IS ONE STEP ENOUGH FOR OFFLINE POLICY IMPROVEMENT?

Anonymous authors
Paper under double-blind review

ABSTRACT

Behavior regularization in offline reinforcement learning limits the exploitation of critic errors, but strong anchoring can also restrict policy improvement. We study how policy improvement is composed through multi-step proximal policy improvement (MPI), which re-centers each proximal objective on the preceding policy. We parameterize the procedure by a nominal total horizon T and K stages with local horizon T/K, distinguishing subdivision at a fixed total horizon from additional refinement at a common local horizon. Our analysis shows that sequential re-centering can reach endpoints unavailable to any single proximal step and characterizes how subdivision reduces the leading local discretization error of ideal updates under a fixed critic. We consider TD3+BC and IQL-based policy extraction to examine how improvement composition interacts with actor objectives and policy geometry. TD3+BC experiments on D4RL locomotion suggest that subdivision can broaden the range of useful total horizons, while adding refinement stages at a fixed small local horizon can improve return. The results identify improvement composition as a design choice alongside regularization strength, with distinct effects from horizon subdivision and additional policy extraction.

1 INTRODUCTION

Offline reinforcement learning improves a policy using a fixed dataset and learned value estimates. Behavior regularization limits the use of actions on which these estimates may be unreliable (Fujimoto et al., 2019; Kumar et al., 2019; Wu et al., 2019). Its coefficient sets the trade-off between value maximization and proximity to dataset actions. Strong anchoring can leave useful improvements unrealized; weak anchoring can lead the policy into regions where the critic extrapolates poorly.

Regularization strength alone does not specify how improvement is composed. In TD3+BC, the actor trades critic value against distance from dataset actions throughout training. We study *multi-step proximal policy improvement* (MPI): its first deterministic actor uses the same data anchor, and each later actor is anchored to the preceding actor. These are separately trained actors; later actors use their freshly updated predecessor as the reference during training.

This construction asks whether a given improvement budget works better as one actor step or as several smaller steps. We also examine longer chains built by adding steps of the same size. Under an ideal fixed critic, a smooth strongly convex example shows that two stages can reach an endpoint unavailable to any single proximal step. For smooth ideal updates and a small total budget, subdivision reduces the leading error relative to the corresponding policy flow.

We test two deterministic actor updates and Gaussian policy extraction with an IQL critic. On nine D4RL locomotion tasks, deeper TD3+BC chains retain high returns over a wider range of tested total horizons. A matched control that anchors later actors to the first actor finds both gains and losses from re-centering across tasks. The Gaussian experiments also show depth-dependent gains and failures across extraction rules.

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.

2 BACKGROUND

2.1 OFFLINE REINFORCEMENT LEARNING

Consider a discounted Markov decision process $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, P, r, \gamma \rangle$. Offline reinforcement learning seeks a policy maximizing

$$J(\pi) = \mathbb{E}_\pi \left[\sum_{t \geq 0} \gamma^t r(s_t, a_t) \right]$$

using only a fixed dataset $\mathcal{D} = \{(s_i, a_i, r_i, s'_i)\}_{i=1}^N$, without further environment interaction (Levine et al., 2020). Because the learned policy may select actions that are weakly represented in the dataset, its value estimates can rely on extrapolation; when such estimates enter Bellman backups, the resulting error can propagate through learning (Fujimoto et al., 2019; Kumar et al., 2019).

A common strategy is therefore to regularize policy improvement toward data-supported actions. Behavior regularization can be imposed directly on the actor, while conservative value-learning methods instead modify the critic to discourage unsupported actions (Wu et al., 2019; Kumar et al., 2020). In this work, we keep the value-learning procedure of the base algorithm and study the actor-side improvement step.

2.2 ACTOR-SIDE POLICY IMPROVEMENT

Let $\widehat{Q}$ denote a critic learned from $\mathcal{D}$, and let $\rho_\mathcal{D}$ be the dataset state marginal. An actor update converts the learned value information into a policy while controlling how far the policy moves from the data-supported region. The particular extraction rule can materially affect offline RL performance even when the value-learning procedure is fixed (Park et al., 2024).

We consider two extraction patterns used later in the paper. TD3+BC directly maximizes a learned critic while penalizing squared deviation from dataset actions (Fujimoto & Gu, 2021). IQL learns Q and V from dataset transitions and extracts a stochastic policy by advantage-weighted regression, with $A(s, a) = Q(s, a) - V(s)$ (Kostrikov et al., 2022). We study how locally regularized actor improvement composes across stages while retaining each base algorithm's value-learning procedure.

2.3 PROXIMAL UPDATES AND POLICY GEOMETRY

A proximal update balances optimization of an objective against movement from a reference point. Given an objective F, a reference ν, a metric d, and a step size $h > 0$, a metric proximal step takes the form

$$x^+ \in \arg\min_x \left\{ F(x) + \frac{1}{2h} d^2(x, \nu) \right\}.$$

Smaller h penalizes movement from ν more strongly. In Euclidean coordinates, a differentiable energy generates a gradient flow with two familiar time discretizations:

$$\dot{x}_t = -\nabla F(x_t), \qquad x^{\mathrm{I}}_{k+1} = x_k - h \nabla F(x^{\mathrm{I}}_{k+1}), \qquad x^{\mathrm{E}}_{k+1} = x_k - h \nabla F(x_k). \tag{1}$$

For Euclidean d, the proximal problem above is the *implicit* (backward Euler) step: its gradient is evaluated at the new point. The *explicit* (forward Euler) step evaluates it at the current reference. Repeating metric proximal steps gives the minimizing-movement construction for gradient flows (Parikh & Boyd, 2014; Ambrosio et al., 2005); the Wasserstein JKO scheme is an example (Jordan et al., 1998). The implicit and explicit rules use the same vector field but evaluate it at different actions; this distinction defines the deterministic actor updates in Section 4.1. The metric fixes policy movement: KL/Fisher geometry yields local trust-region movement (Schulman et al., 2015), and Wasserstein geometry uses transport cost (Zhang et al., 2018; Terpin et al., 2022; Pfau et al., 2025).

3 SINGLE-STEP PROXIMAL IMPROVEMENT

Hold a learned critic $\widehat{Q} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ fixed. We define the critic-induced policy energy

$$\mathcal{E}_{\widehat{Q}}(\pi) = -\mathbb{E}_{s \sim \rho_\mathcal{D},\, a \sim \pi(\cdot|s)} \widehat{Q}(s, a). \tag{2}$$

Minimizing this energy corresponds to maximizing the learned critic over dataset states. For a state distribution ρ and a statewise policy metric d, we lift d to the policy space by

$$\mathsf{d}_\rho^2(\pi, \nu) := \mathbb{E}_{s\sim\rho} d^2(\pi(\cdot \mid s), \nu(\cdot \mid s)) \,. \tag{3}$$

The statewise metric specifies how two action distributions are compared at a fixed state, while d_ρ aggregates that movement over the states relevant to the actor objective.

Given a reference policy ν, a *single proximal improvement* (SPI) stage with coefficient $h > 0$ is

$$\mathcal{P}_h^{\widehat{Q}}(\nu) \in \arg\min_{\pi\in\Pi} \left\{ \mathcal{E}_{\widehat{Q}}(\pi) + \frac{1}{2h} \mathsf{d}_{\rho_\mathcal{D}}^2(\pi, \nu) \right\}. \tag{4}$$

Thus $\mathcal{P}_h^{\widehat{Q}}$ takes a reference policy and returns a policy that trades critic improvement against displacement from that reference. The admissible policy class Π and the geometry determine the operator. We assume a minimizer exists when discussing exact proximal maps.

3.1 TD3+BC as a Wasserstein proximal objective

For deterministic policies, the Wasserstein–2 distance between Dirac measures reduces exactly to Euclidean action distance,

$$W_2^2(\delta_{\mu(s)}, \delta_{\nu(s)}) = \|\mu(s) - \nu(s)\|_2^2.$$

We therefore use the action-mean normalization

$$\mathsf{d}_\rho^2(\mu, \nu) = \mathbb{E}_{s\sim\rho} \frac{\|\mu(s) - \nu(s)\|_2^2}{n_a}, \tag{5}$$

where n_a is the action dimension. TD3+BC optimizes

$$\mathcal{L}_{\text{actor}}(\mu) = -\alpha\, \mathbb{E}_s \bar{Q}(s, \mu(s)) + \mathbb{E}_{(s,a)\sim\mathcal{D}} \frac{\|\mu(s) - a\|_2^2}{n_a}, \tag{6}$$

with $\bar{Q} = Q_1/C$ and a positive detached normalization C (Fujimoto & Gu, 2021). Averaging the squared cloning term over dataset actions leaves a distance to their conditional mean, plus a variance term independent of the actor.

Lemma 1 (Behavior anchoring as a proximal objective)**.** *Assume a fixed critic, a fixed positive scale C, $\alpha > 0$, finite objective expectations, and finite dataset-action second moments. On a closed convex action space, the TD3+BC objective has the same minimizers, when they exist, as Eq.* (4) *with the metric in Eq.* (5)*, deterministic reference $\nu(s) = \mathbb{E}_\mathcal{D}[a \mid s]$, energy $\mathcal{E}_{\bar{Q}}$, and $h = \alpha/2$.*

3.2 Advantage weighting and local Fisher–Rao geometry

For stochastic policies, the connection is different. Let $A(s, a) = Q(s, a) - V(s)$ and consider the KL-regularized distributional improvement problem from a reference $\nu(\cdot \mid s)$:

$$\max_{\pi(\cdot|s)} \left\{ \mathbb{E}_{a\sim\pi(\cdot|s)} A(s, a) - \frac{1}{h} D_{\text{KL}}(\pi(\cdot \mid s) \,\|\, \nu(\cdot \mid s)) \right\}.$$

The unrestricted optimizer is proportional to $\nu(a \mid s) e^{hA(s,a)}$, motivating advantage-weighted extraction in AWAC and IQL (Nair et al., 2021; Kostrikov et al., 2022). Since D_{KL} has second-order term $\frac{1}{2} d_{\text{FR}}^2$, the distributional objective is locally related to a Fisher–Rao proximal step (Appendix A.6).

