DELAYED HEURISTIC-BEHAVIOR-TARGET COUPLING FOR OFF-POLICY REINFORCEMENT LEARNING IN INVENTORY CONTROL

Anonymous authors
Paper under double-blind review

ABSTRACT

Many reinforcement learning methods tie data collection to the task policy, while coordinating independently optimized behavior and target policies remains challenging. Policies learned by these methods can perform unreliably in long-horizon tasks under uncertainty, such as inventory control, where failures can incur extreme costs. We introduce Heuristic Behavior Policy Optimization (HBPO), an off-policy framework that routes heuristic guidance through persistent behavior learning, coupled with target-policy, updating separately with residual-based advantage estimation and emphatic weighting. Under matched environment interaction budgets on DSIS and its risk-augmented variant DSIS-RB, HBPO recovers many high-cost initial policies and improves most non-collapse comparison groups. Cold starts achieve competitive costs without final cost collapse across all tested seeds. We identify a transmission chain in the delayed heuristic–behavior–target coupling as a key driver of recovery and improvement. Generation-level diagnostics trace heuristic uptake and subsequent selective target responses, while component ablations support the contribution of this transmission mechanism to policy recovery and improvement.

1 INTRODUCTION

Learning a policy for long-horizon planning requires choosing how to collect the experience used to improve it (Whitney et al., 2021). In inventory control, an order changes the pipeline of future deliveries, and its consequences depend on subsequent demand (Böttcher et al., 2023). Decisions therefore shape not only future costs but also the states encountered during learning. This makes the behavior policy that collects data an important intervention point: if the behavior policy is itself learned, what should it be optimized to do for the learning process as a whole? Concretely, can heuristic guidance applied to it and behavior policy persistently influence the target policy, which is the policy finally want to optimize, to achieve a better performance?

Existing designs give reason to doubt that this mechanism happens in most RL frameworks. Importantly, failures are observed in long-horizon planning tasks. In a studied multi-echelon inventory network, PPO converged consistently only after inventory transitions were constrained during training (Geevers et al., 2024), and our comparisons show high-cost runs for PPO, Discrete SAC, and adapted GPS (Section 5). These methods share a default: experience is collected by the task policy or a recent snapshot of it (Schulman et al., 2017; Haarnoja et al., 2018). Off-policy learning permits a distinct behavior policy (Degris et al., 2012), but leaves open how that policy should improve and how its deviations should inform target updates. Behavior–target mismatch can impair approximate off-policy updates, and fitting a policy to a teacher's actions need not preserve the teacher's long-horizon performance. Our focus is therefore to construct explicit coupling updating rules for the transmission path itself: which behavioral deviations should a separately updated target reinforce or counteract, and when should planning-derived guidance enter its updates?

We introduce **Heuristic Behavior Policy Optimization (HBPO)**, an off-policy coupled framework that routes heuristic guidance through a persistent behavior policy to a separately and selectively optimized target policy. HBPO forms an explicit delayed heuristic–behavior–target transmission: heuristic guidance, constructed from the current generation target policy, first shapes the successor

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.

behavior policy then reaches two-next target updates through four combined optimizations. This transmission is based on data collected by behavior, residual-based advantage estimation supporting both policies and emphatic weighting carrying heuristic confidence forward. Under same interaction budgets on DSIS and DSIS-RB, HBPO recovers many high-cost initial policies, lowers mean costs across most initialization sources, and improves most non-collapse comparison groups. Cold starts achieve competitive costs without cost collapse across all tested seeds. We identify the transmission of delayed heuristic–behavior–target coupling as a key driver of recovery and improvement. An anonymized implementation is included in the supplementary material and will be made publicly available upon acceptance.

2 Related Work

Behavior learning and selective transfer. Off-policy learning permits the collector to differ from the target (Degris et al., 2012). On-policy methods remove the difference by collecting only from the current policy (policy-gradient, natural-gradient, trust-region, and asynchronous families) (Williams, 1992; Sutton et al., 1999; Schulman et al., 2016; Kakade, 2001; Schulman et al., 2015; 2017; Mnih et al., 2016), so the data source cannot outgrow the target. Common algorithms derive their evolving behavior distributions from learned value functions or actors and exploration rules (Mnih et al., 2015; van Hasselt et al., 2016; Schaul et al., 2016; Lillicrap et al., 2016; Fujimoto et al., 2018; Haarnoja et al., 2018). Correction schemes persistently track the behavior (truncated importance sampling, Retrace, V-trace) reconcile it with target updates without shaping it (Wang et al., 2017; Munos et al., 2016; Espeholt et al., 2018; Gruslys et al., 2018), so correction grows with the gap. Heuristic shaping (optimism, count-based bonuses, curiosity, learned novelty policies, variance-optimized behavior) can improve data collection, but these gains need not translate into consistent target-policy improvement (Auer et al., 2002; Bellemare et al., 2016; Pathak et al., 2017; Burda et al., 2019; Whitney et al., 2021; Xu et al., 2018; Liu et al., 2020; Goodall et al., 2026). HBPO couples three in one loop: a temporary heuristic controller guides a persistent behavior learner, and behavior data plus an advantage-filtered transfer term weighting signed target-advantage scores by absolute behavior–target disagreement feed target updates. ACER uses a one-sided residual to correct importance-weight truncation bias (Wang et al., 2017); HBPO uses absolute behavior–target disagreement to weight an additional advantage-based transfer objective. Our experiments support the contribution of this heuristic–behavior–target coupling (Section 5).

Coupled updates and transmission. With the heuristic, behavior, and target roles fixed, the key question is how improvement travels through their coupled updates. GPS fits local controllers directly into the global policy (Levine & Koltun, 2013); conservative and regularized policy iteration fit an improved action distribution into the actor (Kakade & Langford, 2002; Montgomery & Levine, 2016; Abdolmaleki et al., 2018); TRPO, PPO, and GRPO make each improved policy the immediate data collector (Schulman et al., 2015; 2017; Shao et al., 2024); and TPO fits the policy to a softmax-improved target distribution (Kaddour, 2026). HBPO instead claims a delayed coupling: information and policy impacts enter the next generation separately, rather than the current one, making the transmission $\pi_g \to q_{H,g} \to \mu_{g+1} \to \pi_{g+2}$ explicit; Section 5 examines how this transmission operates during recovery and improvement.

Emphatic state weighting. Emphatic temporal-difference and actor–critic methods relate behavior sampling to the state weighting of an off-policy learning objective (Sutton et al., 2016; Imani et al., 2018; Graves et al., 2023; Zhang et al., 2019), and expected emphatic traces and adaptive interest extend this framework to learned state emphasis (Jiang et al., 2022; Klissarov et al., 2022). HBPO adapts this perspective through separate learned masses for target-return learning and behavior transfer, with heuristic-improvement confidence supplying interest for the transfer mass.

Inventory control as a testbed. Deep RL has proven effective across lost-sales, dual-sourcing, and multi-echelon inventory problems (Gijsbrechts et al., 2022), though training conditions materially affect learning (Geevers et al., 2024). Böttcher et al. (2023) optimize dual-sourcing policies by embedding inventory dynamics in recurrent neural networks. We adapt their problem formulation for our testbed. We choose inventory control as a testbed because poor decisions carry severe downside risk, and the systems admit a compact Markov formulation whose known or learnable dynamics support planning-derived guidance. Section 3 defines our inventory control problem.

3 Problem Settings

We adapt the single-item dual-sourcing model of Böttcher et al. (2023) into DSIS, with regular deliveries after four periods and costlier, immediate expedited deliveries. DSIS has IID demand and uninterrupted supply; our extension, DSIS-RB, adds persistent demand and supply disruptions.

We formulate both benchmarks as underlying Markov decision processes $(\mathcal{X}, \mathcal{A}, \mathcal{P}, r, \gamma)$. At period t, DSIS observations $s_t = (I_t, \mathbf{p}_t, B_t)$ contain net inventory, a four-stage order pipeline, and current supply availability. DSIS-RB adds previous demand d_{t-1}, increasing the dimension from six to seven. Actions $a_t = (q_t^r, q_t^e) \in \{0, \dots, 8\}^2$ specify regular and expedited orders. DSIS-RB training pools trajectories across hidden disruption-risk levels ρ, fixed within each trajectory and withheld from all learners and simulation workers. All learned networks in HBPO use only the current observation s_t, with fixed preprocessing. The full environment state is $x_t = (s_t, \rho)$ for DSIS-RB and $x_t = s_t$ for DSIS. Risk variables and their averaging are omitted for DSIS throughout. The environment kernel $\mathcal{P}$ governs order acceptance, arrivals, demand, supply availability, and inventory carryover through $x_{t+1} \sim \mathcal{P}(\cdot \mid x_t, a_t)$ (Appendix C). Period cost c_t comprises expediting, holding, and backlogging. We maximize expected discounted return using $r_t = -c_t/495$ and $\gamma = 0.995$, where 495 is the per-unit, per-period backlog penalty. All methods share information access and real-interaction budgets. DSIS dynamics are public; HBPO and GPS construct DSIS-RB models from shared public simulation specification $\widehat{\mathcal{P}}_g$ and their observed data, without access to true stochastic parameters (Appendix C). Delayed deliveries, accumulated backlogs, and severe shortage penalties make these tasks a testbed for long-horizon decision-making under uncertainty, where failures can incur extreme costs.

4 Heuristic Behavior Policy Optimization Methods

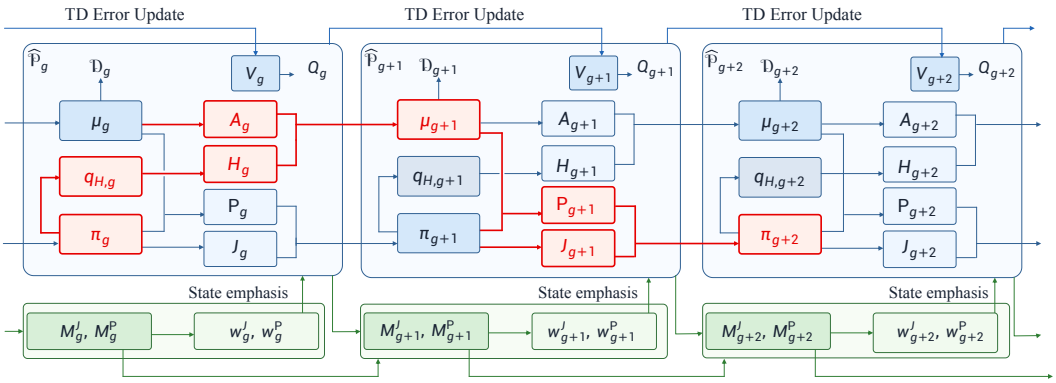

Figure 1: HBPO Learning Framework. Red highlights the delayed heuristic transmission.

HBPO maintains two separately parameterized, persistent policies, $\mu_g = \mu_{\theta_g^\mu}$ and $\pi_g = \pi_{\theta_g^\pi}$. The behavior policy μ_g collects interaction data $\mathcal{D}_g$, while the target policy π_g is evaluated for task performance and defines the continuation policy for model evaluation. Under the simulation specification $\widehat{\mathcal{P}}_g$, frozen within generation g, long-horizon comparisons of alternative first actions followed by π_g construct a temporary heuristic policy $q_{H,g}$. A shared advantage face scores policy changes and behavior–target discrepancies using discounted target-policy advantages, defining the A, J, P objectives. Together with heuristic guidance H, these yield four gradients: behavior combines target-advantage learning g_A with heuristic guidance g_H, while target combines its own advantage learning g_J with selective transfer from behavior g_P.

The frozen reference $v_g(s) = \frac{1}{5}\sum_{j=1}^{5} V_{g,j}(s)$, averaging five separately parameterized scalar critics, supports heuristic evaluation and residual-completed estimation of the discounted model action value $Q_g^{\pi_g}(s, a)$; successor critics are fitted by minimizing squared TD error. Positive masses M_g^χ provide mean-one emphatic weights $w_g^\chi(s)$, $\chi \in \{J, P\}$, for the state-weighted parts of learning; their construction and the critic updates are specified in Section 4.3.

4.1 HEURISTIC OBJECTIVES GUIDANCE

To construct a temporary heuristic policy from π_g, anchor states are sampled from $\mathcal{D}_g$, with each anchor state s defining local time $k = 0$. We compare the long-horizon consequences of candidate first actions a over $T = 1000$ periods, with all subsequent actions drawn from the frozen parent target π_g. Let S_k, A_k, and R_k denote the observation, action, and normalized reward at local decision stage k. Writing $\mathbb{E}_H$ for expectation under $\widehat{\mathcal{P}}_g$, including the synthetic-risk draw for each DSIS-RB tape, we define the expected return $G_{T,g}(s,a)$ and the centered estimated utility $u_g(s,a)$ as

$$G_{T,g}(s,a) = \mathbb{E}_H\left[\sum_{k=0}^{T-1} R_k \,\middle|\, \substack{S_0=s,\ \mathsf{A}_0=a \\ \mathsf{A}_k\sim\pi_g(\cdot|S_k)\ (1\le k<T)}\right], u_g(s,a) = \widehat{G}_{T,g}(s,a) - \sum_{a'\in\mathcal{A}} \pi_g(a' \mid s)\widehat{G}_{T,g}(s,a').$$

Here, $\widehat{G}_{T,g}$ is a Monte Carlo estimate of $G_{T,g}$, constructed using the frozen critic v_g as a control variate and 32 simulation tapes shared across the candidate first-action branches at each distinct query observation. All queries are performed within the model, requiring no additional environment interactions. Appendices B.1.3 and B.4.3 detail the simulation specification and the utility estimator.

For a candidate action distribution q at state s, $\langle q, u_g(s,\cdot)\rangle = \sum_a \big(q_a - \pi_g(a \mid s)\big)\widehat{G}_{T,g}(s,a)$ measures the estimated return gain from replacing the $\pi_g(\cdot \mid s)$ with q and following π_g thereafter. Maximizing this gain alone would concentrate probability on utility-maximizing actions. We therefore construct a temporary stochastic policy for heuristic guidance through Hellinger regularization

$$q_{H,g}(\cdot \mid s) = \arg\max_{q\in\Delta(\mathcal{A})} \left\{ \langle q, u_g(s,\cdot)\rangle - 2\tau \sum_a \left(\sqrt{q_a} - \sqrt{\pi_g(a \mid s)}\right)^2 \right\}. \tag{1}$$

The utility term favors higher estimated returns, while the symmetric, bounded squared Hellinger penalty moderates departures from the target prior to help stabilize heuristic guidance. The coefficient $\tau > 0$ controls this trade-off and remains fixed throughout each run after calibration under the shared hyperparameter-selection protocol HPO. With a strictly positive target prior, scalar root finding yields the temporary guidance policy at each queried state. The resulting policy has no trainable actor parameters. Appendix B.4.7 details the calibration, controller solution, and derivation; Appendix C reports the hyperparameter-selection protocol.

Denote ν_g^J as the empirical distribution of behavior states weighted by w_g^J,introduced in Section 4.3. Starting from anchors drawn from ν_g^J, short model rollouts under $q_{H,g}$ generate the empirical training distribution ξ_g by retaining states at stages $0,\dots,8$. Each retained state receives a controller query with a full $T = 1000$ horizon, so the short rollouts select training locations without shortening the utility horizon. Using $q_{H,g}(\cdot \mid s)$ as the fitting targets at each retained state, define the behavior-guidance gradient by

$$g_{H,g} = -2\,\nabla_{\theta^\mu}\mathbb{E}_{s\sim\xi_g}\left[\left(\arccos\left[\sum_a \sqrt{\mu_{\theta^\mu}(a \mid s)\, q_{H,g}(a \mid s)}\right]\right)^2\right]\Bigg|_{\theta^\mu=\theta_g^\mu}, \tag{2}$$

where $q_{H,g}$ and ξ_g are held fixed during differentiation. Detailed in Appendix B.4.9.

4.2 DELAYED HEURISTIC–BEHAVIOR–TARGET COUPLING

Here we explicitly construct the delayed heuristic–behavior–target coupling and certain the policies updating rules. For the n behavior transitions (s_i, a_i, s_i') in $\mathcal{D}_g$, we now specify the advantage face used by the coupled policy updates. Following the centering used for u_g, we evaluate each first action by its discounted return advantage: $\mathbf{a}_i(a) = Q(s_i,a;\pi_g) - \sum_b \pi_g(b \mid s_i)\, Q(s_i,b;\pi_g)$, where $Q(s,a;\pi_g)$ denotes the discounted model action value, helping advantage estimation, introduced in Equation (11). The vector $\widehat{\mathbf{a}}_i$ records these gains at sampled state s_i compared to π_g, and for policy change ℓ, define advantage face $\Lambda_t[\ell] = \frac{1}{n}\sum_{i=1}^n c_i^t\langle \ell_i, \widehat{\mathbf{a}}_i\rangle$, $(c_i^A, c_i^J, c_i^P) = (1, w_i^J, w_i^P)$. The score rewards adding weight to actions with positive estimated advantage and removing weight from actions with negative estimated advantage. Here $\chi \in \{A, J, P\}$, $w_i^\chi = w_g^\chi(s_i)$, and c_i^χ determines how strongly each sampled state contributes. Appendix B.2.3 details the estimator.

Suppressing generation indices, we hold the parent policies μ, π fixed and use m, p to denote possible updated behavior and target policies. Both actors receive a direct advantage-driven update:

$$A(m) = \tau^{-1}\Lambda_A[m-\mu]. \quad (3) \qquad g_A = \tau^{-1}\Lambda_A\big[\mu\nabla_{\theta^\mu}\log\mu\big]. \quad (4)$$

$$J(p) = \Lambda_J[p-\pi]. \quad (5) \qquad g_J = \tau^{-1}\Lambda_J\big[\pi\nabla_{\theta^\pi}\log\pi\big]. \quad (6)$$

Gradients are evaluated at the frozen parent parameters, with $g_J = \nabla_{\theta^\pi}(J/\tau)$ matching the advantage scale used by A and P. τ is the heuristic regularization coefficient in Equation (1). $A(m)$ directs behavior toward actions favored by target-advantage evidence, while $J(p)$ updates target under its own state emphasis. Together with g_H from Equation (2), behavior combines advantage-driven learning g_A with attraction toward q_H.

Heuristic-guided behavior should also inform target action choices through an explicit policy-level signal. We define

$$P(p) = \tau^{-1}\Lambda_P[|\mu-\pi|(p/\pi-1)], \qquad g_P = \frac{1}{n\tau}\sum_{i,a} w_i^P|\mu_i(a)-\pi_i(a)|\widehat{a}_{ia}\nabla_{\theta^\pi}\log\pi_i(a), \quad (7)$$

where $\mu_i(a) = \mu(a \mid s_i)$ and similarly for π. This advantage-filtered transfer weights disagreements by their magnitude and uses advantage to set their direction: supported behavioral deviations are retained, and opposed deviations are reversed in the scores. P can emphasize actions favored by behavior but rarely selected by target; it can increase disagreement and vanishes when $\mu = \pi$. Behavior's action distribution thus guides target optimization beyond its role as a data source. Theorem 4 shows that the filtered transfer signal admits an auxiliary distribution with nonnegative gain in the frozen advantage face.

Each actor combines its two gradients in its own Fisher geometry, which relates parameter changes to local changes in policy distributions (Kakade, 2001). We use $\mathrm{Fuse}_F(g_1,g_2) = F^{-1}(\alpha_1 g_1 + \alpha_2 g_2)$, with coefficients determined by the natural norms $n_i = \sqrt{g_i^\top F^{-1} g_i}$ to balance their contributions. The updates are

$$\theta^\mu_{g+1} = \theta^\mu_g + t_\mu\,\mathrm{Fuse}_{F_\mu}(g_A, g_H). \quad (8) \qquad \theta^\pi_{g+1} = \theta^\pi_g + t_\pi\,\mathrm{Fuse}_{F_\pi}(g_J, g_P). \quad (9)$$

Shared Fisher geometry makes the two gradients comparable in terms of local policy change, while norm-based fusion balances their contributions. Denote ν^P for the empirical behavior-state distribution weighted by w^P, introduced in Section 4.3, step fractions $t_\mu, t_\pi \in [0,1]$ then constrain endpoint movement on $(\mathcal{D},\xi)$ and (ν^J,ν^P), respectively, keeping each actor's update within both channel budgets. Theorem 7 establishes bounded ascent of the frozen combined objective for the ideal fused direction and characterizes when both channels are first-order nondecreasing. Appendix B.5 specifies the fusion coefficients, Fisher operators, and endpoint checks.

4.3 Critic, State Emphasis and Initialization

The value estimates used above are constructed from the frozen critic reference v_g under $\widehat{\mathcal{P}}_g$. Let $\widehat{\mathcal{P}}_{g,z}$ denote the model kernel at fixed risk z. For a model transition with reward R_t and successor observation S_{t+1}, define the one-step backup and target-policy Bellman residual by

$$\begin{aligned} L_{g,z}v_g(s,a) &= \mathbb{E}_{\widehat{\mathcal{P}}_{g,z}}\left[R_t + \gamma v_g(S_{t+1}) \,\middle|\, S_t = s, \mathsf{A}_t = a\right], \\ e_{g,z}(s) &= \sum_{a'\in\mathcal{A}} \pi_g(a' \mid s) L_{g,z}v_g(s,a') - v_g(s). \end{aligned} \quad (10)$$

Appendix B.2.2 gives the expanded backups. Theorem 1 proves that, under its stated conditions, the expected discounted residual sum exactly equals the action-value error of the one-step backup under the frozen model. Adding this correction therefore recovers the model's exact discounted action value without requiring the reference critic to have converged:

$$Q_g^{\pi_g}(s,a) = \mathbb{E}_{Z\sim m}[L_{g,Z}v_g(s,a)] + \mathbb{E}_{Z\sim m}\mathbb{E}_{\widehat{\mathcal{P}}_{g,Z},\pi_g}\left[\sum_{k=1}^{\infty}\gamma^k e_{g,Z}(S_k) \,\middle|\, S_0 = s, \mathsf{A}_0 = a\right]. \quad (11)$$

where each rollout follows π_g after the first action. To update the critics, we simulate continuations starting from observations sampled from the current batch $\mathcal{D}_g$. The resulting training distribution $\widetilde{\zeta}_g$ describes the simulated continuation states and their retained risks (Appendix B.2.5). Each successor critic starts from its parent and is trained to minimize an expected squared TD error of the form $\mathbb{E}[(V_\theta - \mathcal{T}V_\theta)^2]$, where the expectation is over training samples generated using $\widetilde{\zeta}_g$, and $\mathcal{T}$

denotes the model-based Bellman backup used for critic fitting under π_g. After a finite number of optimization steps, the updated critics are averaged to form v_{g+1}. Appendices B.2.5–B.2.7 specify the sampling and update procedures.

Behavior and target policies can visit states with different frequencies, so behavior-data frequencies alone may not reflect state importance for target learning. HBPO learns two positive state M^J and M^P (raw masses) from $\mathcal{D}_g$. M^J carries state importance through target-policy continuation from behavior states. For selective transfer from behavior, M^P combines confidence in heuristic improvement with existing state importance propagated along behavior-policy transitions, so state emphasis in later target updates reflects both where the heuristic appears helpful and how the behavior policy moves between states. On behavior data $\mathcal{D}_g$ collected by μ_g, full-data normalization and shrinkage of the parent masses give positive, mean-one weights:

$$w_g^\chi(s) = 1 + \frac{M_g^\chi(s) - \mathbb{E}_{D_g}[M_g^\chi]}{\sqrt{\mathbb{E}_{D_g}[(M_g^\chi)^2]}}, \qquad \chi \in \{J, P\}. \tag{12}$$

where $\mathbb{E}_{D_g}$ averages over all current-state occurrences in the data. These weights induce the empirical state distributions ν_g^J and ν_g^P used above for target-return learning and selective transfer from behavior respectively. Appendices B.3.3–B.3.7 specify the mass fitting and weight construction.

To obtain the confidence used by M^P, paired simulations share random inputs to compare T-step undiscounted returns: $q_{H,g}$ for nine stages then π_g, versus π_g throughout. The paired simulations return differences yield confidence scores for heuristic improvement. These scores are interpolated using fixed state features to the successor states in $\mathcal{D}_g$ and centered around one to form the interest signal for fitting M_{g+1}^P. Appendix B.3.5 specifies the confidence and interest construction.

The recursion starts from cold initialization or a warm handoff. Both produce separately stored policies with $\mu_0 = \pi_0$, five prefitted scalar critics, and neutral emphatic masses $M_0^J = M_0^P \equiv 1$, yielding $w_0^J = w_0^P \equiv 1$ (Appendix B.3.9). Cold initialization uses a near-uniform policy and fits critics on simulated states. In DSIS-RB, the model is fitted to the same first-generation real-data batch, included in the common 2M provider-transition budget. Warm handoff copies or distills a source policy and fits fresh scalar critics using centered action information from temporary Q networks trained to evaluate π_0. Both routes finish with constant-offset calibration (Appendix B.2.8).

5 Experiments

5.1 Experimental Setup

We compare HBPO with nine native baselines on DSIS and DSIS-RB: PPO, A2C, Discrete SAC, ACE, PPO-81, TRPO, GRPO, TPO, and GPS. Warm comparisons branch from each baseline's 0.5M-interaction checkpoint into native continuation and HBPO, each using another 1.5M real interactions to assess recovery and improvement across policy sources and initializations. Cold uses 2M real interactions without native pretraining to assess learning without a source policy. Baselines use individually selected configurations; HBPO shares one configuration across all ten initializations per environment, selected on three tuning seeds to balance relative cost reductions from high-cost and non-collapsed starts among seven configuration sets. Final policies from five separate training seeds are evaluated with sampled actions; HBPO reports its target policy π. Appendix C details baseline adaptations, initialization, tuning, evaluation, and separately accounted simulation costs.

5.2 Main Results

Table 1 classifies final costs as Non-collapse ($c \leq 1010$), Collapse ($1010 < c \leq 3980$), or Severe Collapse ($c > 3980$). Thresholds 1010 and 3980 represent one and four periods of backlog-cost accumulation, respectively, indicating economic severity rather than delivery delays (Appendix C.3). Appendix D.1 provides high-precision results and cost-grade calculations.

HBPO lowers mean final cost in 17 of 18 warm source–environment combinations, median cost in 15, and sample standard deviation across seeds in all 18. Discrete SAC on DSIS-RB is the mean-cost exception (1.45% higher), despite a lower median and standard deviation. For one seed, its native endpoint is Collapse while HBPO's is Severe Collapse. **High-cost native endpoints.** Among

Table 1: Final policy costs at a matched budget of 2M environment interactions across five training seeds. H denotes warm-start HBPO; Cold denotes HBPO without a native training prefix. Seed costs are black for Non-collapse, yellow for Collapse, and red for Severe Collapse. Std is the sample standard deviation across seeds; bold marks the lower paired mean or median, computed from unrounded values. Underlining marks HBPO endpoints with costs ≤ 1010 whose seed-matched native endpoints exceed 1010. k/M denote $10^3/10^6$.

	DSIS: Mean cost per period ↓								**DSIS-RB**: Normalized risk-AUC cost ↓							
	Seed cost					Across seeds			Seed cost					Across seeds		
Algorithm	0	1	2	3	4	Mean	Median	Std	0	1	2	3	4	Mean	Median	Std
PPO	21.0	21.2	56.9	13.8k	21.3	2.79k	21.3	6.17k	70.1	75.5	67.0k	75.1	74.9	13.5k	**75.1**	29.9k
H-PPO	20.7	21.3	51.4	159	20.8	**54.6**	**21.3**	59.8	52.7	81.9	5.89k	55.3	112	**1.24k**	81.9	2.60k
A2C	23.7	28.7	23.0k	30.4	23.4	4.62k	28.7	10.3k	74.5	69.6	8.04k	807	69.6	1.81k	**74.5**	3.49k
H-A2C	32.3	20.8	6.64k	20.9	20.7	**1.35k**	**20.9**	2.96k	51.6	51.3	7.40k	126	654	**1.66k**	126	3.22k
Discrete SAC	27.8	36.7	38.6	27.5	345	95.1	36.7	140	1.18k	11.0k	71.2	69.7	102	**2.48k**	102	4.77k
H-Discrete SAC	26.6	34.0	33.1	32.4	55.4	**36.3**	**33.1**	11.1	5.36k	7.03k	61.8	58.9	66.4	2.51k	**66.4**	3.41k
ACE	1.68k	22.6	22.3	7.50k	1.12k	2.07k	1.12k	3.12k	9.96k	12.2k	22.3k	13.8k	15.5k	14.8k	13.8k	4.70k
H-ACE	1.21k	320	21.4	77.8	37.5	**333**	**77.8**	504	4.01k	110	798	121	85.1	**1.03k**	**121**	1.70k
PPO-81	22.4	22.7	21.7	23.0	21.5	22.2	22.4	0.633	191	103	96.1	95.3	95.3	116	96.1	42.1
H-PPO-81	21.4	22.0	21.2	20.9	21.5	**21.4**	**21.4**	0.418	125	80.0	86.6	84.8	84.1	**92.1**	**84.8**	18.5
TRPO	5.03k	12.5k	26.0	29.7	25.4	3.52k	**29.7**	5.46k	428	366	318	2.00k	293	680	366	738
H-TRPO	103	219	25.6	30.8	25.1	**80.7**	30.8	84.1	317	290	303	235	166	**262**	**290**	61.9
GRPO	37.6	32.9	32.3	50.7	30.9	36.9	32.9	8.11	537	1.24k	1.10k	3.23k	1.22k	1.47k	1.22k	1.03k
H-GRPO	22.5	26.8	21.0	23.2	21.9	**23.1**	**22.5**	2.24	80.8	92.3	107	93.2	222	**119**	**93.2**	58.3
TPO	4.21k	4.80k	4.56k	4.17k	4.40k	4.43k	4.40k	262	11.2k	1.62k	2.86k	1.99k	3.39k	4.22k	2.86k	3.98k
H-TPO	45.1	44.8	45.9	44.6	45.7	**45.2**	**45.1**	0.552	238	199	220	215	163	**207**	**215**	28.1
GPS	123	120	107	118	124	118	120	6.93	1.77k	6.22k	2.41k	1.92k	2.01k	2.87k	2.01k	1.89k
H-GPS	39.8	41.7	41.2	44.4	35.5	**40.5**	**41.2**	3.28	234	200	188	240	216	**216**	**216**	21.8
Cold	44.6	47.5	45.1	45.7	42.7	45.1	45.1	1.75	69.7	73.1	74.9	70.0	72.8	72.1	72.8	2.21

36 seed-matched comparisons with native final costs above 1010, HBPO achieves lower costs in 35 and costs ≤ 1010 in 29 (DSIS: 10/12; DSIS-RB: 19/24). **Low-cost native endpoints.** All 54 native endpoints with costs ≤ 1010 have seed-matched HBPO endpoints within this range, with 46 achieving lower costs. **Cold starts.** Cold avoids final cost collapse across all five seeds in both environments, with mean costs of 45.1 on DSIS and 72.1 on DSIS-RB. By median cost over all five seeds, Cold ranks seventh among Cold and the nine native baselines on DSIS and first on DSIS-RB.

5.3 Tracing Heuristic–Behavior–Target Transmission

These diagnostics characterize heuristic uptake and subsequent advantage-aligned target responses during policy recovery and stabilization. Figure 2 examines the delayed transmission chain $\pi_g \to q_{H,g} \to \mu_{g+1} \to \pi_{g+2}$ (red in Section 1). Relative steps measure update size relative to each guidance gap $\varrho_{H\mu} = \frac{\|\mu_{g+1}-\mu_g\|_H}{\|q_{H,g}-\mu_g\|_H}$, $\varrho_{\mu\pi} = \frac{\|\pi_{g+2}-\pi_{g+1}\|_J}{\|\mu_{g+1}-\pi_{g+1}\|_J}$ Here, 0 denotes no update and 1 equal update and gap norms. Larger ratios need not be better: the two updates balance A with H and J with P respectively. Weighted cosines measure alignment with the same gaps: $c_{H\mu} = \cos_H(\mu_{g+1}-\mu_g,\, q_{H,g}-\mu_g)$, $c_{\mu\pi} = \cos_J(\pi_{g+2}-\pi_{g+1},\, \mu_{g+1}-\pi_{g+1})$ Values -1, 0, and 1 indicate opposite, orthogonal, and fully aligned directions. Under the frozen target advantage $\widehat{a}_{g+1}$, $S_\pm = \mathbf{1}\{\pm(\mu_{g+1}-\pi_{g+1})\widehat{a}_{g+1} > 0\}$ identify advantageous and disadvantageous behavioral deviations. Their opportunity magnitudes and net target responses are $B_\pm = \left\langle [\pm(\mu_{g+1}-\pi_{g+1})\widehat{a}_{g+1}]_+ \right\rangle_P$, $G_\pm = \langle(\pi_{g+2}-\pi_{g+1})\widehat{a}_{g+1}S_\pm\rangle_P$. Positive G_+ and G_- indicate adoption of advantageous deviations and correction of disadvantageous ones respectively. Norms, cosines, and averages use the corresponding diagnostic weights.

Together, these observations provide evidence of effective heuristic–behavior–target transmission, whose corrective dynamics help explain recovery from policy collapse, improved average performance, and stable cold-start learning. **Existence.** In Fig. 2, broadly corresponding trends in the short-horizon q_H reference and the behavior and target costs accompany nonzero relative steps and predominantly positive alignment along $q_{H,g} \to \mu_{g+1} \to \pi_{g+2}$. Together, these indicate sustained

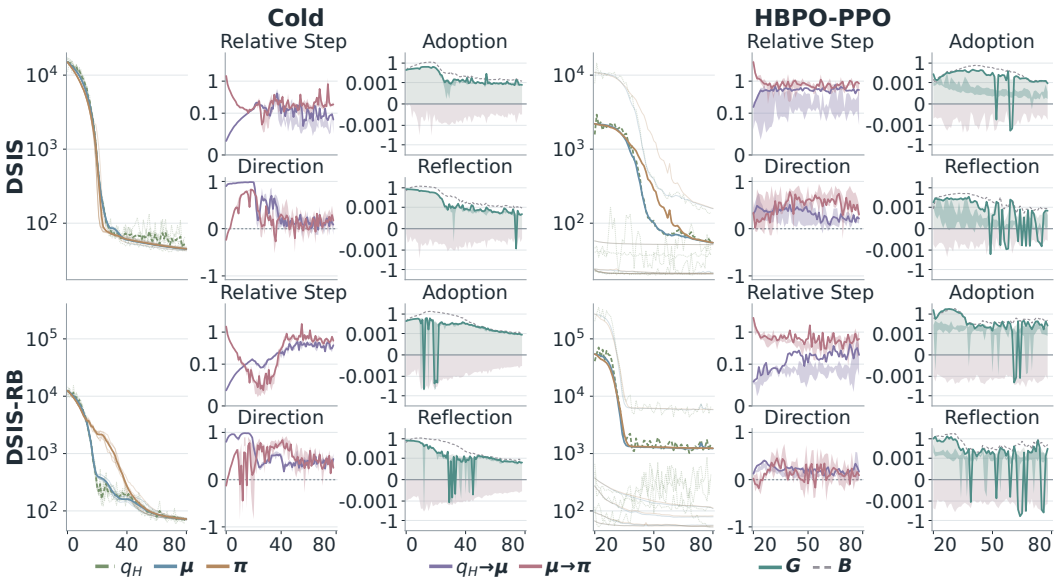

Figure 2: Transmission analysis. Left: standard evaluation costs of μ, π and a q_H reference averaging H-weighted immediate costs over the original 9-stages rollouts (lower is better). Middle: relative step ϱ (top) and direction cosine c (bottom) for $q_H \to \mu$ and $\mu \to \pi$. Right: adoption and reflection, with net responses $G_{\pm}$, dashed opportunities $B_{\pm}$, and shaded favorable/adverse contributions above/below zero.

guidance, with target improvement sometimes following behavior improvement. **Corrective effectiveness.** During the pronounced DSIS–PPO cost reduction, $\mu \to \pi$ exhibits relatively large steps and closer alignment with nonzero $q_H \to \mu$ steps, alongside mostly positive adoption and reflection responses. This connects continued heuristic uptake with target adjustments that, on balance, retain advantageous behavioral deviations and oppose disadvantageous ones. Positive reflection also explains how useful transfer can coexist with weak geometric following. **Interpretability.** In DSIS–Cold, strong early alignment and substantial selection responses give way to weaker following and smaller B, G as costs stabilize. DSIS-RB–Cold shows a complementary pattern: the target lags behind behavior when relative transfer steps are small, then catches up as these steps increase under positive alignment. Changes in transmission thus correspond to identifiable phases of cost improvement. The appendix shows related correspondences: DSIS–TPO exhibits diminishing opportunities and responses after rapid recovery, whereas DSIS–ACE combines positive adoption and reflection with weak or negative alignment and falling costs. In DSIS–TRPO, an interval of negative reflection instead accompanies subsequent cost increases, helping interpret temporary setbacks. These patterns characterize how selective transmission operates during learning. The component interventions below test its contribution to final policy performance.

5.4 Ablation Results

No-A removes target-advantage feedback from behavior learning; **No-P** removes explicit $\mu \to \pi$ transfer while retaining behavior-generated data. **KL-P** replaces P with KL imitation at a matched natural-gradient norm. **Near-on-policy** uses the newly updated behavior policy for immediate KL transfer, then sets $\mu = \pi$. In Figure 3, Full consistently outperforms all four ablations from Cold and avoids the high-cost failures under No-P and Near-on-policy in DSIS–TRPO. Together with broad degradation under No-A, these results support target-informed behavior learning and explicit transfer beyond data collection. KL-P and Near-on-policy often yield slightly lower costs from PPO and SAC starts, indicating a trade-off with protection against high-cost failures. Mean paired relative-cost changes favor Full for all four ablations.

No-H removes H and its weighting and geometry from the Fuse behavior update while retaining behavior–target transmission. **Random-H** generates random guidance targets around μ, matching the statewise distance of the simulation-derived q_H from μ. Both increase mean final cost in most

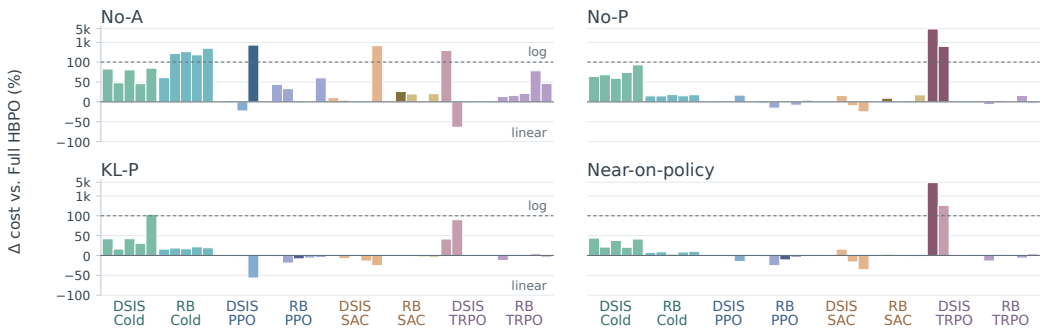

Figure 3: Final target-policy cost changes (%) relative to matched Full HBPO; positive values favor Full. Each color represents one environment–source, with seeds 0–4 from left to right. Dark bars indicate ablation costs above 1010. Shared axes are linear up to 100% and logarithmic above.

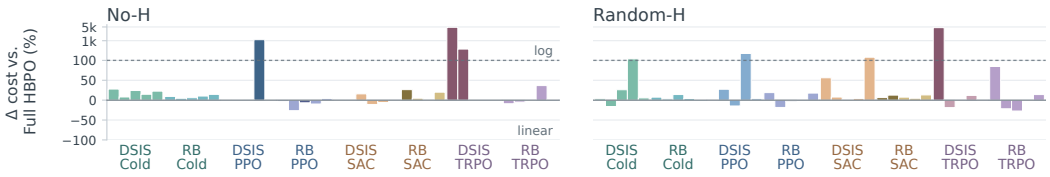

Figure 4: Heuristic-guidance ablations, No-H and Random-H. Plotting conventions follow Figure 3.

tested settings. In these DSIS ablations, No-H and Random-H produce three and one collapsed endpoints, respectively, versus none for Full. Random-H occasionally outperforms Full but costs more in 32 of 40 matched runs, supporting the value of informative guidance directions.

