# AUDITING A LEARNED ADMISSION GATE FOR OFFLINE-TO-ONLINE SAFE RL: DISTINCT CERTIFICATES AND PUBLIC COMPARATOR FAILURES

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

We audit one learned admission gate between BC-All proposals and BC-Safe fallbacks in offline-to-online safe reinforcement learning. We distinguish two certification targets. Fixed-time split-conformal accounting controls an immediate-cost proxy under deployment exchangeability and separately assumed fallback control. A weighted order-statistic rule instead bounds the marginal next-trajectory event $\Pr(G_T > D, \mathcal{Q} \neq \varnothing) \leq \alpha$ for a prespecified family of complete policies; it is not the high-confidence population-risk target of Learn then Test, and neither result protects qualification rollouts or repairs policy-induced shift. A synthetic study shows the intended mechanism and its pre-certificate exposure. On three public Safety-Gymnasium Point tasks, 54 completion/rejection-qualified BC-All/BC-Safe checkpoint pairs fail the frozen joint completion–violation criterion against a pre-treatment, deployable rate-matched mixer. A separate retrospective stress test permutes each gate trajectory's realized fallback mask at equal count and horizon; it is post-treatment and not a causal timing comparison. At the frozen original threshold, gate violation is higher at all task-level point estimates under both comparisons, while completion often increases. Finally, an outcome-inspected diagnostic applying the trajectory rule to 600 retained gate trajectories per selected checkpoint yields an empty qualified set (0/54; rank 600; order statistics 134–810 for budget 10), under an unverified cross-campaign exchangeability premise. Thus the instantiated gate refuses trajectory certification and fails the frozen joint public criterion; this does not generalize to a gate class or establish deployment safety.

## 1 INTRODUCTION

Offline safe-RL systems often place a learned gate between a proposal policy and a conservative fallback. Two questions are usually conflated. The first is what such an admission layer can *certify*: a per-decision proxy bound, a trajectory-level chance constraint, or nothing beyond the fallback's own guarantee. The second is whether a particular learned gate is *useful*, which a favorable gate-versus-proposal contrast cannot show, because switching also changes how often the competent fallback acts. We treat the layer as a measurable admission process $g_\phi : (s, \mathcal{C}) \mapsto (a^{\text{exec}}, z)$ that logs a decision tuple, and we answer both questions with their evidence kept separate.

**Contributions.** We separate fixed-time proxy accounting from complete-trajectory qualification, state their exchangeability and exposure boundaries, and distinguish the latter from LTT's population-risk target (Section 5). Controlled studies check the mechanism without claiming public transport (Section 6.1). The public audit then compares one instantiated gate on 54 selected checkpoints with a pre-treatment mixer and a post-treatment schedule stress test, preserving all adverse and null outcomes (Section 7). The checkpoint pair, not the episode, is the inferential unit.

**Scope.** The evaluated gate is a binary $M = 1$ switch scoring the BC-All proposal with a three-member cost ensemble, disagreement and vote thresholds, a split-conformal margin, and an offline



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.

nearest-neighbor support check; rejection executes the matched BC-Safe fallback. The certificates are marginal and fixed-time or finite-horizon; none controls the infinite-horizon expected cost constraint or qualification-rollout safety. The public result concerns three Point tasks, one instantiated gate, 54 selected checkpoint pairs, and two matched schedules; the count-matched mask is post-treatment, so it does not identify a causal timing effect, define a deployable mixer, hold state occupancy fixed, or rank all schedules.

## 2 Related Work

Cal-QL Nakamoto et al. (2023), CPQ, COptiDICE, and FOSP Xu et al. (2022); Lee et al. (2022); Cao et al. (2025) optimize offline or offline-to-online policies. CAPS, evaluation-time policy switching, and Recovery RL Chemingui et al. (2025); Neggatu et al. (2025); Thananjeyan et al. (2021) make state-dependent policy or recovery choices. Other concrete safeguards include SAILR's advantage-based intervention, predictive safety filters, and SPIBB's baseline bootstrapping Wagener et al. (2021); Wabersich and Zeilinger (2021); Laroche et al. (2019). Their guarantees rely, respectively, on a safe backup and absorbing unsafe states, a dynamics model with terminal-set/backup feasibility, or baseline-relative reward improvement; none transfers automatically to an unverified BC-Safe fallback or to our cost event. We did not reproduce these competitive baselines: V4's state-independent random mixer and V5's outcome-conditioned mask permutation are controls for this gate, not evidence against switching or recovery methods.

Conformal methods also have distinct targets. Learn then Test (LTT) provides a high-probability-over-calibration population-risk statement for any candidate selected from an FWER-valid set Angelopoulos et al. (2025); Conformal Risk Control controls the expected value of an exchangeable monotone bounded loss Angelopoulos et al. (2024). Conformal Policy Control addresses risk under policy-induced shift relative to a safe reference, under its exact weighting, reference-safety, stability, and smoothness premises Prinster et al. (2026). These differ from our marginal next-complete-unit order-statistic event. None automatically protects qualification exposure, and our audit does not estimate the causal value of selector timing.

## 3 Problem: Offline-to-Online Admission

We consider a constrained Markov decision process $(\mathcal{S}, \mathcal{A}, P, r, c, \gamma)$. The offline data are $\mathcal{D}_{\text{off}} = \{(s_i, a_i, r_i, c_i, s_i')\}_{i=1}^n$, collected by an unknown behavior policy. The deployment constraint is the expected discounted cost budget

$$J_c(\pi) = \mathbb{E}_{\pi,P}\left[\sum_{t\geq 0} \gamma^t c(s_t, a_t)\right] \leq d. \tag{1}$$

The fixed-time admission guarantee below does not certify Eq. (1): it controls a per-decision event involving $Y(S_t, a)$. Proposition 3 instead calibrates the realized finite-horizon discounted cost directly. Under the complete-unit exchangeability and prespecification conditions of Proposition 3, it bounds the selected policy's joint event $\Pr(G_T > D,\ \mathcal{Q} \neq \varnothing) \leq \alpha$, not the expected infinite-horizon constraint in Eq. (1). Conditional on qualification, it gives only $\Pr(G_T > D \mid \mathcal{Q} \neq \varnothing) \leq \min\{1, \alpha/q\}$ when $q = \Pr(\mathcal{Q} \neq \varnothing) > 0$. Standard offline-to-online fine-tuning asks for a policy update $\pi_\theta \mapsto \pi_{\theta'}$. We instead insert a per-decision admission layer between a learned proposal $\pi_\theta$ and the environment. At time $t$, a proposal set $\mathcal{C}_t = \{a_t^{(m)}\}_{m=1}^M$ is scored. Candidates must satisfy a nearest-neighbor behavior-support gate,

$$D_{\text{sup}}(s, a) = \min_{(s_i, a_i)\in\mathcal{D}_{\text{off}}} \|\psi(s, a) - \psi(s_i, a_i)\|_2, \qquad b_{\text{sup}}(s, a) = \mathbf{1}[D_{\text{sup}}(s, a) \leq h], \tag{2}$$

The gate in Eq. (2) uses a fixed support feature map $\psi$ chosen before deployment and an offline support radius $h$ estimated from $\mathcal{D}_{\text{off}}$. Admitted actions are ranked by robust cost plus a small proposal-distance penalty $\Delta(a, a_t^{\text{prop}}) = \|a - a_t^{\text{prop}}\|_2$, weighted by $\beta \geq 0$. An admitted action executes when available, otherwise the fallback policy $\pi_{\text{fb}}$ executes:

$$a_t^{\text{exec}} \in \begin{cases} \arg\min_{a\in\mathcal{C}_t\cap\mathcal{A}_{\text{adm}}(s_t)} U_\kappa(s_t, a) + \beta\Delta(a, a_t^{\text{prop}}), & \mathcal{C}_t \cap \mathcal{A}_{\text{adm}}(s_t) \neq \varnothing, \\ \pi_{\text{fb}}(s_t), & \text{otherwise.} \end{cases} \tag{3}$$

The handoff in Eq. (3) is a measurable map $g_\phi : (s_t, \mathcal{C}_t) \mapsto (a_t^{\text{exec}}, z_t)$, where $z_t$ records the acceptance margin, vote fraction, ensemble disagreement, behavior-support flag, proposal distance, and fallback indicator. The empirical question becomes whether this admission map improves the safety–coverage frontier and whether its theoretical guarantees survive the move from one candidate to a deployed trajectory.

## 4 SAFE-O2O ADMISSION RULE

We now instantiate the admission process $g_\phi$ of Eq. (3) with an ensemble cost score, conformal shrinkage, support gating, and a behavior-clone fallback whose control is separately assumed. This gate is our construction from established components, not a reproduction of a specific published gate; the audit concerns this instantiation. Let $\{Q_c^k\}_{k=1}^K$ be an ensemble of conservative cost critics. Each critic induces a thresholded candidate safe set $\mathcal{S}_k(\eta) = \{(s,a) : Q_c^k(s,a) \le \eta\}$. We use the ensemble for both a scalar uncertainty and a set-membership event:

$$p_{\text{safe}}(s,a) = \frac{1}{K}\sum_{k=1}^{K} \mathbf{1}[(s,a) \in \mathcal{S}_k(d')]. \tag{4}$$

With ensemble mean $\mu_c = K^{-1}\sum_k Q_c^k$ and disagreement $\sigma_c = (K^{-1}\sum_k (Q_c^k - \mu_c)^2)^{1/2}$, define the robust cost score $U_\kappa(s,a) = \mu_c(s,a) + \kappa\sigma_c(s,a)$.

*Proxy boundary and estimation.* The calibrated variable $Y(s,a)$ is used only for per-decision calibration; it is not the discounted cost-to-go of the executed trajectory. In our implementation, $Y(s_i,a_i) = c_i$ is the observed immediate cost label of held-out transition $i$, while $U_\kappa$ is the ensemble's conservative cost score trained with bootstrapped cumulative-cost targets. The calibrated threshold $d'$ is obtained by split-conformal shrinkage on a held-out offline split. Given calibration residuals $r_i = Y(s_i,a_i) - U_\kappa(s_i,a_i)$ on $\mathcal{D}_{\text{cal}} \subset \mathcal{D}_{\text{off}}$ with integer $n = |\mathcal{D}_{\text{cal}}| \ge 1$, $0 < \alpha < 1$, and $\alpha \ge 1/(n+1)$, the integer rank $k = \lceil (n+1)(1-\alpha) \rceil$ satisfies $1 \le k \le n$, and

$$\begin{aligned} q_{1-\alpha} &= r_{(k)}, \qquad r_{(1)} \le \cdots \le r_{(n)} \text{ with ties retained,} \\ q_+ &= \max\{q_{1-\alpha}, 0\}, \qquad d' = d - q_+. \end{aligned} \tag{5}$$

The admission set is

$$\mathcal{A}_{\text{adm}}(s) = \{a \in \mathcal{A} : \sigma_c(s,a) \le \varepsilon,\ b_{\text{sup}}(s,a) = 1,\ \mu_c(s,a) + \kappa\sigma_c(s,a) \le d',\ p_{\text{safe}}(s,a) \ge \tau\}. \tag{6}$$

Rejected actions are replaced by the fallback policy $\pi_{\text{fb}}$, instantiated here as a behavior clone trained on $\mathcal{D}_{\text{off}}$. Let $\widetilde{a}_t$ be the selected admitted candidate when one exists, and otherwise the first proposal whose diagnostics are logged and let $F_t = \mathbf{1}[\mathcal{C}_t \cap \mathcal{A}_{\text{adm}}(s_t) = \varnothing]$. Each decision logs

$$z_t = (\mu_c(s_t,\widetilde{a}_t), \sigma_c(s_t,\widetilde{a}_t), p_{\text{safe}}(s_t,\widetilde{a}_t), m_t, b_{\text{sup}}(s_t,\widetilde{a}_t), \Delta(\widetilde{a}_t, a_t^{\text{prop}}), F_t)\,, \tag{7}$$

where $m_t = d' - \mu_c(s_t,\widetilde{a}_t) - \kappa\sigma_c(s_t,\widetilde{a}_t)$. Thus the fields are, in order, critic mean, disagreement, safe-vote fraction, signed robust-cost margin, support indicator, distance from the original proposal, and the unambiguous fallback-execution indicator. The tuple records a decision, not an outcome or safety certificate.

**Algorithm sketch.** *Calibrate* $d'$ by Eq. (5) and specify $\pi_{\text{fb}}$; $\alpha_{\text{fb}}$ is the separate marginal bound in Assumption 2, not an output of this calibration. *For each* $s_t$*:* score $\mathcal{C}_t$ via $(\mu_c, \sigma_c, p_{\text{safe}})$, form $\mathcal{A}_{\text{adm}}(s_t)$ by Eq. (6); if $\mathcal{C}_t \cap \mathcal{A}_{\text{adm}} \ne \varnothing$ execute $\arg\min U_\kappa + \beta\Delta$ over admitted actions, else execute $\pi_{\text{fb}}(s_t)$; log $z_t$.

## 5 THEORY OF ADMISSION PROCESSES

Proposition 1, Lemma 1, and Proposition 2 give fixed-time proxy bookkeeping. Proposition 3 changes the calibration unit and endpoint to obtain a finite-horizon statement for a complete gated policy. Propositions 6–7 state supporting convex geometry.

