WHEN ZERO COST DOES NOT MEAN SAFE: RESPONSE DELAY AND POLICY RETENTION IN RESOURCE-CONSTRAINED REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Safe reinforcement learning under a finite training budget can fail even when standard constraint mechanisms are present: resource degradation can raise future risk before positive cost is observed, response can be delayed, and feasible policies discovered during training can be lost before delivery. We analyze these stages by separating non-detection from post-detection non-recovery through multiplier first-passage analysis and by decomposing final delivery into candidate opportunity and selection loss. Guided by these diagnostics, we develop the ERC family for finite-budget safe RL. Its flagship realization, Enhanced Resource-Coupled Trust Region Policy Optimization with Shadow Generation (ERC-TRPO-SG), integrates resource-aware candidate generation, conditional shadow repair, and policy retention. Across three resource-coupled domains, ERC-TRPO-SG delivers mean-feasible policies for all 20 training seeds in each domain. Under a matched 400,000-interaction cap in thermal management, feasible delivery rises from 4/20 to 16/20 relative to ERC-TRPO. Factorial ablations attribute most of the final-delivery gain to retention. On nine Safety-Gymnasium tasks under matched 10M-interaction budgets, CSPO reaches initial recovery sooner, whereas the ERC-U backbone exhibits fewer late violations and mean-feasible frozen policies on 8/9 tasks versus 6/9. These results distinguish initial recovery, persistent safety, and final policy delivery as separate finite-budget objectives.

1 INTRODUCTION

Reinforcement learning (RL) has achieved strong results in sequential decision-making, yet many control problems couple actions to an observed resource that shapes future safety. We study network defense, solar-powered UAV control, and thermal management, where resource degradation can increase future constraint risk while observed cost remains zero. A finite training budget creates a second difficulty: learning must stop and a policy must be delivered, even when safer candidates appeared earlier. We therefore study safe RL as a temporal response and delivery problem in addition to a constrained optimization problem.

Constrained Markov decision processes (CMDPs) formalize expected-cost constraints (Altman, 1999). CMDP methods address constrained optimization through trust regions, primal–dual updates, primal switching, and response-aware multiplier dynamics (Achiam et al., 2017; Paternain et al., 2019; Stooke et al., 2020; Xu et al., 2021). Recent methods further target recovery speed, informative violations, and last-iterate behavior (Belouadah et al., 2026; Yang et al., 2026; Moskovitz et al., 2023). Together, these lines improve constraint information, corrective dynamics, and iterate behavior. We focus on the temporal chain connecting risk emergence, cost detection, corrective response, persistent feasibility, and final policy delivery within a finite training budget.

We analyze two finite-window failure mechanisms. Response delay arises when resource degradation precedes positive cost or when detected cost fails to build sufficient corrective pressure before the window ends. Delivery degradation arises when useful candidates are generated but later lost or not selected. These mechanisms are distinct from persistence: a policy can recover feasibility and subsequently violate the constraint again. The resulting picture is stage-wise: failure can arise before detection, between detection and response, after recovery, or at final delivery.

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.

Guided by these diagnoses, we develop the ERC family of finite-budget safe-RL methods. Its flagship realization, ERC-TRPO-SG, maps the temporal failure chain to three interventions: resource-aware candidate generation uses observed resource information before positive cost, conditional same-batch shadow repair supplies a corrective alternative when recovery is at risk of missing the remaining budget, and policy retention preserves useful candidates through final delivery. Paired resource interventions separately test whether the observed resource carries predictive information about future constraint trajectories beyond instantaneous cost.

Our contributions are threefold. (i) We identify finite-budget safe RL as a temporal response-and-delivery problem and formalize two corresponding objects: multiplier first passage, which separates non-detection from post-detection non-response, and candidate opportunity versus selection loss, which separates policy discovery from final delivery. (ii) We introduce the ERC family and its flagship ERC-TRPO-SG realization, and use controlled interventions and factorial ablations to attribute the roles of resource information, online repair, and policy retention; under a matched 400,000-interaction D3 budget, feasible delivery increases from 4/20 to 16/20, with retention accounting for most of the endpoint gain. (iii) We test the same temporal distinctions beyond the resource-coupled domains in matched 10M-interaction evaluations on nine Safety-Gymnasium tasks using ERC-U and CSPO, where initial recovery and late-stage persistence separate sharply. Together, these results distinguish initial recovery, persistent safety, and final policy delivery as separate finite-budget objectives.

2 Related Work

Constraint response. CMDP duality, trust-region, primal, and PID-based methods characterize constrained optimization and update dynamics (Altman, 1999; Paternain et al., 2019; Achiam et al., 2017; Xu et al., 2021; Stooke et al., 2020). SCPO, TRC, and CSPO address state-wise, tail-risk, and constraint-sensitive updates (Zhao et al., 2024; Kim & Oh, 2022; Belouadah et al., 2026). CSPO emphasizes rapid recovery after violations; our finite-window analysis additionally separates cost detection from subsequent response and persistence.

Data generation and recovery. FDPI uses a dual policy and importance weighting to increase violating samples (Yang et al., 2026). Consumption MDPs model replenishable resources explicitly (Blahoudek et al., 2020), while Sauté RL augments the state with a safety-budget variable (Sootla et al., 2022). Standard CMDPs can include observed resource variables. Our focus is their temporal role: an environment resource can alter future risk before positive cost is observed, creating a finite-window gap between risk emergence and constraint response. Recovery RL learns a separate recovery policy from offline safety data (Thananjeyan et al., 2021); ERC-TRPO-SG instead forms main and shadow candidates from the same pre-update policy and batch, then compares them through paired validation. The resulting generalization question concerns finite, reused validation panels rather than an offline recovery-zone critic.

Policy delivery. ReLOAD studies last-iterate convergence through optimistic primal–dual updates (Moskovitz et al., 2023). We study the finite candidate pool and its delivery loss, separating whether training generated a useful candidate from whether the delivery rule preserved and selected it. Terminal-policy quality, candidate opportunity, and delivery loss are therefore distinct evaluation quantities. Across these lines of work, our contribution is to connect information arrival, corrective response, persistence, and policy delivery within a single finite-budget temporal framework.

3 Problem Setup

We consider a constrained Markov decision process (CMDP) with nonnegative safety cost and an observed resource state whose evolution can be affected by action-dependent consumption or replenishment and can influence future transitions and costs. For a policy π, the theoretical objective is

$$J_R(\pi) = \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t r_t\right], \qquad J_C(\pi) = \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t c_t\right] \le d. \tag{1}$$

Here $\gamma \in (0,1)$ and $d > 0$; each comparison fixes the initial-state distribution and transition law. Table S1 in the appendix summarizes notation.

Definition 1 (Resource-coupled state)**.** Write the state as $s_t = (x_t, b_t)$, where b_t is an observed resource or safety margin affecting future dynamics or costs, and x_t contains the remaining Markov state. We use "resource-coupled" for standard CMDPs with such an observed resource component.

Its temporal significance is that two states with the same instantaneous cost may nevertheless induce different future-cost distributions through different resource states. The theoretical functionals in Eq. (1), the batch statistic $\widehat{C}_k$, and protocol-specific experimental endpoints are distinct objects; Section 6 specifies the empirical functionals and budgets.

Let $\mathcal{F}_k$ contain the information available before collecting batch k, including observed resources and previous validation outcomes. The nonnegative batch-cost estimate $\widehat{C}_k$ becomes available in $\mathcal{F}_{k+1}$. Cost detection is its first positive value; multiplier recovery is the first crossing of a declared pressure threshold. Policy feasibility and final policy delivery are separate endpoints. Resource information may therefore precede positive observed cost.

4 Finite-Window Response and Delivery

We formalize two finite-window diagnostics corresponding to distinct stages of the temporal chain: response timing through first passage of a persistent projected multiplier, and policy delivery through loss over a finite candidate pool. The first distinguishes failure before and after positive-cost detection; the second distinguishes failure to discover a useful candidate from failure to preserve or select one. Section 5 maps these diagnostics to the corresponding ERC interventions.

4.1 Detection and Multiplier First Passage

Consider a persistent projected multiplier with fixed step size, cost budget, and admissible cap:

$$\nu_{k+1} = \Pi_{[0,\nu_{\max}]}\Big(\nu_k + \alpha_\nu(\widehat{C}_k - d)\Big), \qquad 0 \le \nu_0 < \nu^\star < \nu_{\max}. \tag{2}$$

Here $\alpha_\nu > 0$ is the multiplier step size, $\Pi_{[0,\nu_{\max}]}$ denotes projection onto the admissible interval, and $\nu^\star$ is a prespecified response threshold distinct from the cost budget d. The multiplier is persistent: it is not reset between batches.

Definition 2 (Arrival probability and conditional magnitude)**.** Let

$$Z_k = \mathbf{1}\{\widehat{C}_k > 0\}, \qquad p_k = \Pr(Z_k = 1 \mid \mathcal{F}_k),$$

and, whenever $p_k > 0$,

$$\mu_k = \mathbb{E}[\widehat{C}_k \mid Z_k = 1, \mathcal{F}_k].$$

When $p_k = 0$, we adopt the convention $p_k\mu_k = 0$; μ_k itself need not be defined on that null event.

The product of arrival probability and conditional positive-cost magnitude determines the conditional cost-excess drift:

$$m_k = \mathbb{E}[\widehat{C}_k - d \mid \mathcal{F}_k] = p_k\mu_k - d, \qquad \xi_k = \widehat{C}_k - d - m_k, \quad \mathbb{E}[\xi_k \mid \mathcal{F}_k] = 0. \tag{3}$$

This drift need not equal the excess of the evaluation cost over its budget when the batch estimator targets a different quantity.

A sequence of zero-cost batches erodes accumulated multiplier pressure exactly. If $\widehat{C}_j = 0$ for $j = k, \dots, k+\ell-1$, then

$$\nu_{k+\ell} = [\nu_k - \ell\alpha_\nu d]_+, \qquad [x]_+ = \max\{x, 0\}. \tag{4}$$

Thus, for a fixed initial multiplier and recursion, identical batch-cost inputs induce identical multiplier paths. The resource state affects this baseline recursion only through the cost stream; it does not enter Eq. (2) as an additional term.

For a window of K learner updates, define

$$\tau_\nu = \inf\{k \ge 0 : \nu_k \ge \nu^\star\}, \quad D_K = \{Z_0 = \cdots = Z_{K-1} = 0\}, \quad R_K = D_K^c \cap \{\tau_\nu > K\}. \tag{5}$$

We use $\inf \varnothing = +\infty$. Since $\nu_0 < \nu^\star$ and a zero-cost update cannot increase the multiplier, $D_K \subseteq \{\tau_\nu > K\}$. Hence finite-window response failure decomposes disjointly as

$$\{\tau_\nu > K\} = D_K \,\dot\cup\, R_K. \tag{6}$$

The event D_K captures non-detection, whereas R_K captures detection without recovery by the end of the window. Detection therefore depends on cost arrival, while recovery additionally depends on conditional magnitude and accumulated drift.

Let $g = (\nu^\star - \nu_0)/\alpha_\nu$. The bounded-critic/score condition used for the auxiliary zero-cost gradient analysis is stated as Assumption 1 in Appendix A.2. The first-passage result below uses only Assumptions 2–3.

Assumption 2 (Positive drift before first passage)**.** There exists a deterministic constant $\underline{m} > 0$ such that $m_k \geq \underline{m}$ almost surely on $\{k < \tau_\nu\}$.

For example, $p_k \geq p_{\min}$, $\mu_k \geq \mu_{\min}$, and $p_{\min}\mu_{\min} > d$ imply this assumption with $\underline{m} = p_{\min}\mu_{\min} - d$. Positive conditional drift still permits negative centered increments and nontrivial stochastic delay.

Assumption 3 (Conditional lower-tail bound)**.** There exist deterministic constants $\sigma_- > 0$ and $\theta_- > 0$ such that, almost surely on $\{k < \tau_\nu\}$,

$$\mathbb{E}\big[e^{-\theta \xi_k} \mid \mathcal{F}_k\big] \leq \exp\left(\frac{\sigma_-^2 \theta^2}{2}\right), \qquad 0 \leq \theta \leq \theta_-.$$

The parameter σ_- is a lower-tail scale and is unrelated to either the multiplier ν_k or the discount factor γ.

Proposition 1 (Positive-drift recovery upper bound)**.** *Under Assumptions 2–3, define*

$$\theta^\star = \min\left\{\frac{\theta_-}{2}, \frac{\underline{m}}{2\sigma_-^2}\right\}, \qquad c_- = \frac{\theta^\star \underline{m}}{4}.$$

Then

$$\mathbb{E}[\tau_\nu \mid \mathcal{F}_0] \leq \left\lceil \frac{2g}{\underline{m}} \right\rceil + \frac{1}{1 - e^{-c_-}}. \tag{7}$$

The leading term scales with $g/\underline{m}$: larger threshold gaps delay first passage and stronger positive drift accelerates it. Appendix A.3 gives the proof and tail bounds, and Appendix A.4 covers zero and negative drift. Relating multiplier first passage to policy feasibility additionally requires actor-side conditions; Appendix F gives the corresponding pathwise bridge from multiplier response to policy behavior.

4.2 Resource Information and Candidate Opportunity

Proposition A.2 formalizes the predictive role of resource information: two directly observed resource states can share the same instantaneous cost yet induce different future-cost laws under common subsequent actions and exogenous dynamics. The proposition therefore concerns future-risk information beyond instantaneous cost; Appendix A.6 states the corresponding full-history formulation.

We next turn from response to delivery. Fix a finite candidate pool $\mathcal{C}$ containing the terminal policy π_T, and fix a prespecified scalar endpoint J for which larger values are preferable. We apply the analysis separately to return and negative cost, using a fixed evaluation convention. Thus $J = J_R$ or $J = -J_C$ under Eq. (1), or the corresponding finite-episode evaluation functional. The return-maximizing and cost-minimizing candidates need not coincide.

For an output policy $\pi_{\text{out}} \in \mathcal{C}$, define the candidate opportunity and selection loss by

$$O_J = \max_{\pi \in \mathcal{C}} J(\pi) - J(\pi_T), \qquad L_J = \max_{\pi \in \mathcal{C}} J(\pi) - J(\pi_{\text{out}}). \tag{8}$$

Proposition 2 (Candidate opportunity and delivery)**.** *For any $\pi_{\text{out}} \in \mathcal{C}$,*

$$O_J \geq 0, \qquad L_J \geq 0,$$

and

$$G_J := J(\pi_{\text{out}}) - J(\pi_T) = O_J - L_J. \tag{9}$$

Suppose that the delivery rule selects π_{out} from a nonempty allowed set $\mathcal{A} \subseteq \mathcal{C}$ using an estimated endpoint $\widehat{J}$. Define

$$D_J = \max_{\pi \in \mathcal{C}} J(\pi) - \max_{\pi \in \mathcal{A}} J(\pi), \qquad \delta_J = \max_{\pi \in \mathcal{A}} \widehat{J}(\pi) - \widehat{J}(\pi_{\text{out}}).$$

On any event such that

$$\max_{\pi \in \mathcal{C}} |\widehat{J}(\pi) - J(\pi)| \le \epsilon_J,$$

we have

$$L_J \le D_J + 2\epsilon_J + \delta_J, \qquad G_J \ge O_J - D_J - 2\epsilon_J - \delta_J. \tag{10}$$

The bound separates rule exclusion D_J, estimation error $2\epsilon_J$, and within-set optimization gap δ_J; the factor two reflects estimation error at both the best allowed and delivered candidates. This turns an unsafe final policy into two measurable questions: whether a better candidate ever existed and, if so, how much of that opportunity was lost at delivery. Appendix H.9 specializes the bound to a candidate pool fixed before independent validation, and the empirical analysis evaluates the adaptive validation and selection procedures of ERC-TRPO-SG.

5 Resource-Aware Training and Policy Retention

The ERC family translates the finite-budget diagnoses of Section 4 into stage-specific interventions. ERC-TRPO-SG is the flagship full realization: resource-aware generation acts before positive cost is available, conditional same-batch shadow repair provides an online corrective candidate when projected recovery exceeds the remaining budget, and policy retention protects candidate opportunity through final delivery (Figure 1). Section 4.1 uses multiplier first passage to formalize the generic timing problem; ERC-TRPO-SG operationalizes that temporal question through validation cost excess and recovery-time extrapolation, while Appendix J evaluates a persistent-multiplier ERC realization.

5.1 Candidate Generation and Conditional Shadow Repair

ERC-TRPO uses TRPO (Schulman et al., 2015) with domain-specific training rewards while retaining the observed resource coordinate in the policy state. D1 uses

$$r_t^{\text{train}} = r_t^{\text{raw}} - 20c_t^2.$$

D2 and D3 additionally use a resource-potential difference together with a squared-cost penalty:

$$r_t^{\text{train}} = r_t^{\text{raw}} + w_\Phi\big[\gamma\Phi(s_{t+1}) - \Phi(s_t)\big] - w_f c_t^2, \qquad w_\Phi = w_f = 1. \tag{11}$$

With terminal potential set to zero, the potential term preserves policy ordering for a fixed initial-state distribution (Ng et al., 1999); raw evaluation rewards and costs are unchanged. Potentials and reward details are given in Appendix B.

Every eight updates and at the endpoint, the policy is evaluated on sixteen fixed validation scenarios. After three validations, let $(n_i, \bar{C}_i)$, $i = 1, 2, 3$, denote actor-step counts and mean costs. Define

$$s = \text{median}_{1 \le i < j \le 3} \frac{\bar{C}_j - \bar{C}_i}{n_j - n_i}, \qquad e = [\bar{C}_3 - d]_+, \tag{12}$$

and

$$T_{\text{rec}} = \begin{cases} 0, & e = 0, \\ e/(-s), & e > 0,\ s < -10^{-12}, \\ +\infty, & e > 0,\ s \ge -10^{-12}. \end{cases} \tag{13}$$

If two consecutive over-budget validations predict T_{rec} beyond the remaining actor-step budget, SG schedules repair at the next update. Detailed eligibility conditions are in Appendix B.

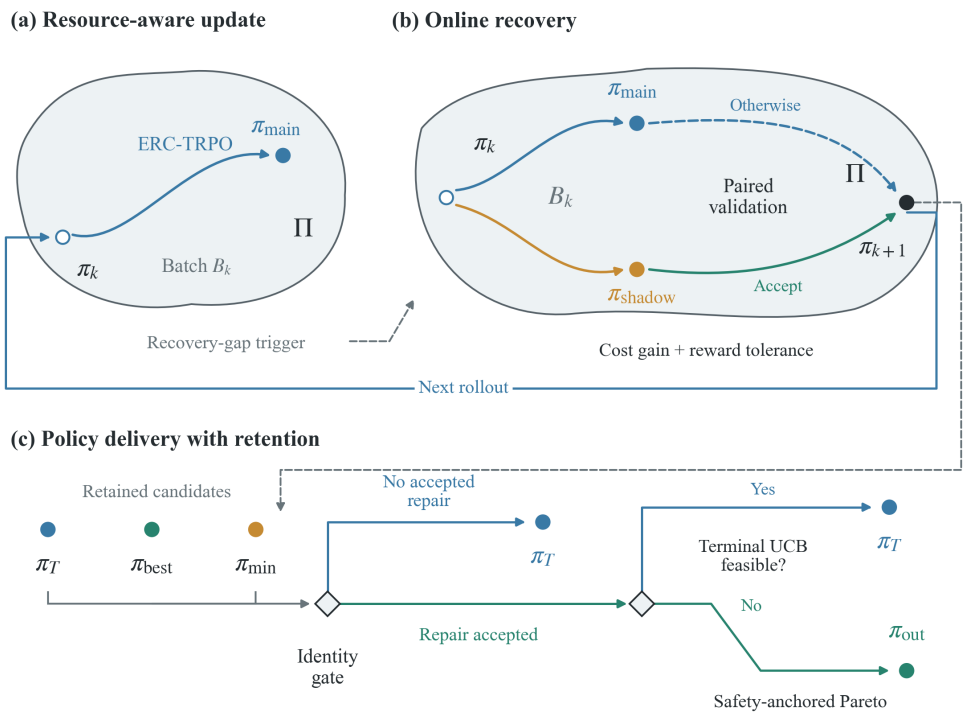

Figure 1: ERC-TRPO-SG maps the finite-budget failure chain to resource-aware candidate generation, conditional same-batch repair, and retained-policy delivery. Paired validation determines the next sampling policy, while the retained pool determines final delivery.

During recovery, both candidates use the same pre-update policy and batch. The main branch performs the ERC update; the shadow uses cost-descent TRPO in D1 and local CPO in D2/D3. Paired validation compares cost and raw return. For scenario i, define

$$\Delta C_i = C_i^{\text{shadow}} - C_i^{\text{main}}, \qquad \Delta R_i = R_i^{\text{shadow}} - R_i^{\text{main}}.$$

For sixteen paired observations X_i, define the nominal one-sided Student-t bounds

$$U_{0.95}(X) = \bar{X} + t_{0.95,15}\frac{s_X}{\sqrt{16}}, \qquad L_{0.95}(X) = \bar{X} - t_{0.95,15}\frac{s_X}{\sqrt{16}},$$

where s_X is the sample standard deviation and $t_{0.95,15}$ is the Student-t quantile. A shadow candidate is accepted only if

$$U_{0.95}(\Delta C) \le -\max\{10^{-8},\ 0.05[\bar{C}_{\text{main}} - d]_+\} \tag{14}$$

and

$$L_{0.95}(\Delta R) \ge -0.05\max\{|\bar{R}_{\text{main}}|,\ 1\}. \tag{15}$$

An accepted shadow becomes the next sampling policy. Full margins, deadbands, recovery termination, and validation reuse are specified in Appendix B. Because the validation panel is reused across decisions, Eqs. (14)–(15) are used for candidate comparison, while final endpoint feasibility is evaluated on scenarios disjoint from validation and selection.

5.2 Retention and Delivery

SG retains the terminal policy π_T, the highest-return cost-UCB-feasible candidate, and the minimum-cost-UCB candidate, omitting missing representatives and duplicate parameter vectors.

Let $\widehat{R}(\pi)$ be validation mean return and $U_C(\pi) = U_{0.95}(C(\pi))$ the nominal cost score. SG returns terminal if no shadow was accepted or $U_C(\pi_T) \le d$. Otherwise, let $\mathcal{P} \subseteq \mathcal{C}$ be the nondominated front for maximizing $\widehat{R}$ and minimizing U_C: domination means another candidate is at least as good

on both scores and strictly better on one. Define $u_{\min} = \min_{\pi \in \mathcal{P}} U_C(\pi)$. The delivered policy is chosen as

$$\pi_{\text{out}} \in \underset{\substack{\pi \in \mathcal{P} \\ U_C(\pi) \le u_{\min} + d}}{\arg\max} \widehat{R}(\pi). \tag{16}$$

The threshold $u_{\min} + d$ defines a relative risk band around the lowest retained cost score. Candidate selection and endpoint assessment remain separate stages: Eq. (16) determines delivery, while final feasibility is reported under the independent evaluation protocol. Appendix B gives numerical details, and Appendix H compares alternative delivery rules.

6 Experiments

The experiments address three questions aligned with the temporal framework. **Q1:** Does the flagship ERC-TRPO-SG improve finite-budget policy delivery in resource-coupled domains? **Q2:** Which gains are attributable to resource-aware generation, shadow repair, and retention? **Q3:** Are response timing, predictive resource information, persistence, and final delivery empirically separable? Sections 6.2–6.3 answer Q1–Q2 with ERC-TRPO-SG and controlled D3 studies; Section 6.4 answers Q3 with controlled response/resource experiments and the matched 10M Safety-Gymnasium evaluation of ERC-U.

6.1 Experimental Setup

We evaluate D1 network defense, D2 solar-powered UAV control, and D3 thermal management with protocol-specific budgets 0.005, 0.0005, and 1. D1 reports undiscounted cost per step; D2 and D3 use discounted episode cost with $\gamma = 0.99$. D3 combines simulated thermal dynamics with public workload and weather traces (Microsoft Azure, 2024; Stojkovic et al., 2025; Yang et al., 2025; Merket & Adhikari, 2020). Figure 2 compares ERC-TRPO-SG with eight learning baselines; complete identities and endpoints are in Appendices B–C.

The primary panel uses 20 seeds, 400,000 requested actor steps, and 50 evaluation scenarios per seed; the matched-total D3 study instead charges training and validation to a common 400,000-interaction cap. Separate D3 cohorts are used for the primary, matched-total, and factorial studies. Seed-bootstrap intervals and designated Holm-corrected paired tests are defined in Appendix C.

6.2 Finite-Budget Policy Delivery

Primary three-domain evaluation. ERC-TRPO-SG delivers mean-feasible policies for all 20 seeds in all three domains (Figure 2). In D3, mean cost decreases from 7.544 to 0.271, mean return increases from 186.996 to 188.745, and feasible delivery rises from 3/20 for ERC-TRPO to 20/20 for ERC-TRPO-SG. Both paired contrasts survive Holm correction. In D3, CSPO and ERC-TRPO-SG occupy the same low-cost, high-return regime. The paired ERC-TRPO study therefore isolates the delivery change introduced by the ERC-TRPO-SG intervention stack.

Matched total interactions. Under the common 400,000-interaction cap, ERC-TRPO-SG reduces mean cost from 5.118 to 0.655, increases mean return from 190.438 to 192.045, and increases feasible delivery from 4/20 to 16/20 relative to ERC-TRPO (Table 1). Both paired contrasts survive Holm correction. ERC-TRPO-SG uses 398,310 acquired interaction steps on average, including 145,554 validation steps. ERC-TRPO-R also reaches 16/20, motivating the component attribution below.