4 Multi-step proximal policy improvement

With a fixed critic, initial reference π_0, and step sizes $h_1, \ldots, h_K$, MPI composes proximal updates:

$$\pi_k \in \mathcal{P}_{h_k}^{\widehat{Q}}(\pi_{k-1}), \qquad k = 1, \ldots, K, \qquad T = \sum_{k=1}^{K} h_k. \tag{7}$$

For equal steps, $h = T/K$. Increasing K at fixed T subdivides the nominal horizon; increasing it at fixed h extends the total horizon to Kh. These coefficients do not measure actual action displacement. In practice, we set $h = T/K$ and train K persistent actors in order: the first uses dataset actions or a base extraction loss, each later actor uses its freshly updated predecessor as a detached reference, and only π_K is deployed. Figure 1 contrasts the ideal single-stage and re-centered paths.

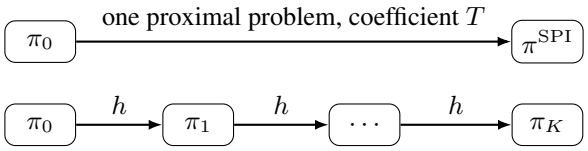

Figure 1: Single-step proximal improvement (SPI) at coefficient T (top) and K re-centered MPI steps at $h = T/K$ (bottom). Each lower arrow uses the preceding policy as its reference.

4.1 Deterministic policy refinement

The two actor updates approximate the implicit and explicit steps in Eq. (1): the implicit loss evaluates the critic at the new actor action, while the explicit update regresses onto a target formed from the critic gradient at the preceding action.

For *implicit-proximal MPI* (I-MPI), ν_1 is the sampled dataset action; for $k \geq 2$, ν_k is the detached, freshly updated predecessor. With $\bar{Q}_k^{\mathrm{I}} = \widehat{Q}/C_k^{\mathrm{I}}$ and a positive detached scale, the minibatch objective is

$$\mathcal{L}_k^{\mathrm{I}}(\mu_k \mid \nu_k) = -2h\,\mathbb{E}_i \bar{Q}_k^{\mathrm{I}}(s_i, \mu_k(s_i)) + \mathbb{E}_i \frac{\|\mu_k(s_i) - \nu_k(s_i)\|_2^2}{n_a}. \tag{8}$$

Here $\mathbb{E}_i$ denotes the minibatch average. Dividing by $2h$ gives the proximal form for later stages. At $K = 1$, I-MPI is TD3+BC with $\alpha = 2T$.

Explicit-target MPI (E-MPI) regresses onto the projected critic-gradient target

$$a_k^{\mathrm{tar}}(s) = \Pi_{\mathcal{A}} \left[\nu_k(s) + n_a h \frac{\nabla_a \widehat{Q}(s, \nu_k(s))}{C_k^{\mathrm{E}}} \right]. \tag{9}$$

The scale $C_k^{\mathrm{E}} > 0$ and the target are detached; $\Pi_{\mathcal{A}}$ projects onto the action space. Appendix B.1 specifies their scales, update order, and regression loss.

In both implementations, the first actor supplies the target actions for critic bootstrapping; later actors only affect policy extraction. We write I-MPI-K and E-MPI-K for chains with K actors using the respective updates. For example, I-MPI-4 trains four actors and evaluates the final actor μ_4.

4.2 Gaussian policy refinement

IQL's value targets use dataset actions and do not depend on the actor (Kostrikov et al., 2022). We can therefore vary extraction while retaining its value-learning procedure. We consider a diagonal-Gaussian actor with a critic-gradient and behavior-cloning loss (Gaussian Q+BC), and advantage-weighted regression (AWR), as first-stage objectives.

Write $A = Q - V$ and $\ell_\pi(s, a) = -\log \pi(a \mid s)$. The first actors minimize

$$\begin{aligned} \mathcal{L}_1^{\mathrm{QBC}} &= -\mathbb{E}_{s \sim \rho_{\mathcal{D}},\, a \sim \pi_1} Q(s, a) + h^{-1} \mathbb{E}_{\mathcal{D}} \ell_{\pi_1}(s, a), \\ \mathcal{L}_1^{\mathrm{AWR}} &= \mathbb{E}_{\mathcal{D}}[e^{hA(s,a)} \ell_{\pi_1}(s, a)]. \end{aligned} \tag{10}$$

Gaussian Q+BC extends DDPG+BC-style extraction (Park et al., 2024) by evaluating expected Q under a learned-variance Gaussian and using dataset negative log-likelihood. The first-stage losses are extraction objectives, whereas each later actor uses a Gaussian proximal loss with its detached, freshly updated predecessor as reference: W_2 after Q+BC and Fisher–Rao after AWR. The first-stage AWR connection to Fisher–Rao is local (Section 3.2). Appendix A.5 gives the distances and local updates; Appendix B.2 gives the training details.

5 Properties of composed updates

5.1 Endpoints beyond one proximal step

Changing the coefficient of one proximal problem retains its reference. A second proximal step uses the first endpoint as its reference and can change the set of reachable endpoints.

For an energy E on $\mathbb{R}^2$, write $\mathcal{P}_h^E(x) = \arg\min_y\{E(y) + \|y-x\|_2^2/(2h)\}$ and define the endpoints of exactly K steps by

$$\mathcal{R}_K(E, x_0) = \{(\mathcal{P}_{h_K}^E \circ \cdots \circ \mathcal{P}_{h_1}^E)(x_0) : h_1, \ldots, h_K > 0\}.$$

Proposition 1 (Strict enlargement of reachable endpoints)**.** *There exist a smooth strongly convex $E : \mathbb{R}^2 \to \mathbb{R}$ and $x_0 \in \mathbb{R}^2$ such that*

$$\overline{\mathcal{R}_1(E, x_0)} \subsetneq \overline{\mathcal{R}_2(E, x_0)}.$$

Appendix A.2 proves the inclusion and gives a quadratic example for which it is strict. The reachable sets range over positive step sizes; this proposition establishes a difference between update compositions, not a fixed-T performance ordering.

5.2 Local horizon subdivision

With a fixed critic, unrestricted statewise optimization, and inactive action constraints, the deterministic flow and its implicit and explicit steps satisfy

$$\begin{aligned} \partial_t \mu_t(s) &= n_a \nabla_a \bar{Q}(s, \mu_t(s)) =: f_{\bar{Q}}(\mu_t)(s), \\ \mu^{\mathrm{I}} &= \nu + T f_{\bar{Q}}(\mu^{\mathrm{I}}), \\ \mu^{\mathrm{E}} &= \Pi_{\mathcal{A}}\left[\nu + T f_{\bar{Q}}(\nu)\right]. \end{aligned} \tag{11}$$

These are the backward and forward steps in Eq. (1), specialized to the deterministic actor. Appendix A.1 derives the relations and distinguishes the fixed reference from a sampled first-stage target.

At a fixed state, write $x \in \mathbb{R}^{n_a}$ for the action and $f(x) = n_a \nabla_a \bar{Q}(s, x)$, with Jacobian $J_f(x)$. Let $\Phi_T(x)$ be its time-T flow. The critic and its normalization stay fixed across steps.

Proposition 2 (Local horizon composition)**.** *Suppose an ideal explicit or implicit update is a local map*

$$\Psi_h(x) = x + hf(x) + h^2 p(x) + O(h^3),$$

with $f \in C^2$, $p \in C^1$, and a remainder uniform on a neighborhood containing the local flow and update paths. Assume action constraints are inactive there. For fixed $K \geq 1$ and $T \to 0$, define

$$\Delta_2(x) := \frac{1}{2} J_f(x) f(x) - p(x).$$

Then

$$\begin{aligned} \Psi_{T/K}^K(x) - \Phi_T(x) &= -\frac{T^2}{K}\Delta_2(x) + O(T^3), \\ \Psi_{T/K}^K(x) - \Psi_T(x) &= \frac{K-1}{K} T^2 \Delta_2(x) + O(T^3). \end{aligned} \tag{12}$$

Explicit Euler has $p = 0$, while backward Euler has $p(x) = J_f(x)f(x)$. Both share the first-order displacement $Tf(x)$; at fixed K and small T, subdivision reduces the leading error relative to the common flow by $1/K$. The second relation measures the difference from one direct step. The expansion assumes exact statewise maps and a fixed critic and normalization throughout the chain. It does not establish convergence as $K \to \infty$ at fixed T; Appendix A.3 gives the proof and further scope conditions.

5.3 Critic and optimization errors

For approximate actor solves and an imperfect critic, the proximal objectives give the following fixed-critic bound.

Proposition 3 (Inexact improvement under a fixed critic)**.** *Let $\widehat{F} = -\mathcal{E}_{\widehat{Q}}$ be fixed throughout a chain with $h_k > 0$. Assume the preceding policy is feasible at each stage and the attained proximal objective is at most its value at that preceding policy plus $\xi_k \geq 0$. For a reference objective F, suppose $|\widehat{F}(\pi) - F(\pi)| \leq \epsilon$ on the visited policies. Then*

$$F(\pi_K) - F(\pi_0) \geq \sum_{k=1}^{K} \frac{\mathsf{d}_{\rho_{\mathcal{D}}}^2(\pi_k, \pi_{k-1})}{2h_k} - \sum_{k=1}^{K} \xi_k - 2\epsilon. \tag{13}$$

Telescoping the stagewise inequalities gives the displacement term minus accumulated optimization error; comparing the two endpoints adds 2ϵ (Appendix A.4). The result bounds the change in F under the stated critic and optimization errors, without assuming those errors are small in a trained actor.

