INITIALIZATION AND SECONDARY-TASK RECOVERY UNDER FIXED PRIORITY CONTROL

Anonymous authors
Paper under double-blind review

ABSTRACT

Learning one secondary-task coefficient offers a narrow interface for fault-tolerant manipulation with given fault reports and priority control. We distinguish changes to a trained policy from useful constrained control. In Isaac Sim, six PPO actors deploy constant full retention despite stochastic training, as established by guarded final-layer logit bounds. A paired intervention removes only their initial full-retention bias, retaining architecture, geometry, optimization, budgets, and the selection rule. Training and selection change the action map and produce variable deployed behavior, but all prespecified constrained-control criteria fail. Across 992 quarter-tolerance tasks per neutral fit, policies have 6–12 primary-crossing cases; a selected margin rule has three and weakly improves recovery count and secondary cost over every fit. A subsequent same-input diagnostic establishes a material replacement of the trained categorical distribution by its mode. A same-weight native deployment intervention then executes both choices: sampling improves some costs and recoveries but introduces 46 primary-crossing episodes across twelve sampled arms, versus zero in their six modal counterparts. None of the three prespecified fixed-penalty fits meets the useful-decoder criterion. Complementary selection and force-budget interventions establish bounded scheduling opportunity; failed gradient, residual, and fixed-probe screens retain their stopping conclusions. Removing constant deployment through initialization is insufficient for constrained-control gains. These findings neither identify a complete explanation of the remaining loss nor establish RL impossibility or a physical safety certificate. Two archived hardware trials from a separately deployed PPO policy document scalar execution.

1 INTRODUCTION

A repaired actuator can regain command effectiveness while the robot still has task error and residual motion. Restoring a suppressed secondary channel quickly can remove error but create a physical transient; retaining too little can prevent recovery. Can feedback selection of one coefficient improve primary-constrained task progress over strong causal scalar rules?

We learn only $\alpha_{\text{sec}} \in [0, 1]$. Masks, tolerances, current fault reports, kinematics, gravity compensation, and the weighted priority allocator are given. The policy acts at 20 Hz, the controller recomputes channels at 100 Hz, and the simulator records every 2 ms state (Figure 1). Learning may act before, during, and after the fault; it does not learn detection or the hierarchy.

Implementation choices can substantially affect on-policy RL (Engstrom et al., 2020; Andrychowicz et al., 2021). Here the deployment interface supplies a specific diagnostic. The six originally selected PPO actors have strictly positive global logit-gap bounds for the full-retention action. Their weights and probabilities change during training, but deterministic deployment cannot use any observation. This does not show that RL is unsuitable, or by itself identify the initial bias as the cause of poor control.

We therefore change only that initial bias and retain both penalty variants, all three fitting seeds, and the same budget and selection rule. Same-input initial-to-selected comparisons verify a changed action map; fresh physical tests verify nonconstant deployed behavior. Yet the primary and secondary evaluation criteria fail, and a selected margin rule weakly improves all three aggregate control criteria over every neutral fit. The intervention changes exploration and selected checkpoints as well

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.

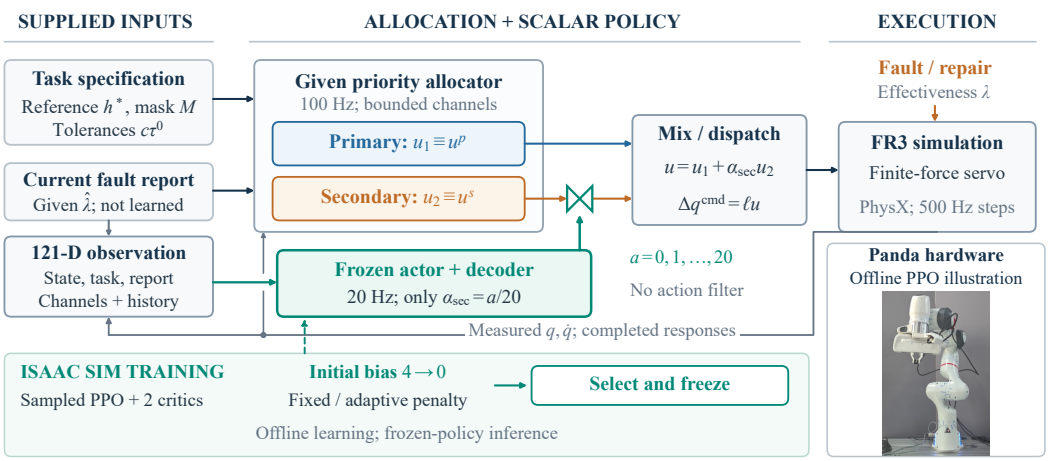


Figure 1: Framework and learning interface. Supplied task specifications, fault reports, and priority allocation produce primary and secondary channels. The frozen actor selects only α_{sec} at 20 Hz; the 100 Hz controller mixes the channels and a finite-force servo executes the command. Isaac Sim training, checkpoint selection, and deployment are separated. Initialization and same-weight decoding are distinct interventions; outcome checks do not filter actions or certify safety. The video frame shows the Panda hardware platform.

as final mappings. Its conclusion is consequently about the complete training/selection outcome: removing constant deployment this way is insufficient for constrained-control gains.

The selected action map also need not represent the trained stochastic law. A same-weight, same-input development diagnostic shows that modal deployment materially replaces the categorical distribution, including substantial probability on distant alpha values. We then change only the deployment rule for those frozen weights and run complete native episodes on four fresh development scenarios. Sampling changes recovery and error, but introduces primary crossings and fails the prespecified useful-decoder criterion for every fixed-penalty fit. Thus distribution replacement identifies a concrete intervention without predicting a beneficial direction. This distinguishes the total training/selection effect of initialization from a same-weight deployment effect.

Two complementary studies keep the scope of that conclusion clear. Fixed-predictor selection interventions restore lost progress without beating stronger simple controls. Matched-prefix force-budget changes alter finite-option value while retaining a common post-repair-admitted recovering option. They establish physical scheduling opportunity and decision limitations in separate controlled settings; neither identifies the residual loss of nominal PPO nor shows that useful feedback has been learned. Our contribution is this separation of deployed-action limits, their intervention, and measured control value, with all unsuccessful learning comparisons retained.

2 Task, Supplied Controller, and Empirical Constraints

For seven joint coordinates q, define $h(q) = (\phi, \theta, \psi, x, y, z)$ using $R = R_z(\psi)R_y(\theta)R_x(\phi)$: literal base-frame extrinsic xyz Euler angles and Cartesian position. Anchored angular branches obey $|\theta| \le \pi/3$ and changes at most $\pi/2$. A supplied nonempty proper six-bit mask M marks primary coordinates; its complement marks secondary coordinates. All 62 masks occur in training and evaluation. With $e = h^\star - h(q)$, base tolerances $\tau^0 = (.02, .02, .02, .01, .01, .01)$, and secondary normalization $\sigma = (.2, .2, .2, .05, .05, .05)$, define

$$P_c(e, M) = \max_{i:M_i=1} \frac{|e_i|}{c\tau_i^0}, \qquad S(e, M) = \frac{\sum_{i:M_i=0}(e_i/\sigma_i)^2}{\sum_i (1 - M_i)}. \tag{1}$$

Angles are in radians and positions in metres. The task provides c; it is not selected from favorable test outcomes. Full-feedback learning uses $c \in \{1, .25, .1\}$, with .25 primary. Earlier option experiments also retain .5 as a rescore. These are experimental tolerances, without an application-specific safety justification.

The primary reference stays at $h_p(q_0)$. A geometrically qualified joint trajectory moves secondary coordinates on that primary level set for 3 s and then holds them fixed. Saved joint witnesses establish sampled geometric reachability, not successful dynamic tracking or recoverability from every disturbed state. Episodes last 6.4 s. Faults attenuate newly rebased position increments on one or two joints; repair restores effectiveness without resetting physical state or history. This model is not a mechanically locked joint or an identified hardware failure model.

A bounded command is not a physical task guarantee. The given damped coordinate allocator constructs u^p, u^s, first bounding the primary channel and then fitting the secondary channel within the remaining command box. Learning only chooses

$$\Delta q^{\mathrm{cmd}} = \ell(u^p + \alpha_{\mathrm{sec}} u^s), \qquad \ell = 1/30 \text{ s}. \tag{2}$$

The current externally supplied effectiveness report enters allocation; the actual injected effectiveness attenuates the increment sent to a finite-force position servo. Five 2 ms PhysX steps implement each 10 ms controller interval. Damping, inertia, nonzero primary error, and force limits mean neither the nullspace identity nor primary invariance is exact. The 1 rad/s bound is a *command-channel* bound: independently audited physical joint speed reaches 1.16358 rad/s in the original feedback test. Appendix A specifies this supplied plant and allocator without relying on a historical experiment.

Outcome definitions. We report sampled primary crossings, chart and joint-limit events, unclipped secondary-error AUC, and recovery separately. At repair time t_R, the next 200 actual controller intervals define the 2 s recovery window; the preceding interval ending at t_R is excluded. A recovery dwell in the full-feedback experiment requires 50 consecutive intervals inside this window with primary, chart, and joint admission and endpoint $S \leq .01$. Its starting time is reported, or 2 s on nonrecovery; an earlier crossing is not erased by later recovery. Full-episode failure counts also check the initial state's primary, chart, and joint admission. The earlier finite-option recovery metric used primary/chart admission; the budget study's feasible-and-recovering sets additionally require joint admission. No joint event occurs in those earlier data, so this distinction does not alter their recovery counts. The full-feedback training cost also includes actual joint-limit violations.

3 Full-Feedback Learning of Secondary Retention

The policy chooses secondary retention throughout the healthy, faulty, and repaired phases using the given controller and current fault report. The experiment tests this feedback interface against causal scalar rules.

Information and action clock. At actor time k, the policy samples $a_k \sim \pi_\theta(\cdot \mid o_k)$ and sets $\alpha_k = a_k/20$, $a_k \in \{0, \dots, 20\}$, for the next five controller intervals. Evaluation uses the largest-logit category. The 121-dimensional observation contains current joint state, the six-coordinate primary mask, supplied reference rate and effectiveness report, prepared channels, and four completed command–response pairs. Its error and tolerance fields use the requested tolerance $\tau = c\tau^0$, namely e/τ and τ/τ^0, for $c \in \{1, .25, .1\}$. This differs from the earlier repair-only bank's base-tolerance encoding. Future fault and repair times and future reference sequences beyond the supplied one-interval feedforward are excluded. The finite history is not assumed to be Markov. There is no action filter and no forced-zero fault phase.