6 Discussion and Limitations

First, generation-level diagnostics and component ablations together support heuristic–behavior–target transmission as a mechanism contributing to policy recovery and improvement. The delayed path follows from the frozen-parent update schedule; the experiments assess the coupled learning mechanism and its component channels. Emphatic weighting adapts the principle of propagating state interest through policy-dependent transitions to HBPO's two learning roles, providing separate state emphasis for target-return learning and heuristic-informed transfer. The ablations in Appendix D.3 show mixed endpoint effects, so a consistent final-cost benefit is not established in the evaluated settings. We adopt this pathway considering its theoretical advantages, which not obviously show in experiment endpoint. Achieving sustained state-distribution correction remains an opportunity to strengthen this framework. Second, HBPO improves 35 of 36 comparisons with high-cost native endpoints. For DSIS-RB SAC seed 0, cost falls from 7.76k at handoff to 5.36k but exceeds native continuation's 1.18k, highlighting limits to recovery across initializations. Third, the coupling rules are not inventory-specific, though HBPO uses structural priors in $\widehat{\mathcal{P}}$. Shared information access and comparison with model-informed GPS support the architectural comparison under common modeling assumptions. Isolating the contribution of structural priors and testing other domains remain future directions.

7 Conclusion

We introduced HBPO, built on an explicit delayed heuristic–behavior–target coupling mechanism. Across DSIS and DSIS-RB, HBPO recovers many high-cost policies, improves overall performance and avoids final collapse in all tested cold starts. Transmission diagnostics trace how heuristic guidance shapes behavior learning and subsequent selective target updates, helping explain these gains. Ablations further support the contributions of informative heuristic guidance, target-informed behavior learning, and explicit behavior–target transfer to performance and robustness.

AI USE STATEMENT

We used OpenAI Codex with GPT-5.5 Extra high, GPT-5.6 Sol Ultra and GPT-6 Astra Ultra, primarily in regular mode and occasionally in fast mode. These were the only generative AI assistants used in this work. Codex assisted formalizing the mathematics of the authors' method, developing and checking proofs, and preparing and inspecting the appendix. It also generated the method implementation and ran and debugged experiments under the authors' supervision. The authors polished the main section, identified and reviewed related work, and checked references, with assistance from Codex. The authors conceived the method and reinforcement-learning framework, designed the experiments, analyzed the results, and produced the figures.

The authors carefully drafted the main section, reviewed the manuscript, read the AI-assisted proofs, and manually worked through selected derivations. Their code review consisted of inspecting key implementation components. The authors take full responsibility for the final manuscript and accompanying code, including the mathematical claims, references, and experimental results.

REPRODUCIBILITY STATEMENT

Appendix A summarizes the notation. Appendix B specifies the algorithm, benchmark dynamics, information access, initialization procedures, and mathematical assumptions and derivations. Appendix C describes hyperparameter selection, interaction budgets, random seeds, and evaluation. Appendix D reports extended results, and Appendix E details the ablation interventions and diagnostic procedures. An anonymized code package is included in the supplementary material, with the implementation, experiment configurations, dependency specifications, and instructions for running training and evaluation. The code will be made publicly available upon acceptance.

REFERENCES

Abbas Abdolmaleki, Jost Tobias Springenberg, Yuval Tassa, Remi Munos, Nicolas Heess, and Martin Riedmiller. Maximum a posteriori policy optimisation. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=S1ANxQW0b.

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2–3):235–256, 2002. doi: 10.1023/A:1013689704352. URL https://doi.org/10.1023/A:1013689704352.

Marc G. Bellemare, Sriram Srinivasan, Georg Ostrovski, Tom Schaul, David Saxton, and Remi Munos. Unifying count-based exploration and intrinsic motivation. In *Advances in Neural Information Processing Systems*, volume 29, 2016. URL https://arxiv.org/abs/1606.01868.

Lucas Böttcher, Thomas Asikis, and Ioannis Fragkos. Control of dual-sourcing inventory systems using recurrent neural networks. *INFORMS Journal on Computing*, 35(6):1308–1328, 2023. doi: 10.1287/ijoc.2022.0136. URL https://pubsonline.informs.org/doi/10.1287/ijoc.2022.0136.

Yuri Burda, Harrison Edwards, Amos Storkey, and Oleg Klimov. Exploration by random network distillation. In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1810.12894.

Petros Christodoulou. Soft actor-critic for discrete action settings. *arXiv preprint arXiv:1910.07207*, 2019. URL https://arxiv.org/abs/1910.07207.

Thomas Degris, Martha White, and Richard S. Sutton. Off-policy actor-critic. In *Proceedings of the 29th International Conference on Machine Learning*, 2012. URL https://arxiv.org/abs/1205.4839.

Lasse Espeholt, Hubert Soyer, Remi Munos, Karen Simonyan, Volodymyr Mnih, Tom Ward, Yotam Doron, Vlad Firoiu, Tim Harley, Iain Dunning, Shane Legg, and Koray Kavukcuoglu. IMPALA: Scalable distributed deep-RL with importance weighted actor-learner architectures. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of*

Machine Learning Research, pp. 1407–1416. PMLR, 2018. URL https://proceedings.mlr.press/v80/espeholt18a.html.

Scott Fujimoto, Herke van Hoof, and David Meger. Addressing function approximation error in actor-critic methods. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1587–1596. PMLR, 2018. URL https://proceedings.mlr.press/v80/fujimoto18a.html.

Kevin Geevers, Lotte van Hezewijk, and Martijn R. K. Mes. Multi-echelon inventory optimization using deep reinforcement learning. *Central European Journal of Operations Research*, 32:653–683, 2024. doi: 10.1007/s10100-023-00872-2. URL https://link.springer.com/article/10.1007/s10100-023-00872-2.

Joren Gijsbrechts, Robert N. Boute, Jan A. Van Mieghem, and Dennis J. Zhang. Can deep reinforcement learning improve inventory management? Performance on lost sales, dual-sourcing, and multi-echelon problems. *Manufacturing & Service Operations Management*, 24(3):1349–1368, 2022. doi: 10.1287/msom.2021.1064. URL https://pubsonline.informs.org/doi/10.1287/msom.2021.1064.

Alex W. Goodall, Edwin Hamel-De Le Court, and Francesco Belardinelli. Behaviour policy optimization: Provably lower variance return estimates for off-policy reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 21324–21332, 2026. doi: 10.1609/aaai.v40i26.39278. URL https://ojs.aaai.org/index.php/AAAI/article/view/39278.

Eric Graves, Ehsan Imani, Raksha Kumaraswamy, and Martha White. Off-policy actor-critic with emphatic weightings. *Journal of Machine Learning Research*, 24(146):1–63, 2023. URL https://jmlr.org/papers/v24/21-1350.html.

Audrunas Gruslys, Will Dabney, Mohammad Gheshlaghi Azar, Bilal Piot, Marc Bellemare, and Remi Munos. The reactor: A fast and sample-efficient actor-critic agent for reinforcement learning. In *International Conference on Learning Representations*, 2018. URL https://arxiv.org/abs/1704.04651.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1861–1870. PMLR, 2018. URL https://proceedings.mlr.press/v80/haarnoja18b.html.

Ehsan Imani, Eric Graves, and Martha White. An off-policy policy gradient theorem using emphatic weightings. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://papers.neurips.cc/paper_files/paper/2018/hash/3ef815416f775098fe977004015c6193-Abstract.html.

Ray Jiang, Shangtong Zhang, Veronica Chelu, Adam White, and Hado van Hasselt. Learning expected emphatic traces for deep RL. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 7015–7023, 2022. doi: 10.1609/aaai.v36i6.20660. URL https://ojs.aaai.org/index.php/AAAI/article/view/20660.

Jean Kaddour. Target policy optimization. *arXiv preprint arXiv:2604.06159*, 2026. URL https://arxiv.org/abs/2604.06159v1.

Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *Proceedings of the 19th International Conference on Machine Learning*, pp. 267–274, 2002. URL https://people.eecs.berkeley.edu/~pabbeel/cs287-fa09/readings/KakadeLangford-icml2002.pdf.

Sham M. Kakade. A natural policy gradient. In *Advances in Neural Information Processing Systems*, volume 14, 2001. URL https://proceedings.neurips.cc/paper/2001/hash/4b86abe48d358ecf194c56c69108433e-Abstract.html.

Martin Klissarov, Rasool Fakoor, Jonas W Mueller, Kavosh Asadi, Taesup Kim, and Alexander Smola. Adaptive interest for emphatic reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 95–108. Curran Associates, Inc., 2022. doi: 10.52202/06843 1-0008. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/008079ec00eec9760ee93af5434ee932-Paper-Conference.pdf.

Sergey Levine and Vladlen Koltun. Guided policy search. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pp. 1–9. PMLR, 2013. URL https://proceedings.mlr.press/v28/levine13.html.

Sergey Levine, Chelsea Finn, Trevor Darrell, and Pieter Abbeel. End-to-end training of deep visuomotor policies. *Journal of Machine Learning Research*, 17(39):1–40, 2016. URL https://jmlr.org/papers/v17/15-522.html.

Timothy P. Lillicrap, Jonathan J. Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. In *International Conference on Learning Representations*, 2016. URL https://arxiv.org/abs/1509.02971.

Anji Liu, Yitao Liang, and Guy Van den Broeck. Off-policy deep reinforcement learning with analogous disentangled exploration. In *Proceedings of the 19th International Conference on Autonomous Agents and Multi-Agent Systems*, 2020. URL https://starai.cs.ucla.edu/papers/LiuAAMAS20.pdf.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015. doi: 10.1038/nature14236. URL https://www.nature.com/articles/nature14236.

Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1928–1937. PMLR, 2016. URL https://proceedings.mlr.press/v48/mniha16.html.

William H. Montgomery and Sergey Levine. Guided policy search via approximate mirror descent. In *Advances in Neural Information Processing Systems*, volume 29, 2016. URL https://papers.neurips.cc/paper_files/paper/2016/hash/a00e5eb0973d24649a4a920fc53d9564-Abstract.html.

Rémi Munos, Tom Stepleton, Anna Harutyunyan, and Marc G. Bellemare. Safe and efficient off-policy reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 29, 2016. URL https://proceedings.neurips.cc/paper/2016/file/c3992e9a68c5ae12bd18488bc579b30d-Paper.pdf.

Deepak Pathak, Pulkit Agrawal, Alexei A. Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 2778–2787. PMLR, 2017. URL https://proceedings.mlr.press/v70/pathak17a.html.

Tom Schaul, John Quan, Ioannis Antonoglou, and David Silver. Prioritized experience replay. In *International Conference on Learning Representations*, 2016. URL https://arxiv.org/abs/1511.05952.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1889–1897. PMLR, 2015. URL https://proceedings.mlr.press/v37/schulman15.html.

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

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. URL https://arxiv.org/abs/1707.06347.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. URL https://arxiv.org/abs/2402.03300.

Richard S. Sutton. Dyna, an integrated architecture for learning, planning, and reacting. *ACM SIGART Bulletin*, 2(4):160–163, 1991. doi: 10.1145/122344.122377. URL https://doi.org/10.1145/122344.122377.

Richard S. Sutton, David A. McAllester, Satinder P. Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems*, volume 12, 1999. URL https://proceedings.neurips.cc/paper/1999/hash/464d828b85b0bed98e80ade0a5c43b0f-Abstract.html.

Richard S. Sutton, A. Rupam Mahmood, and Martha White. An emphatic approach to the problem of off-policy temporal-difference learning. *Journal of Machine Learning Research*, 17(73):1–29, 2016. URL https://jmlr.org/papers/v17/14-488.html.

Hado van Hasselt, Arthur Guez, and David Silver. Deep reinforcement learning with double q-learning. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, volume 30, pp. 2094–2100, 2016. doi: 10.1609/aaai.v30i1.10295. URL https://ojs.aaai.org/index.php/AAAI/article/view/10295.

Ziyu Wang, Victor Bapst, Nicolas Heess, Volodymyr Mnih, Remi Munos, Koray Kavukcuoglu, and Nando de Freitas. Sample efficient actor-critic with experience replay. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=HyM25Mqel.

William F. Whitney, Michael Bloesch, Jost Tobias Springenberg, Abbas Abdolmaleki, Kyunghyun Cho, and Martin Riedmiller. Decoupled exploration and exploitation policies for sample-efficient reinforcement learning. *arXiv preprint arXiv:2101.09458*, 2021. URL https://arxiv.org/abs/2101.09458.

Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3–4):229–256, 1992. doi: 10.1007/BF00992696. URL https://doi.org/10.1007/BF00992696.

Tianbing Xu, Qiang Liu, Liang Zhao, and Jian Peng. Learning to explore via meta-policy gradient. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 5463–5472. PMLR, 2018. URL https://proceedings.mlr.press/v80/xu18d.html.

Shangtong Zhang, Wendelin Boehmer, and Shimon Whiteson. Generalized off-policy actor-critic. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/0e095e054ee94774d6a496099eb1cf6a-Abstract.html.

A NOTATION

Table 2 summarizes the benchmark, model, critic, and policy-update notation. Its main entries follow Appendix B; the final groups record notation used in the main text and the other appendices. Scope-qualified entries distinguish local uses of the same letter rather than identifying different mathematical objects.

Table 2: Notation and scope. The generation-g model, policies, reference critic, and sampling measures are frozen when constructing that generation's learning signals.

Symbol	Meaning and convention
Indices, observations, actions, and physical dynamics	
$g;\ t;\ k$	Learning generation; physical period; simulated period, optimization iteration, or sampled-root index, as specified locally.
$j;\ b;\ \chi$	Critic member $j \in \{1, \dots, 5\}$; simulation-tape index; weighting role (J for target return, P for transmission, and 0 for the unweighted behavior-return role). The risk grid separately uses $j = 0, \dots, 5$.
$\mathcal{S};\ s, S;\ s', S'$	Observation space; an observation and its random counterpart; successor observation and its random counterpart. In the residual theorem, $\mathcal{S}$ is countable and closed under the frozen model's transitions.
$I_t;\ \mathbf{p}_t = (p_{t,1}, \dots, p_{t,4})$	Net inventory (negative values represent backlog) and the four-stage regular-order pipeline.
$B_t;\ B(s);\ u, U$	Current binary regular-supply availability, its observation-coordinate accessor, and realized/random next availability. Zero means unavailable and one means available.
$\mathsf{D}_t, d_t;\ d^-(s)$	Random demand and its realization; the previous-demand coordinate of an RB observation. The demand coordinate is omitted in DSIS.
$\mathcal{A};\ a, \mathsf{A};\ a_t = (q_t^r, q_t^e)$	The 81-action set $\{0, \dots, 8\}^2$; a realized/random action; regular and expedited order quantities. The symbols q^r, q^e denote quantities only.
$F(s, a, d, u)$	Deterministic observation-transition map from the public inventory balance, pipeline shift, current-availability order gate, and next availability u.
$c(s, a, d), c_t;\ r(s, a, d), r_t;\ R, R_k$	Period cost and its realization; normalized reward and its realization; random one-step or step-k model reward. Here $r = -c/495$, the common unit for values and utilities.
$[x]_+;\ \gamma$	Positive part $\max\{x, 0\}$; critic discount factor, equal to 0.995 in the implementation.
$\delta_x;\ \mathbf{1}\{E\};\ \otimes$	Unit point mass at x; indicator of statement E; product of probability measures.
Benchmark laws and risk stratification	
$\rho;\ X_t = (S_t, \rho)$	Actual environment risk, fixed within a trajectory; the corresponding augmented physical state. Risk is not observed by the policy or critic.
$\mathcal{D}_{\mathrm{RB}};\ \omega_D$	True RB demand support $\{0, 1, 2, 3, 8\}$ and stationary demand distribution $(0.3, 0.2, 0.2, 0.2, 0.1)$ in that displayed order.
$K_D(x, d);\ K_{B,\rho}(b_{\mathrm{cur}}, u)$	True demand-transition probability given previous demand x; true supply-transition probability given current availability b_{cur} and actual risk ρ.
$\mathcal{P}_\rho(dr, ds' \mid s, a)$	True joint reward–observation transition kernel conditional on environment risk ρ; it excludes scheduled episode resets.
$H_{\mathrm{env}};\ \rho_j$	Physical episode length, 1000 periods; registered risk grid $\rho_j = 0.1j$, $j = 0, \dots, 5$.
$\omega_{\rho,j};\ \omega_\rho;\ \nu_{\mathrm{env}}$	Risk-grid weight, the weight vector $(0.1, 0.2, 0.2, 0.2, 0.2, 0.1)$, and the target risk-composition measure $\sum_j \omega_{\rho,j}\delta_{\rho_j}$.
$\mathbf{n}^{(0)}, \mathbf{n}^{(1)}$	Alternating 25-trajectory risk-stratum count vectors, $(3, 5, 5, 5, 5, 2)$ and $(2, 5, 5, 5, 5, 3)$; two blocks have exactly the target risk proportions.

Continued on next page

Table 2 (continued)

Symbol	Meaning and convention
$J_\pi(\rho_j);\ \mathcal{R}_{\mathrm{trap}}(\pi)$	Expected episode mean cost under fixed policy π at risk ρ_j; normalized trapezoidal risk score $\sum_j \omega_{\rho,j} J_\pi(\rho_j)$.
Y_ℓ	Realized mean cost of trajectory ℓ in the fixed-policy, 50-trajectory stratification calculation.
Observed-data model and persistent synthetic risk	
$\mathscr{D}^{\mathrm{real}}_{\le g};\ N^{\mathrm{real}}_g$	Multiset of all counted real transitions through generation g, including the initialization prefix when applicable; its size.
$\widehat{\mathcal{D}}_g;\ d_{\max}$	Demand support discovered from current and successor previous-demand coordinates; its maximum, used in the inventory-growth bound.
$\mathcal{C}_g(x,y);\ \mathcal{C}_g(x)$	Number of observed demand transitions $x \to y$; their row total $\sum_y \mathcal{C}_g(x,y)$.
$\widehat{\omega}_g;\ \alpha_D$	Pooled empirical next-demand distribution; total pseudocount mass per fitted demand row, equal to one.
$\widehat{K}_{D,g}(x,y)$	Fitted demand-transition probability, $(\mathcal{C}_g(x,y) + \alpha_D \widehat{\omega}_g(y))/(\mathcal{C}_g(x) + \alpha_D)$, for $x, y \in \widehat{\mathcal{D}}_g$.
$\mathcal{C}^{\downarrow}_g;\ \mathcal{C}^{\mathrm{rec}}_g$	Counts of observed transitions starting in an outage and of those that recover to availability.
$\kappa^{\mathrm{rec}};\ \widehat{\kappa}^{\mathrm{rec}}_g$	Common recovery probability in the fitted supply family; its smoothed estimate $(\mathcal{C}^{\mathrm{rec}}_g + 1)/(\mathcal{C}^{\downarrow}_g + 2)$. The superscript distinguishes it from heuristic hyperparameters.
$Z, z;\ m$	Synthetic risk and its realization; its prior $m = \mathrm{Unif}[0, 1/2]$. Each tape draws risk independently at its query root and retains that risk throughout the tape.
$\widehat{K}_{B,g,z}(b_{\mathrm{cur}}, u)$	Fitted supply-transition probability conditional on synthetic risk z; recovery probability is $\widehat{\kappa}^{\mathrm{rec}}_g$ and available-to-outage probability is $\widehat{\kappa}^{\mathrm{rec}}_g z/(1-z)$.
$\widehat{p}_g(d, u \mid s, a, z)$	Modeled joint demand/next-availability probability mass function, factoring into the fitted demand and supply kernels; one-variable forms denote its marginals.
$\widehat{\mathcal{P}}_{g,z}(dr, ds' \mid s, a)$	Frozen fitted reward–observation kernel conditional on one retained synthetic risk z.
$\widehat{\mathcal{P}}_g$	Full simulation specification: a query observation, an independent root draw $Z \sim m$, and subsequent transitions retaining Z. It does not denote an observation-only Markov kernel with risk redrawn at each step.
$\widetilde{\mathcal{P}}_g(dr, ds', dz' \mid s, z, a)$	Augmented reward–observation–risk kernel $\widehat{\mathcal{P}}_{g,z}(dr, ds' \mid s, a)\delta_z(dz')$, preserving risk.
Policies, scalar critics, and Bellman operators	
$\mu_g;\ \pi_g;\ \pi_0;\ \sigma$	Frozen parent behavior policy; frozen parent target policy; initial categorical policy; a generic fixed policy used to define an augmented transition kernel.
$\phi(s);\ d_\phi$	Fixed normalized network feature vector; its dimension, 7 in RB and 6 in DSIS.
$V_\theta;\ \theta$	A live scalar critic and its parameter vector $(W_1, b_1, W_2, b_2, w_3, b_3)$: two width-128 Tanh layers followed by a scalar linear output.
$W_1, W_2;\ b_1, b_2;\ w_3, b_3$	Hidden-layer weight matrices and biases; output weight vector and scalar bias of V_θ. Their dimensions are specified with the network definition.
$V_{g,j}, \theta_{g,j};\ v_g$	Generation-g critic member and its parameters; frozen reference $v_g = \frac{1}{5}\sum_{j=1}^{5} V_{g,j}$, averaging outputs rather than parameters.
$L^u_g V(s,a)$	Discounted action backup integrating model demand at fixed next availability u: expected reward plus γ times successor value.
$L_{g,z} V(s,a)$	Discounted action backup obtained from L^u_g by also integrating next availability under $\widehat{K}_{B,g,z}$.

Continued on next page

Table 2 (continued)

Symbol	Meaning and convention
$\mathcal{T}_g^u V;\ \mathcal{T}_{g,z}^{\pi_g} V$	Target-policy averages of $L_g^u V$ and $L_{g,z} V$, respectively; the former holds next availability fixed, whereas the latter integrates it conditional on risk z.
$\widehat{P}_{g,z}^a;\ \widehat{P}_{g,z}^{\pi_g}$	One-step state-transition operators acting on scalar functions: first action forced to a, or sampled from π_g, under fixed risk z. They contain neither a reward term nor a discount factor.
$\mathrm{Id};\ (\widehat{P}_{g,z}^{\pi_g})^\ell$	Identity operator on scalar functions; ℓ successive target-policy transition operators, with zeroth power equal to Id.
$\mathbb{E}_z^{s,a};\ \mathbb{E}_z^s$	Frozen-model rollout expectation at fixed risk z from observation s, with first action forced to a, or also sampled from π_g; all later actions follow π_g.
Conditional values, residual completion, and sampled advantages	
$V_{g,z}^{\pi_g};\ \Delta_{g,z}$	Exact infinite-horizon discounted model value at fixed risk z; conditional value error $V_{g,z}^{\pi_g} - v_g$.
$e_{g,z}$	Frozen-reference Bellman residual $\mathcal{T}_{g,z}^{\pi_g} v_g - v_g$ (backup minus prediction), integrating the target action and model disturbances conditional on z.
C_v	Finite uniform bound on $\lvert v_g(s)\rvert$ in the residual-completion theorem; it does not bound exact values or rewards uniformly over all initial states.
$\bar{V}_g;\ c_{\mathrm{val}}$	Root-risk-averaged exact conditional value $\mathbb{E}_{Z\sim m}[V_{g,Z}^{\pi_g}]$; a scalar prediction used in the conditional-value variance calculation.
$Q_g^{\pi_g};\ Q_g^0$	Exact discounted model action value with an independent persistent root risk; one-step base $\mathbb{E}_{Z\sim m}[L_{g,Z} v_g]$.
n_{tr}	Deterministic residual-continuation truncation index used in the proof; the remaining evaluation error is at step $n_{\mathrm{tr}} + 1$.
N, N_b	Geometric residual-continuation length, with $\Pr(N = \ell) = (1-\gamma)\gamma^\ell$ for $\ell \geq 0$; its draw on tape b. Geometric survival supplies the discount.
$M;\ Z_b;\ S_{b,k}^a$	Number of residual tapes per sampled query, $M = 32$; tape-b risk; its step-k observation after forced first action a and target-policy continuation.
$\widehat{Q}_g$	Ideal Monte Carlo action-value estimator: the dense base plus the average geometrically stopped sum of unweighted residuals. The implementation may retain only its centered correction.
$A_g^{\mathrm{adv}};\ A_g^{\mathrm{base}}$	Exact-model and one-step-base advantages, each centered over actions using $\pi_g(\cdot \mid s)$.
$\widehat{A}_g$	Sampled advantage $\widehat{Q}_g(s,a) - \sum_{a'} \pi_g(a' \mid s)\widehat{Q}_g(s,a')$.
$\Delta_g^Q(s,a)$	Actionwise evaluation error $Q_g^{\pi_g}(s,a) - Q_g^0(s,a)$; centering removes it exactly only when it is constant across actions at the given state.
Emphatic root measures and successor-critic fitting	
$n;\ s_i;\ D_g$	Number of generation-batch occurrences, $n = 25{,}000$; occurrence-i observation; empirical measure $D_g = n^{-1}\sum_i \delta_{s_i}$, retaining repetitions.
$w_i^\chi;\ \nu_g^\chi$	Frozen normalized occurrence weight and weighted root measure $n^{-1}\sum_i w_i^\chi \delta_{s_i}$, for $\chi \in \{J, P\}$; $n^{-1}\sum_i w_i^\chi = 1$.
$\widetilde{D}_g;\ \widetilde{\nu}_g^\chi$	Augmented root measures $D_g \otimes m$ and $\nu_g^\chi \otimes m$, initially independent in observation and synthetic risk.
$\widetilde{P}_{g,\sigma};\ \lambda$	Policy-σ transition kernel on observation–risk pairs, preserving risk; a generic augmented-state distribution. The measure $\lambda\widetilde{P}_{g,\sigma}$ is its one-step successor law.
$\widetilde{\eta}_g$	Augmented distribution after one consumer transition: the equal mixture of roots/first-action policies (D_g, μ_g), (ν_g^J, π_g), and (ν_g^P, μ_g).

Continued on next page

Table 2 (continued)

Symbol	Meaning and convention
$\widetilde{\zeta}_g$	Critic-training state–risk law $(1-\gamma)\sum_{\ell\geq 0}\gamma^\ell \widetilde{\eta}_g \widetilde{P}^\ell_{g,\pi_g}$; its zero-index term is already after the first transition.
$G;\ n_\zeta;\ (S_\ell, U_\ell)$	Additional geometric target-policy transition count when sampling a training-bank row; bank size $n_\zeta = 25{,}000$; retained observation and a newly sampled next availability under that row's retained risk.
$\varepsilon_g(\theta; s, u)$	Live fitting error $V_\theta(s) - (\mathcal{T}_g^u V_\theta)(s)$ (prediction minus backup), distinct from the frozen-reference residual $e_{g,z}$.
$\mathcal{J}_{V,g};\ \widehat{\mathcal{J}}_{V,g}$	Population and empirical means of ε_g^2 for one critic member. Actions and demand are integrated before squaring; paired next availability is sampled.
$\theta_j^{(k)};\ V_{g+1,j};\ v_{g+1}$	Live scalar parameters after k updates from $\theta_{g,j}$; member after 1000 updates; successor output mean, available to consumers in the next generation.
$\beta_1, \beta_2;\ \epsilon_{\mathrm{Adam}}$	Adam first/second-moment decay coefficients $(0.9, 0.999)$ and denominator stabilizer 10^{-8}.
Warm/cold initialization and temporary fitted Q evaluation	
$n_0;\ (s_i, u_i)$	Number of initialization pairs, 500,000; observation/next-availability pair, recorded in warm initialization and model-generated in cold initialization.
$d_i;\ s'_{i,a}, r_{i,a}$	Demand realized in recorded source transition i; all-action factual counterfactuals $F(s_i, a, d_i, u_i)$ and $r(s_i, a, d_i)$.
$Q_j^{(k)}, Q_j^{-,(k)};\ Q_j$	Temporary online and target FQE networks at iteration k; final online network $Q_j = Q_j^{(2000)}$. These networks are discarded after warm initialization.
$\psi_j^{(k)}, \psi_j^{-,(k)}$	Online/target FQE parameter vectors. Target parameters are updated after each online step; these are distinct from scalar-critic parameters θ.
$y_{j,i,a}^{(k)}$	Detached FQE reward-plus-bootstrap target using the frozen initial categorical policy π_0 and target network $Q_j^{-,(k)}$.
$\mathcal{I}^{(k)};\ n_{\mathrm{b}};\ h$	Ordered list of uniformly sampled source-occurrence indices, with replacement, at FQE iteration k; its size 256; a position within that list. All five FQE members share the indices.
$\xi_{j,h,a}^{(k)}$	Independent $\mathrm{Bernoulli}(0.8)$ FQE mask for member j, minibatch position h, action a, and iteration k. Repeated source indices may receive different masks.
$\mathcal{L}_j^{\mathrm{FQE},(k)}$	Masked FQE squared-error loss, normalized by $n_{\mathrm{b}}\lvert\mathcal{A}\rvert$, including the masked-out pairs in the denominator.
A_j^{init}	Frozen centered action label from final online FQE member j: $Q_j(s,a) - \sum_{a'}\pi_0(a' \mid s)Q_j(s,a')$.
$L_{\mathrm{init}}^u;\ \mathcal{T}_{\mathrm{init}}^u$	Supply-conditioned action backup using the initialization model; its average over the initial policy π_0.
$\mathcal{J}_j^{\mathrm{warm}};\ \mathcal{J}_j^{\mathrm{cold}}$	Full-gradient scalar initialization objectives: actionwise fitting to A_j^{init} for warm initialization; squared policy-averaged Bellman discrepancy for cold initialization.
$\widetilde{V}_j;\ \overline{e}_j;\ c_{\mathrm{off}}$	Scalar member before offset calibration; its full-bank mean backup-minus-prediction residual; a constant output shift, chosen as $\overline{e}_j/(1-\gamma)$.
$V_{0,j}$	Calibrated initial critic member $\widetilde{V}_j + \overline{e}_j/(1-\gamma)$; its empirical mean residual vanishes in exact arithmetic, not necessarily pointwise.
Heuristic utility and control variates	
$H;\ G_{H,g}(s,a)$	Heuristic horizon, $H = 1000$; expected undiscounted H-period normalized reward after first action a, averaging independent root risk and target-policy continuation.
$v_{g,k}$	Heuristic control function, equal to v_g for $k < H$ and zero at $k = H$.

Continued on next page

Table 2 (continued)

Symbol	Meaning and convention
$L^{\mathrm{H}}_{g,z,k}(s,a)$	Undiscounted control backup $\mathbb{E}[R + v_{g,k+1}(S') \mid s,a,z]$. Unlike $L_{g,z}V$, it is the evaluated backup rather than an operator with a free value-function argument.
$e^{\mathrm{H}}_{g,z,k}(s)$	Heuristic control residual $\sum_{a'} \pi_g(a' \mid s) L^{\mathrm{H}}_{g,z,k}(s,a') - v_{g,k}(s)$; its successor-control coefficient is one, not γ.
$\widetilde{G}_{H,g,z};\ \widehat{G}_{H,g}$	One-tape residual-control utility estimator at fixed risk; average over 32 independently drawn tape risks and their coupled action continuations.
$h_g;\ a_\star$	Estimated heuristic utility centered over actions using π_g; fixed reference action used when accumulating equivalent action differences.
Actor surrogate and shared critic evidence	
$\vartheta, \vartheta_g;\ \pi_\vartheta$	Actor parameters and frozen parent parameters; the parameterized target policy, with $\pi_g = \pi_{\vartheta_g}$.
υ	Fixed state-sampling distribution for the local actor surrogate.
$\mathcal{J}^{\mathrm{actor}}_g(\vartheta; f, \upsilon)$	Frozen-state surrogate $\mathbb{E}_{S\sim\upsilon}[\sum_a \pi_\vartheta(a \mid S) f(S,a)]$, for a frozen scalar action-evaluation function f.
$\mathrm{sc}_g(s,a)$	Parent-policy score vector $\nabla_\vartheta \log \pi_\vartheta(a \mid s)\big\vert_{\vartheta=\vartheta_g}$.
$K_{\mathrm{res}};\ \varpi_i;\ \mathsf{I}_k$	Number of sampled residual roots, 128; root probability $(1 + w^J_i + w^P_i)/(3n)$; independently sampled occurrence index for root k, with replacement.
$w^0_i;\ \mathrm{coef}_i(a)$	Unit weight for the unweighted behavior-return role; prescribed scalar coefficient for occurrence i and action a, independent of the shared sampling when unbiasedness is asserted.
$\widehat{A}^{(k)}_g$	Sampled centered advantage from the independent continuation bank attached to sampled root occurrence k; repeated selections receive distinct banks.
$\widehat{\Lambda}^{\chi}_g[\mathrm{coef}]$	Shared linear critic-evidence functional for role $\chi \in \{0, J, P\}$: a dense weighted base plus an importance-weighted sampled centered correction.
Main-text actors, weighting, and coupled updates	
$\mathcal{X};\ x_t;\ \mathcal{P}$	Full environment state space, state, and transition kernel in the problem formulation. In RB, $x_t = (s_t, \rho)$; in DSIS, $x_t = s_t$. These differ from the observation space $\mathcal{S}$ and fitted model specification $\widehat{\mathcal{P}}_g$.
$\mathcal{Y}_g$	Persistent learner tuple $(\mu_g, \pi_g, V_{g,1:5}, M^J_g, M^P_g)$ in Methods; it is not the physical state X_t.
$\theta^\mu_g, \theta^\pi_g$	Behavior/target actor parameters in Methods. The target parameter corresponds to ϑ_g in Appendix B; unqualified θ in the critic construction denotes scalar-critic parameters.
$\mathcal{D}_g;\ w^t_g(s)$	Methods notation for the behavior-data panel and normalized state weight, $t \in \{J, P\}$. Empirical observation averaging corresponds to D_g in Appendix B, and occurrence i has weight $w^t_i = w^t_g(s_i)$.
M^J_g, M^P_g	Positive learned raw emphatic masses for target-return and transmission weighting; the normalized consumer weights w^J_g, w^P_g are derived from these masses.
$T;\ G_{T,g};\ \widehat{G}_{T,g};\ u_g$	Methods notation for the heuristic horizon, exact utility, estimated utility, and centered estimated utility. These correspond to $H, G_{H,g}, \widehat{G}_{H,g}, h_g$ in Appendix B, respectively.
$\Delta(\mathcal{A});\ q;\ q_{H,g}$	Probability simplex over actions; a candidate distribution in the Methods heuristic optimization; its Hellinger-regularized controller. These local uses of q are distributions, whereas q^r, q^e are order quantities.
$\tau;\ \kappa$	Frozen run-level heuristic utility/regularization scale; the separately tuned heuristic scale hyperparameter in the protocol. Neither is the fitted supply recovery probability $\widehat{\kappa}^{\mathrm{rec}}_g$.

Continued on next page

Table 2 (continued)

Symbol	Meaning and convention
ξ_g	Empirical distribution of states from short heuristic-controller rollouts in Methods. It is distinct from the indexed binary FQE mask $\xi^{(k)}_{j,h,a}$.
$\widehat{P}_{g,z}$ (Methods)	Reward–observation kernel in the Methods backup formulas, written $\widehat{\mathcal{P}}_{g,z}$ in Appendix B; it is distinct from the appendix's state-transition operators with superscripts a or π_g.
$Q(s,a;\pi_g);\ \mathbf{a}_i(a), \widehat{\mathbf{a}}_i$	Methods notation for $Q_g^{\pi_g}$ and for its exact/estimated centered action evidence at observation occurrence s_i. Appendix B spells out A_g^{adv} and its sampled correction.
$\Lambda_A, \Lambda_J, \Lambda_P;\ c_i^t;\ \ell_i$	Methods advantage functionals, role weights $(c_i^A, c_i^J, c_i^P) = (1, w_i^J, w_i^P)$, and actionwise coefficients. The Methods role A corresponds to role 0 in the shared functional $\widehat{\Lambda}_g^{\chi}$.
$A(m), J(p), P(p);\ H$	Local behavior-advantage, target-advantage, and discrepancy-weighted transmission surfaces; heuristic guidance channel in Methods. The bare channel label H differs from the numeric heuristic horizon H in Appendix B.
m, p (Methods)	Candidate updated behavior/target probability vectors in the local surfaces. This local m is distinct from the synthetic-risk probability measure m in Appendix B.
g_A, g_H, g_J, g_P	Frozen-parent channel gradients. The target-return channel uses $g_J = \nabla(J/\tau)$; the A and P surfaces already include the factor $1/\tau$.
$F_\mu, F_\pi;\ \mathrm{Fuse}_F$	Actor-specific damped Fisher matrices; fusion map $F^{-1}(\alpha_1 g_1 + \alpha_2 g_2)$. These matrices are distinct from the physical transition map $F(s,a,d,u)$.
$n_i;\ \alpha_1, \alpha_2;\ t_\mu, t_\pi$	Channel natural norms $n_i = (g_i^\top F^{-1} g_i)^{1/2}$; fusion coefficients; accepted endpoint step fractions. The norms n_i are not sample counts, and the coefficients need not sum to one.
Evaluation and protocol-specific notation	
$C_s(\rho), C_s^{\mathrm{RB}};\ \bar{C}, S_C^2$	Evaluation mean cost at a risk and its risk-weighted aggregate for training seed s; mean and sample variance across training seeds. Here subscript s indexes seeds rather than observations.
X_0 cost	Evaluated cost of the initialized target policy in the result/protocol discussion; it is not the augmented physical state X_t or learner tuple $\mathcal{Y}_g$.
$C, NC;\ n_C, n_{NC};\ \Delta_C, \Delta_{NC}$	Retrospective paired-endpoint cost groups, their sizes, and mean per-seed percentage changes from native to HBPO. Group thresholds follow the corresponding result table.
Q_t	Q_t here denotes an order quantity, not an action-value function.
N, T (evaluation)	Number of evaluation tapes and periods per tape in the protocol. This deterministic sample count N is distinct from the geometric residual length in Appendix B.
$P_1, \ldots, P_4;\ q_r, q_e$ (protocol)	Protocol names for pipeline quantities and regular/expedited orders, written $p_1, \ldots, p_4$ and q^r, q^e in Appendix B.
$D;\ A;\ K_A(\rho)$ (protocol)	Protocol demand, availability, and true supply kernel, written D_t, B_t, and $K_{B,\rho}$ in Appendix B. In Methods model rollouts, A_k instead denotes an action, written A_k in the appendix.
$\mathcal{D}_{\mathrm{demand}};\ p;\ \omega_k$ (protocol)	Protocol true-demand support, stationary demand probabilities, and risk-evaluation weights; these correspond to $\mathcal{D}_{\mathrm{RB}}, \omega_D, \omega_{\rho,k}$, respectively.

B Algorithmic Details and Mathematical Derivations

B.1 Benchmark and Fitted Simulation Model

Scope and the DSIS special case. We first specify DSIS-RB and its fitted simulation model; the critic construction follows in Section B.2. DSIS is recovered as a special case of this framework

by making regular supply permanently available and replacing the demand process with its original public i.i.d. distribution. Its previous-demand coordinate and latent-risk variable can then be omitted. Thus, recovering DSIS requires specializing both supply and demand; setting the RB risk to zero alone does not recover its original demand process.

Throughout, ρ denotes the actual trajectory-specific risk in the environment, whereas z is a realization of the synthetic risk variable Z drawn by the learning algorithm. Neither risk is an input to the policy or value networks. Model-based computations use a declared simulation distribution for Z, independently of the unobserved risk of an observed trajectory. The nonnegative integer g indexes learning generations; model snapshot g includes the real batch collected for that generation. A synthetic tape is one simulated continuation with its own random draws.

B.1.1 Public supply-chain dynamics and rewards.

