-number states and then performs a fresh transition reset under the accepted ranges. The reset and resampling consume the restored random streams, so rollback defines a reproducible recovery transition rather than exact replay of the previously observed trajectory.

A.2 Group Selection, Bilateral Proposals, and Termination

The group queue is shuffled using the selection seed. Each selection advances the unfinished boundaries of one group; the other groups keep the ranges specified in Section 3.1. Groups with unfinished boundaries remain eligible for later queue cycles.

The coarse proposal for an accepted boundary $f_k^\star$ can be written as a contraction of its distance to its allowed limit b,

$$f_{\text{cand}} = \Gamma(f_k^\star, b; \rho_{\text{grow}}) \equiv b - \frac{b - f_k^\star}{\rho_{\text{grow}}}, \qquad \rho_{\text{grow}} > 1. \tag{3}$$

This expression moves either an upper or a lower boundary toward its respective limit. Upper and lower sides are tracked separately, and parameter-specific minimum steps, caps, and refinement factors determine the effective proposal. The growth factor is not a multiplier on a signed physical value. In particular, changing ρ_{grow} by 20% does not mean multiplying $f_k^\star$ by 1.2.

Table 4 gives one recorded multi-parameter actuation update from a curriculum trajectory using $\rho_{\text{grow}} = 1.4$. The upper effort limit moves from 46.96 to 56.40, with an allowed upper limit of 80; Equation 3 produces that update. Velocity, stiffness, and damping also change within the same group proposal, while the lower boundaries remain unchanged. The example illustrates the parameter geometry and is not included in the aggregate performance results.

Table 4: Recorded actuation interval update. Values are rounded for display; all four upper boundaries change within one group. Effort, velocity, stiffness, and damping retain their physical units in the simulator.

Parameter	Lower bound	Previous upper bound	Proposed upper bound
Effort limit	31.2426	46.9600	56.4000
Velocity limit	21.2426	35.4318	42.4513
Stiffness	31.2426	46.9600	56.4000
Damping	0.6959	1.5432	2.2451

After failure, the stored bracket supplies the retry in Equation 2. Acceptance advances the accepted side; rejection tightens the failed side or advances local refinement. Table 5 distinguishes stopping a parameter side, completing a group, and ending the search. A failed proposal alone does not complete a group.

Table 5: Stopping logic for boundary refinement, group selection, and global search.

Scope	Stopping condition	Action
Parameter side	Reaches its search limit or meets its configured resolution/failure stopping condition	Stop proposing on this side
Group	All relevant parameter sides have stopped	Remove from the eligible queue
Global search	No eligible groups remain, or the PPO-update budget has been reached at a phase boundary	Apply the final gate and return the latest accepted policy

A.3 Curriculum Evaluation and Aggregation

The window-feasibility check L_k separately averages the per-update mean normalized episode length and per-update mean tracking error over the most recent 500 iterations, with a phase cap of 1,500 PPO updates. It requires the two window averages to be at least 0.85 and at most 0.5, respectively. This training-window gate is distinct from the joint per-episode success criterion used for fixed evaluation. The one-rollout retention comparison C_k evaluates the current policy under accepted conditions against the last accepted checkpoint, using $\delta_e = \delta_R = 0.05$ in Equation 1. Their product S_k determines acceptance.

The diagnostic return ratio for C_k is $q_k^R = R_{k|\star}/R^\star$, recorded as `compare_reward_ratio` (`mastery_relative_compare_reward_ratio` in phase logs) when $R^\star > 10^{-6}$. Here $R_{k|\star}$ and $R^\star$ are the current and checkpoint rollout return statistics, respectively, exposed by the evaluator as `mean_reward`. For these positive references, the return condition in C_k is equivalent to $q_k^R \geq 1 - \delta_R$; passing C_k additionally requires its tracking condition. For zero or negative references, the absolute-value tolerance in Equation 1 remains the rule, and this diagnostic ratio is not reported.

Each one-rollout retention comparison evaluates the current and accepted-checkpoint policies for at most 1,000 simulator steps in the parallel training environments under the accepted ranges. Both policies begin from the same restored simulator, command, and random-number state. For each environment, return accumulates per-step rewards clipped below at -5 until its first termination or the rollout horizon, after which returns are averaged across environments. Tracking error is averaged over valid steps within each environment and then across environments. Environments without a termination contribute their horizon-truncated return and tracking statistics.

Table 6: Curriculum and evaluation settings. The training checks and final episode success serve different purposes.

Setting	Value	Role
KV-cache window	32 steps	Actor temporal context
Maximum phase length	1,500 updates	Candidate training budget
Candidate window	500 iterations	Training aggregation
Retry interpolation α	0.4	Accepted-to-failed boundary interpolation
L_k length/error thresholds	$0.85/0.5$	Candidate-window feasibility
Episode-success criterion	Section 4.1	Fixed evaluation
Retention tracking tolerance δ_e (C_k)	5%	Retention comparison
Retention return tolerance δ_R (C_k)	5%	Retention comparison

A.4 Actor, Critic, and PPO

The actor is a two-layer decoder-only transformer using causal attention, rotary position embeddings, and a KV-cache window of 32 steps. During rollout and deployment, each step supplies the current proprioceptive observation, command, and previous action. Temporal attention uses cached keys and values, without requiring an explicitly stacked history of raw observations as input. Simulator-only information is provided to the critic but not to the deployed actor. Temporal caches are cleared on resets and curriculum transitions. At deployment the network weights are fixed; changes in history-conditioned actions do not constitute online gradient updates or checkpoint restoration.

The comparisons sharing the HORIZON learner use identical observation components, normalization, and action-to-motor mappings. The HORIZON actor receives no privileged actuation parameters when motor ranges expand. GRAM retains its native training context, with the exposed training quantities and inference-time history adapter described in Appendix C.

A.5 Reward Function

The reward combines command tracking with regularization of posture, motion, and actuation. Table 9 lists the terms and weights used in the experiments. Disabled and zero-weight terms do not

Table 7: Policy and critic settings.

Parameter	Value
Policy class	decoder-only transformer
Initial action noise std	1.2
Activation	GELU
Embedding dimension	128
Transformer layers	2
Attention heads	8
Feedforward dimension	512
KV-cache window (steps)	32
RoPE base	10000
Critic hidden dimensions	$[512, 256, 128]$
Privileged encoder hidden dimension	64

Table 8: PPO settings.

Parameter	Value
Steps per environment	24
Value loss coefficient	1.0
Clipped value loss	enabled
Clip range	0.2
Entropy coefficient	0.01
Learning epochs	8
Minibatches	4
Learning rate	5×10^{-5}
Learning-rate schedule	adaptive
Discount factor γ	0.99
GAE λ	0.95
Target KL	0.01
Max gradient norm	1.0

contribute. Reward is a training quantity and is not substituted for the episode-success criterion. Cross-method return comparisons require matching reward terms and scaling.

Table 9: Reward terms and weights.