Two matched learning objectives. Let S_t be endpoint mean squared secondary error normalized by the fixed task scales, and let v_t indicate any primary-threshold, Euler chart, or joint-limit violation among the five actual 2 ms states in interval t. Primary ratios are compared strictly to one; joint limits allow a 10^{-6} rad numerical tolerance. The declared interval reward and cost are

$$r_t = -.01\{\min(S_t, 100) + .002(\alpha_t - \alpha_{t-1})^2\}, \qquad z_t = .01 v_t. \tag{3}$$

Both variants use the same categorical PPO actor and independent task and cost critics. The fixed-penalty variant uses $\lambda_g = 10$; the adaptive variant uses one multiplier for each of the $62 \times 3 = 186$ mask–tolerance groups g. Separate reward and cost advantages use generalized advantage estimation (Schulman et al., 2016); they are combined as $\widehat{A}^r_k - \lambda_g \widehat{A}^z_k$, normalized over the rollout, and used in

PPO's clipped objective (Schulman et al., 2017). After collecting a group episode, its next multiplier is

$$\lambda_g^+ = \mathrm{clip}_{[0,1000]}\left(\lambda_g + 10\sum_{t=0}^{639} .995^t z_t\right). \tag{4}$$

The actor update uses the pre-update multiplier. Five interval signals are aggregated with weights $.995^j$, and the actor discount is $.995^5$; the finite task ends with zero bootstrap value.

This is a controlled comparison of standard penalty-based policy learning, not a new constrained optimization algorithm. Expected-cost constraints and multiplier adaptation are established approaches (Achiam et al., 2017; Stooke et al., 2020). We use neither CPO's constrained update nor PID multiplier feedback. Finite sampling, bounded multipliers, and the damped physical allocator provide no pointwise safety certificate. The experiment must establish an advantage over causal scalar feedback baselines; an advantage over an earlier fixed recovery option alone would not establish the value of full-feedback learning.

4 Paired Learning and Prospective Evaluation

All learning, ablations, and comparative physical rollouts in this section use Isaac Sim; the two hardware records are identified separately. Each initialization trains two penalty variants with three paired seeds. Every run uses $62 \times 42 \times 640 = 1{,}666{,}560$ physical controller intervals; each mask–scale group receives three healthy and eleven faulted episodes. Geometry, optimization, and the fixed budget are shared; on-policy trajectories can diverge.

Four development scenarios select one checkpoint from updates $\{12, 24, 36, 42\}$ per seed/variant; audit snapshots are not extra candidates. Neutral choices are 24/24/36 for fixed penalties and 24/36/24 for condition dual. The same validation set selects strong 20 Hz controls: five constants, nine phase rules, and six margin settings, with one duplicate leaving nineteen unique candidates. One global winner per family and a mask–tolerance lookup among all nineteen are compared; the primary comparator is the lookup. All seeds are retained.

The neutral experiment reserves sixteen fresh scenarios distinct from the original test. All 62 masks and three tolerances are seen in training; generalization is to scenarios, not masks. Tolerance-independent controls may reuse physical traces for rescoring; PPO and margin controls execute each requested scale. This feedback comparison includes no planner.

The unchanged primary criterion at $c = .25$ requires each dual seed to have zero full primary/chart/joint failures and aborts, no recovery-count loss against the lookup, and a strictly negative paired 16-scenario 95% interval for mean-three-model minus lookup full-horizon AUC. All contexts remain included. A separate phase-control comparison, registered after the original test and neutral training but before the new test, retains both variants with nominal 97.5% intervals for its two contrasts. It additionally requires the phase and every fitted seed to have zero full failures and aborts, and no seed-level recovery-count loss. It cannot replace the primary criterion. Exact rules, registration timing, and uncertainty units appear in Appendix B. Native post-repair comparisons include earlier actions and are total policy effects, not matched-state restoration effects.

5 A Trained Stochastic Policy Can Deploy One Action

The original actors initialize the full-retention logit bias to four. All six selected models use checkpoint 12. Let the final hidden tanh activation be $z(o) \in [-1, 1]^{128}$ and the logits be $l_j(o) = W_j z(o) + b_j$. For every competitor $j < 20$,

$$l_{20}(o) - l_j(o) \geq b_{20} - b_j - \|W_{20} - W_j\|_1 =: B_j. \tag{5}$$

Thus $\min_j B_j > 0$ makes category 20 ($\alpha = 1$) the unique argmax for every valid hidden activation. This elementary bound is a property of deployed actions, not a new theorem or a safety certificate. After conservative float32/float64 roundoff allowances, all six minima remain positive, from 1.459 to 1.986. The claim assumes finite tanh outputs and ordinary float32 head arithmetic; it excludes NaNs, corrupted evaluation, and changed precision or architecture (Appendix C).

Table 1: Same-input initial-to-selected mapping comparison. Ranges cover all three retained fits per variant; each uses its own selected-policy validation inputs. Category counts concern these inputs, not initial-policy physical trials. Exact per-fit values and probability changes appear in Table 6.

Variant	Selected updates	Changed argmax (%)	Categories: initial / selected
Fixed penalty	24/24/36	98.238–99.328	20–21 / 16–18
Condition dual	24/36/24	96.938–97.617	20–21 / 15–17

Independent optimizer replay confirms that training changes weights and stochastic probabilities. Nevertheless, the selected deterministic actions cannot depend on mask, tolerance, fault report, or history. All saved test actions agree with the bound. This distinction is stronger than merely observing an approximately constant policy on a finite test set.

The original test completes 25,792 physical episodes without abort. At quarter tolerance each actor crosses in 4/992 tasks, all of which recover. Mean-three-dual minus lookup full AUC is $-.0012077$ (95% interval $[-.0034493, 0]$), so the primary criterion fails. The phase rule has zero crossings and 992 recoveries, but higher AUC ($.28631$ versus $.14689$). Appendix D reports quarter results and all-scale event/recovery counts; complete window tables and contrasts remain in the source archive (Appendix I). The bound diagnoses the selected initialization/training/decoder combination; it neither isolates the initial bias's effect nor invalidates RL.

6 Initialization Changes the Deployed Map

The original actor initializes the category-20 logit bias to four. The intervention sets only that bias to zero after identical network construction, preserving the other initial tensors, architecture, critics, optimization, objective, geometry, budget, and argmax rule. Six paired fits retain both penalty variants and all seeds. The other nonzero output weights remain; neutral bias is not uniform logits.

Independent execution checks verify the initial tensor difference and 252 matching rollout-start Torch RNG states. Different action distributions still produce different exploration, trajectories, and gradients. The same validation rule now selects updates 24 or 36 rather than 12; this is a mediator of the complete training/selection intervention, not a same-update direct bias effect. Same-input model checks distinguish a changed learned map from an unchanged initial random map (Appendix C.1). They do not estimate an initial policy's physical return or identify useful history features. The six original selected actors share one deterministic $\alpha = 1$ reference on the fresh cohort, not six independent samples.

The mapping changes, but nonconstancy was not created by learning. We compare each initial actor with its selected checkpoint on exactly the same 95,232 validation observations visited by that selected policy: 571,392 correlated inputs in total, not independent experimental units. Initial neutral actors already use 20–21 categories; selected actors use 15–18. Argmax changes on 96.94–99.33% of inputs, with changed probabilities and logit gaps (Table 1). This describes optimization and selection changing a map. It does not measure initial-policy physical performance or prove that history is useful. On the fresh quarter-tolerance test, 451–772 of 992 episodes per neutral fit actually contain action changes, versus zero for the original reference. Nonconstant deployment is established separately from the control outcomes below.

7 Does the Changed Policy Improve Control?

The fresh test completes all 432 physical batches: 26,784 episodes, 17,141,760 controller intervals, and zero aborts. Tolerance rescores produce 32,736 episode records, not additional physical trials. Table 2 retains every fit and simple comparator.

The primary criterion fails, with a stronger simple control visible. Dual fits have 6/7/12 full primary failures and 968/970/987 recoveries; the lookup has 12 and 991. Their mean-three-model AUC difference is $+.56927$ (95% interval $[.37448, .78111]$). Every dual seed fails the zero-event and recovery conditions, and the cost interval is positive. The margin rule has three failures, 991

Table 2: Neutral-initialization feedback test at the prespecified quarter tolerance. Each row uses the same 16 scenarios and 62 masks (992 episodes); all three fitting seeds are retained. The original six actors share one exactly identical deterministic policy; their reference row is one physical sample, not six replicates. Failures include initial and actual 2 ms primary/chart/joint events. Recovery is a separate post-repair dwell; AUC is full-horizon secondary error.

Policy	Full failures	Recoveries	Mean AUC
Selected constant	0	758	4.66312
Selected phase rule	3	992	0.52622
Selected margin rule	3	991	0.25643
Mask–tolerance lookup	12	991	0.22261
Original selected actors (shared)	15	991	0.21916
PPO fixed (1)	10	978	0.74821
PPO fixed (2)	11	958	0.51017
PPO fixed (3)	12	991	0.64370
PPO dual (1)	6	968	0.95623
PPO dual (2)	7	970	0.92683
PPO dual (3)	12	987	0.49260

recoveries, and AUC .25643: it weakly improves all three aggregate criteria over *every* neutral fit, with strictly fewer failures and lower cost. This observed aggregate dominance is not per-context dominance or a population bound. Against the original reference's 15 failures, neutral fits reduce failure counts but have higher cost and no higher recovery count: a tradeoff, not a learning win. Fewer failures do not imply casewise improvement: fixed (3) and dual (2) each introduce one new crossing; fixed (3)'s tied 991 recoveries hide one gain and one loss (Table 11). No chart or joint event occurs.

The secondary phase comparison also fails. The phase rule itself has three full primary failures, although all 992 tasks recover. Mean fixed/dual minus phase AUC is $+.10780/+.26566$; the two nominal 97.5% intervals are $[.05354, .17168]$ and $[.18774, .34995]$. Fitted seeds lose 1–34 recoveries, with no paired gains. Both separately registered criteria therefore fail independently of the lookup comparison. The old cohort's zero phase failures do not transfer to this fresh cohort. Appendix D.1 reports all-scale counts and key paired outcomes. All 45 contrasts and complete recovery/action-dynamics records remain in the source archive identified in Appendix I.

8 Distribution Replacement and Physical Control Value

Variable argmax actions do not describe the full trained categorical policy. We therefore compare the selected distribution and its mode on exactly the same stored development inputs, holding weights and inputs fixed. Let $m(o)$ denote the deployed modal alpha and $p(a \mid o)$ the diagnostic categorical probabilities on the 21-point action grid. Define

$$D(o) = 1 - p(m(o) \mid o), \qquad W(o) = \sum_a p(a \mid o)|a - m(o)|. \tag{6}$$

D is the total-variation distance to deterministic deployment and W is the expected absolute alpha difference. These elementary distribution distances do not measure a physical return difference.

Across the six selected models, we replay 571,392 validation inputs visited by their respective modal policies. The prespecified materiality panel uses the three fixed-penalty fits at quarter tolerance during the causally reported-fault phase. Each scenario first averages within each mask episode, then weights all 62 masks equally. All three fits pass $D \geq .50$ and $W \geq .10$ in each of the four scenarios, exceeding the required two fits with three passing scenarios each. Across these twelve cells, D ranges from .8064 to .9023 and W from .2137 to .4382 (Table 12). The dual fits and other phases/scales are retained, not used to replace the prespecified panel.