At period $t \in \{0, 1, \ldots\}$, let s_t denote the observed inventory configuration. Its net inventory is $I_t \in \mathbb{Z}$; its regular-order pipeline is $\mathbf{p}_t = (p_{t,1}, p_{t,2}, p_{t,3}, p_{t,4})$, where $p_{t,\ell} \in \{0, \ldots, 8\}$ is the outstanding quantity at stage $\ell \in \{1, 2, 3, 4\}$; and its current availability is $B_t \in \{0, 1\}$, with 0 meaning unavailable and 1 available. Negative I_t represents backlog. DSIS-RB additionally observes the previous demand d_{t-1}. The six- and seven-dimensional observations are

$$s_t^{\mathrm{DSIS}} = (I_t, \mathbf{p}_t, B_t), \qquad s_t^{\mathrm{RB}} = (I_t, \mathbf{p}_t, B_t, d_{t-1}). \tag{13}$$

Hereafter, s_t refers to the applicable observation above. We use S_t for the corresponding random observation and D_t for random demand, with realization d_t. Let $\mathcal{A}$ be the finite action set, and let q_t^r and q_t^e denote the regular and expedited order quantities:

$$a_t = (q_t^r, q_t^e) \in \mathcal{A} = \{0, \ldots, 8\}^2. \tag{14}$$

There are 81 actions. The implementation encodes (q^r, q^e) by the integer $9q^r + q^e \in \{0, \ldots, 80\}$; the equations below use the pair.

Given realized demand d_t and next-period availability B_{t+1}, the public inventory and pipeline equations are

$$\begin{aligned} I_{t+1} &= I_t + p_{t,1} + q_t^e - d_t, \\ \mathbf{p}_{t+1} &= (p_{t,2}, p_{t,3}, p_{t,4}, B_t q_t^r). \end{aligned} \tag{15}$$

Define the deterministic RB observation-transition map F by

$$s_{t+1} = F(s_t, a_t, d_t, B_{t+1}) = \big(I_{t+1}, p_{t,2}, p_{t,3}, p_{t,4}, B_t q_t^r, B_{t+1}, d_t\big). \tag{16}$$

DSIS uses the same map with the last coordinate omitted and $B_t = B_{t+1} = 1$. Current availability B_t determines whether the new regular order q_t^r is accepted. Previously accepted pipeline orders $p_{t,1}$ still arrive during an outage. Expedited orders q_t^e arrive immediately, while accepted regular orders $B_t q_t^r$ enter the four-period pipeline.

For a real number x, write $[x]_+ = \max\{x, 0\}$. Let $c(s, a, d)$ and $r(s, a, d)$ denote the period cost and normalized reward functions, respectively; their realized values are c_t and r_t. They satisfy

$$\begin{aligned} c_t &= c(s_t, a_t, d_t) = 10q_t^e + 5[I_{t+1}]_+ + 495[-I_{t+1}]_+, \\ r_t &= r(s_t, a_t, d_t) = -\frac{c_t}{495}, \end{aligned} \tag{17}$$

All values, Bellman residuals, and heuristic utilities below use these normalized reward units. The critic discount factor is $\gamma = 0.995$.

B.1.2 The true benchmark transition law.

For the RB benchmark, let $\mathcal{D}_{\mathrm{RB}}$ denote the demand support:

$$\mathcal{D}_{\mathrm{RB}} = \{0, 1, 2, 3, 8\}. \tag{18}$$

The stationary demand probability row vector ω_D is

$$\omega_D = (0.3, 0.2, 0.2, 0.2, 0.1) = \frac{1}{10}(3, 2, 2, 2, 1) \tag{19}$$

Its entries follow the displayed order in $\mathcal{D}_{\mathrm{RB}}$, and $\omega_D(d)$ denotes the entry associated with demand value d. Thus $\omega_D(d) = \Pr(\mathsf{D}_t = d)$ under stationary initialization, where Pr denotes probability. Let $\mathbf{1}\{E\}$ equal one when statement E holds and zero otherwise. For $x, d \in \mathcal{D}_{\mathrm{RB}}$, define the demand transition matrix K_D by

$$\Pr(\mathsf{D}_t = d \mid \mathsf{D}_{t-1} = x) = K_D(x, d) = 0.75\,\mathbf{1}\{d = x\} + 0.25\,\omega_D(d). \tag{20}$$

Thus, with probability 0.75 demand retains its previous value; otherwise it is independently redrawn from ω_D. This construction satisfies $\omega_D K_D = \omega_D$ and $\sum_{d \in \mathcal{D}_{\mathrm{RB}}} d\,\omega_D(d) = 2$.

An environment trajectory receives a risk $\rho \in [0, 0.5]$, which remains fixed throughout that trajectory. Conditional on ρ, let $K_{B,\rho}(b_{\mathrm{cur}}, u) = \Pr(B_{t+1} = u \mid B_t = b_{\mathrm{cur}}, \rho)$ for $b_{\mathrm{cur}}, u \in \{0, 1\}$. In row and column order $(0, 1)$, this supply transition matrix is

$$K_{B,\rho} = \begin{pmatrix} \frac{3}{4} & \frac{1}{4} \\ \frac{\rho}{4(1-\rho)} & 1 - \frac{\rho}{4(1-\rho)} \end{pmatrix}. \tag{21}$$

Its stationary probability row vector is $(\rho, 1-\rho)$: the flows $\rho/4$ from 0 to 1 and $(1-\rho)\rho/[4(1-\rho)]$ from 1 to 0 agree. Demand and supply innovations are independent conditional on the current observation and ρ. Below we suppress time indices: $s = s_t$, $a = a_t$, $s' = s_{t+1}$, and $u = B_{t+1}$. Write $B(s)$ for the availability coordinate of s and $d^-(s)$ for its previous-demand coordinate. For any point x, δ_x denotes the probability measure with unit mass at x; dr and ds' indicate the reward and next-observation coordinates of a measure. The joint reward–observation transition kernel $\mathcal{P}_\rho$ is

$$\mathcal{P}_\rho(dr, ds' \mid s, a) = \sum_{d \in \mathcal{D}_{\mathrm{RB}}} \sum_{u \in \{0,1\}} K_D(d^-(s), d) K_{B,\rho}(B(s), u)\, \delta_{(r(s,a,d), F(s,a,d,u))}(dr, ds'). \tag{22}$$

At each episode initialization, inventory and pipeline quantities are zero. Writing $\mathrm{Bernoulli}(p)$ for a binary distribution with probability p of one, the initial demand and availability obey

$$\mathsf{D}_{-1} \sim \omega_D, \qquad B_0 \mid \rho \sim \mathrm{Bernoulli}(1-\rho), \tag{23}$$

with independent demand and supply initialization conditional on ρ.

Risk stratification and its statistical target. The registered risk grid is $\rho_j = 0.1j$, $j = 0, \ldots, 5$. An episode contains $H_{\mathrm{env}} = 1000$ physical periods. Let π be a fixed policy assigning a probability $\pi(a \mid s)$ to each $a \in \mathcal{A}$, and let $\mathbb{E}_{\pi,\rho_j}$ denote expectation over its actions and the benchmark disturbances at risk ρ_j, with the initialization above. Its expected episode mean cost is

$$J_\pi(\rho_j) = \mathbb{E}_{\pi,\rho_j}\left[\frac{1}{H_{\mathrm{env}}} \sum_{t=0}^{H_{\mathrm{env}}-1} c_t\right]. \tag{24}$$

Define $\mathcal{R}_{\mathrm{trap}}(\pi)$ as the normalized trapezoidal risk score and $\omega_{\rho,j}$ as its weight at grid point j; ω_ρ collects these six weights in grid order:

$$\begin{aligned} \mathcal{R}_{\mathrm{trap}}(\pi) &= \frac{1}{0.5} \sum_{j=0}^{4} \frac{0.1}{2} \big[J_\pi(\rho_j) + J_\pi(\rho_{j+1})\big] \\ &= \sum_{j=0}^{5} \omega_{\rho,j} J_\pi(\rho_j), \qquad \omega_\rho = \frac{1}{10}(1, 2, 2, 2, 2, 1). \end{aligned} \tag{25}$$

The endpoints receive half the weight of an interior grid point because each endpoint belongs to one trapezoid, whereas each interior point belongs to two. These risk weights are distinct from the demand probabilities ω_D.

The probability measure ν_{env} describing the target risk composition is

$$\nu_{\mathrm{env}} = \sum_{j=0}^{5} \omega_{\rho,j} \delta_{\rho_j}. \tag{26}$$

Writing $\rho \sim \nu_{\mathrm{env}}$ specifies this target composition, not independent sampling of every training trajectory. Training implements proportional stratification in blocks of 25 trajectories. The desired

counts are $25\omega_\rho = (2.5, 5, 5, 5, 5, 2.5)$. Let $\mathbf{n}^{(0)}$ and $\mathbf{n}^{(1)}$ denote the alternating integer count vectors, with coordinates ordered by the risk grid:

$$\begin{aligned} \mathbf{n}^{(0)} &= (3, 5, 5, 5, 5, 2), \\ \mathbf{n}^{(1)} &= (2, 5, 5, 5, 5, 3), \\ \mathbf{n}^{(0)} + \mathbf{n}^{(1)} &= (5, 10, 10, 10, 10, 5) = 50\omega_\rho. \end{aligned} \tag{27}$$

Each block is privately permuted before trajectories are supplied. Alternating the two integer allocations avoids systematically rounding one endpoint upward: every complete pair of blocks has exactly the registered risk proportions, without random fluctuation in its stratum counts.

For a policy held fixed across such a pair, let Y_ℓ be the realized mean cost of trajectory $\ell \in \{1, \ldots, 50\}$, and let $\mathbb{E}$ denote expectation over benchmark and policy randomness. Then

$$\mathbb{E}\left[\frac{1}{50}\sum_{\ell=1}^{50} Y_\ell\right] = \sum_{j=0}^{5} \omega_{\rho,j} J_\pi(\rho_j) = \mathcal{R}_{\mathrm{trap}}(\pi). \tag{28}$$

This is unbiasedness for the registered discrete trapezoidal target. For the actual changing training policies, the applicable guarantee is exact two-block risk balance; it does not assert unbiasedness of a training update for a single fixed-policy objective. Likewise, the trapezoidal score is a quadrature approximation to a continuous-risk integral, not an exactly unbiased estimator of that integral. Formal evaluation holds the policy fixed, evaluates each stratum separately, and combines its sample mean with ω_ρ.

Within an episode, the augmented variable $X_t = (S_t, \rho)$ is a controlled Markov state for the physical dynamics. The observation S_t alone need not be Markov after a persistent hidden risk is marginalized. The displayed physical kernel excludes the scheduled reset after H_{env} periods. The preceding probabilities specify the benchmark for reproducibility; the RB learner does not receive its true stochastic kernels, trajectory risks, risk assignments, or private random tapes.

For DSIS, the same physical and reward equations apply on the reachable set $B_t = B_{t+1} = 1$. Here Unif denotes the uniform distribution, and demand is independent and identically distributed (i.i.d.):

$$\mathsf{D}_t \overset{\mathrm{iid}}{\sim} \mathrm{Unif}\{0, 1, 2, 3, 4\}. \tag{29}$$

Its observation is the six-dimensional vector

$$s_t = (I_t, \mathbf{p}_t, B_t). \tag{30}$$

B.1.3 Constructing the fitted simulation model.

For DSIS-RB, the generation-g model is fitted only to counted real observations. We define its components before summarizing all structural and estimation assumptions in Table 3.

Observed-data demand model. Let $\mathscr{D}^{\mathrm{real}}_{\le g}$ be the multiset of counted, observed transitions available after collecting the generation-g batch, including any counted initialization prefix. Every observed occurrence is counted once, including repeated transitions. Set $N^{\mathrm{real}}_g = |\mathscr{D}^{\mathrm{real}}_{\le g}| > 0$; no fitted RB model is formed without real data. Recall that $d^-(s)$ selects the previous-demand coordinate. Let $\widehat{\mathcal{D}}_g$ be the union of the values of $d^-(s)$ and $d^-(s')$ in this multiset. For $x, y \in \widehat{\mathcal{D}}_g$, define the pair count $\mathcal{C}_g(x, y)$, its row total $\mathcal{C}_g(x)$, and the total count as follows:

$$\mathcal{C}_g(x, y) = \sum_{(s,a,r,s') \in \mathscr{D}^{\mathrm{real}}_{\le g}} \mathbf{1}\{d^-(s) = x,\ d^-(s') = y\}, \tag{31}$$

The row and total counts satisfy

$$\mathcal{C}_g(x) = \sum_{y \in \widehat{\mathcal{D}}_g} \mathcal{C}_g(x, y), \qquad N^{\mathrm{real}}_g = \sum_{x,y \in \widehat{\mathcal{D}}_g} \mathcal{C}_g(x, y). \tag{32}$$

The pooled empirical next-demand probability mass function $\widehat{\omega}_g$ is

$$\widehat{\omega}_g(y) = \frac{\sum_{x \in \widehat{\mathcal{D}}_g} \mathcal{C}_g(x, y)}{N^{\mathrm{real}}_g}. \tag{33}$$

With pseudocount mass $\alpha_D = 1$, define the fitted demand transition matrix $\widehat{K}_{D,g}$ on $\widehat{\mathcal{D}}_g$ by

$$\widehat{K}_{D,g}(x,y) = \frac{\mathcal{C}_g(x,y) + \alpha_D \widehat{\omega}_g(y)}{\mathcal{C}_g(x) + \alpha_D}, \qquad \alpha_D = 1. \tag{34}$$

Each conditional row is smoothed toward the pooled empirical distribution using one total pseudocount. Its entries are nonnegative and sum to one. If $\mathcal{C}_g(x) = 0$, the row equals $\widehat{\omega}_g$; this includes a value observed only as a successor. Queries require $d^-(s) \in \widehat{\mathcal{D}}_g$; the implementation rejects an unseen conditioning value rather than inventing a row.

Supply model and persistent synthetic risk. Recall that $B(s)$ is the availability coordinate. Let $\mathcal{C}_g^{\downarrow}$ count observed transitions starting in an outage and $\mathcal{C}_g^{\mathrm{rec}}$ count those that recover:

$$\begin{aligned} \mathcal{C}_g^{\downarrow} &= \sum_{(s,a,r,s') \in \mathscr{D}^{\mathrm{real}}_{\le g}} \mathbf{1}\{B(s) = 0\}, \\ \mathcal{C}_g^{\mathrm{rec}} &= \sum_{(s,a,r,s') \in \mathscr{D}^{\mathrm{real}}_{\le g}} \mathbf{1}\{B(s) = 0, B(s') = 1\}. \end{aligned} \tag{35}$$

Let $\kappa^{\mathrm{rec}} \in (0,1)$ be the common recovery-probability parameter in the fitted supply family, and let $\widehat{\kappa}_g^{\mathrm{rec}}$ be its estimate:

$$\widehat{\kappa}_g^{\mathrm{rec}} = \frac{\mathcal{C}_g^{\mathrm{rec}} + 1}{\mathcal{C}_g^{\downarrow} + 2}. \tag{36}$$

Under a $\mathrm{Beta}(1,1)$ prior (the uniform prior on $(0,1)$) for κ^{rec}, this is the posterior mean based on the outage-origin transitions. It remains defined when no outage has been observed, in which case $\widehat{\kappa}_g^{\mathrm{rec}} = 1/2$. Let m denote the synthetic-risk distribution. At the root of each new synthetic tape, the algorithm draws Z independently of the queried observation and of the other tapes:

$$Z \sim m = \mathrm{Unif}[0, 0.5]. \tag{37}$$

The tape retains this draw throughout its evolution. Conditional on $Z = z$, let $\widehat{K}_{B,g,z}$ be the modeled supply transition matrix, with rows and columns ordered as $(0,1)$:

$$\widehat{K}_{B,g,z} = \begin{pmatrix} 1 - \widehat{\kappa}_g^{\mathrm{rec}} & \widehat{\kappa}_g^{\mathrm{rec}} \\ \widehat{\kappa}_g^{\mathrm{rec}} \frac{z}{1-z} & 1 - \widehat{\kappa}_g^{\mathrm{rec}} \frac{z}{1-z} \end{pmatrix}. \tag{38}$$

Because $0 < \widehat{\kappa}_g^{\mathrm{rec}} < 1$ and $0 \le z/(1-z) \le 1$, all matrix entries lie in $[0,1]$ and each row sums to one. The identity

$$(1-z)\widehat{K}_{B,g,z}(1,0) = z\widehat{K}_{B,g,z}(0,1) \tag{39}$$

shows that $(z, 1-z)$ is a stationary probability row vector, so the stationary outage probability is z, including the boundary $z = 0$. This parameterization is a modeling assumption; knowing only the risk interval would not imply it.

The combined model and generation-level freezing. Recall that r and F are the public reward and physical-transition maps in Equations (17) and (16). For a query with $d^-(s) \in \widehat{\mathcal{D}}_g$ and $z \in [0, 0.5]$, define the conditional reward–observation kernel $\widehat{\mathcal{P}}_{g,z}$ by

$$\begin{aligned} &\widehat{\mathcal{P}}_{g,z}(dr, ds' \mid s, a) \\ &\quad = \sum_{d \in \widehat{\mathcal{D}}_g} \sum_{u \in \{0,1\}} \widehat{K}_{D,g}(d^-(s), d) \widehat{K}_{B,g,z}(B(s), u)\, \delta_{(r(s,a,d), F(s,a,d,u))}(dr, ds'). \end{aligned} \tag{40}$$

Here δ_x again denotes unit probability mass at x. We write $\widehat{\mathcal{P}}_g$ for the full simulation specification: a query observation, an independent initial risk draw from m, and the conditional dynamics above with that risk retained. To state its Markov transition explicitly, let z' be the successor risk, let dz' indicate its measure coordinate, and define the augmented kernel $\widetilde{\mathcal{P}}_g$ on observation–risk pairs by

$$\widetilde{\mathcal{P}}_g(dr, ds', dz' \mid s, z, a) = \widehat{\mathcal{P}}_{g,z}(dr, ds' \mid s, a) \delta_z(dz'). \tag{41}$$

The model snapshot is frozen before generation-g model-based computations begin. For DSIS, $\widehat{\mathcal{P}}_{g,z}$ reduces to the public nominal kernel independently of g and z; no stochastic-law estimation or latent-risk simulation is required.

Summary of assumptions. Table 3 lists all structural and estimation assumptions; HBPO uses every applicable row. The last column reports explicit baseline use, separately from access to public information. In the table, $\widehat{p}_g$ denotes the conditional joint probability mass function of modeled demand d and next availability u, with its one-variable versions denoting the corresponding marginals. For a synthetic tape, $k \in \{0, 1, \ldots\}$ indexes simulated periods, S_0 is its query observation, and Z_k is its retained risk at period k; $\perp$ means independence. In rows marked Both, DSIS omits the unused previous-demand and risk arguments: its demand is uniform on $\{0, \ldots, 4\}$ and its next availability equals one.

Table 3: Declared structural and estimation priors. Both means DSIS and DSIS-RB. Baseline use refers to explicit use in the learner; all methods interact with the same physical benchmark.

Formula	Description	Physical or modeling role	DSIS / DSIS-RB	Baseline use
Physical rules and stochastic model structure				
$a_t = (q_t^r, q_t^e)$, $q_t^r, q_t^e \in \{0, \ldots, 8\}$	Finite action domain.	Two ordering channels, each accepting quantities from 0 to 8.	Both	All: action interface; GPS: model branches.
$I_{t+1} = I_t + p_{t,1} + q_t^e - d_t$	Deterministic inventory update.	Inventory conservation, scheduled arrivals, and immediate, reliable expediting.	Both	GPS: explicit state model.
$p_{t+1,\ell} = p_{t,\ell+1}, \quad \ell = 1, 2, 3,$ $p_{t+1,4} = B_t q_t^r$	Pipeline shift and order gate.	Four-period regular lead time; outages reject new regular orders without cancelling existing orders.	Both	GPS: explicit state model.
$c_t = 10q_t^e + 5[I_{t+1}]_+ + 495[-I_{t+1}]_+,$ $r_t = -c_t/495$	Known cost and reward map.	Expediting, holding, and backlog costs; a common reward scale.	Both	GPS: model rewards; all: observed rewards.
$\mathsf{D}_t \overset{\text{iid}}{\sim} \text{Unif}\{0, \ldots, 4\}$, $B_t = B_{t+1} = 1$	Public nominal stochastic law.	Memoryless demand and uninterrupted regular supply.	DSIS	GPS: nominal model.
$\widehat{p}_g(d \mid s_t, a_t, z)$ $= \widehat{K}_{D,g}(d^-(s_t), d)$, $\widehat{p}_g(u \mid s_t, a_t, z)$ $= \widehat{K}_{B,g,z}(B_t, u)$	First-order, exogenous disturbance models.	Previous demand and current availability carry temporal dependence; orders do not control these laws.	DSIS-RB	GPS: same conditional model structure.
$\widehat{p}_g(d, u \mid s_t, a_t, z)$ $= \widehat{K}_{D,g}(d^-(s_t), d)$ $\cdot \widehat{K}_{B,g,z}(B_t, u)$	Conditional factorization.	Separate demand and supply innovations; DSIS has i.i.d. demand and degenerate supply.	Both	GPS: factorized disturbance model.
$\widehat{K}_{B,g,z}(0, 1) = \widehat{\kappa}_g^{\text{rec}}$, $\widehat{K}_{B,g,z}(1, 0) = \widehat{\kappa}_g^{\text{rec}} \dfrac{z}{1 - z}$	Common recovery and stationary-risk family.	Binary interruptions with modeled stationary outage share z.	DSIS-RB	GPS: same supply family.

Continued on next page

Table 3 (continued)

Formula	Description	Physical or modeling role	DSIS / DSIS-RB	Baseline use
$Z_0 \sim \mathrm{Unif}[0, 0.5], \quad Z_0 \perp S_0,$ $Z_k = Z_0 \quad \text{for all } k$	Independent query prior; persistent tape risk.	Models sustained risk heterogeneity on the public interval; not the actual risk-assignment law.	DSIS-RB	GPS: same synthetic-risk prior.
Estimation priors (regularization, not benchmark laws)				
$\widehat{K}_{D,g}(x, y)$ $= \dfrac{\mathcal{C}_g(x, y) + \widehat{\omega}_g(y)}{\mathcal{C}_g(x) + 1}$	One pooled pseudocount per row.	Regularizes demand transitions with sparse observed counts.	DSIS-RB	GPS: same demand estimator.
$\kappa^{\mathrm{rec}} \sim \mathrm{Beta}(1, 1),$ $\widehat{\kappa}_g^{\mathrm{rec}} = \dfrac{\mathcal{C}_g^{\mathrm{rec}} + 1}{\mathcal{C}_g^{\downarrow} + 2}$	Smoothed recovery estimate.	Regularizes recovery counts from observed outages.	DSIS-RB	GPS: same recovery estimator.

Reading the baseline column. All refers to PPO, A2C, Discrete SAC, ACE, PPO-81, TRPO, GRPO, TPO, and GPS. Every method has access to the same public inventory rules and cost function, and to the complete nominal stochastic law in DSIS. The declared model-class assumptions are available to all comparison methods. An entry naming GPS identifies explicit use in its predictive model, rather than exclusive access to the corresponding information. The other baselines use their native learning rules on observed transitions; receiving a reward does not itself constitute evaluating the analytical reward map on a hypothetical transition. GPS and HBPO fit unknown RB probabilities to their respective counted data, so sharing an estimator does not imply sharing fitted parameters.

Information boundary. The functional forms in the table are specified in advance, rather than learned solely from data. In particular, the supply parameterization $\widehat{\kappa}_g^{\mathrm{rec}} z/(1 - z)$ does not follow from knowing the risk interval $[0, 0.5]$ alone: it also assumes a common recovery probability and a stationary outage share z. These assumptions do not give the learner the true RB recovery probability 0.25, the true demand transition matrix, the actual trajectory risk ρ, or future random tapes. The RB demand support is discovered from observations, and unknown probabilities are estimated only from the learner's own counted observations. The synthetic distribution of Z is neither the private benchmark risk-assignment law nor a posterior conditioned on the queried state. The DSIS fallback separately grants every method its complete public nominal law.

B.1.4 Use of structural priors and the information-fair comparison.

The priors in Table 3 enter HBPO through initialization, including fitted Q evaluation (FQE) on counterfactuals using observed disturbances; supply-conditioned replay; action-value backups and residual continuations; successor-critic state generation and fitting; and heuristic utility, occupancy, calibration, and confidence/interest calculations. These computations use simulated transitions and conditional expectations for learning, with enumeration over finite action and disturbance branches rather than the inventory state space. They provide no additional real-environment observations.

All methods share public information and a $2{,}000{,}000$ real-interaction budget per formal learning run, with synthetic computation recorded separately. GPS uses the same model class, estimation rules, and tape-fixed uniform synthetic-risk prior as HBPO, fitted to its own observations. Neither learner receives unrestricted environment resets. Fairness therefore concerns information access and real interactions, not equal computation or identical fitted models.

The structural prior is part of HBPO. Comparison with model-informed GPS assesses the overall learning architecture beyond shared model access; it does not isolate an individual transmission component. Mechanism analyses and ablations examine guidance and transmission under common modeling and estimation rules, while their collected data and fitted probabilities may differ. Using

models and inventory structure to support policy learning has established precedents (Sutton, 1991; Levine & Koltun, 2013; Böttcher et al., 2023).

B.2 Critic Construction and Model-Based Residual Completion

B.2.1 The five scalar critics and the frozen value reference.

At generation g, HBPO maintains five separately parameterized scalar networks. For member index $j \in \{1, \dots, 5\}$, let $\theta_{g,j}$ denote its trainable parameters and write

$$V_{g,j}(s) = V_{\theta_{g,j}}(s). \tag{42}$$

The frozen ensemble reference is the mean of their scalar outputs,

$$v_g(s) = \frac{1}{5} \sum_{j=1}^{5} V_{g,j}(s). \tag{43}$$

The networks have no shared trainable layers.

Let d_ϕ denote the feature dimension: $d_\phi = 7$ in RB and $d_\phi = 6$ in DSIS. Recall that I is inventory, $p_1, \dots, p_4$ are pipeline coordinates, B is current availability, and d^- is previous demand. The fixed RB feature map, written as a column vector, is

$$\phi(s) = \left(\operatorname{asinh}(I/4), p_1/8, p_2/8, p_3/8, p_4/8, B, d^-/8\right)^\top \in \mathbb{R}^7, \tag{44}$$

where asinh is the inverse hyperbolic sine. DSIS omits the last coordinate.

Each network has two Tanh hidden layers of width 128 and a scalar linear output. For a generic parameter vector $\theta = (W_1, b_1, W_2, b_2, w_3, b_3)$, the dimensions are $W_1 \in \mathbb{R}^{128 \times d_\phi}$, $W_2 \in \mathbb{R}^{128 \times 128}$, $b_1, b_2, w_3 \in \mathbb{R}^{128}$, and $b_3 \in \mathbb{R}$; $\tanh$ acts coordinatewise. The resulting scalar function is

$$V_\theta(s) = w_3^\top \tanh(W_2 \tanh(W_1 \phi(s) + b_1) + b_2) + b_3. \tag{45}$$

During generation g, v_g is a frozen continuation-value reference. It need not equal the exact value under either the current model or the environment. For $g \geq 1$, its parameters are inherited from the preceding generation even when the newly fitted model has changed; v_0 is supplied by the initialization procedure described below.

B.2.2 One-step action backups and the target-policy residual.

Recall that $\widehat{\mathcal{P}}_{g,z}$ is the frozen model conditional on synthetic risk z, $\widehat{\mathcal{D}}_g$ is its finite demand support, and $\widehat{K}_{D,g}$ and $\widehat{K}_{B,g,z}$ are its demand and availability kernels. Also recall the deterministic successor map F, normalized reward r, and discount $\gamma = 0.995$. Here $d^-(s)$ and $B(s)$ denote the previous-demand and current-availability coordinates of s. For a real-valued function V with finite backups and a prescribed next availability $u \in \{0, 1\}$, define the supply-conditioned action backup by integrating demand:

$$L_g^u V(s, a) = \sum_{d \in \widehat{\mathcal{D}}_g} \widehat{K}_{D,g}(d^-(s), d) \left[r(s, a, d) + \gamma V(F(s, a, d, u)) \right]. \tag{46}$$

Integrating next availability conditional on the fixed tape risk z gives

$$L_{g,z} V(s, a) = \sum_{u \in \{0,1\}} \widehat{K}_{B,g,z}(B(s), u) L_g^u V(s, a). \tag{47}$$

Let R and S' denote the reward and successor from one transition of this model. Then

$$L_{g,z} V(s, a) = \mathbb{E}_{\widehat{\mathcal{P}}_{g,z}}[R + \gamma V(S') \mid s, a]. \tag{48}$$

The frozen parent target policy π_g induces the discounted Bellman operator $\mathcal{T}_{g,z}^{\pi_g}$,

$$\mathcal{T}_{g,z}^{\pi_g} V(s) = \sum_{a \in \mathcal{A}} \pi_g(a \mid s) L_{g,z} V(s, a). \tag{49}$$

Its residual at the frozen reference v_g is

$$\begin{aligned} e_{g,z}(s) &= (\mathcal{T}^{\pi_g}_{g,z} v_g)(s) - v_g(s) \\ &= \sum_{a' \in \mathcal{A}} \pi_g(a' \mid s) \sum_{d \in \widehat{\mathcal{D}}_g} \sum_{u \in \{0,1\}} \widehat{K}_{D,g}(d^-(s), d) \widehat{K}_{B,g,z}(B(s), u) \\ &\qquad \cdot \big[r(s, a', d) + \gamma v_g(F(s, a', d, u)) \big] - v_g(s). \end{aligned} \tag{50}$$

This residual integrates the target action, demand, and next availability conditional on z. It is the conditional mean of a temporal-difference error, rather than a sampled reward increment. Its sign indicates whether the model's one-step evaluation exceeds or falls below the reference value; it does not indicate that the policy itself has improved. Theorem 1 shows why propagating this signed inconsistency recovers the remaining continuation-value error.

Recall the synthetic-risk prior $m = \mathrm{Unif}[0, 1/2]$ in RB. The risk-averaged one-step base is

$$Q^0_g(s, a) = \mathbb{E}_{Z \sim m}[L_{g,Z} v_g(s, a)]. \tag{51}$$

Let $\widehat{\kappa}^{\mathrm{rec}}_g$ denote the fitted recovery probability, as in the supply model above. Since v_g has no risk input, the RB root backup uses the analytic mean supply kernel

$$\mathbb{E}_{Z \sim m}[\widehat{K}_{B,g,Z}(B, 0)] = \begin{cases} 1 - \widehat{\kappa}^{\mathrm{rec}}_g, & B = 0, \\ \widehat{\kappa}^{\mathrm{rec}}_g (2 \log 2 - 1), & B = 1, \end{cases} \tag{52}$$

where $\log$ is the natural logarithm and $\mathbb{E}_{Z \sim m}[Z/(1 - Z)] = 2 \int_0^{1/2} z/(1 - z)\, dz = 2 \log 2 - 1$. Only the one-step root backup is averaged in this way. Multistep simulation retains the sampled z; replacing its kernel at every step by the prior mean would in general define a different process.

B.2.3 Residual completion of the discounted action value.

Recall that m is the synthetic-risk law and that $\widehat{\mathcal{P}}_{g,z}$ is the frozen reward–observation kernel at generation g (Equation (40)). Let S_k, A_k, and R_k denote the observation, action, and normalized reward at simulated step k. Define the model *action-value function* queried by HBPO as

$$Q^{\pi_g}_g(s, a) = \mathbb{E}_{Z \sim m} \mathbb{E}_{\widehat{\mathcal{P}}_{g,Z}, \pi_g} \left[\sum_{k=0}^{\infty} \gamma^k R_k \,\middle|\, S_0 = s, \mathsf{A}_0 = a \right], \tag{53}$$

where the first action is fixed to a, and $\mathsf{A}_k \sim \pi_g(\cdot \mid S_k)$ for $k \geq 1$. The risk is sampled independently at the query root and remains fixed thereafter. This defines a model action-value function under the declared risk prior; z is a simulated scenario, not an estimate of the queried trajectory's true risk ρ.

For the following derivation, fix generation g and write $\mathbb{E}^{s,a}_z$ for expectation under the rollout law with fixed risk z, initial observation s, and forced first action a. Write $\mathbb{E}^s_z$ when the initial action also follows π_g. Define the exact conditional model value $V^{\pi_g}_{g,z}$ and its error $\Delta_{g,z}$ relative to v_g by

$$V^{\pi_g}_{g,z}(s) = \mathbb{E}^s_z \left[\sum_{k=0}^{\infty} \gamma^k R_k \right], \qquad \Delta_{g,z}(s) = V^{\pi_g}_{g,z}(s) - v_g(s). \tag{54}$$

Here $V^{\pi_g}_{g,z}$ is the exact conditional *model* value, whereas v_g is the frozen network reference. For a scalar function f, let the model's one-step state-transition operators be $\widehat{P}^a_{g,z}$ (forced first action a) and $\widehat{P}^{\pi_g}_{g,z}$ (first action sampled from π_g):

$$(\widehat{P}^a_{g,z} f)(s) = \mathbb{E}^{s,a}_z[f(S_1)], \qquad (\widehat{P}^{\pi_g}_{g,z} f)(s) = \sum_{a' \in \mathcal{A}} \pi_g(a' \mid s) (\widehat{P}^{a'}_{g,z} f)(s). \tag{55}$$

Theorem 1 (Residual completion and the omitted continuation error)**.** *Let $\mathcal{S}$ be the countable observation space of the frozen inventory model, closed under its transitions. Fix the model, the observation-based policy π_g, a discount $0 < \gamma < 1$, and a reference v_g satisfying $|v_g(s)| \leq C_v < \infty$ for all $s \in \mathcal{S}$, where C_v is a constant. Assume that, for every $s \in \mathcal{S}$ and $a \in \mathcal{A}$,*

$$\int \mathbb{E}^{s,a}_z \left[\sum_{k=0}^{\infty} \gamma^k |R_k| \right] m(dz) < \infty. \tag{56}$$

The action set is finite, so this also implies the corresponding condition when the first action follows π_g. Recall that $\mathcal{T}_{g,z}^{\pi_g}$ is the Bellman operator in Equation (49)*, and that $e_{g,z} = \mathcal{T}_{g,z}^{\pi_g} v_g - v_g$. Write* Id *for the identity operator on scalar functions. For an integer $\ell \geq 0$, $(\widehat{P}_{g,z}^{\pi_g})^\ell$ denotes ℓ-fold operator composition, with zeroth power* Id*. Then the Bellman residual and conditional value error satisfy*

$$e_{g,z} = (\mathrm{Id} - \gamma \widehat{P}_{g,z}^{\pi_g}) \Delta_{g,z}, \qquad \Delta_{g,z} = \sum_{\ell=0}^{\infty} \gamma^\ell (\widehat{P}_{g,z}^{\pi_g})^\ell e_{g,z}. \tag{57}$$

These identities hold at every $s \in \mathcal{S}$ on a common set of risks of full m-measure. The series converges absolutely pointwise; no bounded-value-space inverse is assumed. In particular, the exact error of the one-step base is

$$Q_g^{\pi_g}(s,a) - Q_g^0(s,a) = \gamma \mathbb{E}_{Z \sim m}[(\widehat{P}_{g,Z}^a \Delta_{g,Z})(s)]. \tag{58}$$

It is recovered by the discounted residual sum:

$$Q_g^{\pi_g}(s,a) = Q_g^0(s,a) + \int \mathbb{E}_z^{s,a} \left[\sum_{k=1}^{\infty} \gamma^k e_{g,z}(S_k) \right] m(dz). \tag{59}$$

Proof. Because $\mathcal{S}$ is countable and $\mathcal{A}$ is finite, Equation (56) provides a common set of risks of full m-measure on which absolute discounted reward integrability holds for every initial state and action. Fix a risk z in this set. Conditional on z, the observation process is Markov. Bellman consistency of the exact conditional value gives

$$\Delta_{g,z} = e_{g,z} + \gamma \widehat{P}_{g,z}^{\pi_g} \Delta_{g,z}. \tag{60}$$

For any integer $n_{\mathrm{tr}} \geq 0$, iterating this relation yields

$$\Delta_{g,z}(s) = \sum_{\ell=0}^{n_{\mathrm{tr}}} \gamma^\ell ((\widehat{P}_{g,z}^{\pi_g})^\ell e_{g,z})(s) + \gamma^{n_{\mathrm{tr}}+1} \mathbb{E}_z^s[\Delta_{g,z}(S_{n_{\mathrm{tr}}+1})]. \tag{61}$$

The absolute terminal term is bounded by

$$\gamma^{n_{\mathrm{tr}}+1} \mathbb{E}_z^s[|\Delta_{g,z}(S_{n_{\mathrm{tr}}+1})|] \leq \mathbb{E}_z^s \left[\sum_{k=n_{\mathrm{tr}}+1}^{\infty} \gamma^k |R_k| \right] + \gamma^{n_{\mathrm{tr}}+1} C_v \longrightarrow 0. \tag{62}$$

Here the reward-tail bound follows from the Markov property and conditional Jensen's inequality. Absolute integrability makes its expectation tend to zero. Moreover, conditional Jensen's inequality applied to the one-step residual gives

$$\sum_{\ell=0}^{\infty} \gamma^\ell \mathbb{E}_z^s[|e_{g,z}(S_\ell)|] \leq \mathbb{E}_z^s \left[\sum_{\ell=0}^{\infty} \gamma^\ell |R_\ell| \right] + \frac{(1+\gamma) C_v}{1-\gamma} < \infty. \tag{63}$$

This proves both identities in Equation (57), including absolute convergence of its series. The same bounds can be integrated against $m(dz)$ by the stated mixture integrability assumption.

Conditioning the discounted return on its first transition yields

$$Q_g^{\pi_g}(s,a) = \int \left\{ L_{g,z} v_g(s,a) + \gamma (\widehat{P}_{g,z}^a \Delta_{g,z})(s) \right\} m(dz). \tag{64}$$

Subtracting $Q_g^0(s,a) = \int L_{g,z} v_g(s,a) m(dz)$ proves Equation (58).

For the residual-completion identity, at every continuation step $k \geq 1$ the tower property gives

$$\mathbb{E}_z^{s,a}[e_{g,z}(S_k)] = \mathbb{E}_z^{s,a}[R_k + \gamma v_g(S_{k+1}) - v_g(S_k)]. \tag{65}$$

The restriction $k \geq 1$ ensures that the action follows π_g; the forced root action is handled by $L_{g,z} v_g(s,a)$. Multiplying by γ^k and summing gives, for each integer $n_{\mathrm{tr}} \geq 0$,

$$L_{g,z} v_g(s,a) + \mathbb{E}_z^{s,a} \left[\sum_{k=1}^{n_{\mathrm{tr}}} \gamma^k e_{g,z}(S_k) \right] = \mathbb{E}_z^{s,a} \left[\sum_{k=0}^{n_{\mathrm{tr}}} \gamma^k R_k + \gamma^{n_{\mathrm{tr}}+1} v_g(S_{n_{\mathrm{tr}}+1}) \right]. \tag{66}$$

The sum is empty when $n_{\mathrm{tr}} = 0$. After integration over $m(dz)$, the remaining reference-value term satisfies

$$\gamma^{n_{\mathrm{tr}}+1} \int \mathbb{E}_z^{s,a}[|v_g(S_{n_{\mathrm{tr}}+1})|]m(dz) \leq \gamma^{n_{\mathrm{tr}}+1} C_v \longrightarrow 0. \tag{67}$$

For the forced-action rollout, the analogue of Equation (63) starts at $k = 1$; its right-hand side remains integrable over the root risk. Consequently the infinite residual sum and its expectation can be interchanged, and taking $n_{\mathrm{tr}} \to \infty$ proves Equation (59). □

The theorem identifies the correction exactly: it is the part of future policy value missing from the one-step bootstrap. The factor γ^k is imposed by the original discounted return and the telescoping cancellation; it is not a separately chosen weight for risk or optimism. At a finite deterministic continuation length n_{tr}, the remaining error is exactly

$$Q_g^{\pi_g}(s,a) - \left\{ Q_g^0(s,a) + \int \mathbb{E}_z^{s,a}\left[\sum_{k=1}^{n_{\mathrm{tr}}} \gamma^k e_{g,z}(S_k)\right] m(dz) \right\} = \gamma^{n_{\mathrm{tr}}+1} \int \mathbb{E}_z^{s,a}[\Delta_{g,z}(S_{n_{\mathrm{tr}}+1})]m(dz). \tag{68}$$

Thus a fixed truncation retains a terminal evaluation error. The geometrically stopped estimator below implements the complete series with an almost surely finite random continuation.