Name	Term ϕ_i	w_i
Linear velocity	$\exp(-\|c_{xy} - v_{xy}\|_2^2/\sigma^2)$	1.25
Yaw velocity	$\exp(-(c_{\omega_z} - \omega_z)^2/\sigma^2)$	1.25
Vertical velocity	v_z^2	-2.0
Roll/pitch angular velocity	$\|\omega_{xy}\|_2^2$	-0.05
Action rate	$\|a_t - a_{t-1}\|_2^2$	-0.02
Feet air time	$\sum_j (T_j^{\text{air}} - 0.5)$	0.2
Flat orientation	$\|g_{xy}^{\text{proj}}\|_2^2$	-2.5
Hip deviation	$\|q - q_0\|_1$	-0.4
Thigh/calf deviation	$\|q - q_0\|_1$	-0.04
Adaptive base height	$(z - z^\star)^2$	-5.0
Normalized action smoothness	$\|\bar{a}_t - \bar{a}_{t-1}\|_2$	-0.02
Normalized joint power	$\sum_j \|\tau_j \dot{q}_j\|/m$	-2.0×10^{-5}
Normalized torque	$\|\bar{\tau}\|_2^2$	-2.0×10^{-4}
Normalized joint acceleration	$\|\ddot{\bar{q}}\|_2^2$	-2.5×10^{-7}

B Physical Ranges and Evaluation

B.1 Base, Search Limits, and Actual Exposure

Task and procedural bodies. The task is velocity-command tracking by procedurally generated quadrupeds with twelve actuated joints, using Go2-derived geometry. At scene construction, each environment index seeds a body sample that remains fixed across episode resets. Relative to the reference geometry, body length and total leg length are sampled uniformly from $[0.84, 1.16]$, body width and height from $[0.88, 1.12]$, and thigh/calf thickness factors from $[0.85, 1.15]$/$[0.90, 1.10]$; the thigh fraction of total leg length has a uniform offset in $[-0.12, 0.12]$. Link masses and inertias follow the geometry scaling before physical randomization. Training and fixed evaluation use the same procedural family; the seven-group tests vary physical conditions rather than define a held-out morphology split. Physics runs at 200 Hz with control at 50 Hz. Commands are resampled every 10 s, with $v_x \in [-1, 1]$ and $v_y \in [-0.5, 0.5]$ m/s; yaw commands use heading-error feedback with gain 0.5 and clipping to $[-1, 1]$ rad/s, with 5% standing commands. Training uses generated terrain, while the common fixed physical-shift evaluation uses flat terrain.

Table 10 specifies the physical intervals used to define Base and the allowed search space. An accepted interval generally occupies only part of that space. In Core, the three non-expanded groups retain the Base intervals. Actual training exposure additionally depends on when a randomization event is applied: a startup event, an episode reset, and an interval disturbance need not sample the same values at the same times.

The seven-group evaluation defines the initial-state group through initial joint position and velocity variation. Knee-joint-limit randomization and additional push events remain separately named auxiliary conditions and do not enter the seven-group test suite.

B.2 Normalized Range Reach

Range reach normalizes the accepted parameter ranges by the fixed search limits configured for each trajectory. For a two-sided parameter j with nominal reference c_j, accepted interval $[l_j, u_j]$, and search limits $[L_j, U_j]$, the score is

$$r_j = \frac{(u_j - c_j) + (c_j - l_j)}{(U_j - c_j) + (c_j - L_j)} = \frac{u_j - l_j}{U_j - L_j}. \tag{4}$$

Table 10: Physical randomization groups and experimental ranges. Base defines near-nominal variation; search limits constrain candidate proposals. Parameter values within a cell follow the parameter order in that row.

Group	Parameters	Base range	Search limit
Actuation	effort, velocity, stiffness, damping	$[39, 41]$, $[29, 31]$, $[39, 41]$, $[0.9, 1.1]$	$[20, 80]$, $[10, 60]$, $[20, 80]$, $[0.4, 4.0]$
Mass	rigid-body mass scale	$[0.9, 1.1]$	$[0.4, 5.0]$
Disturbance	push velocity, external force, external torque	$[-0.5, 0.5]$, $[-0.02, 0.02]$, 0.0	$[-5.0, 5.0]$, $[-1.0, 1.0]$, $[-0.5, 0.5]$
Contact	static/dynamic friction, restitution	$[0.4, 2.0]$, $[0.4, 2.0]$, 0.0	$[0.05, 6.0]$, $[0.05, 6.0]$, 1.0
Inertia	inertia scale	$[0.9, 1.1]$	$[0.5, 2.0]$
COM	base COM offset	$[-0.003, 0.003]$	$[-0.3, 0.3]$
Initial state	initial joint position/velocity	$[0.9, 1.1]$, $[-0.5, 0.5]$	$[0.5, 1.5]$, $[-2.0, 2.0]$

The equality applies when $L_j \le l_j \le c_j \le u_j \le U_j$. The implementation clips each attained side distance to its available extent and excludes sides with zero available extent. For an increase-only parameter, $r_j = (u_j - c_j)/(U_j - c_j)$, and for a decrease-only parameter, $r_j = (c_j - l_j)/(c_j - L_j)$, each clipped to $[0, 1]$. Thus the denominator is the available physical extent from the nominal reference to the configured limits. Numerator and denominator have the same physical units, making r_j dimensionless.

For group G with parameter weights w_j, the reported percentage is $100 \sum_{j \in G} w_j r_j / \sum_{j \in G} w_j$. Equal weights give the arithmetic mean. Base reach uses the initial randomized interval in the same expression and can therefore be nonzero. Range reach summarizes parameter extents rather than the volume of the joint parameter space. Differences between group percentages, including the pairwise quantity F, are reported in percentage points. The range-reach figures use the fixed limits of their respective configurations, whereas common-severity evaluation uses the intervals in Table 10.

B.3 Common Severity and Episode Success

Episode metric. For valid recorded steps V_i of episode i, the fixed evaluator computes

$$e_i = \frac{1}{|V_i|} \sum_{t \in V_i} \frac{|c_{x,t} - v^b_{x,t}| + |c_{y,t} - v^b_{y,t}|}{2}, \tag{5}$$

where $\mathbf{c}_t$ is the velocity command and $\mathbf{v}^b_t$ is base velocity in the body frame. Thus e_i is planar mean absolute velocity error in m/s; yaw error is not included. The episode limit is $L_{\max} = 1{,}000$ control steps (20 s), so success requires at least 950 steps and $e_i < 0.4$ m/s. Reported tracking error averages e_i across valid episodes. These fixed-evaluation statistics are distinct from the training-window and retention-comparison statistics.

For parameter j, write its Base interval as $[l^0_j, u^0_j]$ and its common limit interval as $[l^1_j, u^1_j]$. The interval at severity s is

$$[l_j(s), u_j(s)] = [(1 - s)l^0_j + sl^1_j,\ (1 - s)u^0_j + su^1_j]. \tag{6}$$

The seven-group severity test uses this construction for all groups simultaneously. An isolated-group test uses the limit interval for one group and Base for the others. A joint-subset test uses limit intervals for the named groups and Base for all remaining groups. The command, terrain, horizon, and randomization-event definitions are fixed across policies within each comparison.

Episode success follows the unified definition in Section 4.1. The common evaluator averages tracking error over the episode's recorded valid steps and applies the same denominator and invalidity rule to every method. Falls and tracking failures count as valid control outcomes. Infrastructure-invalid episodes are recorded separately. This criterion is fixed across all reported simulation evaluations; the training-time L_k check instead uses the window averages described in Appendix A.3.

Let p_i be the success percentage at severity s_i. CS-AUC is

$$\text{CS-AUC} = \frac{1}{s_m - s_0} \sum_{i=0}^{m-1} (s_{i+1} - s_i) \frac{p_i + p_{i+1}}{2}, \qquad (s_0, \ldots, s_m) = (0, 0.25, 0.5, 0.75, 1). \quad (7)$$

CS-AUC has a 0–100 range because the success inputs are percentages.

RCS-AUC in Figure 4a uses the fixed coordinate $x \in \{0, 0.25, 0.5, 0.75, 1\}$, corresponding to Base-DR target-success levels of 90, 70, 50, 30, and 10%. Each group maps these coordinates to its own physical path. Unreachable levels are clamped to the frozen endpoints; among the 35 target levels, 6 are interpolated, 7 use the easy endpoint, and 22 use the hard endpoint.