6.3 Component Attribution and Delivery Opportunity

Retention and shadow repair. The 2×2 factorial identifies retention as the dominant source of final-delivery gain: with shadow training present, retention changes return by $+1.164$ $[0.626, 1.735]$, changes cost by -3.271 $[-4.559, -2.112]$, and raises feasibility from 2/20 to 17/20. Adding shadow to a retained pipeline changes return by $+0.020$ $[-0.573, 0.632]$ and cost by -0.072 $[-0.273, 0.107]$, while feasibility remains 17/20 (Table 2). Accepted shadow policies instead act online by changing the subsequent sampling trajectory.

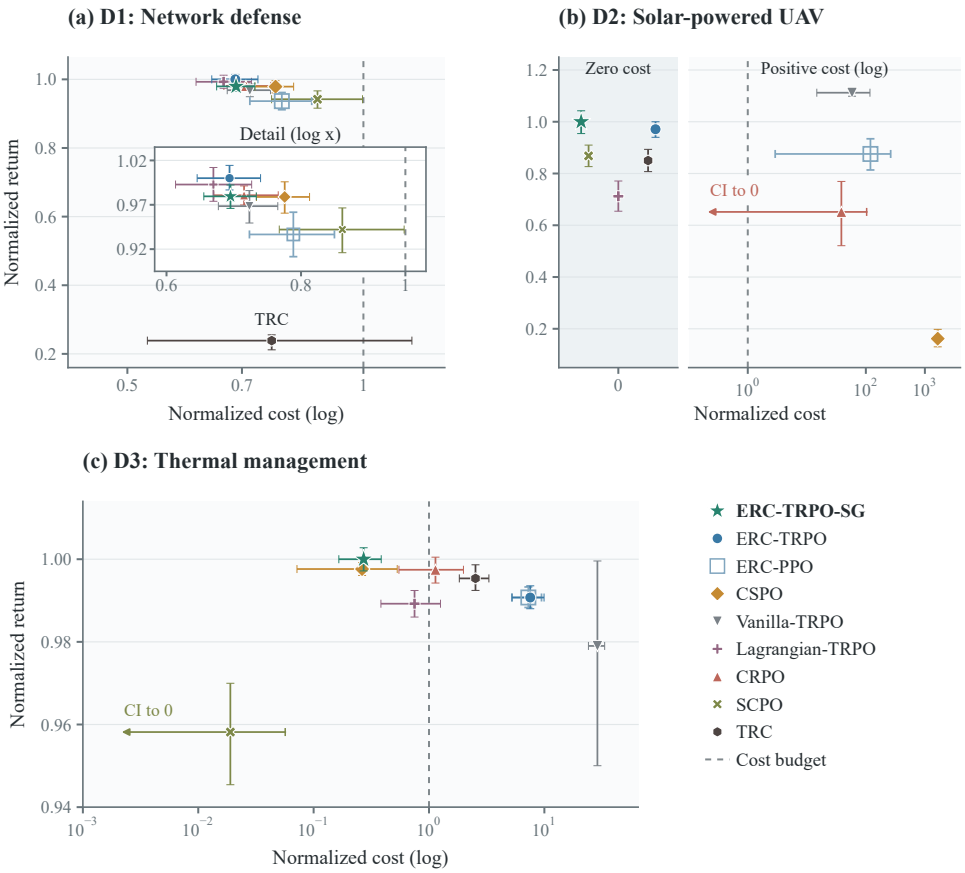

Figure 2: Primary three-domain results across nine methods. Cost is normalized by the domain budget; return by the highest-return mean-feasible method in each panel (ERC-TRPO in D1; ERC-TRPO-SG in D2–D3). Error bars are marginal 95% seed-bootstrap intervals; dashed lines mark the cost budget.

Table 1: D3 matched-total confirmation under a 400,000-interaction cap, seeds 31000–31019; independent 72-scenario evaluation.

Method	Return	Cost	Feasible	Acquired interaction steps
ERC-TRPO	190.438	5.118	4/20	400,000
ERC-TRPO-SG	192.045	0.655	16/20	398,310
ERC-TRPO-R	191.960	0.786	16/20	398,750

Candidate generation. Without shadow repair, retention, monitoring, or intermediate selection, ERC shaping reduces TRPO evaluation cost by 26.815 in D3. The corresponding benefit is not established within CSPO; Appendix I.1 reports the four-arm results, paired intervals, and corrected tests.

Candidate opportunity and selection. A matched two-checkpoint analysis yields cost opportunity 2.456 and selection loss 0.049, directly instantiating $G_J = O_J - L_J$. On the separate SG pool, minimum-UCB selection reaches 18/20 feasibility versus 16/20 for SG; Appendix H reports the selector analyses. The ablation assigns different roles to the two mechanisms: shadow repair changes the online sampling trajectory, whereas retention accounts for most of the measured final-delivery improvement.

Table 2: D3 shadow–retention factorial, seeds 600–619, using the common selector. Differences are on minus off (negative Δ cost means improvement); brackets denote 95% paired intervals.

Component contrast	Δ Return	Δ Cost	Feasible seeds / 20
Add retention with shadow	$1.164\ [0.626, 1.735]$	$-3.271\ [-4.559, -2.112]$	$2 \to 17$
Add shadow with retention	$0.020\ [-0.573, 0.632]$	$-0.072\ [-0.273, 0.107]$	$17 \to 17$
Add retention without shadow	$1.004\ [0.398, 1.613]$	$-7.668\ [-9.770, -5.690]$	$0 \to 17$

6.4 Response, Resource, and Long-Horizon Persistence

Finite-window response. Across 52,800 fixed-policy paths, positive-cost detection and multiplier first passage occur on distinct time scales. In the positive-drift cells, the log–log slope of geometric-mean first-passage time versus inverse calibrated drift is 0.977 with 95% interval $[0.936, 1.017]$. In 900 actor-updating tabular runs, every terminal policy is feasible but 307 multiplier paths never reach the response threshold. These results separate response timing from eventual policy feasibility.

Resource information before positive cost. Across 120 paired D3 interventions, resource replacement lowers subsequent discounted cost by $1.006\ [0.730, 1.227]$ and delays the first subsequent positive cost by $10.033\ [4.158, 17.617]$ steps. Thus observed resource state carries predictive information about future constraint trajectories before instantaneous cost becomes positive.

Public-benchmark persistence across the ERC family. We next evaluate the same temporal questions on nine Safety-Gymnasium tasks with five seeds per task and matched 10M-interaction budgets (Belouadah et al., 2026; Ji et al., 2023). ERC-U carries the ERC resource-aware update into this public continuous-control setting, while ERC-TRPO-SG additionally couples candidate competition, independent validation, and retention in the resource-coupled domains. CSPO reaches first post-violation recovery earlier (median 0.20M versus 0.38M interactions), whereas ERC-U records 14/450 over-budget task-epochs over the final 50 epochs, compared with 161/450 for CSPO. Under independent frozen-terminal evaluation, ERC-U is task-mean feasible on 8/9 tasks versus 6/9 for CSPO and has lower task-mean cost on 7/9 tasks. Initial recovery, persistent safety, and final policy delivery therefore provide a common empirical interpretation across the controlled finite-window studies and long-horizon public tasks. Appendix E reports the complete protocol and all task-level endpoints.

7 Conclusion

Finite-budget safe RL is a temporal problem as well as a constrained-optimization problem: resource degradation can signal future risk before positive cost appears, constraint response must occur within the remaining budget, and useful feasible policies must persist to final delivery. We formalize this structure through multiplier first-passage and candidate-opportunity/selection-loss analyses, and introduce the **ERC family** to operationalize resource-aware response and delivery. Its flagship ERC-TRPO-SG is the only method in our nine-method panel to achieve $20/20$ mean-feasible seeds in all three resource-coupled domains; under a matched 400k-interaction budget, it raises D3 feasible delivery from $4/20$ to $16/20$, with retention identified as the dominant endpoint mechanism. On nine Safety-Gymnasium tasks, ERC-U further separates early recovery from late persistence. Together, these results establish **response, persistence, and delivery as distinct finite-budget objectives** and make the training budget an explicit part of safe-RL design.

AI USE STATEMENT

- Yes, to aid or polish writing. Details are described in the paper.
- Yes, for retrieval and discovery (e.g., finding related work). Details are described in the paper.
- Yes, for research ideation or execution. Details are described in the paper.
- Yes, to draft sections of the paper. Details are described in the paper.
- Yes, for proving mathematical claims. Details are described in the paper.

ETHICS STATEMENT

The study evaluates simulated control using resource dynamics and exogenous workload and weather traces. It involves no human-subject experiments or interventions on physical devices. Reported feasibility concerns the stated cost definitions and evaluation protocols; it is not a certification for safety-critical deployment. The appendices report unsuccessful configurations, observed failures and the scope of each safety claim alongside the successful results.

REPRODUCIBILITY STATEMENT

Section 3 defines the cost conventions and budgets. Appendix A gives the theoretical assumptions and proofs, with extensions in Appendices F and H. Appendix B specifies the algorithm, training settings, validation panels and delivery rule. Appendices C–E provide complete endpoints, statistical definitions, source information and public-task protocols. Appendices H–J distinguish the additional experimental cohorts, selection procedures, interaction accounting and persistent-response extension. These details identify which policies, seeds and evaluation conventions underlie each reported result.

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, 2017.

Eitan Altman. *Constrained Markov Decision Processes*. Chapman & Hall/CRC, 1999.

Ayoub Belouadah, Sylvain KUBLER, and YVES LE TRAON. CSPO: Constraint-sensitive policy optimization for safe reinforcement learning. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=3ySR3TCMRP.

František Blahoudek, Tomáš Brázdil, Petr Novotný, Melkior Ornik, Pranay Thangeda, and Ufuk Topcu. Qualitative controller synthesis for consumption Markov decision processes. In *Computer Aided Verification*, volume 12225 of *Lecture Notes in Computer Science*, pp. 421–447. Springer, 2020. doi: 10.1007/978-3-030-53291-8_22.

J. Ji, B. Zhang, J. Zhou, X. Pan, W. Huang, R. Sun, Y. Geng, Y. Zhong, J. Dai, and Y. Yang. Safety Gymnasium: A unified safe reinforcement learning benchmark. In *Advances in Neural Information Processing Systems*, volume 36, 2023. Datasets and Benchmarks Track.

Dohyeong Kim and Songhwai Oh. TRC: Trust region conditional value at risk for safe reinforcement learning. *IEEE Robotics and Automation Letters*, 7(2):2621–2628, 2022.

N. Merket and R. Adhikari. Weather data for buildings energy simulations, 2020. Data set.

Microsoft Azure. Azure LLM inference trace 2024. Azure Public Dataset, 2024. URL https://github.com/Azure/AzurePublicDataset/blob/master/AzureLLMInferenceDataset2024.md. Data set.

Ted Moskovitz, Brendan O'Donoghue, Vivek Veeriah, Sebastian Flennerhag, Satinder Singh, and Tom Zahavy. ReLOAD: Reinforcement learning with optimistic ascent-descent for last-iterate convergence in constrained MDPs. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 25303–25336, 2023.

Andrew Y. Ng, Daishi Harada, and Stuart Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In *Proceedings of the Sixteenth International Conference on Machine Learning*, pp. 278–287, 1999.

Santiago Paternain, Luiz Chamon, Miguel Calvo-Fullana, and Alejandro Ribeiro. Constrained reinforcement learning has zero duality gap. In *Advances in Neural Information Processing Systems*, volume 32, pp. 7555–7565, 2019.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1889–1897, 2015.

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.

Aivar Sootla, Alexander I Cowen-Rivers, Taher Jafferjee, Ziyan Wang, David H Mguni, Jun Wang, and Haitham Ammar. Saute RL: Almost surely safe reinforcement learning using state augmentation. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 20423–20443, 2022.

Jovan Stojkovic, Chaojie Zhang, Íñigo Goiri, Josep Torrellas, and Esha Choukse. DynamoLLM: Designing LLM inference clusters for performance and energy efficiency. In *IEEE International Symposium on High Performance Computer Architecture*, 2025.

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, 2020.

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.

Tengyu Xu, Yingbin Liang, and Guanghui Lan. CRPO: A new approach for safe reinforcement learning with convergence guarantee. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 11480–11491, 2021.

L. Yang, Y. Wang, Y. Yu, Q. Weng, J. Dong, K. Liu, C. Zhang, Y. Zi, H. Li, Z. Zhang, N. Wang, Y. Dong, M. Zheng, L. Xi, X. Lu, L. Ye, G. Yang, B. Fu, T. Lan, L. Zhang, L. Qu, and W. Wang. GPU-disaggregated serving for deep learning recommendation models at scale. In *22nd USENIX Symposium on Networked Systems Design and Implementation*, pp. 847–863, 2025.

Yujie Yang, Jinglin Teh, Ziyu Lin, Kaicheng Yu, Tao Zhang, and Shengbo Eben Li. Breaking safety paradox with feasible dual policy iteration. In *International Conference on Learning Representations*, 2026.

Weiye Zhao, Rui Chen, Yifan Sun, Feihan Li, Tianhao Wei, and Changliu Liu. State-wise constrained policy optimization. *Transactions on Machine Learning Research*, 2024.

NOTATION

Table S1 summarizes the main symbols. Environment time t, learner update k and training seed are distinct indices. Cost functionals retain the units of their stated protocol.

Table S1: Notation and interpretation.

Symbol	Meaning
$s_t = (x_t, b_t)$	Markov state: task/context x_t and observed resource or safety margin b_t.
$a_t,\ r_t,\ c_t$	Action, reward and nonnegative safety cost.
$\pi_k,\ \pi_T,\ \pi_{\mathrm{out}}$	Policy at update k, terminal policy and delivered policy.
$\gamma,\ d$	Discount and cost budget in the specified units.
$J_R,\ J_C$	Theoretical expected discounted return and cost.
$J_C^{\mathrm{ev}}(\pi)$	Evaluation cost: D1 cost rate or D2/D3 discounted episode cost.
$\mathcal{F}_k,\ \widehat{C}_k$	Information available before batch k and the cost estimate obtained from that batch.
$Z_k,\ p_k,\ \mu_k$	Positive-batch-cost indicator, conditional arrival probability and conditional positive-cost magnitude.
$m_k,\ \underline{m},\ \xi_k$	Conditional cost-excess drift, its positive lower bound and the centered cost-excess increment.
$\nu_k,\ \alpha_\nu,\ \nu_{\max}$	Persistent multiplier, update step size and projection cap.
$\nu^\star,\ \tau_\nu$	Pressure threshold and first crossing time, distinct from policy feasibility.
$K,\ D_K,\ R_K$	Finite update window, non-detection event and post-detection non-crossing event.
$g,\ \sigma_-,\ \theta_-$	Scaled threshold gap and conditional lower-tail parameters.
$\mathcal{C},\ \mathcal{A},\ \mathcal{P}$	Candidate pool, rule-allowed subset and validation Pareto front.
$J,\ \widehat{J},\ \epsilon_J$	Prespecified higher-is-better endpoint, its estimate and a uniform estimation-error radius.
$O_J,\ L_J,\ G_J$	Candidate opportunity, selection loss and realized delivery gain, with $G_J = O_J - L_J$.
$D_J,\ \delta_J$	Rule-exclusion loss and empirical optimization gap within the allowed set.
$U_C, U_{0.95}(\Delta C)$, $L_{0.95}(\Delta R)$	Nominal cost score and paired difference bounds; not adaptive coverage guarantees.
$T_{\mathrm{rec}},\ e,\ s$	Predicted remaining recovery time, current positive cost excess and median validation slope.
$\Delta,\ \eta,\ a_k,\ T_K$	Initial violation margin, terminal tolerance, stopped indicator $\mathbf{1}\{k < \tau_\nu\}$ and $K \wedge \tau_\nu$ in Appendix F.
$O_K,\ s_k,\ L_K$	Cumulative positive cost excess, stopped margin shortfall and its cumulative sum.
$\mathcal{Z},\ z,\ x_{iz}$	Finite evaluation support, scenario index and frozen-policy scenario cost.
$\mathcal{S},\ \widehat{\mathcal{S}},\ \mathcal{S}^-$	True, conservatively estimated and interior feasible sets.
$\nu_k^{\mathrm{prop}},\ \ell_k,\ b_{\mathrm{bias}}$	Standard multiplier proposal, actor price and fixed bias in the extension.
$h_t,\ h_t^+,\ w_t^1,\ w_t^H$	Pre-/post-action resource headroom, local weight and fully observed resource-window weight.
$\mathcal{V}_k,\ v_k$	Observed window indices and credit ratio, distinct from multiplier ν_k.
$\varrho_t(\theta),\ a_k^{\mathrm{ep}}$	Policy likelihood ratio and episode-scaling factor in the extension's actor surrogate.

A Theoretical Conditions and Proofs

A.1 Notation and Sampling Conventions

The observed resource coordinate is part of the Markov state. A scalar consumption–replenishment representation is

$$b_{t+1} = \mathrm{clip}(b_t - C_{\mathrm{phys}}(s_t, a_t) + G_{\mathrm{phys}}(s_t, a_t, \zeta_t), 0, b_{\mathrm{cap}}(s_t)) \,. \tag{A2}$$

Here $C_{\mathrm{phys}}, G_{\mathrm{phys}} \geq 0$ are physical consumption and replenishment; ζ_t is an exogenous disturbance, distinct from the centered batch increment ξ_k. Environment time, update index, actor-interaction count and episode length are t, k, n, H. The updater receives $\widehat{C}_k$, and $\mathcal{F}_k$ contains information available before that batch. The resource b_t, cost budget d, multiplier ν_k and response threshold $\nu^\star$ are separate quantities.

In Figure 3, q controls entry into an absorbing hazardous state and differs from batch arrival probability p_k. Parameter β is the probability of an aggressive action within that state; it does not control exit. Protocol $P \in \{U, F, S\}$ determines cost aggregation. U uses an independent random horizon L with $\Pr(L \geq t) = \gamma^t$ and sums $\sum_{t=0}^{L} c_t$, an unbiased estimator of infinite discounted cost. Over 64 post-transition opportunities, F uses $\sum_{t=1}^{64} \gamma^{t-1} c_t$, whereas S uses $64^{-1} \sum_{t=1}^{64} c_t$. Both fixed-window protocols restart. Their common numerical budget 0.1 has different units. Figure 3 therefore uses protocol-specific analytic boundaries $q_P^\star$.

The controlled model starts in zero-cost state H. A cautious action stays in H; an aggressive action enters absorbing state C with probability q. Cautious and aggressive rewards are 0.7 and 1 in both states. Both actions cost zero in H; their costs in C are 0.001 and 1. The fixed policy always acts aggressively in H and does so with probability β in C. The geometric waiting time gives

$$A(q) = \frac{\gamma q}{1 - \gamma(1-q)}, \qquad D_C(q) = \frac{A(q)}{1-\gamma}, \qquad J_C = D_C(q)(0.001 + 0.999\beta).$$

The Lagrangian action-value difference in C is $0.3 - 0.999\nu$, yielding $\nu^\star = 100/333$. Where the constraint binds and the mixture is admissible, $\beta^\star = (d/D_C(q) - 0.001)/0.999$, capped at one when the budget is slack. An always-cautious C policy is not this binding mixture.

A.2 Zero-Cost Updates and the Raw Gradient Bound

Assumption 1 (Bounded critic and score). Throughout the zero-cost window, the cost critic satisfies the uniform absolute bound $|V_C(s)| \leq \varepsilon_V$ at every sampled and bootstrapped next state, the policy score has norm at most G, and no additional importance weights enter the estimator. For the raw, unnormalized GAE cost gradient,

$$\|g_C^{\mathrm{GAE}}\| \leq \frac{G(1+\gamma)\varepsilon_V}{1 - \gamma\lambda_{\mathrm{GAE}}}. \tag{A7}$$

Here $\lambda_{\mathrm{GAE}} \in [0, 1)$ is the trace-decay parameter (Schulman et al., 2016).

Substitution into Eq. (2) gives Eq. (4): upper projection is inactive and $[[x-a]_+ - a]_+ = [x - 2a]_+$ for $a \geq 0$. An unbaselined Monte Carlo estimator from complete all-zero-cost returns has zero advantages and zero cost gradient. For GAE, $|\delta_t^C| \leq (1+\gamma)\varepsilon_V$. Summing discounted residual magnitudes and multiplying by the score bound proves Eq. (A7). A finite window shortens the geometric sum. Bootstrap values outside the assumed bound require a boundary term; importance weighting requires a weight bound. Advantage normalization is outside this raw-magnitude statement.

A.3 Event Decomposition and Recovery Bounds

Define $\widehat{h}_k = \Pr(Z_k = 1 \mid D_k)$ whenever $\Pr(D_k) > 0$, with D_0 the whole sample space. Successive conditioning gives

$$\Pr(D_K) = \prod_{k=0}^{K-1} (1 - \widehat{h}_k). \tag{A10}$$

If an intermediate D_k has zero probability, the full no-detection event also has probability zero. On D_K, Eq. (4) keeps all multipliers below $\nu^\star$; partitioning $\{\tau_\nu > K\}$ by D_K proves Eq. (6).

For Proposition 1, Assumption 3 is the conditional lower-tail condition

$$\mathbb{E}[e^{-\theta\xi_k} \mid \mathcal{F}_k] \le \exp(\sigma_-^2\theta^2/2), \qquad 0 \le \theta \le \theta_-, \quad k < \tau_\nu. \tag{A11}$$

Set $I_k = \mathbf{1}\{k < \tau_\nu\}$ and $S_n = \sum_{k=0}^{n-1} I_k\xi_k$. Because I_k is $\mathcal{F}_k$-measurable, iterated conditioning gives $\mathbb{E}[e^{-\theta S_n} \mid \mathcal{F}_0] \le \exp(n\sigma_-^2\theta^2/2)$. On $\{\tau_\nu > n\}$ the upper cap has not been active, and lower reflection only increases the multiplier. Thus $S_n < g - n\underline{m}$. For $n \ge \lceil 2g/\underline{m}\rceil$, exponential Markov with $\theta^\star = \min\{\theta_-/2, \underline{m}/(2\sigma_-^2)\}$ yields

$$\Pr(\tau_\nu > n \mid \mathcal{F}_0) \le \exp\bigl(-n\theta^\star\underline{m}/2 + n\sigma_-^2(\theta^\star)^2/2\bigr) \le e^{-c_- n}, \quad c_- = \theta^\star\underline{m}/4. \tag{A20}$$

Summing the tail, bounding its first $\lceil 2g/\underline{m}\rceil$ terms by one, proves Eq. (7). No positive-drift assumption is needed after first passage.

A complementary bound uses an upper bound on mean cost. If $\mathbb{E}[\widehat{C}_k \mid \mathcal{F}_k] \le \bar{C}$ throughout the window, then $W_k = \nu_k/\alpha_\nu$ satisfies $W_{k+1} \le W_k + \widehat{C}_k$. Crossing requires cumulative cost at least $g = (\nu^\star - \nu_0)/\alpha_\nu$. Conditional Markov gives

$$\Pr(\tau_\nu \le K \mid \mathcal{F}_0) \le \min\left\{1, \frac{K\bar{C}}{g}\right\} = \min\left\{1, \frac{\alpha_\nu K\bar{C}}{\nu^\star - \nu_0}\right\}. \tag{A21}$$

The mean bound alone does not imply concentration. For an optional refinement, assume additionally, for every update in the window,

$$\mathbb{E}\Bigl[\exp\{\theta(\widehat{C}_k - \mathbb{E}[\widehat{C}_k \mid \mathcal{F}_k])\} \mid \mathcal{F}_k\Bigr] \le \exp\left(\frac{\theta^2\sigma_+^2}{2(1-b_+\theta)}\right), \quad 0 \le \theta < 1/b_+.$$

Here $\sigma_+, b_+ > 0$ are deterministic. For $a_K = g - K\bar{C} > 0$, iterated conditioning and exponential Markov at $\theta = a_K/(K\sigma_+^2 + b_+a_K)$ give $\Pr(\tau_\nu \le K \mid \mathcal{F}_0) \le \exp[-a_K^2/(2(K\sigma_+^2 + b_+a_K))]$. The lower drift bound controls delayed crossing; the upper mean and upper-tail conditions control early crossing.

A.4 Negative Drift and the Critical Case

Let $X_k = \widehat{C}_k - d$, $W_0 = \nu_0/\alpha_\nu$ and $h = \nu^\star/\alpha_\nu = W_0 + g$. The absolute threshold h and initial distance g coincide only when $\nu_0 = 0$. If, for some $\lambda_+ > 0$, $\mathbb{E}[e^{\lambda_+ X_k} \mid \mathcal{F}_k] \le 1$ throughout the window, then

$$\Pr(\tau_\nu \le K \mid \mathcal{F}_0) \le \min\{1, e^{-\lambda_+ g} + Ke^{-\lambda_+ h}\}. \tag{A13}$$

Before first crossing, the recursion is the reflected walk $W_{k+1} = [W_k + X_k]_+$. The exponential of any cumulative increment sequence is a nonnegative supermartingale. The initial excursion reaches h with probability at most $e^{-\lambda_+(h-W_0)}$. Each subsequent excursion from zero has bound $e^{-\lambda_+ h}$; at most K can start within K updates. A union bound proves Eq. (A13), without independent excursions.

Zero drift alone gives no such strictly positive transform parameter. Independent costs equal to 0 or $2d$ with equal probability yield symmetric increments $\pm d$ and can reach any finite threshold. Constant cost d leaves the multiplier unchanged. Equal zero drift can therefore produce different recovery behavior.

A.5 The Actor Bridge and Resource Invariance

Assumption 4 (Policy-infeasibility persistence). For $\Delta > 0$, $J_C^{\mathrm{ev}}(\pi_0) \ge d + \Delta$, and whenever $\nu_k < \nu^\star$ in the window, $J_C^{\mathrm{ev}}(\pi_{k+1}) \ge d + \Delta$. This is condition B8 in the experimental records. If it holds along a path, $\tau_\nu > K$ implies $J_C^{\mathrm{ev}}(\pi_K) \ge d + \Delta$ and $\nu_K < \nu^\star$, hence the terminal event F_K in Eq. (F14) for $\eta = \Delta/2$. The converse fails: a low terminal multiplier can follow an earlier crossing. Appendix F measures persistence exceptions and retains this distinction.