Same-weight native intervention. To measure the missing physical consequence, we retain all six weights and compare mode deployment with two predetermined categorical action streams on four fresh *development* scenarios at $c = .25$. Every arm restarts the complete episode with the

Table 3: Same-weight native deployment at $c = .25$: four new development scenarios, all 62 masks (248 contexts per arm). F/D denote fixed/dual fits. J is full-horizon secondary AUC; the sampling mean averages two streams within scenario. Every mode arm has zero full failures. Sample failures (V) and recoveries (R) remain separate for streams S0/S1. The last column counts scenarios with at least 10% lower sampling-mean J; no fit meets the full useful-decoder criterion.

Fit	J_{mode}	$\overline{J}_{\text{sample}}$	$V_{\text{S0/S1}}$	R_{mode}	$R_{\text{S0/S1}}$	Cost scenes
F-9201	1.3595	1.2097	2/5	242	248/248	3/4
F-9202	0.7302	1.3072	5/10	239	248/247	0/4
F-9203	0.8772	1.0769	0/4	248	248/248	0/4
D-9201	1.5427	1.2536	2/6	242	248/248	3/4
D-9202	1.1622	1.1796	2/7	242	248/247	1/4
D-9203	0.8369	1.0923	1/2	246	248/248	0/4
Margin	0.2414		0	248	—	—

same prescribed initial, reference, fault, and servo conditions; later states follow its actions. All 62 masks are retained, alongside one shared frozen margin reference: 19 methods, 76 banks, and 4,712 physical episodes. Sampling uses native float32 Categorical probabilities with an explicit float64 inverse CDF; common uniforms are fixed by scenario and stream, not outcome. This is not a replay of training randomness or a matched-repair-state experiment.

The prespecified primary criterion requires a 10% full-AUC reduction both overall and in at least three of four scenarios, with no additional full failures or recovery-count loss in *either* stream. At least two of three fixed-penalty fits must qualify; dual fits cannot rescue the criterion. Means retain all crossing cases. Exact sampling and the separate margin comparison appear in Appendix E.1.

No fixed-penalty fit qualifies. Sampling-mean AUC decreases in two fits and increases in four. F-9201 and D-9201 meet the cost-only condition and recover all 248 contexts in each stream, but both introduce crossings. Across all twelve sampled arms there are 46 crossing episodes; eleven arms have at least one, whereas all six modal arms have zero on this cohort. F-9203/S0 has zero, but S1 has four. The margin reference has zero failures, 248 recoveries, and $J = .24137$, lower than every sampled arm. Both the raw and matched-observed-safety secondary criteria fail. The conditional confirmation study is consequently not activated.

Holding weights fixed identifies a total deployment-policy effect, including induced state-distribution changes. It does not show that every cost worsens, that learned probabilities outperform untrained randomization, or that history features provide value. The large D/W screen establishes a material action-law replacement, not a predictor of constrained-control benefit.

9 Complementary Scheduling Evidence and Its Limits

Scheduling opportunity itself requires evidence. Under static references, initially exact primary tracking, an exact undamped priority nullspace, ideal velocity following, and an autonomous locally unique secondary field, $\dot{q} = \alpha f(q)$ merely reparameterizes one trajectory in time. With exact $\dot{e}_s = -\alpha K_s e_s$ for positive diagonal K_s, full retention weakly dominates slower schedules. The complete chart, rank, joint-domain, and dynamical assumptions appear in Appendix F; this ideal limit is not the simulated plant.

Fixed predictors, changed selection. Separate studies make one choice among four fixed recovery continuations after replaying identical healthy-one/fault-zero prefixes. Immediate restoration crosses in 31/248 quarter-tolerance development cases versus zero for a half-second ramp, while both recover. Outcome regression then loses progress through excessive holding. A prospective fixed-model selection intervention on eight fresh scenarios redirects holding to ramp and removes all 87/91/106 recovery deficits without adding a post crossing. Its anchored ensemble beats ramp AUC (.35008 versus .41763), yet the training-only mask lookup (.22947) and fixed half (.30930) remain stronger. This identifies a selection-induced loss without establishing neural superiority. These one-decision results are not same-interface baselines for PPO. Appendix F.1 summarizes the key results; complete method/scale/context records remain in the source archive (Appendix I).

Identical repair context, changed physical budget. Another prospective study replays full prefixes and changes only the post-repair configured force budget ($\beta = 1$ versus 4), hidden from the repair observation. Across 496 contexts in eight scenarios, immediate quarter-tolerance post crossings fall from 81 to 42, including 40 rescued and one new crossing; this is not pointwise improvement. Although 41/496 contexts have disjoint .01-near-optimal finite-option sets, all 496 retain a common post-repair-admitted recovering option. One shared prefix already crosses, so this is not full-episode safety. Appendix G retains the paired physical contrasts and finite-option boundary; the plot and complete policy comparisons remain in the source archive (Appendix I). These interventions show that physical headroom affects option value without identifying a unique saturation mechanism. They neither explain nominal PPO's remaining loss nor show that nominal feedback resolves hidden-budget ambiguity.

Two archived hardware trials. To connect the task interface to a physical setting, we retain two recorded Panda trials of a horizontal circle with XYZ priority and the J2/J1/J5 fault schedule at 10/20/30 s. Appendix J shows the platform, measured errors, and logged α_{sec}. The archived Panda records come from a separately deployed offline PPO A separate, explicitly simulated task-informed replay is also reported in the appendix.

10 Related Work and Contribution Boundary

Implementation sensitivity in on-policy RL is established: Engstrom et al. (2020) show that implementation choices can substantially change PPO/TRPO comparisons and agent behavior, and Andrychowicz et al. (2021) systematically study such choices in a shared actor–critic framework. We do not claim initialization sensitivity itself as new. Our narrower contribution links a verified bound on deterministic deployed actions to a prospective initial-bias intervention, same-input initial-to-selected mapping comparisons, and separately measured physical constraints, recovery, and cost. This intervenes on the complete training/selection procedure, not only its decoder, and does not establish a general explanation of PPO failure. Stochastic training followed by deterministic deployment is also studied directly in constrained policy gradients (Montenegro et al., 2025). Their C-PG guarantees require specified noise and regularity assumptions; they do not transfer to our categorical PPO argmax or sampled 2 ms constraint checks. We contribute a bounded same-weight physical comparison, not the distinction itself or a general decoder-benefit criterion.

Learning within task hierarchies predates this study. Policy search can operate below safety/alignment tasks (Lundell et al., 2018); priorities can be learned from demonstrations (Silvério et al., 2019), adapted with RL (Karimi & Ahmadi, 2022; Testa et al., 2025), or managed using genetic programming (Adami et al., 2025). Nullspace saturation already combines priority and joint-constrained task scaling (Flacco et al., 2015). Penco et al. (2020) use multiobjective optimization to learn strict/soft priorities and gains. We keep that hierarchy and its gains fixed and learn only a secondary coefficient; neither hierarchical learning nor this restricted channel is claimed as a first contribution.

PPO supplies the implemented clipped policy objective (Schulman et al., 2017). Expected-cost constrained policy search and multiplier dynamics are established in CPO and PID Lagrangian methods (Achiam et al., 2017; Stooke et al., 2020). Our grouped integral multiplier update is neither of those algorithms nor a new safety theorem. Predictive filters require explicit model and feasible-continuation assumptions (Wabersich & Zeilinger, 2021); Recovery RL learns a recovery-zone switch (Thananjeyan et al., 2021). The present detector and allocator are given, and physical cost checks do not filter an unsafe action in advance.

The mismatch between predictive loss and decision value also has clear precedents (Wilder et al., 2019; Elmachtoub & Grigas, 2022). Our finite-option regressors use squared error, not those decision-focused objectives. Likewise, empirical ramp defaults do not inherit baseline-performance guarantees from SPIBB or conservative contextual bandits (Laroche et al., 2019; Kazerouni et al., 2017). Our interventions separate decision behavior from physical value without inheriting these methods' guarantees.

11 Discussion and Limitations

Physical scheduling opportunity, predictor accuracy, changed policy probabilities, and nonconstant actions are distinct from useful constrained control. Fixed-model selection interventions repair holding, while stronger simple controls preclude learning superiority. Matched budget changes alter option value without eliminating a common recovering choice. The initialization intervention then changes the nominal feedback learner's action map and deployed dynamics, but does not repair its constrained-performance deficit. Removing constant deployment through this intervention is insufficient for constrained-control gains; the original collapse does not prove PPO impossibility. This total-effect comparison does not isolate the cause of the remaining loss, and neither more training nor a particular feature can be declared a fix.

A same-input diagnostic additionally separates variable modal actions from the categorical law retained in the trained network. Complete same-weight native rollouts show that sampling can change physical recovery and cost while introducing primary crossings. The failed prospective criterion rejects this useful-decoder rationale on its fixed development cohort; it does not exclude other stochastic policies. Distribution distance alone does not predict benefit. The failed fixed-probe opportunity screen (Appendix H.2) also stops that separate finite family without testing an observer or learner. These interventions discriminate specific hypotheses without supplying a complete explanation of nominal PPO's remaining loss.

Separate development screens did not justify the proposed cost-estimator training change or the tested finite residual family (Appendices H and H.1). These negative screens bound those hypotheses, not the value of all possible alpha feedback policies.

The intervention includes changed exploration and checkpoint selection. Validation mapping is not initial-policy physical performance, native test effects are not matched repair-state effects, and a scalar policy's nominal results do not resolve the separate hidden-budget construction. Evidence concerns one simulated robot/servo family, exact current reports, supplied kinematics/gravity compensation, free-space tasks, and declared fault/tolerance distributions. All masks are seen in training. The native decoder comparison has four development scenario clusters and two action streams within each; these are not extra independent mask-level samples or a population risk estimate. Geometric witnesses do not prove general dynamic recoverability; command bounds are not physical velocity guarantees. Neither sampled constraint admission nor either empirical comparison is a safety certificate.

AI Use Statement

OpenAI Codex and associated language-model agents assisted with research framing; proposing and refining hypotheses; experiment and analysis design; developing conditional mathematical arguments and their proofs; literature search and synthesis; implementing and debugging methods, simulation jobs, analysis and verification scripts; processing recorded data; interpreting results; drafting and editing manuscript text; and preparing tables and figures. AI-assisted code generated training and evaluation datasets through simulator execution; the reported numerical trajectories and metrics were obtained from executed simulations or archived hardware logs, not invented by a language model. Separate AI-agent roles also provided critical feedback and executable cross-checks; these roles are not independent human peer reviews. Protocols, source identities and numerical records document automated checks, but do not replace human author review. The authors remain responsible for the final content, claims and artifacts.

REFERENCES

Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 22–31. PMLR, 2017.