**Definition 1** (Admission process and achievable set)**.** *An* admission process *is a measurable map* $g_\phi : (s, \mathcal{C}) \mapsto (a^{\text{exec}}, z)$*. For population* $\mathcal{P}$*, let* $R_\phi = \Pr(A \in \mathcal{A}_{\text{adm}}(S;\phi), Y(S,A) > d)$ *and* $C_\phi = \Pr(A \in \mathcal{A}_{\text{adm}}(S;\phi))$*. We analyze the closed convex achievable set*

$$\mathcal{H} = \overline{\text{conv}}\{(R_\phi, C_\phi) : \phi \in \Phi\} \subset [0,1]^2, \qquad \kappa(\alpha) = \max\{C : (R,C) \in \mathcal{H},\ R \le \alpha\}.$$

*We require* $\Phi$ *to contain the reject-all rule* $\phi_\varnothing$*, for which* $(R_{\phi_\varnothing}, C_{\phi_\varnothing}) = (0,0)$*. Convexification means an ex ante randomized choice among complete rules, independent of* $(S,A,Y)$*; closure includes limits. Hence every risk slice is nonempty, and compactness of* $\mathcal{H}$ *makes the maximum well defined. Without this requirement, for example* $\mathcal{H} = \{(1/2, 1/2)\}$ *and* $\alpha < 1/2$*, the displayed maximum is undefined.*

**Proposition 1** (Nested admission monotonicity)**.** *Fix the critics* $\{Q_c^k\}_{k=1}^K$ *and the support gate of Eq.* (2)*. Compare two knob settings with* $q'_+ \ge q_+$*,* $\kappa' \ge \kappa$*,* $\varepsilon' \le \varepsilon$*, and* $\tau' \ge \tau$*. Then* $\mathcal{A}'_{\text{adm}}(s) \subseteq \mathcal{A}_{\text{adm}}(s)$ *for every state* $s$*.*

*Proof.* Each change weakens no acceptance constraint: $d - q'_+ \le d - q_+$ tightens the robust-cost margin; larger $\kappa'$ increases $U_\kappa$ because $\sigma_c \ge 0$; smaller $\varepsilon'$ tightens disagreement; larger $\tau'$ tightens the vote threshold. Any action admitted by the stricter tuple therefore satisfies all constraints of the looser tuple. □

**Assumption 1** (Fixed-time marginal calibration)**.** *For a fixed deployment time* $t$ *and every proposal index* $m \in \{1, \dots, M\}$*, the calibration residuals* $\{Y(s_i,a_i) - U_\kappa(s_i,a_i)\}_{i \in \mathcal{D}_{\text{cal}}}$ *and candidate residual* $Y(S_t, a_t^{(m)}) - U_\kappa(S_t, a_t^{(m)})$ *are jointly exchangeable. No conditional-on-the-realized-calibration or sequential exchangeability is asserted.*

**Assumption 2** (Marginal fallback control)**.** *At the same fixed time* $t$*, the fallback satisfies*

$$\Pr(Y(S_t, \pi_{\text{fb}}(S_t)) > d) \le \alpha_{\text{fb}}.$$

*This is a separate assumption; calibrating on the same held-out split does not automatically establish it under deployment shift.*

**Lemma 1** (Per-candidate unsafe-admission control)**.** *Under Assumption 1 and Eq.* (5)*, each prespecified candidate* $A_t = a_t^{(m)}$ *satisfies*

$$\Pr\left(A_t \in \mathcal{A}_{\text{adm}}(S_t) \ \wedge\ Y(S_t, A_t) > d\right) \le \alpha. \tag{8}$$

*Proof.* Joint exchangeability gives the ordinary split-conformal marginal statement $\Pr(r_t > q_+) \le \alpha$, where $r_t = Y(S_t,A_t) - U_\kappa(S_t,A_t)$ and $q_+ \ge q_{1-\alpha}$. If $A_t \in \mathcal{A}_{\text{adm}}(S_t)$, Eq. (6) gives $U_\kappa(S_t,A_t) + q_+ \le d$. Therefore $\{A_t \in \mathcal{A}_{\text{adm}}(S_t), Y(S_t,A_t) > d\} \subseteq \{r_t > q_+\}$, proving the marginal bound. This argument does not condition on the realized calibration sample and does not create a supermartingale. □

Lemma 1 provides the per-candidate marginal premise for the fixed-time accounting result.

**Proposition 2** (Fixed-time executed-action proxy accounting)**.** *Suppose, in addition to the conditions of Lemma 1, that Assumption 2 holds. Let* $a_t^{\text{exec}}$ *be the action executed by Eq.* (3) *from a proposal set* $\mathcal{C}_t$ *of size* $M$*. Then at the fixed time* $t$ *for which these assumptions hold*

$$\Pr\left(Y(S_t, a_t^{\text{exec}}) > d\right) \le \min\{1, M\alpha + \alpha_{\text{fb}}\}. \tag{9}$$

*In particular,* $M = 1$ *(single-proposal mode) gives* $\min\{1, \alpha + \alpha_{\text{fb}}\}$*.*

*Proof.* Partition by admit/fallback. On $\{\mathcal{C}_t \cap \mathcal{A}_{\text{adm}}(S_t) \ne \varnothing\}$, $a_t^{\text{exec}} \in \mathcal{A}_{\text{adm}}(S_t) \cap \mathcal{C}_t$. Then $\{Y(S_t, a_t^{\text{exec}}) > d,\ a_t^{\text{exec}} \in \mathcal{A}_{\text{adm}}\} \subseteq \bigcup_{m=1}^M \{a_t^{(m)} \in \mathcal{A}_{\text{adm}}(S_t),\ Y(S_t, a_t^{(m)}) > d\}$, since $a_t^{\text{exec}}$ equals one of the $M$ admitted candidates. By Lemma 1 applied to each candidate and the union bound, this set has probability at most $M\alpha$. On $\{\mathcal{C}_t \cap \mathcal{A}_{\text{adm}}(S_t) = \varnothing\}$, $a_t^{\text{exec}} = \pi_{\text{fb}}(S_t)$, and the event has probability at most $\alpha_{\text{fb}}$ by hypothesis. Summing the two cases and using the trivial probability bound of one gives Eq. (9). The $M = 1$ specialization is immediate. □

*Boundary and operational envelope.* Lemma 1 is silent when fallback fires; Proposition 2 exposes that term. This route remains fixed-time, depends on separately assumed fallback control, and concerns immediate $Y$. It does not link cumulative-target critic training to trajectory cost. For desired proxy ceiling $\beta > \alpha_{\mathrm{fb}}$, it requires $|\mathcal{D}_{\mathrm{cal}}| \geq M/(\beta - \alpha_{\mathrm{fb}}) - 1$; for $\beta = .10, \alpha_{\mathrm{fb}} = .05$, the minima are 19, 99, and 199 for $M = 1, 5, 10$.

Let $G_T(\pi; W) = \sum_{t=0}^{T-1} \gamma^t c_t$ be the realized discounted cost when complete policy $\pi$ (proposal, gate, and fallback together) acts on deployment unit $W$.

**Proposition 3** (Weighted complete-trajectory qualification)**.** *Fix $0 < \alpha < 1$, complete policies $\Pi = \{\pi_1, \ldots, \pi_L\}$, and weights $w_\ell > 0$ with $\sum_\ell w_\ell \leq 1$ before calibration. For each $\ell$, let $n_\ell \geq 0$ be an integer and suppose its $n_\ell$ calibration costs and next complete-unit cost are exchangeable. Define $r_\ell = \lceil (n_\ell+1)(1-\alpha w_\ell) \rceil$, where $\lceil x \rceil$ is the smallest integer not below $x$. Then $1 \leq r_\ell \leq n_\ell+1$. Sort costs nondecreasing with ties retained; let $q_\ell$ be the $r_\ell$th cost when $1 \leq r_\ell \leq n_\ell$, and set $q_\ell = +\infty$ otherwise. Set $\mathcal{Q} = \{\ell : q_\ell \leq D\}$. Any calibration-measurable $\widehat{\ell} \in \mathcal{Q}$ satisfies*

$$\Pr\Big(G_T(\pi_{\widehat{\ell}}; W_{\widehat{\ell},n_{\widehat{\ell}}+1}) > D,\ \mathcal{Q} \neq \varnothing\Big) \leq \alpha. \tag{10}$$

*Proof.* For fixed $\ell$, split-conformal ranks bound $\Pr(G_T(\pi_\ell; W_{\ell,n_\ell+1}) > q_\ell) \leq \alpha w_\ell$. The selected failure event is contained in the union of these events, whose probability is at most $\alpha \sum_\ell w_\ell \leq \alpha$. No independence across policies is needed. □

The rank exists exactly when $\alpha w_\ell \geq 1/(n_\ell + 1)$. Hence some prespecified allocation can make every policy rank-feasible iff $\sum_\ell (n_\ell + 1)^{-1} \leq \alpha$; choosing $w_\ell \geq [\alpha(n_\ell + 1)]^{-1}$ proves sufficiency, and summing these necessary lower bounds proves necessity. Uniform $w_\ell = 1/L$ therefore requires $L \leq \alpha(n + 1)$ when every policy has $n$ units. Nonuniform weights can prioritize prespecified policies or unequal rollout budgets but cannot remove the need for policy-level evidence. The policy may gate adaptively within a trajectory because proposal, rejection, state feedback, and fallback are inside $G_T$. The result is marginal, not conditional, shifted, or anytime, and controls neither Eq. (1) nor qualification-rollout safety. If $q = \Pr(\mathcal{Q} \neq \varnothing) > 0$, Eq. (10) implies only

$$\Pr(G_T > D \mid \mathcal{Q} \neq \varnothing) \leq \min\{1, \alpha/q\};$$

without a lower bound on $q$, no nontrivial conditional guarantee follows. This is division by $q$, not a new conformal result.

**Distinct target from Learn then Test.** Proposition 3 is an exact, distribution-free marginal prediction statement for the next complete-unit cost. LTT instead tests hypotheses about each candidate's population risk and, with probability at least $1 - \delta$ over calibration, permits any selection from an FWER-valid risk-controlled set Angelopoulos et al. (2025). The latter is often the more operational high-confidence target, but it does not subsume the former without equating risk, confidence, and loss semantics. Conversely, our next-unit statement is not a PAC population-risk certificate. Both require a prespecified candidate and testing procedure plus a calibration law relevant to deployment; neither supplies transport, sequential validity, or safe qualification exposure.

**Proposition 4** (Reachable unsupported-pair identification barrier)**.** *Fix horizon $T$ and candidate $\pi$. Let behavior occupancy of $(s^*, a^*)$ be zero, but let the common pre-hit dynamics give candidate hitting probability $\rho_\pi = \Pr_\pi(\exists t < T : (S_t, A_t) = (s^*, a^*)) > 0$. Suppose the model class permits two continuations that agree on behavior support and before the hit, all non-hit and safe-hit trajectories have $G_T \leq D$, and every unsafe-hit trajectory has $G_T > D$. Then the two models have identical offline-data laws but candidate violation probabilities $0$ and $\rho_\pi$. Therefore no offline-only procedure can distinguish them or uniformly certify a violation bound $\beta < \rho_\pi$ for this class.*

*Proof.* The behavior never visits $(s^*, a^*)$, so changing only its continuation leaves the offline-data law unchanged. Before the first hit, candidate paths and hence hit probability $\rho_\pi$ are common. The stated cost-range conditions make violation impossible in the safe model and equivalent to hitting in the unsafe model, yielding probabilities $0$ and $\rho_\pi$. □

**Proposition 5** (Qualification exposure precedes the certificate)**.** *For policy $\ell$, let $V_\ell = \sum_{j=1}^{n_\ell} \mathbf{1}[G_T(\pi_\ell; W_{\ell j}) > D]$. If its exchangeable calibration and next-unit indicators have marginal*

*violation probability $p_\ell$, then $\mathbb{E}[V_\ell] = n_\ell p_\ell$. Proposition 3 gives no bound on realized $V_\ell$ and no distribution-free upper bound below $n_\ell$ is possible without an additional pre-calibration safety assumption.*

*Proof.* Linearity of expectation gives the identity without independence. The exchangeable law concentrated on indicator one has $V_\ell = n_\ell$ surely, establishing the worst case. Qualification is computed only after observing these endpoints. □

Unsupported behavior occupancy alone is insufficient: if $\rho_\pi = 0$, both violation probabilities are zero; if the permitted continuation range cannot cross $D$, the construction also gives no contrast. The zero-versus-one specialization additionally requires $\rho_\pi = 1$. These are analytic boundary cases of the proposition, not reported executable experiments. Policy-specific rollouts are not the only route, but replacing them requires licensed coverage, a valid dynamics model, an external shield, or another identifying restriction; none is assumed here.

**Proposition 6** (Capacity of the convex achievable set)**.** *For the closed convex set $\mathcal{H}$, $\kappa$ is attained, non-decreasing, and concave on* $[0, 1]$.

*Proof.* The reject-all point belongs to every feasible slice, so each slice is nonempty; it is compact as a closed subset of compact $\mathcal{H}$, proving attainment. Monotonicity follows by nesting of those slices. If $(R_i, \kappa(\alpha_i)) \in \mathcal{H}$ with $R_i \le \alpha_i$, convexity puts their $\theta$-mixture in $\mathcal{H}$ with risk at most $\theta\alpha_1 + (1-\theta)\alpha_2$ and coverage $\theta\kappa(\alpha_1) + (1-\theta)\kappa(\alpha_2)$, proving concavity. □

For $0 < \lambda < \infty$, define $L_\lambda(R, C) = R + \lambda(1 - C)$ directly on $\mathcal{H}$. This parameterization selects efficient points supported by positive finite weights; it need not select endpoints requiring zero or limiting weights.

**Proposition 7** (What Lagrangian tuning does and does not select)**.** *A minimizer of $L_\lambda$ exists and is a supported efficient point of $\mathcal{H}$: $(R^*, C^*)$ minimizes iff*

$$R - \lambda C \ge R^* - \lambda C^* \quad \textit{for every } (R, C) \in \mathcal{H}.$$

*Moreover,*

$$\min_{(R,C)\in\mathcal{H}} L_\lambda(R, C) = \min_{\alpha\in[0,1]} \{\alpha + \lambda[1 - \kappa(\alpha)]\}.$$

*If an optimum occurs at an interior $\alpha^*$, then and only then $1/\lambda$ belongs to the concave superdifferential $\partial\kappa(\alpha^*)$; where one-sided derivatives exist this is $\kappa'_+(\alpha^*) \le 1/\lambda \le \kappa'_-(\alpha^*)$. Endpoints require the corresponding one-sided inequality.*

*Proof.* Compactness gives existence, and the displayed half-space condition is exactly the definition of minimizing the linear functional. For any $(R, C) \in \mathcal{H}$, $C \le \kappa(R)$, giving the "$\ge$" direction of the scalar equality. Conversely, a capacity maximizer at $\alpha$ has some actual risk $r \le \alpha$; evaluating at $r$ gives the reverse direction after minimizing over $\alpha$. The interior condition is $0 \in \partial[\alpha - \lambda\kappa(\alpha)]$, equivalently $1/\lambda \in \partial\kappa(\alpha)$. □

*Boundary.* This is support geometry, not a universal equivalence between knob optimization and every constrained optimum. A scalar Lagrangian selects exposed faces (possibly multiple points); it need not uniquely select a rule, recover unsupported points of a nonconvex unrandomized family, or justify empirical slope estimation from five configurations.

## 6 Experiments and Reporting Protocol

We report each method by the vector

$$\mathcal{M}(\pi) = \left(\widehat{J}_r, \widehat{J}_c, \widehat{\rho}_{\text{viol}}, \widehat{\rho}_{\text{succ}}, \widehat{\rho}_{\text{adm}}, \widehat{\rho}_{\text{fall}}\right), \tag{11}$$

where $\widehat{J}_r$ and $\widehat{J}_c$ are empirical episode return and cost, $\widehat{\rho}_{\text{viol}}$ and $\widehat{\rho}_{\text{succ}}$ are violation and success frequencies, and $\widehat{\rho}_{\text{adm}}$ and $\widehat{\rho}_{\text{fall}}$ are admission/fallback rates available only for admission-layer methods. CQL+Safety denotes a CQL proposal with the local safety penalty; No-Expansion freezes the offline proposal. The experiments report a Pareto diagnostic, not a single-score leaderboard.

Table 1: Untouched procedural test family (30 controller clusters, 4,800 episodes per arm). Violation is cumulative cost $> 1$.

| Policy | Success | Violation |
|---|---|---|
| Calibration-qualified selective (.35) | .984 | .020 |
| Robust-only | .691 | .042 |
| Agile-only | .894 | .401 |
| Forced proposal | .408 | .778 |
| Oracle selector | .985 | .018 |
| Previous gate (.52) | .970 | .045 |
| Selective − robust estimate | $+.293$ | $-.0219$ |
| Simultaneous relevant bound | lower $+.148$ | upper $-.0113$ |
| Acceptance boundary | $> 0$ | $\leq .020$ |

**Reproducibility statement.** The primary study is regenerated by `simulation/run_calibrated_selective_family.py`; its frozen design is the frozen protocol record and complete JSON is under `simulation/results/`. The trajectory study is regenerated by `simulation/run_trajectory_qualified_learned_score.py`. The public replication uses `simulation/execute_v3_shard.py` and `simulation/v3_aggregate.py`. The matched-selector extension uses `simulation/v4_matched_selector.py`, `simulation/v4_matched_selector_aggregate.py`, and `simulation/results/v4_matched_selector_aggregate.json`; the accompanying artifact record documents its 54-shard verification.