Persistent risk and an observation-only critic. RB introduces a further reason for conditional evaluation error: $v_g(s)$ has no z input, whereas $V_{g,z}^{\pi_g}(s)$ can depend on the persistent risk. At any fixed state s, assume $\mathbb{E}_{Z\sim m}[(V_{g,Z}^{\pi_g}(s))^2] < \infty$. Writing $c_{\mathrm{val}} \in \mathbb{R}$ for a scalar prediction and $\mathrm{Var}_{Z\sim m}$ for variance over the root-risk law,

$$\inf_{c_{\mathrm{val}}\in\mathbb{R}} \mathbb{E}_{Z\sim m}[(V_{g,Z}^{\pi_g}(s) - c_{\mathrm{val}})^2] = \mathrm{Var}_{Z\sim m}(V_{g,Z}^{\pi_g}(s)). \tag{69}$$

Indeed, the expected squared error equals this variance plus $(c_{\mathrm{val}} - \mathbb{E}_{Z\sim m}[V_{g,Z}^{\pi_g}(s)])^2$. When the conditional values differ, even an ideal scalar reference cannot represent all of them simultaneously. Moreover, setting $\bar{V}_g(s) = \mathbb{E}_{Z\sim m}[V_{g,Z}^{\pi_g}(s)]$ does not generally make $\mathbb{E}_{Z\sim m}[\widehat{P}_{g,Z}^a(V_{g,Z}^{\pi_g} - \bar{V}_g)](s)$ zero: the same z determines both the successor distribution and the continuation value. This is why HBPO conditions both the residual and the continuation on one persistent z, then averages their completed action values. Averaging the transition kernel anew at every step would describe a different stochastic process.

The theorem's assumptions hold for the implemented inventory models. Let $d_{\max}$ be the largest demand in the frozen model's finite support, and let $(I, \mathbf{p})$ be the initial inventory and pipeline. Since each accepted regular order and each expedited order is at most 8, for every simulated step $k \geq 0$,

$$\begin{aligned} |I_k| &\leq |I| + \sum_{\ell=1}^{4} p_\ell + (16 + d_{\max})k, \\ |R_k| &\leq \frac{80}{495} + |I_{k+1}|. \end{aligned} \tag{70}$$

These bounds are uniform in z; geometric discounting makes their sums finite. A frozen Tanh network with finite parameters is bounded by $\sum_i |w_{3,i}| + |b_3|$, with w_3, b_3 as in Equation (45). Thus the theorem does not require rewards or exact values to be globally bounded across initial states.

Interpretation and scope of the guarantee. The theorem establishes the correction needed to make the one-step reference exact under the frozen model. It does not make residual simulation the only possible policy-evaluation method. Its telescoping basis is closely related to multi-step temporal-difference decompositions, including generalized advantage estimation with its trace parameter set to one (Schulman et al., 2016); here the increments are model-integrated residuals and the RB latent persists throughout each continuation. The reference critic can therefore be used as a control variate without changing the population action-value target. Finite simulated tapes still introduce Monte Carlo error, and the identity alone does not guarantee a variance reduction. Model misspecification and mismatch between the simulation risk prior and the benchmark remain separate sources of error. Claims about real-environment return improvements consequently rest on the experiments, rather than on this model-evaluation identity alone.

B.2.4 FINITE COMPUTATION THROUGH GEOMETRICALLY STOPPED CONTINUATIONS.

We estimate the infinite residual series using a random, almost surely finite continuation length. For a fixed query (s, a), independently of the entire model path and of its risk Z, draw

$$N \in \{0, 1, \ldots\}, \qquad \Pr(N = \ell) = (1-\gamma)\gamma^{\ell}, \quad \ell = 0, 1, \ldots. \tag{71}$$

For any z in the common full-m-measure set of Theorem 1, all path expectations below use $\mathbb{E}_z^{s,a}$, and the outer expectation also averages over N. Since $\Pr(N \geq k) = \gamma^k$ for integers $k \geq 0$, absolute residual integrability in Theorem 1 justifies

$$\mathbb{E}\left[\sum_{k=1}^{N} e_{g,z}(S_k)\right] = \sum_{k=1}^{\infty} \Pr(N \geq k)\mathbb{E}[e_{g,z}(S_k)] = \mathbb{E}\left[\sum_{k=1}^{\infty} \gamma^k e_{g,z}(S_k)\right]. \tag{72}$$

Let $M = 32$ be the number of tapes per sampled query root, and let $b = 1, \ldots, M$ index the tapes. Tape b draws an independent $Z_b \sim m$ and a length N_b with the law above. Write $S_{b,k}^a$ for its step-k observation after forcing first action a and following π_g thereafter. An ideal action-value estimator is

$$\widehat{Q}_g(s, a) = Q_g^0(s, a) + \frac{1}{M}\sum_{b=1}^{M}\sum_{k=1}^{N_b} e_{g,Z_b}(S_{b,k}^a), \qquad M = 32. \tag{73}$$

Each tape samples its own Z_b, which is shared across the 81 alternative first-action branches and held fixed throughout their continuations. The branches also share the addressed random numbers used for geometric stopping, exogenous disturbances, and subsequent policy-action sampling. These shared numbers couple action branches without changing any one branch's marginal rollout law.

Geometric survival already supplies the discount, so the accumulated residuals receive no additional factor γ^k. The $N_b = 0$ sum is empty. These continuations have no fixed 1000-step cutoff or terminal bootstrap. In RB, the first modeled transition also uses $\widehat{\mathcal{P}}_{g,Z_b}$; it is not conditioned on the real transition's next availability.

Under the stated sampling laws and exact arithmetic, $\mathbb{E}[\widehat{Q}_g(s, a)] = Q_g^{\pi_g}(s, a)$. A finite collection of tapes still produces Monte Carlo error. For the policy-learning interface, define the sampled centered advantage by

$$\widehat{A}_g(s, a) = \widehat{Q}_g(s, a) - \sum_{a' \in \mathcal{A}} \pi_g(a' \mid s)\widehat{Q}_g(s, a'). \tag{74}$$

The implementation stores the centered residual correction directly. When all first-action branches on a tape reach the same observation, their remaining common tail cancels under this centering and need not be simulated. Thus it implements the centered estimator without materializing every absolute $\widehat{Q}_g(s, a)$. The selection of query roots and their use by policy updates are described in Section B.5.

B.2.5 THE STATE DISTRIBUTION USED TO TRAIN SUCCESSOR CRITICS.

The next critics are fitted at states reached when the current behavior and target policies use the model. Let μ_g and π_g denote the frozen parent behavior and target policies, respectively. We first specify how starting observations are selected, then how their continuations determine the training distribution. Let $s_1, \ldots, s_n$ be the observations in the generation-g batch, with $n = 25{,}000$. These are transition occurrences: repeated observations retain their multiplicity. Write δ_s for the Dirac probability measure concentrated at s, so that $\delta_s(E) = \mathbf{1}\{s \in E\}$ for any measurable observation set E.

For the training-bank construction, let $D_g = n^{-1}\sum_{i=1}^{n} \delta_{s_i}$ be the empirical observation distribution, and let ν_g^J and ν_g^P be the normalized empirical distributions obtained by the frozen target-return and transmission emphatic weights. Each selects an occurrence from this same batch; their weights are nonnegative and have empirical mean one. Their exact definitions and occurrence-sampling interpretation are in Section B.3. The critic treats these three root distributions as fixed inputs.

For RB, recall the independent synthetic risk in Equation (37). Let $m = \mathrm{Unif}[0, 1/2]$ denote its probability law, with density $m(dz) = 2\mathbf{1}_{[0,1/2]}(z)\,dz$. The symbol $\otimes$ denotes the product of probability measures: for observation and risk sets E and C, $(D_g \otimes m)(E \times C) = D_g(E)m(C)$. For

either weighting role $\chi \in \{J, P\}$, define

$$\widetilde{D}_g = D_g \otimes m, \qquad \widetilde{\nu}_g^{\chi} = \nu_g^{\chi} \otimes m, \quad \chi \in \{J, P\}, \tag{75}$$

These product measures correspond to sampling a starting observation from the indicated distribution and an independent $Z \sim m$.

Let $\widetilde{P}_{g,\sigma}$ be the transition kernel on (S, Z) that samples an action from a fixed policy σ, advances the observation using $\widehat{\mathcal{P}}_{g,Z}$, and preserves Z. For an augmented-state distribution λ, the measure $\lambda\widetilde{P}_{g,\sigma}$ is the distribution after one such transition. For any measurable set E of augmented states,

$$(\lambda\widetilde{P}_{g,\sigma})(E) = \int \widetilde{P}_{g,\sigma}((s, z), E)\, \lambda(ds, dz). \tag{76}$$

For a nonnegative integer ℓ, the power $\widetilde{P}_{g,\sigma}^{\ell}$ denotes ℓ successive transitions; its zeroth power is the identity kernel.

The three uses of the critic select roots and first actions from (D_g, μ_g), (ν_g^J, π_g), and (ν_g^P, μ_g), respectively. Mixing these uses equally gives the distribution after the first transition,

$$\widetilde{\eta}_g = \frac{1}{3}\left(\widetilde{D}_g\widetilde{P}_{g,\mu_g} + \widetilde{\nu}_g^J\widetilde{P}_{g,\pi_g} + \widetilde{\nu}_g^P\widetilde{P}_{g,\mu_g}\right), \tag{77}$$

After this first transition, all three routes follow the frozen target policy π_g. We weight their subsequent states geometrically, giving the normalized critic-training distribution

$$\widetilde{\zeta}_g = (1-\gamma)\sum_{\ell=0}^{\infty}\gamma^{\ell}\widetilde{\eta}_g\widetilde{P}_{g,\pi_g}^{\ell}. \tag{78}$$

The coefficients sum to one. In particular, the term $\ell = 0$ describes the state immediately after the first transition; it does not describe the original root observation.

Equation (78) directly specifies a sampling procedure for a finite training bank. Choose one of the three root-distribution/policy pairs above with probability $1/3$. Draw S_0 from that root distribution and an independent $Z \sim m$, sample the first action A_0 from the selected policy, and simulate the first successor S_1. The pair (S_1, Z) then has distribution $\widetilde{\eta}_g$. Independently draw the number of additional target-policy transitions,

$$G \sim \mathrm{Geom}_0(1-\gamma), \qquad \Pr(G = \ell) = (1-\gamma)\gamma^{\ell}, \quad \ell = 0, 1, \ldots. \tag{79}$$

Starting from S_1, simulate G further transitions under π_g and retain (S_{1+G}, Z). Conditional on $G = \ell$, its distribution is $\widetilde{\eta}_g\widetilde{P}_{g,\pi_g}^{\ell}$; averaging over G therefore gives exactly $\widetilde{\zeta}_g$. When $G = 0$, retain S_1 without an additional transition. The same sampled Z is used for every transition on this path.

For each retained pair (S, z), sample one additional next availability

$$U \sim \widehat{K}_{B,g,z}(B(S), \cdot). \tag{80}$$

This draw uses the same tape risk, rather than a newly sampled risk or the availability from the first transition. The generation contains $n_\zeta = 25{,}000$ such training pairs (S, U), shared by all five critics; n_ζ denotes the training-bank size.

The augmented formulation is essential: although Z is independent of the root observation, it generally becomes statistically dependent on later observations. Replacing the augmented kernel by powers of a risk-averaged observation kernel would change $\widetilde{\zeta}_g$. For DSIS, the latent coordinate disappears and $U = 1$.

B.2.6 Learning the successor critics by minimizing Bellman-consistency error.

Let V_θ denote one live scalar critic, with trainable parameter vector θ; the following objective is optimized separately for each of the five members. We fit it using a model-based form of the one-step temporal-difference (TD) error. As in TD-based critic learning, this error compares the current value prediction with the immediate reward plus a discounted bootstrap value at the successor observation.

Here, we average the TD error over actions under the frozen target policy π_g and over demand under the frozen model, conditional on the sampled next availability U. We then minimize the square of this conditional mean. The resulting population objective is

$$\mathcal{J}_{V,g}(\theta) = \mathbb{E}_{\substack{(S,Z)\sim\widetilde{\zeta}_g \\ U\sim\widehat{K}_{B,g,Z}(B(S),\cdot)}}\left[\left(V_\theta(S) - \sum_{a\in\mathcal{A}}\pi_g(a\mid S)\sum_{d\in\widehat{\mathcal{D}}_g}\widehat{K}_{D,g}(d^-(S),d) \cdot\left[r(S,a,d)+\gamma V_\theta(F(S,a,d,U))\right]\right)^2\right]. \tag{81}$$

The expression inside the square is the negative of the conditional expected TD error; this sign convention does not affect the squared loss. We optimize this objective with full gradients through both the current and successor value predictions, as detailed below.

For the frozen bank of n_ζ pairs sampled above, $\{(S_\ell,U_\ell)\}_{\ell=1}^{n_\zeta}$, define the empirical objective

$$\widehat{\mathcal{J}}_{V,g}(\theta) = \frac{1}{n_\zeta}\sum_{\ell=1}^{n_\zeta}\left[V_\theta(S_\ell) - \sum_{a\in\mathcal{A}}\pi_g(a\mid S_\ell)\sum_{d\in\widehat{\mathcal{D}}_g}\widehat{K}_{D,g}(d^-(S_\ell),d) \cdot\left\{r(S_\ell,a,d)+\gamma V_\theta(F(S_\ell,a,d,U_\ell))\right\}\right]^2. \tag{82}$$

Recall that L_g^u in Equation (46) averages the one-step reward and discounted next value over model demand, holding next availability u fixed. Define the supply-conditioned target-policy Bellman backup $\mathcal{T}_g^u$ and its signed error ε_g by

$$\mathcal{T}_g^u V_\theta(s) = \sum_{a\in\mathcal{A}}\pi_g(a\mid s)L_g^u V_\theta(s,a), \tag{83}$$

$$\varepsilon_g(\theta;s,u) = V_\theta(s) - \mathcal{T}_g^u V_\theta(s). \tag{84}$$

The order of operations is part of the objective: actions and demand are integrated before squaring, while next availability is sampled and retained in the training pair. Recall that $\mathcal{T}_{g,z}^{\pi_g}$ in Equation (49) additionally averages over next availability. For fixed (s,z), let $\mathbb{E}_U$ and Var_U denote expectation and variance under $U\sim\widehat{K}_{B,g,z}(B(s),\cdot)$. Then

$$\mathbb{E}_U\left[\left(V_\theta(s)-\mathcal{T}_g^U V_\theta(s)\right)^2\right] = \left[V_\theta(s)-(\mathcal{T}_{g,z}^{\pi_g}V_\theta)(s)\right]^2 + \mathrm{Var}_U\left[\mathcal{T}_g^U V_\theta(s)\right]. \tag{85}$$

Therefore, the RB fitting objective is generally different from the square of the fully supply-integrated Bellman residual. The distinction vanishes in DSIS, where next availability is deterministic.

Both value terms belong to the same live member and remain differentiable. Write ∇_θ for the gradient with respect to critic parameters θ. In particular,

$$\nabla_\theta\varepsilon_g(\theta;s,u) = \nabla_\theta V_\theta(s) - \gamma\sum_{a\in\mathcal{A}}\pi_g(a\mid s)\sum_{d\in\widehat{\mathcal{D}}_g}\widehat{K}_{D,g}(d^-(s),d)\nabla_\theta V_\theta(F(s,a,d,u)), \tag{86}$$

and

$$\nabla_\theta\widehat{\mathcal{J}}_{V,g}(\theta) = \frac{2}{n_\zeta}\sum_{\ell=1}^{n_\zeta}\varepsilon_g(\theta;S_\ell,U_\ell)\nabla_\theta\varepsilon_g(\theta;S_\ell,U_\ell). \tag{87}$$

The model, policy, and sampled bank are fixed during this optimization. There is no stop-gradient or slowly updated target critic on the successor-value side of this loss.

B.2.7 Finite update and use of the successor.

For member $j\in\{1,\dots,5\}$, write $\theta_{g,j}$ for its frozen parent parameter vector and $\theta_j^{(k)}$ for its live parameters after k optimizer updates. It starts from $\theta_j^{(0)}=\theta_{g,j}$, and performs 1000 Adam updates

with minibatch size 256 and the configured value learning rate. Adam uses first- and second-moment decay coefficients $(\beta_1, \beta_2) = (0.9, 0.999)$, denominator stabilizer $\epsilon_{\mathrm{Adam}} = 10^{-8}$, zero weight decay, and fresh optimizer state; gradients are not clipped. Members share the frozen training bank but use separately addressed minibatch schedules. The successor is the finite optimization output,

$$V_{g+1,j} = V_{\theta_j^{(1000)}}, \qquad v_{g+1} = \frac{1}{5}\sum_{j=1}^{5} V_{g+1,j}. \tag{88}$$

This definition does not presume an exact minimizer or Bellman fixed point.

The subscript $g+1$ records when the fitted critic becomes available; the policy in its fitting objective is still the frozen parent π_g. Throughout this fit, current-generation residual estimation, heuristic queries, and actor updates continue to use v_g. The fitted mean v_{g+1} becomes their value reference in the next generation. This separates the live network being optimized from the frozen reference used to construct the current generation's learning signals.

B.2.8 Warm and cold critic initialization.

Both initialization routes produce five scalar critics $V_{0,j}$, $j \in \{1, \ldots, 5\}$. The warm route first evaluates the initial categorical policy π_0 with temporary Q networks and transfers their centered action information; the cold route directly fits scalar critics on simulated states. Both finish with the same constant-offset calibration.

Warm initialization and temporary Q evaluation. Warm initialization uses the counted 500,000-transition source prefix. After constructing the initial categorical policy π_0, five temporary action-value networks are initialized from the source critic. A scalar source value is copied across action outputs; for SAC, the five initial networks use its two online Q networks, two target Q networks, and an additional copy of the first online Q network. These copies have separate parameter storage.

Write $n_0 = 500{,}000$ for the number of initialization pairs. For source occurrence $i \in \{1, \ldots, n_0\}$, let s_i be its observation, d_i its realized demand, and u_i its next availability, all obtained from the recorded transition. In RB, d_i is recorded in the next observation; in DSIS, it is recovered from the public inventory balance using the recorded action and current and next observations. Public dynamics provide counterfactual successors and rewards for every action:

$$s'_{i,a} = F(s_i, a, d_i, u_i), \qquad r_{i,a} = r(s_i, a, d_i). \tag{89}$$

No unobserved demand or supply realization is needed for this construction.

For fitted Q evaluation (FQE), let $Q_j^{(k)}$ and $Q_j^{-,(k)}$ denote member j's online and target action-value networks at iteration k, with parameter vectors $\psi_j^{(k)}$ and $\psi_j^{-,(k)}$, respectively. Each target network starts as a separate copy of its online network. The initial policy π_0 remains fixed. For every recorded occurrence i and action $a \in \mathcal{A}$, define the detached target

$$y_{j,i,a}^{(k)} = r_{i,a} + \gamma \sum_{a' \in \mathcal{A}} \pi_0(a' \mid s'_{i,a}) Q_j^{-,(k)}(s'_{i,a}, a'). \tag{90}$$

Here "detached" means that differentiation of the FQE loss does not propagate through $y_{j,i,a}^{(k)}$. At each iteration, draw an ordered minibatch $\mathcal{I}^{(k)} = (i_1, \ldots, i_{n_{\mathrm{b}}})$ of $n_{\mathrm{b}} = 256$ occurrence indices uniformly with replacement from $\{1, \ldots, n_0\}$; the five members share this minibatch. For each member j, minibatch position h, and action a, draw a fresh binary mask $\xi_{j,h,a}^{(k)}$, independently across positions, actions, members, and iterations, with probability 0.8 of being one. Thus $\xi_{j,h,a}^{(k)} \sim \mathrm{Bernoulli}(0.8)$. The loss is

$$\mathcal{L}_j^{\mathrm{FQE},(k)} = \frac{1}{n_{\mathrm{b}}|\mathcal{A}|} \sum_{h=1}^{n_{\mathrm{b}}} \sum_{a \in \mathcal{A}} \xi_{j,h,a}^{(k)} \left[Q_j^{(k)}(s_{i_h}, a) - y_{j,i_h,a}^{(k)} \right]^2. \tag{91}$$

The denominator counts every minibatch-action pair, including masked pairs; it is not the number of nonzero masks. After the Adam update of the online parameters to $\psi_j^{(k+1)}$, update the target by

$$\psi_j^{-,(k+1)} = 0.995\, \psi_j^{-,(k)} + 0.005\, \psi_j^{(k+1)}. \tag{92}$$

FQE performs 2000 updates per member. Its Adam learning rate is $1.2020438491997917\times10^{-4}$ for PPO, $4.418119410401287 \times 10^{-5}$ for ACE, and $5.669429427494781 \times 10^{-4}$ for the other source methods. These target Q networks belong only to temporary initialization; generation-level scalar-critic fitting uses the full gradient through both value predictions, as specified above. Write $Q_j = Q_j^{(2000)}$ for the final online Q network used in the next step.

Scalar fitting for the two initialization routes. Five fresh scalar critics are initialized with separately randomized orthogonal hidden weights, zero hidden biases, and zero output weights and biases. In the warm case, let the fitted temporary Q networks define

$$A_j^{\text{init}}(s,a) = Q_j(s,a) - \sum_{a' \in \mathcal{A}} \pi_0(a' \mid s) Q_j(s,a'). \tag{93}$$

The advantage labels A_j^{init} and the policy π_0 are held fixed during scalar fitting. The operator L^u_{init} is the supply-conditioned backup in Equation (46), evaluated with the frozen initialization model. The warm scalar-fitting objective is

$$\mathcal{J}_j^{\text{warm}}(\theta) = \frac{1}{n_0} \sum_{i=1}^{n_0} \sum_{a \in \mathcal{A}} \pi_0(a \mid s_i) \cdot \left[L^{u_i}_{\text{init}} V_\theta(s_i,a) - V_\theta(s_i) - A_j^{\text{init}}(s_i,a) \right]^2. \tag{94}$$

Here u_i is the observed next availability in the source prefix, while $L^{u_i}_{\text{init}}$ integrates demand using the frozen initialization model. Both occurrences of V_θ are differentiated. This objective transfers centered action information from the temporary Q critics into the scalar value representation.

Cold initialization instead creates 500 synthetic trajectories of length 1000 under π_0. Their $n_0 = 500{,}000$ current-observation/ next-availability pairs are again denoted by (s_i, u_i); here both coordinates come from model simulation. Each member minimizes

$$\mathcal{J}_j^{\text{cold}}(\theta) = \frac{1}{n_0} \sum_{i=1}^{n_0} \left[V_\theta(s_i) - \sum_{a \in \mathcal{A}} \pi_0(a \mid s_i) L^{u_i}_{\text{init}} V_\theta(s_i,a) \right]^2. \tag{95}$$

For RB, this simulation starts only after collecting the initial counted 25,000 transitions and fitting their model. Initial observations are sampled uniformly, with multiplicity, from the observed joint reset-state occurrences. Every synthetic trajectory receives an independent $Z \sim m = \text{Unif}[0, 1/2]$, fixed for its entire length. For DSIS, simulation uses the public initial state and known model. The cold bank does not reuse private environment tapes.

Calibrating the constant value offset. For either route, each scalar critic is optimized for two complete epochs with minibatch size 256 and Adam learning rate $5.669429427494781 \times 10^{-4}$. Each epoch visits all n_0 pairs once in a member-specific permutation; the last minibatch contains 32 pairs. Adam state is initialized once for this fit and retained across its two epochs, with the same moment coefficients and denominator stabilizer as in generation-level fitting. Let $\widetilde{V}_j$ be the resulting network. To describe the final calibration, specialize the supply-conditioned Bellman backup in Equation (83) to the initial policy and model:

$$\mathcal{T}^u_{\text{init}} V(s) = \sum_{a \in \mathcal{A}} \pi_0(a \mid s) L^u_{\text{init}} V(s,a). \tag{96}$$

This operator returns the expected immediate reward plus discounted next value, averaging over the action and model demand while holding the next availability u fixed. Its conditioning matches the initialization pairs (s_i, u_i): u_i is observed in the warm bank and generated by the model in the cold bank. It is the conditional counterpart of $\mathcal{T}^{\pi_g}_{g,z}$ in Equation (49), which additionally averages over next availability.

Compute the full-bank mean backup-minus-prediction residual,

$$\overline{e}_j = \frac{1}{n_0} \sum_{i=1}^{n_0} \left[(\mathcal{T}^{u_i}_{\text{init}} \widetilde{V}_j)(s_i) - \widetilde{V}_j(s_i) \right]. \tag{97}$$

Adding a constant c_{off} to every value prediction adds γc_{off} to the discounted next-value term, because the action and demand probabilities each sum to one. Therefore,

$$\mathcal{T}^u_{\text{init}}(\widetilde{V}_j + c_{\text{off}})(s) = \mathcal{T}^u_{\text{init}} \widetilde{V}_j(s) + \gamma c_{\text{off}}, \tag{98}$$

and hence the residual changes by $-(1-\gamma)c_{\text{off}}$ at every pair:

$$\mathcal{T}^u_{\text{init}}(\widetilde{V}_j + c_{\text{off}})(s) - (\widetilde{V}_j(s) + c_{\text{off}}) = \mathcal{T}^u_{\text{init}}\widetilde{V}_j(s) - \widetilde{V}_j(s) - (1-\gamma)c_{\text{off}}. \tag{99}$$

The final initialization chooses

$$V_{0,j}(s) = \widetilde{V}_j(s) + \frac{\overline{e}_j}{1-\gamma}, \tag{100}$$

implemented by shifting the scalar output bias (up to floating-point rounding of that bias). The new mean residual is $\overline{e}_j - (1-\gamma)\overline{e}_j/(1-\gamma) = 0$ in exact arithmetic. This removes a constant offset in the empirical Bellman error; it does not set every statewise residual to zero. After warm initialization, the temporary Q networks are discarded.

B.3 Emphatic Weighting: Mass Learning, Heuristic Interest, and Cross-Generation Use

B.3.1 Meaning of emphatic weighting and the collected data.

Emphatic weighting assigns state-dependent importance to learning updates by propagating an interest signal through transitions. Here interest is an externally specified or data-derived nonnegative function that states which locations should receive additional learning emphasis. This separates how frequently an observation occurs in behavior data from how strongly it contributes to a learning objective. At learning generation g, HBPO uses two weighting roles: target-return learning, denoted by J, and behavior-to-target transmission, denoted by P. We call their positive state scores M^J_g and M^P_g raw masses. They are functions of the observation, not normalized state distributions. Normalization and shrinkage convert these scores into weights and then into empirical probability measures for the corresponding learning objectives.

Let $\mathcal{D}_g$ be the indexed multiset of $n = 25{,}000$ transitions collected along trajectories of the frozen behavior policy μ_g. The index $i = 1, \dots, n$ identifies an occurrence, including repeated observations and transitions:

$$\mathcal{D}_g = \big((s_i, a_i, s'_i)\big)_{i=1}^n. \tag{101}$$

Recall that D_g averages over current observations. Let D'_g average over successor observations and let Γ_g retain the paired transition occurrences. Writing δ_x for unit point mass at x, these empirical measures are

$$D_g = \frac{1}{n}\sum_{i=1}^n \delta_{s_i}, \qquad D'_g = \frac{1}{n}\sum_{i=1}^n \delta_{s'_i}, \qquad \Gamma_g = \frac{1}{n}\sum_{i=1}^n \delta_{(s_i,a_i,s'_i)}. \tag{102}$$

Repeated observations retain their occurrence multiplicities. Masses are fitted at successor observations, whereas the weights consumed in generation g are computed from the parent masses at current observations. The symbol $\chi \in \{J, P\}$ denotes a weighting role.

B.3.2 Emphatic weighting in prior work.

The emphatic TD (ETD) construction of Sutton et al. (2016) propagates user-specified interest through target-policy transitions to weight bootstrapped value updates. Its linear stability theorem concerns the expected update under the paper's assumptions; it is not a stability theorem for nonlinear mass networks. To summarize the follow-on part of this construction, consider a finite Markov model, fixed behavior and target policies μ and π, a constant discount $0 < \gamma < 1$, and a fixed nonnegative interest function ι. Assume that d_μ is a strictly positive stationary state distribution under μ, that $S_0 \sim d_\mu$, and that $\mu(a \mid s) > 0$ whenever $\pi(a \mid s) > 0$. Generate the trajectory under μ. Let k index its steps, and let F^{fo}_k denote the follow-on trace. Initialize $F^{\text{fo}}_0 = \iota(S_0)$; for $k \geq 1$, using the target-to-behavior action ratio ϱ_{k-1}, set

$$F^{\text{fo}}_k = \iota(S_k) + \gamma\varrho_{k-1}F^{\text{fo}}_{k-1}, \qquad \varrho_{k-1} = \frac{\pi(\mathsf{A}_{k-1} \mid S_{k-1})}{\mu(\mathsf{A}_{k-1} \mid S_{k-1})}. \tag{103}$$

Let $P(s' \mid s, a)$ be this model's transition probability and let $P_\pi(s, s') = \sum_a \pi(a \mid s)P(s' \mid s, a)$ be its target-policy transition matrix. Define the expected follow-on state mass by $f_\iota(s) = d_\mu(s)\lim_{k\to\infty}\mathbb{E}[F^{\text{fo}}_k \mid S_k = s]$. The mass satisfies

$$f_\iota = d_\mu \odot \iota + \gamma P_\pi^\top f_\iota, \qquad f_\iota = (\mathrm{Id} - \gamma P_\pi^\top)^{-1}(d_\mu \odot \iota). \tag{104}$$

Thus the conditional mean trace and the state mass obtained by multiplying it by d_μ are different objects. We use this follow-on component to explain the connection, rather than identifying it with the full emphasis of every ETD trace-parameter setting.

OffPAC (Degris et al., 2012) corrects sampled actions with importance ratios but retains behavior-state weighting in its actor semi-gradient. For this literature comparison, let ϑ be the actor parameter vector and let π_ϑ be a differentiable target policy. For the displayed score formula, assume that both $\mu(a \mid s)$ and $\pi_\vartheta(a \mid s)$ are strictly positive for every state and action. Let Q^{π_ϑ} and V^{π_ϑ} be the exact discounted action and state values in the finite Markov model, and let ∇_ϑ denote differentiation with respect to ϑ. Using exact action values to isolate the state-weighting issue, write OffPAC's basic actor direction as $g_{\mathrm{OffPAC}}(\vartheta)$:

$$\begin{aligned} g_{\mathrm{OffPAC}}(\vartheta) &= \mathbb{E}_{S\sim d_\mu,\, \mathsf{A}\sim\mu(\cdot|S)}\left[\frac{\pi_\vartheta(\mathsf{A}\mid S)}{\mu(\mathsf{A}\mid S)} Q^{\pi_\vartheta}(S,\mathsf{A})\nabla_\vartheta \log \pi_\vartheta(\mathsf{A}\mid S)\right] \\ &= \sum_s d_\mu(s)\sum_a \nabla_\vartheta \pi_\vartheta(a\mid s) Q^{\pi_\vartheta}(s,a). \end{aligned} \tag{105}$$

Action importance sampling changes the action average to the target policy, while the outer state average remains weighted by d_μ.

The policy-gradient theorem underlying Actor–Critic with Emphatic weightings (ACE; Imani et al., 2018) supplies the emphatic state weighting for the exact gradient of the interest-weighted excursion objective $\mathcal{J}_\iota$:

$$\mathcal{J}_\iota(\vartheta) = \sum_s d_\mu(s)\iota(s)V^{\pi_\vartheta}(s). \tag{106}$$

Holding μ, d_μ, and ι fixed when differentiating, its policy-gradient theorem gives

$$\nabla_\vartheta \mathcal{J}_\iota(\vartheta) = \sum_s f_\iota(s)\sum_a \nabla_\vartheta \pi_\vartheta(a\mid s) Q^{\pi_\vartheta}(s,a), \tag{107}$$

where f_ι is the follow-on state mass in Equation (104), evaluated at $\pi = \pi_\vartheta$. For unit interest, comparison with Equation (105) isolates the missing follow-on state weighting in the semi-gradient. This equality uses exact values and state weights; finite trace and value estimates introduce additional approximation. Graves et al. (2023, Section 5.2) also study directly learning conditional emphatic weights from predecessor transitions. This motivates the regression viewpoint below. HBPO specifies its own labels, anchored mass functions, shrinkage, and generation schedule; the cited results do not certify these modifications.

B.3.3 Why learn positive masses, and what the regression represents.

The purpose of the mass networks is to store a state-dependent importance field that can be evaluated on a later generation's observations. A recorded transition provides a predecessor carrying existing mass and a successor at which that mass should contribute to learning. Fitting a shared function at successor observations pools this incoming evidence and generalizes it across the fixed observation features. The two roles use different labels, specified below, but share the following parameterization and fitting principle.

Recall that $\chi \in \{J, P\}$ indexes the two raw mass functions. For each role, let $f_{\omega_g^\chi}$ be a separate scalar network with parameter vector ω_g^χ, two width-128 Tanh hidden layers, and a scalar output. Recall the fixed observation features ϕ from Equation (44). Let $s_\circ$ be the reference observation with zero inventory and pipeline and available regular supply; its previous-demand coordinate is also zero when present. Let $\beta_g^\chi > 0$ be the prescribed mass at this reference. Define

$$M_g^\chi(s) = \beta_g^\chi \exp\Big[f_{\omega_g^\chi}(\phi(s)) - f_{\omega_g^\chi}(\phi(s_\circ))\Big], \qquad \beta_g^\chi > 0. \tag{108}$$

Thus $M_g^\chi(s) > 0$ and $M_g^\chi(s_\circ) = \beta_g^\chi$. The reference value is imposed by the parameterization. If $s_\circ$ occurs among the observed successors, this value need not equal the unrestricted regression target there.

For a fixed role χ, let $Y_i^\chi > 0$ be the regression label attached to transition occurrence i; the two roles construct these labels below. For a candidate positive mass function M, define the full-data

loss $\mathcal{L}_g^\chi(M)$ by

$$\mathcal{L}_g^\chi(M) = \frac{1}{n}\sum_{i=1}^{n}\left[M(s_i') - Y_i^\chi \log M(s_i')\right]. \tag{109}$$

This is the Poisson-form loss (a quasi-likelihood objective for positive real labels), with label-only terms omitted. It does not assume that these labels are integer-valued Poisson observations. Recall that Γ_g in Equation (102) samples recorded transition occurrences uniformly. More explicitly, draw an occurrence index I uniformly from $\{1,\dots,n\}$ and set $(S,\mathsf{A},S',Y^\chi) = (s_I,a_I,s_I',Y_I^\chi)$. This defines the joint empirical law including the label; expectations subscripted by Γ_g below use this attached-label law when needed. Its current and successor marginals remain D_g and D_g'.

For an observed successor s', write $x_{\log} = \log M(s')$, with log denoting the natural logarithm. Let $\ell(x_{\log}; s')$ be the conditional loss. Its value and first two derivatives are

$$\begin{aligned}\ell(x_{\log}; s') &= e^{x_{\log}} - \mathbb{E}_{\Gamma_g}[Y^\chi \mid S' = s']\, x_{\log},\\ \frac{\partial \ell}{\partial x_{\log}} &= e^{x_{\log}} - \mathbb{E}_{\Gamma_g}[Y^\chi \mid S' = s'], \qquad \frac{\partial^2 \ell}{\partial x_{\log}^2} = e^{x_{\log}} > 0.\end{aligned} \tag{110}$$

Let $M^\star(s')$ denote the unrestricted pointwise minimizer at an observed successor. The positive conditional mean and strict convexity give

$$M^\star(s') = \mathbb{E}_{\Gamma_g}[Y^\chi \mid S' = s']. \tag{111}$$

A successor receives the average label carried into it by its observed predecessors. This calculation explains the meaning of the labels and the choice of loss. The actual learner approximates this regression with the anchored network and finite optimization; it does not set each prediction equal to the unrestricted minimizer.

B.3.4 TARGET-RETURN MASS M^J.

The J mass addresses the difference between the states represented by behavior data and those reached when the target policy continues from that behavior-distributed source. Action importance ratios alone change the action distribution at a given observation; they do not change how often that observation contributes to the update. Propagating mass through target-policy transitions provides a state-level adjustment for the target-return channel. Its ideal reference is specified below, so this motivation does not presume a stationary target-occupancy ratio.

Recall that π_g and μ_g are the frozen parent target and behavior policies. Define their action-ratio function ϱ_g and its value at recorded occurrence i by

$$\varrho_g(s,a) = \frac{\pi_g(a \mid s)}{\mu_g(a \mid s)}, \qquad \varrho_{g,i} = \varrho_g(s_i,a_i). \tag{112}$$

The implementation evaluates this ratio from parent-policy log probabilities without clipping; the categorical policies have common positive action support. The J learner has an online parameter vector $\omega_{g,k}^J$ and a private slow copy $\overline{\omega}_{g,k}^J$ at inner iteration k, initialized by $\omega_{g,0}^J = \overline{\omega}_{g,0}^J = \omega_g^J$. For $k = 0,\dots,999$, define the slow mass $\overline{M}_{g,k}^J$ and the detached regression label $Y_{g,i,k}^J$ as

$$\overline{M}_{g,k}^J(s) = \beta_g^J \exp\left[f_{\overline{\omega}_{g,k}^J}(\phi(s)) - f_{\overline{\omega}_{g,k}^J}(\phi(s_\circ))\right], \tag{113}$$

$$Y_{g,i,k}^J = (1-\gamma) + \gamma \varrho_{g,i} \overline{M}_{g,k}^J(s_i). \tag{114}$$

The source term $1-\gamma$ injects positive mass; the continuation term carries predecessor mass through the recorded transition with an action correction toward π_g. Recall that β_g^J is the parent reference mass. At iteration k, the online prediction uses the parameterization in Equation (108) with parameters $\omega_{g,k}^J$ and successor reference value $\beta_{g+1}^J = 1-\gamma$. It is fitted at s_i' using Equation (109) with $Y_i^\chi = Y_{g,i,k}^J$, holding the labels fixed during that optimization step. After the step, the slow parameters are updated by

$$\overline{\omega}_{g,k+1}^J = 0.995\,\overline{\omega}_{g,k}^J + 0.005\,\omega_{g,k+1}^J. \tag{115}$$

The slow mass retains β_g^J throughout the generation, while the online mass uses β_{g+1}^J. Only the final online network is committed as M_{g+1}^J.

For a scalar state function h, define the empirical reverse-transition operator $\mathcal{B}_g^J$ at each observed successor by

$$(\mathcal{B}_g^J h)(s') = \mathbb{E}_{\Gamma_g}[\varrho_g(S, \mathsf{A})h(S) \mid S' = s']. \tag{116}$$

For a fixed inner iteration k, let $M_{g,k}^{J,\star}(s')$ be the unrestricted pointwise minimizer with the slow mass $\overline{M}_{g,k}^J$ held fixed. Recalling the regression identity in Equation (111), we obtain

$$M_{g,k}^{J,\star}(s') = (1-\gamma) + \gamma(\mathcal{B}_g^J \overline{M}_{g,k}^J)(s'). \tag{117}$$

This empirical identity holds on the successor-data support and does not require $D_g = D_g'$.