B.4 Cost, Policy Selection, and Repetition

For Table 2, HORIZON (Core) uses approximately 45,000 PPO iterations per training seed, and HORIZON (Full) approximately 60,000. Multi-expert distillation totals approximately 135,000 teacher and student optimization iterations per training seed. The remaining methods, except Linear, use approximately 45,000 PPO iterations per training seed. The archived Linear run had a 26,000-update cap but stopped before completing it; its last saved checkpoint is at 14,000 updates. For HORIZON, these are nominal phase-boundary budgets: the current phase is completed before stopping, so the realized update count can exceed the target by the remaining phase updates (at most 1,500). Base initialization, accepted and failed candidate updates, and retries all count. One-rollout retention comparisons freeze policy weights and perform no learner updates, so they are excluded from the PPO-update budget, as are final tests. Their additional simulator interactions and inference time are evaluation overhead; matching PPO-update counts does not imply matching total simulator interactions or wall-clock time. Table 11 reports these budgets and the returned-policy rules.

For a retention comparison with n_{env} parallel environments and actual rollout lengths $T_{\text{cur}}, T_{\text{ref}} \leq 1{,}000$, the simulator-interaction count is $n_{\text{env}}(T_{\text{cur}} + T_{\text{ref}})$. With 4,096 environments, this is at most 8,192,000 transitions per comparison. This counts all vectorized simulator steps, although each environment contributes to the comparison statistics only until its first termination or the horizon. Runtime overhead includes policy inference, simulator stepping, and checkpoint/state restoration; matching PPO-update counts does not quantify these costs.

HORIZON (Core and Full) and its Active-DR sampler variant return the latest accepted checkpoint after completing and judging the final phase. The no-rollback variant retains its terminal learner state; Base-DR, Direct-wide DR, OpenAI ADR, DORAEMON, and GRAM return their budget-end checkpoints. Multi-expert distillation returns the final student checkpoint. Linear returns its last successful pre-expansion checkpoint, selected without OOD-all inspection. OOD-all is evaluated only after each returned policy is fixed. Shared-checkpoint continuations retain the cost and identity of their shared prefix; they answer a continuation question distinct from independent from-scratch training.

Within each training seed, success is first aggregated across valid evaluation episodes. Except for Linear, we then compute the mean and sample SD of the per-training-seed results. Linear instead reports variation across five evaluation seeds of one trained policy. Evaluation seeds sample the test distribution; they do not supply additional trained policies. Within each restoration comparison, Restore and Range-only are paired at the same parent failure event. The six-event analysis treats the parent event as its unit and reports sample variation across event-level effects; repeated continuations for P1 and P2 additionally quantify variation conditional on those parents. Episode-level intervals, when computed from raw counts, describe uncertainty conditional on the particular policy and are kept distinct from cross-training-seed variation.

Table 11: Training budgets and returned-policy protocol for Table 2. Budgets count optimization iterations per training seed, including all teacher and student training for distillation. HORIZON completes the final phase before applying its return rule. Retention comparisons are inference-only overhead excluded from the training budget. OOD-all is evaluated only after policy selection.

Method	Budget	Charged scope	Returned policy
HORIZON (Core)	$\approx 45{,}000$	Base initialization, accepted and failed phases, and retries	Final accepted
HORIZON (Full)	$\approx 60{,}000$	Same accounting as Core over seven groups	Final accepted
w/o rollback	$\approx 45{,}000$	Same as Core; failed-phase learner updates remain live	Budget end
Base-DR / Direct-wide DR	$\approx 45{,}000$	PPO updates under the fixed distribution	Budget end
OpenAI ADR	$45{,}000$	PPO updates, including boundary-sampled trajectories	Budget end
HORIZON + Active-DR	$\approx 45{,}000$	PPO updates and sampler training	Final accepted
DORAEMON	$\approx 45{,}000$	PPO updates under the adaptive distribution	Budget end
GRAM	$\approx 45{,}000$	Learner and adaptation updates	Budget end
Linear	$26{,}000$ cap; $14{,}000$ last saved	Early-stopped run, including the failed tail	Last successful pre-expansion
Multi-expert distillation	$\approx 135{,}000$	Teacher and student optimization, end to end	Final student

C Baseline Protocols

C.1 Method Comparison and Evaluation Scope

Table 2 reports the unified method comparison and ablations, with mean and sample SD across five independent training seeds except for Linear, which uses five evaluation seeds of one training run. Table 12 reports isolated-group results and the Core/Full composition comparison. Unless labeled Full, methods expanding beyond Base use the same four Core groups. The GRAM row follows this unified protocol; the separate three-training-seed native-context study below uses its own initialization and evaluation protocol.

Unless otherwise stated, comparisons use commanded locomotion of Procedural Bodies in Isaac Lab with PPO through the `rsl_rl` stack (Mittal et al., 2025; Schwarke et al., 2025). HORIZON uses a two-layer decoder-only transformer with a 32-step KV-cache window and a privileged MLP critic. Core expands actuation, mass, disturbance, and COM, whereas Full expands all seven groups; groups outside Core remain randomized at Base. Tables 7, 8, and 10 give the network, optimization, and physical-range settings. The following subsections specify baseline architectures, schedules, and checkpoint-selection rules.

The *w/o rollback* ablation follows the same phase gates, range proposal, failed-boundary update, and retry interpolation ($\alpha = 0.4$) as HORIZON. On failure it still retracts the candidate ranges to their accepted values and queues the refined retry, but it disables checkpoint reload: the actor, critic, optimizer, learning rate, and observation-normalization state produced during the failed phase remain live. Range retraction and boundary refinement are retained. Because HORIZON returns its final accepted checkpoint while this variant returns its terminal learner, Table 2 measures the effect of the complete rollback procedure, including final checkpoint selection. It does not separate restoration during training from filtering a failed final candidate. The fixed-range, matched-update continuations in Appendix D.2 isolate the effect of restoring learning state on subsequent learning.

C.2 Fixed and Scheduled Randomization

Base-DR uses the near-nominal intervals in Table 10 and serves as a nominal-control reference. Direct-wide samples the maximum-limit intervals of actuation, mass, disturbance, and COM from its first training update, while contact, inertia, and initial state remain randomized at Base. It tests

Table 12: Isolated-group and joint success (%) for the Core and Full curricula and the comparison methods. Values are reported as mean ± SD. OOD-all shifts the seven physical groups jointly; isolated columns shift the named group.

Training setting	Actuation	Mass	Disturbance	COM	OOD all
Core-domain	$\mathbf{69.3 \pm 2.8}$	$\mathbf{92.1 \pm 1.1}$	$\mathbf{68.0 \pm 2.0}$	$\mathbf{71.1 \pm 1.8}$	$\mathbf{36.8 \pm 2.5}$
Full-domain	63.4 ± 2.5	87.3 ± 1.2	62.9 ± 1.2	69.0 ± 2.5	28.1 ± 1.8
Multi-expert distill. (Rusu et al., 2016)	45.4 ± 2.2	53.9 ± 1.7	26.4 ± 2.1	49.1 ± 1.2	16.5 ± 1.2
GRAM (Queeney et al., 2026)	60.7 ± 1.3	69.4 ± 2.0	37.9 ± 0.2	51.0 ± 1.0	22.7 ± 1.1
HORIZON + Active-DR sampler (Mehta et al., 2020)	50.0 ± 2.1	47.8 ± 2.3	9.4 ± 1.3	34.9 ± 0.4	14.7 ± 0.7