6 Experiments

We study improvement composition on nine D4RL locomotion tasks (Fu et al., 2020), with four training seeds per observed configuration. Task-level scores average seeds; aggregate curves weight tasks equally. Deterministic scores use the final checkpoint after 10^6 updates. A fixed-T TD3+BC sweep tests horizon subdivision: all depths share the critic-update count, while depth K uses K actor optimizer calls per delayed update. We then test both gradient-flow discretizations and Gaussian extraction with an IQL critic.

6.1 Horizon and depth in TD3+BC

TD3+BC reaches its highest sampled mean at $T = 1.5$. After I-MPI-4's peak at $T = 4$, its mean stays at or above 90% of that peak through $T = 10$ and first drops below at $T = 12$. We define R_1 by $0 < T \leq 1.5$, R_2 by $1.5 < T \leq 10$, and R_3 by $10 < T \leq 40$ on the sampled grid. These boundaries describe the aggregate curve rather than a common threshold for every task.

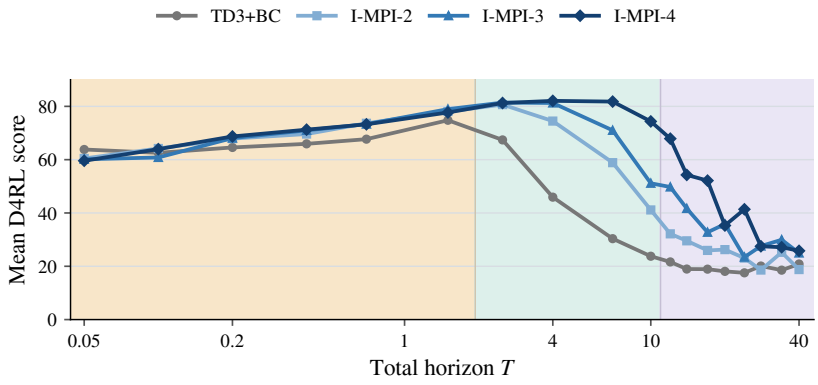

Figure 2: Mean D4RL score by total horizon T. Shading marks R_1, R_2, and R_3 as defined in the text.

In R_1, returns are similar across depths. In R_2 and R_3, I-MPI-4 has a higher mean than TD3+BC. Across R_2 and R_3 combined (108 task–horizon cells), the fraction with four-seed mean below 20 is 25.0% for I-MPI-4 versus 53.7% for TD3+BC; I-MPI-4 exceeds I-MPI-3 by 8.3 points (95% task-bootstrap interval $[2.9, 14.3]$). Its task-level mean exceeds TD3+BC on eight of nine tasks, with HalfCheetah medium as the exception (Appendix C.1).

Table 1: Regional mean D4RL score and percentage of low-return cells. Higher mean scores and lower low-return percentages are better.

	Mean D4RL score			Low-return cells (%)			
Method	R_1	R_2	R_3	R_1	R_2	R_3	$R_2 + R_3$
TD3+BC	66.6	41.9	19.3	**0.0**	27.8	66.7	53.7
I-MPI-2	69.0	63.8	24.9	1.9	8.3	58.3	41.7
I-MPI-3	68.7	71.2	33.2	1.9	5.6	50.0	35.2
I-MPI-4	**69.1**	**79.9**	**41.4**	1.9	**0.0**	**37.5**	**25.0**

R_1, R_2, and R_3 contain 54, 36, and 72 task–horizon cells, respectively; horizons have equal weight. Bold marks the best displayed value in each column, including rounded ties.

At a common small local horizon $h = 0.1$, the two-stage chain $(K, T) = (2, 0.2)$ averages 67.9, versus 62.6 for the one-stage $(K, T) = (1, 0.1)$ actor; here adding a stage also extends the total horizon.

At fixed T, increasing depth reduces the first actor's coefficient $h = T/K$ and changes the critic targets it supplies. Within I-MPI-4 runs, the endpoint scores 9.13 points above its first actor on average in R_2; over the sampled $12 \leq T \leq 20$ part of R_3, their mean difference is -0.23. In a matched $K = 4$ control, re-centering has positive and negative task-level effects relative to a fixed reference (mean -0.73 points; Appendices C.2 and C.3).

6.2 Implicit and explicit actor updates

On the shared $4 \leq T \leq 20$ grid, increasing depth from $K = 2$ to $K = 3$ raises the nine-task mean from 38.4 to 46.4 for E-MPI and from 41.2 to 51.9 for I-MPI. TD3+BC averages 25.4 on the same grid (Figure 3).

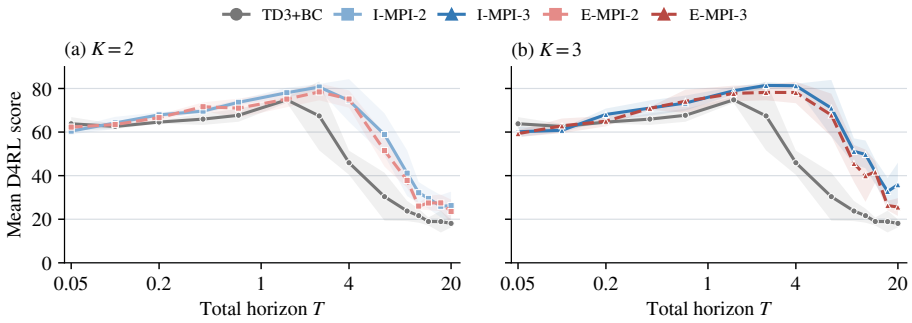

Figure 3: I-MPI and E-MPI at matched depth and total horizon. Bands show one sample standard deviation across seed-specific task means; TD3+BC has $K = 1$.

6.3 Gaussian extraction with IQL critics

IQL's actor-independent value targets let us vary extraction without changing its value-learning procedure. Gaussian AWR–FR responds differently across tasks (Figure 4). Hopper medium-replay improves with depth around $T = 2.5$. On Hopper and HalfCheetah expert at $T = 1.5$, the single-stage policies remain strong where deeper endpoints collapse. Thus FR refinement helps in some task–horizon settings but cannot be read as a uniform depth gain.

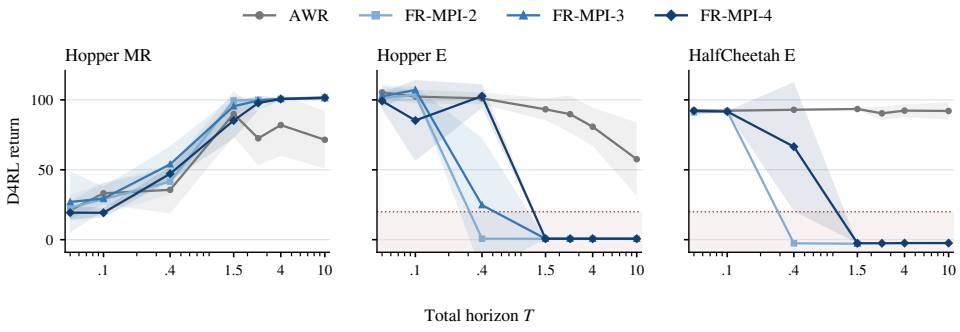

Figure 4: Gaussian AWR and Fisher–Rao refinements with an IQL critic. MR and E denote medium-replay and expert; bands show one sample standard deviation across seeds. Missing configurations are blank; the shaded region in expert panels marks returns below 20.

With Gaussian Q+BC–W_2, greater depth raises the nine-task mean over intermediate horizons; all depths lose the gain at the largest T (Figure 5).

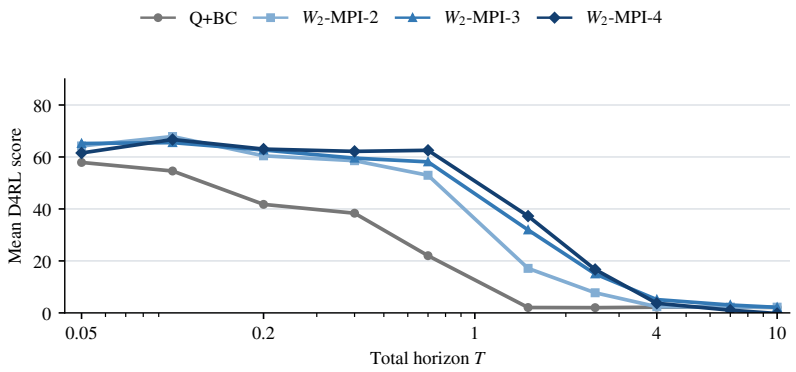

Figure 5: Gaussian Q+BC and W_2 refinements with an IQL critic. The curves use the same measured horizon grid.

On Hopper and HalfCheetah expert at $T = 1.5$, Gaussian Q+BC–W_2 also fails, including at $K = 1$, whereas the AWR single-stage actor retains high return. These configurations change both the extraction loss and the refinement geometry; equal nominal T does not calibrate their movement scales (Appendix B.2).

6.4 TD3+BC TARGET AND ENDPOINT DIAGNOSTICS

A separate matched diagnostic at five horizons compares the I-MPI-4 first actor supplying bootstrap targets, the TD3+BC target actor, and the deployed I-MPI-4 endpoint. The first actor lies closer to observed successor actions than the TD3+BC target actor, while the endpoint moves farther from those actions (Figure 6). Thus a smaller target-action displacement can coexist with a more distant deployed policy. The diagnostic measures movement relative to recorded actions; it does not establish that either actor has a more accurate critic.