Alessandro Adami, Aris Synodinos, Matteo Iovino, Ruggero Carli, and Pietro Falco. Learning Stack-of-Tasks management for redundant robots. *arXiv preprint arXiv:2508.10780v2*, 2025. URL https://arxiv.org/abs/2508.10780v2. Version 2, 23 December 2025.

Marcin Andrychowicz, Anton Raichuk, Piotr Stańczyk, Manu Orsini, Sertan Girgin, Raphaël Marinier, Léonard Hussenot, Matthieu Geist, Olivier Pietquin, Marcin Michalski, Sylvain Gelly, and Olivier Bachem. What matters for on-policy deep actor-critic methods? a large-scale study. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/pdf?id=nIAxjsniDzg.

Adam N. Elmachtoub and Paul Grigas. Smart "predict, then optimize". *Management Science*, 68 (1):9–26, 2022. doi: 10.1287/mnsc.2020.3922. URL https://pubsonline.informs.org/doi/10.1287/mnsc.2020.3922.

Logan Engstrom, Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Firdaus Janoos, Larry Rudolph, and Aleksander Madry. Implementation matters in deep RL: A case study on PPO and TRPO. In *International Conference on Learning Representations*, 2020. URL https://iclr.cc/virtual_2020/poster_r1etN1rtPB.html.

Fabrizio Flacco, Alessandro De Luca, and Oussama Khatib. Control of redundant robots under hard joint constraints: Saturation in the null space. *IEEE Transactions on Robotics*, 31(3):637–654, 2015. doi: 10.1109/TRO.2015.2418582. URL https://www.diag.uniroma1.it/~labrob/pub/papers/TRO_SNS_Jun2015.pdf.

Masoud Karimi and Mojtaba Ahmadi. A reinforcement learning approach in assignment of task priorities in kinematic control of redundant robots. *IEEE Robotics and Automation Letters*, 7 (2):850–857, 2022. doi: 10.1109/LRA.2021.3133934. URL https://doi.org/10.1109/LRA.2021.3133934.

Abbas Kazerouni, Mohammad Ghavamzadeh, Yasin Abbasi-Yadkori, and Benjamin Van Roy. Conservative contextual linear bandits. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/bdc4626aa1d1df8e14d80d345b2a442d-Paper.pdf.

Romain Laroche, Paul Trichelair, and Rémi Tachet des Combes. Safe policy improvement with baseline bootstrapping. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3652–3661. PMLR, 2019. URL https://proceedings.mlr.press/v97/laroche19a.html.

Jens Lundell, Robert Krug, Erik Schaffernicht, Todor Stoyanov, and Ville Kyrki. Safe-To-Explore state spaces: Ensuring safe exploration in policy search with hierarchical task optimization. In *Proceedings of the IEEE-RAS International Conference on Humanoid Robots*, 2018. URL https://arxiv.org/abs/1810.03516.

Alessandro Montenegro, Leonardo Cesani, Marco Mussi, Matteo Papini, and Alberto Maria Metelli. Learning deterministic policies with policy gradients in constrained markov decision processes. *arXiv preprint arXiv:2506.05953*, 2025. URL https://arxiv.org/abs/2506.05953.

Luigi Penco, Enrico Mingo Hoffman, Valerio Modugno, Waldez Gomes, Jean-Baptiste Mouret, and Serena Ivaldi. Learning robust task priorities and gains for control of redundant robots. *IEEE Robotics and Automation Letters*, 5(2):2626–2633, 2020. doi: 10.1109/LRA.2020.2972847. URL https://doi.org/10.1109/LRA.2020.2972847.

John Schulman, Philipp Moritz, Sergey Levine, Michael I. Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. In *International Conference on Learning Representations*, 2016. URL https://arxiv.org/abs/1506.02438.

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

João Silvério, Sylvain Calinon, Leonel Rozo, and Darwin G. Caldwell. Learning task priorities from demonstrations. *IEEE Transactions on Robotics*, 35(1):78–94, 2019. doi: 10.1109/TRO.2018.2878355. URL https://arxiv.org/abs/1707.06791v3.

Adam Stooke, Joshua Achiam, and Pieter Abbeel. Responsive safety in reinforcement learning by PID lagrangian methods. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9133–9143. PMLR, 2020. URL https://proceedings.mlr.press/v119/stooke20a.html.

Andrea Testa, Marco Laghi, Edoardo Del Bianco, Gennaro Raiola, Enrico Mingo Hoffman, and Arash Ajoudani. A stable method for task priority adaptation in quadratic programming via reinforcement learning. *Robotics and Computer-Integrated Manufacturing*, 91:102857, 2025. doi: 10.1016/j.rcim.2024.102857. URL https://www.sciencedirect.com/science/article/pii/S0736584524001443.

Brijen Thananjeyan, Ashwin Balakrishna, Suraj Nair, Michael Luo, Krishnan Srinivasan, Minho Hwang, Joseph E. Gonzalez, Julian Ibarz, Chelsea Finn, and Ken Goldberg. Recovery RL: Safe reinforcement learning with learned recovery zones. *IEEE Robotics and Automation Letters*, 6 (3):4915–4922, 2021. doi: 10.1109/LRA.2021.3070252. URL https://arxiv.org/abs/2010.15920.

Kim P. Wabersich and Melanie N. Zeilinger. A predictive safety filter for learning-based control of constrained nonlinear dynamical systems. *Automatica*, 129:109597, 2021. doi: 10.1016/j.automatica.2021.109597. URL https://www.sciencedirect.com/science/article/pii/S0005109821001175.

Bryan Wilder, Bistra Dilkina, and Milind Tambe. Melding the data-decisions pipeline: Decision-focused learning for combinatorial optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 1658–1665, 2019. doi: 10.1609/aaai.v33i01.33011658. URL https://ojs.aaai.org/index.php/AAAI/article/view/3982.

A PLANT, GIVEN ALLOCATOR, AND OBSERVATION CONTRACT

The common simulator is a fixed-base, seven-arm-link FR3 articulation in CPU PhysX, using the project's geometry, masses and inertias. Tasks are free-space tracking, without a collision-avoidance claim. The controller runs at 100 Hz, the actor at 20 Hz, and each controller interval contains five actual 2 ms steps. Position/velocity solver iterations are 16/4. Drives have asserted SI stiffness/damping $15000/500$ on all joints; at each substep ideal gravity compensation $g(q)$ is applied separately and drive limits are $\max(F_{\max} - |g|, 0)$, where $F_{\max} = (87, 87, 87, 87, 12, 12, 12)$ Nm. These are configured budgets, not measured total motor torque. The 1 rad/s command cap is not a physical speed guarantee.

Coordinates, task, and faults. Let $h(q) = (\phi, \theta, \psi, x, y, z)$ with $R = R_z(\psi)R_y(\theta)R_x(\phi)$ and anchored angular branches. All nonempty proper six-bit primary masks (codes 1–62) are used; their complements are secondary. The chart requires $|\theta| \le \pi/3$ and anchored angular differences at most $\pi/2$. Primary base tolerance is $\tau^0 = (.02, .02, .02, .01, .01, .01)$ in radians/metres, secondary scales are $\sigma = (.2, .2, .2, .05, .05, .05)$, and feedback uses $c \in \{1, .25, .1\}$. Initial joints are $(0, 0, 0, -1.5, 0, 1.5, 0) + U[-.1, .1]^7$ rad, with zero velocity. Joint limits are

$$q_{\min} = (-2.8973, -1.7628, -2.8973, -3.0718, -2.8763, .4398, -2.8973),$$
$$q_{\max} = (2.8973, 1.7628, 2.8973, -.1169, 2.8763, 3.7525, 2.8973).$$

Primary references stay at $h_p(q_0)$; secondary references move for 3 s then remain fixed. Episodes last 6.4 s. One or two uniformly chosen distinct joints have effectiveness in $\{0, .2, .5\}$, with onset $U[1.3, 1.9]$ s and repair $U[3.5, 4.3]$ s. Nominal reports are exact current effectiveness; future fault/repair times are excluded from inputs.

Exact command construction. For spatial angular and Cartesian Jacobians J_ω, J_x,

$$J_h = \begin{bmatrix} E^{-1} J_\omega \\ J_x \end{bmatrix}, \quad E^{-1} = \begin{bmatrix} \cos\psi/\cos\theta & \sin\psi/\cos\theta & 0 \\ -\sin\psi & \cos\psi & 0 \\ \cos\psi\tan\theta & \sin\psi\tan\theta & 1 \end{bmatrix}. \tag{7}$$

With $e = h^\star - h(q)$, $W = \mathrm{diag}(1, 1, 1, 5, 5, 5)$, and current report $\hat{\lambda}$, define $A = W J_h \,\mathrm{diag}(\hat{\lambda})$, $d = W(\dot{h}^\star + 12e)$, where the supplied rate is $\dot{h}^\star = [h^\star(t + .01) - h^\star(t)]/.01$. Write $B^\dagger_\delta = B^\top(BB^\top + \delta^2 I)^{-1}$. First set $v^p = A^\dagger_{p,.015} d_p$, divide by $\max(1, \|v^p\|_\infty)$, and clip to $[l, u]$ to obtain u^p, with

$$l_j = \max\{-1, (q_{\min,j} + .01 - q_j)/\ell\}, \quad u_j = \min\{1, (q_{\max,j} - .01 - q_j)/\ell\}, \quad \ell = 1/30 \text{ s}. \tag{8}$$

Then $N = I - A^\dagger_{p,.015} A_p$ and $v^s = N(A_s N)^\dagger_{.025}(d_s - A_s u^p)$. Divide v^s by $\max(1, \|v^s\|_\infty/2)$ and take the largest $\beta \in [0, 1]$ satisfying $u^p + \beta v^s \in [l, u]$; set $u^s = \beta v^s$. The policy chooses only alpha:

$$\Delta q^{\mathrm{cmd}} = \ell(u^p + \alpha u^s), \qquad q^{\mathrm{target}} = \mathrm{clip}_{[q_{\min}, q_{\max}]}(q + \lambda \odot \Delta q^{\mathrm{cmd}}). \tag{9}$$

The .035 rad increment and .01 rad joint-margin guards may change the bounded command by at most 10^{-7} rad. Targets are renewed every 10 ms; holding alpha for 50 ms does not freeze the allocator or target. Damping means N is not an exact nullspace projector. Zero effectiveness sets a newly rebased increment to zero; it does not weld the joint or remove its finite-force servo and existing motion.

References and reset. A seeded direction is projected into the initial primary nullspace and scaled to maximum component .7 rad. A 200-point joint path uses four corrections per point and quintic time interpolation $s = 10z^3 - 15z^4 + 6z^5$, $z = \mathrm{clip}(t/3, 0, 1)$. Every emitted 2 ms reference is corrected again. Geometric admission requires weighted primary residual $\le 10^{-9}$, joint margin $\ge .08$ rad, the stated chart, and final normalized secondary RMS excursion $\ge .2$; at most eight ordered direction attempts are allowed. Rejected attempts are retained; failure stops the batch. Witnesses establish sampled geometry, not dynamic recoverability. Singular/nonfinite corrector rows are handled individually under a disclosed pre-physics amendment, without replacing scenario IDs or allowing inactive placeholders to qualify.