Training setting	Initial state	Contact	Inertia
Core-domain	92.0 ± 1.4	94.5 ± 0.8	91.0 ± 1.1
Full-domain	$\mathbf{99.0 \pm 0.5}$	$\mathbf{98.5 \pm 0.7}$	$\mathbf{97.8 \pm 0.9}$
Multi-expert distill.	80.9 ± 1.5	78.0 ± 2.0	81.4 ± 1.0
GRAM	70.0 ± 1.8	70.0 ± 1.4	71.2 ± 1.7
HORIZON + Active-DR sampler	54.7 ± 2.4	54.7 ± 1.1	57.6 ± 0.6

immediate full-range randomization of the same four groups expanded by HORIZON (Core), using the same learner and deployment inputs. The seven-group OOD-all evaluation additionally shifts the three non-Core groups. Linear widens the Core-group ranges from Base toward the same target on a fixed update schedule for this locomotion task. Later expansions produce short-episode failure, so Linear returns its last successful pre-expansion checkpoint, selected from training-phase performance without OOD-all inspection. This checkpoint reaches 51.4% CS-AUC; after subsequent expansion fails, episode length falls to about 21 near 2,000 updates and about 12 beyond 10,000. The archived run ended before its planned 26,000-update cap, so a completed-budget checkpoint is unavailable. Its last saved 14,000-update checkpoint is evaluated below under the same fixed success criterion. Each row summarizes five evaluation seeds of one training run, not five independently trained policies.

Linear checkpoint	OOD-all success (%)	Base tracking ↓	CS-AUC (%)
Returned, pre-collapse (1,500 updates)	22.0 ± 1.9	0.196 ± 0.004	51.4 ± 0.8
Last saved, early-stopped run (14,000 updates)	2.6 ± 0.6	0.470 ± 0.003	1.1 ± 0.3

C.3 Automatic and Active Domain Randomization

OpenAI ADR adapts independent uniform parameter intervals using boundary performance (OpenAI et al., 2019). Our locomotion adaptation shares HORIZON's two-layer transformer, 32-step KV-cache window, and PPO learner, starting from Base and adapting the four Core groups within the common search limits. At each reset, boundary sampling occurs with probability 0.5: one parameter and one side are selected uniformly, that parameter is fixed to its current boundary, and the others remain randomized. Each boundary collects 240 complete-episode success indicators using the common length and tracking criteria. A mean success rate of at least 0.80 expands the boundary, while a rate below 0.50 contracts it; each step is 0.02 of that parameter's full allowed span. These are locomotion-specific thresholds and step sizes. Boundary trajectories also train PPO. Range contraction retains the current learner; HORIZON's retention gate, checkpoint rollback, and failed-boundary refinement are disabled. Each of five independent training seeds receives 45,000 PPO updates and returns its budget-end checkpoint.

OpenAI ADR obtains $19.8 \pm 1.9\%$ OOD-all success, 0.206 ± 0.002 Base tracking error, and $61.5 \pm 0.6\%$ CS-AUC. It has the highest mean CS-AUC among the alternative methods, exceeding DORAEMON by about 6.7 percentage points and trailing HORIZON (Core) by about 1.0 point. Yet Core achieves about 17.0 points higher OOD-all success and roughly half the Base tracking error. Similar integrated success can therefore coexist with different performance at the strongest joint shift and different nominal tracking precision. This comparison evaluates the complete adaptive

curricula; the matched continuation studies isolate learning-state restoration, while the no-rollback ablation includes the final policy-selection effect.

Active-DR sampling ablation. Active Domain Randomization uses discrepancies between reference and randomized rollouts to learn a sampler within predefined ranges (Mehta et al., 2020). HORIZON + Active-DR sampler replaces uniform sampling with this performance-informed sampler while retaining HORIZON's range proposals, acceptance checks, checkpoint restoration, and boundary refinement. The comparison therefore tests the sampler within HORIZON rather than an independent Automatic Domain Randomization method.

C.4 DORAEMON

DORAEMON expands distribution entropy subject to a success constraint (Tiboni et al., 2024). Our task adaptation uses the two-layer H32 transformer and PPO learner for approximately 45,000 updates, adapting the four Core groups while the other three remain at Base. Each adaptive physical coordinate is mapped affinely from a Beta variable on $[0, 1]$ to its fixed search-limit interval in Table 10. The distribution is factorized, $q_\phi(\mathbf{z}) = \prod_j \mathrm{Beta}(z_j; a_j, b_j)$, initialized around Base and updated every 500 PPO iterations. The target U is the product of $\mathrm{Beta}(1, 1)$ marginals on the same support.

For completed trajectories sampled under q_k, let y_i denote the binary episode-success indicator from Appendix B.3. The distribution update solves

$$\begin{aligned} \min_{\phi} \quad & D_{\mathrm{KL}}(q_\phi \| U) \\ \text{subject to} \quad & \widehat{S}(\phi) \geq 0.5, \qquad D_{\mathrm{KL}}(q_k \| q_\phi) \leq 0.04, \\ & \widehat{S}(\phi) = \frac{1}{N} \sum_{i=1}^{N} \frac{q_\phi(\mathbf{z}_i)}{q_k(\mathbf{z}_i)} y_i. \end{aligned} \tag{8}$$

The success estimate uses ordinary importance weights, without clipping or self-normalization. If the current distribution fails the success constraint, a backup step maximizes $\widehat{S}$ within the same KL constraint before attempting entropy expansion; if feasibility remains unattainable, the performance-improving backup distribution is retained. An invalid numerical solution leaves the current distribution unchanged. Optimization uses SciPy's `trust-constr` on CPU with float64 arithmetic, analytic Beta KL, at most 500 solver iterations, and gradient tolerance 10^{-6}. Initial Beta concentration is capped at 200 and shape parameters are bounded below by 0.8. The update changes the sampling distribution without HORIZON's retention gate, learner restoration, or boundary refinement. Pretraining updates, if used, count toward the total learner budget. The unified comparison yields $18.9 \pm 0.9\%$ OOD-all success, 0.182 ± 0.001 Base tracking error, and $54.8 \pm 1.0\%$ CS-AUC.

C.5 GRAM AND PRIVILEGED TRAINING CONTEXT

GRAM retains its native actor architecture, training context, and history adapter (Queeney et al., 2026). The five-seed row in Table 2 uses the unified Procedural Bodies task initialization and Core-group support. The separate native-context study uses GRAM's Go2 task initialization and seven-group simulator randomization; neither study adds the randomized groups as new privileged context coordinates. GRAM's native context includes additive base mass, foot friction, motor strength, and joint bias, while other variations enter through simulator events. During fixed evaluation, explicit context is zeroed and actions use the native 16-step history adapter. The adapter also maps the HORIZON roll/pitch/yaw observation convention to GRAM's projected-gravity convention.

Under the unified protocol in Table 2, GRAM obtains $22.7 \pm 1.1\%$ seven-group OOD-all success, 0.407 ± 0.004 Base tracking error, and $34.3 \pm 0.9\%$ CS-AUC across five independent training seeds.

In the separate native-context study, evaluating three checkpoints with the paper-wide episode-success rule yields $19.29 \pm 1.48\%$ OOD-all success, 0.4069 ± 0.0037 Base tracking error, and $34.84 \pm 0.78\%$ CS-AUC, integrated over the five common severity levels. These checkpoints were trained for 1,000 pretraining and 500 fine-tuning iterations; evaluation uses one common seed and 4,096 valid episodes per checkpoint and condition.

An expanded GRAM context containing additional randomized physical quantities and time-varying disturbance parameters is outside these comparisons. Reward differences also preclude interpreting return as a common metric even when episode success is evaluated identically.