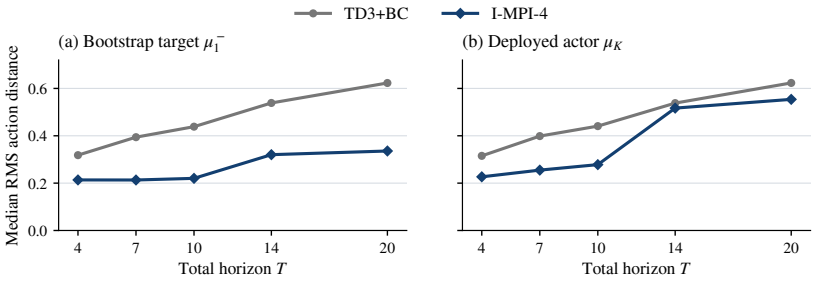

Figure 6: RMS distance to observed successor actions: (a) bootstrap target, (b) deployed actor. Gray: TD3+BC; blue: I-MPI-4. Points are medians across the matched tasks and two diagnostic seeds.

7 RELATED WORK

Behavior regularization and policy extraction. BCQ, BEAR, and BRAC limit departure from offline data (Fujimoto et al., 2019; Kumar et al., 2019; Wu et al., 2019); TD3+BC and ReBRAC use quadratic actor penalties (Fujimoto & Gu, 2021; Tarasov et al., 2023), whereas AWAC and IQL use advantage-weighted extraction (Nair et al., 2021; Kostrikov et al., 2022). Extraction matters even with fixed value learning (Park et al., 2024).

One-step and iterative offline improvement. One-step RL extracts from a behavior-policy value estimate (Brandfonbrener et al., 2021); iterative critic regularization can coincide with it under specific assumptions (Eysenbach et al., 2023). STR uses behavior-supported trust regions (Mao et al., 2023), CFPI gives constrained one-step and iterative operators (Li et al., 2023), and CPI re-centers KL regularization across policy-iteration rounds (Ma et al., 2024). MPI composes actors under a

common critic within each update, with total horizon T and depth K varied separately; its first deterministic actor supplies critic targets.

Policy geometry and expressive actors. Information-geometric trust regions and Wasserstein policy flows provide movement costs (Amari & Nagaoka, 2000; Schulman et al., 2015; Zhang et al., 2018; Moskovitz et al., 2020; Terpin et al., 2022; Pfau et al., 2025). Value Gradient Flow implements Wasserstein minimizing movement through value-gradient particle transport with separate training and evaluation budgets (Xu et al., 2026). MPI learns persistent actor stages, tests subdivision at fixed T, and deploys the final actor without test-time transport. Diffusion-QL and FQL enrich actor parameterization (Wang et al., 2023; Park et al., 2025); MPI changes the update sequence within an actor family.

8 Discussion and conclusion

Splitting a nominal horizon changes the reachable endpoints of ideal proximal updates and the behavior of trained actors. On the measured TD3+BC sweep, I-MPI-4 retains high return over more horizons and has fewer low-return cells in R_2 and R_3. At fixed T, its first actor also uses a smaller coefficient and supplies the bootstrap targets. That branch moves less relative to dataset actions while the endpoint need not stay equally close (Appendix C.4). The fixed-reference control has mixed task-level effects, so the aggregate gain cannot be attributed solely to re-centering. At a common small local horizon, adding a second stage can raise the nine-task mean as the total horizon grows.

Depth helps on some Gaussian extraction tasks and fails on others. The high-return regions were identified retrospectively, and selecting a horizon on two dynamics families gives mixed held-out outcomes (Appendix C.1). The flow analysis assumes ideal updates with a fixed critic and scale; training changes both the critic and its normalization. Composition therefore matters in the tested setting, but depth alone does not select a reliable horizon. Training uses extra actors; inference deploys only the endpoint.

AI use statement

Generative AI tools assisted with drafting and polishing manuscript text, literature retrieval, formulating and checking mathematical proofs, experiment-design feedback, analysis code, interpretation of logged results, and figure preparation. The authors are responsible for the final scientific claims and supporting evidence.

Reproducibility statement

Sections 4–6 define the refinement objectives, analysis assumptions, and experimental protocols. The appendix gives proofs, training details, and the Gaussian evaluation protocol. Analysis inputs and figure generators accompany the manuscript.

References

Shun-ichi Amari and Hiroshi Nagaoka. *Methods of Information Geometry*. American Mathematical Society, 2000.

Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient Flows in Metric Spaces and in the Space of Probability Measures*. Birkhäuser, 2005.

David Brandfonbrener, William Whitney, Rajesh Ranganath, and Joan Bruna. Offline RL without off-policy evaluation. In *Advances in Neural Information Processing Systems*, 2021.

Benjamin Eysenbach, Matthieu Geist, Sergey Levine, and Ruslan Salakhutdinov. A connection between one-step RL and critic regularization in reinforcement learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 9485–9507, 2023. URL https://proceedings.mlr.press/v202/eysenbach23a.html.

Justin Fu, Aviral Kumar, Ofir Nachum, George Tucker, and Sergey Levine. D4RL: Datasets for deep data-driven reinforcement learning, 2020.

Scott Fujimoto and Shixiang Shane Gu. A minimalist approach to offline reinforcement learning. In *Advances in Neural Information Processing Systems*, 2021.

Scott Fujimoto, David Meger, and Doina Precup. Off-policy deep reinforcement learning without exploration. In *Proceedings of the 36th International Conference on Machine Learning*, 2019.

Matthias Gelbrich. On a formula for the L2 Wasserstein metric between measures on Euclidean and Hilbert spaces. *Mathematische Nachrichten*, 147:185–203, 1990. doi: 10.1002/mana.19901470121.

Ernst Hairer, Syvert P. Nørsett, and Gerhard Wanner. *Solving Ordinary Differential Equations I: Nonstiff Problems*. Springer, 2 edition, 1993. doi: 10.1007/978-3-540-78862-1.

Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the Fokker–Planck equation. *SIAM Journal on Mathematical Analysis*, 29(1):1–17, 1998. doi: 10.1137/S0036141096303359.

Ilya Kostrikov, Ashvin Nair, and Sergey Levine. Offline reinforcement learning with implicit Q-learning. In *International Conference on Learning Representations*, 2022.

Aviral Kumar, Justin Fu, George Tucker, and Sergey Levine. Stabilizing off-policy Q-learning via bootstrapping error reduction. In *Advances in Neural Information Processing Systems*, 2019.

Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative Q-learning for offline reinforcement learning. In *Advances in Neural Information Processing Systems*, 2020.

Sebastian Kurtek and Karthik Bharath. Bayesian sensitivity analysis with the Fisher–Rao metric. *Biometrika*, 102(3):601–616, 2015. doi: 10.1093/biomet/asv026.

Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.

Jiachen Li, Edwin Zhang, Ming Yin, Qinxun Bai, Yu-Xiang Wang, and William Yang Wang. Offline reinforcement learning with closed-form policy improvement operators. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 20485–20528, 2023. URL https://proceedings.mlr.press/v202/li23av.html.

Yi Ma, Jianye Hao, Xiaohan Hu, Yan Zheng, and Chenjun Xiao. Iteratively refined behavior regularization for offline reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/663bce02a0050c4a11f1eb8a7f1429d3-Paper-Conference.pdf.

Yixiu Mao, Hongchang Zhang, Chen Chen, Yi Xu, and Xiangyang Ji. Supported trust region optimization for offline reinforcement learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 23829–23851, 2023. URL https://proceedings.mlr.press/v202/mao23c.html.

Ted Moskovitz, Michael Arbel, Ferenc Huszár, and Arthur Gretton. Efficient wasserstein natural gradients for reinforcement learning. *arXiv preprint arXiv:2010.05380*, 2020.

Ashvin Nair, Abhishek Gupta, Murtaza Dalal, and Sergey Levine. AWAC: Accelerating online reinforcement learning with offline datasets. In *International Conference on Learning Representations*, 2021.

Neal Parikh and Stephen Boyd. Proximal algorithms. *Foundations and Trends in Optimization*, 1 (3):127–239, 2014.

Seohong Park, Kevin Frans, Sergey Levine, and Aviral Kumar. Is value learning really the main bottleneck in offline RL? In *Advances in Neural Information Processing Systems*, 2024.

Seohong Park, Qiyang Li, and Sergey Levine. Flow Q-learning. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025.

David Pfau, Ian Davies, Diana L. Borsa, João G. M. Araújo, Brendan D. Tracey, and Hado van Hasselt. Wasserstein policy optimization. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 49128–49149, 2025.

John Schulman, Sergey Levine, Philipp Moritz, Michael I. Jordan, and Pieter Abbeel. Trust region policy optimization. In *Proceedings of the 32nd International Conference on Machine Learning*, 2015.

Denis Tarasov, Alexander Nikulin, Dmitry Akimov, Vladislav Kurenkov, and Sergey Kolesnikov. CORL: Research-oriented deep offline reinforcement learning library. NeurIPS Offline Reinforcement Learning Workshop, 2022.

Denis Tarasov, Vladislav Kurenkov, Alexander Nikulin, and Sergey Kolesnikov. Revisiting the minimalist approach to offline reinforcement learning. In *Advances in Neural Information Processing Systems*, 2023.

Antonio Terpin, Nicolas Lanzetti, Batuhan Yardim, Florian Dörfler, and Giorgia Ramponi. Trust region policy optimization with optimal transport discrepancies: Duality and algorithm for continuous actions. *arXiv preprint arXiv:2210.11137*, 2022.

Zhendong Wang, Jonathan J. Hunt, and Mingyuan Zhou. Diffusion policies as an expressive policy class for offline reinforcement learning. In *International Conference on Learning Representations*, 2023.

Yifan Wu, George Tucker, and Ofir Nachum. Behavior regularized offline reinforcement learning. *arXiv preprint arXiv:1911.11361*, 2019.