### 6.1 Calibration/test-separated selective admission

**Reported protocol.** Each controller parameterizes an 18-step progress CMDP with balanced low/medium/high hazard strata, low/high score-noise strata, and independently sampled score shift, goal, and action progress. Robust fallback is the default. Candidate thresholds $\{.20, .25, .30, .35, .40\}$ admit the proposal only below the noisy hazard score. On 24 calibration controllers with 120 paired episodes per arm, a threshold is eligible only if its controller-cluster one-sided lower success bound exceeds zero and upper violation-difference bound is at most the stated .02 margin. Bonferroni adjustment covers five thresholds and two endpoints ($\alpha = .05/10 = .005$ per bound); among eligible thresholds, the largest success lower bound selects .35. The final reported design uses family-keyed shock streams and evaluation seed 40401; this describes the evaluated design, not an independently authenticated chronology.

**Untouched test and inference.** The test family contains 30 newly generated controller environments and 160 paired episodes per arm. Six fixed policies use identical environment/controller budgets: calibrated selective, robust-only, agile-only, forced proposal, an oracle that observes latent hazard, and the previous .52 gate. Simultaneous one-sided test bounds use $\alpha = .05/2 = .025$ per endpoint. Table 1 reports the evaluated held-out execution; the paper does not independently authenticate chronology.

The reported held-out table satisfies both numerical criteria, and robust-only has .691 success. Of 48 threshold×hazard×noise regions in the separate pointwise audit, 14 fail its stated criterion. These are synthetic generator draws with an informative noisy score; they do not establish public-task or learned-score transport, and no abbreviated digest is offered as submission-visible authentication.

### 6.2 Trajectory qualification with a finite-data learned score

The learned-score trajectory study preserves the post-qualification boundary: only threshold .20 qualifies, achieves success .9830 and violation .0206 on 5,000 disjoint units, and records 199/3,996 unsafe qualification trajectories before any certificate. Matched-rate and transport controls expose the synthetic mechanism without validating realistic critics or shifted deployment. A separate full-component slice likewise does not establish fallback Pareto dominance. Appendix A preserves the complete protocols, numerical results, and limitations.

Table 2: Theory-to-public-evidence status. "No" means unestablished, not empirically false.

| Object | Certificate status | Valid public interpretation |
|---|---|---|
| Fixed-time proxy | Exchangeability: no; fallback bound: no | Gate diagnostics only; no Proposition 2 deployment claim. |
| Complete trajectory | Prospective qualification: no; retrospective set empty | Outcome-inspected refusal for the selected gates, not a certificate. |
| V4 mixer | Pre-treatment, deployable | Complete-policy gate-versus-mixer contrast conditional on frozen V3 calibration. |
| V5 shuffle | Post-treatment | Conditional schedule stress test, not timing causality or a deployable policy. |

Table 3: Gate-on minus exact-count shuffled placement. Intervals use checkpoint-pair $t$ inference. "Signs" counts positive violation contrasts; criterion is the frozen joint rule.

| Task | Gate–shuffle completion | Gate–shuffle violation | signs | criterion |
|---|---|---|---|---|
| PointGoal1 | +0.0710 [+0.0593, +0.0827] | +0.2372 [+0.2118, +0.2627] | 20/20 | no |
| PointButton1 | +0.0343 [+0.0172, +0.0513] | +0.0600 [+0.0290, +0.0910] | 11/14 | no |
| PointPush1 | +0.0345 [+0.0131, +0.0559] | +0.0198 [-0.0005, +0.0400] | 12/20 | no |

## 7 Public Safety-Gymnasium Audit: Pre-Treatment and Retrospective Comparators

**Checkpoint population.** Independently trained BC-All/BC-Safe actors (seeds 100–119, three public datasets) enter confirmation only after disjoint completion qualification and a counterfactual rejection interval within [.05,.95]. Of 60 pairs, 20 PointGoal, 14 PointButton, and 20 PointPush qualify. Inference concerns these selected pairs, not failed pairs, arbitrary training procedures, or a gate class.

**V4: pre-treatment deployable comparator.** For each checkpoint, the state-independent Bernoulli mix probability is fixed from prior V3 data; arm decisions depend only on assignment seed, reset, and step. On 200 fresh common resets per arm, V4 therefore estimates a complete-policy gate-versus-mixer contrast conditional on frozen calibration, not a timing mediation effect. Gate-on violation is higher by .2430, .0164, and .0375 on PointGoal, PointButton, and PointPush. The task-level checkpoint-pair $t$ intervals are marginal, not simultaneous across tasks; PointButton's unrounded lower endpoint is only $2.6580 \times 10^{-6}$. A two-sided sign-test sensitivity analysis with Holm correction supports an adverse effect only on PointGoal, so we make no family-wise all-task significance claim. No task passes the frozen joint rule. Full rows are in Table 6.

**V5: retrospective conditional schedule stress test.** For each pair, 200 common resets compare gate-on with a SHA-256-seeded uniform permutation of its realized fallback mask. All 10,800 reset pairs match fallback count and administrative horizon exactly. The mask is observed on the gate trajectory before permutation, so V5 is post-treatment: it neither defines a deployable matched policy nor identifies a causal timing effect. Success is native completion, violation is cost above 10, and the frozen joint rule requires a negative upper 95% violation endpoint and completion lower endpoint above $-.05$. Appendix B gives the exact estimand and termination rule.

**Results.** At the frozen original threshold, no task passes the joint rule. Gate/shuffle completion rates are .9805/.9095 on PointGoal, .7932/.7589 on PointButton, and .6695/.6350 on PointPush; corresponding violation rates are .7560/.5188, .5854/.5254, and .4945/.4748. The checkpoint-$t$ violation interval is strictly positive on Goal and Button and includes zero on Push. The distribution-free sign sensitivity preserves the adverse conclusion only for PointGoal; Button and Push do not survive Holm correction. Large episode counts are not treated as independent replication.

**Interpretation.** Matching exact realized fallback count removes action-dose and administrative horizon differences, but intentionally changes placement and induced trajectories. The observed

schedule does not pass the frozen joint rule against its assigned count shuffle on the qualified population. This does not identify whether score error, state-distribution feedback, or another schedule explains the trade-off.

**Retrospective trajectory-certificate refusal.** We applied Proposition 3's arithmetic without tuning to the 200 gate-on rows per checkpoint from each of V3, V4, and V5: $\alpha = .10$, $L = 54$, uniform weights, $n = 600$, $D = 10$, $\gamma = 1$, and $T \leq 1000$. Rank $\lceil 601(1 - .1/54) \rceil = 600$ is feasible, but all 54 order statistics exceed 10 (range 134–810), so $\mathcal{Q} = \varnothing$. A V4-only gate-by-gate sensitivity ($L = 1, n = 200$, rank 181) also refuses 54/54 (range 54–105), but is not a simultaneous family claim. This is an outcome-inspected diagnostic: cross-campaign exchangeability was neither prespecified nor established, the population was selected by V3 completion/rejection, and cumulative rows require $\gamma = 1$. It supplies no untouched test, no estimate of $\Pr(\mathcal{Q} \neq \varnothing)$, and no safety falsification. The deterministic recomputation and input hashes are in `simulation/retrospective_trajectory_refusal.py` and `simulation/results/retrospective_trajectory_refusal.json`.

**Component diagnostic.** V6, calibrated on disjoint episodes, establishes no clean safety–efficiency decomposition; no-vote decisions coincide with the full gate. Appendix P preserves all adverse and null contrasts. These selected-pair results imply neither deployment safety nor a universal switching-policy ranking.

## 8 Conclusion

For one M=1 neural gate on 54 selected BC checkpoint pairs, the strongest public conclusion is refusal: neither the deployable V4 comparison nor the post-treatment V5 stress test passes the frozen joint criterion, and the retrospective trajectory calculation qualifies no gate. Proposition 3 remains a valid next-unit result under its own complete-unit exchangeability premise, distinct from LTT's high-confidence population-risk target. The remaining limits are substantive: no public prospective trajectory certificate, no established fixed-time fallback or transport premise, no executed competitive learned switching or recovery baseline, and no inference beyond three Point tasks and the selected checkpoints.

## References

P. Nakamoto, Y. Zhai, A. Singh, M. Sobol Mark, Y. Ma, C. Finn, A. Kumar, and S. Levine. Cal-QL: Calibrated Offline RL Pre-Training for Efficient Online Fine-Tuning. NeurIPS, 2023.

Z. Liu, Z. Guo, H. Lin, Y. Yao, J. Zhu, Z. Cen, H. Hu, W. Yu, T. Zhang, J. Tan, and D. Zhao. Datasets and Benchmarks for Offline Safe Reinforcement Learning. Journal of Data-centric Machine Learning Research, 2024.

A. N. Angelopoulos and S. Bates. A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification. Foundations and Trends in Machine Learning, 16(4):494–591, 2023.

A. N. Angelopoulos, S. Bates, E. J. Candès, M. I. Jordan, and L. Lei. Learn then Test: Calibrating Predictive Algorithms to Achieve Risk Control. Annals of Applied Statistics, 19(2), 2025. DOI: `https://doi.org/10.1214/24-AOAS1998`.

D. Prinster, C. Fannjiang, J. W. Park, K. Cho, A. Liu, S. Saria, and S. Stanton. Conformal Policy Control. International Conference on Machine Learning (ICML), 2026. arXiv v4: `https://arxiv.org/abs/2603.02196v4`.

A. N. Angelopoulos, S. Bates, A. Fisch, L. Lei, and T. Schuster. Conformal Risk Control. International Conference on Learning Representations (ICLR), 2024. `https://arxiv.org/abs/2208.02814`.

S. R. Howard, A. Ramdas, J. McAuliffe, and J. Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. Annals of Statistics, 49(2):1055–1080, 2021.

V. Vovk and R. Wang. E-values: Calibration, combination, and applications. Annals of Statistics, 49(3):1736–1754, 2021.

I. Gibbs and E. Candès. Adaptive Conformal Inference Under Distribution Shift. NeurIPS, 2021.

L. Lindemann, M. Cleaveland, G. Shim, and G. J. Pappas. Safe Planning in Dynamic Environments using Conformal Prediction. IEEE Robotics and Automation Letters, 8(8):5116–5123, 2023.

Y. Geifman and R. El-Yaniv. Selective Classification for Deep Neural Networks. NeurIPS, 2017.

H. Xu, X. Zhan, and X. Zhu. Constraints Penalized Q-Learning for Safe Offline Reinforcement Learning. AAAI, 2022.

J. Lee, C. Paduraru, D. J. Mankowitz, N. Heess, D. Precup, K.-E. Kim, and A. Guez. COptiDICE: Offline Constrained Reinforcement Learning via Stationary Distribution Correction Estimation. ICLR, 2022.

C. Cao, Y. Xin, S. Wu, L. He, Z. Yan, J. Tan, and X. Wang. FOSP: Fine-tuning Offline Safe Policy through World Models. International Conference on Learning Representations (ICLR), 2025. URL: https://openreview.net/pdf?id=dbuFJg7eaw.

Y. Chemingui, A. Deshwal, H. Wei, A. Fern, and J. R. Doppa. Constraint-Adaptive Policy Switching for Offline Safe Reinforcement Learning. Proceedings of the AAAI Conference on Artificial Intelligence, 39(15):15722–15730, 2025.

N. S. Neggatu, J. Houssineau, and G. Montana. Evaluation-Time Policy Switching for Offline Reinforcement Learning. Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS), 2025.

B. Thananjeyan, A. Balakrishna, S. Nair, M. Luo, K. Srinivasan, M. Hwang, J. E. Gonzalez, J. Ibarz, C. Finn, and K. Goldberg. Recovery RL: Safe Reinforcement Learning with Learned Recovery Zones. IEEE Robotics and Automation Letters, 6(3):4915–4922, 2021.

N. Wagener, B. Boots, and C.-A. Cheng. Safe Reinforcement Learning Using Advantage-Based Intervention. NeurIPS, 2021. https://arxiv.org/abs/2106.09110.

K. P. Wabersich and M. N. Zeilinger. A Predictive Safety Filter for Learning-Based Control of Constrained Nonlinear Dynamical Systems. Automatica, 129:109597, 2021. https://arxiv.org/abs/1812.05506.

R. Laroche, P. Trichelair, and R. Tachet des Combes. Safe Policy Improvement with Baseline Bootstrapping. ICML, 2019. https://proceedings.mlr.press/v97/laroche19a.html.

R. J. Tibshirani, R. Foygel Barber, E. Candès, and A. Ramdas. Conformal Prediction Under Covariate Shift. NeurIPS, 2019.

# APPENDIX

## A DETAILED SYNTHETIC VALIDATION PROTOCOLS

We fit an affine least-squares hazard score from 4,000 independent labeled synthetic transitions, freeze it, and compare robust-only with learned-score thresholds $.20, .35, .50$. Each complete policy includes all state feedback and fallback calls. Following Proposition 3, $L = 4$, familywise $\alpha = .10$, $n = 999$ independent calibration trajectories, $T = 18$, $\gamma = .99$, and budget $D = 1$ give per-policy level $.025$. Qualification and selection use calibration data only; 5,000 disjoint trajectory units estimate performance. This post-review controlled study was not preregistered and supplies synthetic mechanism evidence only.

Only threshold $.20$ qualifies: its trajectory-cost quantile is $.9105 \leq 1$; robust-only, $.35$, and $.50$ have quantiles $1.1075, 1.1611, 2.0000$. The selected policy obtains test success $.9830$ and budget violation $.0206$, versus $.6306$ and $.0282$ for robust-only. Paired unit-level Wald intervals give success difference $.3524$ (95% CI $[.3390, .3658]$) and violation difference $-.0076$ ($[-.0114, -.0038]$).

The qualification ledger itself contains 29, 24, 38, and 108 budget violations for robust-only and thresholds $.20, .35, .50$: 199/3,996 trajectories are unsafe by the study endpoint before any certificate exists. Proposition 5 makes this temporal boundary explicit.

A calibration-only grid selects a random switcher with matched test proposal rate $.7335$ versus $.7308$ for the learned gate. On the same 5,000 paired units, random switching has $.4762$ success and $.7116$ violation; learned-minus-random differences are $.5068$ (95% CI $[.4929, .5207]$) and $-.6910$ ($[-.7038, -.6782]$). A latent-hazard oracle has $.9830$ success and $.0190$ violation. The matched-rate result rejects this state-independent random-switching comparator, but does not isolate score information from all timing- or state-dependent switching policies. Near-oracle performance is not evidence of realistic critic learning. These comparisons quantify the synthetic mechanism, not the conformal guarantee.

A post-review, non-preregistered nested-family audit uses the same frozen score and calibration generator. With uniform weights, family sizes $L = 1, 2, 4, 8, 16$ have ranks $900, 950, 975, 988, 994$; the $.20$ gate is selected for $L = 1, 2, 4$, while no policy qualifies for $L = 8, 16$. The exact feasibility ceiling is $L = 100$ for $n = 999, \alpha = .10$, but feasibility of a rank does not imply qualification. Composition therefore changes the decision well before the arithmetic ceiling.

Two sensor interventions probe transport without refitting. Under a $+.18$ offset, the policy remains qualified (quantile $.8973$; success $.9822$, violation $.0234$). Under inversion, qualification fails (quantile $3.8452$; success $.2260$, violation $.8796$). Nominal qualification cannot be transported to the inverted law.