The existence of a resource-recovery action is weaker than uniform safe invariance. A scalar recovery branch admits a distance-over-increment bound only when its net increment is at least $\epsilon > 0$ at every reachable full state and disturbance. Bursts and damage in D1, and consumption without sunlight in D2, can violate this premise. D3 requires an invariant set in the full temperature space. Observed recovery rates do not establish these uniform conditions.

A.6 Visible History and Two-Point Identification

Let H_k contain all information available to the learner before batch k, including resources, actions, rewards, costs, validation results and internal randomization. Two environments induce history laws P_0, P_1 and output laws Q_0, Q_1 through a common decision kernel. Define $\mathrm{TV}(P, Q) = \sup_B |P(B) - Q(B)|$.

Lemma A.1 (Information-limited decisions). If successful decision sets B_0, B_1 are disjoint, then

$$Q_0(B_0^c) + Q_1(B_1^c) \geq 1 - \mathrm{TV}(P_0, P_1). \tag{A1}$$

Proof. Data processing gives $\mathrm{TV}(Q_0, Q_1) \leq \mathrm{TV}(P_0, P_1)$. Hence $Q_0(B_0) \leq Q_1(B_0) + \mathrm{TV}(P_0, P_1)$, while disjointness implies $Q_1(B_0) + Q_1(B_1) \leq 1$. Rearrangement proves the claim. Identical history laws force failure probability at least one half in one environment. This concerns full visible information, not cost alone. A common successful decision invalidates disjointness; observed resource differences can distinguish histories even when both current costs are zero.

A.7 Conditional Separation of Resource States and Future Cost

Fix the non-resource state x, current action a, subsequent open-loop actions and conditional exogenous law. Write current cost as $c(x, b, a)$ and the future-cost-vector law as K_b. For interventions on the next thermal state, these are intervention laws rather than observational conditionals.

Proposition A.2 (Predictive insufficiency of instantaneous cost). If $b \neq b'$ satisfy $c(x, b, a) = c(x, b', a)$ but $K_b \neq K_{b'}$, no single-valued kernel of the fixed quantities and instantaneous cost can represent both future laws. Different expectations of an integrable future-cost functional suffice to distinguish the laws.

Proof. A kernel maps identical inputs to identical distributions, contradicting the premise; identical laws also give identical integrable expectations. For a concrete CMDP, let $b_{t+1} = \max\{0, b_t - a_t\}$ and $c_t = [0.25 - b_{t+1}]_+$. Starting from $b_0 = 0.6$ and $b'_0 = 1.0$, action $a_0 = 0.2$ gives zero current cost in both. A second action $a_1 = 0.2$ yields $c_1 = 0.05$ versus $c'_1 = 0$, so the future laws are different point masses. The construction neither requires hidden resources nor asserts separation in every resource CMDP.

Figure 4 intervenes on D3's two post-action temperatures while holding the current action and recorded zero cost fixed. It shares subsequent open-loop actions and exogenous traces. The resulting differences establish separation within this simulated intervention family. Each arm follows its own termination, so the total effect includes exposure duration.

B Algorithms and Reproduction

This appendix gives the training rewards, recovery predictor, full algorithm and validation protocol. It also records baseline settings and experimental stages.

B.1 Validation and Output Decisions

ERC-TRPO updates the policy with TRPO (Schulman et al., 2015) and shapes finite-batch learning with resource-related reward components. Raw evaluation rewards and costs remain unchanged. In D1, the deployed training reward is $r_t^{\mathrm{train}} = r_t^{\mathrm{raw}} - 20c_t^2$. The raw task reward includes a recovery bonus with coefficient 30 for all methods. That term belongs to the common task specification, not to our contribution. D2 and D3 use

$$r_t^{train} = r_t^{raw} + w_\Phi \left[\gamma \Phi(s_{t+1}) - \Phi(s_t)\right] - w_f \, c_t^2, \qquad w_\Phi = w_f = 1, \tag{B15}$$

Here Φ is a resource-derived potential, w_Φ its weight, and w_f the cost weight. Both domains set the next-state potential to zero at termination and at fixed-horizon truncation. The formal coefficient of the additional shutdown penalty in D3 is zero.

The frozen potentials are

$$\Phi_{\mathrm{D2}}(s) = \rho(s)^2,$$
$$\Phi_{\mathrm{D3}}(s) = \left[\min\left\{\mathrm{clip}_{[0,1]}\left(\frac{100 - T_{\mathrm{chip}}(s)}{80}\right), \mathrm{clip}_{[0,1]}\left(\frac{35 - T_{\mathrm{inlet}}(s)}{20}\right)\right\}\right]^2.$$

Here $\rho(s) \in [0,1]$ is the observed UAV resource reserve, and T_{chip} and T_{inlet} are temperatures in degrees Celsius. The clipping operator is $\mathrm{clip}_{[0,1]}(x) = \min\{1, \max\{0, x\}\}$. D3 normalizes chip and inlet headroom by $100 - 20 = 80$ and $35 - 15 = 20$, respectively, using the frozen trip and emergency thresholds. Both domains use shaping discount $\gamma = 0.99$ and the terminal-potential convention above.

The recovery prediction uses the median of three pairwise validation slopes and the current positive cost excess:

$$T_{\mathrm{rec}} = \begin{cases} 0, & e = 0, \\ e/(-s), & e > 0,\ s < -10^{-12}, \\ +\infty, & e > 0,\ s \geq -10^{-12}. \end{cases} \tag{B16}$$

Here $e = [\bar{C}_3 - d]_+$ and s is the median pairwise slope of the last three validations, as in Section 5.1. The deadband has cost-per-actor-step units; $+\infty$ means no finite recovery is predicted.

ALGORITHM 1 ERC-TRPO-SG

Require: cost budget d; initial policy; training-step target; fixed validation panel of 16 scenarios; validation period of eight updates.

1: Initialize the online policy, accepted-repair count, validation history, recovery mode, and retained representatives.

2: While the training protocol permits another batch:

3: Record the pre-update policy and the recovery mode at the start of the update.

4: Collect a batch using the online policy and form the domain-specific training reward of Section 5.1.

5: Compute the main TRPO candidate from this policy and batch.

6: If the recorded mode is recovery:

7: Compute the shadow from the same pre-update policy and batch: cost-descent TRPO in D1, local CPO in D2/D3.

8: Evaluate both candidates on the paired validation panel.

9: If Eqs. (14)–(15) pass, adopt the shadow and increment the total and per-phase acceptance counts; otherwise adopt the main candidate.

10: Exit recovery after rejection, after the accepted policy's mean cost meets the budget, or after three accepted repairs in the phase.

11: Otherwise adopt the main candidate; evaluate it if eight updates have elapsed or training has reached its endpoint.

12: Whenever validation occurs, append the adopted policy's mean cost to the three-point history and update the retained representatives.

13: For a non-recovery validation, compute Eq. (B16) once three observations exist; update the consecutive-trigger count.

14: If two consecutive eligible validations predict insufficient recovery, schedule recovery for the next update and reset the phase count.

15: End while; ensure the terminal policy has its final validation statistics.

16: Form the distinct retained set containing the terminal, best-feasible-UCB, and minimum-UCB representatives.

17: If no shadow was accepted, or the terminal cost UCB meets the budget, return the terminal policy.

18: Otherwise return the safety-anchored Pareto choice of Eq. (16), using the stated tie-breaking order.

The local CPO shadow in line 7 uses the global expected discounted episode-cost budget d, passed as `d_exp`: 0.0005 in D2 and 1.0 in D3. Its batch surrogate equals the sampled mean episode cost at the pre-update policy and adds an episode-scaled, discount-weighted cost-advantage correction using the likelihood ratio minus one. If no episode completes in the batch, the mean cost-critic return target supplies the initial estimate. The local linearized constraint uses this surrogate minus d; it does not introduce a separate batch-local budget or a zero-cost target.

When the initial surrogate cost $\widehat{C}_{\mathrm{old}} > d$, the internal line search requires a surrogate decrease of at least $\max\{10^{-8}, 0.01(\widehat{C}_{\mathrm{old}} - d)\}$, rather than feasibility in a single step. Otherwise it requires surrogate cost at most $d + 10^{-8}$ and no surrogate reward decrease beyond 10^{-8}. Both branches enforce the KL limit 0.01. A failed mixed recovery direction may fall back to natural cost descent; if neither is accepted, the pre-update policy is restored. These internal tests generate the shadow; adopting it still requires the separate paired-validation conditions in Eqs. (14)–(15).

The primary protocol fixes actor-training steps and records validation interactions separately. The matched-total-interaction protocol reserves the required decision validation before admitting another batch and charges both types of interaction to its cap. For the separately evaluated extension only, compute the multiplier proposal, release and actor price after collecting the batch and before the main update in line 5; modify that main surrogate by Eq. (J2). Shadow acceptance does not reset the multiplier.

The controller validates every eight updates on 16 scenarios, estimates its three-point trend after at least three validations, and enters recovery after two consecutive triggers. D1/D2 validation seeds are 4300–4315. D3 development selectors are 3, 11, 18, 26, 33, 41, 48, 56, 63, 71, 78, 86, 93, 101, 108 and 116. Evaluation scenarios 2000–2049 are excluded from acceptance and selection.

The one-sided t upper bound is $\overline{x} + t_{.95,n-1} s_x / n^{1/2}$; the lower bound changes the plus to minus, with n paired scenarios. Cost and return differences use the same scenarios. The 95% level is nominal for one comparison and endpoint. Repeated panel reuse and checkpoint selection introduce selection effects; the final 50-scenario panel supplies independent endpoint evaluation. All-zero validation cost yields a zero t-bound, which cannot rule out rare off-panel risk.

Candidate π_{best} has the highest validation return among cost-UCB-feasible policies; π_{min} has the lowest cost UCB. Ties use return and checkpoint time. The terminal policy always exists, so the pool is nonempty. Parameters are deduplicated before forming the Pareto front. The identity rule returns terminal if no repair was accepted, otherwise prioritizes a UCB-feasible terminal, then applies Eq. (16). The final Pareto choice breaks return ties by lower cost UCB and then the later checkpoint, with a 10^{-15} numerical boundary tolerance.

Table B1: Additional validation interactions and recovery activity.

Domain	Training steps	Mean validation steps	Validation range	Validation / training	Seeds accepting repair	Mean acceptances
D1	401408	96,970.1	80,606–109,531	24.2%	1/20	0.05
D2	400000	69,576.2	60,079–76,949	17.4%	0/20	0.00
D3	400000	246,733.7	161,280–308,736	61.7%	20/20	12.85

Table B1 uses the last cumulative validation counter in each training log and excludes formal evaluation. D3 accepts 12.85 repairs on average, reflecting repeated recovery phases. With zero acceptance, the identity gate retains that run's terminal policy. Monitoring calls, random-number management and run conditions can still differ from historical backbone runs, so D2 endpoint differences are not shadow gains.

Policy and critic gradients use only collected training batches. Validation runs deterministic policies in separate instances without backpropagation; its outcomes feed recovery and retention. These interactions provide decision information rather than extra gradient batches. D1 uses 401,408 actor steps and D2/D3 use 400,000. Training, decision validation and final evaluation are charged separately.

B.2 Baseline Identity and Training Settings

D1 uses discrete policies; its TRC-Discrete-LQCLP adaptation does not represent the original continuous-action implementation directly. D2/D3 use continuous actions, hidden widths 64 and 64, discount 0.99, GAE parameter 0.95, 2,048 samples per update, maximum KL 0.01, damping 0.1, value learning rate 0.001 and 40 value-fit steps per update. Their last batch is truncated to budget; D1 retains its rollout boundary.

Vanilla-TRPO optimizes raw reward; ERC-TRPO adds the domain reward terms in Eq. (B15); ERC-PPO changes the policy optimizer. Lagrangian-TRPO learns a persistent projected multiplier. CRPO switches reward and cost updates and retains feasible iterates. SCPO uses state-conditioned local maximum-cost constraints and safe initialization; TRC uses a Gaussian-CVaR surrogate; CSPO combines a learned multiplier, cost-gradient sensitivity correction and a fixed linear cost term. A shared evaluation budget does not equate these internal objectives or initializations.

D1 replenishment includes $0.12(1-b)^{1.5}$ and damage scaling. Its recovery reward uses nominal replenishment, not net growth after bursts and clipping. Cost is recorded before the damage update, while the final resource log is recorded afterwards; the latter cannot reconstruct cost pointwise. D2/D3 treat fixed-horizon truncation as terminal in the potential. Historical D3 backbone chip temperatures stay below their soft threshold, with safety events primarily driven by inlet temperature.

Historical D1 uses SB3 2.9.0 with Torch 2.5.1, outside SB3's declared dependency range. Models have been replayed in their actual environment, with the failed dependency check retained. Historical D3 formal runs use GPU; the new four-arm reward controls all use CPU. Component effects are estimated within that CPU cohort, not interpreted as hardware effects relative to historical models.

B.3 Names and Experimental Stages

ERC-TRPO denotes the backbone, ERC-TRPO-SG the principal method and ERC-TRPO-U the Safety-Gymnasium adaptation (Ji et al., 2023). Archived V5P-B maps to SG; ERCTRPOSGUCB maps to U. Earlier V5F shares online training but differs in output choice, and is a historical selection ablation rather than a tenth independent learning baseline.

The principal rule and 600–619 component panel were fixed before the 200–219 matched panel. Relative to the earlier output rule, D1 return rises by 63.967 and feasibility from 18/20 to 20/20, but the return difference does not survive the selector-family Holm correction. D2/D3 endpoints are identical. The output-rule revision remains in the record.

The theoretical text uses stage-one v6 and stage-two v2. Controlled experiments originally followed stage-one v5; v6 subsequently clarified compatibility and tightened statements. The present wording was therefore not entirely preregistered.

C Complete Endpoints and Statistical Definitions

This appendix reports the full endpoints and the statistical definitions used in the main text. Section C.1 gives the nine learning methods. Section C.2 lists the non-learning controllers. Section C.3 defines the paired comparisons and feasibility criteria. Section C.4 reports episode-cost occurrence and budget exceedance.

C.1 Nine Learning Methods

All values below come from the matched per-seed panel. Return is the raw episode mean. Cost follows the domain definition. Brackets give marginal 95% seed-bootstrap intervals, not a joint return-cost region. The final column counts training seeds whose evaluation mean satisfies budget.

Table C1: Complete D1 learning-method results.

Method	Return [95% CI]	Cost [95% CI]	Feasible seeds
ERC-TRPO-SG	2983.524 [2942.115, 3022.619]	0.00344 [0.003249, 0.003635]	20/20
ERC-TRPO	3045.924 [3005.459, 3089.807]	0.003433 [0.003203, 0.003668]	20/20
CSPO	2981.441 [2925.813, 3033.412]	0.003863 [0.003636, 0.004074]	20/20
Vanilla-TRPO	2949.836 [2891.961, 3003.406]	0.003582 [0.003352, 0.003806]	20/20
Lagrangian-TRPO	3024.089 [2966.534, 3082.571]	0.003316 [0.003059, 0.003599]	20/20
CRPO	2987.229 [2953.803, 3019.423]	0.003541 [0.003319, 0.00381]	19/20
SCPO	2869.496 [2790.023, 2943.748]	0.004369 [0.003818, 0.004987]	14/20
TRC	726.736 [645.126, 778.181]	0.003819 [0.002651, 0.005765]	19/20
ERC-PPO	2852.728 [2776.427, 2929.018]	0.003935 [0.003582, 0.004296]	19/20

Table C2: Complete D2 learning-method results.

Method	Return [95% CI]	Cost [95% CI]	Feasible seeds
ERC-TRPO-SG	5.207 [4.969, 5.427]	0 [0, 0]	20/20
ERC-TRPO	5.056 [4.893, 5.206]	0 [0, 0]	20/20
CSPO	0.843 [0.678, 1.030]	0.825 [0.728, 0.923]	0/20
Vanilla-TRPO	5.791 [5.720, 5.852]	0.029 [0.007383, 0.059]	10/20
Lagrangian-TRPO	3.708 [3.406, 4.014]	0 [0, 0]	20/20
CRPO	3.392 [2.713, 4.006]	0.019 [0, 0.052]	17/20
SCPO	4.519 [4.304, 4.737]	0 [0, 0]	20/20
TRC	4.426 [4.203, 4.651]	0 [0, 0]	20/20
ERC-PPO	4.559 [4.236, 4.864]	0.060 [0.001457, 0.132]	16/20

Table C3: Complete D3 learning-method results.

Method	Return [95% CI]	Cost [95% CI]	Feasible seeds
ERC-TRPO-SG	188.745 [188.211, 189.269]	0.271 [0.165, 0.385]	20/20
ERC-TRPO	186.996 [186.483, 187.530]	7.544 [5.241, 9.936]	3/20
CSPO	188.297 [188.005, 188.594]	0.263 [0.072, 0.530]	19/20
Vanilla-TRPO	184.783 [179.313, 188.665]	28.822 [24.094, 33.309]	0/20
Lagrangian-TRPO	186.713 [186.103, 187.316]	0.749 [0.383, 1.258]	17/20
CRPO	188.263 [187.657, 188.839]	1.138 [0.549, 1.989]	11/20
SCPO	180.850 [178.445, 183.079]	0.019 [0, 0.057]	20/20
TRC	187.869 [187.320, 188.490]	2.529 [1.831, 3.305]	4/20
ERC-PPO	186.997 [186.540, 187.477]	7.262 [5.235, 9.445]	0/20

Figure 2 fixes return reference values at 3045.923582, 5.206529 and 188.745038 for D1–D3. These are selected only among mean-feasible methods in the nine-method panel. Error bars condition on these references rather than re-bootstrap them. The D2 rule-controller return 6.164 is not used to renormalize the learning panel. Normalized scores do not establish optimality among all controllers.

C.2 Non-Learning Controllers

Fixed controllers use the same 50 evaluation scenarios: four D1, five D2 and six D3 controllers, totaling 750 records. Without training randomness, D1/D2 intervals resample scenarios. D3 clusters by five workload-weather combinations. These intervals have a different sampling unit from learning-method seed intervals. Controller names label rules; a "safe" label does not guarantee the present budget.

Table C4: D1 non-learning controllers.

Fixed policy	Return [95% CI]	Cost [95% CI]	Mean feasible
Aggressive	635.840 [487.013, 795.191]	0.036 [0.032, 0.041]	No
Heuristic	380.786 [338.103, 424.920]	0.021 [0.018, 0.023]	No
Preset safe rule	418.198 [260.754, 587.054]	0.041 [0.037, 0.045]	No
Random	413.986 [310.896, 529.593]	0.022 [0.019, 0.026]	No

Table C5: D2 non-learning controllers.

Fixed policy	Return [95% CI]	Cost [95% CI]	Mean feasible
Aggressive	6.172 [5.968, 6.347]	0.001502 [0, 0.004312]	No
Full thrust	0.206 [0.187, 0.227]	1.577 [1.505, 1.652]	No
Heuristic	6.164 [5.954, 6.333]	0 [0, 0]	Yes
Random	0.215 [0.192, 0.238]	1.738 [1.640, 1.841]	No
Conservative	1.775 [1.483, 2.068]	0 [0, 0]	Yes

Table C6: D3 non-learning controllers.

Fixed policy	Return [95% CI]	Cost [95% CI]	Mean feasible
Conservative	75.138 [33.933, 132.880]	0 [0, 0]	Yes
Energy-saving	191.194 [182.937, 202.670]	19.923 [19.249, 20.545]	No
Max service / max cooling	174.642 [168.009, 183.816]	0 [0, 0]	Yes
Max service / moderate cooling	183.062 [176.436, 192.230]	0.489 [0.114, 0.765]	Yes
Random	170.207 [154.583, 192.149]	0.233 [0.064, 0.384]	Yes
Thermostatic	189.858 [180.773, 202.493]	17.106 [16.232, 17.797]	No

C.3 Paired Comparisons and Feasibility

The main test family contains twelve contrasts: SG against ERC and CSPO, across three domains and two endpoints. Output-selection and D3 component contrasts use their own families. A bootstrap interval excluding zero and a corrected permutation test are different criteria. Both are retained.

Zero cost means no positive cost observed in the fixed panel. Seed-level mean feasibility, feasibility of the method mean, and feasibility of its mean-cost upper interval are separate statistics. D1 TRC and D3 Lagrangian-TRPO have means below budget but upper intervals above it. SG meets all three panel criteria in each domain. These do not imply zero population risk or trajectorywise safety.

C.4 Episode Cost Occurrence and Budget Exceedance

This analysis reaggregates 27,000 frozen endpoints: nine methods, three domains, 20 seeds and 50 episodes. No episodes are added. Positive-cost occurrence means at least one positive cost in the episode. Exceedance compares the episode's domain cost against budget: undiscounted cost divided by episode steps in D1, discounted finite-episode cost in D2/D3. Occurrence is not a hard-shutdown or irreversible-damage rate.

Each seed's 50-episode rate is computed first, then the 20 seeds are averaged equally. Intervals use 100,000 seed bootstraps, conditional on the saved scenarios. The 1,000 episodes are not treated as independent training replicates. A zero-event cell has a degenerate zero bootstrap interval, not a zero upper bound for unknown scenarios.

Table C7: D1 episode-risk percentages, with 95% seed-bootstrap intervals.

Method	Positive-cost episodes	Over-budget episodes
ERC-TRPO-SG	83.5 [81.3, 85.8]	22.4 [20.3, 24.6]
ERC-TRPO	81.8 [79.8, 83.7]	22.1 [19.6, 24.5]
ERC-PPO	87.7 [85.9, 89.6]	27.0 [23.8, 30.2]
Vanilla-TRPO	83.9 [81.5, 86.2]	24.9 [22.7, 27.0]
Lagrangian-TRPO	84.2 [82.1, 86.3]	22.8 [20.5, 25.2]
CRPO	84.0 [81.9, 86.2]	24.2 [22.4, 26.2]
SCPO	86.7 [84.5, 88.8]	27.2 [24.6, 29.9]
TRC	82.0 [78.9, 85.2]	26.8 [19.6, 36.5]
CSPO	85.8 [83.3, 88.1]	26.3 [23.9, 28.7]

Table C8: D2 episode-risk percentages, with 95% seed-bootstrap intervals.

Method	Positive-cost episodes	Over-budget episodes
ERC-TRPO-SG	0.0 [0.0, 0.0]	0.0 [0.0, 0.0]
ERC-TRPO	0.0 [0.0, 0.0]	0.0 [0.0, 0.0]
ERC-PPO	7.8 [0.1, 17.4]	7.8 [0.1, 17.4]
Vanilla-TRPO	7.5 [3.1, 12.6]	7.4 [3.0, 12.6]
Lagrangian-TRPO	0.0 [0.0, 0.0]	0.0 [0.0, 0.0]
CRPO	2.0 [0.0, 5.0]	2.0 [0.0, 5.0]
SCPO	0.0 [0.0, 0.0]	0.0 [0.0, 0.0]
TRC	0.0 [0.0, 0.0]	0.0 [0.0, 0.0]
CSPO	73.7 [63.8, 83.6]	73.7 [63.8, 83.6]

Table C9: D3 episode-risk percentages, with 95% seed-bootstrap intervals.

Method	Positive-cost episodes	Over-budget episodes
ERC-TRPO-SG	71.6 [56.0, 85.7]	4.2 [1.4, 7.5]
ERC-TRPO	94.8 [85.9, 100.0]	79.9 [63.7, 93.6]
ERC-PPO	98.9 [96.7, 100.0]	83.9 [74.6, 92.3]
Vanilla-TRPO	99.7 [99.1, 100.0]	97.4 [92.2, 100.0]
Lagrangian-TRPO	78.9 [65.4, 90.9]	12.9 [4.6, 23.6]
CRPO	74.4 [59.4, 87.7]	27.8 [16.4, 39.9]
SCPO	3.7 [0.0, 10.8]	0.0 [0.0, 0.0]
TRC	97.3 [93.5, 100.0]	73.7 [59.1, 86.7]
CSPO	13.5 [6.5, 21.1]	6.9 [2.6, 12.1]

Across SG seeds, D1 positive-cost occurrence has median 84% and range 76%–94%. Budget exceedance has median 23% and range 16%–32%. Both are zero in D2. D3 values are 88% (4%–100%) and 0% (0%–22%), respectively. These episode distributions coexist with mean-feasible outputs for every seed. They illustrate the distinct information in occurrence, cumulative cost and seed means.