C.6 Policy Capacity and Group Diagnostics

The capacity diagnostic contrasts an MLP and transformer under multi-group training. Figure 6 separates the logged range-reach summary from isolated-group success. The ordering differs across the two quantities and across physical groups. For example, the transformer has larger displayed disturbance range reach and higher disturbance success, whereas the MLP has larger mass range reach with similar mass success. The figure motivates evaluating capability independently of curriculum range size. This architecture comparison has a training lineage and cost distinct from the common algorithm study.

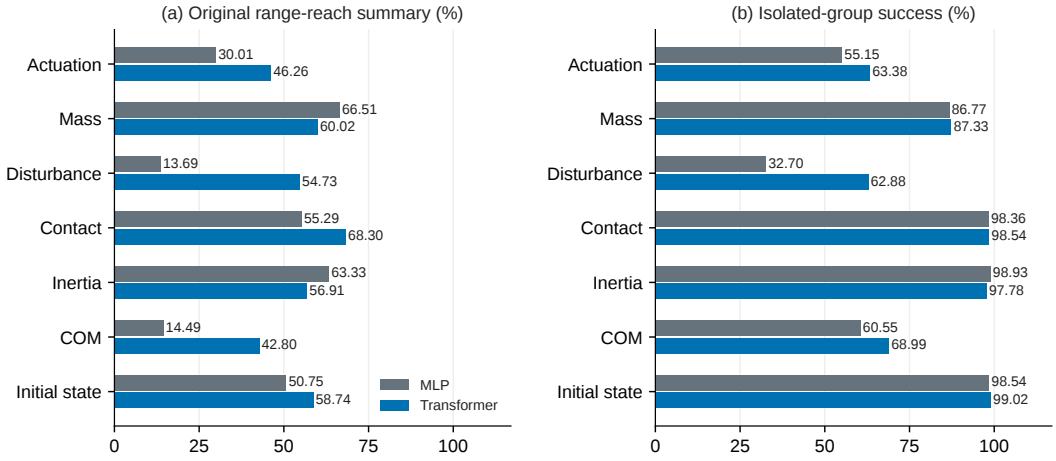


Figure 6: MLP and transformer diagnostics shown in separate range and success panels. Each success row tests the named group under its trajectory-specific protocol; range reach and success have different meanings.

C.7 Multi-Expert Distillation

The distillation comparison uses seven frozen HORIZON checkpoints, one trained for each physical group, as teachers. Each teacher collects observation–action sequences under its corresponding single-domain setting. A training sequence retains its source-teacher identity g, and its stored teacher actions supply the supervision targets. The student is a two-layer decoder-only transformer with a 32-step history window and the same deployment inputs as HORIZON. Sequence behavior cloning minimizes

$$\mathcal{L}_{\mathrm{BC}}(\theta) = \mathbb{E}_{(h_t, a_t^{(g)}, g)\sim\mathcal{D}} \left[\frac{1}{12} \left\| \mu_\theta(h_t) - a_t^{(g)} \right\|_2^2 \right], \tag{9}$$

where h_t contains the student's deployment-available observation history and $a_t^{(g)}$ is the action recorded from the source teacher. Teacher identity selects the supervision target but is not an input to the student; actions from different teachers are not averaged. At evaluation, the student acts alone without a domain identifier. The approximately 135,000-iteration budget in Table 11 includes training all seven single-domain teachers and the student, rather than 135,000 student iterations. Larger single-domain accepted ranges describe teacher exposure but do not establish coverage of the student's joint test distribution. Figure 7 shows the frontier summary.

The tested distillation procedure achieves 16.5% joint success, below the 36.8% of HORIZON (Core). This result is specific to the evaluated teacher checkpoints, state-distribution coverage, teacher assignment, distillation objective, and student training budget.

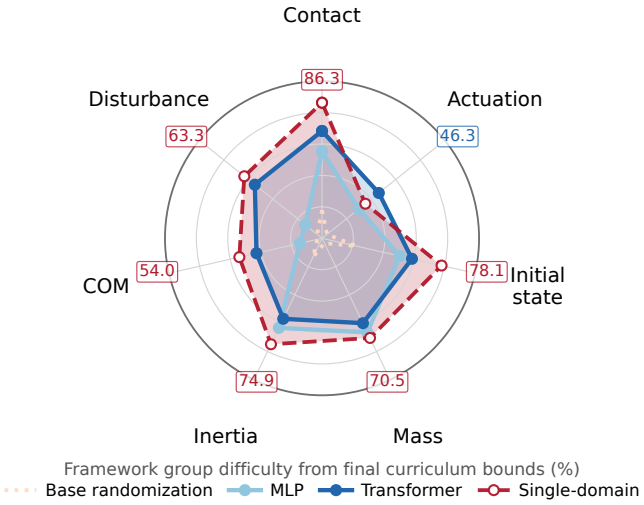


Figure 7: Single-domain and multi-domain frontier summary. The quantities describe training-range reach on the respective axes and remain distinct from the joint behavior of the distilled student.

D Mechanism and Sensitivity Diagnostics

D.1 Phase Outcomes and the L_k and C_k Ablations

The phase diagnostic contains 1,503 recorded phases from 65 curriculum runs. It associates the joint acceptance outcome S_k with the change in logged normalized frontier difficulty over the next three phase decisions. Table 13 reports the category summaries. Phases with $L_k = 1, C_k = 0$ (only L_k passes) have an episode fraction of 0.891, compared with 0.865 among phases with $L_k = C_k = 1$ (both checks pass), while their subsequent average frontier changes are 0.0285 and 0.0635. The Spearman correlations with subsequent frontier change are 0.35 for episode fraction and 0.52 for the composite score.

Table 13: Association between phase outcome and subsequent logged frontier progress. Episode fraction averages normalized training episode length over the entire phase, whereas L_k uses the most recent 500 iterations at the decision point. Lower episode lengths early in training can therefore keep the phase average below 0.95 even when the final window passes the gate. The last two columns describe the next three phase decisions. These observations are not a randomized intervention on restoration.

Phase category	Phases	Episode fraction	Reward ratio	Frontier change	Progress (%)
$L_k = C_k = 1$	656	0.865	1.039	+0.0635	85.8
$L_k = 1,\ C_k = 0$	414	0.891	0.775	+0.0285	55.6
Low episode length	433	0.682	0.387	+0.0100	16.2

The joint acceptance rule participates in accepting ranges and determining subsequent proposals. Its association with future frontier movement therefore partly reflects the search procedure itself. Phases from a run are also dependent, so the number of phases is not a count of independently trained policies. The matched L_k-only ablation below directly isolates C_k, while the controlled restoration branches in Section 4.2 isolate learning-state restoration.

The matched ablation compares HORIZON, with $S_k = L_k C_k$, against an L_k-only variant, with $S_k = L_k$. Both branches start from the same accepted checkpoint and use five paired continuation-training seeds, with the same restoration and boundary-refinement procedures. In the paired phase trace, both variants pass L_k during the target actuation expansion. HORIZON measures $q_k^R = 0.8943$, obtains $C_k = 0$, and rejects the candidate; the L_k-only variant measures $q_k^R = 0.8632$ but

accepts because C_k is omitted from S_k. On fixed Base evaluations with matched evaluation seeds across branches, HORIZON achieves $96.93 \pm 1.23\%$ Base success with 0.1133 ± 0.0056 tracking error, compared with $87.56 \pm 1.91\%$ and 0.1804 ± 0.0073 for L_k-only. We aggregate results within each training seed and report the mean and sample SD across the five continuation-training seeds. This intervention shows that C_k is not redundant with L_k: it prevents a candidate that remains feasible from being accepted after degradation under the accepted conditions.