Reset reads back actual joint state, recomputes error/channels, sets previous alpha to one and four completed-transition histories to zero. Feedback sets $\tau = c\tau^0$ before reset. Option banks instead

Table 4: The shared 121-dimensional observation, in its exact one-based field order. All fields are clipped to $[-20, 20]$ and stored as float32. Only the finite-option predictor applies its training-only standardization and appends a four-way option indicator; PPO receives the clipped observation directly.

Indices	Field	Encoding at the current control boundary
1–7	Joint position	$(q - q_{\mathrm{mid}})/q_{\mathrm{halfwidth}}$, using joint limits.
8–14	Joint velocity	Measured $\dot{q}$, in rad/s; no additional scale.
15–20	Coordinate error	e/τ: $\tau = \tau^0$ for option banks; requested $c\tau^0$ for PPO.
21–26	Supplied reference rate	Supplied one-interval finite difference divided by $(.5, .5, .5, .15, .15, .15)$; zero after the reference stops.
27–32	Primary mask	The six binary entries of M.
33–38	Tolerance encoding	τ/τ^0: ones in option banks, $c\mathbf{1}$ for full-feedback PPO.
39–45	Effectiveness report	Current supplied $\hat{\lambda}$; all ones at initial/repair boundaries, current attenuation during a reported fault.
46–52	Primary channel	Prepared, bounded u^p in rad/s.
53–59	Secondary channel	Prepared, bounded u^s in rad/s.
60	Primary conditioning	Smallest singular value of $[W J_h \operatorname{diag}(\hat{\lambda})]_p$.
61	Previous retention	Coefficient from the last completed interval.
62–121	Four completed transitions	Oldest to newest, 15 fields each: seven issued increments divided by .035 rad, seven actual joint displacements divided by .035 rad, and the executed coefficient.

encode e/τ^0 and six tolerance ones; requested c only enters selection and rescoring. Finite history is not assumed Markov, and clipping can discard error magnitude. The given reference, report, gravity model, kinematics and allocator are not learned contributions.

B Training, Selection, and Evaluation Protocol

Both initializations train separate actors and two critics with Tanh layers 121–128–128; actors have 21 logits for $\alpha \in \{0, .05, \ldots, 1\}$. The sole initial tensor change is category-20 bias $4 \to 0$, retaining output gain .01 and all other seeded draws. Neutral logits are not exactly uniform. Each initialization has fixed-penalty and condition-dual variants at seeds 9201/9202/9203. Every run executes $62 \times 42 \times 640 = 1{,}666{,}560$ controller intervals (333,312 actor decisions); all six runs complete 9,999,360 intervals. For zero-based update u and mask index i, training scenario seed is $10^5 s + 62u + i$, mask is $i + 1$, scale index is $(i + u) \bmod 3$. Updates with $u \bmod 5 = 0$ are healthy: each mask–scale condition gets three healthy and eleven faulted episodes. Paired geometry and initial RNG states do not imply equal actions or trajectories after policies differ.

Optimization and physical signals. Separate Adam optimizers use learning rate 3×10^{-4}, $\epsilon = 10^{-5}$, and gradient-norm bound .5. PPO uses four epochs, minibatches of 1,024, ratio clip .2, entropy coefficient .001, and GAE parameter .95. Define endpoint secondary cost $S_t = |Q|^{-1} \sum_{i \in Q} (e_{t+1,i}/\sigma_i)^2$. An interval violation is the union of any actual 2 ms primary crossing, chart exit, or joint-limit excess above 10^{-6} rad. The physical signals are those in Eq. 3; a violating interval incurs .01 s cost, not an estimate of continuous crossing time. For $t = 5k$, aggregate either interval signal x as

$$X_k = \sum_{j=0}^{4} \gamma_p^j x_{5k+j}, \qquad \gamma_p = .995, \quad \gamma_a = \gamma_p^5. \tag{10}$$

Thus actor-level discounting preserves the full physical discounted sum. At the true 6.4 s terminal boundary both bootstraps are zero. Separate reward/cost TD errors form truncated GAE; critics fit unnormalized return targets. Actor advantages subtract the old group multiplier times cost GAE, then normalize globally. PPO maximizes the clipped surrogate plus entropy; separate optimizers prevent a critic gradient from setting actor clipping but do not establish adequate optimization.

Each of 186 mask–scale multipliers starts at 10. Fixed penalties retain 10; adaptive multipliers use step 10, zero cost budget and cap 1,000, updating only sampled groups from their discounted physical cost. Zero-cost episodes do not decrease this nonnegative-cost multiplier. The finite budget

implies $J_z < 1.920$ and $\lambda_g \leq 10 + 14 \cdot 10 \cdot 1.919128 < 278.678$; the cap cannot bind. This integral rule is neither CPO nor a safety guarantee.

Eligible checkpoints and controls. Only post-update checkpoints 12/24/36/42 enter selection; other saved states are audit data. Validation uses scenarios 1040000–1040003, all masks and scales, and selects every seed separately by: aborts; chart/joint failures; maximum mask–scale failure count; total then quarter failures; all-scale then quarter integrated primary excess; quarter then all-scale AUC; earlier update. Original choices are all 12; neutral choices are fixed 24/24/36 and dual 24/36/24. There is no best-seed or test-driven checkpoint choice. Changed exploration and checkpoint selection are mediators of the initialization intervention.

The same 20 Hz causal control grid contains constants $\{0, .25, .5, .75, 1\}$, nine distinct fault-retention/repair-ramp rules, and six margin rules. The phase grid uses fault retention $0, .5, 1$ and ramps $0, .1, .5, 1$ s, with one canonical full-retention rule. Margin threshold pairs are $(.25, .6), (.4, .8), (.6, .9)$, upward rates 2/5 per second, and .1 s positive-slope lookahead. One duplicate leaves 19 unique controls. The mask–scale lookup chooses among all 19, minimizing aborts, chart/joint events, any-constraint failures, integrated excess, AUC, and canonical tie order. Validation also selects global family winners: constant .25; phase fault .5 with .1 s repair ramp; margin $(.6, .9)$ with upward rate 2. The lookup is the fixed primary comparator.

Outcomes and uncertainty. Original test IDs are 1050000–1050015; neutral IDs are 1060000–1060015. Each method/scale has 992 contexts. All masks/scales are seen in training; the tests concern new scenarios. Constant/phase trajectories may be rescored at different tolerances; these are not extra physical episodes. Full, fault, and actual repaired two-second windows retain all contexts, including crossings and initial admission. Recovery requires 50 consecutive repaired intervals with all substeps primary/chart/joint-admitted and endpoint $S \leq .01$, wholly within 200 intervals. Native policies can have different repair states; post comparisons are total policy effects.

The primary quarter criterion requires every dual seed to have zero full primary/chart/joint/abort events and recovery count at least lookup, plus a strictly negative upper endpoint for the mean-three-minus-lookup 95% full-AUC interval. Paired bootstrap uses 10,000 draws of 16 scenario clusters, preserving every mask and fitted model, seeds 1050999/1060999 for original/neutral cohorts. Intervals condition on the retained fits. Both primary criteria fail.

The neutral phase addendum was registered after original test and neutral training, before new test outcomes. It requires phase and every seed to have zero full failures, no seed-level recovery-count loss, and a strictly negative nominal 97.5% interval for each of two variant-minus-phase contrasts (Bonferroni coverage under the percentile-bootstrap approximation). All 992 contexts and paired gains/losses are retained; it cannot replace the lookup criterion. Both phase criteria fail. Neither finite zero counts nor expected-cost optimization certifies physical safety.

Table 5: Independent selected-actor bound, before reading test outcomes. All models are validation-selected update 12. Positive conservative logit-gap bounds establish constant $\alpha = 1$ under the stated arithmetic conditions; they are not task-performance or safety bounds.

Actor	Real-valued bound	Float32-guarded bound
Fixed penalty 9201	1.794743	1.794642
Fixed penalty 9202	1.459367	1.459260
Fixed penalty 9203	1.986402	1.986301
Condition dual 9201	1.645383	1.645278
Condition dual 9202	1.585256	1.585152
Condition dual 9203	1.935837	1.935736

Table 6: Descriptive initial-to-selected mapping on validation inputs, not initial-policy physical outcomes. Each row uses 95,232 observations visited by its selected policy. Categories and top-two gaps list initial/selected values. KL is $D_{\mathrm{KL}}(\pi_{\mathrm{selected}} \| \pi_{\mathrm{initial}})$ in nats. Correlated decision inputs are not independent replications.

Variant / seed	Update	Changed (%)	Categories	Mean KL	Mean gaps
Fixed / 9201	24	98.238	21/16	0.20986	0.0026/0.5043
Fixed / 9202	24	98.615	21/18	0.08397	0.0021/0.3191
Fixed / 9203	36	99.328	20/16	0.40172	0.0021/0.2837
Dual / 9201	24	96.938	21/15	0.16695	0.0026/0.5023
Dual / 9202	36	97.617	21/15	0.29622	0.0022/0.3681
Dual / 9203	24	97.211	20/17	0.35067	0.0021/0.1869

C Deployed-Action Bound and Mapping Diagnostics

For finite hidden activations $z \in [-1, 1]^{128}$, logits $l_j = W_j z + b_j$ obey

$$l_{20} - l_j \geq b_{20} - b_j - \|W_{20} - W_j\|_1 =: B_j, \tag{11}$$

because every coordinate product is at least its negative absolute weight. Hence $\min_{j<20} B_j > 0$ makes alpha one the unique modal action for every valid input. For ordinary float32 head arithmetic the audit subtracts

$$[\gamma_{260}(2^{-24}) + \gamma_{522}(2^{-53})]A_j + 10^{-12}, \quad A_j = |b_{20}| + |b_j| + \|W_{20}\|_1 + \|W_j\|_1, \tag{12}$$

where $\gamma_k(u) = ku/(1 - ku)$, and rounds the lower endpoint outward. This covers both 128-term head dot products, biases and audit arithmetic; it excludes invalid activations, NaNs and changed precision/architecture. All six selected update-12 bounds stay positive. This elementary action bound is not a new theorem or a physical safety bound. Optimizer replay and changed probabilities remain compatible with constant argmax.

C.1 Initial-to-selected map

The true pre-update-one neutral actor and selected actor are evaluated on the same inputs visited by that selected policy: 95,232 observations per fit (four scenarios, three scales, 62 masks, 128 decisions), totaling 571,392. The initial maps already emit 20–21 categories; selected maps emit 15–18, while changing 96.94–99.33% of actions. Probabilities, gaps and KL also change. These are fixed correlated inputs, not physical initial-policy rollouts or proof that training created nonconstancy, made history useful, or improved control.