**Finite full-component execution audit.** An aggregate finite synthetic check reports 2/199 calibration budget violations before policy qualification ($q = .1$ trajectory quantile $1.641 \leq D = 2.5$), preserving the temporal limitation. Because traces and per-mechanism counts are not included, this result does not substantiate execution frequency or coverage for every Section 4 component.

On 500 paired test units, the full slice has success $.994$ and violation $.006$ versus $.912$ and $.016$ for fallback-only; differences are $.082$ (95% CI $[.057, .107]$) and $-.010$ ($[-.0217, .0017]$). Thus simultaneous fallback Pareto dominance is not established. Forced proposal has zero success and unit violation. This is execution coverage on a hand-specified finite process, not validation of neural critics, public efficacy, qualification-rollout safety, or external relevance. The runner is `simulation/run_full_admission_slice.py`.

## B EXACT-COUNT SCHEDULE PROTOCOL DETAIL

Each qualified pair receives 200 new common reset seeds. Gate-on first records its complete binary fallback mask. The environment is reset to the same seed, and a SHA-256-seeded uniform permutation of that mask supplies the count-shuffled arm. If the gate rollout has length $L$, the shuffled rollout is administratively capped at $L$; any natural termination before $L$ fails the exact-match gate. All 10,800 pairs execute exactly $L$ steps and the same fallback count. The shuffled placement is

Table 4: Executed controlled study. Each arm has 960 episodes in 16 controller clusters. "Viol." is $\Pr(\text{cost} > 1)$; NR is normalized return.

| Policy arm | Success | Viol. | NR | Cost | Fallback |
|---|---|---|---|---|---|
| Gate off (robust label) | 0.336 | 0.802 | 0.309 | 1.730 | 0.000 |
| Gate off (agile label) | 0.356 | 0.779 | 0.337 | 1.667 | 0.000 |
| Gate on + robust | 0.797 | 0.261 | 0.763 | 0.802 | 0.199 |
| Gate on + agile | 0.804 | 0.349 | 0.770 | 0.931 | 0.199 |
| Fallback only, robust | 0.954 | 0.107 | 0.843 | 0.585 | 1.000 |
| Fallback only, agile | 0.888 | 0.311 | 0.825 | 0.857 | 1.000 |
| Gate ITT difference | $+0.454$ | $-0.485$ | $+0.444$ | $-0.832$ | – |
| 95% cluster CI | $[.424, .484]$ | $[-.515, -.456]$ | $[.412, .476]$ | $[-.874, -.790]$ | – |

assigned without consulting the shuffled rollout's state, score, reward, cost, or outcome; those variables may depend on the assigned schedule. Success uses the environment's native completion field; violation is episode cost above 10. The frozen joint benefit rule requires the upper 95% endpoint of gate-minus-shuffle violation below zero and the lower completion endpoint above $-.05$.

**Estimand.** For checkpoint $j$ and reset $e$, let $M_{je}$ and $L_{je}$ be the mask and length realized by gate-on, and let $\Pi_{je}$ be the pre-seeded uniform permutation applied after resetting the environment. For endpoint $Y$, we report checkpoint means of

$$Y_{je}(\text{gate}) - Y_{je}(\Pi_{je}(M_{je}); L_{je}) .$$

This finite randomized schedule contrast asks whether the observed gate mask outperforms its assigned permutation at equal dose and horizon. Because $M_{je}$ is generated on the first trajectory and actions alter later states, it is not a pre-treatment causal estimand for "timing alone" and the permuted schedule is not available before observing gate-on.

## C Matched episode-randomized gate-effect study

**Process, randomization, and endpoint.** A stochastic 22-step progress task removes the earlier competence floor. Six arms cross gate enforcement with robust/agile fallback labels and add both fallback-only controls; 16 controller clusters contribute 60 randomly assigned episodes per arm. Success is goal attainment without catastrophe and violation is cumulative cost $> 1$. The primary ITT contrast averages gate-on cells minus duplicate randomized proposal controls; fallback-specific contrasts are identified, but a fallback main effect is not identified when the gate is off. Inference uses 16 controller contrasts and $t_{15}$ intervals. Appendix F specifies the final generator, seed, and endpoint used for these reported outcomes; it does not authenticate design chronology.

The proposal-only cells complete 34.6% pooled, and fallback-only cells complete 95.4%/88.8%; these are descriptive rates, not a prespecified competence classification. Primary controller SEs are .0142 (success) and .0140 (violation), and leave-one-controller-out estimates remain .449–.460 and $-.492$–$-.480$. Fallback-specific success effects are $+.460$ (CI .418–.503) robust and $+.448$ (.408–.488) agile; violation effects are $-.541$ (CI $-.581$–$-.500$) and $-.430$ ($-.462$–$-.399$). Against fallback-only, however, gated success is $-.157$ (CI $-.189$–$-.126$) robust and $-.083$ ($-.109$–$-.057$) agile; violation is $+.154$ (.119–.189) and $+.038$ ($-.013$–.088). Thus no measured success–cost benefit over either fallback-only arm is established. Episode budget-violation rates for gate-on robust/agile (.261/.349) and fallback-only robust/agile (.107/.311), together with gate-on fallback rates (.199/.199), are deployment diagnostics, not proxy, sequential, or trajectory guarantees; the simulator did not log a separate realized $Y > d$ stream.

**Adverse robustness audit.** We additionally crossed base hazard probability $\{.08, .13, .18\}$, gate threshold $\{.35, .52, .70\}$, and proposal hazard-failure probability $\{.35, .52, .68\}$, retaining 16 controllers and 60 episodes per arm/controller in each of 27 cells. Gate-on minus forced-proposal success ranges $+.137$–$+.605$ and violation ranges $-.567$–$-.278$; both directions improve in 27/27 cells. Yet gate-on success remains below robust fallback by .034–.391 and below agile fallback by .002–.303. The gate Pareto-dominates robust or agile fallback in 0/27 cells each. This grid tests three hand-constructed parameters, not learned scores, realistic shift, or external tasks. Full cell rows are in `simulation/results/final_meta_robustness.csv`.

### C.1 FROZEN PUBLIC-TASK REPLICATION AND IDENTIFIABILITY BOUNDARY

The earlier eight-task study completes all 120 cells, 2,400 episodes, and 1.2M decisions, but fallback-only, forced-admit, and gated arms each have success 0/800; its completion endpoint cannot attribute a gate effect. Two 3-seed proposal packages also reverse the cost direction (223.467 to 156.533 versus .200 to 93.667), but proposal and training change, so this is not an interaction estimate.

A native Safety-Gymnasium PointGoal protocol instead fixes `info["goal_met"]`, three 10,000-update BC-All/BC-Safe checkpoint pairs, and a one-candidate gate. V1 passes competence but rejects only .540%–1.546%, so its 5% floor prohibits confirmation. The paper-local protocol record designates V2 as frozen before outcome inspection, but is not external authentication of that chronology. V2 changes only the scalar threshold using offline folds to target 20% rejection; checkpoints, endpoint, assignment, and reset schedules are recorded as fixed. On disjoint qualification resets, proposal/fallback complete 146/150 and 130/150 and all three rejection intervals lie inside [.05,.95]. Thus 1,200 block-randomized confirmation episodes execute: gate-on uses 424,950 proposal and 175,050 fallback decisions, while gate-off uses 600,000 proposal decisions.

Both randomized arms complete .9867 of episodes. Gate-on minus gate-off completion is 0 (checkpoint $t$ 95% CI $[-.02151, .02151]$, randomization $p = 1$). Violation is .7817 versus .7967, a difference of $-.0150$ ($[-.06914, .03914]$, $p = .567$); episode cost is 35.297 versus 39.500, a difference of $-4.203$ ($[-9.694, 1.287]$). Mean return is 19.026 versus 19.661. The disjoint 300-episode fallback-only baseline has completion .8833, violation .3600, cost 9.877, and return 8.769, exposing a utility–safety trade-off rather than dominance. The preregistered rule required the violation interval below zero and the completion interval above $-.05$; only the latter holds, so `beneficial_gate_decision=false`. This is an identified public learned gate-on/off null result, not evidence of benefit.

## D V5 EXACT-COUNT SCHEDULE STRESS RESULTS

Table 5: V5 retrospective schedule stress test. Differences are gate-on minus a uniform permutation of the gate arm's realized fallback mask, with checkpoint-pair $t$ intervals. Every reset matches fallback count and administrative horizon. The criterion requires violation's upper endpoint below zero and completion's lower endpoint above $-.05$.

| Task | Gate–shuffle completion | Gate–shuffle violation | signs | criterion |
|---|---|---|---|---|
| PointGoal1 | +0.0710 [+0.0593, +0.0827] | +0.2372 [+0.2118, +0.2627] | 20/20 | no |
| PointButton1 | +0.0343 [+0.0172, +0.0513] | +0.0600 [+0.0290, +0.0910] | 11/14 | no |
| PointPush1 | +0.0345 [+0.0131, +0.0559] | +0.0198 [-0.0005, +0.0400] | 12/20 | no |

All 54 V3-qualified checkpoint pairs contribute 200 fresh matched reset pairs: 10,800 exact count/length matches and 21,600 episode records. The population verifier recomputes every arm pair and schedule hash; the 162-file shard ledger and compact task results are bound in the accompanying artifact record. The conditional permutation compares placements at the gate arm's realized intervention count. Because its mask is post-treatment, it neither identifies a timing-only causal effect nor defines a deployable policy.

## E V4 MATCHED-SELECTOR RESULTS

All 54 V3-qualified checkpoint pairs contribute 200 fresh matched reset seeds per arm. The complete 43,200-episode ledger, protocol and source hashes, and nine-file runtime closure are verified in the accompanying artifact record. The matched rate is an expected action-level fallback fraction, not exact realized action count. No cross-task heterogeneity or all-switching-policy claim is made.

Table 6: V4 matched-selector test. Differences are gate-on minus state-independent rate-matched mixing with checkpoint-pair $t$ intervals. The criterion requires violation's upper endpoint below zero and completion's lower endpoint above $-.05$.

| Task | Gate–random completion | Gate–random violation | benefit |
|---|---|---|---|
| PointGoal1 | +0.0797 [+0.0685, +0.0910] | +0.2430 [+0.2097, +0.2763] | no |
| PointButton1 | -0.0096 [-0.0234, +0.0041] | +0.0164 [+0.0000, +0.0329] | no |
| PointPush1 | +0.0295 [+0.0157, +0.0433] | +0.0375 [+0.0173, +0.0577] | no |

## F CONTROLLED-STUDY GENERATOR AND RANDOMIZATION

For controller $j$, $H_{jt} \sim \text{Bernoulli}(p_j)$ with $p_j = \text{clip}(.13 + N(0, .018^2), .08, .18)$ and score $X_{jt} = .22 + .56H_{jt} + N(0, \sigma_j^2)$, $\sigma_j = \text{clip}(.25 + N(0, .025^2), .18, .32)$; the gate rejects above .52. Proposal progress is $\max\{.5, v_j + N(0, .12^2)\}$, $v_j = \text{clip}(2 + N(0, .08^2), 1.75, 2.25)$, with hazard failure .68. Robust/agile fallback progress is $\max\{.3, 1.18 + N(0, .08^2)\}$ / $\max\{.4, 1.52 + N(0, .10^2)\}$ and hazard failure $.015/.055$; safe-state failure is $.001/.002$ versus .003 for proposals. Step costs are $.18/.35$ on fallback hazards and $.01/.015$ otherwise, versus $1/.02$ for proposal hazard/safe states; catastrophe adds 1. Return is progress $-81[\text{catastrophe}] - .5\,\text{cost} - .15\,\text{steps}$ and normalized return divides by 20. Controller, assignment, and episode streams are SHA-256 derivations of seed 40040 and their indices. Within each controller, a seeded shuffle assigns 60 episodes to each of six arms. We average within controller and arm before forming contrasts.

## G EXACT PROCEDURAL-GENERATOR SPECIFICATIONS

**Calibration/test-separated family.** Controller $j$ uses hazard bin $(.08, .14, .20)_{j \bmod 3} + U(0, .025)$ and score-noise bin $(.10, .16)_{\lfloor j/3 \rfloor \bmod 2} + U(0, .025)$, with independent score shift $U(-.025, .025)$, goal $U(16.15, 17.55)$, and robust/agile/proposal progress means $U(.90, 1.00)$, $U(1.20, 1.34)$, and $U(1.68, 1.86)$. At each of 18 steps, $H \sim \text{Bernoulli}(p_j)$ and $X = (.84 \text{ if } H \text{ else } .16) + \text{shift} + N(0, \sigma_j^2)$. Proposal/robust/agile progress has Gaussian standard deviation $.08/.05/.06$ and floors $.4/.25/.3$. Their safe/hazard costs are $(.015, 1)/(.008, .14)/(.012, .34)$; safe/hazard failure probabilities are $(.001, .58)/(.0005, .004)/(.0015, .045)$; failure adds cost 1. Success is reaching the goal before failure and violation is total cost $> 1$. Environments use the first eight bytes of SHA-256 on `seed|environment|controller`; episodes use `episode|family-seed|controller|episode`. Calibration seed 40101 has 24 controllers and 120 episodes; untouched seed 40401 has 30 and 160. For controller differences $d_j$, every reported one-sided bound is $\bar{d} \pm t_{m-1,1-\alpha} s_d/\sqrt{m}$: calibration uses $(m, \alpha, t) = (24, .005, 2.8073357)$ for each of ten threshold–endpoint bounds, and test uses $(30, .025, 2.0452296)$ for each of two endpoints. Threshold selection and the .02 noninferiority margin are frozen in the named preregistration artifact.

**Learned-score trajectory family.** Training draws 4,000 independent $p \sim U(.08, .225)$, $H \sim \text{Bernoulli}(p)$, and $X = (.84 \text{ if } H \text{ else } .16) + N(0, .17^2)$ from SHA-256 key `learned-score|training`. Least squares of $H$ on $X$ gives frozen clipped score $\widehat{h}(x) = \text{clip}(-.1085379872 + 1.0011547118x, 0, 1)$. Complete-unit key `trajectory-unit|unit-id|shift` draws $p \sim U(.08, .225)$, $\sigma \sim U(.10, .185)$, goal $U(16.15, 17.55)$, and robust/proposal means $U(.90, 1.00)/U(1.68, 1.86)$. Progress floors, costs, and failure laws equal the robust/proposal values above; $T = 18, \gamma = .99, D = 1$. Calibration IDs are 0–998 and disjoint test IDs 100000–104999. Offset adds .18 to both sensor centers; inversion swaps .16 and .84. The coverage comparator chooses rates $.50, .51, \ldots, .90$ using calibration proposal frequency only, then draws each decision from SHA-256 key `random-gate|unit|shift|step`; the oracle proposes iff $H = 0$. These constants, fitted coefficients, split IDs, and rank convention are also serialized in the result JSON.