For an ideal population reference, now replace the empirical transition law Γ_g by the stationary transition law of a finite Markov model. Fix policies μ and π, assume a strictly positive stationary behavior distribution d_μ and target-action support covered by μ, and write $P(s' \mid s, a)$ for the model's transition probability. Define the population joint law Γ_μ and target transition matrix P_π by

$$\Gamma_\mu(s, a, s') = d_\mu(s)\mu(a \mid s)P(s' \mid s, a), \qquad P_\pi(s, s') = \sum_a \pi(a \mid s)P(s' \mid s, a). \tag{118}$$

Both state marginals of Γ_μ are d_μ. This replacement is a population idealization, not a consequence of equality of empirical marginals. Remove the neural parameterization and its reference-anchor constraint for this idealization. Suppress the generation and inner-iteration indices, and let M^J be the unrestricted positive fixed point of the population conditional equation

$$M^J(s') = (1-\gamma) + \gamma \mathbb{E}_{\Gamma_\mu}\left[\frac{\pi(\mathsf{A} \mid S)}{\mu(\mathsf{A} \mid S)} M^J(S) \,\middle|\, S' = s'\right]. \tag{119}$$

Define its induced state mass e^J by

$$e^J(s) = d_\mu(s)M^J(s), \qquad \text{or} \qquad e^J = d_\mu \odot M^J. \tag{120}$$

This multiplication converts a per-occurrence importance multiplier into a state mass: behavior visits s with frequency $d_\mu(s)$, and each such occurrence contributes mass $M^J(s)$. For any scalar test function h, it follows that $\mathbb{E}_{d_\mu}[M^J(S)h(S)] = \sum_s e^J(s)h(s)$. Multiplying the population conditional fixed-point equation by $d_\mu(s')$ gives

$$\begin{aligned} e^J(s') &= (1-\gamma)d_\mu(s') + \gamma \sum_s \sum_{a:\mu(a|s)>0} d_\mu(s)\mu(a \mid s)P(s' \mid s, a)\frac{\pi(a \mid s)}{\mu(a \mid s)} M^J(s) \\ &= (1-\gamma)d_\mu(s') + \gamma \sum_s P_\pi(s, s')e^J(s). \end{aligned} \tag{121}$$

Here P is the population transition probability just defined. Summing over s' gives $\sum_s e^J(s) = 1$. Recall that state masses are column vectors, $\top$ denotes transpose, and Id is the identity matrix. Expanding the inverse in powers indexed by the nonnegative integer ℓ yields

$$e^J = (1-\gamma)(\mathrm{Id} - \gamma P_\pi^\top)^{-1} d_\mu = (1-\gamma)\sum_{\ell=0}^{\infty} \gamma^\ell (P_\pi^\top)^\ell d_\mu. \tag{122}$$

The ideal reference is the discounted visitation distribution of target-policy excursions beginning from d_μ. The multiplier $M^J(s) = e^J(s)/d_\mu(s)$ adjusts behavior-state frequencies toward this reference. The reference distribution generally differs from a stationary target distribution d_π satisfying $P_\pi^\top d_\pi = d_\pi$, and from a finite evaluation trajectory's occupancy. Uniqueness of d_π is not needed for this reference construction.

B.3.5 From heuristic improvement to confidence and interest.

The P role receives interest from the heuristic's estimated improvement over the parent target policy. Recall that w_i^J is the positive, mean-one weight of current observation occurrence i, obtained from

the parent mass M_g^J by Equation (138). Draw 16 anchor occurrences with replacement, selecting occurrence i with probability w_i^J/n; denote them by $\bar{s}_{g,f}$ for $f = 1, \ldots, 16$. At each anchor, run 32 paired simulations, indexed by $b = 1, \ldots, 32$, with horizon $H = 1000$ and period index $k = 0, \ldots, H-1$. Recall the heuristic controller $q_{H,g}$ defined in Equation (1). Superscripts H and π denote the intervention and baseline arms, while $S^{\mathrm{H}}_{g,f,b,k}, S^{\pi}_{g,f,b,k}$ and $\mathsf{A}^{\mathrm{H}}_{g,f,b,k}, \mathsf{A}^{\pi}_{g,f,b,k}$ denote their observations and actions. Both arms start at $S^{\mathrm{H}}_{g,f,b,0} = S^{\pi}_{g,f,b,0} = \bar{s}_{g,f}$, with action laws

$$\mathsf{A}^{\mathrm{H}}_{g,f,b,k} \sim \begin{cases} q_{H,g}(\cdot \mid S^{\mathrm{H}}_{g,f,b,k}), & 0 \le k < 9, \\ \pi_g(\cdot \mid S^{\mathrm{H}}_{g,f,b,k}), & 9 \le k < H, \end{cases} \qquad \mathsf{A}^{\pi}_{g,f,b,k} \sim \pi_g(\cdot \mid S^{\pi}_{g,f,b,k}), \qquad 0 \le k < H. \tag{123}$$

Each arm advances from its own observation using the frozen fitted reward–observation kernel in Equation (40). Within each pair, the arms use common random inputs, including the root risk draw in RB (Equation (37)); that risk is retained throughout the pair as in Equation (41).

Let $C^{\mathrm{H}}_{g,f,b,k}$ and $C^{\pi}_{g,f,b,k}$ be the realized period costs of the two arms, and let $R^{\mathrm{H}}_{g,f,b,k}$ and $R^{\pi}_{g,f,b,k}$ be their normalized rewards. Using $R = -C/495$ from Equation (17), define the complete paired improvement $\Delta_{g,f,b}$ by

$$\Delta_{g,f,b} = \sum_{k=0}^{H-1} \left(C^{\pi}_{g,f,b,k} - C^{\mathrm{H}}_{g,f,b,k} \right) = 495 \sum_{k=0}^{H-1} \left(R^{\mathrm{H}}_{g,f,b,k} - R^{\pi}_{g,f,b,k} \right). \tag{124}$$

Positive $\Delta_{g,f,b}$ means that the finite heuristic intervention reduced total cost over the H simulated periods. These are complete undiscounted H-period returns. The common factor 495 cancels from the standardized statistic. Let $\overline{\Delta}_{g,f}$ and $\widehat{\sigma}^2_{g,f}$ be the paired sample mean and sample variance:

$$\overline{\Delta}_{g,f} = \frac{1}{32} \sum_{b=1}^{32} \Delta_{g,f,b}, \qquad \widehat{\sigma}^2_{g,f} = \frac{1}{31} \sum_{b=1}^{32} (\Delta_{g,f,b} - \overline{\Delta}_{g,f})^2. \tag{125}$$

Let $\widehat{\sigma}_{g,f}$ be the nonnegative square root of this sample variance and let $\mathrm{CDF}_{t_{31}}$ denote the cumulative distribution function of the standard Student-t distribution with 31 degrees of freedom. For $\widehat{\sigma}_{g,f} > 0$, define the confidence score $c_{g,f}$ by

$$c_{g,f} = \mathrm{CDF}_{t_{31}} \left(\frac{\sqrt{32}\, \overline{\Delta}_{g,f}}{\widehat{\sigma}_{g,f}} \right). \tag{126}$$

At zero sample standard deviation, set $c_{g,f}$ to 1, $1/2$, or 0 according as the mean improvement is positive, zero, or negative. This Student-t transformation supplies a bounded directional score; no calibrated posterior-probability interpretation is required.

Recall the fixed feature vector ϕ. Define the normalized Gaussian interpolation coefficients $a_{g,f}(s)$ and the resulting confidence field $\widehat{c}_g(s)$ below, where $\| \cdot \|_2$ is the Euclidean norm and h in the denominator is a dummy anchor index:

$$\begin{aligned} a_{g,f}(s) &= \frac{\exp[-\|\phi(s) - \phi(\bar{s}_{g,f})\|_2^2/2]}{\sum_{h=1}^{16} \exp[-\|\phi(s) - \phi(\bar{s}_{g,h})\|_2^2/2]}, \\ \widehat{c}_g(s) &= \sum_{f=1}^{16} a_{g,f}(s) c_{g,f}. \end{aligned} \tag{127}$$

Let i_g^P be the transmission interest. Recall that D'_g is the empirical measure of all successor observations s'_i. Centering the confidence field over this complete successor data gives

$$i_g^P(s) = 1 + \widehat{c}_g(s) - \mathbb{E}_{D'_g}[\widehat{c}_g], \qquad \mathbb{E}_{D'_g}[\widehat{c}_g] = \frac{1}{n} \sum_{i=1}^{n} \widehat{c}_g(s'_i). \tag{128}$$

Hence $\mathbb{E}_{D'_g}[i_g^P] = 1$. For finite features, all interpolation coefficients are strictly positive. If the anchor scores differ, each interpolated score lies strictly between their minimum and maximum; if all scores coincide, the field is constant. Together with $c_{g,f} \in [0,1]$, this proves $i_g^P(s) > 0$ in exact arithmetic. The interest emphasizes locations with confidence above the successor-data average. Constant confidence, including uniformly high confidence, gives $i_g^P \equiv 1$.

B.3.6 TRANSMISSION MASS M^P.

The transmission channel requires a different notion of state importance. Behavior visitation describes where experience was collected, but it does not indicate where a finite heuristic intervention has stronger evidence of improving on the parent target. The P mass converts that spatial variation in confidence into persistent state emphasis for transmission. Its comparison is between behavior visitation and a confidence-weighted reference propagated under the behavior policy. It does not estimate visitation under $q_{H,g}$ or directly measure behavior–target action disagreement; the latter enters the separate transmission objective.

Recall that M_g^P is the frozen parent transmission mass and that D_g averages over current observations s_i. Define its normalized multiplier R_g^P by

$$R_g^P(s) = \frac{M_g^P(s)}{\mathbb{E}_{D_g}[M_g^P]}, \qquad \mathbb{E}_{D_g}[R_g^P] = 1. \tag{129}$$

Before optimization, construct an unnormalized label $\widetilde{Y}_{g,i}^P$ and its full-data mean-one normalization $Y_{g,i}^P$ for every transition occurrence i. Freeze this complete label vector throughout the fit:

$$\widetilde{Y}_{g,i}^P = i_g^P(s_i') + \gamma R_g^P(s_i), \qquad Y_{g,i}^P = \frac{\widetilde{Y}_{g,i}^P}{n^{-1}\sum_{j=1}^n \widetilde{Y}_{g,j}^P}. \tag{130}$$

Because interest and normalized parent mass have unit mean on their respective data marginals, exact arithmetic gives

$$Y_{g,i}^P = \frac{i_g^P(s_i') + \gamma R_g^P(s_i)}{1+\gamma}. \tag{131}$$

The implementation computes the full-data denominator explicitly. This label contains no action importance ratio and no evolving slow network: it combines successor interest with normalized predecessor mass carried along behavior transitions. For a scalar state function h, define the empirical reverse-transition operator $\mathcal{B}_g^P$ at each observed successor by

$$(\mathcal{B}_g^P h)(s') = \mathbb{E}_{\Gamma_g}[h(S) \mid S' = s']. \tag{132}$$

Let $M_g^{P,\star}$ be the unrestricted conditional regression target. By Equation (111), on the successor-data support it is

$$M_g^{P,\star}(s') = \frac{i_g^P(s') + \gamma(\mathcal{B}_g^P R_g^P)(s')}{1+\gamma}. \tag{133}$$

The online network starts from the parent parameters and is fitted with Equation (109) using the fixed labels $Y_i^\chi = Y_{g,i}^P$. Recall the reference observation $s_\circ$ in Equation (108); the successor reference value is set to $\beta_{g+1}^P = i_g^P(s_\circ)$. This is an imposed anchor. If $s_\circ$ occurs among the observed successors, it need not equal the unrestricted conditional target there.

For the ideal population reference, recall the stationary law Γ_μ and positive behavior distribution d_μ in Equation (118). Define the behavior-policy transition matrix $P_\mu(s,s') = \sum_a \mu(a \mid s)P(s' \mid s,a)$. Freeze a positive interest function i^P normalized by $\mathbb{E}_{d_\mu}[i^P] = 1$, and replace the empirical conditional expectations by expectations under Γ_μ. For any normalized predecessor multiplier, the resulting unrestricted conditional target has mean $(1+\gamma)/(1+\gamma) = 1$ under d_μ. At an exact fixed point, therefore, the raw output equals its normalized multiplier, which we denote by R^P. Define the associated state mass $e^P(s) = d_\mu(s)R^P(s)$. Multiplication of the population conditional fixed-point equation by $d_\mu(s')$ gives

$$e^P = \frac{d_\mu \odot i^P + \gamma P_\mu^\top e^P}{1+\gamma}. \tag{134}$$

Define the effective propagation factor $\lambda_P = \gamma/(1+\gamma)$, so $1-\lambda_P = 1/(1+\gamma)$. This is the factor induced by the P-label normalization, not an ETD eligibility-trace parameter. Solving with the identity matrix Id gives

$$e^P = (1-\lambda_P)(\mathrm{Id} - \lambda_P P_\mu^\top)^{-1}(d_\mu \odot i^P). \tag{135}$$

This reference starts from interest-weighted behavior states and propagates under μ. In this idealization, $R^P = e^P/d_\mu$ describes the corresponding change in state importance. Unit interest has the neutral fixed point $R^P \equiv 1$, irrespective of behavior–target policy disagreement.

Both mass learners use $\gamma = 0.995$ and 1000 Adam updates with minibatches of 256 uniformly sampled transition occurrences, with replacement. Each role has fresh optimizer state and the configured emphatic learning rate; the Adam coefficients are those specified in Section B.2.7. All labels are detached during differentiation. The J labels evolve through its private slow network, while the P labels remain fixed throughout its fit.

B.3.7 Deriving consumer weights from the masses.

Recall the frozen parent raw masses M_g^χ for $\chi \in \{J, P\}$ and the current-observation empirical measure D_g. Let $\overline{m}_g^\chi$ be the full-data mean mass, R_g^χ the normalized mass multiplier (agreeing with Equation (129) for P), and $v_{M,g}^\chi$ its mean squared deviation from one:

$$\overline{m}_g^\chi = \mathbb{E}_{D_g}[M_g^\chi], \qquad R_g^\chi(s) = \frac{M_g^\chi(s)}{\overline{m}_g^\chi}, \qquad v_{M,g}^\chi = \mathbb{E}_{D_g}[(R_g^\chi - 1)^2]. \tag{136}$$

Since $\mathbb{E}_{D_g}[R_g^\chi] = 1$, its second moment equals $1 + v_{M,g}^\chi$. Define the shrinkage coefficient α_g^χ and the consumer weight function w_g^χ by

$$\alpha_g^\chi = (1 + v_{M,g}^\chi)^{-1/2}, \qquad w_g^\chi(s) = 1 + \alpha_g^\chi(R_g^\chi(s) - 1). \tag{137}$$

Substitution gives the main-text expression (Equation (12)):

$$w_g^\chi(s) = 1 + \frac{M_g^\chi(s) - \mathbb{E}_{D_g}[M_g^\chi]}{\sqrt{\mathbb{E}_{D_g}[(M_g^\chi)^2]}}, \qquad \chi \in \{J, P\}. \tag{138}$$

The moments are computed once over all current-state occurrences and remain fixed during subsequent minibatch computations.

For finite positive masses, $0 < \alpha_g^\chi \le 1$, so

$$w_g^\chi = (1 - \alpha_g^\chi) + \alpha_g^\chi R_g^\chi > 0, \qquad \mathbb{E}_{D_g}[w_g^\chi] = 1. \tag{139}$$

Multiplying M_g^χ by a positive constant leaves R_g^χ and w_g^χ unchanged. Moreover,

$$\mathbb{E}_{D_g}[(w_g^\chi - 1)^2] = \frac{v_{M,g}^\chi}{1 + v_{M,g}^\chi} < 1. \tag{140}$$

Let $w_i^\chi = w_g^\chi(s_i)$ and let $w^\chi = (w_1^\chi, \dots, w_n^\chi)$ be the vector of occurrence weights. Its algebraic effective sample size, denoted by $\mathrm{ESS}(w^\chi)$, is

$$\mathrm{ESS}(w^\chi) = \frac{(\sum_i w_i^\chi)^2}{\sum_i (w_i^\chi)^2} = n\frac{1 + v_{M,g}^\chi}{1 + 2v_{M,g}^\chi} > \frac{n}{2}. \tag{141}$$

These identities establish positivity, normalization, and controlled dispersion of the consumer weights. Scale invariance here concerns the consumer map; the J fitting rule is not scale invariant because its additive source term $1 - \gamma$ remains fixed when raw mass is rescaled.

B.3.8 Induced measures and their use in updating the critic.

Recall that $w_i^\chi = w_g^\chi(s_i)$ is the frozen weight of current occurrence i. Let ν_g^χ be the probability measure induced by these weights on the observed roots. Together with the unweighted empirical measure D_g, it is

$$D_g = \frac{1}{n}\sum_{i=1}^n \delta_{s_i}, \qquad \nu_g^\chi = \frac{1}{n}\sum_{i=1}^n w_i^\chi \delta_{s_i}, \quad \chi \in \{J, P\}. \tag{142}$$

Sampling from ν_g^χ means selecting occurrence i with probability w_i^χ/n. For any real-valued function f that is finite on the observed roots,

$$\mathbb{E}_{S\sim\nu_g^\chi}[f(S)] = \frac{1}{n}\sum_{i=1}^{n} w_i^\chi f(s_i). \tag{143}$$

Recall the normalized mass multiplier R_g^χ and shrinkage coefficient α_g^χ in Equations (136) and (137). Define the normalized-mass probability measure $\overline{\nu}_g^\chi = n^{-1}\sum_i R_g^\chi(s_i)\delta_{s_i}$. The consumer measure then satisfies

$$\nu_g^\chi = (1-\alpha_g^\chi)D_g + \alpha_g^\chi\overline{\nu}_g^\chi. \tag{144}$$

The induced measure is an explicit mixture of the empirical behavior measure and the normalized-mass measure.

Emphasis enters the successor-critic update through the distribution of its training states. Recall from Section B.2.5 that a root-policy pair is chosen uniformly from (D_g, μ_g), (ν_g^J, π_g), and (ν_g^P, μ_g). The augmented root measures add an independent root risk in RB (Equation (75)). Recall that $\widetilde{\eta}_g$ is the augmented state law after the first consumer transition, and $\widetilde{\zeta}_g$ is the law after an additional geometric number of target-policy transitions (Equations (77) and (78)). The resulting training bank contains $n_\zeta = 25{,}000$ pairs (S_ℓ, U_ℓ), indexed by $\ell = 1, \ldots, n_\zeta$, where S_ℓ is the retained observation and U_ℓ is the additional next-availability draw in Equation (80). Recall that V_θ is a live scalar critic with parameter vector θ, and that $\mathcal{T}_g^u$ denotes its target-policy Bellman backup conditional on next availability u. The prediction-minus-backup fitting error is $\varepsilon_g(\theta; s, u) = V_\theta(s) - (\mathcal{T}_g^u V_\theta)(s)$, as defined in Equation (84). The empirical fitting objective $\widehat{\mathcal{J}}_{V,g}(\theta)$ is

$$\widehat{\mathcal{J}}_{V,g}(\theta) = \frac{1}{n_\zeta}\sum_{\ell=1}^{n_\zeta}\varepsilon_g(\theta; S_\ell, U_\ell)^2, \tag{145}$$

which is Equation (82). Running the prescribed finite optimization produces each successor member $V_{g+1,j}$, for $j = 1, \ldots, 5$, and their output mean v_{g+1} (Equation (88)). Thus emphatic weights affect where successor critics are fitted. Their effect is incorporated through root selection and continuation sampling; the training-bank loss is not multiplied by w_i^χ a second time.

The same frozen weights enter the target-return and transmission update channels. The induced measures ν_g^J and ν_g^P also enter the target actor's shared Fisher geometry and its two channelwise endpoint constraints (Section 4.2). Recall from Equation (184) that $\varpi_i = (1 + w_i^J + w_i^P)/(3n)$ is the occurrence probability for their shared residual-root proposal; the corresponding importance factors are given in Equation (186). Finally, ν_g^J selects the heuristic anchors used above, linking current target-return emphasis to future confidence-based transmission emphasis.

B.3.9 Neutral mass and weight initialization.

Both initialization routes start from neutral masses, $M_0^J = M_0^P = 1$, so $w_0^J = w_0^P = 1$ and $\nu_0^J = \nu_0^P = D_0$.

B.3.10 Cross-generation evolution and two-generation transmission.

During generation g, parent masses are evaluated on the newly collected data $\mathcal{D}_g$. Their consumer weights remain frozen while successor actors, critics, and masses are constructed; actor differentiation treats the weights as constants. The confidence-to-transmission-mass pathway is

$$q_{H,g} \longrightarrow (c_{g,f})_{f=1}^{16} \longrightarrow i_g^P \longrightarrow M_{g+1}^P \longrightarrow (w_{g+1}^P, \nu_{g+1}^P) \text{ on } \mathcal{D}_{g+1} \longrightarrow \pi_{g+2}. \tag{146}$$

These arrows indicate computational dependence, including the paired comparison with π_g used to obtain $c_{g,f}$. Interest i_g^P helps fit M_{g+1}^P on $\mathcal{D}_g$, while the update producing π_{g+1} uses M_g^P. After the joint successor commit, M_{g+1}^P is evaluated and normalized on $\mathcal{D}_{g+1}$; its weights can first affect the target update producing π_{g+2}.

Weights change both because mass functions are learned and because the behavior data used to compute their moments change. Even an unchanged mass can induce different weights on a new data set. Generation-g training labels and normalization constants are therefore not reused as generation-$(g+1)$ consumer weights. The J mass follows the same parent-consumption/successor-fitting schedule, although its fitting labels evolve within a generation through the private slow network.

B.3.11 DESIGN INTERPRETATION AND SCOPE OF THE CLAIMS.

The two roles allocate learning effort to different state references: target-policy continuation from behavior-distributed states for J, and confidence-weighted behavior continuation for P. In the implemented algorithm, these references motivate the labels rather than identify the fitted networks with exact state-distribution ratios. The resulting weights enter the target-update objectives, shared Fisher geometry, and endpoint constraints; they also determine critic-training root selection and select heuristic-confidence anchors. Equations (139)–(141) prove positivity, unit mean, and bounded occurrence-weight concentration. Equation (146) specifies the earliest target update reachable through newly learned confidence-based mass.

The stationary references in Equations (122) and (135) are population fixed points under Γ_μ in Equation (118). They require fixed policies, sufficient action support, a sufficient Markov state, a positive stationary behavior distribution, and unrestricted exact regression; the P reference also requires fixed normalized interest. Equality of two empirical state marginals alone does not imply these population equations. Finite trajectory data, anchored neural approximation, finite optimization, and evolving generations need not realize these references. The empirical conditional regression identities remain valid on the observed successor support for their declared frozen labels, without implying that the fitted networks attain those targets.

These results establish controlled reweighting and the stated computational dependencies. They do not by themselves establish exact occupancy correction, monotone reduction of visitation mismatch, or improved return in every generation. The ESS bound concerns algebraic weight concentration and does not remove temporal dependence in the data. The confidence score summarizes standardized simulated improvement under the frozen fitted kernel in Equation (40); its calibration and transfer to actual returns require separate evidence. ETD's linear expected-update stability result and ACE's exact-gradient theorem do not automatically extend to HBPO's fitted, shrunken, and delayed weights.

B.4 HEURISTIC GUIDANCE: CONSTRUCTION, ESTIMATION, AND BEHAVIOR FITTING

B.4.1 ROLE OF HEURISTIC GUIDANCE AND FROZEN INPUTS

At generation g, heuristic guidance converts long-horizon action comparisons under the frozen simulation model $\widehat{\mathcal{P}}_g$ into a temporary stochastic controller $q_{H,g}$. The parent target policy π_g supplies both the continuation policy used to evaluate actions and the reference distribution used to regularize the controller. Recall the five frozen scalar critic members $V_{g,j}$, $j \in \{1,\dots,5\}$, and their output mean $v_g(s) = \frac{1}{5}\sum_{j=1}^{5} V_{g,j}(s)$, which supplies a control variate. The emphatic probability measure ν_g^J, recalled below, selects the starting observations. The model, parent policies, reference critic, and emphatic weights remain fixed while the controller is constructed.

Heuristic guidance acts on the behavior policy μ, which collects the next generation of interaction data. We therefore define a behavior-fitting objective H_g that attracts μ toward $q_{H,g}$. The target policy π_g determines the controller's preferences; the parameters differentiated in H_g are those of θ^μ.

Throughout this subsection, $T = 1000$ is the heuristic evaluation horizon, denoted by H in the notation table's heuristic-utility entries. In DSIS-RB, abbreviated RB below, each simulated continuation retains one synthetic supply-risk draw. DSIS uses the corresponding construction without a latent-risk coordinate.

The resulting controller provides action-distribution targets for the behavior policy. It also supplies evidence for the finite-intervention confidence used to fit the next generation's transmission emphatic mass M_{g+1}^P. Here, *paired returns* compare two T-step simulations starting from the same anchor observation (a starting state selected as described below) and sharing their exogenous and action-sampling random inputs, including the root risk in RB: one follows $q_{H,g}$ at stages $0,\dots,8$ and π_g thereafter, whereas the other follows π_g throughout. Their return difference measures the simulated gain from this finite intervention. The confidence and interest construction is specified in Appendix B.3.5. Here we develop the action preferences and the behavior-guidance gradient. The controller has no separately trained actor parameters.

B.4.2 SELECTING ANCHOR STATES

Let n be the number of collected observation occurrences and s_i the observation at occurrence $i \in \{1, \dots, n\}$. Write δ_x for unit point mass at observation x. Recall the empirical measure $D_g = n^{-1} \sum_{i=1}^{n} \delta_{s_i}$ and the positive consumer weights w_i^J for the target-return role, normalized so that $n^{-1} \sum_i w_i^J = 1$. The heuristic uses the already defined measure

$$\nu_g^J = \frac{1}{n} \sum_{i=1}^{n} w_i^J \delta_{s_i}. \tag{147}$$

At each generation, we sample 16 observation occurrences independently with replacement. Let $f \in \{1, \dots, 16\}$ index the selections, $I_{g,f} \in \{1, \dots, n\}$ be the selected occurrence index, and $\bar{s}_{g,f}$ its anchor observation:

$$\Pr(I_{g,f} = i) = \frac{w_i^J}{n}, \qquad \bar{s}_{g,f} = s_{I_{g,f}}, \qquad f = 1, \dots, 16. \tag{148}$$

Anchors may occur anywhere in the collected trajectories; they are not restricted to episode starts. Repeated observations and repeated selections retain their multiplicity. These are the same anchor occurrences used by the confidence construction.

Every controller query, whether at an anchor or at a subsequently simulated state, starts a new local evaluation clock at $k = 0$. Its utility horizon is always $T = 1000$, independently of the queried observation's position in its originating trajectory.

B.4.3 THE RETURN, CENTERED UTILITY, AND SIMULATION TAPES

We first expand the expectation $\mathbb{E}_H$ used in Section 4.1. For a fixed synthetic risk $z \in [0, 1/2]$, recall that $\widehat{\mathcal{P}}_{g,z}$ is the frozen joint reward–observation kernel conditional on that risk. Let S_k, A_k, and R_k denote the observation, action, and normalized reward at local stage k, and let $\mathbb{E}_z^{s,a}$ denote expectation under the rollout law

$$\begin{gathered} S_0 = s, \qquad \mathsf{A}_0 = a, \\ (R_k, S_{k+1}) \sim \widehat{\mathcal{P}}_{g,z}(\cdot \mid S_k, \mathsf{A}_k), \qquad 0 \le k < T, \\ \mathsf{A}_k \sim \pi_g(\cdot \mid S_k), \qquad 1 \le k < T. \end{gathered} \tag{149}$$

Let $m = \mathrm{Unif}[0, 1/2]$ be the synthetic-risk prior. For any integrable rollout functional $\mathcal{F}$, the RB expectation is

$$\mathbb{E}_H[\mathcal{F} \mid s, a] = \int_{[0,1/2]} \mathbb{E}_z^{s,a}[\mathcal{F}]\, m(dz). \tag{150}$$

The root risk is drawn independently of the query observation and retained for the entire continuation; it is not inferred from that observation or redrawn at each stage. The kernel and its risk convention are specified in Appendix B.1.3. For DSIS, omit z and the integral over m, and use its frozen simulation kernel. The finite return in the main text is therefore

$$G_{T,g}(s, a) = \mathbb{E}_H \left[\sum_{k=0}^{T-1} R_k \,\middle|\, s, a \right]. \tag{151}$$

For RB, this includes the risk integral in Equation (150); for DSIS it uses the risk-free rollout law. Write c_k for the period cost on the simulated trajectory. Rewards retain the task's normalization $R_k = -c_k/495$, and the return is undiscounted.

Recall that $\mathcal{A} = \{0, \dots, 8\}^2$ is the set of 81 regular/expedited order pairs. Let $\widehat{G}_{T,g}(s, a)$ denote the sample estimate of this return constructed below. The centered estimated utility, denoted by u_g in the main text and h_g in the notation table's heuristic-utility entries, is

$$u_g(s, a) = h_g(s, a) = \widehat{G}_{T,g}(s, a) - \sum_{a' \in \mathcal{A}} \pi_g(a' \mid s) \widehat{G}_{T,g}(s, a'). \tag{152}$$

Thus $\sum_{a \in \mathcal{A}} \pi_g(a \mid s) u_g(s, a) = 0$. We next construct the coupled simulations and the critic-based estimator that supplies $\widehat{G}_{T,g}$.

A *simulation tape* records the random inputs used to generate a continuation. A *trajectory* records the resulting observations, actions, and rewards. A trajectory thus depends on the starting state, policy, and model as well as the tape. We use tapes so that alternative first actions can be compared using common random inputs, rather than generating an independent collection of inputs for every alternative. This couples the simulation noise in the comparisons and permits identical continuation work to be shared.

Under the ideal sampling law, each fixed query observation uses 32 independent tapes, indexed by $b \in \{1, \dots, 32\}$. In RB, let $Z_b \sim m$ be tape b's root risk, and let $U^D_{b,k}, U^B_{b,k} \sim \mathrm{Unif}(0,1)$ be its demand and next-availability inputs. For each action $a' \in \mathcal{A}$, let $\epsilon_{b,k,a'}$ be a standard Gumbel variable, with cumulative distribution function $x \mapsto \exp(-e^{-x})$, and write $\boldsymbol{\epsilon}_{b,k} = (\epsilon_{b,k,a'})_{a' \in \mathcal{A}}$. These primitive inputs are mutually independent across tapes, stages, action coordinates, and input types, including independence from the risks. A utility-evaluation tape $\mathfrak{t}_b$ is

$$\mathfrak{t}_b = \left(Z_b, \{U^D_{b,k}, U^B_{b,k}, \boldsymbol{\epsilon}_{b,k}\}_{k=0}^{T-1}\right). \tag{153}$$

The action noise at $k = 0$ is unused because the first action is forced; the final action and transition need not be sampled when evaluating the control estimator below. DSIS omits the risk and stochastic-availability inputs. At each distinct query state, we enumerate all 81 first actions in $\mathcal{A}$ and evaluate them using these 32 tapes in total. Each tape supports 81 coupled trajectories, one for each forced first action. Write $S^a_{b,k}$ and $\mathsf{A}^a_{b,k}$ for the stage-k observation and action on tape b's branch whose forced first action is a. Thus $S^a_{b,0} = s$ and $\mathsf{A}^a_{b,0} = a$. The finite-logit categorical target policy has $\pi_g(a' \mid s') > 0$ for every action and queried observation. For $1 \le k < T$, branch a samples its action as

$$\mathsf{A}^a_{b,k} = \operatorname*{arg\,max}_{a' \in \mathcal{A}} \left\{\log \pi_g(a' \mid S^a_{b,k}) + \epsilon_{b,k,a'}\right\}. \tag{154}$$

Demand and next availability are obtained by applying the model's inverse cumulative distribution functions to $U^D_{b,k}$ and $U^B_{b,k}$, respectively, using the branch's current observation and, in RB, its fixed risk Z_b. The branches share the tape's risk and random inputs, but their realized actions and successors can differ because their conditional distributions can differ. Sharing a tape preserves the prescribed marginal rollout law for every first action.

B.4.4 Estimating the return with the frozen critic

Let R and S' denote a one-step reward and successor observation under $\widehat{\mathcal{P}}_{g,z}$. Define $Q^{v,H}_{g,z,k}$ as the undiscounted control backup at local stage k, where the superscripts indicate the frozen reference v_g and the heuristic evaluation:

$$Q^{v,H}_{g,z,k}(s,a) = \begin{cases} \mathbb{E}_{\widehat{\mathcal{P}}_{g,z}}[R + v_g(S') \mid s, a], & 0 \le k \le T-2, \\ \mathbb{E}_{\widehat{\mathcal{P}}_{g,z}}[R \mid s, a], & k = T-1. \end{cases} \tag{155}$$

This is exactly the evaluated backup $L^{\mathrm{H}}_{g,z,k}$ listed in the notation table, with $T = H$: it uses v_g at nonterminal successor stages and zero control at the horizon. This quantity is distinct from the completed, discounted action value $Q^{\pi_g}_g$ in Equation (11). The one-step expectations integrate the model's demand support and, in RB, both next-availability outcomes.

For branch a on tape b, let $\widetilde{G}^{(b)}_{T,g,Z_b}(s,a)$ denote the complete one-tape control-variate estimator, defined by

$$\widetilde{G}^{(b)}_{T,g,Z_b}(s,a) = Q^{v,H}_{g,Z_b,0}(s,a) + \sum_{k=1}^{T-1} \left[\sum_{a' \in \mathcal{A}} \pi_g(a' \mid S^a_{b,k}) Q^{v,H}_{g,Z_b,k}(S^a_{b,k}, a') - v_g(S^a_{b,k})\right]. \tag{156}$$

The first term evaluates the forced first action. Each later term is the target-policy average of the one-step backup, minus the reference value at the visited observation. The sampled continuation determines where these control residuals are evaluated.

The same conditional-expectation and telescoping argument used in the proof of Theorem 1 yields the finite-horizon identity below. The theorem itself concerns an infinite discounted return; here the return is undiscounted and the terminal control contribution is zero. Fix s, a, z, and write

$\widetilde{G}_{T,g,z}(s,a)$ for one generic complete tape estimator conditional on risk z, suppressing the tape and branch indices on both the estimator and the trajectory. The tower property gives

$$\begin{aligned}\mathbb{E}_z^{s,a}[\widetilde{G}_{T,g,z}(s,a)] &= \mathbb{E}_z^{s,a}\left[\begin{array}{l} R_0 + v_g(S_1) \\ \quad + \sum_{k=1}^{T-2}\{R_k + v_g(S_{k+1}) - v_g(S_k)\} + R_{T-1} - v_g(S_{T-1})\end{array}\right] \\ &= \mathbb{E}_z^{s,a}\left[\sum_{k=0}^{T-1} R_k\right].\end{aligned} \tag{157}$$

All reference values cancel. For each fixed starting observation, bounded actions and finite demand support give finitely many reachable observations over this finite horizon. Period costs and the finite-parameter frozen networks are finite on that set, uniformly over $z \in [0, 1/2]$. Integrating this identity against $m(dz)$ therefore recovers $G_{T,g}(s,a)$.

The critic need not equal the finite-horizon value function: its role is to supply terms whose expectations cancel. This is an identity in expectation, so the estimate need not equal the realized reward sum on an individual tape. Rewards already enter through $Q^{v,H}_{g,z,k}$; an additional sampled reward sum would count them again.

There are two changes from the discounted critic backup. Recall that $\gamma = 0.995$ is the critic discount factor. Before the last period, the successor value has coefficient 1 instead of γ, because every reward in this return receives equal weight. At the last period, no successor value is included, because rewards after period $T-1$ are outside the objective. Recall the discounted backup $L_{g,z}v_g$ and the state-transition operator $\widehat{P}^a_{g,z}$, evaluated on v_g:

$$\begin{aligned}L_{g,z}v_g(s,a) &= \mathbb{E}_{\widehat{\mathcal{P}}_{g,z}}[R + \gamma v_g(S') \mid s,a], \\ (\widehat{P}^a_{g,z}v_g)(s) &= \mathbb{E}_{\widehat{\mathcal{P}}_{g,z}}[v_g(S') \mid s,a].\end{aligned} \tag{158}$$

Writing $\mathbf{1}_{\{k<T-1\}}$ for the indicator of $k < T-1$, the required adjustment is

$$Q^{v,H}_{g,z,k}(s,a) = L_{g,z}v_g(s,a) + \big(\mathbf{1}_{\{k<T-1\}} - \gamma\big)(\widehat{P}^a_{g,z}v_g)(s). \tag{159}$$

Keeping the discounted backup unchanged would break the cancellation. Keeping a terminal value would add $\mathbb{E}_z^{s,a}[v_g(S_T)]$ to the expected estimate.

B.4.5 Averaging and computing action differences

Average the 32 complete tape estimates to obtain the return estimate used in Equation (152):

$$\widehat{G}_{T,g}(s,a) = \frac{1}{32}\sum_{b=1}^{32} \widetilde{G}^{(b)}_{T,g,Z_b}(s,a). \tag{160}$$

Recall that $u_g = h_g$ is this estimate centered under $\pi_g(\cdot \mid s)$, rather than under a uniform action law. For a fixed query observation independent of its utility tapes, hold the model, policy, and reference critic fixed. In the next display, $\mathbb{E}$ averages only over the 32 utility tapes, including their risks in RB. The preceding expectation identity gives

$$\mathbb{E}[\widehat{G}_{T,g}(s,a)] = G_{T,g}(s,a), \quad \mathbb{E}[u_g(s,a)] = G_{T,g}(s,a) - \sum_{a'} \pi_g(a' \mid s) G_{T,g}(s,a'). \tag{161}$$

These statements concern the complete fixed-state return and utility estimators in exact arithmetic. They do not assert conditional unbiasedness after selecting a state using the same cached utility estimates. Nonlinear transformations of their estimates, including the controller below, do not inherit unbiasedness automatically. Utilities remain in cumulative normalized-reward units, without division by T.

Only action differences are needed to construct u_g. For any fixed reference action $a_\star \in \mathcal{A}$,

$$u_g(s,a) = [\widehat{G}_{T,g}(s,a) - \widehat{G}_{T,g}(s,a_\star)] - \sum_{a'} \pi_g(a' \mid s)[\widehat{G}_{T,g}(s,a') - \widehat{G}_{T,g}(s,a_\star)]. \tag{162}$$

The implementation either accumulates tape-level differences against the reference order pair $a_\star = (0, 0)$, or accumulates coupled branch estimates and centers them afterwards. On the same tape, branches reaching the same observation at the same stage have identical future contributions under their shared inputs. Their subsequent work can be shared, and a tail common to all branches can be omitted. These operations preserve the centered utilities in exact arithmetic; an accumulator with an omitted common tail need not represent an absolute return.

The identities above use exact arithmetic and the same evaluated frozen policy and control function throughout. On RB graphics processors, the heuristic service evaluates bfloat16 copies of the frozen actor and critics, with 32-bit floating-point accumulation. Discounted residual evaluation uses 32-bit critic inference and 64-bit residual algebra. Equation (159) therefore compares backups of the same mathematical functions; it does not assert bitwise equality between these numerical services.

B.4.6 THE REGULARIZATION COEFFICIENT

The coefficient $\tau > 0$ controls how strongly the controller is penalized for departing from π_g: larger values give more weight to staying near the target prior. The shared hyperparameter-selection protocol tunes a positive, dimensionless multiplier κ, distinct from the fitted supply-recovery parameter $\widehat{\kappa}_g^{\mathrm{rec}}$. At initialization, let $\bar{s}_{0,f}^{\mathrm{cal}}$, $f = 1, \ldots, 16$, denote calibration anchors sampled with replacement from the initial emphatic measure ν_0^J of Equation (147), using a separate random-number domain. Let u_0^{cal} be their centered utility estimates, computed with the same 81 forced first actions, 32 tapes, and $T = 1000$ horizon. Define the calibration scale S_{cal} and set

$$\begin{aligned} S_{\mathrm{cal}} &= \mathrm{median}_{f=1,\ldots,16} \left[\max_{a \in \mathcal{A}} u_0^{\mathrm{cal}}(\bar{s}_{0,f}^{\mathrm{cal}}, a) - \min_{a \in \mathcal{A}} u_0^{\mathrm{cal}}(\bar{s}_{0,f}^{\mathrm{cal}}, a) \right], \\ \tau &= \kappa S_{\mathrm{cal}}. \end{aligned} \tag{163}$$

For these 16 ranges, the median is the mean of the eighth and ninth ordered values. The construction requires $0 < S_{\mathrm{cal}} < \infty$, hence $0 < \tau < \infty$. The implementation rejects a zero or nonfinite calibration scale; no positive floor is substituted. Thus S_{cal} supplies a one-time scale conversion; it is not an additional tuned parameter. The resulting τ stays fixed throughout the run, with no separate normalization of later states' utilities.