Haoran Xu, Kaiwen Hu, Somayeh Sojoudi, and Amy Zhang. Reinforcement learning via value gradient flow, 2026.

Ruiyi Zhang, Changyou Chen, Chunyuan Li, and Lawrence Carin. Policy optimization as wasserstein gradient flows. In *Proceedings of the 35th International Conference on Machine Learning*, 2018.

APPENDIX

A PROOFS AND GEOMETRIC DERIVATIONS

A.1 BEHAVIOR ANCHORING: PROOF OF LEMMA 1

$$\mathbb{E}_{(s,a)} \frac{\|\mu(s) - a\|_2^2}{n_a} = \mathsf{d}^2_{\rho_\mathcal{D}}(\mu, b) + \text{const.} \tag{S1}$$

Let $b(s) = \mathbb{E}[a \mid s]$. The conditional bias–variance identity

$$\mathbb{E}[\|\mu(s) - a\|_2^2 \mid s] = \|\mu(s) - b(s)\|_2^2 + \mathbb{E}[\|a - b(s)\|_2^2 \mid s]$$

proves Eq. (S1) after averaging over states and dividing by n_a. The second term is independent of μ. Dividing the anchored objective $-\alpha\mathbb{E}_s\bar{Q}(s, \mu(s)) + \mathsf{d}^2_{\rho_\mathcal{D}}(\mu, \nu)$ by α gives Eq. (4) with coefficient $h = \alpha/2$ and normalized energy. Under the metric inner product $n_a^{-1}\mathbb{E}_s\langle\cdot,\cdot\rangle$, the negative energy gradient is $f_{\bar{Q}}(\mu)(s) = n_a\nabla_a\bar{Q}(s, \mu(s))$. The usual proximal optimality condition (Parikh & Boyd, 2014) gives the corresponding update: for unrestricted statewise optimization, an interior differentiable minimizer satisfies $(\mu^+ - \nu)/T = f_{\bar{Q}}(\mu^+)$; evaluating the same vector field at ν gives the explicit step in Eq. (11).

The conditional-mean anchor is admissible on a closed convex action space. The result assumes a nonempty minimizer set; continuity on a compact admissible class suffices. The same rescaling applies to a ReBRAC-style squared actor penalty with fixed critic and coefficient (Tarasov et al., 2023).

Sampled first hop of E-MPI. For a fixed positive scale C, unrestricted population MSE regression to the sampled explicit targets yields

$$\mu_1^{\mathrm{E}}(s) = \mathbb{E}\left[\Pi_\mathcal{A}\left(a + \frac{n_a h}{C}\nabla_a\widehat{Q}(s,a)\right) \,\middle|\, s\right].$$

Without clipping, this is $b(s) + (n_a h/C)\mathbb{E}[\nabla_a\widehat{Q}(s,a) \mid s]$, which generally differs from a gradient step at $b(s)$. Thus the sampled first E-MPI hop differs from an explicit step at b.

A.2 PROOF OF PROPOSITION 1

Using the standard proximal-map definition (Parikh & Boyd, 2014), for any smooth strongly convex E, the proximal maps are unique and $\mathcal{P}^E_\epsilon(y) \to y$ as $\epsilon \downarrow 0$. Indeed, optimality and the lower bound on E give

$$\|\mathcal{P}^E_\epsilon(y) - y\|_2^2 \le 2\epsilon\big(E(y) - \inf E\big).$$

Every $z = \mathcal{P}^E_h(x_0) \in \mathcal{R}_1(E, x_0)$ is therefore the limit of $\mathcal{P}^E_\epsilon(z) \in \mathcal{R}_2(E, x_0)$, proving $\overline{\mathcal{R}_1(E, x_0)} \subseteq \overline{\mathcal{R}_2(E, x_0)}$.

Consider $E(x) = \frac{1}{2}(x_1^2 + 2x_2^2)$ and $x_0 = (1, 1)$. Direct minimization gives

$$\mathcal{P}^E_h(x_0) = \left(\frac{1}{1+h}, \frac{1}{1+2h}\right), \qquad (\mathcal{P}^E_1 \circ \mathcal{P}^E_1)(x_0) = \left(\frac{1}{4}, \frac{1}{9}\right). \tag{S2}$$

If a sequence of single-step endpoints converged to $(1/4, 1/9)$, its first coordinate would force the corresponding coefficients to converge to 3; its second coordinate would then converge to $1/7$, a contradiction. Hence $(1/4, 1/9) \in \mathcal{R}_2(E, x_0) \setminus \overline{\mathcal{R}_1(E, x_0)}$, and the inclusion is strict.

A.3 PROOF OF PROPOSITION 2

We use the local truncation expansion of forward and backward Euler steps (Hairer et al., 1993). Under the proposition's smoothness, uniform-remainder, and interiority assumptions, write $f_0 =$

$f(x_0)$, $J_0 = J_f(x_0)$, and $p_0 = p(x_0)$. For fixed K, consistency gives $x_j = x_0 + jhf_0 + O(h^2)$ and hence $f(x_j) = f_0 + jhJ_0f_0 + O(h^2)$. Summing the increments gives

$$\begin{aligned} x_K - x_0 &= h\sum_{j=0}^{K-1} f(x_j) + h^2 \sum_{j=0}^{K-1} p(x_j) + O(Kh^3) \\ &= Khf_0 + h^2\frac{K(K-1)}{2}J_0f_0 + Kh^2p_0 + O(h^3). \end{aligned}$$

Substituting $h = T/K$ yields

$$\Psi^K_{T/K}(x_0) = x_0 + Tf_0 + T^2\left[\frac{K-1}{2K}J_0f_0 + \frac{1}{K}p_0\right] + O(T^3).$$

The exact flow has expansion $\Phi_T(x_0) = x_0 + Tf_0 + \frac{1}{2}T^2J_0f_0 + O(T^3)$. Subtracting it, or the single-step expansion $\Psi_T(x_0) = x_0 + Tf_0 + T^2p_0 + O(T^3)$, proves the two identities in Eq. (12). Explicit Euler has $p_0 = 0$; Taylor expansion of the local backward-Euler branch gives $p_0 = J_0f_0$.

At fixed T, convergence as $K \to \infty$ additionally requires a unique flow on $[0, T]$, sufficient smoothness around the full flow and numerical paths, and stable steps. For ReLU critics, the action gradient is constant inside an activation region. There the explicit update and same-region stationary implicit branch coincide, although a global proximal minimizer can lie in another region. The smooth expansion does not apply across an activation boundary.

A.4 Proof of Proposition 3

Let π_{k-1} belong to the feasible class of stage k. It suffices that the approximate solve satisfies

$$-\widehat{F}(\pi_k) + \frac{\mathsf{d}^2_{\rho_\mathcal{D}}(\pi_k, \pi_{k-1})}{2h_k} \le -\widehat{F}(\pi_{k-1}) + \xi_k.$$

An objective value within ξ_k of the proximal infimum implies this inequality, the attained-objective condition used in the proof. Summing it telescopes the fixed learned objective. Applying the error bound at π_0 and π_K gives Eq. (13). Only endpoint critic errors enter the bound, while optimization errors accumulate across stages. A uniform action-value error bound implies the assumed objective error under the fixed dataset state marginal. The bound also applies to feasible stochastic policies when the objective and normalization remain fixed throughout the chain. The live TD3+BC sweep changes normalization scales across stages.

A.5 Gaussian geometry for MPI refinement

For diagonal Gaussians $\pi = \mathcal{N}(m, \mathrm{diag}(\sigma^2))$ and $\nu = \mathcal{N}(\widetilde{m}, \mathrm{diag}(\widetilde{\sigma}^2))$, the Wasserstein distance follows from the Gaussian formula (Gelbrich, 1990):

$$W_2^2(\pi, \nu) = \|m - \widetilde{m}\|_2^2 + \|\sigma - \widetilde{\sigma}\|_2^2. \tag{S3}$$

Its Dirac limit gives deterministic refinement. The square-root-density representation gives the ambient Fisher–Rao distance (Amari & Nagaoka, 2000; Kurtek & Bharath, 2015); with the normalization whose infinitesimal squared line element is the full Fisher information metric, for positive Gaussian densities it takes the form

$$d_{\mathrm{FR}}(\pi, \nu) = 2\arccos\left(\int \sqrt{\pi(a)\nu(a)}\, da\right). \tag{S4}$$

This is the ambient density-space distance evaluated on Gaussians, rather than the intrinsic geodesic distance within the Gaussian family.

For diagonal Gaussians with positive standard deviations, the Bhattacharyya coefficient in Eq. (S4) is

$$B(\pi, \nu) = \prod_{j=1}^{n_a}\left(\frac{2\sigma_j\widetilde{\sigma}_j}{\sigma_j^2 + \widetilde{\sigma}_j^2}\right)^{1/2} \exp\left[-\frac{(m_j - \widetilde{m}_j)^2}{4(\sigma_j^2 + \widetilde{\sigma}_j^2)}\right]. \tag{S5}$$

Thus $d^2_{\mathrm{FR}} = 4 \arccos^2 B$ is evaluated analytically. The shortest ambient path need not be Gaussian. Both distances use sums over action coordinates; rescaling a distance changes its matching coefficient h.

The diagonal-Gaussian proximal minimizer may differ from the unrestricted distributional solution; optimizing within the family is generally not equivalent to projecting the unrestricted solution.