## H V3 PUBLIC IMPLEMENTATION AND COMPLETE LEDGERS

**Actors and checkpoint selection.** Each attempted pair independently trains a BC-All proposal and BC-Safe fallback with two 256-unit hidden layers, Adam at $10^{-3}$, batch size 512, and 10,000 updates on the task's public offline data. BC-Safe retains trajectories with cost at most 10; BC-All retains all trajectories. The optional `code/requirements-public-benchmark.txt` lists DSRL, OSRL and Safety-Gymnasium for Python 3.10/3.11 but does not pin their versions. Restoration lists exist separately; their binding to all 54 V3 pairs, the original data-release digest, and filtered-split hashes is not established here. Thus the complete historical environment and data transformation chain cannot be reconstructed from this account alone. `model_best.pt` is the lower-cost evaluation checkpoint, with reward as the tie-breaker, among evaluations at updates 5,000 and 10,000. Seeds 100–119 are fresh relative to V1/V2.

**Gate instantiated in V3.** For each checkpoint pair, 20,000 transitions train a three-member neural cost ensemble with hidden width 64 for 200 steps (batch 128); 5,000 disjoint transitions calibrate it. The instantiated settings are immediate-cost limit 1, CQL weight .05, disagreement threshold .20, vote threshold .80, $\kappa = 2$, conformal slack .02, target coverage .95, and behavior-support coverage .99. A single BC-All action is accepted only when the neural gate accepts and the nearest-neighbor support check does not reject; otherwise the BC-Safe action executes. V3 changes the scalar calibrated margin using offline data only to target 20% rejection. Thus the public experiment executes the ensemble score, disagreement, voting, conformal margin, support gate, and fallback, but has no multi-candidate ranking effect because $M = 1$.

**Qualification, assignment, and estimand.** Each pair receives 50 disjoint qualification episodes per actor. Proposal and fallback must each complete at least 5/50 episodes and have a one-sided Wilson lower bound above .05; the counterfactual rejection estimate and its episode-level 95% interval must both lie within [.05,.95]. Failed pairs run no confirmation episodes. Each qualified pair then receives 200 gate-on and 200 gate-off episodes randomized in blocks of 20 with common reset schedules, horizon 1,000, and episode-cost budget 10. The checkpoint pair is the inferential unit. A separate 100-episode fallback-only baseline uses a disjoint reset range and is descriptive, not a same-reset randomized contrast. The generated ledgers below show all attempted pairs; "–" denotes the six qualification failures that correctly produced no confirmation or baseline outcomes.

Table 7: V3 task-level absolute outcomes. Bundles are completion/violation/cost/return/fallback rate.

| Task | Gate on | Gate off | Fallback only |
|---|---|---|---|
| PointGoal1 | 0.980/0.761/35.72/18.45/0.279 | 0.983/0.766/36.60/19.09/0.000 | 0.850/0.385/11.50/8.91/1.000 |
| PointButton1 | 0.791/0.582/34.69/4.67/0.097 | 0.785/0.586/37.07/4.56/0.000 | 0.597/0.309/13.96/1.36/1.000 |
| PointPush1 | 0.666/0.493/23.41/3.04/0.159 | 0.703/0.540/25.57/3.24/0.000 | 0.411/0.249/10.59/1.41/1.000 |

Table 8: V3 PointGoal1 qualification ledger; every attempted pair is shown.

| Seed | Proposal | Fallback | Rejection | 95% interval | Result |
|---|---|---|---|---|---|
| 100 | 50/50 | 40/50 | 0.33562 | [0.31203, 0.35921] | pass |
| 101 | 49/50 | 42/50 | 0.33294 | [0.30567, 0.36021] | pass |
| 102 | 50/50 | 44/50 | 0.31616 | [0.29108, 0.34124] | pass |
| 103 | 49/50 | 44/50 | 0.33722 | [0.30687, 0.36757] | pass |
| 104 | 50/50 | 46/50 | 0.30416 | [0.28232, 0.32600] | pass |
| 105 | 47/50 | 44/50 | 0.25830 | [0.22462, 0.29198] | pass |
| 106 | 48/50 | 48/50 | 0.28246 | [0.24569, 0.31923] | pass |
| 107 | 50/50 | 45/50 | 0.29004 | [0.25783, 0.32225] | pass |
| 108 | 49/50 | 43/50 | 0.32176 | [0.29359, 0.34993] | pass |
| 109 | 48/50 | 44/50 | 0.27826 | [0.23844, 0.31808] | pass |
| 110 | 50/50 | 47/50 | 0.29968 | [0.27664, 0.32272] | pass |
| 111 | 50/50 | 44/50 | 0.28290 | [0.25191, 0.31389] | pass |
| 112 | 48/50 | 42/50 | 0.22516 | [0.20058, 0.24974] | pass |
| 113 | 48/50 | 39/50 | 0.26356 | [0.22630, 0.30082] | pass |
| 114 | 47/50 | 40/50 | 0.27520 | [0.24403, 0.30637] | pass |
| 115 | 50/50 | 40/50 | 0.35014 | [0.31882, 0.38146] | pass |
| 116 | 49/50 | 45/50 | 0.30698 | [0.27616, 0.33780] | pass |
| 117 | 50/50 | 40/50 | 0.35894 | [0.33605, 0.38183] | pass |
| 118 | 50/50 | 45/50 | 0.28350 | [0.25441, 0.31259] | pass |
| 119 | 49/50 | 45/50 | 0.29196 | [0.26161, 0.32231] | pass |

Table 9: V3 PointGoal1 absolute outcomes. Bundles are completion/violation/cost/return/fallback rate.

| Seed | Gate on | Gate off | Fallback only |
|---|---|---|---|
| 100 | 0.965/0.735/34.80/18.61/0.305 | 0.995/0.785/37.56/20.37/0.000 | 0.870/0.370/8.85/9.43/1.000 |
| 101 | 0.985/0.765/38.54/19.10/0.300 | 1.000/0.790/35.94/20.01/0.000 | 0.880/0.440/12.37/8.90/1.000 |
| 102 | 0.995/0.825/39.17/20.08/0.302 | 0.995/0.840/44.85/20.01/0.000 | 0.910/0.430/14.62/9.94/1.000 |
| 103 | 0.985/0.795/38.15/18.49/0.300 | 0.990/0.805/38.48/19.20/0.000 | 0.830/0.380/11.66/8.05/1.000 |
| 104 | 0.985/0.840/40.84/19.40/0.292 | 0.985/0.845/43.19/19.80/0.000 | 0.880/0.430/15.24/11.38/1.000 |
| 105 | 0.970/0.730/33.29/17.67/0.243 | 0.995/0.755/35.70/19.42/0.000 | 0.850/0.400/12.20/9.38/1.000 |
| 106 | 0.970/0.765/36.44/18.82/0.284 | 0.955/0.665/31.83/17.92/0.000 | 0.880/0.430/14.60/10.94/1.000 |
| 107 | 0.990/0.790/37.67/18.71/0.278 | 0.990/0.700/33.09/18.82/0.000 | 0.820/0.290/7.26/6.92/1.000 |
| 108 | 0.970/0.730/33.37/17.06/0.270 | 0.990/0.705/34.62/18.93/0.000 | 0.900/0.400/10.51/9.32/1.000 |
| 109 | 0.960/0.725/33.84/17.58/0.278 | 0.970/0.760/36.78/18.77/0.000 | 0.830/0.370/8.90/8.90/1.000 |
| 110 | 0.980/0.710/30.30/18.58/0.252 | 0.995/0.765/34.09/19.52/0.000 | 0.910/0.470/14.06/9.45/1.000 |
| 111 | 0.995/0.740/35.94/19.09/0.268 | 0.985/0.795/38.88/19.78/0.000 | 0.800/0.320/8.43/8.95/1.000 |
| 112 | 0.975/0.755/38.89/18.51/0.249 | 0.970/0.780/35.71/18.72/0.000 | 0.810/0.310/7.35/6.98/1.000 |
| 113 | 0.990/0.740/31.25/17.91/0.266 | 0.970/0.750/33.30/16.76/0.000 | 0.850/0.460/14.75/9.46/1.000 |
| 114 | 0.965/0.760/32.87/17.24/0.260 | 0.970/0.720/33.59/17.84/0.000 | 0.810/0.280/8.72/6.85/1.000 |
| 115 | 1.000/0.835/40.27/18.81/0.315 | 0.970/0.785/40.18/18.74/0.000 | 0.820/0.400/12.84/7.30/1.000 |
| 116 | 0.975/0.645/27.79/17.81/0.266 | 0.985/0.730/32.70/18.75/0.000 | 0.810/0.470/17.24/11.32/1.000 |
| 117 | 0.985/0.785/35.08/18.46/0.287 | 0.985/0.800/38.66/19.38/0.000 | 0.840/0.340/9.76/7.56/1.000 |
| 118 | 0.985/0.805/41.81/19.03/0.294 | 0.985/0.790/37.73/19.66/0.000 | 0.890/0.400/10.70/9.06/1.000 |
| 119 | 0.970/0.740/34.12/18.11/0.276 | 0.985/0.765/35.18/19.32/0.000 | 0.810/0.310/9.92/8.18/1.000 |

Table 10: V3 PointButton1 qualification ledger; every attempted pair is shown.

| Seed | Proposal | Fallback | Rejection | 95% interval | Result |
|---|---|---|---|---|---|
| 100 | 43/50 | 26/50 | 0.09120 | [0.06285, 0.11955] | pass |
| 101 | 42/50 | 29/50 | 0.10346 | [0.07306, 0.13386] | pass |
| 102 | 39/50 | 32/50 | 0.09206 | [0.05963, 0.12449] | pass |
| 103 | 36/50 | 31/50 | 0.07484 | [0.04900, 0.10068] | gate_seed_0_rejection_interval_outside_0.05_0.95 |
| 104 | 35/50 | 35/50 | 0.05178 | [0.02708, 0.07648] | gate_seed_0_rejection_interval_outside_0.05_0.95 |
| 105 | 37/50 | 31/50 | 0.08790 | [0.05527, 0.12053] | pass |
| 106 | 42/50 | 33/50 | 0.12156 | [0.08127, 0.16185] | pass |
| 107 | 38/50 | 28/50 | 0.07708 | [0.04562, 0.10854] | gate_seed_0_rejection_interval_outside_0.05_0.95 |
| 108 | 39/50 | 35/50 | 0.12942 | [0.09417, 0.16467] | pass |
| 109 | 42/50 | 33/50 | 0.09048 | [0.06553, 0.11543] | pass |
| 110 | 35/50 | 33/50 | 0.06344 | [0.03958, 0.08730] | gate_seed_0_rejection_interval_outside_0.05_0.95 |
| 111 | 36/50 | 25/50 | 0.13186 | [0.09447, 0.16925] | pass |
| 112 | 44/50 | 33/50 | 0.09936 | [0.07094, 0.12778] | pass |
| 113 | 37/50 | 31/50 | 0.09142 | [0.06362, 0.11922] | pass |
| 114 | 38/50 | 27/50 | 0.09932 | [0.06795, 0.13069] | pass |
| 115 | 37/50 | 34/50 | 0.07868 | [0.04772, 0.10964] | gate_seed_0_rejection_interval_outside_0.05_0.95 |
| 116 | 39/50 | 29/50 | 0.05470 | [0.03079, 0.07861] | gate_seed_0_rejection_interval_outside_0.05_0.95 |
| 117 | 39/50 | 30/50 | 0.09914 | [0.06800, 0.13028] | pass |
| 118 | 40/50 | 27/50 | 0.10382 | [0.07285, 0.13479] | pass |
| 119 | 42/50 | 35/50 | 0.08830 | [0.05902, 0.11758] | pass |

Table 11: V3 PointButton1 absolute outcomes. Bundles are completion/violation/cost/return/fallback rate.

| Seed | Gate on | Gate off | Fallback only |
|---|---|---|---|
| 100 | 0.755/0.575/32.32/4.68/0.097 | 0.790/0.605/39.30/5.65/0.000 | 0.690/0.320/16.74/2.39/1.000 |
| 101 | 0.810/0.575/35.60/5.46/0.089 | 0.835/0.630/42.91/5.59/0.000 | 0.570/0.310/13.85/0.52/1.000 |
| 102 | 0.830/0.645/32.90/5.43/0.087 | 0.825/0.545/36.62/5.14/0.000 | 0.540/0.260/9.06/0.02/1.000 |
| 103 | – | – | – |
| 104 | – | – | – |
| 105 | 0.855/0.670/42.92/5.94/0.104 | 0.815/0.605/39.72/4.60/0.000 | 0.630/0.290/12.38/1.57/1.000 |
| 106 | 0.735/0.590/37.78/3.53/0.094 | 0.750/0.560/34.82/3.74/0.000 | 0.720/0.410/20.59/3.82/1.000 |
| 107 | – | – | – |
| 108 | 0.825/0.605/36.09/5.28/0.107 | 0.810/0.670/41.94/5.92/0.000 | 0.620/0.320/12.49/2.42/1.000 |
| 109 | 0.720/0.465/25.93/3.21/0.090 | 0.785/0.570/43.51/3.73/0.000 | 0.580/0.330/13.49/1.18/1.000 |
| 110 | – | – | – |
| 111 | 0.810/0.595/34.61/3.93/0.098 | 0.800/0.600/34.51/4.01/0.000 | 0.560/0.280/15.88/1.22/1.000 |
| 112 | 0.760/0.610/33.28/4.08/0.092 | 0.780/0.585/35.50/5.03/0.000 | 0.600/0.320/12.35/1.41/1.000 |
| 113 | 0.755/0.570/35.23/3.33/0.107 | 0.710/0.565/31.61/2.00/0.000 | 0.580/0.260/15.62/-0.11/1.000 |
| 114 | 0.805/0.525/33.66/4.64/0.100 | 0.805/0.555/31.09/4.17/0.000 | 0.520/0.230/8.09/-1.17/1.000 |
| 115 | – | – | – |
| 116 | – | – | – |
| 117 | 0.820/0.580/37.30/5.49/0.101 | 0.745/0.575/35.37/4.98/0.000 | 0.570/0.310/14.70/1.25/1.000 |
| 118 | 0.780/0.575/33.66/4.38/0.088 | 0.780/0.535/36.83/4.06/0.000 | 0.540/0.300/11.33/2.36/1.000 |
| 119 | 0.810/0.565/34.32/6.07/0.102 | 0.755/0.600/35.28/5.28/0.000 | 0.640/0.380/18.88/2.13/1.000 |

Table 12: V3 PointPush1 qualification ledger; every attempted pair is shown.

| Seed | Proposal | Fallback | Rejection | 95% interval | Result |
|---|---|---|---|---|---|
| 100 | 29/50 | 15/50 | 0.13990 | [0.11356, 0.16624] | pass |
| 101 | 37/50 | 27/50 | 0.20046 | [0.16310, 0.23782] | pass |
| 102 | 41/50 | 20/50 | 0.17044 | [0.14333, 0.19755] | pass |
| 103 | 34/50 | 18/50 | 0.19164 | [0.16535, 0.21793] | pass |
| 104 | 33/50 | 20/50 | 0.15808 | [0.13201, 0.18415] | pass |
| 105 | 32/50 | 20/50 | 0.17892 | [0.14725, 0.21059] | pass |
| 106 | 37/50 | 14/50 | 0.20262 | [0.17844, 0.22680] | pass |
| 107 | 36/50 | 21/50 | 0.16164 | [0.13538, 0.18790] | pass |
| 108 | 37/50 | 21/50 | 0.18176 | [0.15229, 0.21123] | pass |
| 109 | 32/50 | 23/50 | 0.15886 | [0.12817, 0.18955] | pass |
| 110 | 32/50 | 22/50 | 0.16968 | [0.13848, 0.20088] | pass |
| 111 | 39/50 | 20/50 | 0.17418 | [0.15021, 0.19815] | pass |
| 112 | 40/50 | 18/50 | 0.19108 | [0.16617, 0.21599] | pass |
| 113 | 33/50 | 18/50 | 0.17586 | [0.14619, 0.20553] | pass |
| 114 | 37/50 | 17/50 | 0.22486 | [0.19973, 0.24999] | pass |
| 115 | 38/50 | 24/50 | 0.16736 | [0.13946, 0.19526] | pass |
| 116 | 34/50 | 29/50 | 0.16650 | [0.13675, 0.19625] | pass |
| 117 | 38/50 | 21/50 | 0.16516 | [0.13809, 0.19223] | pass |
| 118 | 29/50 | 18/50 | 0.12406 | [0.10127, 0.14685] | pass |
| 119 | 34/50 | 26/50 | 0.13538 | [0.11148, 0.15928] | pass |