D Mechanism and Resource Experiments

This appendix reports the controlled experiments behind the mechanism claims. Section D.1 describes the fixed-policy study and persistence opportunities. Section D.2 covers the D1 visibility and resource-coupling supplement. Section D.3 reports reward weights and sensor degradation. Section D.4 gives

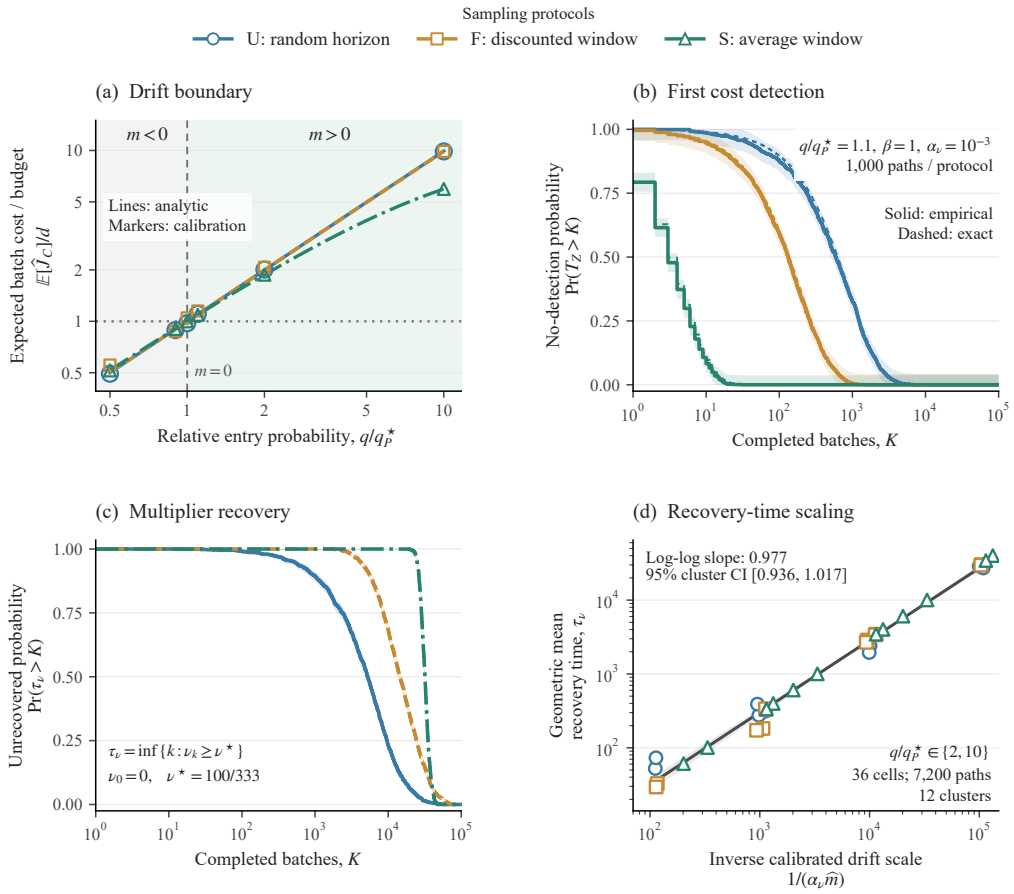

Figure 3: Controlled cost-arrival and multiplier-recovery experiments: estimator calibration, non-detection, recovery survival and positive-drift scaling. Protocol-specific cost units and analytic boundaries are defined in Appendix A; Appendix D.1 specifies sample sizes and uncertainty. The positive-drift scaling slope is 0.977 [0.936, 1.017].

feedback bundles and target-environment retraining. Section D.5 details the resource intervention and source structure. Section D.6 reports recovery actions and four-arm reward controls. Section D.7 compares frozen policies across trace sources. Section D.8 diagnoses D2 action efficiency and value estimation. Section D.9 audits shadow and retention usage.

D.1 Controlled Experiments and Persistence Opportunities

The fixed-policy study sets

$$\gamma = 0.99, \quad \nu_0 = 0, \quad \nu^\star = 100/333, \quad \nu_{\max} = 100,$$

$$\beta \in \{0.5, 1\}, \quad q/q_P^\star \in \{0.5, 0.9, 1, 1.1, 2, 10\}, \quad \alpha_\nu \in \{10^{-4}, 10^{-3}, 10^{-2}\}.$$

Figure 3(a)'s eighteen calibration cells fix $\beta = 1$ and $\alpha_\nu = 10^{-3}$. U uses 2,000,000 independent calibration batches per cell. F and S use 200,000. The original calibration retained no uncertainty interval to propagate.

Figure 3(b) uses 10,000 empirical bootstraps and the maximum whole-curve deviation within each cell. Radii are 0.042, 0.042 and 0.037. These are not jointly covering across protocols. Every selected Figure 3(c) path recovers before administrative censoring, so empirical survival agrees

with Kaplan–Meier. Wilson intervals are pointwise. Figure 3(d) uses 200 paths in each of 36 cells, resampling twelve protocol–β–q combinations while retaining their three step sizes.

Among 900 initially infeasible actor paths, persistence holds over windows of 1, 5, 25, 100, 1,000, 5,000 and 20,000 updates on 713, 505, 275, 144, 91, 83 and 9 paths. The counts also satisfying no multiplier recovery are 713, 505, 275, 144, 54, 0 and 0. Both counts are needed. Unvisited-state tabular parameter invariance passes where an opportunity exists. Other cells are not applicable.

A shared-actor positive control explicitly constrains updates to preserve infeasibility. Its 120 long paths contain 2,400,000 opportunities, 133,832 guard triggers and 20,800 prevented counterfactual exits. Here an exit means leaving the stipulated infeasible region, not violating safety. This control checks a logical condition. It is not evidence of the same long-run property for an ordinary shared actor.

D.2 D1 Visibility and Resource Coupling

The independent D1 supplement uses seeds 300–319. The visibility intervention's detection coefficient is 1.6403, with Holm p approximately 3.70×10^{-7}. The algorithm does not read the visibility parameter. The local resource-coupling/drift slope is 0.0004047 [0.0003557, 0.0004504]. Observed drift remains between -0.0007911 and -0.0003553 throughout, without a zero-crossing transition.

The full-cost training windows in this supplement contain no all-zero batch. Reduced visibility creates zero windows through an observation-channel mask. The experiment therefore supports a detection-channel intervention, not a claim that every natural D1 failure comes from an all-zero-cost batch.

D.3 Reward Weights and Sensor Degradation

The reward-weight expansion contains 760 new trainings and 160 reused models, with 20 seeds per cell. Reused endpoints are not independent confirmation. D3 weights 0, 0.1, 0.3, 0.5, 0.7 and 1 yield costs approximately 26.901, 21.334, 12.332, 8.528, 7.559 and 5.054. The descriptive decrease does not make adjacent contrasts Holm-significant. D1 and D2 show no uniform monotonic gain either.

The linear-versus-squared-penalty cost difference in D2 is zero, with return difference 0.0342 $[-0.1825, 0.2548]$. D3 cost difference is -2.322 $[-4.933, 0.357]$, Holm p = 0.223. Squared penalties thus have no established advantage over linear penalties in this comparison. D3 is sensitive to critic learning rate: reducing the full configuration's rate to 0.1 times its original value changes cost by -4.065 $[-6.273, -2.026]$, Holm p = 0.00689.

At sensor-noise level 0.2, the difference between full-reward degradation and raw-reward degradation is -0.3705 $[-0.6774, -0.1258]$ in D2, Holm p = 0.000488, and 7.896 [4.937, 10.958] in D3, Holm p = 0.001282. These are differences in degradation from each native endpoint, not direct method-cost differences. Resource signals carry domain-dependent noise costs.

A separate four-arm reward study compares raw reward, potential only, squared penalty only and their combination. The D3 squared-penalty mean cost effect is -22.455 $[-26.706, -18.095]$, domain-wise Holm p = 0.0000114. The potential effect interval crosses zero. In D2, the squared penalty reduces return by 0.360, while cost improvement does not survive that family correction. This panel concerns the training reward, not the retention factorial.

D.4 Feedback Bundles and Target-Environment Retraining

The D1/D2 feedback intervention fixes prespecified resource-dependent terms at reference values rather than randomizing individual edges. D1 changes observation, consumption, replenishment and effectiveness terms. D2 changes the consumption denominator, saturated solar recovery and task efficiency. An analogous D3 intervention would remove thermal inertia and was not run as an allegedly equivalent control.

Low, native and high recovery settings alter the D1 recovery coefficient, D2 solar charging rate and D3 maximum cooling conductance. Cooling also changes energy consumption, so D3 is not a pure abstract recovery-speed intervention. Frozen-policy counterfactuals, target retraining and component difference-in-differences estimate different objects.

The prespecified D2 feedback difference-in-differences is -1.200 $[-1.977, -0.508]$ for return, Holm p = 0.00423, and -7.940 $[-8.821, -6.985]$ for cost, Holm p = 0.00000381. D2 high-recovery cost interaction is -0.0712 $[-0.1261, -0.0261]$, Holm p = 0.00305. Corresponding D1/D3 families have no corrected confirmatory finding.

These supplements contain 320 target-environment retrainings, reuse 120 native checkpoints and evaluate 1,400 units over 70,000 episodes. Frozen-policy edge cutting increases the two reward configurations' D1 costs by 0.270 and 0.289, and D2 costs by 0.817 and 0.293. D2 return also rises. Relative to direct transfer, retraining reduces D2 low-charging costs by 2.204 and 1.988, and D3 low-cooling ERC cost by 42.999. Extremely low recovery and episode duration are reported alongside these improvements.

A transition audit finds five disagreements between cost timing and final resource labels among 3,094,636 D1 transitions. Recalculation using final state does not change the main direction or conclusion. A floating-point-tolerance recheck alters one small p-value among sixteen contrasts: the D3 low-recovery cost-interaction Holm p becomes 0.132454 and remains nonsignificant. Frozen trajectories are not rewritten to erase these discrepancies.

D1 hard-failure rate is approximately 0.98 under extremely low recovery. Some D3 configurations survive longer and accumulate more cost, so duration and hard events are reported jointly. All 41,650 run-window checks of the event decomposition pass in the original 5,950 actor runs. There are 7,736 opportunities satisfying persistence and no recovery, including 1,691 in the core Lagrangian group.

D.5 Resource Intervention and Source Structure

Both arms execute action $(0.6, -1.0)$ at the intervention step. Counterfactual protective action $(0.6, 1.0)$ supplies only the two replacement temperatures. From step one onward, both follow open-loop action $(0.6, -0.5)$. When episode length L includes step zero, the subsequent endpoints are

$$C_{\text{post}} = \sum_{t=1}^{L-1} 0.99^{t-1} c_t, \qquad V_{\text{post}} = \sum_{t=1}^{L-1} \mathbf{1}\{c_t > 0\},$$
$$T_{\text{first}} = \inf\{t \geq 1 : c_t > 0\}, \qquad \inf \varnothing = +\infty. \tag{D1}$$

Both arms retain the original stopping rule and 288-step cap. Lengths differ in 28/120 pairs. No common exposure window is imposed afterwards. Every first-cost event is observed. Termination-induced exposure differences are part of the total trajectory effect, not a fixed-exposure incidence-rate effect.

Code/15 and Code/16 denote Azure code-service traces for May 15 and 16, 2024. Conv/17 and Conv/18 denote conversation-service traces for May 17 and 18. DLRM/01 and DLRM/21 identify the corresponding Alibaba GPU 2025 workload days. NY is New York–LaGuardia. VA is Leesburg Executive. PA is Port Angeles–Fairchild, not Pennsylvania. SEA is Seattle–Tacoma. TUS is Tucson–Davis–Monthan.

Within-source intervals use 100,000 paired resamples of twelve months, preserving both arms and all three endpoints. The pooled estimate averages within each source and then equally across ten sources. Source-clustered intervals also use 100,000 resamples. DLRM day_01/VA has zero mean positive-cost-duration difference because monthly signs cancel. day_21/NY has $-1/6$ step. Zero means retain their error bars.

With the prespecified source unit, maxT-adjusted p-values are 0.001953 for cost, 0.007813 for positive-cost duration and 0.035156 for detection delay. Aggregating by six workload days gives Holm p-values 0.09375, 0.125 and 0.09375. Aggregating by five weather sites gives 0.1875 for all three. The ten sources are not ten wholly independent datasets. Five new sources follow metadata rules, while the existing five remain in the cumulative estimate.

For the five new sources alone, cost difference is 0.8485 [0.3797, 1.2758], with 5/5 in the expected direction. Positive-cost duration is 2.4333 [0.1833, 5.0833], with 3/5. Detection delay is -5.0833 $[-11.0667, -1.5500]$, with 5/5. Corresponding maxT p-values are 0.0625, 0.25 and 0.25. Directions and magnitudes support the effect, with lower power than the cumulative ten-source study.

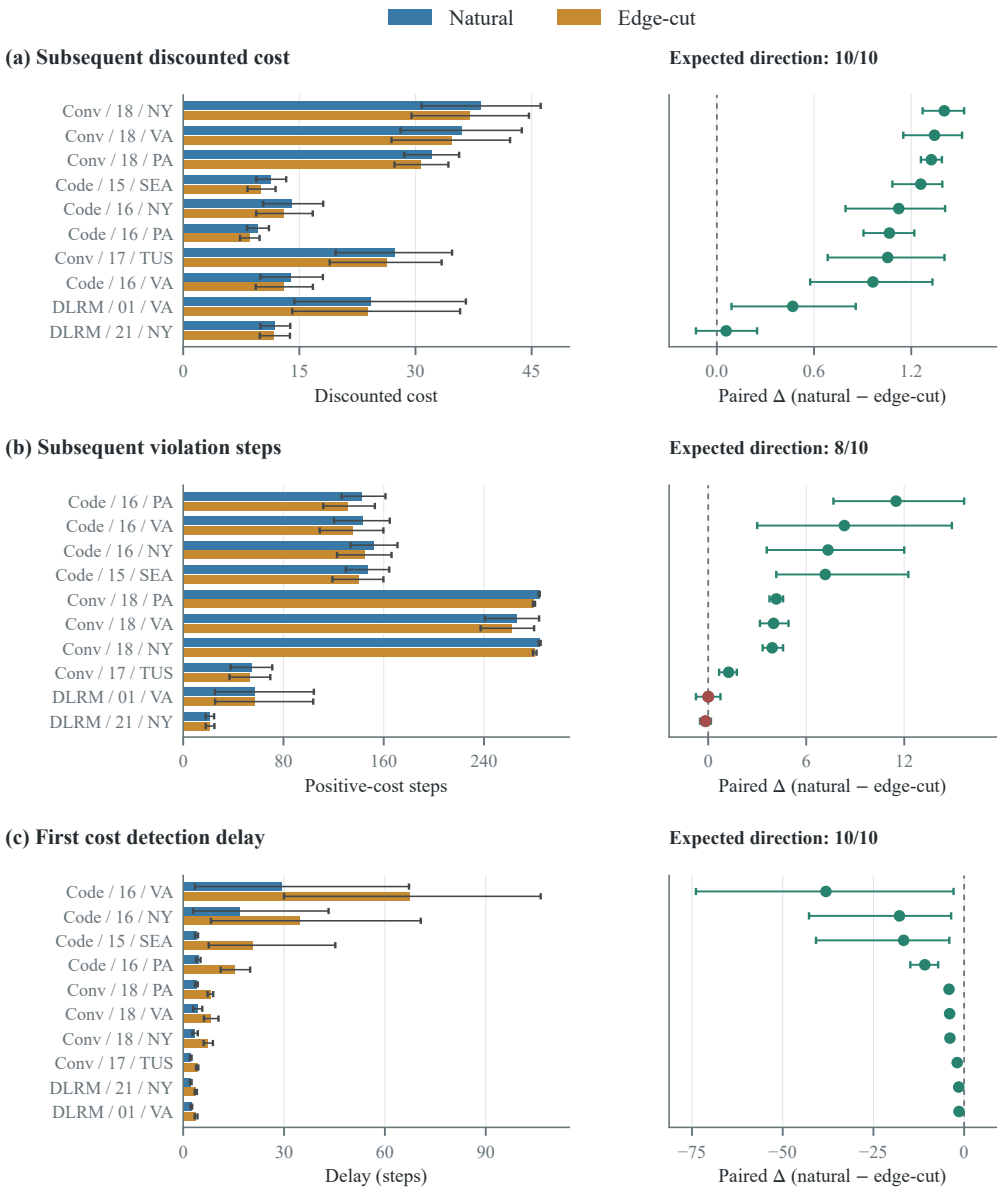

Figure 4: Paired thermal-resource interventions with the same current action and zero current cost. Subsequent cost, positive-cost duration and first detection differ under arm-specific termination. Intervals preserve paired months within sources; pooled intervals resample the ten sources. Appendix D.5 gives the intervention and source definitions.

Two earlier designs had different identification outcomes. A generative resource-free control failed its holdout-matching gate. A natural-resource-tape control passed 96,768 transition comparisons over fourteen policies and 24 trajectories. A 4,096-step engineering trial then reproduced identical learning paths under common initialization, randomness and resource tape. With identical cost inputs the standard dual paths also coincide. This pathwise equivalence differs from the generative control's empirical matching failure. Figure 4 obtains an actual future-risk contrast by changing the next thermal state while retaining current cost and action.

D.6 Recovery Actions and Four-Arm Reward Controls

Episode resets do not count as trajectory recovery. In D1, 806 of 818 backbone episodes entering risk subsequently complete five consecutive safe steps, conditional recovery rate 98.53%. 54 later collapse again. The D2 formal backbone panel has no risk entry, so its conditional recovery rate is undefined. Because visited damaged states and remaining horizons depend on policy, these rates are descriptive. Paired interventions identify local action effects separately.

Action controls start from 5,891 D1, 172 D2 and 5,477 D3 states, drawn from 20, 13 and 20 training seeds. D1 compares recovery against deep defense, retaining monitoring as a third arm. D2 compares low and high thrust at zero climb. D3 compares cooling with service fixed. Estimates average within source seed, then equally across seeds. First-step resource gains favor recovery by 0.10358 [0.10343, 0.10382], 0.08251 [0.08161, 0.08335] and 0.12766 [0.12740, 0.12792], respectively, in normalized resource/margin units.

The D2/D3 four-arm reward study comprises 160 new 400,000-step runs and 8,000 evaluation episodes. D3 costs for raw, potential-only, squared-penalty-only and full rewards are 28.1084, 32.2519, 8.8698 and 6.5804. The squared-penalty main effect is -22.4550 $[-26.7058, -18.0949]$. The potential effect is 0.9270 $[-1.6226, 3.4841]$. Cost interaction is -6.4328 $[-10.6940, -2.3972]$, with mean sign-flip Holm p = 0.03780 but directional sign-test p = 0.1153. Effect magnitude and sign consistency differ. The later weight sweep finds no universal three-domain synergy, which is compatible with this fixed-weight local interaction.

D2's squared-penalty return effect is -0.3597 $[-0.5370, -0.1967]$, Holm p = 0.003010. Its cost effect is -0.005157 $[-0.009879, -0.001234]$, Holm p = 0.1782. The new full configuration has 4/1,000 soft-violation episodes, unlike the historical zero-event panel. D3 five-step local recovery is 698/967 for full reward and 860/1,000 for squared penalty alone. Policy-dependent risk entry prevents interpreting this as a randomized causal effect of the potential on post-entry recovery.

D.7 Frozen Policies across Trace Sources

The initial D3 source study reuses 140 frozen models and evaluates 14,000 episodes on original and balanced 50-trajectory panels. The balanced panel covers 31 days, three weather sites and twelve months. ERC and Vanilla returns are 448.8924 and 449.3690, difference -0.4766 $[-2.3476, 1.4450]$. Costs are 3.5449 and 20.2297, difference -16.6848 $[-21.7336, -11.8168]$, with Holm p = 0.000103 in the 24-test family. The original-panel cost difference -16.6065 $[-21.9289, -11.4732]$ has the same direction.

The initial comparison covered the backbone and other methods. A later evaluation adds the actual SG outputs from seeds 200–219, aligned by checkpoint identity. Its 2,000 episodes bring the eight-method total to 16,000, retaining original and balanced panels separately. Table D1 lists return, cost and feasible seeds for every method. CSPO is absent from this source panel.

Table D1: Eight-method D3 frozen-output comparison across original and balanced trace panels.

Source panel	Method	Return	Cost	Feasible
Balanced 50	CRPO	434.645	1.205	14/20
Balanced 50	ERC-PPO	448.462	3.860	5/20
Balanced 50	ERC-TRPO	448.892	3.545	7/20
Balanced 50	ERC-TRPO-SG	437.782	2.431	12/20
Balanced 50	Lagrangian-TRPO	445.822	0.354	18/20
Balanced 50	SCPO	401.068	0.940	18/20
Balanced 50	TRC	444.973	2.436	7/20
Balanced 50	Vanilla-TRPO	449.369	20.230	0/20
Original 50	CRPO	450.075	1.627	14/20
Original 50	ERC-PPO	467.813	4.003	5/20
Original 50	ERC-TRPO	467.915	3.714	7/20
Original 50	ERC-TRPO-SG	455.060	3.488	11/20

Table D1 (continued)

Source panel	Method	Return	Cost	Feasible
Original 50	Lagrangian-TRPO	464.392	0.329	18/20
Original 50	SCPO	412.831	1.254	18/20
Original 50	TRC	463.786	2.701	7/20
Original 50	Vanilla-TRPO	467.966	20.320	0/20

On the balanced panel, SG-minus-backbone return is -11.110479 $[-18.898420, -5.013570]$, Holm p = 0.000423431. Cost is -1.114266 $[-3.786581, 1.967112]$, Holm p = 0.483576. Feasible seeds are 12/20 for SG and 7/20 for ERC, versus 18/20 for both Lagrangian-TRPO and SCPO. The cost direction and return loss are retained together. Feasible counts do not substitute for the mean-cost test.

D.8 D2 Action Efficiency and Value Estimation

This frozen-policy diagnosis reuses the twenty SG terminal checkpoints from seeds 200–219 without updating the network or environment. Each runs its original deterministic policy, zero-climb intervention and stochastic policy on the same fifty exogenous scenarios 2000–2049. A heuristic is evaluated separately on those scenarios. Of 3,050 episodes, 1,000 exactly replay the deterministic main panel. The new post-hoc diagnosis neither selects nor replaces main-table policies.

Observations contain resource, remaining distance, wind, $\sin\theta$, $\cos\theta$, speed and payload. The rule uses the same observations: thrust 0.15 when resource exceeds 0.45 and sunlight exceeds 0.20, otherwise zero, with zero climb throughout. Its extra structure is the prescribed control law. SG's mean daytime/nighttime thrust is 0.0610/0.0371, versus 0.1251/0 for the rule, indicating different energy allocation and action use.

Table D2: Frozen D2 policy diagnosis. Costs are discounted episode costs; aggregation is first within training seed.

Configuration	Policies × scenarios	Raw return	Discounted cost	Mean absolute climb
SG deterministic	20 × 50	5.207	0.000000	0.0338
SG zero-climb intervention	20 × 50	5.831	0.000000	0.0000
SG stochastic	20 × 50	5.070	0.005042	0.0588
Heuristic	1 × 50	6.164	0.000000	0.0000

Original SG has mean absolute climb 0.0338 and speed 0.3372, versus rule-controller speed 0.4111. Zero climb raises SG speed to 0.3619, lowers resource consumption from 1.8690 to 1.8050, raises mean resource from 0.7147 to 0.7220 and improves return by 0.6249. Subsequent states and thrust respond naturally, so this is a closed-loop action-replacement effect, not a one-step direct effect or proof of zero-climb optimality. Intervals use 100,000 paired seed resamples conditional on the fifty scenarios. The 65.2% gap closure is descriptive.

At genuine termination and the 200-step truncation, the environment zeros next-state potential and GAE masks next-value bootstrapping. All 3,050 replays satisfy

$$\sum_{t=0}^{H-1} \gamma^t [\gamma \Phi_{t+1} - \Phi_t] = -\Phi_0 = -0.64,$$

with numerical error below 10^{-10}. No residual potential or unmasked terminal bootstrap is found. The observation includes phase but not explicit remaining horizon.

The 1,000 stochastic 200-step episodes also record the frozen reward critic and the subsequent discounted shaped Monte Carlo return. Seed-equal mean prediction errors at steps 0–19, 80–119 and 180–199 are -0.3363, -0.0223 and 1.2054: terminal error is predominantly overestimation.

Seed-level RMSE averages are 0.4154, 0.4541 and 1.3628. Day/night mean squared errors are 0.5179 and 0.4973, without the same separation. The pattern is compatible with missing horizon information but also contains sampling and state-distribution effects. Value error is not identified as the causal source of suboptimal action use.

Training cost becomes sparse: mean completed-episode cost is 1.5130 in the first twenty updates and 0.005785 in the last twenty, with 85% zero records in the latter. Frozen stochastic replay has 26/1,000 positive-cost episodes and 133/200,000 positive-cost steps; deterministic output has 0/1,000 episodes. Six original shadow validations all reject, so the identity gate returns terminal. Completed-episode records exclude unfinished cross-update episodes and cannot reconstruct whole-batch zero-cost probability. The diagnosis separates exploration feedback from deterministic delivery and explains why low return of an already feasible output does not itself trigger safety repair.

D.9 Shadow and Retention Usage

Controller logs and selection audits for all sixty primary runs distinguish a satisfied trigger, recovery entry, shadow validation, paired acceptance and historical-policy output. Training can end after recovery entry but before another update, producing no new shadow. Table D3 keeps these events separate.

Table D3: Primary-panel component events, with twenty training seeds per domain.

Event	D1	D2	D3
Trigger condition holds	297	22	347
Recovery entries	103	6	134
Shadow validations	96	6	286
Paired acceptances	1	0	257
Seeds accepting repair	1	0	20
Seeds delivering history	1	0	17

D1's only accepted-repair seed is 204, which delivers a feasible historical candidate at actor step 182,272. The other nineteen return terminal. Every D2 seed has zero acceptance and returns terminal. On these 39 zero-acceptance paths, candidates all come from the unmodified main branch. Recorded checkpoints are therefore not evidence of realized selection gain.

Disabling all shadow acceptance while preserving the identity gate makes the output terminal by construction. This differs from unconditional checkpoint tracking and constraint-aware selection. Existing D1/D2 records identify the 39 zero-acceptance outputs, but D1 seed 204 changes trajectory after its first acceptance. Its no-shadow counterfactual cannot be reconstructed from the final checkpoint alone.

Archived seeds 600–619 contain twenty paired trainings per domain, with shadow attempts enabled or repair disabled while monitoring remains. D1 validates 78 candidates and D2 five; none is accepted. All forty terminal-parameter pairs have maximum absolute tensor difference zero. Applying the current identity gate returns the same policy in each pair under the same deterministic evaluation. This reuses eighty runs, independently of the 200–219 panel, and verifies main-path preservation under no acceptance rather than unconditional retention.

Recomputing the D3 factorial with the common selector gives shadow-added return difference 0.020032 $[-0.572804, 0.631760]$ and cost difference -0.072471 $[-0.272814, 0.107109]$, both Holm p = 0.990234. Both arms are feasible in 17/20 seeds. Both branches bypass the identity gate to avoid confounding shadow history with output rules. Online costs and the shadow-removal noninferiority test appear in Appendix H.3.