The complementary C_k-only intervention repeats both branches with five continuation-training seeds from the same accepted four-group Core checkpoint, matching the seed within each branch pair. Each run uses 4,096 parallel environments, a 6,000-update budget, and the same proposal and restoration mechanisms. In the paired phase trace, the branch with $S_k = L_k C_k$ completes four phases: disturbance, mass, and COM are accepted, while actuation is rejected. The branch with $S_k = C_k$ accepts ten phases. All ten pass C_k and fail the candidate episode-length condition, directly separating retention of accepted behavior from mastery of the candidate range.

Evaluation uses the same frozen ranges for both branches. Each single-boundary case expands one Core group by 20% of the distance from its Base range to the shared accepted range, while the other Core groups remain at the shared accepted ranges; the composed case applies all four expanded boundaries together. For each continuation-training seed, we evaluate the latest accepted and budget-end checkpoints with 1,024 valid episodes per case and matched evaluation seeds across branches. Success is aggregated within each training seed before computing the mean and sample SD across the five continuation-training seeds. Tables 14 and 15 report the aggregate and per-group success rates, respectively.

Table 14: Aggregate L_k ablation results on fixed new-support evaluations. Candidate feasibility uses the window-average L_k thresholds, whereas fixed-evaluation success uses the episode-level criterion in Section 4.1. Single-boundary results average actuation, mass, COM, and disturbance within each training seed. Entries are mean $\pm$ SD across five continuation-training seeds from a shared parent checkpoint, with 1,024 valid episodes per condition and training seed.

Checkpoint	Evaluation	$L_k C_k$	C_k only
Latest accepted	Single new boundary	$\mathbf{83.4 \pm 0.6}$	69.2 ± 0.5
Latest accepted	Four-group new composition	$\mathbf{77.1 \pm 0.8}$	63.1 ± 1.9
Budget endpoint	Single new boundary	$\mathbf{83.4 \pm 0.4}$	64.7 ± 0.4
Budget endpoint	Four-group new composition	$\mathbf{77.4 \pm 1.1}$	58.1 ± 1.8

Table 15: Complete fixed new-support results for the L_k ablation. Entries are success percentages, mean $\pm$ SD across five continuation-training seeds from a shared parent checkpoint. Training seeds are matched between the two branches.

Checkpoint	Fixed evaluation	$L_k C_k$	C_k only
Latest accepted	Actuation boundary	$\mathbf{84.41 \pm 0.99}$	71.07 ± 1.66
Latest accepted	Mass boundary	$\mathbf{84.65 \pm 1.60}$	71.29 ± 1.10
Latest accepted	COM boundary	$\mathbf{85.76 \pm 0.63}$	71.11 ± 1.14
Latest accepted	Disturbance boundary	$\mathbf{78.83 \pm 1.01}$	63.42 ± 1.28
Latest accepted	Four-group composition	$\mathbf{77.11 \pm 0.79}$	63.09 ± 1.95
Budget endpoint	Actuation boundary	$\mathbf{84.47 \pm 1.70}$	65.68 ± 1.31
Budget endpoint	Mass boundary	$\mathbf{85.06 \pm 0.96}$	66.80 ± 0.47
Budget endpoint	COM boundary	$\mathbf{84.12 \pm 1.24}$	67.42 ± 1.25
Budget endpoint	Disturbance boundary	$\mathbf{80.14 \pm 0.64}$	58.81 ± 0.74
Budget endpoint	Four-group composition	$\mathbf{77.38 \pm 1.11}$	58.11 ± 1.76

The per-condition ordering agrees with the aggregate in Table 14: removing L_k lowers mean success at every new boundary and under their joint composition. The latest accepted C_k-only checkpoints already trail the joint gate, and their budget-end mean performance falls further. The paired phase trace shows how this degradation can occur despite every acceptance decision passing C_k. The result identifies the missing signal precisely: C_k limits excessive forgetting on accepted domains, while L_k is needed to establish relatively stable mastery of the current candidate domain before expansion continues.

D.2 Controlled Restoration Continuations

Study structure and event selection. The six parent events are organized along two complementary axes. P1–P2 provide a cross-type contrast between actuation collapse and disturbance without immediate collapse, while P1 and P3–P6 replicate actuation failure across five parent trajectories. For each parent trajectory, we scan the curriculum in temporal order and select the first natural failure in a Core group after at least 500 updates in the current phase, provided that the paired retention comparison is complete and all recorded metrics are finite. At the selected event, both branches retract to the same accepted ranges. Range-only retains the learner state produced during the failed phase, whereas Restore reloads the latest accepted learner state. Each pair is evaluated at its matched endpoint using the paper-wide episode-success criterion, with 1,024 valid episodes per endpoint condition.

Table 16: Detailed matched restoration results. For P1–P2, immediate entries are point estimates at the paired branch points, and endpoint entries are mean ± SD across five continuation-training seeds initialized from a shared pre-failure checkpoint. P3–P6 report endpoint estimates for four additional actuation parent trajectories. Bold denotes higher performance.

Failure group	Evaluation point	Metric	Range-only	Restore
Actuation P1	Immediate	Base	2.87	**98.11**
Actuation P1	Immediate	CS-AUC	1.05	**75.68**
Actuation P1	Endpoint	Base	86.70 ± 1.01	$\mathbf{97.95 \pm 0.40}$
Actuation P1	Endpoint	CS-AUC	63.02 ± 0.93	$\mathbf{70.62 \pm 0.40}$
Disturbance P2	Immediate	Base	92.72	**98.77**
Disturbance P2	Immediate	CS-AUC	81.31	**84.68**
Disturbance P2	Endpoint	Base	88.57 ± 0.93	$\mathbf{97.85 \pm 0.48}$
Disturbance P2	Endpoint	CS-AUC	75.41 ± 0.69	$\mathbf{80.62 \pm 0.37}$
Actuation P3	Endpoint	CS-AUC	74.68	**76.64**
Actuation P4	Endpoint	CS-AUC	77.60	**80.59**
Actuation P5	Endpoint	CS-AUC	**72.85**	72.47
Actuation P6	Endpoint	CS-AUC	75.76	**79.70**

Actuation P1. The actuation continuation uses the first natural actuation proposal failure obtained from a parent checkpoint after replaying its remaining warmup from iteration 1,000. The failure occurred at global iteration 2,001 after 500 proposal updates, when the one-rollout retention comparison C_k had return ratio $q_k^R = 0.1763$ (defined in Appendix A.3). Before the proposal, the accepted actuation intervals were effort limit $[39, 41]$, velocity limit $[29, 31]$, stiffness $[39, 41]$, and damping $[0.9, 1.1]$. The failed proposal changed their upper endpoints to 54, 40.667, 54, and 2.067, respectively. The paired continuation retracts both branches to the same accepted intervals. We repeat both branches from this shared checkpoint with five continuation-training seeds.

Range-only retains the failed phase's actor, critic, optimizer, observation-normalization state, and learning rate. Restore reloads those quantities from the latest accepted checkpoint. The two branch checkpoints otherwise match in iteration, curriculum runtime and evaluator state, random-number state, environment snapshot, command-manager state, and scheduler state. Both branches use 4,096 parallel environments, keep the same accepted physical-randomization ranges, and continue for 2,000 PPO updates. Immediate recovery is evaluated at the paired branch points; the matched-budget endpoint uses the checkpoints at iteration 3,498 after the full continuation.

We evaluate Base with the paper-wide episode-success rule. Each continuation-training seed is evaluated with 1,024 valid episodes on flat terrain. At the paired branch points, Range-only and Restore have immediate Base success of 2.87% and 98.11%, respectively. Across the five continuation-training seeds, their matched-budget endpoint values are $86.70 \pm 1.01\%$ and $97.95 \pm 0.40\%$.