All selected actions reproduce their recorded decisions. Six exported TorchScript actors also match original inference on 571,392 validation inputs and 4,536 batch-one checks; a separate physical audit checks 142,848 held-out inputs and 1,134 batch-one cases. These checks establish software inference parity on the recorded inputs.

Table 7: Original bias-four cohort: native full-feedback test at the prespecified quarter tolerance. Each row uses the same 16 scenarios and 62 masks (992 episodes); all three fitting seeds are retained. Failures include initial and actual 2 ms primary/chart/joint events. Recovery is a separate post-repair dwell; AUC is full-horizon secondary error.

Policy	Full failures	Recoveries	Mean AUC
Selected constant	0	804	3.67805
Selected phase rule	0	992	0.28631
Selected margin rule	2	992	0.15699
Mask–tolerance lookup	4	992	0.14810
PPO fixed (1)	4	992	0.14689
PPO fixed (2)	4	992	0.14689
PPO fixed (3)	4	992	0.14689
PPO dual (1)	4	992	0.14689
PPO dual (2)	4	992	0.14689
PPO dual (3)	4	992	0.14689

Table 8: All-scale full failures / recoveries (F/R) on the 992 matched episode slots per method and tolerance. Recovery does not erase an earlier failure. All six PPO runs remain visible despite identical deterministic behavior.

Policy	$c = 1$: F/R	$c = .25$: F/R	$c = .1$: F/R
Selected constant	0 / 804	0 / 804	0 / 804
Selected phase rule	0 / 992	0 / 992	12 / 992
Selected margin rule	0 / 992	2 / 992	40 / 992
Mask–tolerance lookup	0 / 992	4 / 992	19 / 979
PPO fixed (1)	0 / 992	4 / 992	28 / 992
PPO fixed (2)	0 / 992	4 / 992	28 / 992
PPO fixed (3)	0 / 992	4 / 992	28 / 992
PPO dual (1)	0 / 992	4 / 992	28 / 992
PPO dual (2)	0 / 992	4 / 992	28 / 992
PPO dual (3)	0 / 992	4 / 992	28 / 992

D Original and Neutral Feedback Results

Original-bias evaluation completes 416 physical banks / 25,792 episodes, with no aborts; tolerance rescores produce 29,760 records. All six selected actors share identical deterministic behavior, but are retained as six fitted models. The primary mean-three-dual-minus-lookup AUC effect is $-.0012077$ with 95% interval $[-.0034493, 0]$, failing the strictly negative upper-endpoint condition as well as zero-failure admission. Quarter primary counts per actor are four full, one fault, three post; recovery in all cases does not erase crossings. These original-cohort means are not paired initialization effects on the later cohort.

D.1 Neutral initialization on fresh scenarios

All 432 banks / 26,784 physical episodes complete without abort, yielding 32,736 records including declared tolerance rescores. The main text retains all eleven quarter rows. The unchanged primary effect is $+.569273886$ with 95% interval $[.374480630, .781111777]$. Every dual seed also has crossings and fewer recoveries than lookup. Both secondary phase comparisons fail; phase itself has three quarter crossings despite recovering all 992 cases. There are no chart or joint events in the complete test. Fault/post counts can overlap and cannot be summed to obtain full failures.

Paired outcomes and physical scope. The margin rule weakly improves all three aggregate quarter criteria over every neutral fit, not every individual context. Against original alpha one, fewer aggregate crossings can conceal new failures; fixed (3) and dual (2) each introduce one. Fixed (3)'s tied recovery count hides one gain and one loss. The following table preserves that distinction.

All raw banks and logged-error metrics are checked. Independent deep FK/controller/observation/action reconstruction covers the prespecified 26 original and 27

Table 9: Complete fresh-cohort full failures/recoveries at all scales. Every cell retains 992 tasks; counts refer to physical admission and the separate recovery dwell. Three scale scores of a tolerance-independent trace are not three physical episodes.

Policy	$c = 1$: F / R	$c = .25$: F / R	$c = .1$: F / R
Selected constant	0 / 758	0 / 758	6 / 758
Selected phase	0 / 992	3 / 992	39 / 992
Selected margin	0 / 991	3 / 991	49 / 990
Primary lookup	4 / 991	12 / 991	33 / 973
Original actors (one policy)	4 / 991	15 / 991	40 / 991
PPO fixed (1)	0 / 989	10 / 978	48 / 964
PPO fixed (2)	1 / 976	11 / 958	42 / 918
PPO fixed (3)	2 / 992	12 / 991	44 / 989
PPO dual (1)	0 / 987	6 / 968	47 / 949
PPO dual (2)	1 / 987	7 / 970	55 / 965
PPO dual (3)	4 / 991	12 / 987	45 / 980

Table 10: Separately frozen secondary comparison with the selected phase rule at quarter tolerance. Counts list all three fitting seeds, each over all 992 contexts. Intervals resample 16 paired scenarios and use nominal Bonferroni adjustment for the two secondary contrasts. This analysis does not replace the original lookup primary gate.

Variant	Failures	Recoveries	Δ AUC	97.5% interval
PPO fixed	10/11/12	978/958/991	0.10780	[0.05354, 0.17168]
PPO dual	6/7/12	968/970/987	0.26566	[0.18774, 0.34995]

neutral test banks, rather than claiming full FK reconstruction of every test state. Original actual speed reaches 1.16358 rad/s; no actual drive torque is recorded. Full windows, all 36 original and 45 neutral AUC contrasts, all 30 neutral binary comparator pairs, and action-dynamics records remain in the unchanged source tables and JSONs named in Appendix I. Only the declared primary and phase intervals have their stated confirmatory roles; the additional comparisons are descriptive.

E Categorical-to-Mode Diagnostic and Native Intervention

For the six selected neutral actors, the same-input diagnostic replays 72 validation banks / 571,392 modal-policy observations. Native float32 logits reproduce all stored modes. Descriptive float64 softmax gives $D = 1 - p(m \mid o)$ and $W = \sum_a p(a \mid o)|a - m|$. The quarter reported-fault panel averages within each mask episode then equally over all 62 masks. All three fixed fits meet $D \geq .50, W \geq .10$ in all four scenarios, exceeding the prescribed two-fit/three-scenario criterion. This establishes material distribution replacement, not benefit. Native float32 Categorical weights renormalized in float64 differ from diagnostic softmax by at most 3.38×10^{-8} per category and 6.37×10^{-8} TV; this is not bit-identical sampling.

E.1 Same-weight native physical comparison

The six checkpoint tensors remain fixed. Four fresh development scenarios 1080000–1080003 execute all 62 seen masks at $c = .25$: mode and two sample streams per actor, plus one shared margin reference, giving 76 banks / 4,712 complete physical episodes. Every arm restarts its full 6.4 s episode from the same prescribed initial/reference/fault/servo conditions. Later states follow its actions; there is no q/v clone or forced common repair state. No refitting, temperature selection, probability-mean action, filter, or new checkpoint is introduced.

At each actual 20 Hz decision, native float32 `Categorical(logits=...).probs` is cast to float64 and renormalized. Its cumulative sum is clipped at one; the last positive-mass category and all later entries are set to one. The first CDF entry strictly above the fixed uniform selects the action, excluding zero-mass plateaus. Endpoint handling was fixed before data generation. This is not a replay of training-time random draws or `torch.multinomial`. Each scenario/stream has a fixed

Table 11: Posthoc paired outcomes versus the original constant-$\alpha = 1$ reference, all 992 quarter-tolerance contexts. A rescue removes an original full constraint failure; a new failure loses an originally admitted context. Recovery gains and losses are separate binary outcomes. Native histories and repair states differ. No new criterion or uncertainty interval is introduced.

Neutral fit	Rescues	New failures	Recovery gains	Recovery losses
Fixed (1)	5	0	1	14
Fixed (2)	4	0	0	33
Fixed (3)	4	1	1	1
Dual (1)	9	0	0	23
Dual (2)	9	1	0	21
Dual (3)	3	0	0	4

Table 12: Same-input decoder diagnostic at quarter tolerance during the causally reported-fault phase. Each cell is D/W, computed by time averaging within each mask episode and then equally weighting all 62 masks. The prespecified panel is the three fixed-penalty fits; the dual fits are retained descriptively, not substituted for that panel. Each fixed fit passes all four scenario thresholds $D \geq .50$, $W \geq .10$. These are distribution summaries, not physical performance or independent decision samples.

Variant / seed	1040000	1040001	1040002	1040003
Fixed / 9201	0.8486/0.2137	0.8730/0.2520	0.8607/0.2297	0.8744/0.2596
Fixed / 9202	0.9023/0.4382	0.8810/0.3695	0.8929/0.3910	0.8646/0.3314
Fixed / 9203	0.8264/0.2234	0.8064/0.2161	0.8206/0.2246	0.8214/0.2341
Dual / 9201	0.8578/0.2304	0.8791/0.2593	0.8686/0.2433	0.8758/0.2753
Dual / 9202	0.8370/0.2543	0.8220/0.2350	0.8300/0.2416	0.8300/0.2560
Dual / 9203	0.8343/0.1992	0.8568/0.2409	0.8443/0.2158	0.8564/0.2536

128×62 uniform array indexed by decision and mask, shared across actors without redraws. Alpha holds five intervals while the allocator updates at 100 Hz.

The primary panel requires at least two of three fixed-penalty fits to reduce AUC by 10% overall and in at least three of four scenarios, with no more full failures and no recovery-count loss in either stream. Any abort or unknown tail fails admission. Dual fits cannot rescue the primary result. A separate margin comparison requires each stream to have zero full failures, at least margin's recovery, and mean AUC at most 90% of margin; margin must itself have zero failures for the matched-observed-safety claim. Both criteria fail for all fixed fits.

Sampling-mean AUC improves for F-9201/D-9201 and worsens for four fits. Eleven sampled arms introduce crossings: 46 crossing episodes across twelve arms versus zero in six mode arms. F-9203/S0 is the zero-failure exception, while S1 has four. Recovery gains are retained, not allowed to offset crossing counts. The margin has zero crossings, 248 recoveries, and AUC .241366790, lower than every sampled arm. The conditional confirmation is not activated. All actual 2 ms states, commands, observations and decision inputs are independently reconstructed; maximum actual speed is 1.2094 rad/s, not bounded by the command cap.

The four scenarios, not masks or the two streams, are the independent scene units. No narrow mask-level interval is used. Complete paired new/rescued failures and gained/lost recoveries remain in `revision/round09/design/v16_complete_admission.json`, together with all 19 methods/windows and scene-specific contingencies. This native same-weight intervention identifies total deployment-policy effects, not usefulness of history, superiority to untrained randomization, population safety, or impossibility of other stochastic policies.

Table 13: All 19 native methods, 248 complete contexts each. V counts primary crossings in the full, fault, and post-repair two-second windows; all chart/joint failures and aborts are zero. Recovery uses the unchanged dwell criterion. Full AUC includes every context, including crossing cases. S0 and S1 are retained action streams, not additional scenario clusters.