On the no-repair 600–619 paths, the frozen actor and pool are delivered either terminal-only or through the terminal-UCB/Pareto rule after bypassing the zero-acceptance branch. The latter removes both repair and the identity branch relative to full SG. Both outputs use the same archived deterministic scenarios 4400–4449, excluded from selection. All forty terminals are replayed, with eleven extra D1 historical outputs, totaling 2,550 policy-scenario units without retraining.

Table D4: Output-rule comparison on no-repair paths, seeds 600–619.

Domain	Output rule	Return	Cost	Feasible
D1	Terminal (identity gate)	2818.382	0.003175	20/20
D1	Identity gate bypassed	2753.805	0.003600	20/20
D2	Terminal (identity gate)	4.960	0.000000	20/20
D2	Identity gate bypassed	4.960	0.000000	20/20

D1 bypass-minus-terminal return is -64.577 $[-113.057, -20.641]$, with cost-rate difference 0.000425 [0.000211, 0.000657]. D2 policy identities and both endpoints are unchanged. Intervals use 100,000 paired seed resamples conditional on the archived fifty scenarios. This same-trajectory post-hoc comparison supports identity retention in this panel, separately from overall algorithm gains in new training.

The shadow–retention factorial contrasts are reported in Table 2 of the main text. Retention is applied through the common selector so that the shadow-count identity gate does not confound the component comparison. This appendix retains the paired component audit; the corresponding online-cost diagnostics and shadow-removal controls are reported in Appendix H.3.

E Nine Public Tasks and Adaptation Scope

Table E1: Historical training-tail endpoints after 1M interactions, five seeds per task and cost budget 25. Table E3 evaluates frozen terminal policies independently.

Task	CSPO return / cost	ERC-U return / cost	Feasible seeds C / U
AntVelocity-v1	3.789 / 1.125	1003.309 / 10.630	5/5; 5/5
CarButton1-v0	-1.461 / 22.985	-1.656 / 25.021	4/5; 3/5
CarGoal1-v0	3.112 / 18.362	4.479 / 18.071	5/5; 5/5
HalfCheetahVelocity-v1	569.765 / 9.063	1034.078 / 14.631	5/5; 5/5
HopperVelocity-v1	832.157 / 24.561	991.639 / 17.438	2/5; 4/5
HumanoidVelocity-v1	467.406 / 1.373	659.481 / 4.196	5/5; 5/5
PointButton1-v0	0.187 / 18.647	0.239 / 20.867	5/5; 5/5
PointGoal1-v0	1.807 / 15.688	4.943 / 21.331	5/5; 4/5
SwimmerVelocity-v1	36.131 / 15.036	33.029 / 27.318	5/5; 2/5

This appendix reports the public-task adaptation and its empirical analyses. Section E.1 covers detection gates and recovery dynamics. Section E.2 gives independent stochastic evaluation of frozen terminals. Section E.3 reports action semantics and recovery implementation. Section E.4 describes controlled repairs and last-update degradation. Section E.5 studies public-task selection on identical candidate pools. Section E.6 reports the audited matched 10M comparison.

Both methods use seeds 0–4, the same environment, 1M interaction budget, network size and cost limit 25. Learning rates and constraint parameters are method-specific. The configuration audit finds no mismatch in required common fields and 270 differences in informational algorithm fields. All ninety runs finish. The final twenty training epochs characterize late training, not the independent deterministic fifty-episode endpoint used in the three-domain study.

ERC-TRPO-U leaves source observations, dynamics, raw reward, cost and termination unchanged. It maintains an internal reserve $z \in [0, 1]$, initialized at one. At each epoch it reads mean logged episode cost $\overline{C}$, updates

$$z \leftarrow \mathrm{clip}\left(z + \frac{0.1(d - \overline{C})}{\max\{d, 10^{-8}\}}, 0, 1\right),$$

and sets penalty $\lambda_z = 2(1 - z)$. This reserve is constructed from cost feedback rather than a physical resource supplied by the environment.

Ordinary and recovery modes use advantages $(A_R - \lambda_z A_C)/(1 + \lambda_z)$ and $-A_C$, respectively, both with TRPO. From epoch eight, the one-sided 95% episode-cost t upper bound triggers recovery above 25 and exits at 0.8 times budget. Each epoch collects 20,000 steps. Observations are normalized; rewards and costs are not. Both GAEs use discount 0.99 and parameter 0.95 and are separately standardized. Actor and critics use two 64-unit tanh layers, with maximum KL 0.01.

There are no same-batch competing candidates, independent validation panel or terminal Pareto retention, hence the separate ERC-TRPO-U name.

In the historical tail, SwimmerVelocity exceeds mean budget and has 2/5 feasible ERC-U seeds; CarButton1 has 3/5. CSPO has 5/5 and 4/5 on these tasks. On HopperVelocity, CSPO's method mean is below 25 despite only 2/5 feasible seeds, illustrating the distinction between method and seed feasibility.

An earlier eighteen-run adaptation study did not progress to longer-budget confirmation. Its archived summary is retained, although some original logs could not be relocated. Eighteen later 1M controlled repairs use new development seeds and are reported separately in Section E.4, not pooled with those historical runs or the formal nine-task panel.

The historical summaries also compare lower exit thresholds and smaller recovery steps. Seeds 127 and 128 at 400,000 steps have unchanged main endpoints with a lower exit threshold. A smaller recovery trust region raises CarButton cost from 46.283 to 55.377, with 0/2 feasible seeds in both configurations. Because the eighteen original logs could not be relocated, these are summary-level records, not newly audited trajectories.

Historical tail costs for ERC-U are 27.318 on SwimmerVelocity and 25.021 on CarButton1. With five paired seeds, the smallest two-sided exact sign-flip p-value is 0.0625; no individual-task exact 0.05 significance is established.

E.1 Detection Gates and Recovery Dynamics

The audit reconstructs 2,250 epochs from all 45 formal ERC-U runs. Costs use a rolling window of up to 100 completed episodes, carried across epochs. UCB entry starts at epoch eight and exit requires UCB at most 20. Reconstructed gates match every log, and reserve recursion agrees within floating-point tolerance. After warm-up, no task has a recorded mean above 25 with recovery off. This version uses UCB hysteresis, not a recovery-trend estimator.

Table E2: Recovery dynamics of the two historical over-budget tasks. Metrics are first aggregated within seed, then equally across five seeds.

Metric	**CarButton1**	**SwimmerVelocity**
Last-20 mean cost	25.021	27.318
Last-20 recovery fraction	96%	52%
Total mode switches	2.2	7.0
Longest recovery segment	34.8	5.8
Last-20 mean UCB margin	9.335	6.666
Last-20 feasible mean / failing UCB	38%	7%

All five CarButton seeds enter recovery at epoch eight. Seeds 1 and 3, which finish over budget, remain in recovery for 43 consecutive updates and have tail costs 32.1305 and 25.8760. Their failures require neither repeated exits nor an inactive recovery gate. Swimmer averages 7.0 switches and a longest recovery segment of 5.8 updates. Its over-budget seeds 0, 2 and 3 have tail costs 29.3150, 30.4775 and 39.8735. These are sustained-but-insufficient and intermittent-response patterns, respectively.

Recovery optimizes negative cost advantage, with TRPO line search checking surrogate improvement and KL. All recovery updates with a following log are accepted, yet the next rolling cost remains above 25 in seed-equal proportions 63.5% and 64.3% for CarButton and Swimmer. Local surrogate acceptance is therefore distinct from budget compliance. Row i records pre-update samples; its mode

controls update i. The next window may still mix older-policy episodes, so this is a descriptive post-response record, not an independently evaluated causal policy effect.

Mean tail UCB-minus-mean margins are 9.335 for CarButton and 6.666 for Swimmer. At sample size 100 and the original t level, these imply mean within-window episode-cost standard deviations about 56.224 and 40.149. In CarButton, 38% of late records have feasible means but over-budget UCBs. The window mixes episode randomness and policy change, and archived episode data are insufficient to separate them; the dispersion is not identified as environmental noise alone.

E.2 Independent Stochastic Evaluation of Frozen Terminals

All ninety epoch-fifty checkpoints use training seeds 0–4 and fifty episodes each, with environment seeds 700000–700049 and action seeds 800000–800049. Policy parameters and observation mean/standard deviation are frozen; normalization is not updated and no historical model is selected. Evaluation retains source termination and the 1,000-step cap. Return and cost are undiscounted episode sums, with budget 25. Aggregation is first by seed, then equally across seeds; the exact two-sided five-seed p-value floor is 0.0625.

Table E3: Independent frozen stochastic-policy evaluation. Task-mean feasibility and feasible seed counts are separate.

Task	CSPO return / cost	ERC-U return / cost	Feasible seeds C / U
AntVelocity	27.098 / 1.272	1574.171 / 31.640	5/5; 1/5
CarButton1	-1.448 / 15.792	-1.585 / 31.700	5/5; 3/5
CarGoal1	4.253 / 17.488	6.309 / 17.792	3/5; 3/5
HalfCheetahVelocity	837.214 / 24.464	1276.064 / 18.824	2/5; 4/5
HopperVelocity	821.278 / 32.736	1064.739 / 4.836	3/5; 5/5
HumanoidVelocity	515.101 / 2.316	832.136 / 6.068	5/5; 5/5
PointButton1	0.710 / 16.452	1.306 / 17.552	4/5; 5/5
PointGoal1	2.796 / 16.012	6.076 / 22.772	5/5; 3/5
SwimmerVelocity	50.823 / 16.352	58.161 / 85.412	4/5; 3/5

ERC-U is both mean-feasible and higher-return than CSPO on 6/9 tasks. Its Ant, CarButton and Swimmer costs are 31.640, 31.700 and 85.412. The historical 7/9 count uses the last twenty rolling training windows, mixing policy changes and online observation normalization. It is not the same frozen output. The difference cannot be attributed uniquely to one factor or interpreted as seven stable task wins.

E.3 Action Semantics and Recovery Implementation

Table E4: Deterministic mean-action evaluation on the same fifty scenarios, reported separately from Table E3.

Task	CSPO return / cost	ERC-U return / cost	Feasible seeds C / U
AntVelocity	789.360 / 1.892	2302.989 / 36.960	5/5; 2/5
CarButton1	-12.322 / 23.916	-6.664 / 59.788	3/5; 1/5
CarGoal1	-5.967 / 13.960	6.621 / 22.692	4/5; 4/5
HalfCheetahVelocity	1432.769 / 146.692	1684.656 / 44.532	0/5; 2/5
HopperVelocity	971.645 / 8.264	933.269 / 0.000	4/5; 5/5
HumanoidVelocity	499.781 / 0.956	1335.413 / 3.528	5/5; 5/5
PointButton1	-1.184 / 26.008	0.049 / 27.984	2/5; 2/5
PointGoal1	1.957 / 26.204	4.877 / 34.316	3/5; 1/5
SwimmerVelocity	31.882 / 0.000	31.762 / 8.348	5/5; 4/5

Under mean actions, only CarGoal and Humanoid jointly favor ERC-U in mean feasibility and return. Stochastic policies and mean-action policies are therefore different delivery semantics. Together the two modes use 9,000 episodes and 7,687,736 evaluation interactions, including 3,706,234 for

stochastic evaluation. All ninety policies and normalization states stay fixed; configuration and model audits pass.

The recovery objective is negative cost advantage directly, so changing the ordinary penalty under a fixed mode and batch leaves it unchanged. Historical Train/KL also includes internal Fisher-vector-product zeros and cannot represent actual update KL. Later development records actual pre/post-update KL, cost-surrogate change, newly completed episodes and rolling cost separately, distinguishing detection, accepted recovery and budget compliance.

E.4 Controlled Repairs and Last-Update Degradation

Development seeds 8301–8303 cover Ant, CarButton and Swimmer, with original and one repaired configuration per task: eighteen 1M-step runs. CarButton replaces cost credit on complete 1,000-step paths with undiscounted realized cost-to-go minus the original cost baseline, leaving reward unchanged. Ant/Swimmer veto same-batch cost-surrogate-increasing proposals during five ordinary updates after recovery exit. Environment, network, original gate and budget stay fixed. No shadow, retention or resource-linked multiplier is added.

The expansion gate is fixed before running: repair must achieve 3/3 mean-feasible seeds, lower cost in at least two pairs and by at least 10% on average, limit return loss to 10% of the larger of absolute control return and one, keep online cost within 105% of control, and actually activate its mechanism. All three tasks fail the joint gate after all eighteen trainings finish; independent-seed confirmation is not started. Terminals use fifty stochastic episodes on scenarios 710000–710049, separate from the later selection study.

Table E5: Terminal evaluation of the 1M development repairs, three training seeds per configuration, without historical selection.

Task	Training variant	Return	Cost	Feasible	Online cost / step
Ant	Original	1355.591	6.007	3/3	0.011788
Ant	Post-exit guard	1249.922	7.660	3/3	0.010145
CarButton	Original	-1.598	24.027	2/3	0.035014
CarButton	Full-episode cost credit	-2.486	22.347	2/3	0.043360
Swimmer	Original	48.662	12.807	3/3	0.029390
Swimmer	Post-exit guard	42.541	120.620	2/3	0.035989

Ant online cost falls by 13.94%, but terminal cost rises and return falls. CarButton terminal cost falls by 6.99%, with feasibility still 2/3 and online cost rising 23.84%. Swimmer online cost rises by 22.45%; repaired seed 8303 has terminal cost 361.66. Mechanisms execute: Ant and Swimmer veto eleven and 24 proposals, and CarButton replaces 3,000 full-episode targets. These fixed configurations do not qualify for expansion based on isolated favorable seeds.

A post-failure diagnosis selects Swimmer repair seed 8303 and compares saved epochs 49/50 on new paired scenarios 730000–730049. Return is 45.911/94.623 and cost 11.100/327.240, with 0/50 versus 50/50 budget-exceeding episodes. Before update fifty, newly sampled mean cost is 16.05 and rolling UCB is 4.795; training then ends after an ordinary update, leaving no next batch to detect the changed policy. Snapshots include both actor and normalization changes. This single-seed post-hoc diagnosis identifies terminal degradation, not a unique cause or selector gain.

The study uses 18,000,000 training interactions, 881,118 terminal-evaluation interactions and an additional 80,000 implementation-check training steps. The epoch-49/50 diagnosis adds 100 episodes and 100,000 evaluation steps. All 51 snapshots per run are retained; paired initialization and configuration checks pass. Parallel wall time is approximately 96.3 minutes. Parent-process CPU excludes environment subprocesses and is not a complete algorithm speed comparison.

E.5 Public-Task Selection on Identical Candidate Pools

All eighteen development runs supply ten fixed snapshots at epochs 5, 10, . . . , 50, excluding the post-hoc-discovered epoch 49. Every candidate uses sixteen common stochastic validation scenarios 740000–740015 and action seeds 840000–840015. Ninety decisions are frozen before the forty distinct selected policies are each evaluated for fifty episodes on scenarios 750000–750049 and action seeds 850000–850049. Policy identity includes actor and observation normalization. Training seeds have already informed development; selection and evaluation panels are independent of one another.

The rules are terminal, minimum cost UCB, highest return among UCB-feasible candidates, original SG, and Pareto without the identity gate. Minimum-UCB ties favor higher validation return then later time; maximum-return ties favor lower UCB then later time. With no feasible candidate, maximum-return falls back to minimum UCB and records validation failure. Original SG returns terminal in all 18/18 runs because no shadow was accepted. The no-identity ablation retains the original pool and minimum-UCB-plus-one-budget risk band. Nominal 95% t scores are not post-selection certificates.

Table E6: Independent evaluation of AntVelocity-v1 same-pool selection. Original SG equals terminal entry by entry.

Training variant	Selector	Return	Cost	Feasible	Episode exceedance
Original	Feasible max-return	1369.442	5.827	3/3	2.7%
Original	Minimum UCB	1018.163	0.807	3/3	0.0%
Original	No-identity Pareto	1369.442	5.827	3/3	2.7%
Original	Terminal	1369.442	5.827	3/3	2.7%
Post-exit guard	Feasible max-return	1287.638	10.913	3/3	14.7%
Post-exit guard	Minimum UCB	725.692	0.373	3/3	0.0%
Post-exit guard	No-identity Pareto	1252.673	7.660	3/3	12.7%
Post-exit guard	Terminal	1252.673	7.660	3/3	12.7%

Table E7: Independent evaluation of CarButton1-v0 same-pool selection. Original SG equals terminal entry by entry.

Training variant	Selector	Return	Cost	Feasible	Episode exceedance
Original	Feasible max-return	-1.056	25.553	2/3	28.7%
Original	Minimum UCB	-1.152	23.387	2/3	26.0%
Original	No-identity Pareto	-1.152	23.387	2/3	26.0%
Original	Terminal	-1.488	20.600	2/3	22.0%
Full-episode cost credit	Feasible max-return	-2.048	37.173	1/3	28.7%
Full-episode cost credit	Minimum UCB	-2.226	35.427	1/3	27.3%
Full-episode cost credit	No-identity Pareto	-2.393	30.433	2/3	27.3%
Full-episode cost credit	Terminal	-2.393	30.433	2/3	27.3%

Table E8: Independent evaluation of SwimmerVelocity-v1 same-pool selection. Original SG equals terminal entry by entry.

Training variant	Selector	Return	Cost	Feasible	Episode exceedance
Original	Feasible max-return	61.175	15.267	3/3	14.7%
Original	Minimum UCB	22.586	0.300	3/3	0.0%
Original	No-identity Pareto	47.969	12.927	3/3	14.0%
Original	Terminal	47.969	12.927	3/3	14.0%
Post-exit guard	Feasible max-return	38.138	9.507	3/3	7.3%
Post-exit guard	Minimum UCB	10.757	0.067	3/3	0.0%
Post-exit guard	No-identity Pareto	15.767	0.267	3/3	0.0%
Post-exit guard	Terminal	41.913	95.267	2/3	33.3%

Simple feasible retention raises original Swimmer return and restores mean feasibility in the post-exit-guard branch. Minimum UCB and no-identity Pareto often reduce cost further at larger return loss. CarButton full-episode-credit seed 8301 has epoch-35 UCB 14.505 but independent mean cost 35.240. Both simple rules select it, lowering group feasibility from terminal 2/3 to 1/3. This is one candidate selected twice, not two independent failures.

For repaired Swimmer seed 8303, feasible maximum return and no-identity Pareto both select epoch forty, with independent return 17.333 and cost 0.360, versus terminal 95.771 and 285.360. They recover a low-risk output without attaining the separate post-hoc epoch-49 result. Not all candidates receive independent test evaluation; test maxima are therefore not reported as true oracle opportunity.

No retraining is added. Validation costs 2,880 episodes and 2,410,790 interactions, 13.39% of the prior 18M training budget. Ant configuration means are approximately 8.13%–8.23% and the other tasks use 16%. Independent evaluation adds 2,000 episodes and 1,951,797 interactions. A six-process two-stage queue takes approximately 26.3 minutes. Terminal-only delivery needs none of this monitoring. All ninety endpoints reproduce episode by episode, and the validation-decision-freeze-evaluation ordering passes audit. Frozen-rule independent training-seed confirmation remains unperformed.

E.6 Matched 10M Public-Benchmark Evaluation

We extend the public-task comparison to a matched 10M-interaction horizon on the same nine Safety-Gymnasium tasks and seeds 0–4. ERC-U carries the ERC resource-aware update into this public continuous-control setting; ERC-TRPO-SG additionally combines same-batch candidate competition, independent validation, and terminal retention in the resource-coupled domains. Each terminal checkpoint is evaluated as a stochastic policy for 50 episodes using environment seeds 700000–700049 and action seeds 800000–800049; observation normalizers are frozen. The protocol therefore evaluates 90 terminal policies and 4,500 independent episodes under cost budget 25.

Table E9: Matched 10M frozen stochastic terminal evaluation. Values are task means across five training seeds; feasibility in the last column counts seed-level mean-feasible terminal policies.

Task	CSPO		ERC-U		Feasible seeds
	Return	Cost	Return	Cost	C / U
AntVelocity-v1	3191.995	21.000	3080.735	11.820	5/5; 4/5
CarButton1-v0	0.765	14.448	-0.152	22.628	5/5; 4/5
CarGoal1-v0	25.238	27.792	18.111	25.928	2/5; 2/5
HalfCheetahVelocity-v1	2252.679	22.000	2399.885	7.436	4/5; 5/5
HopperVelocity-v1	1582.872	40.524	1601.343	18.316	2/5; 4/5
HumanoidVelocity-v1	6505.864	21.904	6326.863	16.040	3/5; 4/5
PointButton1-v0	10.916	22.748	4.188	18.048	4/5; 5/5
PointGoal1-v0	22.840	25.236	17.208	24.484	2/5; 3/5
SwimmerVelocity-v1	104.433	1.268	80.956	23.164	5/5; 2/5

At the task-mean level, ERC-U is feasible on 8/9 tasks versus 6/9 for CSPO and has lower mean cost on 7/9 tasks; at the seed level, the corresponding counts are 33/45 and 32/45. The two aggregation levels report complementary properties of the delivered policies. ERC-U improves both return and cost on HalfCheetahVelocity and HopperVelocity, while the remaining tasks map the task-specific safety–utility trade-offs.

The archive audit enumerates all 90 frozen-evaluation summaries and all 90 episode files, each with 50 episodes. Every summary reports status `pass`, frozen observation normalization, a stochastic policy, and the common budget 25. Independently, the training archive contains 90 passing audit records and one complete 500-epoch trajectory for every method–task–seed combination; two additional partial duplicate progress files are excluded by the 500-epoch completeness rule. Reaggregating these records reproduces every value in Table E9.

For the long-horizon trajectories, each task–epoch is first averaged across its five training seeds. Both methods have an over-budget epoch in every task. CSPO reaches first post-violation recovery earlier (median 0.20M interactions versus 0.38M for ERC-U), whereas the final 50 epochs contain 161 over-budget task-mean epochs for CSPO and 14 for ERC-U. The trajectory quantities characterize response and persistence over the completed 10M window; Table E9 reports the frozen-terminal delivery endpoint.

F Long Windows, Output Verification and Parameter Shifts

This appendix separates multiplier crossing, policy persistence and verified delivery. Section F.1 gives cumulative and terminal bounds with persistence exceptions. Sections F.2–F.4 cover fixed-support verification, parameter shifts and conditional long-run cost.

F.1 Policy Persistence with Exceptions

Use the multiplier recursion and first-passage time of Section 4. Write $J_k = J_C^{\mathrm{ev}}(\pi_k)$ for the protocol-specific population cost. Fix $\Delta > 0$, set $\eta = \Delta/2$, and start from $J_0 \geq d + \Delta$. Define

$$a_k = \mathbf{1}\{k < \tau_\nu\}, \qquad T_K = K \wedge \tau_\nu, \qquad s_k = a_k \min\{\Delta, [d + \Delta - J_k]_+\},$$

Then $L_K = \sum_{k<K} s_k$, $s_0 = 0$ and $0 \le s_k \le \Delta$. Each protocol retains its own population cost and budget units. The cumulative violation and truncated gap satisfy

$$\begin{aligned} O_K &= \sum_{k=0}^{K-1} [J_k - d]_+ \ge \Delta T_K - L_K, \\ L_K &= \sum_{k=0}^{K-1} s_k = \sum_{k=0}^{K-1} \mathbf{1}\{k < \tau_\nu\} \min\{\Delta, [d + \Delta - J_k]_+\}. \end{aligned} \tag{F0}$$

For the terminal statement, use half the initial margin as the terminal tolerance:

$$F_K = \{J_C^{\mathrm{ev}}(\pi_K) > d + \eta\} \cap \{\nu_K < \nu^\star\}. \tag{F14}$$

Proposition F.1 (Cumulative margin gap). On every path and fixed window K, Eq. (F0) holds without all-path B8 persistence or independent updates. The proof uses

$$[x - d]_+ \ge \Delta - \min\{\Delta, [d + \Delta - x]_+\}, \qquad x \ge 0. \tag{F1}$$

Proof. Checking $x \le d$, $d < x < d + \Delta$, and $x \ge d + \Delta$ proves the inequality. Multiply by a_k, sum, and retain nonnegative costs at other times to obtain Eq. (F0). All-path B8 gives zero pre-crossing gap. With exceptions, both their duration and magnitude enter the bound. □

The terminal event is separate. Splitting by whether the multiplier has ever crossed gives

$$\begin{aligned} \Pr(F_K) = \Pr(\tau_\nu > K) - \Pr(\tau_\nu > K, J_K \le d + \eta) \\ + \Pr(\tau_\nu \le K, \nu_K < \nu^\star, J_K > d + \eta). \end{aligned} \tag{F2}$$

The second term counts policies leaving the terminal violation region before multiplier recovery. The third counts an earlier crossing followed by a low terminal multiplier. This preserves the distinction between cumulative violation and a specified terminal event.

If the actual update process additionally satisfies a uniform gap contraction

$$\mathbb{E}[s_{k+1} \mid \mathcal{F}_k] \le (1 - \lambda) s_k + u a_k$$

with deterministic constants $0 < \lambda \le 1$ and $u \ge 0$, then

$$\mathbb{E}[O_K] \ge [\Delta - u/\lambda]_+ \mathbb{E}[T_K]. \tag{F3}$$

Proof. Subtract s_k from the contraction recurrence, sum, and use $s_0 = 0$ to obtain

$$\lambda \mathbb{E}[L_K] \le u \mathbb{E}[T_K] - \mathbb{E}[s_K].$$

Substitute into Eq. (F0) and use nonnegativity. The coefficient is informative when $u < \lambda\Delta$. This condition is on the true update given full history. Log reanalysis verifies the pathwise decomposition but does not certify these uniform dynamic constants. □

The 900 tabular paths cover U/F/S and three initially violating parameter cells, one hundred seeds per cell and 20,000 updates per path: 18,000,000 updates total. Opportunity-level persistence is 869,151/872,832, or 99.5783%. Under the original opportunity protocol 815 paths fail at least once; literal all-path persistence fails on 891. Normalized cumulative gaps have mean 0.371910, median 0.069585 and maximum 0.999950, giving retained lower-bound fractions 0.628090, 0.930415 and minimum 0.000050. These are post-hoc path summaries. All pathwise inequality checks pass.