B.4.7 THE HELLINGER-REGULARIZED CONTROLLER

Recall that u_g is the centered utility in Equation (152), and that the positive scale $\tau = \kappa S_{\mathrm{cal}}$ is fixed by Equation (163). At a fixed query observation, write $p = (p_a)_{a \in \mathcal{A}}$ for the target prior with $p_a = \pi_g(a \mid s) > 0$, and $u_a = u_g(s, a)$. Let $q = (q_a)_{a \in \mathcal{A}}$ be a candidate distribution for the controller's action at s, where q_a is the probability assigned to action a. The optimization ranges over the entire simplex $\Delta(\mathcal{A}) = \{q : q_a \geq 0,\ \sum_{a \in \mathcal{A}} q_a = 1\}$, the set of all probability distributions on the 81 actions. The prior p supplies the regularization reference; no initial candidate q is required, because the maximizing distribution is computed by the scalar root solve derived below. The controller is defined by

$$q_{H,g}(\cdot \mid s) = \operatorname*{arg\,max}_{q \in \Delta(\mathcal{A})} \left\{ \sum_a q_a u_a - 2\tau \sum_a (\sqrt{q_a} - \sqrt{p_a})^2 \right\}. \tag{164}$$

The displayed penalty uses the unnormalized squared Hellinger discrepancy $\sum_a (\sqrt{q_a} - \sqrt{p_a})^2$; the convention with a factor $1/2$ gives half this value.

Since p and q have unit mass, denote the same objective by $\Phi(q)$ and expand it as

$$\Phi(q) = \sum_a q_a u_a + 4\tau \sum_a \sqrt{p_a q_a} - 4\tau. \tag{165}$$

At an interior point $q_a > 0$ for all a, its Hessian entries are

$$\frac{\partial^2 \Phi}{\partial q_a^2} = -\tau \sqrt{p_a}\, q_a^{-3/2} < 0, \qquad \frac{\partial^2 \Phi}{\partial q_a \partial q_{a'}} = 0 \quad (a \neq a'). \tag{166}$$

The objective is strictly concave. Its maximizer has full support: if $q_a = 0$, transfer a small mass $\varepsilon > 0$ from any positive coordinate to action a. Because $p_a > 0$, the gain in the square-root term

is of order $\sqrt{\varepsilon}$, whereas the utility change and the loss at the donating coordinate are of order ε. Compactness gives existence, and strict concavity gives uniqueness.

Introducing a multiplier $\lambda_g(s)$ for the unit-mass constraint yields

$$u_a + 2\tau\sqrt{\frac{p_a}{q_a}} - \lambda_g(s) = 0. \tag{167}$$

Therefore,

$$q_{H,g}(a \mid s) = \pi_g(a \mid s)\left(\frac{2\tau}{\lambda_g(s) - u_g(s,a)}\right)^2, \qquad \lambda_g(s) > \max_a u_g(s,a), \tag{168}$$

where $\lambda_g(s)$ is determined by

$$\sum_a \pi_g(a \mid s)\left(\frac{2\tau}{\lambda_g(s) - u_g(s,a)}\right)^2 = 1. \tag{169}$$

The left-hand side is continuous and strictly decreasing in $\lambda_g(s)$, tends to infinity at the lower boundary, and tends to zero as $\lambda_g(s) \to \infty$. Hence a unique scalar root determines the entire action distribution.

For a bounded numerical bracket, let $u_{\max}$ be the largest utility, $p_{\max}$ the total prior mass of all maximizing actions (not the largest prior probability), and δ the positive multiplier gap:

$$u_{\max} = \max_a u_a, \qquad p_{\max} = \sum_{a:u_a=u_{\max}} p_a, \qquad \delta = \lambda_g(s) - u_{\max}. \tag{170}$$

At the normalization root, the maximizing actions contribute $(2\tau/\delta)^2 p_{\max}$, and every denominator is at least δ. Hence $(2\tau/\delta)^2 p_{\max} \le 1 \le (2\tau/\delta)^2$, giving

$$2\tau\sqrt{p_{\max}} \le \delta \le 2\tau. \tag{171}$$

The implementation uses 64 bisection iterations in $\log\delta$, retaining log prior probabilities. This solves for one scalar normalizer per query; the resulting probabilities satisfy the controller formula up to floating-point and root-solve error, without adding a probability floor.

Since $q = p$ is feasible and $\sum_a p_a u_a = 0$, optimality of the exact controller gives

$$\sum_{a\in\mathcal{A}} q_{H,g}(a \mid s) u_a \ge 2\tau \sum_{a\in\mathcal{A}} \left(\sqrt{q_{H,g}(a \mid s)} - \sqrt{p_a}\right)^2 \ge 0. \tag{172}$$

This is a property of the estimated first-action replacement objective under the frozen model. It neither guarantees improvement in the exact model return after Monte Carlo action selection nor establishes improvement from repeatedly executing the controller or in the real environment.

B.4.8 Selecting fitting locations through short controller rollouts

Recall the 16 anchor occurrences $\bar{s}_{g,f}$ from Equation (148). From each, generate 8 outer model trajectories under the controller. The superscript O labels these outer trajectories: let $S^{\mathrm{O}}_{g,f,b,k}$ and $\mathsf{A}^{\mathrm{O}}_{g,f,b,k}$ denote their stage-k observations and actions, with $b \in \{1,\dots,8\}$ indexing outer paths and $S^{\mathrm{O}}_{g,f,b,0} = \bar{s}_{g,f}$. In RB, each path draws a risk $Z^{\mathrm{O}}_{g,f,b}$ independently from $m = \mathrm{Unif}[0, 1/2]$, the prior recalled in Equation (150), and retains it throughout. At stages $k = 0,\dots,8$, sample

$$\mathsf{A}^{\mathrm{O}}_{g,f,b,k} \sim q_{H,g}(\cdot \mid S^{\mathrm{O}}_{g,f,b,k}) \tag{173}$$

and then sample the reward and successor from $\widehat{\mathcal{P}}_{g,Z^{\mathrm{O}}_{g,f,b}}(\cdot \mid S^{\mathrm{O}}_{g,f,b,k}, \mathsf{A}^{\mathrm{O}}_{g,f,b,k})$ in RB, or from the risk-free model in DSIS. Let ξ_g be the empirical fitting measure of the pre-action observations at these nine stages. Recalling that δ_x denotes unit point mass at x, set

$$\xi_g = \frac{1}{16\cdot 8\cdot 9}\sum_{f=1}^{16}\sum_{b=1}^{8}\sum_{k=0}^{8}\delta_{S^{\mathrm{O}}_{g,f,b,k}}. \tag{174}$$

This panel contains exactly 1152 observation occurrences, including repetitions. Each occurrence receives equal fitting weight; the emphatic weights have already entered through anchor selection.

At every retained observation, the fitting target is the full distribution $q_{H,g}(\cdot \mid s)$, obtained from a utility query with a new local horizon of 1000 periods. The nine-stage rollout selects fitting locations without reducing the utility horizon at later stages.

Utility estimation and outer trajectory generation use separate random-number domains. In RB, an inner controller query samples its own utility-tape risks and does not read the outer trajectory's latent risk, so the controller remains a function of the visible observation. Repeated queries at the same observation reuse the generation's cached estimate. Separate outer paths use separately addressed random inputs, including when their anchors coincide. Thus 1152 counts fitting occurrences, not necessarily distinct utility queries; observations along a path are also temporally dependent.

B.4.9 Deriving the behavior-guidance gradient

Recall that ξ_g is the equal-weight empirical measure of the 1152 pre-action observations $S^{\mathrm{O}}_{g,f,b,k}$ in Equation (174); O denotes the outer controller trajectories. Let θ^μ parameterize the candidate behavior policy μ_{θ^μ}, and recall the parent policy $\mu_g = \mu_{\theta^\mu_g}$ at parameters θ^μ_g. Assume that the candidate probabilities are strictly positive and differentiable in a neighborhood of those parent parameters, as for the finite-logit categorical actor. For two action distributions $p = (p_a)$ and $q = (q_a)$, define the categorical Fisher–Rao distance d_{FR} by

$$d_{\mathrm{FR}}(p,q) = 2\arccos\left(\sum_{a\in\mathcal{A}} \sqrt{p_a q_a}\right). \tag{175}$$

With ξ_g and all controller targets fixed, define the behavior-fitting objective

$$H_g(\mu_{\theta^\mu}) = -\frac{1}{2}\,\mathbb{E}_{s\sim\xi_g}\left[d^2_{\mathrm{FR}}\big(\mu_{\theta^\mu}(\cdot \mid s), q_{H,g}(\cdot \mid s)\big)\right]. \tag{176}$$

Maximizing H_g reduces the distance between the behavior policy and the controller at the retained observations. Define its behavior-guidance gradient $g_{H,g}$ at the parent behavior parameters by

$$\begin{aligned} g_{H,g} &= \nabla_{\theta^\mu} H_g(\mu_{\theta^\mu})|_{\theta^\mu=\theta^\mu_g} \\ &= -2\,\nabla_{\theta^\mu}\mathbb{E}_{s\sim\xi_g}\left[\arccos^2\left(\sum_a \sqrt{\mu_{\theta^\mu}(a \mid s) q_{H,g}(a \mid s)}\right)\right]\Bigg|_{\theta^\mu=\theta^\mu_g}, \end{aligned} \tag{177}$$

recovering Equation (2).

For an explicit expression, define the affinity $\mathcal{B}_g(s)$, angle $\varphi_g(s)$, and behavior-policy score vector $\mathrm{sc}^\mu_g(s,a)$ at the parent parameters:

$$\begin{aligned} \mathcal{B}_g(s) &= \sum_a \sqrt{\mu_g(a \mid s) q_{H,g}(a \mid s)}, \quad \varphi_g(s) = \arccos \mathcal{B}_g(s), \\ \mathrm{sc}^\mu_g(s,a) &= \nabla_{\theta^\mu} \log \mu_{\theta^\mu}(a \mid s)|_{\theta^\mu=\theta^\mu_g}. \end{aligned} \tag{178}$$

Let $\mathcal{B}_{\theta^\mu}(s)$ denote the same affinity with μ_g replaced by μ_{θ^μ}. Differentiation gives

$$\nabla_{\theta^\mu}\mathcal{B}_{\theta^\mu}(s)|_{\theta^\mu=\theta^\mu_g} = \frac{1}{2}\sum_a \sqrt{\mu_g(a \mid s) q_{H,g}(a \mid s)}\ \mathrm{sc}^\mu_g(s,a). \tag{179}$$

Apply the chain rule to $-2\arccos^2 \mathcal{B}_{\theta^\mu}(s)$ and use the score identity $\sum_a \mu_g(a \mid s)\,\mathrm{sc}^\mu_g(s,a) = 0$, which follows by differentiating $\sum_a \mu_{\theta^\mu}(a \mid s) = 1$. This gives

$$g_{H,g} = \mathbb{E}_{s\sim\xi_g}\left[\frac{2\varphi_g(s)}{\sin\varphi_g(s)}\sum_a\left(\sqrt{\mu_g(a \mid s) q_{H,g}(a \mid s)} - \mathcal{B}_g(s)\mu_g(a \mid s)\right)\mathrm{sc}^\mu_g(s,a)\right]. \tag{180}$$

The expectation is the arithmetic average over the 1152 retained occurrences, and the action sum includes all 81 actions. The subtracted term $\mathcal{B}_g(s)\mu_g(a \mid s)$ contributes zero after the score sum. It is

a tangent-space centering of the same Fisher–Rao gradient; there is no additional action-importance ratio or state-weight correction in this fitting objective.

At $\mu_g(\cdot \mid s) = q_{H,g}(\cdot \mid s)$, use $\lim_{\varphi\to 0} \varphi/\sin\varphi = 1$; the centered action coefficients are zero, so the contribution is zero. For stable evaluation, let $\mathbf{r}_g(s)$ be the residual between the square-root controller vector and its projection onto the square-root behavior vector. Square roots below act coordinatewise, $\|\cdot\|_2$ is the Euclidean norm, and $\mathrm{atan2}(y,x)$ is the two-argument angle function:

$$\begin{aligned} \mathbf{r}_g(s) &= \sqrt{q_{H,g}(\cdot \mid s)} - \mathcal{B}_g(s)\sqrt{\mu_g(\cdot \mid s)}, \\ \sin\varphi_g(s) &= \|\mathbf{r}_g(s)\|_2, \\ \varphi_g(s) &= \mathrm{atan2}\big(\|\mathbf{r}_g(s)\|_2, \mathcal{B}_g(s)\big). \end{aligned} \tag{181}$$

This evaluates the same gradient without directly differentiating an inverse cosine at coincident distributions.

During differentiation, the sampled observations, controller targets, utility estimates, and parent simulation components remain fixed. No gradient passes through action selection, model trajectories, the critic, or the controller's scalar root. Recall that $\tau = \kappa S_{\mathrm{cal}}$, with tuned multiplier κ and initialization scale S_{cal}, is fixed by Equation (163). It affects $g_{H,g}$ through the controller and its induced fitting locations, and adds no multiplicative factor to this gradient. The resulting $g_{H,g}$ enters the behavior update through the coupled-update rule defined in the main text.

B.5 Delayed Heuristic–Behavior–Target Coupling and Policy Updates

This subsection assembles the shared advantage evidence, the direct behavior and target objectives, the transmission objective, and the fused finite actor updates, in that order. All updates use the same frozen generation-g parents: behavior policy μ_g, target policy π_g, reference critic v_g, simulation model, and normalized consumer weights w_i^J, w_i^P defined below. The simulation model is the public kernel in DSIS and the fitted kernel in DSIS-RB. Once generated, residual samples, controller targets, and heuristic fitting locations are also held fixed during actor differentiation. In particular, the target successor uses μ_g, not the simultaneously constructed μ_{g+1}: guidance that changes behavior in generation g can first enter the discrepancy used to construct π_{g+2}. The corresponding delay through the transmission mass is specified in Appendix B.3.10.

The behavior actor combines the advantage objective A with the heuristic objective H_g; the target actor combines its advantage objective J with the transmission objective P. Recall that H_g is the negative half squared Fisher–Rao distance to the frozen heuristic controller, averaged over its fitting observations, and write $g_H = g_{H,g}$ for its parent gradient from Equations (176) and (180). We differentiate the candidate behavior policy in H_g, holding the controller and its fitting measure fixed. The results below concern these frozen policy-update objectives and their induced policy changes.

B.5.1 Constructing the shared advantage face

We first specify the exact, base, and sampled advantages used by all three advantage channels. Recall that the generation batch contains $n = 25{,}000$ observation occurrences s_i, indexed by $i = 1, \dots, n$, and that $D_g = n^{-1}\sum_i \delta_{s_i}$ is their empirical measure, with δ_{s_i} the unit point mass. The frozen weights satisfy $w_i^J, w_i^P > 0$ and $n^{-1}\sum_i w_i^J = n^{-1}\sum_i w_i^P = 1$; their weighted measures are $\nu_g^\chi = n^{-1}\sum_i w_i^\chi \delta_{s_i}$ for $\chi \in \{J, P\}$. All action sums below range over the finite action set $\mathcal{A}$. Write $\pi_i(a) = \pi_g(a \mid s_i)$ and $\mu_i(a) = \mu_g(a \mid s_i)$. For action-indexed arrays x, y, use $\langle x, y\rangle = \sum_{a\in\mathcal{A}} x(a)y(a)$, componentwise if one array is vector-valued. Recall the exact discounted model action value $Q_g^{\pi_g}$ and its one-step base Q_g^0 from Equations (53) and (51). Their centered advantages are

$$\begin{aligned} \mathbf{a}_i(a) &= A_g^{\mathrm{adv}}(s_i, a) = Q_g^{\pi_g}(s_i, a) - \sum_b \pi_i(b) Q_g^{\pi_g}(s_i, b), \\ \mathbf{a}_i^0(a) &= A_g^{\mathrm{base}}(s_i, a) = Q_g^0(s_i, a) - \sum_b \pi_i(b) Q_g^0(s_i, b). \end{aligned} \tag{182}$$

All three channels use target-policy centering. The completed evidence evaluates a first action followed by π_g, while the one-step base approximates that continuation through v_g, including when the evidence is used to update behavior.

Centering does not generally remove the critic's continuation error. Recall that $\gamma \in (0,1)$ is the critic discount, $Z \sim m$ is a synthetic risk with prior $m = \mathrm{Unif}[0, 1/2]$, drawn independently at the query root and retained along the continuation, and $\Delta_{g,z} = V^{\pi_g}_{g,z} - v_g$ is the conditional model-value error. Here $V^{\pi_g}_{g,z}$ is the exact discounted value at fixed risk z, and $\widehat{P}^{a}_{g,z}, \widehat{P}^{\pi_g}_{g,z}$ are the one-step observation-transition operators at that risk, with the action fixed to a or averaged under π_g, respectively. In DSIS the risk variable is omitted. Theorem 1, followed by subtraction of the π_i-weighted action mean, gives

$$\mathbf{a}_i(a) - \mathbf{a}^0_i(a) = \gamma \mathbb{E}_{Z\sim m}\Big[\big((\widehat{P}^{a}_{g,Z} - \widehat{P}^{\pi_g}_{g,Z})\Delta_{g,Z}\big)(s_i)\Big]. \tag{183}$$

Here the operators act on the function $\Delta_{g,Z}$ before evaluation at s_i. Consequently, the base recovers the entire centered action vector exactly if and only if the omitted continuation $Q^{\pi_g}_g(s_i, a) - Q^0_g(s_i, a)$ is constant across actions. On-policy sampling does not impose this condition. The shared advantage face therefore includes the centered residual correction derived earlier.

The base is evaluated at every batch occurrence. For residual completion, the implementation draws $K_{\mathrm{res}} = 128$ occurrences with replacement. Let ϖ_i denote the probability of occurrence i and I_k the occurrence index in draw k; the draws are independent and identically distributed (iid) conditional on the frozen inputs:

$$\varpi_i = \frac{1 + w^J_i + w^P_i}{3n}, \qquad \mathsf{I}_k \overset{\mathrm{iid}}{\sim} \mathrm{Categorical}(\varpi_1, \ldots, \varpi_n), \quad k = 1, \ldots, K_{\mathrm{res}}. \tag{184}$$

Recall that $\widehat{A}^{(k)}_g$ denotes the sampled discounted target-policy advantage from Equation (74), using the independent bank of 32 continuation tapes assigned to draw k. It is centered under π_g at the selected occurrence. Repeated selections of the same occurrence receive distinct banks. Define the centered residual correction by

$$\Delta\widehat{\mathbf{a}}^{(k)}(a) = \widehat{A}^{(k)}_g(s_{\mathsf{I}_k}, a) - \mathbf{a}^0_{\mathsf{I}_k}(a). \tag{185}$$

Let $\chi \in \{0, J, P\}$ index the behavior-advantage, target-advantage, and transmission roles, respectively, and set $w^0_i = 1$. For finite actionwise coefficients $\ell = (\ell_i(a))_{i,a}$, define the sampled linear evidence functional

$$\widehat{\Lambda}^{\chi}_g[\ell] = \frac{1}{n}\sum_{i=1}^{n} w^{\chi}_i \langle \ell_i, \mathbf{a}^0_i\rangle + \frac{1}{K_{\mathrm{res}}}\sum_{k=1}^{K_{\mathrm{res}}} \frac{w^{\chi}_{\mathsf{I}_k}}{n\varpi_{\mathsf{I}_k}} \left\langle \ell_{\mathsf{I}_k}, \Delta\widehat{\mathbf{a}}^{(k)}\right\rangle. \tag{186}$$

All three channels share the same root draws and continuation banks. Since each weight vector has empirical mean one, $\sum_i \varpi_i = 1$ and $0 < w^{\chi}_i/(n\varpi_i) \le 3$ in exact arithmetic. In floating-point arithmetic, the implementation normalizes $1 + w^J_i + w^P_i$ by its computed sum and reuses that same probability vector for root sampling and inverse-proposal factors. For vector-valued coefficients, the functional acts componentwise.

Define the occurrence-level estimated advantage vector $\widehat{\mathbf{a}}_i$ by the following algebraic representation of the main-text evidence:

$$\widehat{\mathbf{a}}_i = \mathbf{a}^0_i + \frac{1}{K_{\mathrm{res}}\varpi_i}\sum_{k:\mathsf{I}_k = i} \Delta\widehat{\mathbf{a}}^{(k)}. \tag{187}$$

An empty sum is zero, so an occurrence not selected for residual completion retains its base advantage. Therefore,

$$\widehat{\Lambda}^{\chi}_g[\ell] = \frac{1}{n}\sum_i w^{\chi}_i \langle \ell_i, \widehat{\mathbf{a}}_i\rangle, \qquad \langle \pi_i, \widehat{\mathbf{a}}_i\rangle = 0. \tag{188}$$

This identity does not require materializing a completed action-value table at all roots: the implementation retains the dense base and sampled corrections separately. Distinct occurrences of the same observation may have different realized corrections. The main-text abbreviations are $\Lambda_A = \widehat{\Lambda}^0_g$, $\Lambda_J = \widehat{\Lambda}^J_g$, and $\Lambda_P = \widehat{\Lambda}^P_g$.

Theorem 2 (Conditional mean of the shared advantage evidence)**.** *Let $\mathscr{G}_g$ be the sigma-field generated by the frozen generation inputs before residual-root and continuation-bank sampling. Expectations below are taken under the residual-sampling law conditional on these inputs. Assume the integrability conditions of Theorem 1. Conditionally on $\mathscr{G}_g$, draw roots with Equation* (184) *and use the unbiased centered continuation estimators from Equations* (72) *and* (74)*. In exact arithmetic, for every finite $\mathscr{G}_g$-measurable coefficient array ℓ and every role $\chi \in \{0, J, P\}$,*

$$\mathbb{E}[\widehat{\mathbf{a}}_i \mid \mathscr{G}_g] = \mathbf{a}_i, \qquad \mathbb{E}[\widehat{\Lambda}_g^{\chi}[\ell] \mid \mathscr{G}_g] = \frac{1}{n} \sum_i w_i^{\chi} \langle \ell_i, \mathbf{a}_i \rangle. \tag{189}$$

Proof. For a root draw k, let $\mathbf{1}\{\mathsf{I}_k = i\}$ be the indicator of selecting occurrence i. Conditional unbiasedness of the continuation bank gives

$$\mathbb{E}[\mathbf{1}\{\mathsf{I}_k = i\} \Delta \widehat{\mathbf{a}}^{(k)} \mid \mathscr{G}_g] = \varpi_i (\mathbf{a}_i - \mathbf{a}_i^0),$$

Substitution in Equation (187) cancels the proposal factor and proves the first identity. The second follows from Equation (188), conditional linearity, and measurability of the frozen weights and coefficients. □

The result applies componentwise to frozen parent-policy score coefficients. Nonlinear transformations, gradient norms, fusion coefficients, and adaptively selected candidates do not inherit this unbiasedness statement.

Further consequences of critic evaluation error. The following consequence also applies to fixed query observations outside the empirical batch. For such observations, $A_g^{\mathrm{adv}}(s, \cdot)$ and $A_g^{\mathrm{base}}(s, \cdot)$ denote the $\pi_g(\cdot \mid s)$-centered versions of $Q_g^{\pi_g}(s, \cdot)$ and $Q_g^0(s, \cdot)$, respectively, as in Equation (182).

Corollary 1 (Action-dependent evaluation error survives centering)**.** *Under Theorem 1, fix an observation s satisfying its assumptions. The one-step base recovers all centered action values at s exactly if and only if $Q_g^{\pi_g}(s, a) - Q_g^0(s, a)$ is constant across actions. At batch occurrences, Equation* (183) *gives the corresponding centered correction.*

Proof. Subtract the π_g-weighted mean over actions from (58). Writing $\Delta_g^Q(s, a) = Q_g^{\pi_g}(s, a) - Q_g^0(s, a)$ for the actionwise error, equality of the centered vectors means $\Delta_g^Q(s, a) = \sum_{a' \in \mathcal{A}} \pi_g(a' \mid s) \Delta_g^Q(s, a')$ for every action a. The right-hand side is independent of a, which proves necessity; a constant vector also satisfies this equality, proving sufficiency. □

On-policy sampling does not impose this constancy condition or make $\Delta_{g,z}$ vanish. To express its consequence for a policy update, let υ be any fixed state-sampling distribution, including one obtained on-policy. Let ϑ denote actor parameters, distinct from critic parameters θ, and let ϑ_g be the parent actor parameter, so that $\pi_g = \pi_{\vartheta_g}$. For a frozen scalar action-evaluation function f, define the local actor surrogate $\mathcal{J}_g^{\mathrm{actor}}$ and the parent-policy score vector sc_g by

$$\mathcal{J}_g^{\mathrm{actor}}(\vartheta; f, \upsilon) = \mathbb{E}_{S \sim \upsilon}\left[\sum_a \pi_\vartheta(a \mid S) f(S, a)\right], \qquad \mathrm{sc}_g(s, a) = \nabla_\vartheta \log \pi_\vartheta(a \mid s)\big|_{\vartheta = \vartheta_g}. \tag{190}$$

In the expectation below, $S \sim \upsilon$ and $\mathsf{A} \mid S \sim \pi_g(\cdot \mid S)$; the shorthand Q in $\mathcal{J}_g^{\mathrm{actor}}$ denotes the frozen function $Q_g^{\pi_g}$. Assume the policy is differentiable with positive action probabilities, the surrogate expectations are finite, and differentiation can be interchanged with the state expectation. Also assume absolute score integrability, where $\|\cdot\|_2$ denotes the Euclidean norm:

$$\mathbb{E}_{\upsilon, \pi_g}\left[\|\mathrm{sc}_g(S, \mathsf{A})\|_2 \max_{a' \in \mathcal{A}} \big(|Q_g^{\pi_g}(S, a')| + |Q_g^0(S, a')|\big)\right] < \infty. \tag{191}$$

For a finite empirical state measure these integrability and interchange conditions hold whenever the pointwise scores and action values are finite. Then

$$\nabla_\vartheta \mathcal{J}_g^{\mathrm{actor}}(\vartheta; Q, \upsilon)\big|_{\vartheta_g} - \nabla_\vartheta \mathcal{J}_g^{\mathrm{actor}}(\vartheta; Q_g^0, \upsilon)\big|_{\vartheta_g} = \mathbb{E}_{\upsilon, \pi_g}[\mathrm{sc}_g(S, \mathsf{A})(A_g^{\mathrm{adv}}(S, \mathsf{A}) - A_g^{\mathrm{base}}(S, \mathsf{A}))]. \tag{192}$$

This follows from $\nabla_\vartheta \pi_\vartheta = \pi_\vartheta \nabla_\vartheta \log \pi_\vartheta$ and $\sum_a \nabla_\vartheta \pi_\vartheta(a \mid s) = 0$, the same baseline cancellation used in policy-gradient methods (Sutton et al., 1999). Residual completion supplies the missing population score term without requiring the scalar critics to have converged. This is a statement about the frozen, fixed-state surrogate. Interpreting it as the gradient of a trajectory-return objective additionally requires that objective's correct state and risk occupancy measure.

Why the supply-chain structure makes the correction relevant. A regular order changes the pipeline immediately, but its inventory and cost consequences occur after the four-period lead time. For example, with $v_g \equiv 0$ and fixed expedited quantity,

$$Q_g^0(s,(q^r,q^e)) = \int \mathbb{E}_z^{s,(q^r,q^e)}[R_0] m(dz) \quad \text{is independent of } q^r. \tag{193}$$

Immediate cost contains no regular-order charge, and the current regular quantity does not change end-of-period inventory. When supply is available, however, different regular quantities enter different successor pipelines and can produce different future holding and backlog costs. The residual continuation recovers precisely the long-term action differences omitted by this base. An approximate nonzero critic can already capture part of those differences; the same identity completes the part it misses. This issue exists in DSIS as well as DSIS-RB, regardless of whether data collection is on-policy.

B.5.2 Direct advantage learning for behavior and target

Suppress generation indices on the parent distributions and write $m_{\theta^\mu}(a \mid s) = \mu_{\theta^\mu}(a \mid s)$ and $p_{\theta^\pi}(a \mid s) = \pi_{\theta^\pi}(a \mid s)$ for the two parameterized candidate policies. Let $\theta_g^\mu, \theta_g^\pi$ be their frozen parent parameter vectors, so $m_{\theta_g^\mu} = \mu$ and $p_{\theta_g^\pi} = \pi$. The symbols m, p in the local objectives denote candidate distributions; this use of m is distinct from the synthetic-risk prior above. Write $m_i(a) = m_{\theta^\mu}(a \mid s_i)$ and $p_i(a) = p_{\theta^\pi}(a \mid s_i)$ for their occurrence-level rows. Assume positive, differentiable candidate probabilities near the parents and finite values and derivatives at the retained observations. Recall that $\tau = \kappa S_{\mathrm{cal}} > 0$ is the fixed run-level reward scale from Equation (163); $\kappa > 0$ is the tuned dimensionless multiplier and $S_{\mathrm{cal}} > 0$ is the initialization utility-range scale.

The *score function* of a policy is its log-probability gradient. The parent behavior and target scores are

$$\begin{aligned} \mathrm{sc}_g^\mu(s,a) &= \nabla_{\theta^\mu} \log m_{\theta^\mu}(a \mid s)|_{\theta_g^\mu}, \\ \mathrm{sc}_g(s,a) &= \nabla_{\theta^\pi} \log p_{\theta^\pi}(a \mid s)|_{\theta_g^\pi}. \end{aligned} \tag{194}$$

The target score is the sc_g of the notation table, with $\vartheta_g = \theta_g^\pi$. Write $\mathrm{sc}_i^\mu(a) = \mathrm{sc}_g^\mu(s_i,a)$ and $\mathrm{sc}_i(a) = \mathrm{sc}_g(s_i,a)$.

Recall that $\widehat{\Lambda}_g^\chi$ is the sampled linear advantage functional in Equation (186); role 0 uses unit weights and role J uses w_i^J. The two direct surfaces are

$$A(m) = \frac{1}{\tau}\widehat{\Lambda}_g^0[m-\mu], \qquad J(p) = \widehat{\Lambda}_g^J[p-\pi]. \tag{195}$$

During differentiation, the subtracted parents μ, π and the weights, base advantages, sampled root indices, and continuation corrections defining $\widehat{\Lambda}_g^\chi$ remain fixed. In particular, the appearance of π_g in the centering and continuation estimator does not create an additional derivative. Only the candidate distribution depends on the differentiated actor parameters.

Theorem 3 (Parent gradients of the direct advantage surfaces)**.** *Under the positivity and differentiability assumptions above, define $g_A = \nabla_{\theta^\mu} A(m_{\theta^\mu})|_{\theta_g^\mu}$ and $g_J = \nabla_{\theta^\pi}(J(p_{\theta^\pi})/\tau)|_{\theta_g^\pi}$. Inside the functionals, the arrays $\mathrm{sc}^\mu, \mathrm{sc}$ have rows $\mathrm{sc}_i^\mu, \mathrm{sc}_i$; multiplication by a policy probability is action-wise. With all evidence and parents frozen,*

$$g_A = \frac{1}{\tau}\widehat{\Lambda}_g^0[\mu\ \mathrm{sc}^\mu] = \frac{1}{n\tau}\sum_{i,a} \mu_i(a)\widehat{\mathbf{a}}_i(a)\,\mathrm{sc}_i^\mu(a), \tag{196}$$

and

$$g_J = \frac{1}{\tau}\widehat{\Lambda}_g^J[\pi\ \mathrm{sc}] = \frac{1}{n\tau}\sum_{i,a} w_i^J \pi_i(a)\widehat{\mathbf{a}}_i(a)\,\mathrm{sc}_i(a). \tag{197}$$

Proof. For behavior, linearity and the chain rule give

$$\begin{aligned}\nabla_{\theta^\mu} A(m_{\theta^\mu}) &= \frac{1}{n\tau}\sum_{i,a} \widehat{\mathbf{a}}_i(a)\nabla_{\theta^\mu} m_{\theta^\mu}(a \mid s_i) \\ &= \frac{1}{n\tau}\sum_{i,a} \widehat{\mathbf{a}}_i(a) m_{\theta^\mu}(a \mid s_i)\nabla_{\theta^\mu} \log m_{\theta^\mu}(a \mid s_i).\end{aligned} \tag{198}$$

Evaluation at θ^μ_g proves Equation (196). For target, J is expressed in unscaled advantage units, so

$$\begin{aligned}\nabla_{\theta^\pi} \frac{J(p_{\theta^\pi})}{\tau} &= \frac{1}{n\tau}\sum_{i,a} w_i^J \widehat{\mathbf{a}}_i(a)\nabla_{\theta^\pi} p_{\theta^\pi}(a \mid s_i) \\ &= \frac{1}{n\tau}\sum_{i,a} w_i^J \widehat{\mathbf{a}}_i(a) p_{\theta^\pi}(a \mid s_i)\nabla_{\theta^\pi} \log p_{\theta^\pi}(a \mid s_i).\end{aligned} \tag{199}$$

Evaluation at θ^π_g proves Equation (197). □

Behavior therefore learns whether its first-action changes are supported by target-continuation evidence; no separate Q^{μ_g} is estimated. Target uses the same action evidence under ν_g^J. For both surfaces, the coefficient rows $m_i - \mu_i$ and $p_i - \pi_i$ sum to zero. For any positive differentiable policy q, differentiating its probability normalization gives $\sum_a q(a \mid s)\nabla \log q(a \mid s) = 0$, where the gradient is taken with respect to that policy's parameters. Thus action-independent baselines cancel in these two channels. The transmission coefficients below do not generally have this cancellation property.

B.5.3 Advantage-filtered behavior-to-target transmission

Recall that $\mu_i(a), \pi_i(a)$ are the frozen parent probabilities, $\widehat{\mathbf{a}}_i$ is the sampled discounted target-centered advantage, and $\widehat{\Lambda}^P_g$ is their linear functional with transmission weights w_i^P in Equation (186). Write $\delta_i(a) = \mu_i(a) - \pi_i(a)$ for the parent probability discrepancy, distinct from the point mass δ_{s_i} and the earlier controller multiplier gap. Write $\delta = (\delta_i(a))_{i,a}$ for the full array; products, absolute values, and ratios inside the functional below act actionwise. The transmission surface is

$$P(p) = \frac{1}{\tau}\widehat{\Lambda}^P_g\left[|\delta|\left(\frac{p}{\pi} - 1\right)\right], \tag{200}$$

For an arbitrary candidate p_{θ^π},

$$\nabla_{\theta^\pi} P(p_{\theta^\pi}) = \frac{1}{\tau}\widehat{\Lambda}^P_g\left[\frac{|\delta|}{\pi}\nabla_{\theta^\pi} p_{\theta^\pi}\right] = \frac{1}{\tau}\widehat{\Lambda}^P_g\left[|\delta|\frac{p_{\theta^\pi}}{\pi}\nabla_{\theta^\pi}\log p_{\theta^\pi}\right]. \tag{201}$$

The parent denominator π, the discrepancy δ, and the evidence are constants here. Evaluating at $p_{\theta^\pi_g} = \pi$ cancels the probability ratio. Define $g_P = \nabla_{\theta^\pi} P(p_{\theta^\pi})|_{\theta^\pi_g}$; then

$$g_P = \frac{1}{\tau}\widehat{\Lambda}^P_g[|\delta|\,\mathrm{sc}] = \frac{1}{n\tau}\sum_{i,a} w_i^P |\delta_i(a)|\widehat{\mathbf{a}}_i(a)\,\mathrm{sc}_i(a). \tag{202}$$

Disagreement supplies the magnitude and advantage supplies the sign. Unlike g_J, this score coefficient is not multiplied by $\pi_i(a)$; an action rarely selected by target can still receive substantial transmission when behavior assigns it appreciably more mass.

Centered evidence is essential because, in general, $\sum_a |\delta_i(a)|(p_i(a)/\pi_i(a) - 1) \neq 0$. Replacing advantages by uncentered action values changes P. Both the dense base and sampled correction are therefore centered before entering this channel.

Theorem 4 (Selective transmission in the advantage face)**.** *Fix an occurrence on the finite action set $\mathcal{A}$. Write $\pi = (\pi_a)_a$ for a strictly positive target distribution, $\mu = (\mu_a)_a$ for a behavior distribution, and $e = (e_a)_a$ for finite centered evidence with $\sum_a \pi_a e_a = 0$. The evidence can be the exact vector $\mathbf{a}_i$ or the realized sampled vector $\widehat{\mathbf{a}}_i$. Let $\tau > 0$. Define the discrepancy δ_a, its relative magnitude h_a, and the signed score coefficient b_a by*

$$\delta_a = \mu_a - \pi_a, \qquad h_a = \frac{|\delta_a|}{\pi_a}, \qquad b_a = \frac{|\delta_a| e_a}{\tau}. \tag{203}$$

Write $\operatorname{sign}(x) = 1$ for $x > 0$ and -1 for $x < 0$. For $\delta_a e_a \neq 0$,

$$b_a = \frac{\delta_a |e_a|}{\tau} \operatorname{sign}(\delta_a e_a). \tag{204}$$

Thus the sign of the discrepancy is retained in the coefficient b_a when $\delta_a e_a > 0$ and reversed when $\delta_a e_a < 0$. For a nonnegative auxiliary step η, define the distribution q_η^R (with R labeling reflection) and its normalizing constant Z_η by

$$q_{\eta,a}^R = \frac{\pi_a \exp(\eta h_a e_a/\tau)}{Z_\eta}, \qquad Z_\eta = \sum_b \pi_b \exp(\eta h_b e_b/\tau). \tag{205}$$

It satisfies

$$\langle q_\eta^R - \pi, e\rangle = \frac{1}{Z_\eta} \sum_a \pi_a e_a \left[\exp(\eta h_a e_a/\tau) - 1\right] \geq 0. \tag{206}$$

For $\eta > 0$, the inequality is strict if and only if $|\delta_a||e_a| > 0$ for at least one action. In addition,

$$\frac{d}{d\eta} \langle q_\eta^R - \pi, e\rangle \Big|_{\eta=0} = \frac{1}{\tau} \sum_a |\delta_a| e_a^2. \tag{207}$$

Proof. For one occurrence, the transmission contribution of action a to a candidate p is

$$T_a(p) = \frac{|\delta_a|}{\tau \pi_a} (p_a - \pi_a) e_a. \tag{208}$$

The omitted occurrence weight is positive and does not change signs. Whenever $\delta_a \neq 0$, $T_a(p)$ is positive precisely when the candidate's coordinate change is supported by the evidence. We examine the four nonzero sign combinations separately.

(i) *Positive advantage and increased behavioral mass:* $e_a > 0$, $\delta_a > 0$. Behavior assigns more probability to a favorable action. Here $b_a = \delta_a e_a/\tau > 0$, so increasing p_a contributes positively to P. The behavioral deviation is retained.

(ii) *Negative advantage and decreased behavioral mass:* $e_a < 0$, $\delta_a < 0$. Behavior assigns less probability to an unfavorable action. Here $b_a = (-\delta_a) e_a/\tau < 0$, and decreasing p_a makes both $(p_a - \pi_a)$ and e_a negative, so $T_a(p) > 0$. This beneficial reduction is also retained.

(iii) *Positive advantage and decreased behavioral mass:* $e_a > 0$, $\delta_a < 0$. Behavior underweights a favorable action. Following that decrease would make $T_a(p) < 0$. Instead, $b_a = (-\delta_a) e_a/\tau > 0$ rewards an increase, reflecting the behavioral deviation.