The severity path linearly interpolates both endpoints of the failed-group intervals from Base to the search limits at $s \in \{0, 0.25, 0.5, 0.75, 1\}$. Trapezoidal integration gives immediate CS-AUC point estimates of 1.05% for Range-only and 75.68% for Restore. Across the five continuation-training seeds, the endpoint CS-AUC values are $63.02 \pm 0.93\%$ and $70.62 \pm 0.40\%$, respectively, a

7.60-point Restore advantage. These comparisons distinguish restoring accepted learning state from merely retracting the physical ranges.

Disturbance P2. The disturbance continuation starts from a natural failure with accepted push, force, and torque magnitudes of 2.0, 0.3467, and 0.1667 and a failed proposal of 3.0, 0.5644, and 0.2778, respectively. Its retention return ratio, $q_k^R = 0.9767$, indicates a milder failure than the actuation case. Range-only and Restore start from the same parent checkpoint, use the same accepted ranges, terrain state, reset procedure, and 2,000-update continuation budget, and are repeated with five continuation-training seeds. At the paired branch points, their immediate Base success rates are 92.72% and 98.77%, and their immediate CS-AUC values are 81.31% and 84.68%. Across the five continuation-training seeds, endpoint Base success is $88.57 \pm 0.93\%$ versus $97.85 \pm 0.48\%$, while endpoint CS-AUC is $75.41 \pm 0.69\%$ versus $80.62 \pm 0.37\%$. Restore therefore retains a 9.28-point Base-success and 5.21-point CS-AUC advantage at the matched endpoint.

At the matched endpoint, Restore improves the strongest failed-group severity success by 11.07 ± 1.80 points for actuation and 8.30 ± 2.25 points for disturbance, and reduces Base tracking error by 0.0428 ± 0.0021 and 0.0543 ± 0.0026, respectively. Each failure study contains two branches continued with five training seeds, with 1,024 valid evaluation episodes per seed and endpoint condition; the immediate values come from the two paired branch points before these seeded continuations. Both continuations passed their 2,000-update audits: the accepted ranges remained fixed and no NaNs occurred. The shared parent checkpoint fixes the pre-failure learner state and accepted ranges; the endpoint variation is across independently seeded continuation runs from that checkpoint, not across independently trained parent policies.

Table 17: Endpoint effects of learning-state restoration across six parent failure events. Values are Restore minus Range-only in percentage points; positive values favor Restore. P1–P2 provide the cross-type contrast, and P1 with P3–P6 forms the actuation cross-parent replication. Summary rows report mean ± sample SD across parent-event estimates.

Event	Failure group	Stratum	ΔBase	ΔCS-AUC	ΔSeverity-1
P1	Actuation	Collapse	$+11.25$	$+7.60$	$+11.07$
P2	Disturbance	No collapse	$+9.28$	$+5.21$	$+8.30$
P3	Actuation	Collapse	-1.17	$+1.95$	$+3.12$
P4	Actuation	No collapse	$+2.93$	$+2.99$	$+6.05$
P5	Actuation	Collapse	$+1.27$	-0.38	-0.78
P6	Actuation	Collapse	$+2.25$	$+3.94$	$+7.62$
Collapse ($n = 4$)			$+3.40 \pm 5.43$	$+3.28 \pm 3.38$	$+5.26 \pm 5.18$
No collapse ($n = 2$)			$+6.10 \pm 4.49$	$+4.10 \pm 1.57$	$+7.18 \pm 1.59$
All events ($n = 6$)			$+4.30 \pm 4.86$	$+3.55 \pm 2.74$	$+5.90 \pm 4.19$

Each endpoint metric favors Restore in five of the six parent failure events. The positive mean differences in both strata show a shared restoration benefit across failures with and without immediate capability collapse.

D.3 Interpolation and Evaluation Consistency

The boundary interpolation study compares $\alpha = 0.2, 0.4, 0.6$ using in-training joint evaluations. The $\alpha = 0.6$ run branches from the $\alpha = 0.2$ pre-failure checkpoint rather than using an independent identical initialization. The summaries use phases 18, 18, and 15, respectively. They are therefore identified by their actual phase rather than treated as a fully phase-matched sweep. The default has the highest observed in-training score in this diagnostic. Because the runs do not share independent identical initializations or fully matched phases, this result does not establish a stable optimum for α. Minimum-step and refinement rules also affect the realized intervals independently of the nominal growth factor.

The fixed-evaluation consistency check applies the paper-wide episode-success rule to the same 1,280 episodes. The resulting seven-group and Core-joint success percentages are 26.41 and 28.75, respectively. This evaluation is separate from the training-time L_k gate and does not perturb L_k or C_k or supply additional training seeds.

Table 18: In-training four-group joint success (%) in the boundary interpolation diagnostic. The $\alpha = 0.6$ run is a continuation branch; phases and initializations are not fully matched.

α	In-training phase	Four-group joint success
0.2	18	25.98
0.4	18	33.98
0.6	15	28.08

D.4 Frozen-Policy Behavioral Probe

Checkpoint restoration during training is distinct from recovering locomotion after a sudden deployment perturbation. A frozen-policy probe compares checkpoints at logged iterations 3,000 and 5,000 under an abrupt actuation change, a sham condition, and a static altered-actuation condition. Each condition uses 256 paired evaluation episodes. The recovery rule combines root height, roll, pitch, and tracking thresholds held continuously for 0.5 s within a 3 s deadline. A Base sham calibration record defines the thresholds: root height at least 0.27545 m, absolute roll at most 0.09713 rad, absolute pitch at most 0.11574 rad, and tracking error at most 0.4.

Table 19: Frozen-policy probe using the primary recovery rule. Counts are recovered episodes / 256. Both checkpoint evaluations use fixed weights. The sham control measures changes that also occur without the abrupt actuation intervention.

Condition	Checkpoint 3,000	Checkpoint 5,000	Difference (pp)
Abrupt actuation change	156/256	215/256	+23.05
Sham	176/256	242/256	+25.78
Static altered actuation	139/256	210/256	+27.73

The increase under the abrupt change is accompanied by a larger increase in the sham condition. The difference in differences is -2.73 percentage points, with a paired-episode bootstrap 95% interval of $[-8.59, 3.13]$. Thus this probe does not isolate an improvement specific to recovering from the abrupt intervention. Its episodes are repeated evaluations within a training lineage, not independent training runs. The result supports using precise training-state terminology for HORIZON's rollback operation rather than claiming a demonstrated behavioral recovery mechanism.

E Domain Composition Details

E.1 Development Evidence and Core Selection

The development analysis includes seven single-group curricula, twelve two-group combinations, and two three-group combinations. Table 20 summarizes the isolated evaluations along the single-group trajectories. Earliest measurements follow initial training, and peaks are selected within each trajectory. The single-group development evaluations use the paper-wide episode-success rule. They describe progress under that development protocol rather than repeated training-seed estimates for the common comparison.

Figure 4b shows all ten pairs among actuation, mass, disturbance, COM, and inertia. The diagnostic quantities compare each pair with the corresponding single-group references. Let $d_{i|\{i,j\}}$ be the logged normalized frontier summary for group i in a pair, and $p_{i|\{i,j\}}$ its isolated-group success. The displayed changes average over the two groups:

$$\begin{aligned} F_{ij} &= \tfrac{1}{2}\big[(d_{i|\{i,j\}} - d_{i|\{i\}}) + (d_{j|\{i,j\}} - d_{j|\{j\}})\big], \\ O_{ij} &= \tfrac{1}{2}\big[(p_{i|\{i,j\}} - p_{i|\{i\}}) + (p_{j|\{i,j\}} - p_{j|\{j\}})\big]. \end{aligned} \tag{10}$$

Values are percentage-point changes on the respective normalized quantities. Frontier and success changes measure different objects and are not summed into a capability score. Each pair is compared with the checkpoint-selected single-group references used in the development analysis. A best isolated-evaluation reference is not automatically the policy at the final accepted frontier.