At 20,000 updates, every tabular policy is feasible under its protocol population cost, although 307 multiplier paths never cross. There are 3,338 completed pre-crossing excursions outside the high-violation region, lasting up to 4,317 updates, plus 326 censored excursions. Duration and magnitude therefore require separate measurement.

Table F1: Neural-network development panels at 20,000 updates. The terminal event is Eq. (F14).

Round	Architecture	Paths	All-path B8 passes	No recovery	Terminal event
First	Separate	15	0	15	11
First	Shared	15	0	15	7
Second	Separate	10	1	6	5
Second	Shared	10	0	7	1

The two rounds' 0/30 and 1/20 all-path successes combine shared and separate architectures. Shared networks alone have 0/15 and 0/10 successes, with normalized mean gaps 0.495359 and 0.384688. Explicit freezing or projected-persistence positive controls pass on all 120 paths. Their purpose is to construct the diagnosed infeasible region, not to demonstrate successful safety learning.

F.1.1 Cumulative and Terminal Bounds under Return Conditions

To distinguish the resource b_t, write the outside-region indicator as

$$\chi_k = a_k \, \mathbf{1}\{J_k < d + \Delta\}.$$

Suppose fixed $\varepsilon \in [0, 1]$ and $\kappa \in (0, 1]$ bound full-history conditional transitions. When $a_k = 1$ and $\chi_k = 0$, the next policy leaves the high-violation region with probability at most ε. When $\chi_k = 1$, it returns with probability at least κ. Set $\zeta = \varepsilon + \kappa$. Here "return" is to the diagnostic high-violation region, opposite to safety recovery.

$$\mathbb{E}[\chi_{k+1} | \mathcal{F}_k] \le \varepsilon a_k + (1 - \zeta)\chi_k \tag{F3a}$$

Proposition F.2 (Conditional bounds with repeated exceptions). For fixed $K \ge 1$ and $\alpha \in (0, 1)$, these conditions imply

$$\mathbb{E}[O_K] \ge \frac{\Delta\kappa}{\zeta}\mathbb{E}[T_K] \tag{F3b}$$

$$\Pr\left(O_K \ge \frac{\Delta}{\zeta}\left[\kappa T_K - \sqrt{\frac{K}{2}\log\frac{1}{\alpha}}\right]_+\right) \ge 1 - \alpha \tag{F3c}$$

Proof. Let $N_K^\chi = \sum_{k<K} \chi_k$ and define the martingale

$$M_K = \sum_{k<K} \big(\chi_{k+1} - \mathbb{E}[\chi_{k+1} \mid \mathcal{F}_k]\big).$$

Summing Eq. (F3a) gives $\zeta N_K^\chi \le \varepsilon T_K - \chi_K + M_K$, with $\chi_0 = 0$. Combine with $O_K \ge \Delta(T_K - N_K^\chi)$ and take expectations for Eq. (F3b). Each martingale difference has conditional support width at most one. Conditional Hoeffding gives $\Pr(M_K \ge u) \le e^{-2u^2/K}$, proving Eq. (F3c). First multiplier crossing zeros the indicator, so actor changes need not be independent of crossing. □

A terminal bound also controls $\mathbb{E}[\chi_K]$. Define $q_0 = 0$. For $0 < \zeta \le 1$, take

$$q_K = \frac{\varepsilon}{\zeta}[1 - (1 - \zeta)^K].$$

For $1 < \zeta \le 2$ and $K \ge 1$, take $q_K = \varepsilon$. Then

$$\Pr(F_K) \ge [\Pr(\tau_\nu > K) - q_K]_+ \tag{F3d}$$

Proof. In the first case, take expectations in Eq. (F3a) and iterate its order-preserving recurrence. In the second, the coefficient is negative, so bound that term by zero rather than incorrectly iterating it as order-preserving. Since $\eta < \Delta$, event $\{\tau_\nu > K, J_K \le d + \eta\}$ is contained in $\{\chi_K = 1\}$. Eq. (F2) completes the argument. The return-probability condition also bounds the tail of an outside-region segment by $(1-\kappa)^\ell$. Exit includes return to high violation or multiplier recovery. Prespecified error allocation extends fixed-window statements to multiple windows. □

F.1.2 Concentration and Terminal Bounds from Gap Contraction

The contraction assumption in Eq. (F3) also gives the following bounds for fixed $0 < \eta < \Delta$ and K:

$$\Pr(F_K) \ge \left[\Pr(\tau_\nu > K) - \frac{u[1-(1-\lambda)^K]}{\lambda(\Delta-\eta)}\right]_+ \tag{F3e}$$

$$\Pr\left(O_K \ge \left[\left(\Delta - \frac{u}{\lambda}\right) T_K - \frac{\Delta}{\lambda}\sqrt{\frac{K}{2}\log\frac{1}{\alpha}}\right]_+\right) \ge 1-\alpha \tag{F3f}$$

Proof. The recurrence and $a_k \le 1$ give $\mathbb{E}[s_K] \le (u/\lambda)[1-(1-\lambda)^K]$. On the event $\{\tau_\nu > K,\ J_K \le d+\eta\}$, we have $s_K \ge \Delta - \eta$. Markov's inequality and Eq. (F2) yield Eq. (F3e). Summing the gap martingale differences M_K^s gives $\lambda L_K \le uT_K - s_K + M_K^s$. Their conditional support width is at most Δ, hence with probability at least $1-\alpha$,

$$M_K^s \le \Delta\sqrt{(K/2)\log(1/\alpha)}.$$

Combining with Eq. (F0) proves Eq. (F3f). A nontrivial expected coefficient requires $u < \lambda\Delta$. □

These extensions are conditional mathematical results. Uniform exit, return and contraction constants for actual networks remain uncertified. Every pathwise decomposition passes on the 900 tabular paths, but completed outside-region segments last up to 4,317 updates and 326 are censored. Opportunity-level success mixes histories and excludes already-outside updates, so it cannot replace the required conditional probabilities. High opportunity success and eventual self-correction can coexist.

F.2 Verification on Fixed Evaluation Support

The objects are actual SG outputs from seeds 200–219, on the fifty fixed formal evaluation scenarios. Let $\mathcal{Z} = \{z_1, \dots, z_{50}\}$ include each evaluator's full random seed. A frozen policy then produces deterministic scenario cost. D2/D3 use

$$J_{\mathcal{Z}}(\pi) = \frac{1}{50}\sum_{j=1}^{50} X(\pi, z_j) = \mathbb{E}_{Z\sim\mathrm{Unif}(\mathcal{Z})}[X(\pi, Z)]. \tag{F4}$$

where X is finite-episode discounted cost. In D1, let $X_1(\pi, z_j)$ be undiscounted episode cost and $L(\pi, z_j) > 0$ its length, with budget $d_1 = 0.005$. Its cost-rate target is

$$J_{1,\mathcal{Z}}(\pi) = \frac{\sum_{j=1}^{50} X_1(\pi, z_j)}{\sum_{j=1}^{50} L(\pi, z_j)} \le d_1. \tag{F5}$$

Positive total duration makes this equivalent to

$$\frac{1}{50}\sum_{j=1}^{50}\left[X_1(\pi, z_j) - d_1 L(\pi, z_j)\right] \le 0.$$

Every support point has been evaluated, yielding deterministic verification on the stated uniform finite support. This reuses endpoint evaluation, adds no independent experiment and does not reselect policies. Unknown-distribution sampling certification is separate.

Table F2: Verification on the fixed fifty-scenario support; D1 retains cost-rate units.

Domain	Evaluation target	Range over 20 policies	Budget	Passed
D1	Cost rate	0.002636 to 0.004493	0.005	20/20
D2	Discounted cost	0.000000 to 0.000000	0.0005	20/20
D3	Discounted cost	0.000025 to 0.794452	1.0	20/20

D1's mean policy cost rate is 0.003440. Every D2 policy has zero support cost. D3's maximum is 0.7945. All meet budget. The check links the gate, retained candidate and final evaluation to the same output. A separate distribution-free sampling certificate does not reach budget with the available samples and prior tail bounds and is not pooled with these passes.

An enlarged-sample attempt adds two hundred records for each of twenty D1 and twenty D2 outputs, totaling 8,000, but still obtains no certificate with the fixed mixture-e-process and Hoeffding-type procedures. Cost-structure audits give D1 margin variable $Y \in [-0.9, 44.1]$ and D2 episode envelope $(1 - 0.99^{200})/0.01$. D3 can overshoot on its terminating transition, so instantaneous cost is not bounded by one. The corrected episode envelope is 94.49072464 under ambient temperature at most $32^{\circ}\mathrm{C}$, fixed dynamics and legal actions.

Worst-case finite-support reweighting yields total-variation radii 0.02176 for D1's fifty-scenario rate, one for D2's fifty-scenario target, and 0.15889 for D3's full 72-scenario library, simultaneously over the original twenty outputs. Allowing off-support contamination with only global cost bounds gives conservative tolerable proportions approximately 0.2045%, 0.0005773% and 0.1961%. Parameter-OOD settings are not proved to belong to these cost-defined distribution sets.

F.3 Parameter Shifts and Checkpoint Alignment

Frozen-policy parameter-OOD evaluation uses seeds 200–219 for the original seven methods and 600–619 for SG. Each policy uses evaluation seeds 60000–60049, whose parameter draws are shared across methods. SG-versus-seven comparisons are descriptive across different training-seed panels.

D1 draws burst probability, burst consumption and single damage uniformly from [0.10, 0.14], [0.12, 0.16] and [0.024, 0.032]. D2 draws wind lower/upper endpoints from [0.70, 0.76]/[0.90, 0.95], payload endpoints from [0.80, 0.86]/[0.94, 1.00], and gust probability from [0.20, 0.30]. Remaining settings use the existing evaluation interface.

Outputs are aligned with SG's identity gate: none of seeds 600–619 accepts a shadow, so terminal policies are evaluated. D1 terminal checkpoints have been reloaded, and both D1/D2 SCPO interfaces restore their running maximum-cost state. The aligned replay covers all 8,000 D1 episodes and 1,000 D2 SCPO episodes. The remaining 7,000 D2 records are retained. D1's six unaffected methods reproduce all 6,000 existing episodes. Tables F3–F4 use these completed identity- and estimator-aligned results.

Table F3: Complete aligned D1 parameter-OOD endpoints, twenty seeds and fifty scenarios per seed.

Method	Return	Cost rate	Feasible
CRPO	908.652	0.028553	0/20
ERC-PPO	851.737	0.031079	0/20
ERC-TRPO	901.910	0.028514	0/20
ERC-TRPO-SG	901.537	0.028226	0/20
Lagrangian-TRPO	894.193	0.028942	0/20
SCPO	866.875	0.030517	0/20
TRC	206.488	0.023363	0/20
Vanilla-TRPO	912.936	0.028477	0/20

Table F4: Complete aligned D2 parameter-OOD endpoints, twenty seeds and fifty scenarios per seed.

Method	Return	Discounted cost	Feasible
CRPO	1.171	0.343310	8/20
ERC-PPO	1.629	0.291241	5/20
ERC-TRPO	2.186	0.348406	8/20
ERC-TRPO-SG	2.407	0.230136	9/20
Lagrangian-TRPO	0.895	0.020438	16/20
SCPO	1.625	0.076615	15/20
TRC	1.676	0.042760	15/20
Vanilla-TRPO	2.461	0.956202	2/20

D1 costs are recomputed per seed as undiscounted total cost divided by total steps. SG has rate 0.028226 and return 901.537, versus backbone 0.028514 and 901.910. All eight methods have 0/20 feasible seeds at budget 0.005. Primary-panel feasibility is not retained under this stronger parameter distribution.

D2 SG has return 2.407, discounted cost 0.230136 and 9/20 feasible seeds, versus backbone 2.186, 0.348406 and 8/20. Lagrangian-TRPO, SCPO and TRC achieve 16/20, 15/20 and 15/20 feasibility with lower mean returns than SG. Aligned SCPO return is 1.625 and cost 0.076615. The panel displays distinct safety-return trade-offs.

F.4 Conditional Long-Run Cost of Frozen Outputs

The original twenty D3 outputs from seeds 200–219 are enumerated over the full 72-scenario library: 1,440 new evaluation episodes. Mean cost is 0.254483 and maximum policy mean is 0.816257, with 20/20 below one. This differs from both the original fifty-scenario endpoint and the 31000–31019 matched-budget cohort.

Fix deterministic policies and the original episode reset. Let x_{iz} be discounted cost of policy i in scenario z. Let p be uniform weights on 72 points, and let

$$\mathcal{Q}_\rho = \{q : \mathrm{TV}(q, p) \le \rho\}.$$

If

$$\eta_{\mathrm{out}} = 1 - \max_{i,\, q \in \mathcal{Q}_\rho} \sum_z q_z x_{iz} > 0,$$

where q ranges over probability vectors on this support, put $a = \eta_{\mathrm{out}}/2$. A positive $\lambda_{\mathrm{e}} > 0$ then exists such that

$$\max_{i,\, q \in \mathcal{Q}_\rho} \sum_z q_z \exp\{\lambda_{\mathrm{e}}(x_{iz} - 1 + a)\} \le 1. \tag{F6}$$

Proof. Finite support uniformly bounds the second-order exponential remainder. The exponential moment has derivative at zero at most $-\eta_{\mathrm{out}}/2$, so sufficiently small positive λ_{e} satisfies Eq. (F6). Moving at most ρ probability mass from low to high values solves the worst linear reweighting. Exponentiation preserves order. Mean and exponential-moment reweighting are checked separately and verified by linear programming. Numerical bounds use double precision and inward margins. □

Proposition F.3 (Time-uniform frozen-output bound). Before each episode, choose a covered policy I_n predictably from the past, and let $X_n = x_{I_n, Z_n}$ be its realized discounted cost under scenario Z_n. If every history-conditional scenario law belongs to $\mathcal{Q}_\rho$, then

$$M_N = \exp\left\{ \lambda_{\mathrm{e}} \sum_{n=1}^{N} (X_n - 1 + a) \right\}$$

is a nonnegative supermartingale starting at one. For any $\delta \in (0, 1)$,

$$\Pr\left(\forall N \geq 1 : \frac{1}{N}\sum_{n=1}^{N} X_n \leq 1 - a + \frac{\log(1/\delta)}{\lambda_e N}\right) \geq 1 - \delta. \tag{F7}$$

Proof. Eq. (F6) bounds the conditional mean multiplicative increment by one. Bounded optional stopping at first crossing of $1/\delta$, followed by a limit in the truncation time, gives $\Pr(\sup_N M_N \geq 1/\delta) \leq \delta$. Take logarithms for Eq. (F7). Predictable switching is allowed for any given rule, not simultaneously over every post-hoc rule on the same sample. For twenty fixed-policy sequences, allocate $\delta_i = 0.05/20$ to each. □

Table F5: Sufficient deployment-prefix lengths for the original twenty D3 outputs, not certification sample counts.

Conditional scenario laws	Minimum expected margin	95% prefix, one sequence	Worst joint-95% prefix, 20 outputs
Uniform 72 scenarios	0.183743	33	40
Within-library TV ≤ 0.1	0.064394	177	353

Positive margin first establishes conditional expected-cost feasibility. Table F5 marks when the time-uniform cumulative-average upper bound drops below budget. A width-only bound is looser, requiring up to 4,979 episodes at TV = 0.1. Unknown scenarios, extra action noise, non-reset physical dynamics and continuing unverified policy updates require a new Eq. (F6) premise.

To include a finite training and monitoring prefix, let its charged complete episodes and nonempty terminal fragment have cumulative budget excess

$$S_0 = \sum_{k=1}^{K_0}(X_k - 1).$$

Conditional on that training history, Eq. (F7) gives

$$S_0 + \sum_{n=1}^{N}(X_n - 1) \leq S_0 + \frac{\log(1/\delta)}{\lambda_e} - aN. \tag{F8}$$

The same probability event covers every later prefix. Initial excess averages 3,428.603085 across the twenty training/monitoring ledgers. At TV = 0.1 with joint 95% coverage, the worst sufficient threshold for the combined average is 115,598 further deployment episodes. This retains the exploration excess when quantifying eventual cumulative recovery.

G Materials and Reporting Scope

The archive separately retains derivations, frozen protocols, seed endpoints, validation logs, transition replays and statistical summaries. The nine-method three-domain panel is reconstructed from thirty method–domain entries after removing three historical output-rule entries. Text, tables and Figure 2 use the same twenty-seed aggregation rather than inflating sample size with episode counts.

Appendix F covers long-window persistence, certification and distribution changes. Appendix H covers six selector/budget studies and selection-loss proofs. Appendix I covers shaping without retention and matched checkpoint opportunity. Appendix J covers multiplier development, new-seed confirmation and terminal diagnosis. Failed criteria, original outputs and engineering fixes remain in the record regardless of result direction.

Experimental inputs are simulator states, actions and exogenous traces, not language-model prompts. Reproduction materials include reward/cost composition, validation scenarios, stopping conditions,

checkpoints and complete interaction trajectories. The main text concentrates on the research questions, mechanisms and decisive results.

The experiments concern expected cost, local recovery and policy delivery in simulated control, not unverified physical deployment. Workload and weather traces drive simulation; there are no real-device interventions or human-subject experiments. The accompanying evidence retains fixed-rule controllers, failed tasks and negative component results.

H Controlled Comparisons of Selection, Budget and Release

This appendix collects the added selector, strong-baseline, stability, certification, monitoring-frequency and matched-budget studies. The training seed is the statistical unit. When rules select the same candidate, its evaluation is reused exactly. Candidate sets, protocols and test families are fixed for each study, and failed seeds are retained. The evaluation library had been used earlier in the project; these additions are not project-wide unseen blind tests.

H.1 Selection Rules on a Common Candidate Pool

SG's Pareto branch maximizes validation return over

$$\{\pi \in \mathcal{P} : U_C(\pi) \leq u_{\min} + d\}.$$

For a finite pool, this set is nonempty and its output is nondominated. The threshold is $u_{\min} + d$, not the absolute budget d. For example, consider exact costs 1.1, 0.8 and 0.2, returns 3, 2 and 1, budget one, and an accepted repair. A terminal cost of 1.1 fails the terminal gate but can still be selected by the Pareto branch. Reproducing this example with the original code separates relative risk anchoring from absolute-budget certification.

The D3 comparison fixes the twenty SG pools from seeds 31000–31019 and their original sixteen-scenario validation scores. It compares terminal, minimum UCB, highest return among UCB-feasible candidates, and original SG. Each pool contains at most the three actually retained representatives, deduplicated by parameter fingerprint. If no candidate is UCB-feasible, feasible max-return falls back to minimum UCB and records failure of the screen. Unsaved checkpoints are not reconstructed. Selection is locked before full 72-scenario evaluation.

Table H1: D3 common-pool comparison; twenty seeds per rule.

Output rule	Return	Cost	Feasible	Episode exceedance
Terminal	191.758874	2.841600	4/20	60.0000%
Minimum UCB	191.644849	0.244335	18/20	9.3750%
Feasible max-return	192.413042	0.497838	17/20	21.4583%
SG	192.045122	0.654706	16/20	25.9028%

SG uses eighteen Pareto exits and two UCB-feasible terminal exits. Five outputs differ from feasible max-return: three Pareto exits and two terminal-priority exits. SG minus terminal has cost difference -2.186894, 95% interval [-3.251311, -1.277010], Holm p = 0.000366. SG minus minimum UCB has cost difference 0.410371 [0.199532, 0.666279], Holm p = 0.001221. SG minus feasible max-return has return difference -0.367920 [-0.729964, -0.079135], Holm p = 0.25. These six comparisons form one test family; intervals are marginal paired-bootstrap intervals with 100,000 resamples.

H.2 Independent Common-Pool Confirmation in D1 and D2

For each domain, four rules are replayed on the retained pools of twenty existing seeds 600–619. Output fingerprints are frozen before evaluation on 100 new scenarios, seeds 935000–935099. These archives come from the paired shadow study, with the original SG rule applied to the retained pools; this is neither new SG training nor parameter OOD. D1 uses 401,408 training steps and D2 uses 400,000. These cohorts are distinct from D3's matched-interaction study.

Table H2: D1/D2 common-pool results on 100 new scenarios.

Domain	Rule	Return	Cost	Feasible
D1	Terminal / SG	2978.459070	0.003817906	20/20
D1	Minimum UCB	2909.974747	0.004244965	16/20
D1	Feasible max-return	2929.766651	0.004226232	16/20
D2	Terminal / SG	4.898469	0	20/20
D2	Minimum UCB	5.093360	0	20/20
D2	Feasible max-return	5.093360	0	20/20

Both domains retain terminal under original SG because no shadow is accepted. In D1, SG minus minimum UCB has cost difference -0.000427059, 95% interval [-0.000649721, -0.000216962], Holm p = 0.011169. Against feasible max-return, the difference is -0.000408326 [-0.000642907, -0.000185307], Holm p = 0.024170. Return differences are not established. In D2, the two simple rules select identical outputs, switching to history in 15/20 seeds and increasing return by 0.194890 [0.071583, 0.337013], Holm p = 0.024170. Twelve comparisons across both domains are corrected together; identical outputs are not independent evidence.

Episode budget-exceedance rates for the three distinct D1 outputs are 27.35%, 28.80% and 29.15%; all are zero in D2. The evaluation adds 7,500 episodes and 1,412,075 steps: 160 rule labels correspond to 75 distinct evaluated policies. Neither simple rule passes the cross-domain retraining gate: no domain may lose feasible seeds, mean return must retain at least 99% of SG, normalized mean cost must not rise, and at least one domain must improve strictly. This is a resource-allocation gate, not a statistical noninferiority test. The original algorithm is unchanged; no domain-specific mixture of rules is introduced.

H.3 Shadow Simplification and Matched-Budget Confirmation

Experimental ERC-TRPO-R disables shadow repair and removes the identity shortcut that depends on accepted shadows, retaining monitoring and terminal-UCB/Pareto delivery. This differs from disabling shadow while leaving that shortcut intact. In the earlier ten-seed pilot, SG has return 188.721913, cost 0.408995 and 9/10 feasibility; the backbone has 186.790264, 6.680084 and 2/10. This pilot is not pooled with the later independent twenty-seed confirmation, seeds 31000–31019.

All sixty new runs underlying Table 1 are complete, without seed replacement. For R minus SG, return differs by -0.085149, 95% interval [-0.779693, 0.556950], and cost by 0.130965 [-0.125110, 0.483905]. The prespecified joint noninferiority criterion permits return loss at most two and cost increase at most 0.1, assessed with one-sided 97.5% approximate bootstrap bounds. The return criterion passes; the cost criterion does not.

Table H3: Actual interactions and stage times under the common cap. Timing comes from six concurrent single-thread CPU processes on one machine.

Method	Training steps	Validation steps	Total acquisition	Stage time (s)
ERC-TRPO	400000	0	400000	351.61
ERC-TRPO-SG	252756.35	145554.05	398310.40	273.08
ERC-TRPO-R	311296	87453.50	398749.50	304.84

Every run stays within the 400,000-step cap; final evaluation uses another 72 episodes per seed. Episode exceedance rates are 25.9028% for SG and 22.7778% for R, although SG has lower mean cost. SG's four over-budget evaluation seeds are 31005, 31014, 31015 and 31016. Seeds 31014 and 31015 passed the selection-time UCB screen; the other two had no passing candidate. Full-library verification passes 16/20 for both SG and R, or 14/20 under TV ≤ 0.1; corresponding backbone counts are 4/20 and 1/20. All failed outputs remain in the summaries.

In twenty existing paired D3 branches with equal training volume, mean complete-training-episode cost is 3.388568 without shadow and 2.311695 with shadow: difference -1.076873, 95% interval [-1.281343, -0.880634]. Cumulative discounted cost including the final episode prefix is 4715.500813 versus 3216.123153. Both exploratory Holm p-values are approximately 0.000003815. This post-hoc training-endpoint analysis is separate from the unresolved independent final-delivery increment; average training cost still exceeds budget. The additional matched-budget SG/R comparison, online means 2.512508 and 3.544009, also includes differences in training volume.

Table H4: Independent reevaluation of earlier frozen training. D1/D2 use scenarios 934000–934099; D3 uses the 72-scenario library. This is distinct from Table H2.

Domain	Output	Return	Cost	Feasible
D1	SG	2904.489271	0.004362	17/20
D1	R	2837.042880	0.004683	13/20
D2	SG / R	4.983633	0.000008	20/20
D3	SG	192.286692	0.339212	19/20
D3	R	192.386698	0.418370	17/20

These remain reevaluations of existing seeds 600–619. In D1, R minus SG has return difference -67.446391, 95% interval [-118.622824, -20.479122], and cost difference 0.000321 [0.000084, 0.000607]; six-comparison Holm p-values are 0.093750 and 0.117188. D3's cost difference is 0.079158 [-0.095372, 0.275589]. D1/D2 terminal parameters match; D1 differs through historical selection, while D2 actually delivers identical policies. The original 4400–4449 evaluation remains in Table D4; all three panels are retained.

H.4 Equal Retention Opportunities for a Strong Baseline

CSPO is selected from the frozen D3 frontier, retaining its official optimizer and native 20,000-step batches; ERC uses native 2,048-step batches. Shared retention nodes are 40,000, 80,000, . . . , 320,000 and 358,528 training steps, each evaluated on the same sixteen scenarios. Both no-shadow retention arms bypass the identity shortcut and share terminal-UCB priority and Pareto selection. A separate CSPO arm trains for 400,000 steps without retention. The study contains sixty new runs.

Table H5: Strong-baseline retention control; twenty seeds and full 72-scenario evaluation.