Method	V_{full}	V_{fault}	V_{post}	R	J_{full}	J_{fault}	J_{post}
F-9201 mode	0	0	0	242	1.359493	1.166298	0.177578
F-9201 S0	2	1	0	248	1.249915	1.162497	0.070810
F-9201 S1	5	3	1	248	1.169432	1.094463	0.056952
F-9202 mode	0	0	0	239	0.730212	0.655501	0.067292
F-9202 S0	5	2	0	248	1.345584	1.243178	0.079027
F-9202 S1	10	5	0	247	1.268719	1.174842	0.068187
F-9203 mode	0	0	0	248	0.877177	0.824706	0.045417
F-9203 S0	0	0	0	248	1.108604	1.039682	0.056151
F-9203 S1	4	1	1	248	1.045152	0.986989	0.044139
D-9201 mode	0	0	0	242	1.542697	1.297898	0.221977
D-9201 S0	2	1	0	248	1.294188	1.204974	0.071217
D-9201 S1	6	5	0	248	1.212976	1.132034	0.061412
D-9202 mode	0	0	0	242	1.162197	1.039705	0.099825
D-9202 S0	2	1	0	248	1.199275	1.112732	0.068086
D-9202 S1	7	3	0	247	1.159908	1.078918	0.060773
D-9203 mode	0	0	0	246	0.836921	0.741473	0.073930
D-9203 S0	1	1	0	248	1.138916	1.060697	0.067360
D-9203 S1	2	1	0	248	1.045611	0.984389	0.049631
Margin	0	0	0	248	0.241367	0.239723	0.001251

Table 14: All scenario-level full-AUC differences, sampling mean minus comparator. Two streams and all 62 masks are averaged within each scenario. Negative values favor sampling. These four development clusters do not support a narrow mask-level confidence interval; all four differences are shown.

Fit	Comparator	1080000	1080001	1080002	1080003
F-9201	mode	+0.146973	-0.308376	-0.254281	-0.183592
F-9201	margin	+1.656808	+0.777624	+1.213001	+0.225796
F-9202	mode	+1.120956	+0.544865	+0.559815	+0.082122
F-9202	margin	+1.835760	+0.913800	+1.241606	+0.271974
F-9203	mode	+0.308908	+0.240242	+0.182107	+0.067547
F-9203	margin	+1.478798	+0.615909	+1.076989	+0.170348
D-9201	mode	-0.151924	-0.350741	-0.386541	-0.267251
D-9201	margin	+1.716035	+0.826700	+1.268000	+0.238127
D-9202	mode	+0.029567	+0.031290	+0.166353	-0.157635
D-9202	margin	+1.698553	+0.754252	+1.072907	+0.227185
D-9203	mode	+0.198276	+0.235635	+0.504824	+0.082636
D-9203	margin	+1.519463	+0.583082	+1.142181	+0.158860

F IDEAL RETIMING AND FINITE-OPTION EVIDENCE

Let static smooth task coordinates have constant full-row-rank J_p and $B = J_s N_p$, with exact undamped $N_p = I - J_p^\dagger J_p$ and positive diagonal gains. Suppose primary error starts at zero, effectiveness is restored, following is ideal, and there is no disturbance or saturation. All trajectories must stay in the same chart, rank region and joint domain. For $u^p = J_p^\dagger K_p e_p$ and $u^s = N_p B^\dagger (K_s e_s - J_s u^p)$, exact priority gives $\dot{e}_p = -K_p e_p$, hence $u^p = 0$. With a locally unique autonomous secondary field f, measurable $\alpha \in [0, 1]$ yields

$$q_\alpha(t) = \bar{q}(s(t)), \quad s(t) = \int_0^t \alpha(\xi) d\xi \le t, \qquad e_{s,i}(t) = e_{s,i}(0) e^{-k_i s(t)}. \tag{13}$$

The chain rule and local uniqueness establish the first identity; full row rank gives $\dot{e}_s = -\alpha K_s e_s$ and the second. Alpha one weakly dominates componentwise error and its integral, and cannot finish a qualifying dwell later. This elementary conditional property is not a new theorem or the simulated plant's identified model. Nonzero primary error instead gives $\dot{e}_s = -(1-\alpha) J_s u^p - \alpha K_s e_s$; damping, moving references, inertia and saturation can also break the assumptions. For a hypothetical $\dot{q} = v$, $\dot{v} = A(u^p + \alpha u^s - v)$,

$$\ddot{e}_p = -\dot{J}_p v - J_p A(u^p - v) - \alpha J_p A u^s. \tag{14}$$

Geometric nullspace cancellation need not cancel the dynamic term; isotropic response still leaves curvature. A model $\dot{v} = Gu - Av + d$ would instead use $J_p G u^s$. Neither is claimed as the actual drive model.

F.1 MATCHED RECOVERY OPTIONS AND CONTEXTUAL REGRESSION

Each repair-time choice uses an identical replayed healthy-one/fault-zero prefix and four fixed continuations: $0, .5, 1$, or $\alpha_k = \min(1, (k+1)/50)$ at 100 Hz. No learned replanning follows. On four development scenarios, quarter immediate restoration crosses in 31/248 contexts versus ramp zero; both recover all contexts, with AUC .15929/.42565 and restricted recovery time .28351/.52629 s. All repair boundaries are initially admitted. Different transient options are not pooled into this finite action library. A hindsight minimum over these four physical continuations is not a global oracle or a lower bound for later-feedback policies.

Full-information fitting uses every option's actual two-second base-normalized primary peak and AUC. Training/validation/test use 16/4/8 scenarios (IDs 960000–960015, 960100–960103, 960200–960207): 992/248/496 repair contexts and four physical continuations each. The 121 base-tolerance inputs are training-standardized and augmented with an option indicator; requested tolerance is not a duplicated training sample. MLP seeds 9701/2/3 fit two log outcomes with AdamW 3×10^{-4}, weight decay 10^{-4}, batches 256, up to 200 epochs. Validation MSE selects among 50/100/200; all select 200, with no best seed. Option-specific ridge, same-mask neighbors, fixed actions, training-only mask lookup and the declared kinematic rule are retained baselines. Predicted peak plus empirical margin admits candidates; minimum predicted AUC selects, with minimum-peak empty fallback. This is contextual outcome regression, not PPO, Q^*, or a safety certificate.

F.2 PREDICTION AND SELECTION FAILURES

On the first eight-scenario test, the three MLPs recover 383/387/393 of 496, with 1/1/2 quarter post crossings and AUC 1.6551/1.5746/1.4878. Ramp has zero crossings, 496 recoveries and AUC .62584; ridge has zero, 434, and 1.3237. Holding counts 113/109/103 equal the MLP recovery deficits, but only 7/11/11 arise from empty predicted sets. In every holding case, actual ramp is admitted and recovers. Prediction MSE alone therefore does not establish decision quality.

A second eight-scenario cohort retains weights and tests holding redirection, no-hold reranking and a ramp-default rule. Redirection alone removes all 87/91/106 recovery deficits without new quarter crossings. The ensemble improves AUC over ramp by .06756 (paired eight-scenario 95% interval [.04628,.08940]), with zero post crossings and 496 recoveries each, yet stronger simple controls remain:

Table 15: Prespecified paired physical-budget contrasts: $\beta = 4$ minus $\beta = 1$. Eight scenario clusters, each retaining all 62 masks; 10,000 paired bootstrap draws. Units are base-normalized primary error, not radians or metres.

Endpoint	Mean difference	95% interval
Immediate peak	-0.03125	$[-0.03693, -0.02589]$
Immediate minus ramp peak gap	-0.03126	$[-0.03695, -0.02589]$

Table 16: Finite hidden-budget ambiguity; 496 contexts at each scale. $\mathcal{G}_b$ requires initial/post primary, chart, and joint admission plus recovery. Undefined minimax cases are excluded with an explicit denominator, not assigned finite regret.

c	Different $\mathcal{F}_b$	Disjoint near-optima	Common $\mathcal{G}_b$	Mean regret
1	2	2/496	496	0.00074
0.5	12	12/496	496	0.00750
0.25	43	41/496	496	0.01983
0.1	41	38/491	491	0.01198

Policy	Two-second AUC	Recovery time (s)
Ramp	.41763	.51560
Half	.30930	.56044
Training mask lookup	.22947	.41167
Anchored ensemble	.35008	.48615

Anchoring is slightly worse than redirection due to exact floored-zero cost ties. At scale .1, ensemble and ramp each have eight post crossings, and seven contexts lack a safe-recovering option. A default can execute ramp even when prediction excludes it; false-safe admission counts do not certify execution. All 30 methods/four scales and paired contrasts remain in the archived complete tables. Neither cohort demonstrates RL or richer-state superiority; neither is a same-interface PPO baseline.

G Matched Physical-Budget Intervention

Eight scenarios 990200–990207, all 62 masks, four fixed options and $\beta \in \{1, 4\}$ produce 64 banks / 3,968 physical episodes. Complete healthy-one/fault-zero prefixes and repair observations match exactly. Only from reported repair does total configured force budget become βF_{nom}, with per-substep drive limit $\max(\beta F_{\mathrm{nom}} - |g(q)|, 0)$ and unchanged gravity compensation. Gains, reference, allocator and inputs remain fixed; beta is hidden from the repair observation. This tests configured headroom, not measured torque or a uniquely identified saturation mechanism.

Both prespecified physical contrasts have negative paired eight-scenario 95% bootstrap intervals (10,000 draws, seed 990299): Immediate quarter post crossings change from 81 to 42, comprising 40 rescues, one new crossing and 41 common crossings. The new crossing is scenario 990202 mask 4. Ramp has zero post crossings and 496 recoveries in both budgets, but one shared prefix crossing means one full-episode failure. Increasing budget therefore improves the mean, not every context. Actual speed reaches 1.06627 rad/s despite the 1 rad/s command cap; chart and joint-limit counts are zero.

For common repair context o, define $\mathcal{F}_b(o; c)$ as the four options admitted at repair and throughout the next 2 s in primary, chart and joint checks; $\mathcal{G}_b$ additionally requires recovery. Earlier prefix events remain in full-episode outcomes but not these sets. When $\mathcal{G}_b$ is nonempty, let $J_b^* = \min_{a \in \mathcal{G}_b} J_b(a)$ and $\mathcal{N}_b = \{a \in \mathcal{G}_b : J_b(a) - J_b^* \le .01\}$. The finite deterministic common-admissible regret is

$$D(o; c) = \min_{a \in \mathcal{G}_1 \cap \mathcal{G}_4} \max_{b \in \{1,4\}} [J_b(a) - J_b^*]. \tag{15}$$

Absent optima or intersections are undefined/infinite, never assigned finite success. This is not a bound for arbitrary schedules, randomized policies or feedback observing later responses.