Table 13: V3 PointPush1 absolute outcomes. Bundles are completion/violation/cost/return/fallback rate.

| Seed | Gate on | Gate off | Fallback only |
|---|---|---|---|
| 100 | 0.645/0.420/24.17/2.57/0.145 | 0.640/0.570/31.27/2.71/0.000 | 0.450/0.240/9.43/1.30/1.000 |
| 101 | 0.765/0.565/25.51/3.82/0.181 | 0.725/0.570/28.77/3.41/0.000 | 0.450/0.220/8.33/1.52/1.000 |
| 102 | 0.650/0.550/23.93/3.22/0.138 | 0.750/0.525/28.82/3.58/0.000 | 0.380/0.210/6.68/1.19/1.000 |
| 103 | 0.720/0.435/20.25/3.45/0.156 | 0.775/0.560/24.42/3.81/0.000 | 0.490/0.320/13.87/1.86/1.000 |
| 104 | 0.605/0.515/21.84/2.72/0.131 | 0.695/0.580/29.84/3.30/0.000 | 0.420/0.270/13.85/1.62/1.000 |
| 105 | 0.595/0.430/24.67/2.50/0.155 | 0.690/0.490/19.45/3.01/0.000 | 0.310/0.220/9.13/1.25/1.000 |
| 106 | 0.670/0.585/26.80/3.18/0.178 | 0.710/0.600/27.39/3.51/0.000 | 0.450/0.260/12.69/1.43/1.000 |
| 107 | 0.675/0.580/28.75/3.20/0.169 | 0.755/0.565/24.80/3.52/0.000 | 0.360/0.170/6.41/1.12/1.000 |
| 108 | 0.725/0.530/27.34/3.37/0.163 | 0.715/0.580/26.66/3.44/0.000 | 0.340/0.310/10.69/1.26/1.000 |
| 109 | 0.610/0.445/23.47/2.39/0.143 | 0.660/0.505/23.77/2.85/0.000 | 0.390/0.170/7.63/1.41/1.000 |
| 110 | 0.565/0.450/20.53/2.67/0.153 | 0.650/0.470/22.05/3.01/0.000 | 0.440/0.260/10.10/1.38/1.000 |
| 111 | 0.725/0.520/22.77/3.64/0.176 | 0.660/0.545/23.44/3.14/0.000 | 0.530/0.330/15.40/1.92/1.000 |
| 112 | 0.680/0.510/29.33/3.25/0.171 | 0.745/0.555/21.20/3.68/0.000 | 0.380/0.230/11.85/1.29/1.000 |
| 113 | 0.680/0.445/19.61/3.25/0.172 | 0.725/0.545/28.34/3.33/0.000 | 0.370/0.210/10.95/1.32/1.000 |
| 114 | 0.640/0.470/20.27/2.91/0.181 | 0.690/0.495/21.98/2.97/0.000 | 0.320/0.200/7.81/1.21/1.000 |
| 115 | 0.670/0.470/19.66/2.87/0.150 | 0.760/0.615/28.50/3.30/0.000 | 0.490/0.240/11.29/1.63/1.000 |
| 116 | 0.670/0.460/21.15/2.71/0.149 | 0.740/0.535/25.55/3.44/0.000 | 0.420/0.310/13.45/1.67/1.000 |
| 117 | 0.705/0.500/25.84/2.95/0.165 | 0.690/0.470/27.69/3.18/0.000 | 0.380/0.280/10.81/1.04/1.000 |
| 118 | 0.625/0.460/18.77/2.73/0.136 | 0.595/0.465/20.52/2.34/0.000 | 0.400/0.220/10.40/1.31/1.000 |
| 119 | 0.705/0.525/23.59/3.33/0.168 | 0.685/0.565/26.98/3.19/0.000 | 0.460/0.320/11.04/1.50/1.000 |

# I Extended Proofs

We expand the fixed-time accounting proposition. The closed-convex capacity and support arguments are complete in Section 5.

## I.1 Proposition 2 (fixed-time executed-action accounting)

**Fixed-time decomposition.** Fix $t \geq 1$. Define $E^{\mathrm{adm}} = \{\mathcal{C}_t \cap \mathcal{A}_{\mathrm{adm}}(S_t) \neq \varnothing\}$ and $E^{\mathrm{fb}} = \{\mathcal{C}_t \cap \mathcal{A}_{\mathrm{adm}}(S_t) = \varnothing\}$ (mutually exclusive, exhaustive). Then

$$\Pr(Y(S_t, a_t^{\mathrm{exec}}) > d) = \Pr(Y > d, E^{\mathrm{adm}}) + \Pr(Y > d, E^{\mathrm{fb}}).$$

**Bounding the admit term.** On $E^{\mathrm{adm}}$, $a_t^{\mathrm{exec}} = \arg\min_{a \in \mathcal{C}_t \cap \mathcal{A}_{\mathrm{adm}}} \{U_\kappa(s_t, a) + \beta\Delta(a, a_t^{\mathrm{prop}})\}$ is one of the $M$ proposed actions. Therefore

$$\{Y(S_t, a_t^{\mathrm{exec}}) > d, E^{\mathrm{adm}}\} \subseteq \bigcup_{m=1}^{M} \{a_t^{(m)} \in \mathcal{A}_{\mathrm{adm}}(S_t),\, Y(S_t, a_t^{(m)}) > d\}. \tag{12}$$

By Assumption 1, Lemma 1 supplies a marginal bound of $\alpha$ for each $a_t^{(m)}$. The union bound gives $\Pr(Y > d, E^{\mathrm{adm}}) \leq M\alpha$; no conditional claim is used.

**Bounding the fallback term.** On $E^{\mathrm{fb}}$, $a_t^{\mathrm{exec}} = \pi_{\mathrm{fb}}(S_t)$. Hence

$$\Pr(Y > d, E^{\mathrm{fb}}) \leq \Pr(Y(S_t, \pi_{\mathrm{fb}}(S_t)) > d) \leq \alpha_{\mathrm{fb}} \tag{13}$$

by the separate marginal fallback assumption. Summing the terms and truncating at one gives Eq. (9).

**Operational $M$-factor witness.** The factor cannot be justified merely by writing down a union of events; here is a construction satisfying the stated admission and selection rules. At one fixed time let $S = (B_1, \ldots, B_M)$ have independent $\mathrm{Bernoulli}(q)$ coordinates. Action $a$ has immediate cost $Y(S, a) = B_a$, the fallback has cost zero, and $d = 1/2$. Use one critic $U_\kappa(S, a) = (1 - B_a)/4$, zero ensemble disagreement, a constant support feature map with radius zero, and proposal-distance weight $\beta = 0$. Calibration uses independent states and uniform actions; its residuals, and each prespecified candidate residual, are i.i.d. on $\{-1/4, 1\}$ with upper mass $q$, satisfying Assumption 1. For feasible rank $k = \lceil (n+1)(1-\alpha) \rceil \leq n$, the calibration quantile is $-1/4$ exactly when $Z \sim \mathrm{Bin}(n, q)$ satisfies $Z \leq n - k$. Then $q_+ = 0$ and every candidate passes both vote and cost gates. Equation (3) chooses a bad action whenever any $B_a = 1$, because its critic value is smaller. Otherwise the quantile is one, all candidates fail admission, and the safe fallback executes. Thus

$$\Pr(Y(S, a^{\mathrm{exec}}) > d) = \Pr(Z \leq n - k)\{1 - (1 - q)^M\}.$$

Each individual unsafe-admission probability is only $q \Pr(Z \leq n - k)$. Taking $q = \alpha/2$ and $n$ sufficiently large gives order $M\alpha$ when $M\alpha$ is small; this is order-level sharpness, not equality with the finite bound $M\alpha$. At $n = 1000$, $\alpha = .1$, $q = .05$, and $M = 4$, exact enumeration gives individual probability .0500 and executed probability .185494. This is a witness within the guarantee assumptions, not evidence about trained critics or complete-trajectory qualification.

# J Local and Official OSRL Public-Task Baseline Logs

The first table records local CarCircle OSRL logs referenced from Section C.1. The summary is generated by `simulation/summarize_osrl_local_logs.py` from a local OSRL log directory. The second table records the official PointCircle BC-safe checkpoint evaluation generated by `simulation/run_osrl_official_baseline.py`. These rows document baseline progress only; they do not change the zero-success public-task limitation.

Table 14: Local OSRL public-task baseline log summary on OfflineCarCircle1Gymnasium-v0 (delay $= 2$). Three completed BC-Safe / BC-All / CPQ seeds, plus a 50k-step COptiDICE seed and a 25k-step partial seed. Reward/cost are eval means at the final logged step.

| Baseline | Seeds | Steps | EpReward | EpCost | Status |
|---|---|---|---|---|---|
| BC-Safe | 3 | 50k | $11.96 \pm 1.88$ | 135.6 | complete |
| BC-All | 3 | 50k | $13.49 \pm 0.56$ | 343.6 | complete |
| CPQ Xu et al. (2022) | 3 | 50k | $-14.53 \pm 0.93$ | 266.7 | complete local |
| COptiDICE Lee et al. (2022) | 1 (+1) | 50k | $-11.25$ | 355.4 | 1 seed complete, 1 seed 25k partial |

Table 15: Official OSRL BC-safe checkpoint evaluation on OfflinePointCircle1Gymnasium-v0 from the `osrl_official_baseline_eval.csv` artifact. Ready rows aggregate 3 official checkpoint runs $\times$ 3 evaluation seeds $\times$ 20 episodes. Cost-10 remains blocked because the official Drive PointCircle cost-10 folder did not expose a BC-safe run in the downloaded listing. Cost limit identifies the checkpoint training budget. Here ViolRate is the fraction of evaluation episodes containing any positive-cost step, not the fraction whose cumulative cost exceeds that budget.

| Cost limit | Ckpts | Episodes | EpRet | EpCost | ViolRate | Success |
|---|---|---|---|---|---|---|
| 10 | 0 | 0 | – | – | – | – |
| 20 | 3 | 180 | 34.84 | 3.53 | 0.156 | 0.0 |
| 40 | 3 | 180 | 39.42 | 4.14 | 0.233 | 0.0 |
| 80 | 3 | 180 | 39.30 | 18.82 | 0.583 | 0.0 |

## K Public-Task Negative-Evidence Ledger Figure

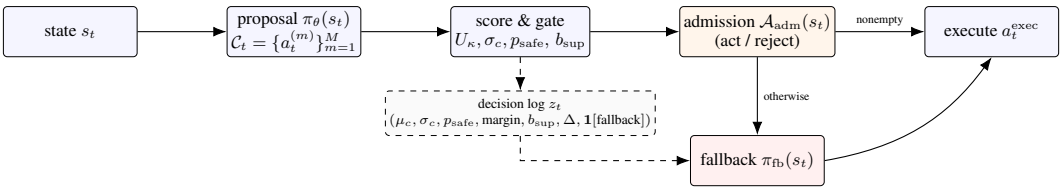
The appendix figure is the rendering of the public negative-evidence ledger summarized in Section C.1. It is generated by `simulation/plot_public_negative_ledger.py` from existing per-row CSV files and adds no new claim: the bottom panel shows `SuccessRate`$= 0.0$ everywhere, while the top panel shows the accompanying costs.

**Numeric failure-mode associations.** The following table aggregates offline terminal rate, admission, fallback, support, and rollout length statistics from `simulation/results/negative_ledger_failure_diagnosis.csv`. Audited offline datasets rarely reach terminal states (0.001–0.002), and fallback is frequent in some probes. These co-occurrences motivate controlled follow-up but do not explain the all-zero outcomes or attribute them to the gate.

## L Public-task claim audit and component ablation

The prior five-seed component audit remains adverse and exploratory. No-Ensemble has $+1.25$ cost and $+0.63$ violation points, but No-Margin has $-0.50$ and $-0.34$, while No-Gate is nearly unchanged ($+0.08, -0.02$). It does not support necessity or benefit of the margin or disagreement-gate components; only the No-Ensemble direction is consistent with benefit, without causal identification.

**Admission-process schematic.**




**Admission-process schematic.** This navigation aid shows the handoff $g_\phi : (s_t, \mathcal{C}_t) \mapsto (a_t^{\text{exec}}, z_t)$ with act / reject / fallback ternary decision; it is not a separate algorithmic claim. Acceptance follows Eq. (6); the executed action follows Eq. (3). The tuple $z_t$ is an operational decision log; the propositions use their separately stated proxy, trajectory, offline-law, or achievable-set objects.

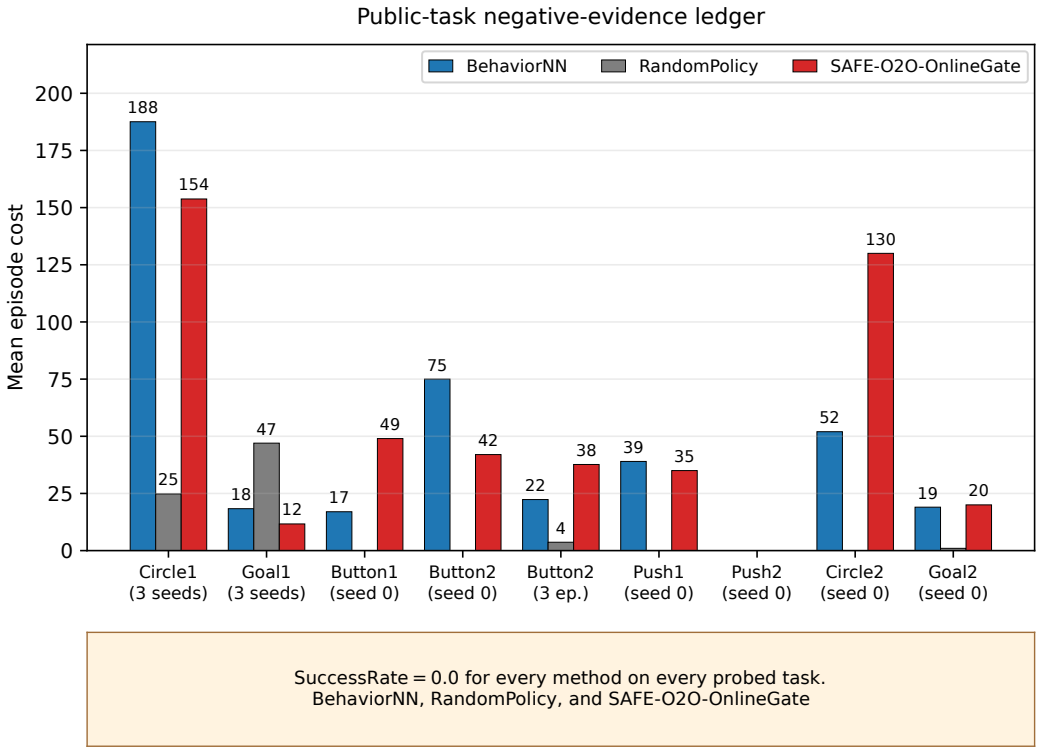



Figure 1: Public-task negative-evidence ledger across nine SAFE-O2O DSRL probes (PointCircle1/2, PointGoal1/2, PointButton1/2 and an extended seed-0/3-episode PointButton2 run, PointPush1/2). Top: mean episode cost; bottom: mean `SuccessRate`. `SuccessRate` is 0.0 in every row for every method. The figure is included as failure-mode evidence and as the visual companion to Section C.1; it is not a benchmark validation result.