Method / output	Return	Cost	Feasible
CSPO + retention	192.265133	0.011889	20/20
CSPO paired terminal	192.265133	0.011889	20/20
CSPO, 400k training	192.070727	0.163222	19/20
ERC + retention	191.744583	0.664720	16/20
ERC paired terminal	190.727708	5.594527	7/20

ERC retention minus CSPO retention has cost difference 0.652831, 95% interval [0.337356, 1.084718], Holm p = 0.000046, and return difference -0.520549 [-1.184393, 0.173810]. CSPO retention minus pure training has cost difference -0.151333 [-0.299602, -0.039004], Holm p = 0.043945. However, all twenty CSPO retention outputs are terminal, so this difference cannot be attributed to historical selection. Shared arms train for 358,528 steps; CSPO uses 41,472 validation steps and ERC averages 41,402.95, charging early termination as observed. Shared nodes truncate one CSPO batch and nine ERC batches; training volume and the adjusted learning-rate schedule are part of the comparison.

H.5 Routine Validation Frequency

The SG trigger, three-point trend, sixteen scenarios, paired shadow validation and output rule remain fixed. Only routine validation spacing changes from eight to four or sixteen updates. Each new

spacing adds twenty runs; spacing eight reuses matched-budget confirmation. Recovery may still add paired validation every update.

Table H6: Validation frequency under the common acquisition cap.

Spacing	Return	Cost	Feasible	Training steps	Validation steps
4	192.295574	0.488203	18/20	199036.20	198289.20
8	192.045122	0.654706	16/20	252756.35	145554.05
16	191.827207	0.670784	16/20	302674.30	96276.10

Relative to spacing eight, spacing four changes return by 0.250452, 95% interval [-0.401034, 0.841085], and cost by -0.166503 [-0.495239, 0.124607]. Spacing sixteen changes return by -0.217915 [-0.878438, 0.410418] and cost by 0.016078 [-0.213382, 0.262495]. All four Holm p-values are one. Episode exceedance rates are 17.7778%, 25.9028% and 24.0278%. Spacing sixteen reduces validation steps by about 34% with similar descriptive endpoints. No equivalence or noninferiority test was prespecified, so nonsignificance does not establish budget independence.

H.6 Independent Validation-Panel Stability

Twenty SG candidate pools, repair counts and representative roles are fixed. Excluding the original sixteen validation scenarios, a frozen permutation partitions 96 of the remaining 104 scenarios into six disjoint panels of sixteen. Neither scenario IDs nor complete workload/weather arrays duplicate the original panel. Only candidate scores are recomputed: there is no retraining or reconstruction of history. Offline diagnosis costs 4,896 episodes and 1,410,048 steps.

Table H7: Output and feasibility-screen changes across validation panels; 120 instances equal twenty seeds times six panels.

Rule	Output switches	Selected UCB passes	Evaluation failures among passes	Evaluation cost
Terminal	0/120	14	3/14	2.841600
Minimum UCB	10/120	97	5/97	0.302099
Feasible max-return	8/120	97	6/97	0.482744
SG	7/120	90	8/90	0.603246

Analysis preserves dependence among the six panels within a training seed. SG's 8/90 is a descriptive frequency combining panel perturbation and validation-to-evaluation distribution differences, not ninety independent IID coverage trials. Fingerprints, scores and evaluation endpoints for all 480 rule outputs are retained. An auxiliary comparison uses the mean of the other five panels to separate selection changes from feasibility-screen changes.

H.7 Release After Strict Finite-Library Certification

The certification target is the uniform 72-scenario library. Each proposal is fixed by its selector. After evaluating scenarios in S_n, deterministic bounds B_z for unevaluated scenarios give

$$L_n = \frac{1}{72}\sum_{z\in S_n} x_z, \qquad U_n = \frac{1}{72}\left(\sum_{z\in S_n} x_z + \sum_{z\notin S_n} B_z\right). \tag{H1}$$

Because $0 \le x_z \le B_z$, every inspection satisfies $L_n \le J_{\mathcal{Z}}(\pi) \le U_n$, where $\mathcal{Z}$ is the 72-scenario target library. The thermal envelope propagates maximum IT power and zero cooling through the frozen dynamics; monotonicity controls every legal policy. Costs after episode termination are zero-padded and remaining bounds are summed. Inspections at 16, 32, 48, 64 and 72 scenarios

therefore need no repeated-testing correction. Release occurs only if $U_n \leq 1$; $L_n > 1$ permits early rejection. A failed proposal is rejected, without searching replacement candidates on the same certification data or assuming an already certified fallback.

Each environment step has five Euler substeps, $h = \Delta t/5$. Let T_c, T_i denote chip and inlet temperatures; C_c, C_i are thermal capacities, $R_{\mathrm{ci}}, R_{\mathrm{ia}}$ thermal resistances, A outdoor temperature, P IT power, and Q_{cool} cooling power. The frozen substep is

$$T_{\mathrm{c}}' = \left(1 - \frac{h}{R_{\mathrm{ci}}C_{\mathrm{c}}}\right) T_{\mathrm{c}} + \frac{h}{R_{\mathrm{ci}}C_{\mathrm{c}}} T_{\mathrm{i}} + \frac{hP}{C_{\mathrm{c}}} \tag{H2}$$

$$T_{\mathrm{i}}' = \frac{h}{R_{\mathrm{ci}}C_{\mathrm{i}}} T_{\mathrm{c}} + \left[1 - \frac{h}{C_{\mathrm{i}}}\left(\frac{1}{R_{\mathrm{ci}}} + \frac{1}{R_{\mathrm{ia}}}\right)\right] T_{\mathrm{i}} + \frac{hA}{R_{\mathrm{ia}}C_{\mathrm{i}}} - \frac{hQ_{\mathrm{cool}}}{C_{\mathrm{i}}} \tag{H3}$$

Temperature coefficients are nonnegative in the frozen configuration; legal actions ensure $Q_{\mathrm{cool}} \geq 0$ and $0 \leq P \leq P_{\mathrm{idle}} + P_{\mathrm{dynamic}}$. Starting from the same initial temperatures, maximum IT power, zero cooling and the scenario's actual outdoor temperature yield the upper envelope $(\overline{T}_c, \overline{T}_i)$ by substep induction. The implemented cost $c(T_c, T_i) = \max\{0, (T_c - 85)/15, (T_i - 27)/8\}$ is coordinatewise monotone, hence

$$0 \leq x_z \leq B_z = \sum_{t=0}^{287} 0.99^t c(\overline{T}_{c,z,t+1}, \overline{T}_{i,z,t+1}) \tag{H4}$$

Every envelope step completes all five substeps. Continuing the envelope to 288 steps covers zero-padded cost even when the actual episode ends early. Eq. (H1) therefore holds deterministically at all inspections. The bound B_z comes from the physical model and known outdoor temperature, not the proposed policy's observed sample maximum. Implementation retains an upward numerical margin and checks every evaluated cost against its envelope; the real-arithmetic proof and floating-point checks are recorded separately.

Table H8: Finite-library release coverage and cost. Returns among released proposals must be read alongside rejection counts.

Proposal rule	Released / rejected	Released return	Released cost	Mean certification steps
Terminal	4 / 16	191.197737	0.536050	13248.0
Minimum UCB	18 / 2	191.531187	0.099471	20044.8
Feasible max-return	17 / 3	192.483951	0.325260	19929.6
SG	16 / 4	192.091479	0.339583	19699.2

Every passing proposal requires all 72 scenarios, or 20,736 steps; some failures are rejected earlier. This is a certification replay on existing evaluations, adding no interactions, but Table H8 charges the decision cost required to execute the protocol. That cost is outside the original 400,000-step acquisition confirmation. Released-policy costs are full enumerations of the same target library, not additional independent post-certification tests. Rejected proposals are not zero-cost policies. Certification is a separate delivery audit, not part of SG training.

H.8 Additional-Study Ledger and Evidence Versions

Matched-budget confirmation adds sixty complete runs, followed by sixty strong-baseline runs and forty frequency runs. Common-pool comparisons, panel stability, reevaluation and certification replay are charged for actual added interactions, not counted as new training seeds. Confirmation and subsequent audit reuse the same sixty runs without double counting. D1/D2 common-pool confirmation adds no training.

The six-study package totals 32,375,330 new training steps, 8,958,853 validation/offline-diagnostic steps and 4,790,016 evaluation steps. The 1,410,048 stability steps are offline information, not

information available during training. The separately accounted twenty-seed matched-budget package adds 23,941,198 acquired interaction steps: 19,281,047 training and 4,660,151 decision-validation steps. The primary panel, old ten-seed pilot and additions remain separate; protocols are not pooled to enlarge significance samples.

Tables trace to seed endpoints, candidate fingerprints, raw episodes, frozen protocols and source hashes. Factorial results use the common selector's six paired comparisons. Table H1 uses its frozen six-test family; its resampling version is retained separately from earlier exploratory analyses. Source and parameter OOD are aggregated after checkpoint alignment. Earlier incorrect candidates and duplicate evaluations are excluded from formal comparisons but retained in the audit.

H.9 Candidate Opportunity and Finite-Validation Proofs

Fix a pool $\mathcal{C}$ containing π_T before acquiring n IID validation scenarios. Let J be either return or negative cost, with known episode range width M_J. For error allocation $\alpha \in (0,1)$, simultaneous radii are

$$\epsilon_J = M_J \sqrt{\frac{\log(4|\mathcal{C}|/\alpha)}{2n}}. \tag{H0}$$

Subtracting Eq. (8)'s definitions proves Eq. (9). Both O_J and L_J are nonnegative because the terminal and output policies belong to the pool. Positive gain occurs exactly when $L_J < O_J$. The statement applies separately to return and negative cost; their maximizers need not coincide.

For a nonempty allowed set $\mathcal{A}$ containing π_{out}, choose $\pi^\star_{\mathcal{A}} \in \arg\max_{\mathcal{A}} J$. On the uniform-error event,

$$\begin{aligned} \max_{\mathcal{A}} J - J(\pi_{\text{out}}) &\le |J(\pi^\star_{\mathcal{A}}) - \widehat{J}(\pi^\star_{\mathcal{A}})| + \widehat{J}(\pi^\star_{\mathcal{A}}) - \widehat{J}(\pi_{\text{out}}) \\ &\quad + |\widehat{J}(\pi_{\text{out}}) - J(\pi_{\text{out}})| \le 2\epsilon_J + \delta_J. \end{aligned} \tag{H5}$$

Adding D_J proves Eq. (10). The allowed set may depend on validation because the error event covers the entire fixed pool. SG's early exits correspond to singleton allowed sets. Its Pareto exit has zero return-score optimization gap, but generally nonzero cost-score gap. Exact empirical optimization removes δ_J, not estimation error.

For completeness, let $\psi(\lambda) = \log \mathbb{E} e^{\lambda(Y - \mathbb{E}Y)}$ for a bounded episode variable of range width M_J. Exponential tilting preserves the range, giving $\psi''(\lambda) \le M_J^2/4$. Integrating from $\psi(0) = \psi'(0) = 0$ yields $\psi(\lambda) \le \lambda^2 M_J^2/8$. Independence across scenario rows and exponential Markov then give

$$\Pr(|\widehat{J} - J| > \epsilon) \le 2\exp(-2n\epsilon^2/M_J^2). \tag{H6}$$

A union bound over two endpoints and $|\mathcal{C}|$ candidates proves Eq. (H0) with simultaneous probability at least $1-\alpha$. Candidates may share scenarios. Range bounds must be valid in advance, and the pool must be fixed independently of the validation sample. Continual candidate generation on a reused panel needs separate adaptive coverage.

Feasibility-first selection has an additional boundary loss. Within $\mathcal{C}$, define

$$\begin{aligned} \mathcal{S} &= \{\pi : J_C(\pi) \le d\}, \\ \widehat{\mathcal{S}} &= \{\pi : \widehat{J}_C(\pi) + \epsilon_C \le d\}, \\ \mathcal{S}^- &= \{\pi : J_C(\pi) \le d - 2\epsilon_C\}. \end{aligned}$$

Use the same cost functional and budget units for all three sets. On the uniform-error event, $\mathcal{S}^- \subseteq \widehat{\mathcal{S}} \subseteq \mathcal{S}$. If $\mathcal{S}^- \neq \varnothing$, an approximate empirical-return maximizer from $\widehat{\mathcal{S}}$ satisfies

$$\begin{aligned} \max_{\mathcal{S}} J_R - J_R(\pi_{\text{out}}) &\le B_R + 2\epsilon_R + \delta_R, \\ B_R &= \max_{\mathcal{S}} J_R - \max_{\mathcal{S}^-} J_R. \end{aligned} \tag{H7}$$

Proof. The cost error bounds give both inclusions. Apply Eq. (H5) inside $\widehat{\mathcal{S}}$ and use $\max_{\widehat{\mathcal{S}}} J_R \ge \max_{\mathcal{S}^-} J_R$. Any output from a nonempty $\widehat{\mathcal{S}}$ is feasible on this event; the nonempty interior additionally ensures selection is possible and bounds return loss. An empty-set minimum-cost fallback is uncertified.

Sharper validation can reduce estimation error and enlarge the interior, but cannot eliminate structural exclusion. For example, two policies with true and estimated costs zero and returns two and one still deliver one under terminal priority if the latter is terminal. Return opportunity and exclusion loss are both one, even with $\epsilon_R = 0$.

These identities apply to population expectations or complete finite-library endpoints. A finite-sample estimate of the pool optimum adds maximum-selection noise. Appendix I enumerates deterministic policies on all 72 target scenarios to obtain exact library references. Selection is frozen before evaluation; the reference optimum does not enter delivery and is not an unknown-population optimum.

I Shaping Effects and Checkpoint Opportunity

This appendix studies shaping without retention and checkpoint opportunity under matched validation budgets. Section I.1 isolates shaping within optimizers. Section I.2 gives the matched two-checkpoint protocol. Section I.3 reports selection, stability and statistical strength.

I.1 Controlled Shaping Without Retention

Four arms use seeds 41000–41019 and 400,000 training steps each: eighty complete runs and 32,000,000 training interactions. There is no shadow, monitoring validation, retention or intermediate model selection. Terminal deterministic policies are enumerated on the original 72-scenario library, totaling 5,760 episodes. Within TRPO and within CSPO, initialization, environment, optimizer and schedule match. Only ERC training-reward terms change.

D3 shaping adds 0.99 times next potential minus current potential and subtracts squared instantaneous cost, both with unit weight. Potential is squared thermal headroom. Next potential is zero at termination and fixed-horizon truncation. Additional shutdown penalty is zero. Both CSPO arms retain the native 0.30 linear-cost term and cost-sensitivity optimizer. Safety cost, raw task return and dynamics are unchanged. Table I1 reports raw return rather than shaped reward.

The primary six-test family contains return and cost effects of shaping within TRPO, within CSPO, and their difference. Four overall algorithm comparisons form a separate secondary family. Each seed is first aggregated over 72 scenarios, then analyzed by paired bootstrap, exact sign flips and Holm correction. The earlier four-reward panel isolates potential and squared-cost terms. This panel isolates shaping within optimizers, and they are not pooled.

Table I1: Main shaping effects without retention. Differences are shaping minus no shaping; intervals are pointwise 95%.

Comparison	Return difference [95% CI]	Cost difference [95% CI]	Holm p return / cost
Shaping within TRPO	3.908 [-0.760, 10.189]	-26.815 [-32.754, -20.861]	0.668604 / 0.0000114
Shaping within CSPO	-0.233 [-0.461, -0.013]	0.087 [-0.105, 0.330]	0.237709 / 0.668604
Shaping interaction	4.140 [-0.591, 10.209]	-26.902 [-32.726, -20.940]	0.668604 / 0.0000114
TRPO minus CSPO			

ERC minus original CSPO has return difference -1.414 [-1.966, -0.814], and cost difference 4.317 [2.262, 6.846]. Secondary-family Holm p-values are 0.000713 and 0.0000229. ERC also does not dominate shaped CSPO. No corrected shaping gain on CSPO does not establish equivalence or identify the explanation that the optimizer has already internalized constraint structure. The result establishes an independent cost reduction for this TRPO configuration and its configuration dependence.

I.2 Matched Two-Checkpoint Protocol

The study reuses twenty original seeds 200–219 for each of four methods. Each seed contributes only its actual near-midpoint and terminal checkpoints. Unsaved trajectories are not reconstructed.

Midpoints are 200,704 steps for ERC, Lagrangian and SCPO, and 200,000 for CSPO. The 704-step difference is 0.176% of the total budget. All terminals are 400,000. Candidate counts and validation budgets match, while save times are approximate rather than exact matches.

Each candidate uses three disjoint sixteen-scenario validation panels. Panel zero is primary. Panels one and two are sensitivity checks. Methods share each panel and evaluation count. All selection files are frozen before any formal evaluation. The 72-scenario library has no ID overlap with the three validation panels. Identical outputs reuse measurements. The unit remains twenty training seeds, not sixty panel instances.

Five rules are compared: terminal; SG's original identity rule; common retention bypassing the shadow-count shortcut; minimum cost UCB; and maximum validation return among UCB-feasible candidates. The last rule falls back to minimum UCB with a failed-screen label when none is feasible. None of the four methods accepts shadows, so the original identity rule returns terminal. Common retention is an offline output-layer control, not full SG training. Evaluation uses D3's discounted budget one. SCPO's original maximum-cost training target is unchanged.

A feasibility-recovery opportunity means midpoint cost at most one and terminal cost above one. A jointly feasible return opportunity means both are feasible and midpoint return exceeds terminal beyond numerical tolerance. Minimum-cost and maximum-return opportunities are reported separately, not combined into an attainable joint optimum. Practical-effect thresholds are also prespecified. None of CSPO's seventeen jointly feasible return opportunities reaches 1% of terminal return.

Table I2: Midpoint, terminal and common retention on the fixed evaluation library. Every feasibility denominator is twenty.

Method	Midpoint return / cost	Terminal return / cost	Retention return / cost	Feasible mid / terminal / retained
ERC-TRPO	191.591 / 7.904	190.627 / 6.843	191.140 / 4.436	2/3/5
CSPO	192.270 / 0.000	191.948 / 0.197	191.885 / 0.094	20/19/20
Lagrangian-TRPO	191.502 / 0.624	190.356 / 0.639	190.522 / 0.344	16/17/19
SCPO	168.391 / 0.000	185.901 / 0.014	184.453 / 0.014	20/20/20

Table I4 gives primary-panel cost opportunity, loss and gain. ERC's pool has 5/20 feasible coverage. Common retention recovers two seeds whose terminal policies became infeasible, reaching 5/20. Coverage is 20/20 for CSPO, 19/20 for Lagrangian and 20/20 for SCPO. CSPO cost rises from zero at midpoint to 0.197 at terminal, so terminal does not dominate every historical candidate. Lagrangian has twelve jointly feasible return opportunities; terminal-priority skips eleven on the primary panel. SCPO has three small jointly feasible return opportunities, while aggregate return rises from 168.391 to 185.901. Continued learning and degradation can coexist across seeds.

I.3 Selection, Stability and Statistical Strength

On the primary panel, feasible max-return delivers CSPO return 192.161, cost 0.032 and 20/20 feasibility. For Lagrangian, it delivers 191.303, 0.262 and 19/20. Both have better descriptive cost-return combinations than common retention. Early exits and risk-band restrictions can therefore lose opportunities even when candidates exist. Common retention reduces SCPO return by 1.448 while lowering cost by only 0.000143. All four methods are retained.

Table I3: Common-retention outputs under three validation panels; denominators are twenty original training seeds.

Method	Panel 0 feasible	Panel 1 feasible	Panel 2 feasible
ERC-TRPO	5/20	5/20	4/20
CSPO	20/20	20/20	20/20
Lagrangian-TRPO	19/20	19/20	19/20
SCPO	20/20	20/20	20/20

ERC's third panel recovers the same two opportunities but replaces another feasible terminal with an infeasible candidate. None of the 240 common-retention outputs has selected validation UCB at most one and evaluation-library mean above one. Some outputs exceed budget under both panels and evaluation, reflecting risk-band fallback rather than certification. This count and the separate cohort's 8/90 in H.6 are not pooled into a coverage estimate.

The prespecified family contains eight return/cost comparisons of common retention against terminal. ERC's pointwise bootstrap intervals favor retention, but both Holm p-values are 0.125. No other method passes this family either. Table I4 reports opportunity decomposition and observed realization, not a confirmed population improvement for this cohort. In a separate eight-test midpoint-versus-terminal family, CSPO's lower midpoint cost passes correction ($p = 0.011719$). Lagrangian's higher midpoint return and SCPO's higher terminal return also pass (each $p = 0.000122$).

The evaluation library is the designated complete finite support. Pool optima are obtained by 72-scenario enumeration to measure opportunity, never to select outputs. The three validation panels have different partitions. Eq. (H0)'s same-target IID premise cannot simply be assigned to these data. Finite-library endpoints, across-seed uncertainty and validation-to-evaluation shift are interpreted separately. No full-trajectory opportunity is inferred for unsaved checkpoints or unknown environments.

All 160 policies, source hashes, native action equivalence and selector code paths are checked, including SCPO's augmented state. Eighty validation jobs and eighty formal-evaluation jobs finish, totaling 19,200 episodes and 5,526,519 steps. Eight loading-consistency episodes add 2,304 steps, for 5,528,823 study interactions and zero new training. All 1,200 outputs reproduce, 2,400 cost/return decomposition identities hold, and the independent audit passes 1,457 checks.

The first export reused one argument name for training and evaluation steps, causing a summary error. The fix only renames training steps to checkpoint steps. Frozen analysis code and the repair record are retained. Raw episodes, candidates, choices and statistics are unchanged, with no reevaluation. Branch attribution and proposition formalization followed the results and are reported as post-hoc diagnosis.

Table I4: Cost opportunity, selection loss and realized reduction on the fixed 72-scenario library. Return and feasibility opportunities are reported separately.

Method	Feasibility-recovery opportunity	Jointly feasible return opportunity	Cost opportunity	Selection loss	Realized cost reduction
ERC-TRPO	2/20	0/20	2.456	0.049	2.407
CSPO	1/20	17/20	0.197	0.094	0.104
Lagrangian-TRPO	2/20	12/20	0.410	0.116	0.294
SCPO	0/20	3/20	0.014	0.014	0.000

J Persistent-Response Extension and Development Audit

This appendix evaluates a supplementary persistent-response extension that is not part of the base ERC-TRPO-SG method used for the primary results. It reports the 160,000-step development study, the full three-domain extension, and the response-event analysis using their own cohorts and evaluation protocols.

J.1 Multiplier and Resource-Window Definitions

Base SG has no persistent multiplier. The extension preserves ERC rewards, main TRPO, domain-specific shadow repair, validation and delivery, and changes only main-branch cost pressure. Shadow acceptance does not reset its multiplier. We describe the D3 configuration below; Section J.6 states cross-domain estimator calibration.

Write the persistent multiplier as ν_k, its standard proposal as ν_k^{prop}, the actor price as ℓ_k, and its fixed bias as b_{bias}. These differ from the resource b_t and arrival probability p_k in the main text. D3 initializes $\nu_0 = 0$ and uses cap 100, step size 0.05 and budget 1:

$$\nu_k^{\mathrm{prop}} = \Pi_{[0,100]}\{\nu_k + 0.05(\widehat{C}_k - 1)\}, \qquad \ell_k = \min\{\nu_{k+1} + b_{\mathrm{bias}}, 100\}. \tag{J0}$$

If no complete episode occurs, the proposal remains ν_k; a missing estimate is not zero cost. Standard and fixed-bias controls set $\nu_{k+1} = \nu_k^{\mathrm{prop}}$, with $b_{\mathrm{bias}} = 0$ and 0.025, respectively. The bias preserves pressure after the persistent state decays to zero.

The resource-linked variant also uses bias 0.025. Its credit ratio $\upsilon_k \in [0, 1]$ slows downward release:

$$\nu_{k+1} = \begin{cases} \nu_k - \upsilon_k(\nu_k - \nu_k^{\mathrm{prop}}), & \nu_k^{\mathrm{prop}} < \nu_k, \\ \nu_k^{\mathrm{prop}}, & \nu_k^{\mathrm{prop}} \geq \nu_k. \end{cases}$$

A window-only arm uses the same actor weights but standard release, isolating the release increment.

Let $h_t, h_t^+ \in [0, 1]$ be pre- and post-action resource headroom. In D3 this is the minimum normalized distance of chip and inlet temperatures to their hard thresholds. Let A_t^C be the original cost advantage and ω_t the original discount weight. Define

$$w_t^1 = \begin{cases} \mathrm{clip}\big(h_t^+ / \max\{h_t, 10^{-8}\}, 0, 1\big), & A_t^C < 0, \\ 1, & A_t^C \geq 0. \end{cases}$$

For a fully observed window W_t beginning at the current action and ending after 64 steps or episode termination, without crossing a reset, use

$$w_t^H = \begin{cases} \min\left\{ w_t^1, \mathrm{clip}\left(\dfrac{\min_{j \in W_t} h_j^+}{\max\{h_t, 10^{-8}\}}, 0, 1 \right) \right\}, & A_t^C < 0, \\ 1, & A_t^C \geq 0. \end{cases}$$

Censored batch-tail windows without observed termination retain w_t^1. No future values are imputed. With $\mathcal{V}_k$ the fully observed windows,

$$\upsilon_k = \frac{\sum_{t \in \mathcal{V}_k} \omega_t [-A_t^C]_+ w_t^H}{\sum_{t \in \mathcal{V}_k} \omega_t [-A_t^C]_+ w_t^1}. \tag{J1}$$

A zero denominator gives $\upsilon_k = 0$; the implementation clips the ratio to $[0, 1]$.

To avoid confusion with rewards and states, write the likelihood ratio as $\varrho_t(\theta) = \pi_\theta(a_t \mid s_t)/\pi_{\mathrm{old}}(a_t \mid s_t)$ and the original episode-scaling factor as a_k^{ep}. Before the main actor update, subtract

$$\ell_k \Delta_C(\theta), \qquad \Delta_C(\theta) = a_k^{\mathrm{ep}} \, \mathrm{mean}_t\big[\omega_t(\varrho_t(\theta) - 1) A_t^C w_t\big] \tag{J2}$$

from its ERC surrogate. Standard and fixed-bias arms have $w_t = 1$; window-only and resource-linked arms have $w_t = w_t^H$. All window states were collected before this update; no future deployment access, predictor or model rollout is used. The weighted surrogate is a design choice, not an unbiased-gradient guarantee.