At quarter tolerance all 496 contexts retain a common post-repair-admitted recovering option, while 41 have disjoint near-optima, distributed 3/7/6/8/4/6/5/2 across scenes. Mean regret is .01983, max-

imum 1.28858. At scale .1, the shared finite denominator is 491. A single repair-time observation cannot select different options for the two deliberately hidden settings, but this does not show a common safe-recovering option is absent or explain loss on a fixed nominal plant. The nominal-only PPO experiment does not test resolution of this constructed ambiguity.

The first geometry job stopped on an intermediate Euler singularity before physics. A disclosed amendment rejects individual singular rows, retains rejected attempts and the original eight scenarios/direction order, and never qualifies an inactive placeholder. Configured-budget and all-policy tables, including unfavorable transfer outcomes, remain in the archived full appendix and generated sources; no evidence is deleted or reclassified as a learned success.

H Stopped Development Hypotheses

Cost-estimator screen. Original-policy gradients from six existing runs at updates 1/24/42 compare GAE with same-baseline Monte Carlo cost credit. The declared joint materiality screen fails, so no such training intervention runs. Advantage variability does not establish a harmful gradient direction; zero realized future cost does not imply zero expected risk or baselined advantage. Only update 1 has a four-epoch optimizer replay. Paired variants share that initial rollout, not independent evidence; phase bins of healthy rollouts refer to recorded schedule labels, not actual fault or repair. This does not identify the remaining control loss.

H.1 Fixed residual family

Two reused development scenarios 1040000/1040001, all masks and three scales execute five arms: 30 banks / 1,860 episodes. The selected margin rule remains $(.6, .9)$, upward rate 2/s and .1 s lookahead. Four arms add $\pm.25$ for five 20 Hz decisions, beginning .30 s after reported fault or .10 s after reported repair; the fifth is unchanged. Alpha is clipped and rounded down on the .05 grid, and margin is recomputed in each arm's actual state. Entire episodes replay the prefix; there is no q/v clone.

A useful context must improve full AUC by at least $\max(.001, .01J_{\text{base}})$ while preserving full admission/recovery, or gain recovery/remove a failure without another criterion worsening. The quarter launch rule requires four useful contexts over two masks in each scenario. Neither scene supplies one. All five arms admit and recover all 124 contexts. Negative fault offsets increase mean AUC by .04355018; positive fault offsets are no-ops in 122 cases and harmful in two. Positive repair offsets alter four cases but their mean decrease .00002303 and all individual decreases miss the threshold. At scale .1, positive offsets supply some useful cases only in one scene and add two fault-arm crossings. No residual learner is trained; another family or RL in general is not excluded.

H.2 Fixed probe and hidden-budget options

Four fresh development scenarios replay healthy-one/fault-zero, then at repair change budget to beta one or four and hold $\alpha = .10$ for five 10 ms intervals. Four continuations follow: $0, .5, 1$, or a 20 Hz ramp starting .2 and adding .1 per decision. This is a constant probe, not the earlier ramp $.02, .04, .06, .08, .10$. All 32 banks / 1,984 episodes complete. The original full and repaired two-second clocks include the probe. All 248 contexts have a common admitted-recovering continuation. Efficiency contexts require disjoint .01-near-optimal sets across budgets; scene counts are 1/6/3/3, so only one meets the four-context threshold where two are required. This family stops without an observer or learner; it proves neither history value nor global optimality.

I Supplementary Materials and Execution Details

The anonymous supplementary ZIP contains `latex_source/`, `evidence/tables/`, `evidence/numeric/`, `data/`, and one `video/iclr_experiments.mp4`. The source archive compiles this paper independently. Retained tables include all original/neutral windows and contrasts, all 30 neutral binary comparisons, full option/budget policy tables, and native-decoder paired changes. In particular, `neutral_test_contrasts.tex`, `feedback_all_windows_tables.tex`, `decoder_physics_pairs_table.tex`, and `boundary_all_`

Table 17: Complete fixed-probe development outcomes: all 248 planned episodes per row complete. Each row crosses four scenarios with all 62 masks. P is the full primary-crossing count; chart/joint events and aborts are zero. Full AUC includes the original prefix and probe; recovery uses the unchanged original 2 s post-repair window, including the probe. There is no fitted selector.

β	Continuation	P	Recovered	Full AUC	Post-2s AUC
1	0	0	0	8.837143	4.074321
1	$1/2$	0	248	4.183031	0.630410
1	1	34	248	3.920751	0.368130
1	Ramp from probe	0	248	4.192493	0.639872
4	0	0	0	8.837143	4.074321
4	$1/2$	0	248	4.182629	0.630008
4	1	21	248	3.917542	0.364922
4	Ramp from probe	0	248	4.192474	0.639853

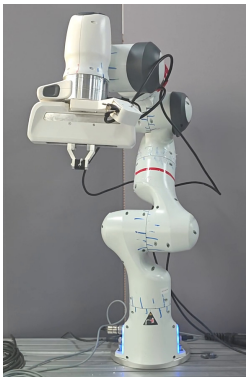
`policy_tables.tex` retain detail omitted from this compact PDF. Numerical summaries and experiment protocols accompany them. Packaging anonymizes filesystem locations in metadata without changing numerical results. The package provides compact numerical summaries and selected traces. Removing repetitive printed inventories changes no metric, model, comparator or scientific gate.

The recorded runtime is Python 3.10.20, Isaac Sim 4.5.0.0, Kit 106.5.0.162521, PyTorch 2.5.1+cpu, NumPy 2.2.6 and CPU PhysX. Exact arithmetic matters: an independent residual-rule replay under NumPy 1.21.5 disagreed at four quantization boundaries; actual-runtime replay agreed without data or formula changes. Original failed attempts remain preserved. Likewise a baseline-validation exit 250 occurred only after all 40 banks were saved; the narrow post-save-close admission rule was frozen before selection/test, with no retry or outcome filtering. A residual collector's serialization repair retained its first complete raw banks and executed only the remaining roster; it was disclosed after those first banks, not retroactively called wholly prospective.

Audit inventories verify implementation and scope, not scientific value or independent replications. Geometric witnesses, recorded substep admission and software export parity support their respective geometric, sampled-trajectory and software claims; they do not certify continuous-time safety.

J TWO ARCHIVED HARDWARE TRIALS AND TASK-INFORMED SIMULATION

The two records in Figure 2 are actual Panda hardware measurements of the same floor-circle task, with position (x, y, z) prioritized and a supplied fault schedule for J2, J1, and J5 at 10, 20, and 30 s.



Recorded hardware context. The video frame shows the physical Panda platform; it is not synchronized to the plotted trials. Each trial lasts approximately 40 s, sampled at approximately 200 Hz. The two panels retain all recorded samples, including position-error peaks of 37.54 and 37.45 mm. The curves describe two trials of one task; they are not independent trajectory families or a hardware ablation.

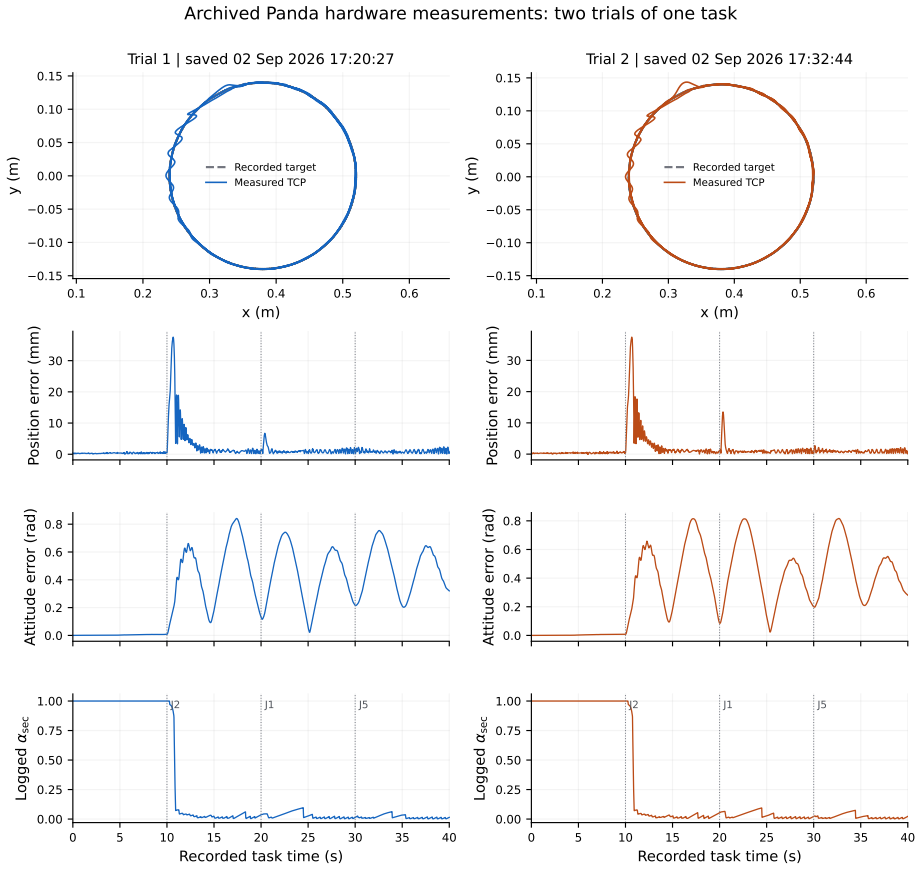


Figure 2: Two archived hardware trials. Column identifiers are the record save times, not synchronized camera times. The top row shows recorded target and measured TCP positions projected onto the xy plane. Subsequent rows show $\|p_d - p\|_2$, $\|\log(R_d R^\top)^\vee\|_2$, and the logged α_{sec}. Dashed vertical lines mark the supplied experiment schedule, not fault-event times inferred from the video.

Separate task-informed simulation. We additionally execute two explicitly simulated illustrations using the recorded Cartesian reference displacements and the same fault order. The paths are translated to the nominal FR3 initial pose $q_0 = (0, 0, 0, -1.5, 0, 1.5, 0)$, with its orientation held fixed. XYZ is primary, using nominal position tolerances of 10 mm per coordinate. The first selected

fixed-penalty fit (F-9201) is fixed before execution; no checkpoint or trajectory is chosen by its replay outcome. Both parallel rollouts use the original 20/100/500 Hz actor/controller/physics clocks for 39.99 s. Cumulative zero command effectiveness replaces a mechanical joint lock; the finite-force servo remains active. This is a shifted-task test with different geometry and dynamics, not a calibrated reconstruction of the recorded Panda controller.

Task source	Mean α	Max. primary ratio	Max. position error	Chart events
Trial 1	.48748	18.6834	195.32 mm	234/3999
Trial 2	.48705	18.7277	195.83 mm	232/3999

Table 18: Primary ratios use all 2 ms states. Chart-event counts are 10 ms intervals with any out-of-chart state; they are not independent episodes.

The simulations complete without a joint-limit event, but have 1,887/1,890 primary-crossing intervals and the chart events shown above. These retained failures preclude a safety or hardware-matching claim. Raw traces, the fixed pre-execution setup, and process exit code zero accompany the supplementary source archive.