Table 16: Numeric failure-mode associations in the public-task negative-evidence ledger. `Term`, `Acc`, `FB`, `Sup`, `MSD`, and `Len` abbreviate terminal rate, accepted-candidate fraction, fallback-decision rate, support-block flag rate, support distance, and mean episode length. `SuccessRate` is 0.0 in every row. Acc divides accepted candidates by all proposed candidates; FB divides fallback decisions by executed steps, so they need not sum to one. Sup is the fraction of steps whose logged candidate has distance above the support threshold, not a support-pass fraction; Sup=0 is compatible with acceptance. The CSV combines indicated per-seed probes and means of those summaries. No new public data is generated.

| Environment | Term | Acc | FB | Sup | MSD | Len |
|---|---|---|---|---|---|---|
| PointCircle1 (3 seeds, summary) | 0.002 | 0.110 | 0.819 | 0.405 | 5.12 | 500 |
| PointGoal1 (3 seeds, summary) | 0.001 | 0.847 | 0.132 | 0.132 | 6.48 | 500 |
| PointButton1 (seed-0 probe) | 0.001 | 0.092 | 0.885 | 0.040 | 8.98 | 200 |
| PointButton2 (seed-0 probe) | – | 0.348 | 0.600 | 0.005 | 9.34 | 200 |
| PointButton2 (seed-0, 3 ep) | – | 0.322 | 0.635 | 0.056 | 8.10 | 500 |
| PointPush1 (seed-0 probe) | 0.001 | 0.362 | 0.590 | 0.000 | 8.06 | 200 |
| PointPush2 (seed-0 probe) | – | 0.019 | 0.970 | 0.175 | 8.11 | 200 |
| PointCircle2 (seed-0 probe) | – | 0.029 | 0.965 | 0.680 | 8.05 | 200 |
| PointGoal2 (seed-0 probe) | – | 0.409 | 0.455 | 0.000 | 6.20 | 200 |

**Claim-to-evidence audit map.** This final map is the non-overclaim checklist for reviewers: each row names the supported claim and the inference that must not be drawn from it.

- **Calibration-qualified selector.** The reported criterion and held-out table establish value over robust-only only on the procedural held-out family; they do not authenticate chronology or establish public-task or learned-score transport.

Table 17: Decision boundary: completed exposure with zero success mass permits no completion attribution or efficacy claim.

| Audit stage | Observable | Result | Licensed inference |
|---|---|---|---|
| Prospective rule | Frozen 8×5×3 grid, endpoint, all-zero veto | preregistration hash matched | No post-hoc success definition or arm exclusion |
| Completeness | cells / episodes / decisions / failed | 120/120; 2,400; 1.2M; 0 | Exposure complete; not efficacy |
| Resolution | completion per environment–arm | 0/100; model-qualified upper 2.9513% | Rules out common rates above 2.9513% at one-sided 95% under the binomial model |
| Arm outcome | fallback / forced / gated | each 0/800 | Tested protocol fails positive validation |
| Attribution veto | environments eligible for gate attribution | 0/8 | No gate benefit, harm, preservation, or failure mechanism from completion |
| Controlled identification | matched policy arms; randomized gate×fallback | executed separately (Table 4) | Gate-policy effect only in the synthetic process |

- **Controlled gate-policy effect.** Supported by 5,760 episode records, balanced assignment within 16 controllers, two nonzero-success fallback-only arms, the $2 \times 2$ factorial, and controller-level contrasts in Table 4. It identifies only the gate-on versus gate-off policy-package effect in the specified synthetic process.
- **Zero-count falsification boundary.** The reported 120-cell, 2,400-episode summaries and zero-count sensitivity analysis reject public-task validation under the tested protocol: the randomized completion contrast is defined but estimated as zero with weak event-rate resolution, and the failure mechanism is not identified.
- **Admission process object.** Supported by Definition 1, Eq. (3), logged tuple $z_t$, and the positioning in Section 2. This does not claim a new offline-RL optimizer or replace public safe-RL baselines.
- **Executed-action marginal cost-proxy bound.** Proposition 2 remains an immediate-$Y$ result with $M\alpha + \alpha_{\mathrm{fb}}$ accounting; it is not a trajectory certificate.
- **Complete-trajectory qualification.** Proposition 3 and the finite-data learned-score runner align calibration with discounted trajectory cost, include fallback behavior, and survive nominal held-out testing. The matched-rate comparator rejects one state-independent random switcher, while the latent-hazard oracle bounds this engineered slice. They require exchangeable complete units; sensor inversion fails qualification and forbids a transport claim.
- **Risk–coverage support geometry.** Supported by Propositions 6–7. It characterizes supported points of the convexified achievable set, not optimizer equivalence, benchmark dominance, or empirical slope validity.
- **Public attribution outcomes.** The eight-environment ledger has zero completion in every arm, and PointGoal V1 fails its intervention floor. Offline-calibrated V2 and the 54-pair V3 replication fail their joint benefit rules. V4 is the pre-treatment deployable comparator and shows higher gate-on violation than an expected-rate mixer. V5 is a post-treatment stress test that exactly matches realized fallback count and episode length; no task passes its joint rule. None supports public gate efficacy or a universal switching-policy comparison.

**Proposition-aligned synthetic validation.** This appendix checks the proposition surfaces in controlled synthetic settings. The reproduction script writes a CSV under `simulation/results/`. These checks support internal consistency only; they are not public-task evidence and do not affect the zero-success DSRL limitation.

- **Proposition 6.** The two-state example from Appendix M is evaluated on grids of 201 and 2,001 points. The closed-form envelope has maximum formula error 0 and minimum concavity gap $-1.11 \times 10^{-16}$ at both resolutions, i.e. numerical roundoff around the exact concave frontier.
- **Proposition 2.** A controlled proxy process with $M = 3$, $\alpha = 0.05$, $\alpha_{\mathrm{fb}} = 0.02$, and $10^5$ trials gives empirical executed-action proxy violation $0.0398$, below the union-bound target

$M\alpha + \alpha_{\text{fb}} = 0.17$. A stricter setting with $M = 5$, $\alpha = 0.02$, $\alpha_{\text{fb}} = 0.01$, and $5 \times 10^5$ trials gives empirical proxy violation $0.0167$, below the corresponding union-bound target $0.11$.

### L.1 Component ablation on the 54 qualified pairs (V6)

V6 freezes the 54 V3-qualified checkpoint pairs and evaluates five variants with identical reset seeds per checkpoint: the full gate; no-margin (replace $\mu + \kappa\sigma$ by $\mu$ in the margin predicate); no-vote (drop $\sigma$ and confidence predicates); no-support (drop the support facade's veto); and robust-only (margin only, no support). Each variant's mixer rate is recalibrated on 40 disjoint calibration episodes before any test observation. Test rows are 200 episodes per variant per checkpoint on both gate-on and the per-variant random-mix arm. Variant-versus-full contrasts are paired by identical reset seed; intervals use checkpoint-level $t$ inference with Bonferroni family-8 adjustment over four variants times two endpoints.

Table 18: V6 variant-minus-full paired contrasts, Bonferroni family-8 intervals. Zero-variance contrasts are marked unassessed rather than assigned a zero-width interval.

| Variant | Completion diff | Violation diff | Status |
|---|---|---|---|
| No-margin | $+.0147\,[.0022, .0272]$ | $+.0105\,[-.0020, .0229]$ | assessed |
| No-support | $-.0027\,[-.0080, .0026]$ | $+.0002\,[-.0075, .0078]$ | assessed |
| No-vote | 0 | 0 | unassessed zero variance |
| Robust-only | $-.0027\,[-.0080, .0026]$ | $+.0002\,[-.0075, .0078]$ | assessed |

Per-variant gate-minus-mixer completion and violation contrasts are respectively $.0331/.0970$ (full), $.0283/.0505$ (no-margin), $.0321/.0949$ (no-support), $.0331/.0970$ (no-vote), and $.0322/.0947$ (robust-only), all with family-24 intervals in the aggregate receipt. Mixer rates are $.29795/.16185/.28913/.29795/.28913$ for full/no-margin/no-support/no-vote/robust-only. The no-vote variant's identical decisions on all 54 checkpoints show the disagreement predicates never flip a margin decision on this population; the zero effect is not evidence of redundancy outside it. The margin ablation's detectable completion increase is paired with a statistically unresolved violation change, so the margin's role on this population is not a clean safety/efficiency decomposition. Runner `simulation/v6_component_ablation.py`; the aggregate is in the accompanying artifact record.

## M Extended discussion and limitations

**Comparator boundary.** V4 is the pre-treatment deployable comparator: it asks whether the observed state-dependent gate improves over state-independent mixing at the same expected fallback rate. It does not. V5 separately permutes each gate arm's realized fallback mask. Completion rises on all three tasks, but violation also rises; its frozen joint rule fails on all three. Exact intervention count and administrative horizon are matched only after the gate trajectory is observed, while induced state occupancy and all alternative schedules are not. Gate-on violation is higher on all three tasks, while completion is higher on PointGoal and PointPush and unresolved on PointButton. This rules out the frozen joint benefit claim for that comparator, not every state-, time-, history-, or uncertainty-dependent switching rule.

**PointGoal V2 protocol and exact audit.** The V2 threshold is selected independently per checkpoint from only 20,000 training and 5,000 calibration transitions to minimize distance from 20% offline rejection; online endpoints are excluded. Qualification rejection rates and episode-clustered 95% intervals are $.31758\,[.28423, .35093]$, $.28862\,[.25224, .32500]$, and $.28030\,[.25034, .31026]$. The confirmatory checkpoint effects for completion are $(-.010, .005, .005)$, for violation $(-.005, -.040, 0)$, and for cost $(-6.475, -4.075, -2.060)$. The create-only run is `.repair/randomized_pointgoal_gate_runs/confirmatory_v2_offline_rejection_20260902_r1`; its 1,500 episode records, 1,500,000 decision records, preregistration, qualification output, failed-cell ledger (empty), and SHA-256 manifest are retained without pooling across evidence classes.

**Zero-count sensitivity.** Under independent homogeneous episodes, $(1 - p_U)^{100} = 0.05$ gives the exact one-sided 95% upper endpoint 2.9513%. Yet 20 episodes are nested in each of five controllers. We therefore also fit a transparent beta-binomial sensitivity model: controllers are independent; within a controller the 20 Bernoulli outcomes are exchangeable with mean $p$ and intracluster correlation $\rho$. Inverting the exact all-zero beta-binomial likelihood gives upper endpoints 2.95, 3.23, 4.24, 5.41, 8.92, 16.03, and 45.07% for $\rho = 0, .01, .05, .10, .25, .50, 1$. The last value is the five-independent-cluster limit. This is sensitivity analysis, not an ICC estimate from all-zero data; without a dependence model, five zero clusters provide no useful tight episode-rate bound. We do not substitute heterogeneous-task pooled 0/800 or 0/2,400 values.

**Multiplicity and failed attribution.** The confirmatory family contains eight environment-level gated-versus-forced safety-frontier hypotheses at familywise $\alpha = .05$ (Bonferroni per-environment $\alpha = .00625$). Native-success McNemar statistics are ties in every environment: raw and adjusted $p = 1$. In PointCircle1, PointGoal1, PointButton1, PointPush1, PointCircle2, PointGoal2, PointButton2, and PointPush2 order, one-sided controller-sign $p$ values for lower violation are $.500, .938, .313, .969, .125, .500, 1, .0625$; Bonferroni-adjusted values are $1, 1, 1, 1, 1, 1, 1, .500$. No joint frontier hypothesis passes. Each arm is a compound policy and state trajectories change after execution, so cost/return contrasts cannot identify a completion mechanism. All eight environments retain the no-attribution verdict.

**Design that would identify the intended effect.** First establish, on a preregistered native success endpoint, that the fixed proposal and fallback each have nonzero completion on the same checkpoint/task/horizon. Then randomize *episode-level* gate enforcement versus forced admission using the same candidate, fallback, trained checkpoint, reset-seed schedule, and evaluation budget; episode-level assignment preserves the policy-level estimand under state feedback. Report an intention-to-treat risk difference with seed-clustered uncertainty and a prespecified noninferiority margin for completion, alongside cost. A factorial replication crossing gate enforcement with fallback identity would separate gate–fallback interaction from proposal competence. This design identifies the effect of the gated policy package relative to forced admission; identifying a per-decision mediation effect would require an additional sequentially randomized intervention at gate-discordant states and explicit time-varying-confounding assumptions. Table 17 records what the present protocol does and does not establish.

**Evidentiary status.** Table 1 is confirmatory only for the frozen procedural-family generator: calibration and test controller parameters are disjoint, and test outcomes do not select the threshold. Table 4 and the 27-cell grid diagnose the previous gate; the eight-task public grid is a confirmatory falsification, and PointGoal V2 is an identified learned-policy null under its frozen benefit rule. Component analyses remain exploratory. These evidence classes are not pooled to rescue one another. The separated-family experiment closes its prespecified fallback-value criterion only for its procedural generator: robust-only clears the prespecified success floor, and simultaneous held-out bounds support selective success uplift with non-inferior violation. The trajectory experiment replaces the constructed score by a finite-sample affine synthetic score and aligns calibration with discounted rollout cost, but it does not execute or validate the full Section 4 pipeline. Its inversion failure shows that fitting a score does not establish transport. All dynamics and labels remain procedurally hand specified; neither experiment establishes realism or DSRL/OSRL efficacy.

The formal limits are structural. Ordinary split conformal does not imply conditional or anytime validity; the fixed-time $Y$ route still mismatches cumulative-target critic training, grows with $M$, and separately assumes fallback control. Proposition 3 repairs the endpoint mismatch only by changing the calibration unit to complete trajectories; Proposition 5 records that those policy-specific rollouts occur before protection. Weighted allocation can focus calibration power but does not create counterfactual costs; Proposition 4 shows the barrier only when candidate occupancy is positive and the admissible continuation range crosses budget. The $L = 8, 16$ null qualification and inversion failure expose multiplicity and transport costs rather than hiding them. Older stress files lack fixed-time $Y$ and fallback-isolated outcomes, so their one-minus-rollout-violation rates .818–.896 remain non-theorem-aligned adverse diagnostics. Convex support geometry provides no optimizer or unique knob.

The public grid remains negative: its endpoint is termination-derived, 100 episodes are nested in five controllers, and beta-binomial sensitivity expands the apparent 2.95% upper endpoint to 45.07%. Neither eight fixed tasks nor 1.2M dependent decisions enlarge the independent sample. PointGoal V2 supplies native metrics, same-checkpoint competent learned policies, and checkpoint-level uncertainty over three distinct training seeds; stochastic independence beyond that seed distinction is not externally established. Its safety effect remains unresolved and fallback-only exposes a utility–safety trade-off. Per-decision mechanism attribution would additionally require sequential randomization and time-varying-confounding assumptions.