For a downward proposal, $0 \leq \upsilon_k \leq 1$ implies $\nu_k^{\mathrm{prop}} \leq \nu_{k+1} \leq \nu_k$. This is a same-proposal pressure relation, not a first-passage ordering across different closed-loop runs. The identical-cost/identical-multiplier statement in Section 4 concerns the standard recursion, whereas resources enter both release and the actor surrogate here.

J.2 New-Seed Confirmation: Budget and Statistics

Five arms use frozen seeds 56100–56119, giving one hundred complete runs without intermediate-result exclusions. Earlier five-seed results, 53620–53624, remain development records. Each run has a 160,000-step actor-plus-validation cap and sixteen validation scenarios. Delivery uses eighty dev_select scenarios disjoint from this run's validation IDs, but previously used in development. New training seeds and a reused evaluation distribution are distinguished; the library is not a wholly unseen formal test.

A single study worker runs independent single-thread processes in a fixed rotating arm order. Training CPU includes collection, updates, validation and storage; delivery evaluation is timed separately. The hundred runs use 12,141,824 actor and 3,818,978 validation steps, totaling 15,960,802 with 39,198 unused against the 16,000,000 cap. Evaluation uses another 5,042,366 steps. SG averages 158,040.1 acquired interaction steps because of recovery reservations; the other arms use 160,000. Caps match, not every realized count.

Primary online comparisons are paired by training seed. Linked-versus-bias and linked-versus-window criteria are frozen, corrected and require at least 1% practical improvement, together with 2% return noninferiority and feasible-seed safeguards. Overall enhancement against SG and standard dual is reported separately. Training cost per step is raw accumulated cost divided by actor steps; delivery cost uses D3's discounted budget one. These are distinct safety quantities.

The full training audit passes 262,333 checks; post-hoc diagnosis verifies 1,859 per-run file hashes. First-positive-multiplier time is inapplicable to SG and serialized as null. That repair changes no trajectory and causes no retraining. Frozen execution records override a legacy terminal-only metadata field: actual selection always uses the original V5PB identity gate and risk band, with branches and output fingerprints checked run by run.

Table J1: Earlier D3 common-160,000-step study, seeds 56100–56119. All original multiplier-table endpoints are retained.

Method	Delivered return	Delivered cost	Feasible seeds	Training cost per step	Training CPU seconds
SG	264.596	1.292982	10/20	0.030992	89.13
+ standard dual	267.086	0.951035	16/20	0.010393	93.48
+ fixed bias	266.678	0.129747	19/20	0.003105	93.69
+ resource window	266.716	0.135316	18/20	0.003083	96.74
+ resource-linked	266.647	0.143406	18/20	0.003085	96.30

Table J2: Interactions and episode risk, averaged within seed first; twenty seeds and eighty episodes per seed.

Method	Mean actor steps	Mean validation steps	Episode exceedance	Positive-cost episodes
SG	114547.2	43492.9	28.50%	75.62%
+ standard dual	123136.0	36864.0	18.88%	58.31%
+ fixed bias	123136.0	36864.0	4.00%	21.69%
+ resource window	123136.0	36864.0	3.94%	22.94%
+ resource-linked	123136.0	36864.0	4.44%	23.75%

Linked/SG training-CPU ratios average 1.082832, with one-sided 95% upper bound 1.102627. Against fixed bias, the mean is 1.027993 and upper bound 1.032680. The overall $\leq$10% overhead gate fails, while the simple-control overhead gate passes. Training and delivery evaluation are timed separately. Equal interaction caps do not establish compute fairness, and historical parallel wall time cannot establish acceleration.

J.3 Overall Enhancement and Resource-Specific Increment

Table J3: Linked-minus-control training cost per step; frozen-analysis 97.5% two-sided paired-bootstrap percentile intervals (1.25th–98.75th percentiles).

Control	Mean difference	Interval	Seeds with strict decrease
SG	-0.02790763	[-0.03417247, -0.02134210]	20/20
+ standard dual	-0.00730809	[-0.00857680, -0.00600152]	20/20
+ fixed bias	-0.00002057	[-0.00004708, 0.00000297]	11/20
+ resource window	0.00000168	[-0.00002761, 0.00003314]	7/20

Against SG, linked online cost has ratio 0.099531 and decreases in every seed. Delivered return differs by 2.051, 95% interval [-0.736, 5.925], supporting the prespecified 2% noninferiority margin rather than a significant increase. Online cost also decreases in all twenty pairs against standard dual. The overall-enhancement gate passes.

Against fixed bias, linked online cost decreases only 0.6625%, improves in 11/20 seeds, and has corrected one-sided direction-test p = 0.823803. It misses the 1% practical-effect gate, with 18/20 feasible seeds versus 19/20 for fixed bias. Against window, online cost rises 0.0544%, with seven improvements and seven ties; the release increment is likewise unestablished. Fixed bias is not renamed as an original resource mechanism, and favorable five-seed effects do not replace twenty-seed evidence.

The release mechanism executes: among 1,220 updates, thirteen seeds have 252 delayed releases. In seven other seeds the multiplier stays zero, and window/linked trajectories and final fingerprints are identical. Among active seeds, online cost decreases in seven and rises in six, with mean difference +0.00000257947, delivered-cost difference +0.0124456 and return difference -0.106872. This post-hoc subgroup does not support dilution by inactive seeds as the explanation for the missing increment.

For each delayed release, the next two batches are united to avoid double counting, then paired within seed. Across thirteen active seeds, local accumulated cost differs by -4.020604 and return by -167.806532; other batches differ by +4.338229 and -45.735638, giving total cost difference +0.317625. These intervals are defined by linked-arm events. They describe locally more conservative behavior and lower cost, without identifying a causal rebound.

Strictly positive multipliers include tiny residuals. Active seeds average only 10.15 updates where the linked/window multiplier difference exceeds 1% of the 0.025 bias. This scale check explains operation without introducing a favorable selection gate. Evidence supports persistent-response enhancement, not linked superiority over bias or universal gains across three domains or nine public tasks.

J.4 Terminal Failures Despite Available Candidates

Failed linked seeds 56116 and 56117 have identically zero multipliers and identical window/linked trajectories. At update 56, actor step 114,688, a historical candidate has validation UCB zero. At update 61, actor step 123,136, terminal UCB exceeds budget; the trigger condition holds but persistence count is only one rather than two, and interaction budget is exhausted. No repair has been accepted, so the identity gate returns terminal.

Table J4: Original delivery and existing historical candidates. Candidates are diagnostic and do not replace Table J1 results.

Seed	Terminal validation UCB	Terminal evaluation cost	Candidate evaluation cost	Terminal return	Candidate return
56116	1.592474	1.112781	0.0000187373	266.683479	265.069339
56117	2.102479	1.153266	0	265.936750	266.620887

These cases are neither false feasible terminal screens nor absent feasible candidates. A structural identity exit excludes an available candidate, corresponding to rule-exclusion loss in Eq. (10). Last-update detection also leaves insufficient budget to complete the trigger. Online risk, detection, candidate supply and delivery are distinct endpoints.

Fixed-bias seed 56116 has the same issue. Seed 56117 has terminal cost 0.979028 under bias versus 1.153266 under window/linked; this contrast is in actor resource weighting, not release. One seed does not establish general harm from the weight. Diagnosis adds no interactions, changes no selector and performs no retraining; post-hoc candidate substitution is not a new 20/20 algorithm result.

J.5 Development Path and Unpromoted Designs

Table J5 retains the development path and each stage's original gate. Training, frozen-policy replay, local updates and conditional evaluation are distinguished. Multiple states from one source are not independent training seeds. Starting a later question does not reverse an earlier failed gate.

Table J5: Complete multiplier-development index. Development evidence and new-seed confirmation remain separate.

Stage	Evidence	Outcome
Prediction-difference dual	6 arms × 5 complete runs	Standard dual helps; resource-prediction increment fails its gate.
Prediction / local-price diagnosis	15 checkpoints; 75 local updates	Prediction gate fails; fixed extra pressure helps. Zero-cost subset signals do not replace the gate.
Response timing	3,000,000 training steps	Calibration fails; decisions match fixed control. Design stopped.
Release decisions	3,000,000 training steps	Development gate fails; 73,728 interruption-related steps charged separately.
SG plus dual	6 arms × 5 seeds; 400k cap	Overall enhancement, without established resource-specific increment.
Resource-compatible weights	New development seeds and separate stress case	Return/feasibility safeguards pass; online-cost and independent-resource gates fail.
Early-actor check	Exact replay of one failed seed	20 updates reproduce; conditional rollouts diagnose behavior, not across-seed efficacy.
Risk direction / robustness	Frozen policies; paired-noise evaluation	Direction-correlation gates fail; no new independent training.
Conditional-state intervention	Frozen-policy evaluation	Separates physical risk, controllability and policy ability from learning gains.
Recovery and release	Conditional-control evaluation	Control gate fails; stopped before learning experiments.
Timing and duration	Conditional evaluation	Evaluation gate fails; no universal earlier-is-better interpretation.
Post-recovery handoff	40 states, 10 trajectories, 1 source seed	Allows learning-stage design, not confirmation of algorithm performance.
Persistent-credit linkage	5 arms × 5 seeds; 160k cap	Independent-component gate fails; all arms retained.
SG baseline completion	15 new runs, 5 reused	Overall-enhancement screen passes; simultaneous return gain and cost reduction unestablished.
20-seed confirmation	5 arms × 20 new seeds; 160k cap	Overall enhancement passes; resource increment and overall low-overhead gates fail.
Terminal / release diagnosis	100 existing runs; 0 new interactions	Local cost reduction does not yield full-run increment; output exits explain two failures.

In the first six-arm, thirty-run study, development costs are 2.403884 for ERC, 0.293841 for standard dual and 0.300692 for resource-prediction differences. Over the primary 24-step prediction horizon, a quadratic resource model reduces mean MSE by 7.17% against a capacity-matched context model, improving in only 3/5 seeds; a strong linear context model remains better. A local 0.025 pressure increase reduces cost by 49.42% relative to standard proposals, in the same direction for 5/5 seeds, but all fifteen standard checkpoints already meet budget on that panel. This supports actor response to price, not necessity of resource prediction or earlier triggering.

An early combination study required evaluation-file I/O recovery; 345,600 duplicate evaluation steps remain charged. Release-decision interruptions and smoke tests are recorded separately. Five-seed SG baselines have different execution histories, so fixed-batch microbenchmarks describe overhead, while compute comparisons use same-round process timing in Table J1.

The early twenty-seed confirmation and terminal diagnosis support persistent-response enhancement without establishing an independent resource-linked increment. The later three-domain extension runs all five frozen arms on the reused seed panel, excluding no domain or seed on performance. Sections J.6–J.8 retain its results and process measurements separately.

J.6 Complete Three-Domain Extension

Five arms in three domains each use twenty seeds 200–219, evaluated on seeds 2000–2049 to match existing baselines. These are reused seed identities and evaluation panels, not new independent generalization confirmation. D1 uses 401,408 actor steps; D2/D3 use 400,000, with validation additional. All 300 runs finish. Five-arm initializations match within seed, and SG triggers, acceptance and delivery remain fixed. Final evaluation occurs after output selection.

Across domains, batch cost is divided by budget. The standard proposal adds 0.05 times batch-cost/budget minus one to the old multiplier, clipped to [0,100]; the actor cost surrogate is also divided by budget. Bias, window and linked arms add price 0.025, capped at 100. For D3 budget one, this matches J.1. Continuous domains retain the multiplier when there is no complete episode; D1 uses mean step cost. Resource signals are observed cognitive reserve in D1, battery in D2 and thermal headroom in D3. Window length, termination and censoring remain unchanged.

Base D1 SG has no cost critic. Multiplier arms use the shadow interface's zero-baseline cost GAE, centered within batch and scaled by one minus discount times the GAE parameter; reward advantages retain original normalization. This is a fixed cross-domain surrogate interface. D2/D3 use learned cost GAE and episode scaling. Resource weights act only on the main branch, shadow acceptance does not reset pressure, and raw environment cost is unchanged.

Table J6: D1 five-arm extension. Acquired interaction steps includes training and validation, excluding final evaluation.

Method	Return	Delivered cost	Feasible	Online cost / step	CPU (s)	Acquired interaction steps steps
SG (same cohort)	3021.111	0.003060	20/20	0.006006	221.83	497,791
+ standard dual	2937.851	0.002871	20/20	0.005149	221.58	490,871
+ fixed bias	3022.605	0.002904	20/20	0.005050	221.73	492,700
+ resource window	3005.503	0.002900	20/20	0.005244	224.43	490,366
+ resource-linked	2937.474	0.002556	20/20	0.005084	222.49	491,091

Table J7: D2 five-arm extension. Acquired interaction steps includes training and validation, excluding final evaluation.

Method	Return	Delivered cost	Feasible	Online cost / step	CPU (s)	Acquired interaction steps steps
SG (same cohort)	5.244	0.000426	19/20	0.015208	271.91	468,944
+ standard dual	0.615	0.010773	19/20	0.017251	272.07	468,579

Table J7 (continued)

Method	Return	Delivered cost	Feasible	Online cost / step	CPU (s)	Acquired interaction steps steps
+ fixed bias	0.538	0.003952	19/20	0.017244	274.73	468,709
+ resource window	0.877	0.003197	17/20	0.023333	273.27	462,031
+ resource-linked	0.709	0.007662	16/20	0.023340	271.57	463,205

Table J8: D3 five-arm extension. Acquired interaction steps includes training and validation, excluding final evaluation.

Method	Return	Delivered cost	Feasible	Online cost / step	CPU (s)	Acquired interaction steps steps
SG (same cohort)	188.483	0.344145	19/20	0.034564	346.62	657,566
+ standard dual	187.199	0.640812	15/20	0.012815	313.25	522,573
+ fixed bias	186.596	0.095654	20/20	0.002492	310.18	515,200
+ resource window	186.673	0.157888	19/20	0.002334	310.21	515,200
+ resource-linked	186.629	0.088055	20/20	0.002318	312.15	515,200

Twenty paired training seeds form the statistical units. Prespecified linked-versus-SG, bias and window comparisons cover return, delivered cost, online cost and CPU: twelve Holm tests per domain. Pointwise 95% intervals use 20,000 paired bootstrap samples; two-sided p-values use 100,000 Monte Carlo sign flips. A pointwise interval excluding zero is not corrected significance. Feasible counts remain descriptive, without selecting favorable extra tests.

Table J9: All prespecified three-domain linked-versus-control comparisons. Differences are linked minus control; intervals are pointwise 95%.

Domain / control	Metric	Mean difference	95% CI	Holm p
D1 / SG	Delivered cost	-0.000504486	[-0.00082122, -0.00019485]	0.055399
D1 / SG	Return	-83.6362	[-166.67, 9.2518]	0.71047
D1 / SG	Online cost	-0.000922191	[-0.0010533, -0.00077387]	0.00036
D1 / SG	CPU	0.660937	[-1.3727, 2.7711]	1
D1 / bias	Delivered cost	-0.000348657	[-0.00078637, 5.2077e-05]	0.72617
D1 / bias	Return	-85.1305	[-158.51, -13.34]	0.35604
D1 / bias	Online cost	3.3652e-05	[-0.00011171, 0.00019059]	1
D1 / bias	CPU	0.764062	[-1.1602, 2.6969]	1
D1 / window	Delivered cost	-0.000344365	[-0.00076461, 6.1512e-05]	0.72617
D1 / window	Return	-68.0288	[-144.92, 10.755]	0.72617
D1 / window	Online cost	-0.000160024	[-0.00024247, -7.9848e-05]	0.01452
D1 / window	CPU	-1.93437	[-4.4156, 0.10395]	0.72617
D2 / SG	Delivered cost	0.0072359	[0.0006088, 0.016588]	1
D2 / SG	Return	-4.53507	[-4.9398, -4.037]	0.00012
D2 / SG	Online cost	0.0081317	[0.0046484, 0.012415]	0.00044
D2 / SG	CPU	-0.346094	[-2.9704, 1.9281]	1
D2 / bias	Delivered cost	0.0037091	[-0.0080629, 0.015386]	1
D2 / bias	Return	0.171184	[-0.28594, 0.75101]	1
D2 / bias	Online cost	0.00609559	[0.0024696, 0.010354]	0.0236
D2 / bias	CPU	-3.16328	[-6.5978, -0.19918]	0.66545
D2 / window	Delivered cost	0.00446484	[-0.0035538, 0.013734]	1
D2 / window	Return	-0.168088	[-0.54085, 0.12155]	1
D2 / window	Online cost	6.73454e-06	[-1.5419e-05, 3.2443e-05]	1
D2 / window	CPU	-1.70547	[-3.8485, 0.21641]	1
D3 / SG	Delivered cost	-0.25609	[-0.43509, -0.089753]	0.075039
D3 / SG	Return	-1.85395	[-2.319, -1.383]	0.00012
D3 / SG	Online cost	-0.0322457	[-0.035606, -0.029037]	0.00012
D3 / SG	CPU	-34.4703	[-38.702, -30.248]	0.00012

Table J9 (continued)

Domain / control	Metric	Mean difference	95% CI	Holm p
D3 / bias	Delivered cost	-0.00759868	[-0.053288, 0.034722]	1
D3 / bias	Return	0.0336874	[-0.14441, 0.20349]	1
D3 / bias	Online cost	-0.000174239	[-0.00033833, -4.6273e-05]	0.063539
D3 / bias	CPU	1.96484	[-1.6812, 6.0579]	1
D3 / window	Delivered cost	-0.069833	[-0.19517, 0.014224]	1
D3 / window	Return	-0.043395	[-0.18726, 0.079092]	1
D3 / window	Online cost	-1.62385e-05	[-6.9302e-05, 3.9664e-05]	1
D3 / window	CPU	1.94141	[-0.90785, 5.5805]	1

D3 linked and fixed bias both achieve 20/20 feasibility. Return differs by 0.0337, 95% interval [-0.1444, 0.2035], and cost by -0.00760 [-0.05329, 0.03472]; both Holm p-values are one, while online-cost p = 0.0635. D1's linked-versus-SG online reduction passes correction, but an independent delivery increment over bias is unestablished. In D2, linked return falls from 5.244 to 0.709, feasibility from 19/20 to 16/20, and online cost rises; other multiplier arms also lose substantial return. D3's favorable trade-off does not establish a uniform three-domain advantage.

J.7 Response Events and Cumulative Violations in Deep Training

All 300 runs export 6,300 events and 1,500 prefixes at 10%, 25%, 50%, 75% and 100% of training. All sixty domain-seed combinations have five arms. The measurement plan was added and locked after 67 runs completed, making it a descriptive measurement added after training began. Actor interactions are the main clock, with concurrent validation counts; same-time order is transition, update, validation, delivery.

Events are first positive cost, first over-budget batch cost, first positive multiplier, first accepted main update with a positive multiplier, first UCB-feasible candidate and final delivery. SG's multiplier is inapplicable; D1 main-update acceptance is unmeasurable because its flag was not saved. Events absent by budget exhaustion are right-censored, never assigned zero delay. Event times accompany observed counts. The chain need not be monotone, and bias-induced actor price is separate from the multiplier itself.

Table J10: Event counts and medians of observed times. Counts are out of twenty; times are actor steps, not censoring-adjusted survival estimates.

Domain / method	First positive cost	First positive dual	First UCB-feasible candidate	Recovery entry
D1 / SG (same cohort)	20/20; 28	N/A	12/20; 188,416	20/20; 98,304
D1 / + standard dual	20/20; 28	20/20; 2,048	20/20; 49,152	20/20; 131,072
D1 / + fixed bias	20/20; 28	20/20; 2,048	20/20; 40,960	20/20; 114,688
D1 / + resource window	20/20; 28	20/20; 2,048	20/20; 49,152	20/20; 122,880
D1 / + resource-linked	20/20; 28	20/20; 2,048	18/20; 49,152	20/20; 131,072
D2 / SG (same cohort)	20/20; 16	N/A	20/20; 106,496	5/20; 65,536
D2 / + standard dual	20/20; 16	20/20; 2,048	20/20; 131,072	5/20; 98,304
D2 / + fixed bias	20/20; 16	20/20; 2,048	20/20; 131,072	5/20; 98,304
D2 / + resource window	20/20; 16	20/20; 2,048	20/20; 155,648	7/20; 98,304
D2 / + resource-linked	20/20; 16	20/20; 2,048	19/20; 147,456	9/20; 98,304
D3 / SG (same cohort)	20/20; 189	N/A	20/20; 49,152	20/20; 98,304
D3 / + standard dual	20/20; 189	20/20; 4,096	20/20; 16,384	12/20; 237,568
D3 / + fixed bias	20/20; 189	13/20; 2,048	20/20; 16,384	0/20; Not observed
D3 / + resource window	20/20; 189	13/20; 2,048	20/20; 16,384	0/20; Not observed
D3 / + resource-linked	20/20; 189	13/20; 2,048	20/20; 16,384	0/20; Not observed

All four D1/D2 multiplier arms first become positive at batch one, 2,048 steps. D2 mean multipliers are about 98.96–99.33, near cap 100; first-batch cost/budget averages about 3319.7. This motivates separate scale and persistence diagnosis without identifying one causal explanation. D3 bias, window and linked arms all obtain UCB-feasible candidates at first validation, step 16,384, and accept no

shadows; only thirteen linked seeds ever have positive multipliers. These endpoints do not support universal improvement through earlier activation.

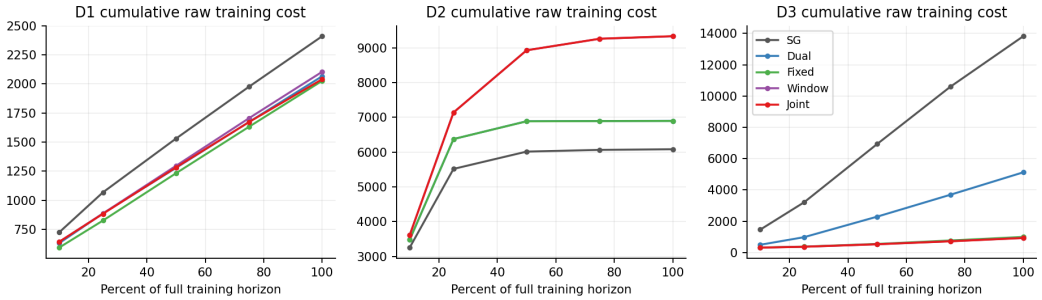

Figure J1: Cumulative raw training-cost prefixes for all three domains. Lines average twenty seeds per arm; the horizontal axis is the planned training-horizon fraction. Cost units are domain-specific. These are prefixes of completed full-horizon runs, not separately trained short-budget experiments; the recovery gate uses the planned remaining horizon.

Figure J1 plots cumulative raw training-cost prefixes for all three domains. Lines average twenty seeds per arm; the horizontal axis is the planned training-horizon fraction. Cost units are domain-specific. These are prefixes of completed full-horizon runs, not separately trained short-budget experiments; the recovery gate uses the planned remaining horizon.

The archive retains cumulative raw cost, positive-cost transitions, complete-episode counts, over-budget complete episodes and summed positive excess. The last unfinished episode contributes raw cost but not complete-episode exceedance rates. D1 uses episode per-step cost; D2/D3 use discounted episode cost, so excess sums are not pooled across domains. Historical weights are incomplete for the 300-run extension. Best-validation curves do not replace independently evaluated oracles; dense-pool oracle analysis and independently retrained short-budget runs are not completed here.

J.8 Historical-Panel Reproduction and CSPO Comparison

Same-round SG achieves 20/20, 19/20 and 19/20 feasibility in D1/D2/D3, versus 20/20 in each historical primary panel. D2 seed 204 has cost 0.008513 and D3 seed 202 cost 1.182739, judged by original budgets. Historical Figure 2 endpoints remain reproducible from their checkpoints and are not overwritten. Extension differences use same-round SG, never splice in historical SG.

Historical D2/D3 used CUDA; the extension uses CPU, with six core source files unchanged from frozen versions. A zero-update first-batch check reproduces the new batch on CPU and approaches the historical batch on CUDA without meeting the 10^{-10} exact-match threshold. D2 values are historical 1.670102612896, CPU 1.677886389114, CUDA 1.670102419137; D3 values are 4.244426573067, 3.352094378477 and 4.244424398431. Device-related sampling diverges before policy updates, but this does not establish device as the sole cause of final 19/20 outcomes. Two domains and devices use 8,192 diagnostic steps, not new training.

Historical D3 CSPO has return 188.296877, cost 0.262942 and 19/20 feasibility. After checking identities and sources of fifty evaluation trajectories, linked minus CSPO has exploratory return difference -1.667576, 95% interval [-2.157643, -1.178873], and cost difference -0.174887 [-0.416489, 0.003719]. This post-hoc cross-batch comparison is outside the 300-run frozen test family and lacks matched-hardware timing. It describes a risk-return trade-off, not overall CSPO dominance.

J.9 Added-Evidence Coverage and Accounting

Added evidence is stored in four groups: the 300-run extension and process reanalysis; 9,000 frozen-evaluation episodes for ninety historical public-task terminal models; eighteen public-task repair-development runs and one hundred last-step diagnostic episodes; and 2,880 selection-validation episodes for 180 candidates plus 2,000 independent evaluations for forty outputs. Earlier shaping,

opportunity, matched-budget, conditional long-run certification, OOD and multiplier development remain in their original appendix sections.

The extension separately charges actor, validation and delivery evaluation. Prefix reconstruction and audits add no interactions. Public-task evaluation and selection monitoring are charged in Appendix E. Earlier nine-task development summaries and the eighteen complete raw runs remain distinct. Failed development gates, unstarted conditional confirmation and incomplete dense-oracle work are reported in their actual states. This manuscript revision starts no training and changes no raw results.