To describe local motion at a fixed state, set $G(m, \sigma) = \mathbb{E}_\epsilon \widehat{Q}(s, m + \sigma \odot \epsilon)$, where $\epsilon \sim \mathcal{N}(0, I)$. Under differentiation beneath the expectation,

$$g_m = \nabla_m G = \mathbb{E}_\epsilon \nabla_a \widehat{Q}(s, m + \sigma \odot \epsilon), \qquad g_\sigma = \nabla_\sigma G = \mathbb{E}_\epsilon[\epsilon \odot \nabla_a \widehat{Q}(s, m + \sigma \odot \epsilon)].$$

The local squared line elements are

$$ds^2_{W_2} = \sum_j (dm_j^2 + d\sigma_j^2), \qquad ds^2_{\mathrm{FR}} = \sum_j \frac{dm_j^2 + 2d\sigma_j^2}{\sigma_j^2}.$$

Lemma S1 (Local Gaussian refinement)**.** *Assume G is smooth with differentiation beneath the expectation, all reference variances are positive, and action and parameter constraints are inactive. For a smooth local proximal branch within the diagonal-Gaussian family converging to the reference as $h \to 0$, using the coordinate-sum distances above, the first-order updates are*

$$\begin{aligned} W_2: &\quad \Delta m = h g_m + O(h^2), && \Delta\sigma = h g_\sigma + O(h^2), \\ \mathrm{FR}: &\quad \Delta m = h\sigma^2 \odot g_m + O(h^2), && \Delta\sigma = \tfrac{h}{2}\sigma^2 \odot g_\sigma + O(h^2). \end{aligned} \tag{S6}$$

All quantities on the right are evaluated at the reference policy, with the same critic and coefficient.

Proof. For $\theta = (m, \sigma)$, write the local squared-distance expansion as $d^2(\theta + \delta, \theta) = \delta^\top M(\theta)\delta + O(\|\delta\|^3)$. The stationary equation for $-G + d^2/(2h)$ on the local branch gives $\delta = hM^{-1}\nabla G + O(h^2)$. For W_2, $M = I$; for FR, $M = \mathrm{diag}(\sigma^{-2}, 2\sigma^{-2})$. Substitution proves Eq. (S6). □

A small variance attenuates the FR mean velocity locally. Using $\widehat{Q}(s, m)$ gives $g_\sigma = 0$. For Gaussian W_2 refinement with a mean-only energy, the optimal standard deviation is the reference standard deviation.

A.6 Local Fisher–Rao connection to advantage weighting

The local Fisher metric and KL expansion are standard (Amari & Nagaoka, 2000). For a regular parametric density family and a small parameter increment δ, let $I(\theta)$ denote its Fisher information matrix. Normalization cancels the linear term in a Taylor expansion of either KL orientation, giving

$$\begin{aligned} D_{\mathrm{KL}}(p_{\theta+\delta} \| p_\theta) &= \tfrac{1}{2}\delta^\top I(\theta)\delta + o(\|\delta\|^2), \\ D_{\mathrm{KL}}(p_\theta \| p_{\theta+\delta}) &= \tfrac{1}{2}\delta^\top I(\theta)\delta + o(\|\delta\|^2), \\ d^2_{\mathrm{FR}}(p_{\theta+\delta}, p_\theta) &= \delta^\top I(\theta)\delta + o(\|\delta\|^2). \end{aligned}$$

These expansions require smooth positive densities and a regular parameterization near the reference. Replacing $d^2_{\mathrm{FR}}/(2h)$ locally by $D_{\mathrm{KL}}(\pi\|\nu)/h$ yields the distributional problem

$$\max_\pi \ \mathbb{E}_\pi A(s, a) - h^{-1} D_{\mathrm{KL}}(\pi\|\nu), \qquad \pi_h(a \mid s) = \frac{\nu(a \mid s) e^{hA(s,a)}}{Z_h(s)}.$$

The formula assumes a finite normalizer and support within ν. Gaussian weighted maximum likelihood fits the advantage-weighted distribution, while finite sampling and clipped weights further distinguish the implemented AWR stage from an exact distributional update (Nair et al., 2021; Kostrikov et al., 2022). Later FR stages use the exact ambient distance in Eqs. (S4) and (S5).

B Implementation and evaluation protocols

B.1 Deterministic algorithms and schedules

The horizon–depth implementation builds on TD3+BC and CORL (Fujimoto & Gu, 2021; Tarasov et al., 2022). It uses $T = \tau = \alpha/2$ and $h = T/K$. The K actors are independently initialized

persistent networks with separate Adam states. Each iteration updates the critic. Every second iteration, actors take one optimizer step each on the same minibatch in chain order. After actor one updates, its target actor and the target critics receive soft updates with coefficient 0.005; later actors use the updated references in Eq. (S7). At actor update t, the exact minibatch references are

$$\nu_{1,t}(s_i) = a_i, \qquad \nu_{k,t}(s_i) = \operatorname{sg} \mu^{+}_{k-1,t}(s_i), \quad k \geq 2, \tag{S7}$$

where $+$ denotes the freshly updated output and sg stops gradients. These time indices are suppressed in the main-text objectives. E-MPI uses the target in Eq. (9) with $\widehat{Q}_t, \nu_{k,t}, C^{\mathrm{E}}_{k,t}$ and minimizes

$$\mathcal{L}^{\mathrm{E}}_{k,t}(\mu_k \mid \nu_{k,t}) = \mathbb{E}_i \frac{\|\mu_k(s_i) - a^{\mathrm{tar}}_{k,t}(s_i)\|_2^2}{n_a}. \tag{S8}$$

Actor losses use $\widehat{Q} = Q_1$; Bellman targets use the minimum of the twin target critics. E-MPI treats its projected regression targets as fixed.

Writing $\mu^{\mathrm{pre}}_{1,t}$ for actor one's output before its update, the detached normalization scales are

$$\begin{aligned} C^{\mathrm{I}}_{1,t} &= \operatorname{sg} \mathbb{E}_i |\widehat{Q}_t(s_i, \mu^{\mathrm{pre}}_{1,t}(s_i))|, \\ C^{\mathrm{E}}_{1,t} &= \operatorname{sg} \mathbb{E}_i |\widehat{Q}_t(s_i, a_i)|, \\ C^{r}_{k,t} &= \operatorname{sg} \mathbb{E}_i |\widehat{Q}_t(s_i, \nu_{k,t}(s_i))|, \quad k \geq 2, \quad r \in \{\mathrm{I}, \mathrm{E}\}. \end{aligned} \tag{S9}$$

Each scale includes an additive 10^{-6} safeguard. E-MPI's first targets use sampled dataset actions, as described in Appendix A.1.

We use a 256–256 ReLU actor MLP with tanh output and twin 256–256 ReLU critic MLPs, batch size 256, Adam learning rate 3×10^{-4}, discount 0.99, target noise 0.2, and noise clipping at 0.5 times the action bound. States are normalized by the dataset mean and standard deviation plus 10^{-3}. Timeout rows are omitted from transition construction; terminal flags stop bootstrapping.

The nine datasets combine `halfcheetah`, `hopper`, and `walker2d` with `medium-v2`, `medium-replay-v2`, and `expert-v2`. We use

$$T \in \{0.05, 0.1, 0.2, 0.4, 0.7, 1.5, 2.5, 4, 7, 10, 12, 14, 17, 20, 24, 28, 34, 40\}.$$

Each deterministic sweep score averages ten evaluation episodes per seed. E-MPI is reported through $T = 20$. In tables, IK and EK denote I-MPI-K and E-MPI-K.

B.2 Gaussian IQL implementation

In Eq. (10), Q is the minimum target critic, without the detached normalization used by TD3+BC, and $\pi_k = \mathcal{N}(m_k, \operatorname{diag}(\sigma_k^2))$. The likelihood-BC weight is $1/h$, and the AWR exponent coefficient is h. In the reported AWR runs, the exponential weight in Eq. (10) is capped at 100: $w_h(s,a) = \min\{e^{hA(s,a)}, 100\}$. Negative log-likelihood (NLL) sums action dimensions. The two first-stage losses differ from their later proximal objectives; Appendix A.6 derives the local AWR–FR relation.

Gaussian Q+BC uses an unsquashed mean and learned state-dependent standard deviation, initialized to one. AWR uses a tanh mean and learned state-dependent standard deviation. Both subsequent refinement losses evaluate expected Q with reparameterized Gaussian samples; the implementation uses eight antithetic samples by default and stops gradients through the updated predecessor. The current implementation also updates an independent full-T control when $K > 1$, so actor-update cost need not equal that of the TD3+BC chain. Geometry is computed between the underlying Gaussians. Mean-action simulator evaluation clips actions to the environment bounds.

Equal h does not calibrate action displacement or critic scales: Gaussian distances and negative log-likelihood sum over action coordinates, whereas the deterministic squared penalty averages them.

B.3 Action and target-value diagnostics

Distances to dataset actions and Bellman targets. On m fixed nonterminal validation transitions, we measure RMS distance from the observed successor action a_{i+1}:

$$D(\pi) = \left(\frac{1}{m n_a} \sum_{i=1}^{m} \left\| \pi(s'_i) - a_{i+1} \right\|_2^2 \right)^{1/2}. \tag{S10}$$

The first target actor μ_1^- and endpoint μ_K are evaluated on the same next states. The squared distance includes conditional action variance.

To measure the corresponding value-space perturbation, we use a matched TD3+BC target critic $F(s', a) = \min_j Q_j^-(s', a)$, held common across methods:

$$\Delta_y^{\mathrm{RMS}} = \gamma\Big(\mathbb{E}\big[\big(F(s', \widetilde{\mu}_1^-(s')) - F(s', \widetilde{a}'_{\mathcal{D}})\big)^2\big]\Big)^{1/2}. \tag{S11}$$

The paired actions receive identical smoothing noise and box clipping. Their RMS value difference measures bootstrap-target perturbation under the shared learned critic.