(iv) *Negative advantage and increased behavioral mass:* $e_a < 0$, $\delta_a > 0$. Behavior overweights an unfavorable action. Following that increase would again make $T_a(p) < 0$. Here $b_a = \delta_a e_a/\tau < 0$ rewards a decrease, reflecting this deviation.

These cases establish Equation (204). If $e_a = 0$ or $\delta_a = 0$, both b_a and the coordinate contribution are zero. Coordinate statements describe contributions to the surface; feasible distributions must also obey the shared unit-mass constraint.

We next show that the auxiliary distribution respects this constraint while preserving a nonnegative total advantage. Positivity of π and finiteness of the evidence give $0 < Z_\eta < \infty$. Consequently q_η^R

is a positive probability distribution. Centering gives

$$\langle q_\eta^R - \pi, e\rangle = \frac{1}{Z_\eta}\sum_a \pi_a e_a \exp(\eta h_a e_a/\tau) = \frac{1}{Z_\eta}\sum_a \pi_a e_a[\exp(\eta h_a e_a/\tau) - 1].$$

If $e_a > 0$, then $h_a \geq 0$ makes the bracket nonnegative. If $e_a < 0$, the bracket is nonpositive, so its product with e_a is again nonnegative. If $e_a = 0$ or $h_a = 0$, the product is zero. For $\eta > 0$, it is strictly positive exactly when $h_a|e_a| > 0$. Since all $\pi_a > 0$, summation proves the inequality and its strictness condition.

Finally, let N_η be the numerator on the right-hand side of Equation (206). At $\eta = 0$, $N_0 = 0$ and $Z_0 = 1$. The quotient rule therefore gives

$$\frac{d}{d\eta}\frac{N_\eta}{Z_\eta}\bigg|_0 = N_0' = \frac{1}{\tau}\sum_a \pi_a h_a e_a^2 = \frac{1}{\tau}\sum_a |\delta_a| e_a^2,$$

which proves Equation (207). □

Here favorable and unfavorable mean positive and negative centered evidence relative to π. The theorem does not require behavior to be better than target in aggregate: beneficial deviations in the first two cases are retained even when behavior has harmful deviations elsewhere. The latter two cases use the disagreement magnitude while reversing its unsupported direction.

The auxiliary distribution also explains the surface itself. For any candidate action distribution p at the same occurrence, setting $\eta = 1$ and using probability normalization gives

$$\begin{aligned}\sum_a (p_a - \pi_a)\log\frac{q_{1,a}^R}{\pi_a} &= \sum_a (p_a - \pi_a)\left(\frac{h_a e_a}{\tau} - \log Z_1\right)\\ &= \frac{1}{\tau}\sum_a |\mu_a - \pi_a|\left(\frac{p_a}{\pi_a} - 1\right)e_a.\end{aligned} \tag{209}$$

The normalizer cancels because $\sum_a (p_a - \pi_a) = 0$. The implementation evaluates the right-hand side directly; q_η^R is a proof device, not an additional actor or a policy-mixing step.

With exact evidence $e = \mathbf{a}_i$, the theorem concerns the frozen model's first-action evaluation. With $e = \widehat{\mathbf{a}}_i$, it is a pathwise statement about the realized occurrence-level evidence; repeated states need not define a single common auxiliary policy. The specific auxiliary path has nonnegative advantage gain, but an arbitrary increase of P need not increase the unweighted advantage face. Nor does the theorem assign signs to every coordinate of a parameterized actor update. For independent categorical logits at this occurrence, the actual parent logit gradient for this occurrence is the positive occurrence weight times the following vector, where $b = (b_a)_a$ recalls the coefficients in Equation (203):

$$b - \pi\sum_a b_a, \tag{210}$$

because differentiating the softmax couples all action probabilities. Shared network parameters further couple states. The channel can increase behavior–target disagreement. It vanishes identically when $\mu = \pi$; a zero parameter gradient alone does not imply a constant finite surface.

B.5.4 Why the advantage channels use the scale $1/\tau$

Recall that $g_H = g_{H,g}$ is the parent gradient of the heuristic objective H_g in Equation (180). The heuristic objective is a dimensionless squared statistical distance; its utility dependence enters through the regularized controller. The unscaled advantage surfaces instead carry reward units. Since $\tau = \kappa S_{\mathrm{cal}}$ in Equation (163) has the same units as the heuristic utility, division by τ expresses the three advantage channels in compatible units. The next result also quantifies the utility-induced scale of guidance while accounting for any behavior–target mismatch already present at the parent.

Theorem 5 (Scale of heuristic guidance)**.** *Recall the frozen centered heuristic utility u_g, the Hellinger-regularized controller $q_{H,g}$ from Equation* (168)*, and its finite empirical fitting measure ξ_g from Equation* (174)*. The utility is the undiscounted finite-horizon quantity, distinct from the discounted advantages $\widehat{\mathbf{a}}_i$. Recall also $d_{\mathrm{FR}}(p,q) = 2\arccos\sum_a \sqrt{p_a q_a}$ for action distributions*

p, q. Assume finite utilities and actor scores on this panel, positive parent probabilities, and $\tau > 0$. Define the utility range $R_g(s)$ and the root-mean-square parent distance $D_{\mu,\pi}$ by

$$R_g(s) = \max_a u_g(s,a) - \min_a u_g(s,a), \qquad D_{\mu,\pi} = \left(\mathbb{E}_{\xi_g} d_{\mathrm{FR}}^2(\mu_g, \pi_g)\right)^{1/2}. \tag{211}$$

For symmetric matrices, write $X \succeq Y$ when $X - Y$ is positive semidefinite and $X \succ 0$ when X is positive definite; a superscript $\top$ denotes transpose. Let the behavior-actor matrix $F_\mu \succ 0$ satisfy

$$F_\mu \succeq \frac{1}{2}\mathbb{E}_{s\sim\xi_g} \sum_a \mu_g(a \mid s)\, \mathrm{sc}_g^\mu(s,a)\, \mathrm{sc}_g^\mu(s,a)^\top. \tag{212}$$

The actor matrix defined below has this property. Use $\|x\|_G = (x^\top G x)^{1/2}$ for a positive-definite G. The numerical factor π in the following bounds is the circle constant. Then

$$d_{\mathrm{FR}}(\pi_g(\cdot \mid s), q_{H,g}(\cdot \mid s)) \le \frac{\pi R_g(s)}{4\tau}, \tag{213}$$

and

$$\|g_H\|_{F_\mu^{-1}} \le \sqrt{2}\left[D_{\mu,\pi} + \frac{\pi}{4\tau}\left(\mathbb{E}_{\xi_g} R_g(S)^2\right)^{1/2}\right]. \tag{214}$$

In particular, at matched parents $\mu_g = \pi_g$, the natural norm is $O(\tau^{-1})$ as $\tau \to \infty$ for fixed fitting locations, utility evidence, parent policies, and score vectors.

Proof. At one fitting location, abbreviate $p = \pi_g(\cdot \mid s)$, $q = q_{H,g}(\cdot \mid s)$, $u_a = u_g(s,a)$, and $\lambda = \lambda_g(s)$, the controller's unit-mass multiplier. Set $B_{\pi,H} = \sum_a \sqrt{p_a q_a}$. Here $\mathbb{E}_q$ and Var_q denote expectation and variance over actions drawn from q. This affinity is between target and controller, distinct from the behavior–controller affinity in Equation (178). The stationarity equation (167) gives $\lambda - u_a = 2\tau\sqrt{p_a/q_a}$. Its first two moments under q are

$$\mathbb{E}_q(\lambda - u) = 2\tau B_{\pi,H}, \qquad \mathbb{E}_q[(\lambda - u)^2] = 4\tau^2.$$

Thus $\mathrm{Var}_q(u) = 4\tau^2(1 - B_{\pi,H}^2)$. Set $\psi = \arccos B_{\pi,H} \in [0, \pi/2]$. Since $2\psi \le \pi \sin\psi$ and a variable of range $R_g(s)$ has variance at most $R_g(s)^2/4$,

$$d_{\mathrm{FR}}(p,q) = 2\psi \le \frac{\pi}{2\tau}\sqrt{\mathrm{Var}_q(u)} \le \frac{\pi R_g(s)}{4\tau}.$$

For the guidance gradient, apply Cauchy–Schwarz to the centered formula in Equation (180). Recall the behavior–controller affinity $\mathcal{B}_g(s) = \sum_a \sqrt{\mu_g(a \mid s) q_{H,g}(a \mid s)}$ and angle $\varphi_g(s) = \arccos \mathcal{B}_g(s)$. At each state, the centered factor $\sqrt{q_{H,g}/\mu_g} - \mathcal{B}_g$ has second moment $1 - \mathcal{B}_g^2 = \sin^2\varphi_g$ under μ_g; this cancels the denominator $\sin\varphi_g$ in that formula. The coincident case follows by continuity. For any parameter direction v, a second Cauchy–Schwarz inequality over the fitting locations gives

$$|g_H^\top v| \le \left(\mathbb{E}_{\xi_g} d_{\mathrm{FR}}^2(\mu_g, q_{H,g})\right)^{1/2} \times \left(\mathbb{E}_{\xi_g} \sum_a \mu_g(a \mid S)[\mathrm{sc}_g^\mu(S,a)^\top v]^2\right)^{1/2}. \tag{215}$$

Equation (212) bounds the last factor by $\sqrt{2}\|v\|_{F_\mu}$. Taking the dual norm, then using the metric triangle inequality and Minkowski's inequality, proves Equation (214). □

For a more direct interpretation, hold the finite fitting locations, utility values, parent policies, and score vectors fixed and let $\tau \to \infty$. The remainders below are uniform over this finite panel and action set; the vector remainder is measured in the Euclidean parameter norm. At each state, put $\epsilon_\tau = 1/(2\tau)$ and $y_\tau = \lambda_g(s)/(2\tau)$. The controller normalization is $\sum_a \pi_g(a \mid s)(y_\tau - \epsilon_\tau u_g(s,a))^{-2} = 1$. At $(y_\tau, \epsilon_\tau) = (1, 0)$ its derivative with respect to y_τ is -2, so the implicit function theorem and target centering give $y_\tau = 1 + O(\epsilon_\tau^2)$. Expanding the inverse square therefore yields

$$q_{H,g}(a \mid s) = \pi_g(a \mid s)\left[1 + \frac{u_g(s,a)}{\tau} + O(\tau^{-2})\right]. \tag{216}$$

At $\mu_g = \pi_g$, this implies, pointwise on the fixed panel, $\sqrt{\pi_g q_{H,g}} = \pi_g[1 + u_g/(2\tau) + O(\tau^{-2})]$ and $\mathcal{B}_g = 1 + O(\tau^{-2})$. Also $2\varphi_g/\sin\varphi_g = 2 + O(\tau^{-2})$, with the continuous value 2 at $\varphi_g = 0$. Substitution in the centered guidance gradient of Equation (180) yields

$$g_H = \frac{1}{\tau}\mathbb{E}_{\xi_g}\sum_a \pi_g(a \mid S)u_g(S,a)\,\mathrm{sc}_g^{\mu}(S,a) + O(\tau^{-2}). \tag{217}$$

This fixed-panel expansion does not differentiate through the controller rollouts that originally selected ξ_g. In general, the whole gradient need not decay as $1/\tau$: if $u_g \equiv 0$, then $q_{H,g} = \pi_g$ for every τ, while guidance can remain nonzero when $\mu_g \neq \pi_g$. The mismatch term in Theorem 5 accounts for this component.

There is also an exact unit-conversion interpretation. Multiplying all frozen return evidence and calibration utilities by $c > 0$ multiplies $u_g, \widehat{\mathbf{a}}_i$, and τ by c. The controller, H_g, and the scaled surfaces $A, J/\tau, P$ are unchanged. This is an invariance of the frozen update interface, not a claim that arbitrary critic optimization is invariant to a change of reward units. It does not make the four gradient norms equal; their relative contributions are handled by the fusion below.

B.5.5 Fisher geometry and the Fuse construction

The score-based Fisher metric and inverse-Fisher preconditioning follow the natural-policy-gradient construction of Kakade (2001). Here we specify the frozen consumer measures and the fusion rule. Recall the four empirical state measures: D_g is the unweighted generation batch, ν_g^J and ν_g^P use the positive mean-one weights w_i^J and w_i^P, and ξ_g is the heuristic fitting measure from Equation (174). For a parent actor q and a frozen probability measure ν on observations, let $\mathrm{sc}_q(s,a)$ be its log-probability gradient at its own parent parameters. Define its empirical Fisher matrix $\mathsf{F}(q,\nu)$ by

$$\mathsf{F}(q,\nu) = \mathbb{E}_{s\sim\nu}\sum_a q(a \mid s)\,\mathrm{sc}_q(s,a)\,\mathrm{sc}_q(s,a)^{\top}, \tag{218}$$

Let I denote the identity matrix on the relevant actor parameter space and $\lambda_F > 0$ the damping coefficient. The common matrix for each actor's two channels is

$$\begin{aligned} F_\mu &= \tfrac{1}{2}\left[\mathsf{F}(\mu_g, D_g) + \mathsf{F}(\mu_g, \xi_g)\right] + \lambda_F I, \\ F_\pi &= \tfrac{1}{2}\left[\mathsf{F}(\pi_g, \nu_g^J) + \mathsf{F}(\pi_g, \nu_g^P)\right] + \lambda_F I. \end{aligned} \tag{219}$$

Thus the common matrix contains a single ridge term $\lambda_F I$. The matrices are positive definite and actor-specific; the two policies do not share a parameter-space update.

For either $(g_1, g_2) = (g_A, g_H)$ or $(g_1, g_2) = (g_J, g_P)$, use the corresponding fixed matrix $F = F_\mu$ or $F = F_\pi$. In this subsection, $i \in \{1, 2\}$ indexes channels, rather than data occurrences. Define their natural norms n_i, their maximum $n_{\max}$, and the scalar denominator Δ_F by

$$n_i = \sqrt{g_i^{\top} F^{-1} g_i}, \quad n_{\max} = \max(n_1, n_2), \quad \Delta_F = n_1^2 + 6n_1n_2 + n_2^2. \tag{220}$$

For $n_{\max} > 0$, define the fusion coefficients α_1, α_2 and the ideal parameter direction d by

$$\begin{aligned} \alpha_1 = \frac{n_{\max}(n_1 + 3n_2)}{\Delta_F}, \qquad \alpha_2 = \frac{n_{\max}(n_2 + 3n_1)}{\Delta_F}, \\ d = \mathrm{Fuse}_F(g_1, g_2) = F^{-1}(\alpha_1 g_1 + \alpha_2 g_2). \end{aligned} \tag{221}$$

If both norms are zero, set both coefficients and d to zero. For $n_{\max} > 0$, the normalized contributions

$$\omega_1 = \frac{n_1(n_1 + 3n_2)}{\Delta_F}, \qquad \omega_2 = \frac{n_2(n_2 + 3n_1)}{\Delta_F} \tag{222}$$

satisfy $\omega_1 + \omega_2 = 1$ and $\alpha_i n_i = n_{\max}\omega_i$. The α_i themselves need not sum to one.

Theorem 6 (Variational characterization of the fusion weights)**.** *Let $n_1, n_2 > 0$, and define the two-coordinate probability simplex $\Delta_2 = \{\omega \in \mathbb{R}^2 : \omega_i \geq 0,\ \omega_1 + \omega_2 = 1\}$. For probability vectors x, y with $y_i > 0$, define the Pearson divergence $\chi^2(x\|y) = \sum_i (x_i - y_i)^2/y_i$. The weights in Equation* (222) *are the unique solution of*

$$\min_{\omega\in\Delta_2}\left\{\chi^2\left(\omega \,\middle\|\, \frac{(n_1, n_2)}{n_1 + n_2}\right) + \chi^2(\omega \,\|\, (1/2, 1/2))\right\}. \tag{223}$$

Proof. Write $r_i = n_i/(n_1+n_2)$ for the normalized natural strengths. The objective equals $\sum_i \omega_i^2(1/r_i+2)-2$. Its constrained minimizer obeys $\omega_i \propto (1/r_i+2)^{-1}$, yielding Equation (222). The positive diagonal quadratic coefficients make the objective strictly convex, proving uniqueness. □

For one zero norm, use the extended convention: a summand with $y_i = 0$ is zero if $x_i = 0$ and $+\infty$ if $x_i > 0$. The only finite feasible weight then assigns all mass to the nonzero channel, agreeing with Equation (222). The double-zero case uses the separately defined zero direction and requires no normalized votes. The two terms balance relative natural-gradient strength and equal channel contributions.

Equal positive norms give $\alpha_1 = \alpha_2 = 1/2$. If $n_1 > 0, n_2 = 0$, the coefficients are $(1, 3)$ and $d = F^{-1}g_1$. The second gradient contributes zero, but its frozen finite objective need not be constant away from the parent. A vanishing channel has a continuously vanishing contribution, and the fused direction extends continuously to the double-zero case without a threshold or denominator floor.

Theorem 7 (Bounded fusion and channel compatibility)**.** *Let $\theta_\star$ denote the parent parameter vector of the actor under consideration, and let U_1, U_2 be two differentiable channel objectives on a neighborhood of $\theta_\star$. Let $F \succ 0$ be fixed and $g_i = \nabla U_i(\theta_\star)$ for $i \in \{1, 2\}$. Construct n_i, α_i, d as in Equations* (220)–(221)*; the votes ω_i in Equation* (222) *are used only when $n_{\max} > 0$. Recall $\|x\|_F = (x^\top Fx)^{1/2}$. When $n_1 n_2 > 0$, define the unit natural directions v_i and their Fisher cosine c_F by*

$$v_i = \frac{F^{-1}g_i}{n_i}, \qquad c_F = \frac{g_1^\top F^{-1} g_2}{n_1 n_2} \in [-1, 1]. \tag{224}$$

For each $i \in \{1, 2\}$, let $j = 3 - i$ denote the other channel. Then

$$\begin{aligned} d &= n_{\max}(\omega_1 v_1 + \omega_2 v_2), \\ n_{\max}|\omega_1 - \omega_2| \le \|d\|_F &\le n_{\max}, \\ g_i^\top d &= n_{\max} n_i(\omega_i + c_F\omega_j), \qquad j \ne i. \end{aligned} \tag{225}$$

Both directional derivatives are nonnegative if and only if

$$c_F \ge -\min\left\{\frac{\omega_1}{\omega_2}, \frac{\omega_2}{\omega_1}\right\}. \tag{226}$$

The channel with the larger natural norm always has a nonnegative directional derivative. If $n_1 \ge n_2$, the weaker channel satisfies

$$g_2^\top d \ge -n_{\max} n_2 \frac{n_1^2 - n_2^2}{\Delta_F}. \tag{227}$$

For any norms, hold α_1, α_2 fixed and set $U = \alpha_1 U_1 + \alpha_2 U_2$. Then $\nabla U(\theta_\star)^\top d = \|d\|_F^2$. If U is twice continuously differentiable along the update segment and $\nabla^2 U \succeq -LF$ there for a nonnegative curvature constant L, then for a parameter-step length $t \ge 0$,

$$U(\theta_\star + td) - U(\theta_\star) \ge t\left(1 - \frac{Lt}{2}\right)\|d\|_F^2, \qquad t \ge 0. \tag{228}$$

The bound applies whenever the entire segment is in the stated domain.

Proof. The identity $\alpha_i n_i = n_{\max}\omega_i$ gives the representation of d. Each v_i has unit F-norm, so the triangle and reverse-triangle inequalities give the two norm bounds. Since $g_i^\top v_i = n_i$ and $g_i^\top v_j = n_i c_F$, taking the inner product with g_i gives its slope. Both slopes are nonnegative exactly when $c_F \ge -\omega_1/\omega_2$ and $c_F \ge -\omega_2/\omega_1$, which is Equation (226).

Furthermore, $\omega_1 - \omega_2 = (n_1^2 - n_2^2)/\Delta_F$. If $n_1 \ge n_2$, then $\omega_1 \ge \omega_2$ and $c_F \ge -1$ gives $g_1^\top d \ge n_{\max} n_1(\omega_1 - \omega_2) \ge 0$ and the stated lower bound for $g_2^\top d$. The zero-norm cases follow directly from positive definiteness of F and the definition of d.

Finally, with the coefficients frozen, $\nabla U(\theta_\star) = \alpha_1 g_1 + \alpha_2 g_2 = Fd$. Taylor's formula with integral remainder yields

$$U(\theta_\star + td) - U(\theta_\star) = t\|d\|_F^2 + \int_0^t (t - s) d^\top \nabla^2 U(\theta_\star + sd) d \, ds.$$

The Hessian lower bound makes the integral at least $-Lt^2\|d\|_F^2/2$, proving the result. □

Likewise, suppose each U_i is twice continuously differentiable on the segment and $\nabla^2 U_i \succeq -L_i F$ there for constants $L_i \geq 0$. For the same nonnegative step length t,

$$U_i(\theta_\star + td) - U_i(\theta_\star) \geq t g_i^\top d - \frac{L_i t^2}{2}\|d\|_F^2. \tag{229}$$

When both norms are positive, the slope is $g_i^\top d = n_{\max} n_i(\omega_i + c_F \omega_j)$; the displayed form also covers zero-norm channels. Strictly positive initial slopes therefore give simultaneous improvement for sufficiently small positive steps. Behavior uses $(U_1, U_2) = (A, H_g)$ and target uses $(U_1, U_2) = (J/\tau, P)$. No derivative passes through the parent Fisher matrices, norms, or fusion coefficients.

These are conditional compatibility guarantees, not unconditional common ascent. For example, $F = 1$, $g_1 = 2$, and $g_2 = -1$ give $(\alpha_1, \alpha_2) = (10/17, 14/17)$ and $d = 6/17$, so the second channel decreases to first order. A larger natural norm describes gradient strength in this geometry, not statistical confidence.

B.5.6 Numerical updates and endpoint feasibility control

Fisher damping and its selection. The ridge λ_F is a fixed positive hyperparameter that stabilizes the linear systems and makes the common Fisher matrices invertible. We use

$$\lambda_F = 0.0067202476327024195 \tag{230}$$

for both actors, all initialization sources, and all generations. This value was inherited from an earlier cold-start study of the precursor implementation that jointly varied the critic learning rate, emphatic learning rate, and Fisher damping. The study evaluated 12 fixed Sobol configurations, mapping damping logarithmically over $[10^{-4}, 10^{-2}]$, with three training seeds per configuration. Each run used 60 generations of 25,000 transitions. Configurations were ranked by the mean stochastic target-policy cost across the three seeds on common tuning tapes, with ties resolved by configuration index. The selected damping was then fixed and inherited by the present implementation. The current seven hyperparameter configurations vary the critic learning rate, emphatic learning rate, and heuristic multiplier κ; they do not retune λ_F. It is not adapted from within-run returns or gradient norms.

Numerical natural-gradient directions. Two conjugate-gradient (CG) solves estimate the channel natural norms. Let $\widetilde{\alpha}_1, \widetilde{\alpha}_2$ be the coefficients obtained by applying Equation (221) to those estimated norms. After freezing these coefficients, a separate CG solve uses the right-hand side $\widetilde{\alpha}_1 g_1 + \widetilde{\alpha}_2 g_2$. The final direction is not formed by adding the two approximate channel solutions. Fisher-vector products avoid constructing a dense parameter-space matrix. The J14 r3 implementation uses a base limit of 256 iterations and Euclidean relative residual tolerance 10^{-4}. The retained J12 implementation uses the same base limit and tolerance 5×10^{-5}. Numerical recovery retains the same frozen equations and data rather than resampling evidence.

Theorem 8 (Residual bound for a numerical fused solve)**.** *Let $F \succ 0$ and let $\widetilde{\alpha}_1, \widetilde{\alpha}_2$ be the fixed coefficients actually passed to the fused CG solve, possibly computed from approximate channel norms. Define its exact solution $d_{\widetilde{\alpha}} = F^{-1}(\widetilde{\alpha}_1 g_1 + \widetilde{\alpha}_2 g_2)$. For a computed direction $\widetilde{d}$, define the linear-system residual r_{CG} and its dual norm $\varepsilon_{\mathrm{CG}}$ by*

$$r_{\mathrm{CG}} = (\widetilde{\alpha}_1 g_1 + \widetilde{\alpha}_2 g_2) - F\widetilde{d}, \qquad \varepsilon_{\mathrm{CG}} = \|r_{\mathrm{CG}}\|_{F^{-1}}. \tag{231}$$

With the exact channel norms $n_i = \|g_i\|_{F^{-1}}$,

$$\|\widetilde{d} - d_{\widetilde{\alpha}}\|_F = \varepsilon_{\mathrm{CG}}, \qquad |g_i^\top(\widetilde{d} - d_{\widetilde{\alpha}})| \leq n_i \varepsilon_{\mathrm{CG}}. \tag{232}$$

If d and α_i denote the ideal exact-norm fusion from Equation (221)*, then also*

$$\|\widetilde{d} - d\|_F \leq \varepsilon_{\mathrm{CG}} + \sum_{i=1}^{2} |\widetilde{\alpha}_i - \alpha_i| n_i. \tag{233}$$

Proof. The residual identity gives $d_{\widetilde{\alpha}} - \widetilde{d} = F^{-1} r_{\mathrm{CG}}$. Taking the F-norm proves the equality, and dual Cauchy–Schwarz proves the slope bound. Finally, $d_{\widetilde{\alpha}} - d = \sum_i (\widetilde{\alpha}_i - \alpha_i) F^{-1} g_i$; the triangle inequality and $\|F^{-1} g_i\|_F = n_i$ prove the last bound. □

Here $\varepsilon_{\mathrm{CG}}$ is an analytical residual norm; the runtime stopping tolerance uses a Euclidean relative residual and does not directly equal this quantity. The coefficient-error term must also be controlled to transfer the exact-norm guarantees of Theorem 7 to a numerical direction. Endpoint control below evaluates the numerical candidate directly.

Finite updates and the two channel budgets. Recall that the Fuse directions in Equation (221) use exact solves at the frozen parents. Let $t_\mu, t_\pi \in [0,1]$ denote behavior and target step fractions. Because all channels are maximization objectives, the ideal updates use the ascent sign:

$$\begin{aligned}\theta^{\mu}_{g+1} &= \theta^{\mu}_{g} + t_\mu \,\mathrm{Fuse}_{F_\mu}(g_A, g_H),\\ \theta^{\pi}_{g+1} &= \theta^{\pi}_{g} + t_\pi \,\mathrm{Fuse}_{F_\pi}(g_J, g_P).\end{aligned} \tag{234}$$

For either actor, let $\theta_\star$ be its parent parameters and $\widetilde{d}$ its computed direction. Endpoint feasibility control (EFC) chooses a step fraction $t \in [0,1]$. Let FP32 denote rounding to the stored single-precision floating-point parameter representation. EFC evaluates the actual candidate $\theta(t) = \mathrm{FP32}(\theta_\star + t\widetilde{d})$. Let q be the parent policy, q_t the policy represented by $\theta(t)$, and ν one of the actor's channel probability measures. Define the averaged Pearson discrepancy $\mathcal{D}_{2,\nu}$ by

$$\begin{aligned}\mathcal{D}_{2,\nu}(q_t\|q) &= \mathbb{E}_{s\sim\nu}\sum_a \frac{(q_t(a\mid s)-q(a\mid s))^2}{q(a\mid s)}\\ &= \mathbb{E}_{s\sim\nu}\sum_a \frac{q_t(a\mid s)^2}{q(a\mid s)} - 1.\end{aligned} \tag{235}$$

Let $\varepsilon_{\mathrm{EFC}} = 0.01$ be the per-channel endpoint budget. The numerical constraint $\mathcal{D}_{2,\nu}(q_t\|q) \le \varepsilon_{\mathrm{EFC}}$ is checked separately for $\nu = D_g, \xi_g$ for behavior and $\nu = \nu_g^J, \nu_g^P$ for target. Checking both measures separately prevents one channel's small movement from concealing a large movement under the other.

EFC first tests $t = 1$. If infeasible, it initializes the feasible lower endpoint at the parent $t = 0$, the upper endpoint at 1, and performs 24 midpoint probes. A feasible probe replaces the lower endpoint and its cached parameter vector; an infeasible probe replaces the upper endpoint. The algorithm commits that exact cached feasible vector, without reconstructing it from the returned scalar. Thus the committed parameters are exactly those that passed the implemented numerical checks. This search does not require global monotonicity of feasibility and need not return the largest feasible step. The mathematical bounds below assume the stated discrepancy inequality for that endpoint; finite-precision checks alone are not interval-arithmetic certificates of the exact inequality.

Corollary 2 (Finite policy-change bounds under each channel measure)**.** *Let ν be a probability measure on observations, let q be positive on the finite action set, and let q_t be a candidate policy. Recall that t labels the endpoint step fraction. Define, at each observation, the total variation and forward Kullback–Leibler divergence by* $\mathrm{TV}(q_t, q) = \frac{1}{2}\sum_a |q_t(a) - q(a)|$ *and* $\mathrm{KL}(q_t\|q) = \sum_a q_t(a)\log(q_t(a)/q(a))$*, using natural logarithms and the convention* $0\log 0 = 0$*. If* $\mathcal{D}_{2,\nu}(q_t\|q) \le \varepsilon_{\mathrm{EFC}}$*, then*

$$\mathbb{E}_\nu\,\mathrm{TV}(q_t, q) \le \frac{\sqrt{\varepsilon_{\mathrm{EFC}}}}{2}, \qquad \mathbb{E}_\nu\,\mathrm{KL}(q_t\|q) \le \log(1+\varepsilon_{\mathrm{EFC}}). \tag{236}$$

For any frozen real action payoff $f(s,a)$ with finite integrated conditional variance $\mathbb{E}_{S\sim\nu}\,\mathrm{Var}_{a\sim q(\cdot|S)}\, f(S,a) < \infty$*,*

$$\begin{aligned}&\left|\mathbb{E}_\nu \sum_a (q_t(a\mid S) - q(a\mid S)) f(S,a)\right|\\ &\qquad \le \sqrt{\varepsilon_{\mathrm{EFC}}}\left[\mathbb{E}_\nu\,\mathrm{Var}_{a\sim q(\cdot|S)}\, f(S,a)\right]^{1/2}.\end{aligned} \tag{237}$$

Proof. Set $\bar{f}(s) = \sum_a q(a\mid s) f(s,a)$. Probability normalization allows replacing f by $f - \bar{f}$ in the policy difference. Applying Cauchy–Schwarz under the joint measure $\nu(ds)q(a\mid s)$ gives the following bound, where $\mathbb{E}_{\nu,q}$ denotes expectation under that measure:

$$\left|\mathbb{E}_{\nu,q}\left[\left(\frac{q_t}{q} - 1\right)(f - \bar{f})\right]\right| \le \sqrt{\mathcal{D}_{2,\nu}(q_t\|q)}\sqrt{\mathbb{E}_\nu\,\mathrm{Var}_q(f)}.$$

Similarly, $\mathrm{TV}(q_t, q) = \frac{1}{2}\mathbb{E}_q|q_t/q - 1|$ gives the total-variation bound. Recall the pointwise Pearson divergence $\chi^2(q_t\|q) = \sum_a (q_t(a) - q(a))^2/q(a)$, whose ν-average is $\mathcal{D}_{2,\nu}(q_t\|q)$. At each state, Jensen's inequality under q_t yields

$$\mathrm{KL}(q_t\|q) = \mathbb{E}_{q_t} \log(q_t/q) \le \log \mathbb{E}_{q_t}(q_t/q) = \log\left(1 + \chi^2(q_t\|q)\right).$$

Averaging the displayed inequality over ν and applying Jensen again proves the KL bound. □

Under the corollary's discrepancy premise with the implemented budget, the average total-variation change is at most 0.05 under each channel measure. These constraints are empirical averages, not uniform bounds at every state. EFC controls distribution movement and does not check objective improvement or the curvature-dependent step conditions in Theorem 7. The fusion result characterizes the ideal proposed direction; the residual theorem describes numerical direction error, and the endpoint corollary bounds finite policy change whenever its discrepancy premise holds.

C Experiments Protocol

We compare HBPO with nine native methods: PPO (Schulman et al., 2017), A2C (Mnih et al., 2016), Discrete SAC (Haarnoja et al., 2018; Christodoulou, 2019), ACE (Imani et al., 2018), PPO-81 (Schulman et al., 2017), TRPO (Schulman et al., 2015), GRPO (Shao et al., 2024), TPO (Kaddour, 2026), and Guided Policy Search (GPS) (Levine & Koltun, 2013; Levine et al., 2016). Each method provides a warm initialization; Cold supplies a tenth HBPO initialization.

C.1 Native baseline selection

Each method is tuned independently in DSIS and DSIS-RB using twelve joint configurations from an eight-dimensional scrambled Sobol sequence (seed 1413164337). Table 4 gives the search ranges. Learning rates use logarithmic sampling on $[10^{-5}, 10^{-3}]$ unless specified otherwise; clipping and GAE intervals use linear sampling, and discrete choices occupy equal intervals.

Each candidate trains for 500,000 real interactions. The eight non-GPS methods use tuning seed 0; GPS uses seeds $5, 6, 7$. Tuning and formal training use separate random streams even when seed labels coincide. Candidates are evaluated with sampled actions on 100 trajectories of 1,000 periods per risk level, using the cost measure defined below. Within each method and tuning seed, candidates share evaluation tapes and action-randomness streams. We select the lowest validation cost, averaged over the three training seeds for GPS, and break ties by trial identifier. Tables 5 and 6 give the selected parameters, rounded to six significant figures. These settings remain fixed for all formal seeds and their paired native/HBPO comparisons.

Table 4: Native search spaces. The DSIS-RB collection multiplier applies only to GRPO and TPO.

Method	Searched dimensions
PPO, PPO-81	Learning rate; clip ratio $[.1, .3]$; GAE $\lambda \in [.90, .99]$; epochs $\{5, 10\}$; rollout length $\{256, 512, 1024\}$; entropy coefficient $\{0, .001, .01\}$.
A2C	Learning rate; GAE $\lambda \in [.90, .99]$; rollout length $\{20, 50, 100\}$; entropy coefficient $\{0, .001, .01\}$.
Discrete SAC	Learning rate; minibatch $\{128, 256, 512\}$; target update rate $\{.005, .01, .02\}$; target entropy ratio $\{.1, .25, .5\}$.
ACE	Actor and critic learning rates; ETD $\lambda \in \{0, .5, .9\}$; behavior temperature $\{.5, 1, 2\}$.
TRPO	KL limit: log $[.003, .03]$; value learning rate; GAE $\lambda \in [.90, .99]$; rollout length $\{512, 1024, 2048\}$.
GRPO	Actor and value learning rates; clip ratio $[.1, .3]$; group size $\{4, 8\}$; actor epochs $\{1, 2, 4\}$; KL coefficient $\{0, .04\}$; RB collection multiplier $\{4, 8, 16\}$.
TPO	Actor and value learning rates; temperature η: log $[.001, 1]$; group size $\{4, 8\}$; actor epochs $\{1, 2, 4\}$; RB collection multiplier $\{4, 8, 16\}$.
GPS	Actor and dual learning rates; guide learning rate: log $[10^{-4}, 10^{-2}]$; penalty: log $[.01, 10]$; per-step KL limit: log $[.001, .1]$; inner iterations $\{2, 4, 8\}$; model horizon $\{64, 256, 1000\}$; model paths $\{8, 16, 32\}$.

Table 5: Selected native configurations for DSIS.

Method	Trial	Selected parameters
PPO	03	Learning rate 2.80429×10^{-4}; clip .214343; GAE $\lambda = .938579$; 5 epochs; rollout 512; entropy 0.
A2C	04	Learning rate 1.20204×10^{-4}; GAE $\lambda = .915092$; rollout 20; entropy .001.
Discrete SAC	11	Learning rate 1.65339×10^{-4}; minibatch 512; target update rate .02 (Polyak coefficient .98); entropy ratio .1.
ACE	01	Actor learning rate 2.09424×10^{-5}; critic/follow-on learning rate 6.92692×10^{-4}; ETD $\lambda = 0$; behavior temperature .5.
PPO-81	04	Learning rate 1.20204×10^{-4}; clip .133537; GAE $\lambda = .914021$; 5 epochs; rollout 512; entropy .001.
TRPO	03	KL limit .0158867; value learning rate 1.39131×10^{-4}; GAE $\lambda = .938579$; rollout 1024.
GRPO	07	Actor learning rate 7.73584×10^{-4}; value learning rate 3.35597×10^{-4}; clip .263714; group size 8; 4 actor epochs; KL coefficient 0.
TPO	04	Actor learning rate 1.20204×10^{-4}; value learning rate 2.16455×10^{-5}; $\eta = .00293339$; group size 4; 2 actor epochs.
GPS	01	Actor learning rate 2.09424×10^{-5}; guide learning rate .00692692; dual learning rate 1.76382×10^{-5}; penalty .0387192; KL limit .00179181; 8 inner iterations; model horizon 64; 16 model paths.

Table 6: Selected native configurations for DSIS-RB.

Method	Trial	Selected parameters
PPO	08	Learning rate 5.66943×10^{-4}; clip .119578; GAE $\lambda = .927511$; 10 epochs; rollout 256; entropy 0.
A2C	11	Learning rate 1.65339×10^{-4}; GAE $\lambda = .984611$; rollout 100; entropy 0.
Discrete SAC	04	Learning rate 1.20204×10^{-4}; minibatch 128; target update rate .005 (Polyak coefficient .995); entropy ratio .25.
ACE	10	Actor learning rate 7.52789×10^{-5}; critic/follow-on learning rate 5.70075×10^{-5}; ETD $\lambda = 0$; behavior temperature 2.
PPO-81	03	Learning rate 2.80429×10^{-4}; clip .214343; GAE $\lambda = .938579$; 5 epochs; rollout 512; entropy 0.
TRPO	00	KL limit .0173037; value learning rate 9.30087×10^{-5}; GAE $\lambda = .954778$; rollout 2048.
GRPO	07	Actor learning rate 7.73584×10^{-4}; value learning rate 3.35597×10^{-4}; clip .263714; group size 8; 4 actor epochs; KL coefficient 0; collection multiplier 4.
TPO	00	Actor learning rate 3.32686×10^{-4}; value learning rate 9.30087×10^{-5}; $\eta = .0669775$; group size 8; 4 actor epochs; collection multiplier 16.
GPS	01	Same selected numerical configuration as DSIS GPS.

C.2 HBPO SELECTION

In each environment, we evaluate the seven configurations in Table 7 on all ten initializations (nine warm sources and Cold), using tuning seeds $5, 6, 7$. Candidates share the initial learner state, first data batch, and calibration scale within each initialization and seed. Each candidate runs for 15 generations of 25,000 real interactions, including the first batch, and is validated with sampled target actions on 100 trajectories of 1,000 periods per risk level. The grid varies value learning rate, emphatic learning rate, and the heuristic scale multiplier κ, with other settings fixed. The calibration of κ is specified in Appendix B.4.6.

Table 7: The complete HBPO candidate set. H5 is used in the final experiments.

Candidate	Value learning rate	Emphatic learning rate	κ
H0	5.66943×10^{-4}	1.56958×10^{-5}	$1/4$
H1	1.79283×10^{-4}	1.56958×10^{-5}	$1/4$
H2	1.79283×10^{-3}	1.56958×10^{-5}	$1/4$
H3	5.66943×10^{-4}	4.96345×10^{-6}	$1/4$
H4	5.66943×10^{-4}	4.96345×10^{-5}	$1/4$
H5	5.66943×10^{-4}	1.56958×10^{-5}	$1/32$
H6	5.66943×10^{-4}	1.56958×10^{-5}	2

For each tuning seed, split initializations into initial-cost groups $C_0 \leq 1010$ and $C_0 > 1010$. For each candidate, compute percentage improvement $100(1-C_{15}/C_0)$, average within each group, and then average these group means across seeds with nonempty groups. Denote the two scores by P and R, respectively. Rank by each score, breaking ties by the other score and then candidate ID. Select the highest-P candidate in the intersection of the two top-three lists; if the intersection is empty, select the highest-R candidate among the P top three. If one group is empty for every seed, use the other ranking. This rule selects H5 in both environments: value learning rate 5.66943×10^{-4}, emphatic learning rate 1.56958×10^{-5}, and $\kappa = 1/32$. One configuration is shared across all initializations and formal seeds within each environment; ablations use it without retuning.