**Worked example: two-state capacity frontier.** To illustrate the geometry of Proposition 6, consider a toy admission population on two state-action atoms $(s_1, a_1)$ and $(s_2, a_2)$ with $\Pr = 1/2$ each. Suppose $Y(s_1, a_1) = 0$ (always safe), $Y(s_2, a_2) = 2d$ (always unsafe). A parameterized admission rule $\phi = p \in [0, 1]$ admits $(s_2, a_2)$ with probability $p$ and always admits $(s_1, a_1)$. Then $C_p = 1/2 + p/2$ and $R_p = p/2$, giving the explicit envelope $\kappa(\alpha) = \min\{1/2 + \alpha, 1\}$ for $\alpha \in [0, 1/2]$ and $\kappa(\alpha) = 1$ for $\alpha \geq 1/2$. This $\kappa$ is concave non-decreasing as predicted, with right-derivative $\kappa'_+(\alpha) = 1$ on $(0, 1/2)$ and 0 thereafter. Since $\alpha - \lambda\kappa(\alpha) = \alpha(1 - \lambda) - \lambda/2$ on $[0, 1/2]$, the Lagrangian-optimal $\alpha^*$ is $1/2$ for $\lambda > 1$ and 0 for $\lambda < 1$ (with every $\alpha \in [0, 1/2]$ optimal at $\lambda = 1$). The single-knob example illustrates the same capacity geometry used by the multi-knob deployment rule of Section 4.

**Why the fixed-time route still does not imply a trajectory budget.** Proposition 3 avoids, rather than repairs, the immediate-proxy route by calibrating $G_T$ itself. The gap for Proposition 2 is not merely terminological. Let $d = 1$, $\gamma = .99$, and consider a deterministic six-step trajectory with $Y(S_t, a_t) = c(S_t, a_t) = .2$ at every step. Then $\Pr(Y(S_t, a_t) > d) = 0$ at every fixed time, satisfying even a zero proxy-violation bound, while

$$J_c = \sum_{t=0}^{5} .99^t(.2) = 1.170397 > 1 = d.$$

Thus no theorem can derive Eq. (1) from the fixed-time event alone without an additional bridge relating $Y$, horizon/discounting, state feedback, and the deployed trajectory. Cumulative-target critic training does not supply that bridge because residuals use immediate labels. Proposition 2 therefore remains proxy accounting; Proposition 3 obtains a finite-horizon chance statement from directly observed complete-trajectory costs under a distinct exchangeability assumption.

**When does Assumption 1 hold?** Assumption 1 requires joint exchangeability of the calibration residuals and each fixed-time candidate residual. Relevant regimes and boundaries are:

- **i.i.d. deployment**: if $(S_t, a_t^{(m)})$ are sampled i.i.d. from the same distribution that generated $\mathcal{D}_{\text{cal}}$, exchangeability holds trivially. This is the cleanest setting but rules out adaptive online policies.
- **Stationary policy without state shift**: a fixed policy and stationary dynamics may support a same-distribution approximation, but stationarity alone does not prove exchangeability with the offline calibration sample.
- **Weighted CP correction**: under a known covariate-shift model, weighted conformal methods such as Tibshirani et al. (2019) can replace ordinary split conformal. That extension is not implemented or claimed here.

Adaptive online policies that condition on past rewards generally violate Assumption 1. In that regime Proposition 2 does not apply; restoring rigor requires a separately validated online or sequential calibration construction.

## N COMPLETE-VECTOR CALIBRATION AND ADMISSION LOSS

Complete-vector calibration changes the calibration unit, not the ranking objective. Freeze critics, candidate generation, and all gate knobs other than the calibrated margin before calibration. Let $n \geq 1$ count calibration states, $M \geq 1$ count candidates per state, and $0 < \alpha < 1$ be the target error level. Assume finite, exchangeable complete residual vectors $V_i = (r_{i1}, \ldots, r_{iM})$, $i = 1, \ldots, n+1$, where $r_{im} = Y(S_i, a_i^{(m)}) - U_\kappa(S_i, a_i^{(m)})$, $S_i$ is a state, and $Y$ is its candidate's immediate-cost label.

The last vector is the fixed-time test unit. This premise is stronger than per-candidate exchangeability against the manuscript's scalar calibration sample; it requires no independence within vectors.

Define $R_i = \max_{1 \le m \le M} r_{im}$ and $k = \lceil (n+1)(1-\alpha) \rceil$. Let $Q = R_{(k)}$ for $k \le n$, using the ordered calibration maxima with ties retained, and $Q = +\infty$ otherwise. Set $Q_+ = \max\{Q, 0\}$. Substitute $d - Q_+$ for the admission threshold, keeping the other admission constraints, robust-cost-plus-distance ranking, and fallback unchanged. An infinite margin explicitly rejects every candidate. Let $A$ denote execution of an admitted candidate and $E = \{Y(S_{n+1}, a^{\text{exec}}) > d\}$ denote unsafe execution.

**Conditional proposition.** Under these premises, $\Pr(E \cap A) \le \alpha$. If the separate marginal fallback assumption holds at this same test-state distribution with bound $\alpha_{\text{fb}}$, then $\Pr(E) \le \min\{1, \alpha + \alpha_{\text{fb}}\}$.

**Proof.** Exchangeability is preserved by taking vector maxima. The conformal rank argument, with strict exceedance and ties retained, gives $\Pr(R_{n+1} > Q) \le (n+1-k)/(n+1) \le \alpha$ when $k \le n$; the infinite case has zero exceedance probability. Since $Q_+ \ge Q$, its exceedance probability cannot increase. On $E \cap A$, the selected candidate satisfies $U_\kappa + Q_+ \le d < Y$, so its residual exceeds $Q_+$ and hence $R_{n+1} > Q_+$. The fallback joint event is contained in the marginal unsafe-fallback event. Adding these disjoint execution branches proves the claim.

These probabilities average over calibration and test units, not over test units conditional on a realized split. Writing $p_A = \Pr(A) > 0$, conditional admitted risk is bounded only by $\min\{1, \alpha/p_A\}$; it is undefined when $p_A = 0$. Marginal fallback control likewise supplies no conditional fallback guarantee.

The local artifact `F40_CANDIDATE_SET.json` reports 96 grid cases, each with 1,000 calibration splits and 1,000 test sets per split. The following $M = 8$ examples report marginal unsafe admitted execution $\widehat{r}$ and decision admission $\widehat{p}_A$, not total executed risk. Original and Bonferroni denote scalar calibration at $\alpha$ and $\alpha/M$; vector denotes maximum-residual calibration. These rows use $n = 99$, $\alpha = .1$, zero distance weight, and equal numbers of calibration units, not equal label budgets.

| Law | Rule | $\widehat{r}$ | $\widehat{p}_A$ |
|---|---|---|---|
| Independent | Original | 0.277136 | 0.996 |
| Independent | Bonferroni | 0.003039 | 0.011 |
| Independent | Vector | 0 | 0 |
| Correlated | Original | 0.039758 | 0.995 |
| Correlated | Bonferroni | 0.000596 | 0.015 |
| Correlated | Vector | 0.039758 | 0.995 |

Dependence materially affects admission; zero joint risk with zero admission is not successful deployment. All three rules coincide in the $M = 1$ cases. Complete calibration costs $nM$ counterfactual labels, unavailable merely from observed offline transitions or critic predictions. The independent reference implementation passed exact small-case and selection checks, but is not a production-adapter test or a public-data improvement. The public $M = 1$ gate cannot benefit from removing multiplicity.

The eight-task zero ledger, proposal-package reversals, PointGoal V1 failure, V2 null, all V3 qualification failures and outcomes, V4 rate-matched adversity, V5 count-matched adversity and sign-test limitations, and exploratory adverse ablation remain unchanged. Deployment exchangeability, induced-occupancy fallback safety, sequential validity, infinite-horizon guarantees, and historical provenance remain unresolved. This is local preparation, not publicly released data or independent human validation.

**Reproduction record.** The entry point is `simulation/candidate_set_calibration_oracle.py`. JSON result paths in this appendix are relative to the project evidence bundle at the repository root. Source and configuration bindings accompany the local records; this does not establish independent external replication.

## O POST-HOC COMPLETION–VIOLATION CONTRASTS

This descriptive reanalysis combines two already executed campaigns without rerunning or modifying their production shards. The mixer sweep retains all four offline-defined margin points (original, q10, q50, q90) for every one of the 54 V3-qualified BC-All/BC-Safe checkpoint pairs: 20 PointGoal, 14 PointButton, and 20 PointPush. The shuffle comparison contributes only the original threshold because no threshold sweep was run for that comparator. Thus the analysis does not impute an unobserved shuffle frontier.

**Frozen data and inferential unit.** At each mixer point, 40 calibration-only gate episodes fixed a state-independent Bernoulli mixer rate before 200 common test resets per arm. The count-shuffle campaign used 200 common resets and exactly matched each gate trajectory's realized fallback count and administrative horizon. For both comparators, paired episode differences are averaged within checkpoint pair; the checkpoint pair, not an episode, is the inferential unit. We recompute two-sided Bonferroni Student bands jointly over all 15 observed task–comparator–point rows and both endpoints (30 contrasts). A row is completion-noninferior (CM, using a five-percentage-point margin) only when its simultaneous completion lower endpoint exceeds $-.05$; the frozen joint criterion additionally requires the simultaneous violation upper endpoint to be below zero.

| Task | Comparator | Point | Δ comp. | Δ viol. | CM | Frontier | Joint |
|---|---|---|---|---|---|---|---|
| Goal | mixer | original | +7.55 | +22.83 | yes | yes | no |
| Goal | mixer | q10 | -0.55 | +0.00 | yes | yes | no |
| Goal | mixer | q50 | +3.43 | +12.93 | yes | yes | no |
| Goal | mixer | q90 | +6.33 | +19.05 | yes | yes | no |
| Goal | shuffle | original | +7.10 | +23.72 | yes | yes | no |
| Button | mixer | original | +0.43 | +0.82 | yes | yes | no |
| Button | mixer | q10 | -12.50 | -14.54 | no | no | no |
| Button | mixer | q50 | -3.61 | -1.54 | no | no | no |
| Button | mixer | q90 | +0.39 | +1.96 | yes | no | no |
| Button | shuffle | original | +3.43 | +6.00 | yes | yes | no |
| Push | mixer | original | +1.10 | +2.80 | yes | no | no |
| Push | mixer | q10 | -3.02 | -3.30 | no | no | no |
| Push | mixer | q50 | +0.95 | +1.25 | yes | no | no |
| Push | mixer | q90 | +2.92 | -0.83 | yes | yes | no |
| Push | shuffle | original | +3.45 | +1.98 | yes | yes | no |

**Table 19:** Gate minus comparator differences in percentage points. Positive completion is favorable; positive violation is unfavorable. "Frontier" is point-estimate nondominance among completion-noninferior rows within the same task and comparator. The shuffle singleton is vacuously nondominated and provides no frontier evidence. These are comparator-relative contrasts, not an absolute gate operating frontier: even the mixer reference changes with its calibration-estimated rate. The label is descriptive and outcome-inspected, not a threshold-selection rule. "Joint" applies the frozen simultaneous-band criterion.

No observed row passes the frozen joint criterion. This is a failure of that criterion for this instantiated gate and these comparators, not evidence that gating cannot help. The sweep also rules out the broader phrase "no benefit": PointPush q90 has favorable point estimates against its mixer ($+2.925$ completion and $-0.825$ violation points), but its simultaneous violation band $[-3.879, 2.229]$ includes zero. Conversely, PointButton q10 reduces violation by $14.536$ points but fails completion noninferiority because its completion lower endpoint is below $-5$ points. PointGoal q10 is nearly equal in mean completion ($-0.55$ points) with zero mean violation change, but does not establish improvement.

**No test-set threshold selection.** All four mixer points and the original shuffle point are reported. Frontier labels were computed after observing test outcomes and are not used to choose, recommend, or certify a threshold. Any future threshold choice must be fixed using separate calibration data and evaluated on new test units. The mixer matches a calibration-estimated rate rather than the exact realized test fallback count; only the original-threshold shuffle comparison matches count and horizon, and it is post-treatment. Therefore neither campaign isolates a general causal value of decision timing or ranks all switching policies.

**Reproduction.** `simulation/threshold_frontier_reanalysis.py` consumes the immutable mixer and count-shuffle aggregates, verifies checkpoint identities, and writes `simulation/results/threshold_frontier_reanalysis.json` plus Table 19. The JSON records every checkpoint effect, input hash, simultaneous band, and label. The source campaigns remain reproducible through `simulation/v5_threshold_mixture.py` and `simulation/v5_count_matched_aggregate.py`; no production shard was overwritten.

## P MATCHED-RATE AND COMPONENT DIAGNOSTIC DETAILS

**Matched-rate V4 selector test.** V4 retains the 54 V3-qualified pairs documented in Appendix H and reconstructs each proposal, fallback, gate, support test, and threshold. Each pair receives 200 fresh common resets under gate-on, gate-off, fallback-only, and a state-independent mixer using its observed V3 fallback rate; arm order is shuffled within reset. Complete task-level intervals are in Appendix E, Table 6. Gate-on raises violation by .2430 on PointGoal, .0164 on PointButton, and .0375 on PointPush; these are marginal task-level intervals, not a simultaneous all-task claim. The Holm-adjusted sign sensitivity is significant only for PointGoal. Reset seeds are $5{,}000{,}000 + 100{,}000s + e$, so distinct checkpoint seeds use disjoint reset ranges within each task, while arms within a pair share resets. The intervals treat checkpoint-pair effects as independent replicates conditional on the shared dataset, frozen training procedure, and V3 qualification; they do not quantify variation across new datasets or training procedures. Completion rises on PointGoal and PointPush but is unresolved on PointButton. No task passes the frozen joint rule. The result fails the frozen joint criterion against this one rate-matched state-independent mixer; it does not compare all switching policies or support deployment safety.

**V6 component diagnosis.** Each component variant is calibrated on 40 disjoint episodes per checkpoint and evaluated on 200 fresh paired resets. Against the full gate, no-margin changes completion by $+.0147$ (family-8 CI $[.0022, .0272]$) and violation by $+.0105$ ($[-.0020, .0229]$); no-support and robust-only each change completion by $-.0027$ ($[-.0080, .0026]$) and violation by $+.0002$ ($[-.0075, .0078]$). No-vote makes identical decisions and is unassessed because its checkpoint effects have zero variance. Thus V6 localizes no clean safety–efficiency decomposition and does not generalize beyond this qualified population.

**What the public evidence supports.** Across 54 qualified learned checkpoint pairs and 43,200 fresh matched-reset episodes, state-dependent gate-on has higher violation than rate-matched state-independent mixing on every public task; no task passes the joint rule. This supports a falsification boundary, not public gate efficacy, deployment safety, or a universal switching-policy ranking (Appendix M).

**Reproducibility.** The frozen protocol, 54 shard outputs, all 21,600 episode rows, 10,800 assignment records, population verifier, aggregate code, and exact sign-sensitivity analysis are paper-local. The verifier recomputes exact count/length and schedule-hash closure for every reset; the analysis uses stored checkpoint effects rather than episode rows as independent units.