Matched comparison. The matched comparison uses nine medium, medium-replay, and expert tasks, $T \in \{4, 7, 10, 14, 20\}$, and two seeds. Target and endpoint actions are evaluated on identical next states and paired successor dataset actions. Rows with a timeout or true terminal flag are excluded. The target-noise diagnostic uses 4096 next-state transitions and 64 draws of clipped smoothing noise per checkpoint. The value-space comparison in Eq. (S11) uses the same-T TD3+BC target critic and identical smoothing noise for both compared actions.

B.4 Aggregation and uncertainty

The region means in Table 1 average four seeds, then tasks and sampled horizons with equal weight. The means refer to the measured grid. A low-return cell has a four-seed task–horizon mean below 20; Table S1 varies this threshold over the large-horizon range.

For 95% percentile intervals, we average the specified seeds and horizons within each task and resample the nine task-level contrasts 100,000 times. Low-return frequencies are also aggregated within task before resampling. These intervals retain the within-task seed and horizon structure; the nine datasets share only three dynamics families. The first/endpoint comparison resamples nine complete tasks 100,000 times, retaining within-run actor pairs, horizons, and training seeds. Evaluation episodes are averaged within each run.

B.5 Gaussian evaluation and coverage

Coverage. The Gaussian curves report final-actor mean-action scores. Gaussian Q+BC–Wasserstein covers the same nine datasets. AWR–FR covers seven datasets on an incomplete horizon–depth grid; Figure 4 shows the observed cells for Hopper medium-replay, Hopper expert, and HalfCheetah expert.

Comparison and aggregation. The $K = 1$ curves use separately trained full-T actors; the first actor of a deeper chain has coefficient T/K. Bands show one sample standard deviation. The main-text curves include only complete configurations; missing configurations are left blank. The nine-task Q+BC–Wasserstein mean weights tasks equally at each shared horizon.

C Deterministic task results and controls

C.1 Task-level horizon responses

Figure S1 resolves the mean trends of Figure 2 by task. Across R2 and R3, I-MPI-4 exceeds TD3+BC on eight of nine task means. Its mean gain over I-MPI-3 is 8.3 points ($[2.9, 14.3]$). Cells whose four-seed mean rises from below 20 with I-MPI-3 to at least 20 with I-MPI-4 contribute 65% of the aggregate signed score gain. Table S1 shows the response to other low-return thresholds. Figure S2 gives E-MPI's task-level responses through $T = 20$.

Held-out family horizon selection. For a retrospective horizon-selection check, we choose each method's highest-mean horizon on two dynamics families and evaluate it on the third. The held-out family scores average 78.4 for I-MPI-4 and 74.8 for TD3+BC. I-MPI-4 minus TD3+BC is $+17.1$ on Hopper, -8.9 on HalfCheetah, and $+2.8$ on Walker2d.

Table S1: Collapse-rate sensitivity across R2 and R3 ($1.5 < T \leq 40$). Entries are percentages of task–horizon cells whose four-seed mean falls below each score threshold.

Threshold	TD3+BC	I2	I3	I4
0	3.7	2.8	2.8	1.9
10	47.2	40.7	33.3	20.4
20	53.7	41.7	35.2	25.0
30	56.5	46.3	37.0	28.7
40	59.3	51.9	38.0	33.3

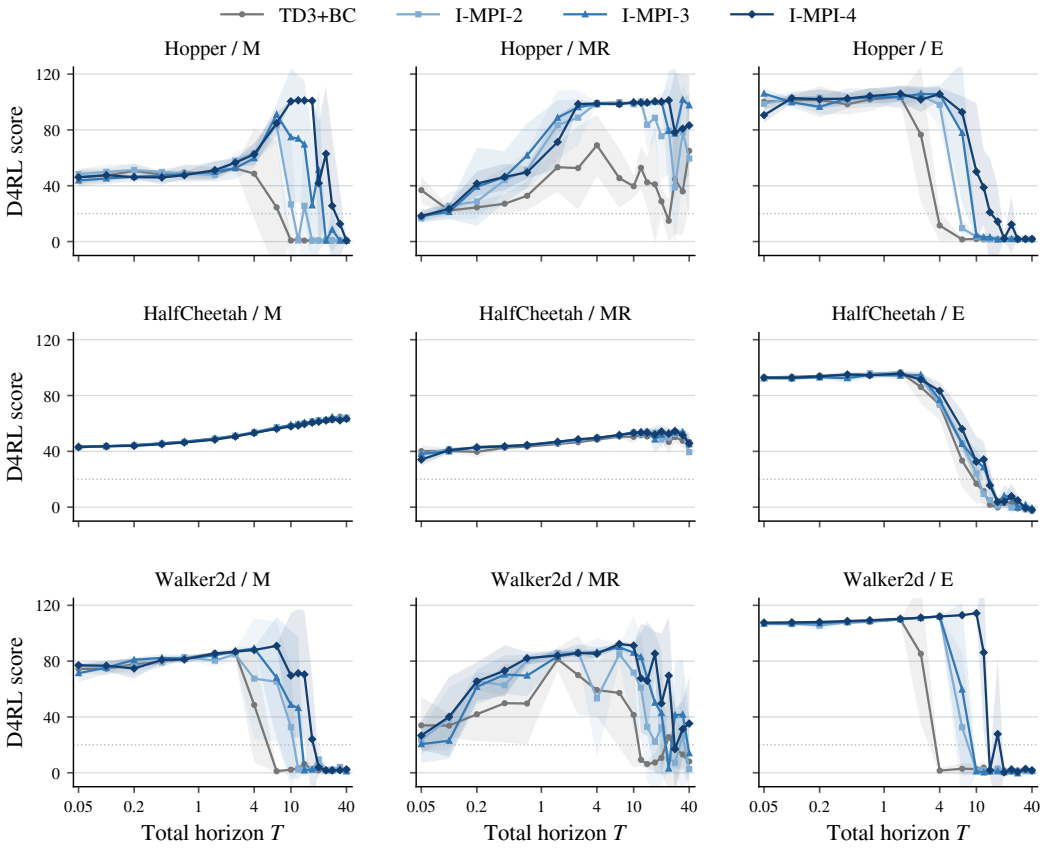

Figure S1: I-MPI task-level sweeps. Bands show one sample standard deviation across seeds; the dotted line marks score 20.

C.2 First actor versus endpoint

Figure S3 compares I-MPI-4's first actor and endpoint at the same checkpoints. Across all fourteen observed horizons through $T = 20$, the endpoint gains 5.76 points on average. Within R2 ($T \in \{2.5, 4, 7, 10\}$), its mean rises from 70.43 to 79.56, a gain of 9.13 points with a 95% task-bootstrap interval of $[-5.4, 23.6]$. Within the observed part of R3 ($T \in \{12, 14, 17, 20\}$), the means are 52.42 and 52.19, a difference of -0.23 points with interval $[-16.8, 15.0]$.

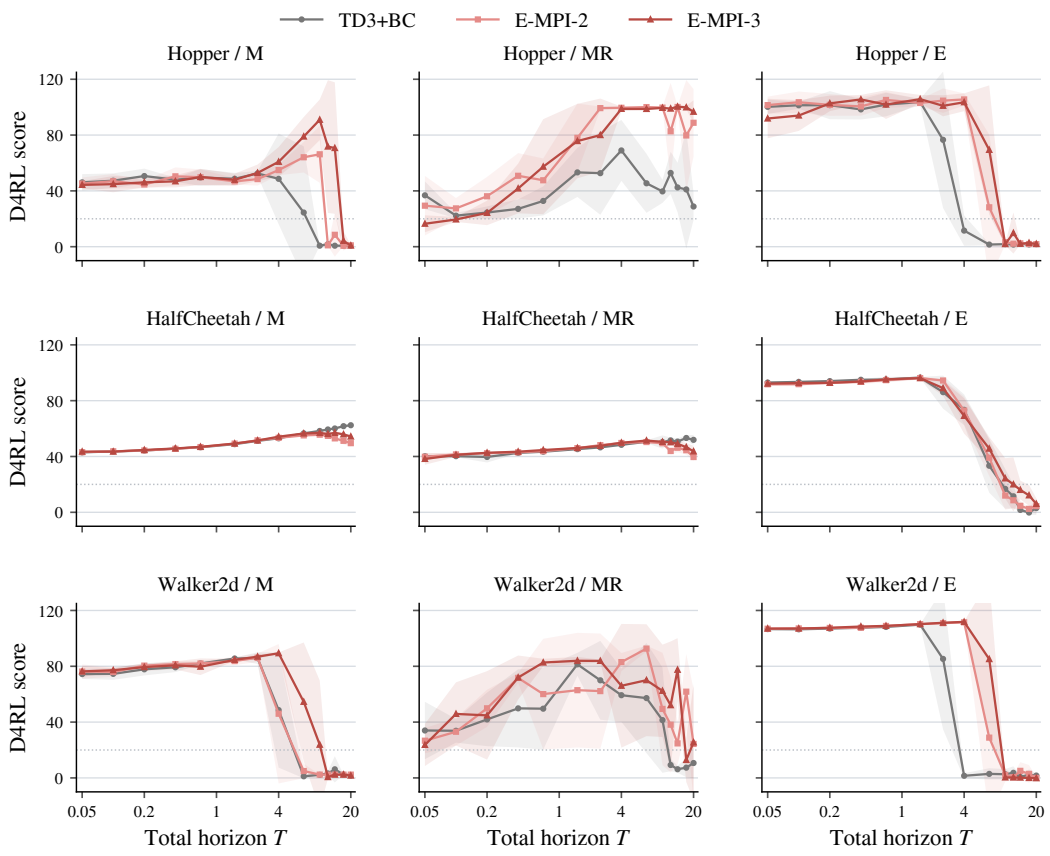

Figure S2: E-MPI task-level sweeps through $T = 20$. Bands show one sample standard deviation across seeds; TD3+BC uses the same grid.