C.3 EVALUATION PROTOCOL

Budgets and seeds. Each formal run uses 2M real interactions and one of five training seeds, $0, \ldots, 4$. Warm comparisons branch from the same native 0.5M checkpoint into native continuation and HBPO, each receiving another 1.5M interactions. Cold receives 2M without a native prefix. Tuning and formal-evaluation data are separate, and formal results do not influence hyperparameter or checkpoint selection. Simulation and evaluation are accounted for separately from real training interactions.

Information access. All methods have the same public-information permissions and visible observations. DSIS's stochastic law is public; DSIS-RB's true stochastic kernels, episode risks, and future random tapes are hidden. In DSIS-RB, HBPO and GPS fit the same model class to their own budgeted observations. Modeling assumptions are given in Appendix B.1.4.

Test cost and reporting. We freeze the final policy and sample its actions; HBPO always reports the target policy π. Each policy is tested on 500 independent trajectories of 1,000 periods, including all periods from the initial state. DSIS reports mean undiscounted raw cost per period. DSIS-RB uses 500 trajectories at each $\rho \in \{0, .1, .2, .3, .4, .5\}$ and reports $C_s^{\mathrm{RB}} = \sum_{k=0}^{5} \omega_k C_s(0.1k)$, where $\omega = (.1, .2, .2, .2, .2, .1)$ and $C_s(\rho)$ is the mean cost for training seed s at risk ρ. This is normalized trapezoidal risk-AUC. We report the mean, median, and sample standard deviation across the five training seeds, using unrounded costs for comparisons. Ablations are paired with their corresponding Full run by source, seed, and evaluation bank, as detailed in Appendix E.

Cost-collapse thresholds. Mean demand is two units per period. Expedited replacement costs $2 \times 10 = 20$, and k periods of mean demand in backlog incur $2k \times 495 = 990k$ per period. The reference cost $20 + 990k$ gives 1010 for one period and 3980 for four periods. We classify costs as Non-collapse ($C \leq 1010$), Collapse ($1010 < C \leq 3980$), or Severe Collapse ($C > 3980$). These

are economic severity thresholds, not estimates of actual delivery delay. Endpoint recovery means a native 2M cost above 1010 paired with an HBPO 2M cost at or below 1010.

D Extended Results

D.1 Detailed Main Results

Tables 8 and 9 expand Table 1 using the same 190 final-policy costs: 90 native endpoints, 90 warm HBPO endpoints, and 10 Cold endpoints. Each policy is evaluated after a total budget of 2M environment interactions; warm HBPO includes the 0.5M native prefix and reports the final target policy π. H denotes warm HBPO, and Cold has no native prefix or paired native endpoint. All summaries retain all five training seeds. Evaluation protocols and cost grades follow Appendix C.3.

Precision and table conventions. Seed costs retain their full effective decimal precision: five places for DSIS and six for DSIS-RB, without k/M abbreviations. Only binary floating-point representation tails are removed. Means and medians are exact at this precision; sample standard deviations use the $n - 1 = 4$ denominator and are displayed to nine decimal places. Black, ochre, and red seed entries indicate $c \leq 1010$, $1010 < c \leq 3980$, and $c > 3980$, respectively. Bold marks the lower paired mean or median; underlining marks HBPO costs ≤ 1010 whose seed-matched native endpoints exceed 1010. Comparisons use the full-precision values, not the abbreviated main-table entries.

DSIS. HBPO lowers mean cost in all nine warm-source comparisons, median cost in eight (TRPO is the exception), and sample standard deviation in all nine. It lowers cost in 40 of 45 seed-matched comparisons. All 12 native endpoints above 1010 have lower-cost HBPO counterparts, and 10 cross to costs ≤ 1010. All 33 native endpoints at or below 1010 remain within this range under HBPO, with 28 achieving lower costs. The two HBPO endpoints still above 1010 are A2C seed 2 (6638.13434) and ACE seed 0 (1207.57231).

DSIS-RB. HBPO lowers mean cost in eight of nine warm-source comparisons, median cost in seven (PPO and A2C are the exceptions), and sample standard deviation in all nine. It lowers cost in 41 of 45 seed-matched comparisons. Among the 24 native endpoints above 1010, 23 have lower-cost HBPO counterparts and 19 cross to costs ≤ 1010. All 21 native endpoints at or below 1010 remain within this range under HBPO, with 18 achieving lower costs. Discrete SAC is the mean-cost exception: 2479.0853802 increases to 2514.9923076 (1.4484%), although its median and standard deviation decrease. Its seed 0 worsens from 1183.646256 to 5359.057081, moving from Collapse to Severe Collapse. The five HBPO endpoints still above 1010 are PPO seed 2, A2C seed 2, Discrete SAC seeds 0 and 1, and ACE seed 0.

Cold starts and scope of the comparisons. All five Cold endpoints are at or below 1010 in each environment. DSIS has mean 45.103804 and range [42.67906, 47.51533]; DSIS-RB has mean 72.1150812 and range [69.675982, 74.916383]. Warm comparisons pair native and HBPO endpoints by environment, source, training seed, and interaction budget; they do not assume identical evaluation random banks across campaigns. Crossing the threshold relative to a native 2M endpoint is distinct from recovery relative to the 0.5M initialization. These descriptive results characterize the evaluated seeds and do not establish statistical significance or a task-independent stability guarantee.

Table 8: Detailed final-policy costs on DSIS (mean cost per period; lower is better), expanding Table 1. Panel (a) retains five decimal places for every seed; panel (b) reports exact means and medians, with sample standard deviations displayed to nine decimal places. Formatting follows the conventions in Appendix D.1.

(a) Individual training seeds

Policy	Seed 0	Seed 1	Seed 2	Seed 3	Seed 4
PPO	21.00574	21.16667	56.89301	13817.30298	21.33039
H-PPO	20.73032	21.25826	51.41226	158.87953	20.75846
A2C	23.65804	28.68256	23012.57348	30.43475	23.38418
H-A2C	32.28773	20.76730	6638.13434	20.88466	20.74205
Discrete SAC	27.82865	36.65902	38.56467	27.45829	345.08867
H-Discrete SAC	26.56887	33.98975	33.06797	32.39846	55.35698
ACE	1679.44062	22.57928	22.34593	7504.31927	1116.35635
H-ACE	1207.57231	319.84945	21.40372	77.84574	37.53063
PPO-81	22.35240	22.70225	21.70040	22.96933	21.49762
H-PPO-81	21.37053	22.01671	21.16084	20.89490	21.48099
TRPO	5026.80023	12492.90475	26.03542	29.66418	25.40624
H-TRPO	102.58326	219.28574	25.62822	30.75901	25.13818
GRPO	37.64513	32.93183	32.32662	50.67199	30.85771
H-GRPO	22.46585	26.81074	20.99287	23.23004	21.89809
TPO	4207.30851	4804.99275	4560.12880	4171.18351	4399.49216
H-TPO	45.11618	44.84068	45.91967	44.60019	45.65733
GPS	123.18311	120.45602	106.86388	117.79971	124.03338
H-GPS	39.81263	41.73971	41.15723	44.38921	35.46236
Cold	44.56601	47.51533	45.09705	45.66157	42.67906

(b) Across all five seeds

Policy	Mean	Median	Std
PPO	2787.539758	21.33039	6165.844492000
H-PPO	**54.607766**	**21.25826**	59.767170003
A2C	4623.746602	28.68256	10279.667193796
H-A2C	**1346.563216**	**20.88466**	2958.082369644
Discrete SAC	95.119860	36.65902	139.827358107
H-Discrete SAC	**36.276406**	**33.06797**	11.055155831
ACE	2069.008290	1116.35635	3121.642702425
H-ACE	**332.840370**	**77.84574**	503.623773687
PPO-81	22.244400	22.35240	0.632540611
H-PPO-81	**21.384794**	**21.37053**	0.418009117
TRPO	3520.162164	**29.66418**	5463.192396034
H-TRPO	**80.678882**	30.75901	84.111535246
GRPO	36.886656	32.93183	8.114700906
H-GRPO	**23.079518**	**22.46585**	2.240041009
TPO	4428.621146	4399.49216	262.226888622
H-TPO	**45.226810**	**45.11618**	0.552130068
GPS	118.467220	120.45602	6.932895316
H-GPS	**40.512228**	**41.15723**	3.276596300
Cold	45.103804	45.09705	1.753226456

Table 9: Detailed final-policy costs on DSIS-RB (normalized risk-AUC cost; lower is better), expanding Table 1. Panel (a) retains six decimal places for every seed; panel (b) reports exact means and medians, with sample standard deviations displayed to nine decimal places. Formatting follows the conventions in Appendix D.1.

(a) Individual training seeds

Policy	Seed 0	Seed 1	Seed 2	Seed 3	Seed 4
PPO	70.061774	75.460918	66970.754322	75.131193	74.882238
H-PPO	52.679426	81.929137	5893.013522	55.296788	112.014382
A2C	74.484857	69.637995	8036.258056	807.481221	69.608571
H-A2C	51.574127	51.333329	7397.851829	125.837686	653.867411
Discrete SAC	1183.646256	10969.145360	71.162997	69.717788	101.754500
H-Discrete SAC	5359.057081	7028.880845	61.773085	58.874363	66.376164
ACE	9955.781036	12184.555697	22312.909190	13801.992411	15519.232290
H-ACE	4013.167881	109.714766	797.749040	121.330187	85.117775
PPO-81	191.330923	103.245154	96.113375	95.326541	95.308221
H-PPO-81	124.935134	80.015061	86.630289	84.824489	84.067253
TRPO	427.649648	365.795013	317.922635	1996.454276	292.815271
H-TRPO	316.810653	289.684982	303.129556	235.322284	166.282749
GRPO	537.104180	1243.313035	1102.386111	3234.895086	1215.766356
H-GRPO	80.778016	92.341207	107.239713	93.158126	222.144895
TPO	11220.857853	1616.815238	2856.305500	1994.682157	3392.571200
H-TPO	238.316432	199.456230	219.981811	214.844334	163.411692
GPS	1767.627848	6215.901052	2410.230847	1922.649659	2013.114622
H-GPS	233.994143	200.358208	188.219671	239.709108	215.895433
Cold	69.675982	73.103120	74.916383	70.041407	72.838514

(b) Across all five seeds

Policy	Mean	Median	Std
PPO	13453.2580890	**75.131193**	29917.189972738
H-PPO	**1238.9866510**	81.929137	2601.790788363
A2C	1811.4941400	**74.484857**	3494.322436684
H-A2C	**1656.0928764**	125.837686	3219.614352805
Discrete SAC	**2479.0853802**	101.754500	4770.066181326
H-Discrete SAC	2514.9923076	**66.376164**	3409.927238739
ACE	14754.8941248	13801.992411	4697.178735828
H-ACE	**1025.4159298**	**121.330187**	1696.948305507
PPO-81	116.2648428	96.113375	42.095435700
H-PPO-81	**92.0944452**	**84.824489**	18.517339470
TRPO	680.1273686	365.795013	737.643085494
H-TRPO	**262.2460448**	**289.684982**	61.918285399
GRPO	1466.6929536	1215.766356	1029.108387150
H-GRPO	**119.1323914**	**93.158126**	58.344879455
TPO	4216.2463896	2856.305500	3977.588834515
H-TPO	**207.2020998**	**214.844334**	28.139939341
GPS	2865.9048056	2013.114622	1887.677624216
H-GPS	**215.6353126**	**215.895433**	21.804066539
Cold	72.1150812	72.838514	2.213379602

D.2 Additional Transmission Analyses

Figures 5–8 complete the transmission analysis in Section 5.3. They show the remaining 16 source–environment combinations: A2C, Discrete SAC, ACE, PPO-81, TRPO, GRPO, TPO, and GPS on DSIS and DSIS-RB. Together with the four Cold and PPO combinations in Figure 2, these cover all ten initializations in both environments. Each combination contains five training seeds; HBPO-SAC denotes the Discrete SAC initialization.

Reading the panels. Within each combination, the left panel shows standard evaluation costs for μ (blue) and π (ochre), together with the green dashed q_H reference. The latter averages H-weighted immediate costs over the original nine-step rollouts and is not a comparable long-horizon policy evaluation. The middle panels show relative step ϱ and direction cosine c, with purple for $q_H \to \mu$ and rose for $\mu \to \pi$. The right panels show adoption and reflection: solid teal curves are net responses $G_{\pm}$, gray dashed curves are opportunities $B_{\pm}$, and pale green/pink backgrounds are mean favorable and adverse contributions above/below zero. Positive G_+ and G_- denote favorable local responses in the adoption and reflection regions, respectively.

Aggregation and timing. Curves are arithmetic means across seeds; cost panels also retain individual seed traces, while mechanism bands show the 25th–75th percentiles. The signed contribution backgrounds are not uncertainty bands. Statistics are computed before axis transformations, without smoothing or subsampling. Costs use logarithmic axes; relative steps and selection responses use inverse-hyperbolic-sine axes with linear scales 10^{-2} and 10^{-6}, respectively. Cost-axis limits are shared within each environment; mechanism limits are shared across all combinations. These limits include the extrema from all 20 combinations. Costs retain checkpoint time t; mechanism panels pair the $q_H \to \mu$ update at g with the $\mu \to \pi$ update and selection responses at $g + 1$, following $\pi_g \to q_{H,g} \to \mu_{g+1} \to \pi_{g+2}$. All panels here are warm starts and display generation plus 20. These are descriptive diagnostics of the recorded trajectories, rather than additional intervention experiments.

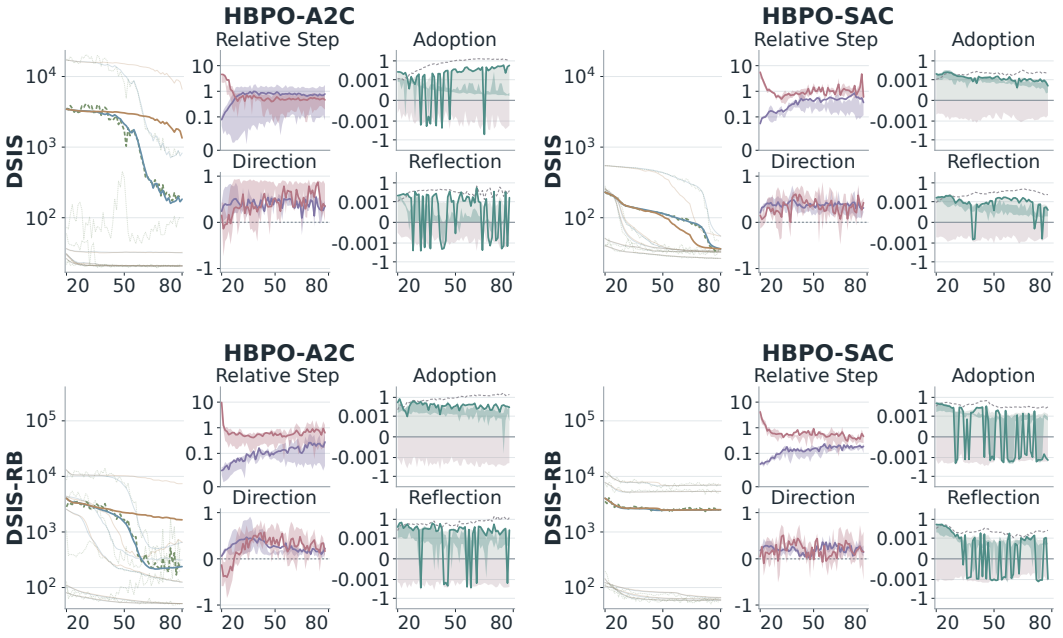

Figure 5: Additional transmission analysis for A2C (left) and Discrete SAC (right). Top: DSIS; bottom: DSIS-RB. Conventions follow Appendix D.2.

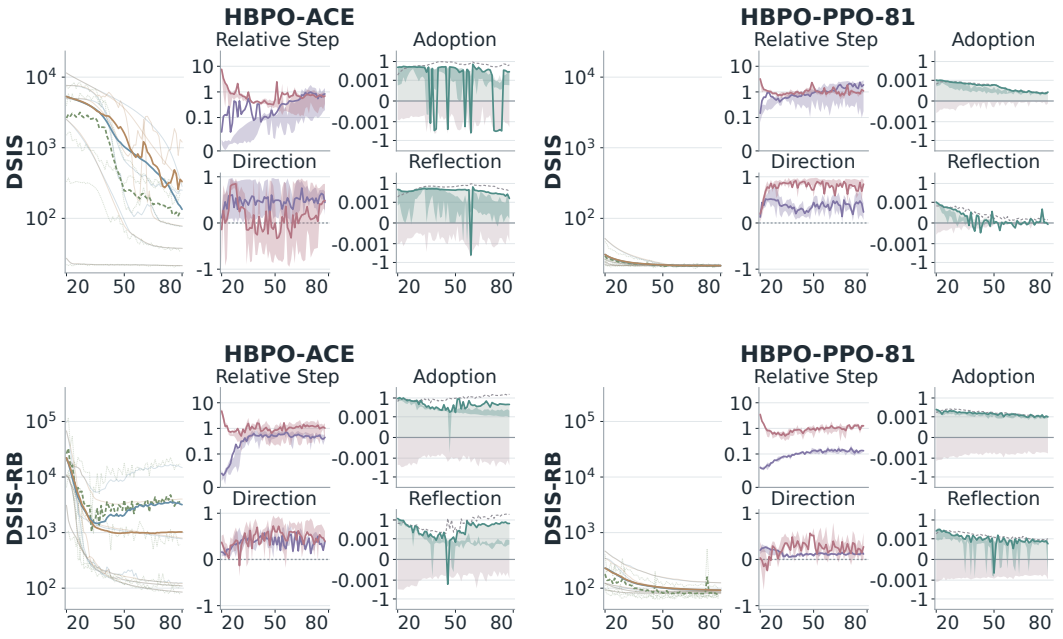

Figure 6: Additional transmission analysis for ACE (left) and PPO-81 (right). Top: DSIS; bottom: DSIS-RB. Conventions follow Appendix D.2.

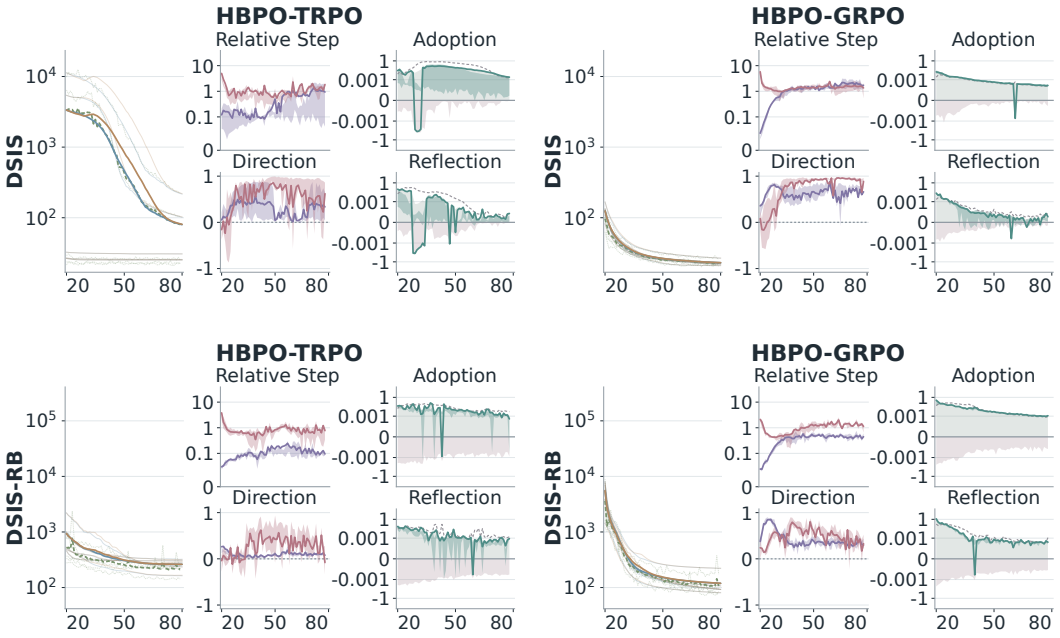

Figure 7: Additional transmission analysis for TRPO (left) and GRPO (right). Top: DSIS; bottom: DSIS-RB. Conventions follow Appendix D.2.

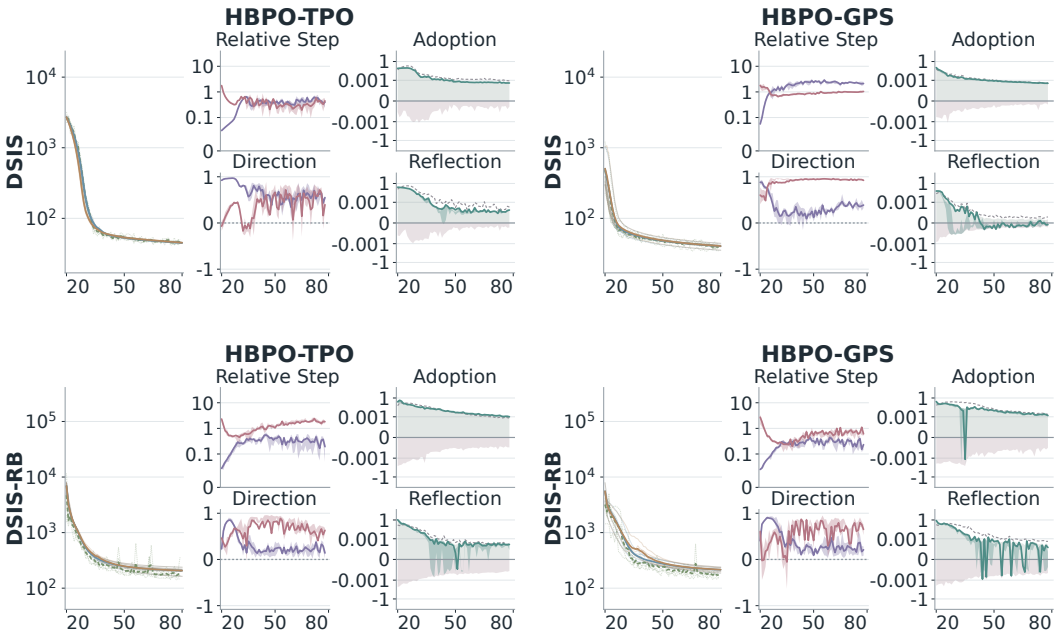

Figure 8: Additional transmission analysis for TPO (left) and GPS (right). Top: DSIS; bottom: DSIS-RB. Conventions follow Appendix D.2.

D.3 Emphatic Weighting and Final Cost

Figure 9 shows that emphatic weighting does not provide a clear, consistent improvement in final target-policy cost after the full 2M-interaction training budget. We compare Full HBPO with No-E, J-only E, and P-only E across Cold, PPO, Discrete SAC, and TRPO initializations in both environments, retaining all five training seeds. The interventions change the emphatic weighting branches while retaining the J and P policy objectives (Appendix E.4). Against No-E, Full achieves a lower final cost in only 19 of 40 matched runs and a lower mean in four of the eight environment–initialization settings. The DSIS Cold mean favors Full, but this difference is driven mainly by one seed; J-only E and P-only E likewise show mixed endpoint effects. Thus, these results do not support a general long-run final-cost advantage from emphatic weighting in the evaluated settings. This is a descriptive comparison, not a formal significance or equivalence test, and it does not determine whether weighting affects earlier learning.

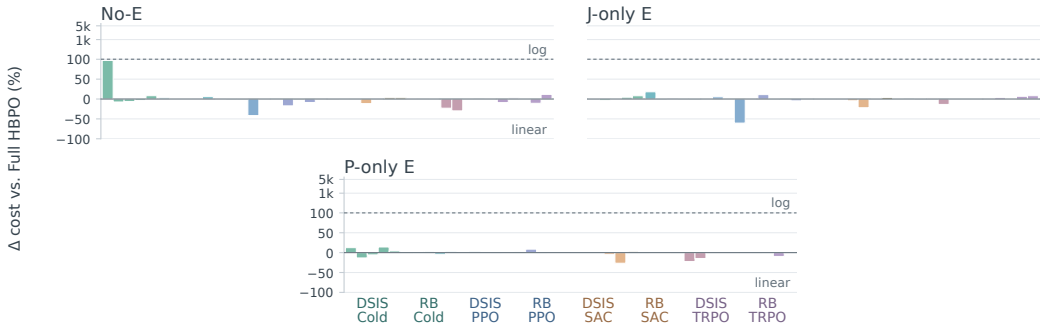

Figure 9: Emphatic-weight ablations at the final target-policy endpoint. No-E disables both emphatic branches; J-only E and P-only E retain only the indicated weighting branch. Each bar gives the percentage cost change relative to matched Full HBPO; positive values favor Full. Each environment–initialization block contains seeds 0–4 from left to right; darker bars mark ablation costs above 1010. All panels share the group order shown below and the H/P ablation scale, linear up to 100% and logarithmic above. All emphatic-ablation bars lie in the linear region.

E Ablation Details and Additional Diagnostics

This appendix specifies the nine implemented interventions and the additional diagnostics underlying the component and transmission analyses. The interventions measure changes in the complete learning process at a fixed setting; the observational diagnostics describe actual updates.

E.1 Matched ablation panel and reporting

The ablation panel comprises PPO, Discrete SAC, TRPO, and Cold initializations in each environment, with training seeds 0–4. Each intervention therefore has 40 matched runs. We evaluate two heuristic interventions, four coupling interventions, and three emphatic interventions, giving 360 intervention endpoints and 40 shared Full references. These are separate changes to Full, not a factorial combination of multiple intervention families. Every variant uses the selected H5 learning rates and $\kappa = 1/32$ without retuning. Warm runs use the same source checkpoint after 0.5M interactions and another 60 generations of $25{,}000$ real transitions; Cold uses 80 generations. Thus the final comparison is at 2M real interactions. The source policy, initialization procedure, scientific seed, and initial real-data batch are matched to the corresponding Full run. Subsequent policies, collected data, fitted models, and simulation work can differ as a consequence of the intervention.

The final target policy is evaluated by sampled actions under the formal protocol of Appendix C: 500 episodes of $1{,}000$ periods in DSIS, and the same number at each of the six risk levels in DSIS-RB, followed by normalized trapezoidal risk aggregation. These measurements do not feed back into training. The endpoint record contains all five seeds in every cell. It combines qualified historical DSIS runs for SAC, TRPO, and Cold with the subsequent DSIS PPO and DSIS-RB runs. Each variant is compared with the Full reference evaluated on its own matched bank. In particular, later Full reruns or reevaluations used for the main comparison and generation diagnostics do not replace these ablation denominators. This preserves within-comparison pairing when the historical and subsequent evaluation banks differ.

For an environment–initialization–seed coordinate i, the plotted change for variant v is

$$\Delta_{v,i} = 100 \left(\frac{c_{v,i}}{c_{\mathrm{Full},i}} - 1 \right). \tag{238}$$

Positive values favor Full. An average paired percentage is the arithmetic mean of these ratios; it is distinct from the percentage change between two group means. Dark bars in the main ablation figures indicate that the *variant's* final cost exceeds 1010. Across-seed dispersion uses the sample standard deviation of the five training-seed costs, rather than the standard error over evaluation episodes. The ablation outcome evidence is an endpoint comparison; it does not establish learning-curve smoothness, faster convergence, or a training-curve area under the curve. Real interaction budgets are matched, while model simulation and numerical work are accounted for separately and need not be equal after a component is removed.

E.2 Heuristic-guidance interventions

No-H. This intervention deletes the entire finite heuristic objective, so that $H(\theta) \equiv 0$ and $g_H = 0$, and removes the associated H-occupancy construction and confidence calibration. The H state channel in the behavior Fisher and endpoint constraints is replaced by the complete real-data panel D_g with unit weights. Consequently, the two behavior channels use the same data measure and their averaged Fisher reduces to the A-channel Fisher. The guidance-derived P interest is set to one at the root and successor states. The A, J, and P objectives, complete residual corrections, and remaining emphatic learning are retained; in particular, setting P interest to one does not disable the P emphatic learner. The initial utility-scale calibration for τ can still require model queries. No-H therefore tests the total contribution of the guidance module, including its state weighting and geometry, rather than only setting a single gradient vector to zero.

Random-H. At each queried public state, we construct a random target around this variant's own frozen parent behavior policy. Let $q_{H,g}^{\mathrm{shadow}}$ be the simulation-derived controller at that same state under the same parent policies and model. Draw a Gaussian action vector $z_g(s, \cdot)$ and choose

$\beta_g(s) \geq 0$ so that

$$q_{R,g}(a \mid s) = \frac{\mu_g(a \mid s) \exp(\beta_g(s) z_g(s,a))}{\sum_b \mu_g(b \mid s) \exp(\beta_g(s) z_g(s,b))}, \tag{239}$$

$$d_{\mathrm{FR}}(\mu_g, q_{R,g}) = d_{\mathrm{FR}}(\mu_g, q_{H,g}^{\mathrm{shadow}}). \tag{240}$$

For zero radius, the target is μ_g. For a positive radius, directions are rejected only if their limiting maximum-action vertex cannot reach that radius; returns are not used to select a random direction. A bracketed scalar solve finds the radius-matching $\beta_g(s)$. The random field is addressed by environment, initialization, training seed, generation, and public state, so repeated visits to the same state within a generation use the same direction. Hidden risk does not enter the address.

The random controller replaces the informative controller in both occupancy rollouts and the paired calibration rollouts. The original confidence-to-interest construction is retained. Because the shadow controller still sets the local radius, Random-H retains model and return information about the magnitude of guidance. The matching is local to this arm's parent policy and visited states; it does not force the Random-H and Full trajectories to visit the same states or to have identical future exploration. This comparison tests the value of the informative guidance direction within the resulting learning process.

E.3 Behavior–target coupling interventions

No-A. We remove the complete A channel from behavior learning, including both the dense base term and sampled residual correction, and set its finite objective to zero. The remaining H channel is passed through the same fusion rule. The original behavior Fisher, both state-channel endpoint constraints, target update, and critic construction are retained. The removed signal is target-continuation advantage consumed by the behavior actor.

No-P. We set the explicit P receiver gradient and finite P objective to zero and recompute the target fusion with the unchanged J channel. The original J and P state measures still define the target Fisher and endpoint constraints. Behavior continues to generate the real training data and can influence the target through data, critics, state weights, and geometry. Thus No-P removes the explicit transfer objective; it does not remove every effect of behavior on target learning.

KL-P. We replace P by forward-KL imitation of the parent behavior policy. Up to a constant independent of the candidate target parameters, the maximization objective is

$$K_g(\theta) = \frac{1}{|D_g|} \sum_{s \in D_g} w_g^P(s) \sum_a \mu_g(a \mid s) \log \pi_\theta(a \mid s). \tag{241}$$

This replacement has no advantage filter. Write $g_K = \nabla K_g$, let g_P be the complete original P gradient including the residual correction, and retain the original target Fisher F_π. For positive norms, the receiver is

$$\widetilde{g}_K = \frac{\sqrt{g_P^\top F_\pi^{-1} g_P}}{\sqrt{g_K^\top F_\pi^{-1} g_K}} g_K. \tag{242}$$

The same frozen scale multiplies the finite KL objective. If the P reference norm is zero, the replacement is zero; a zero KL gradient with a nonzero reference norm is recorded as an unmatchable direction and also yields zero. The scaled receiver and J are fused using the usual target geometry and endpoint constraints. This matches the parent natural-gradient norm, not the finite policy displacement.

Near-on-policy. This arm maintains $\mu_g = \pi_g$ at committed generation boundaries. It first computes the accepted AH behavior successor as a temporary guide $\widetilde{\mu}_{g+1}$, then updates the target using J and norm-matched KL imitation of this fresh guide. For this matching, the reference P objective also uses $\widetilde{\mu}_{g+1} - \pi_g$ as its teacher gap. The real-data panel, critic evidence, and residual corrections remain those of the frozen parent generation; no new real batch is collected between the two actor updates. After the target update, an independent copy of π_{g+1} is committed as μ_{g+1}. This

intervention jointly changes the receiver objective, transmission timing, and persistent separation of the policies. It is therefore an alternative coupling architecture rather than a pure delay-only intervention, and it is not an implementation of GPS.

E.4 EMPHATIC-WEIGHT INTERVENTIONS

The three emphatic interventions retain the H, A, J, and P objectives but change which learned state-weight branches are consumed:

Variant	J emphatic branch	P emphatic branch
Full	enabled	enabled
No-E	disabled	disabled
J-only E	enabled	disabled
P-only E	disabled	enabled

A disabled branch uses a neutral mass network with zero output head and unit root mass, forces both its raw mass and consumed weight to one, and skips that branch's learner update. These unit weights are applied before the residual-root sampling, heuristic construction, critic evidence, actor objectives, and Fisher readers consume them. Disabling P emphatic additionally sets its root and successor interest to one. J-only E and P-only E refer to weighting branches; both J and P policy objectives remain present in either case. The endpoint effects include all subsequent changes in sampling and optimization, so they do not independently identify a distribution correction mechanism.

E.5 TRANSMISSION DIAGNOSTICS AND THEIR TIMING

The transmission analysis follows the actual committed policy sequence, $\pi_g \to q_{H,g} \to \mu_{g+1} \to \pi_{g+2}$. Both actors in an ordinary generation read frozen parents: the target update that produces π_{g+1} uses μ_g, not the simultaneously produced μ_{g+1}. We therefore pair the heuristic–behavior measurements from update $g \to g+1$ with the behavior–target measurements from update $g+1 \to g+2$. The latter use the residual evidence, weights, and advantage of generation $g+1$. Parent and child checkpoint identities are checked before joining the two updates; no link crosses runs or training seeds.

Evidence population. The complete diagnostic population contains both environments, all nine warm sources and Cold, and five seeds. Warm runs have 60 updates and Cold runs have 80, giving 6,200 generation packets, 6,300 checkpoints, and 6,100 complete two-generation links. Each generation retains the original 1,152 heuristic fitting occurrences, the 128 residual-root draws, all 81 actions, and all 32 residual tapes per draw. Offline contraction of these arrays introduces no further environment interactions or model rollouts.

Occurrence measures. The two links are measured on different panels. Write $\{s_h^H, \omega_h^H\}_{h=1}^{1152}$ for the original heuristic fitting occurrences and their training weights. For residual draw k, let i_k be its sampled batch index and $\varpi_i = (1 + w_i^J + w_i^P)/(3n)$ its proposal probability. The retained importance factors are $\iota_k^J = w_{i_k}^J/(n\varpi_{i_k})$ and $\iota_k^P = w_{i_k}^P/(n\varpi_{i_k})$. Define the diagnostic contractions

$$\langle x, y\rangle_H = \frac{1}{1152}\sum_h \omega_h^H \sum_a x_h(a) y_h(a), \tag{243}$$

$$\langle x, y\rangle_J = \frac{1}{128}\sum_k \iota_k^J \sum_a x_k(a) y_k(a), \qquad \langle z\rangle_P = \frac{1}{128}\sum_k \iota_k^P \sum_a z_k(a). \tag{244}$$

The weights are those supplied to the corresponding training packet. In particular, the finite residual sample is not self-normalized by its realized sum of importance weights. These diagnostics are occurrence-based Euclidean contractions in action-probability space, not the parameter-space Fisher norm used by the actor optimizer. Repeated states and repeated root draws retain their original multiplicities.

Step size and direction. On the heuristic panel of generation g, set $d_H = q_{H,g} - \mu_g$, $u_H = \mu_{g+1} - \mu_g$. On the residual panel of generation $g+1$, set $d_J = \mu_{g+1} - \pi_{g+1}$, $u_J = \pi_{g+2} - \pi_{g+1}$. For either link $L \in \{H, J\}$, compute

$$\varrho_L = \frac{\sqrt{\langle u_L, u_L\rangle_L}}{\sqrt{\langle d_L, d_L\rangle_L}}, \qquad c_L = \frac{\langle d_L, u_L\rangle_L}{\sqrt{\langle d_L, d_L\rangle_L \langle u_L, u_L\rangle_L}}. \tag{245}$$

The original projection statistic is $\langle d_L, u_L\rangle_L / \langle d_L, d_L\rangle_L = \varrho_L c_L$, whenever both factors are defined. We show the factors separately to distinguish update magnitude from directional agreement. A zero guidance gap makes both displayed statistics undefined; a zero update with a nonzero gap gives $\varrho_L = 0$ and an undefined cosine. No small-gap threshold or zero-filling is introduced. Stable, scaled contractions preserve very small nonzero gaps. Larger relative steps and positive cosines describe geometric following, not necessarily better policy value.

Adoption and reflection. At each residual draw of generation $g+1$, reconstruct the full frozen sampled advantage from the dense base and every saved residual tape:

$$\widehat{a}_k(a) = a_k^0(a) + \frac{1}{32}\sum_{b=1}^{32} \Delta a_{kb}(a). \tag{246}$$

Here Δa_{kb} is the stored action-centered residual sum over a geometrically stopped continuation; discounting is represented by that stopping law. No tape selection, refitting, clipping, or additional centering is applied during postprocessing. With $d = d_J$, $u = u_J$, define actionwise masks $S_+ = \mathbf{1}\{d\widehat{a} > 0\}$ and $S_- = \mathbf{1}\{d\widehat{a} < 0\}$. Opportunities and responses are

$$B_+ = \langle [d\widehat{a}]_+ \rangle_P, \qquad B_- = \langle [-d\widehat{a}]_+ \rangle_P, \tag{247}$$

$$G_+ = \langle u\widehat{a} S_+ \rangle_P, \qquad G_- = \langle u\widehat{a} S_- \rangle_P. \tag{248}$$

Positive G_+ denotes favorable target movement on advantageous behavioral deviations; positive G_- denotes favorable correction of adverse deviations. Negative responses are retained. In each region, the plotted favorable and adverse components are

$$F_\pm = \langle [u\widehat{a} S_\pm]_+ \rangle_P, \qquad L_\pm = \langle [-u\widehat{a} S_\pm]_+ \rangle_P, \qquad G_\pm = F_\pm - L_\pm. \tag{249}$$

The shaded areas show $F_\pm$ above zero and $-L_\pm$ below zero. The neutral region $d\widehat{a} = 0$ is retained in the audit, so its response together with $G_+ + G_-$ recovers the total sampled target advantage change. When $B_\pm = 0$, that opportunity is inactive; no response-to-opportunity ratio is imputed. These are signed sums under a frozen sampled advantage, without division by τ. They are neither raw-cost improvements nor estimates of a separately identified causal mediation effect.

Costs and presentation. The μ and π curves use the standard stochastic 500-tape, 1,000-period evaluator, at every saved checkpoint. In contrast,

$$C_H(g) = \frac{1}{1152}\sum_h \omega_h^H c_h \tag{250}$$

is the realized immediate raw cost on the original controller occupancy panel: 16 anchors, eight paths per anchor, and nine stages per path. For DSIS-RB these paths use the generation-frozen fitted model; for DSIS they use the public model. Thus C_H is a short-rollout reference from training, not a standard-reset, long-run evaluation of a third policy. It uses neither a new rollout nor an additional bootstrap value, and has no fabricated terminal-generation point.

Costs are indexed by their own checkpoint generations. Link geometry and responses are indexed by the origin g of the complete delayed chain; the last heuristic update has no subsequent target update and is omitted from these linked curves. Warm axes add 20 generation equivalents for the 500,000-interaction native prefix; Cold axes begin at zero. All statistics are computed per seed before aggregation. The main figure shows arithmetic means over all five seeds, pointwise 25th–75th percentile bands for the link and response curves (linear interpolation), and individual seed cost curves. These bands describe seed dispersion, not confidence intervals. Undefined seed-level values leave a gap in the five-seed mean; they do not reduce its denominator. There is no smoothing or temporal subsampling. Logarithmic cost axes and inverse-hyperbolic-sine step/response axes alter only display coordinates, not the underlying means or signed components.