Table 20: Single-group development trajectories. Values are isolated OOD success (%) at the earliest recorded evaluation, the peak, and the last recorded evaluation of each selected trajectory. Each row trains the named group.

Training group	Earliest recorded	Peak	Last recorded
Actuation	42.55	56.59	50.61
Mass	65.99	84.55	84.55
Disturbance	46.68	69.38	68.07
COM	55.08	60.06	60.06
Contact	93.38	95.95	95.04
Inertia	83.11	91.55	91.55
Initial state	96.56	98.41	98.07

Table 21 reports the complete development diagnostic. Actuation–COM, COM–disturbance, and mass–disturbance have positive changes in both accepted range reach and isolated-group success. These observations of facilitated expansion, together with the single-group learning demands, guided the study toward the four-domain Core. The pair averages summarize net changes across the two groups; the groups may benefit unequally. Actuation–mass has negative F and O despite both groups belonging to Core, while COM–inertia has positive O and negative F. The selected configuration therefore reflects the combined development evidence, rather than a rule requiring every within-Core pair to improve both metrics.

Table 21: Complete pairwise development diagnostics. Each joint test shifts the named pair, with other groups at Base, so joint success columns across different rows refer to different test distributions. F and O average changes relative to the two corresponding single-group references, as defined in Equation 10. Values are descriptive summaries; changes are percentage points.

Training pair	Joint success (%)	F (pp)	O (pp)
Actuation + Disturbance	34.16	$+12.61$	-0.78
Mass + Disturbance	63.77	$+4.59$	$+6.02$
Disturbance + Inertia	77.49	$+7.85$	$+7.72$
COM + Disturbance	62.33	$+14.91$	$+17.51$
Mass + COM	60.11	$+4.32$	$+8.84$
Mass + Inertia	82.28	-1.77	$+1.62$
COM + Inertia	77.81	-10.05	$+12.44$
Actuation + COM	67.99	$+33.21$	$+13.35$
Actuation + Mass	61.62	-12.49	-8.38
Actuation + Inertia	48.00	-24.02	-1.42

The additional actuation–contact pair has isolated peaks of 50.78% for actuation and 98.46% for contact. Disturbance–initial-state reaches 77.15% and 99.61% on the respective isolated tests. These results complement the ten-pair diagnostic: an easily controlled partner can either leave a difficult group unresolved or accompany improved control, depending on the pair.

The actuation–mass–disturbance three-group curriculum has isolated peaks of 53.34%, 92.11%, and 67.99%, respectively. Together with the actuation–COM and COM–disturbance pair results, its remaining actuation difficulty motivates the four-group composition. The separate mass–COM–inertia study has isolated peaks of 86.16%, 72.22%, and 98.27%. These are per-test trajectory peaks and may occur at different checkpoints. Joint tests in the pair studies shift the corresponding pair, while the mechanics three-group test shifts mass, COM, and inertia together. Such tests measure different physical combinations from the common seven-group comparison.

E.2 Core–Full Joint-Composition Evaluation

Three targeted extensions test the selected Core together with one group that remains at Base during Core training. Table 22 reports mean $\pm$ SD across the corresponding repeated evaluations.

Table 22: Core and Full under three targeted extensions of the selected Core. Rows specify evaluation conditions rather than additional training configurations. Diff. is Core minus Full in percentage points.

Active OOD groups	Core	Full	Diff.
Selected core + contact	46.8 ± 1.4	34.2 ± 0.9	$+12.6$
Selected core + inertia	43.2 ± 1.2	33.2 ± 0.7	$+10.0$
Selected core + initial state	46.1 ± 1.4	35.9 ± 0.4	$+10.2$

Core leads by 10.0–12.6 percentage points in all three targeted extensions. We additionally re-evaluate the current Core and Full checkpoints on 91 joint OOD combinations. Each row aggregates combinations with the stated number of active groups; Core and Full are evaluated on the same combination in every comparison. The reported success values are means across combinations, and the final column counts combinations for which one checkpoint has higher success. These are checkpoint-level evaluation results, not additional independent training runs.

Table 23: Fixed-checkpoint Core–Full sweep over 91 joint OOD combinations. Success is averaged across combinations within each group-count stratum. Diff. is Core minus Full in percentage points.

Active groups	Combinations	Core success (%)	Full success (%)	Diff. (pp)	Core wins / Full wins
3	35	57.27	56.39	$+0.88$	20 / 15
4	35	46.55	42.71	$+3.84$	22 / 13
5	21	39.06	32.31	$+6.75$	15 / 6
All	91	**48.94**	**45.57**	$+3.37$	**57 / 34**

Core has the higher success value on 57 of the 91 combinations. Its aggregate mean exceeds Full's by 3.37 percentage points; by group count, the corresponding differences are 0.88, 3.84, and 6.75 points for three, four, and five active groups. The counts describe this fixed-checkpoint grid and do not provide uncertainty across independently trained policies.

F Hardware and Additional Transfer

F.1 Trial Protocol and Failure Outcomes

Each quantitative condition has fifteen trials per policy. The first 5 s require balance, followed by 15 s under forward command $(0.3, 0, 0)$. Success requires completing both stages without intervention, and all attempts count toward the denominator, including failures during initial balance. HORIZON completes 49/60 trials and Base-DR completes 0/60; their initial-balance counts are 57/60 and 31/60.

HORIZON completes all thirty payload trials. Among the fifteen FL-locked trials, eight succeed, two end in delayed falls, and five exhibit uncontrolled motion. With four-leg extensions, eleven succeed, one ends in immediate imbalance, and three exhibit uncontrolled motion. These failures account for all eleven unsuccessful HORIZON trials.

For Base-DR, the 3 kg condition produces one delayed fall, one failure to start, and thirteen uncontrolled-motion outcomes. The 5 kg condition produces two immediate imbalances, four delayed falls, and nine failures to start. FL locking produces seven immediate imbalances and eight failures to start. The four-leg extension condition produces thirteen immediate imbalances and two delayed falls. Each condition totals fifteen attempts.

The four-leg extension condition attaches 15 cm calf extensions to all four legs. The FL-locked quantitative condition remains separate from other front- or hind-knee demonstrations. Dynamic two-ball loading and payload-plus-extension sequences in Figure 5 do not enter the repeated-trial denominators. Policy weights remain fixed throughout these demonstrations, including the temporal sequences annotated in the montage.

F.2 Cross-Robot Simulation Control

Additional simulation rollouts use Go2, Go1, A1, Spot, CyberDog, B1, and B2 under base locomotion commands (Figure 8). They examine the deployment-style control interface on other bodies, separately from Procedural Bodies training and the seven-group success evaluations.

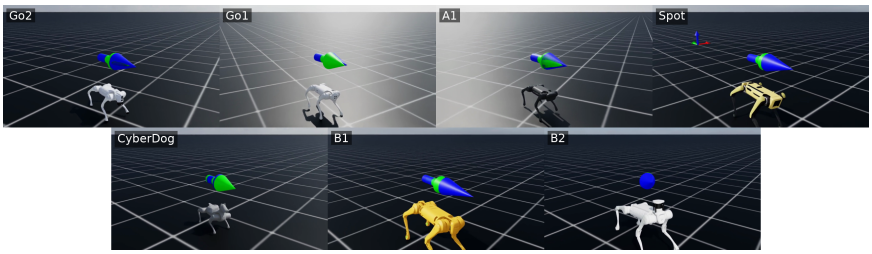


Figure 8: Cross-robot simulation rollouts under base commands. These qualitative frames imply neither success rates nor an additional training-body distribution.