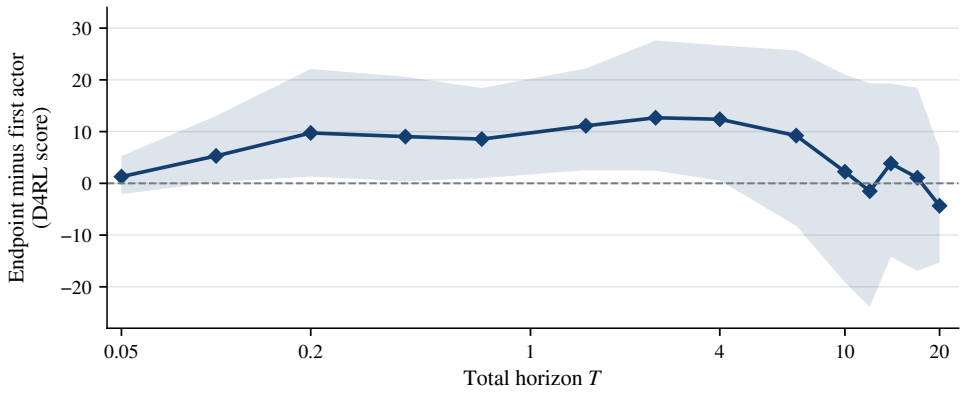

Figure S3: I-MPI-4 endpoint minus first-actor return at sampled horizons. The band is a 95% task-bootstrap interval retaining paired runs within each task.

C.3 Fixed-reference control for re-centering

The first-to-endpoint comparison combines additional actor stages with a changing reference. We therefore compare two $K = 4$ branches with the same actor-update count and local coefficient $h = T/4$ at $T \in \{4, 7, 10\}$ on the nine tasks. In the fixed-reference branch, actors π_2, π_3, π_4 use π_1

as their reference; in the re-centered branch, each uses its predecessor. Figure S4 shows a mean re-centered-minus-fixed difference of -0.73 points, with a task-bootstrap interval of $[-10.6, 8.1]$. The task means range from $+24.9$ on Hopper-medium to -32.8 on HalfCheetah-expert, showing that the return effect depends strongly on the environment. In a stagewise diagnostic on two HalfCheetah-expert seeds, the matched branches share the critic and π_1. The training references differ beginning at π_3. At $T = 7$ and 10, the fixed-reference branch recovers after low returns at π_2, while re-centering continues to reduce return.

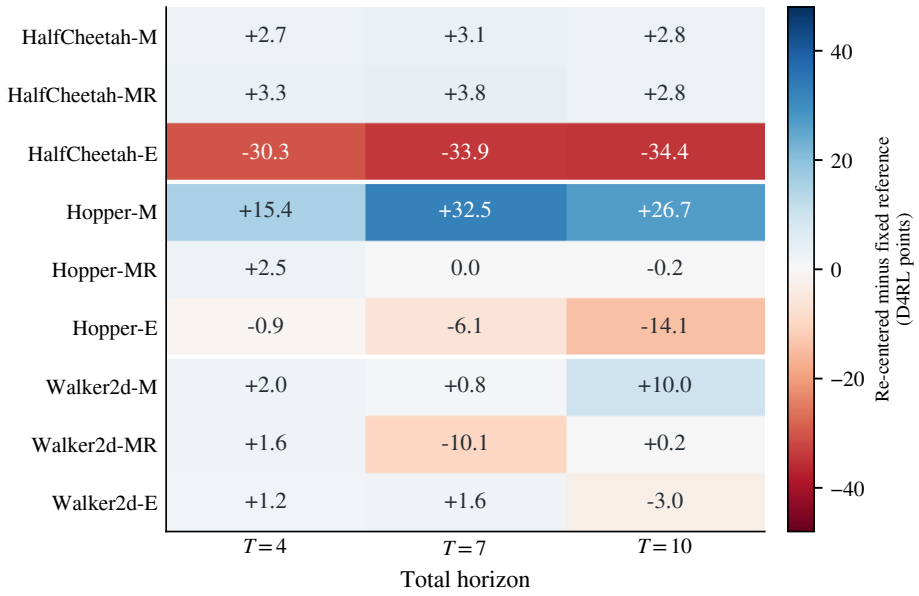

Figure S4: Paired I-MPI-4 endpoint differences: re-centered reference minus fixed reference. M, MR, and E denote medium, medium-replay, and expert. Positive values favor the re-centered endpoint.

C.4 Task-level target and endpoint diagnostics

Figure S5 shows the task–horizon–seed ratios underlying the medians in Figure 6 of Section 6.4. Across I-MPI-3 and I-MPI-4, the first target actor has smaller action displacement and target-value perturbation than TD3+BC on most matched configurations. For I-MPI-4, the median reduction in target-value perturbation under the shared TD3+BC critic is 38%. The final actor lies farther from observed successor actions than its first target actor: the median final-to-target RMS ratios are 1.297 for I-MPI-3 and 1.365 for I-MPI-4.

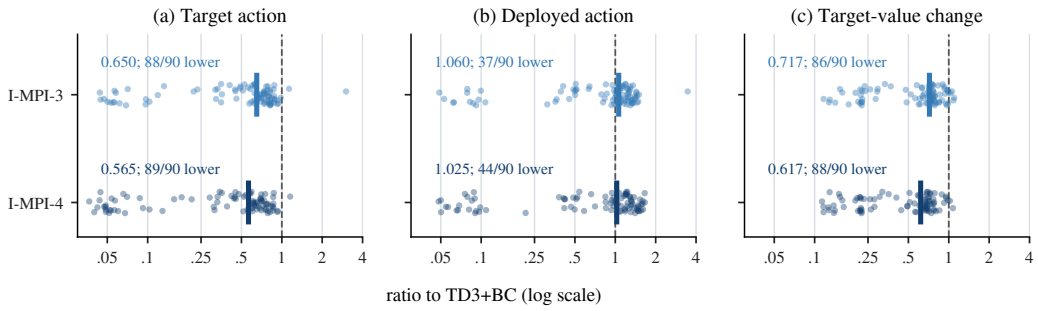

Figure S5: Paired task–horizon–seed ratios to TD3+BC at the same total horizon. Action panels use squared RMS displacement from the observed next dataset action; the value panel uses the RMS Bellman-target perturbation under the common TD3+BC critic and identical smoothing noise. Dots are individual task–horizon–seed cells, vertical bars mark medians, and labels give the median and number of ratios below one. The logarithmic axis retains the spread across runs. Ratios include the conditional-action-variance floor.

D GAUSSIAN Q+BC–W_2 RESULTS BY TASK

Figure S6 shows all nine tasks at the ten horizons observed for every depth $K = 1, \ldots, 4$. The selection protocol is given in Appendix B.5.

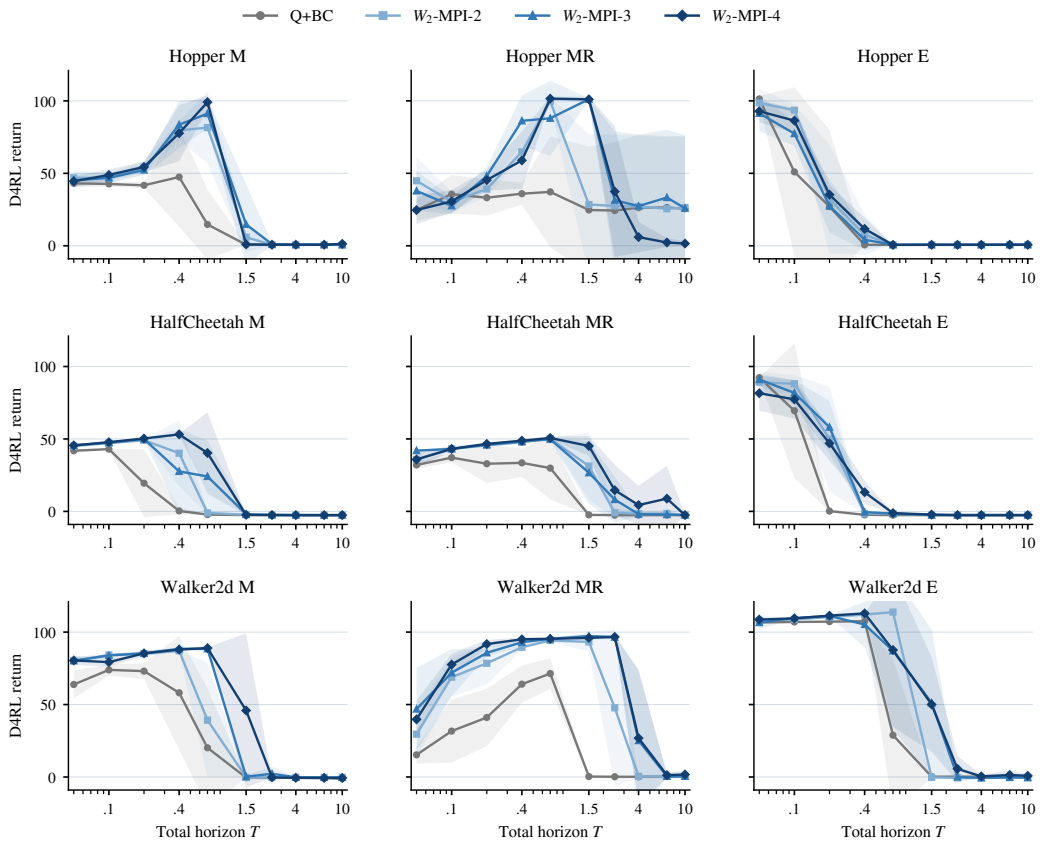

Figure S6: Gaussian Q+BC–Wasserstein, $K = 1$–4. Lines connect observed points; bands show one sample standard deviation across seeds. M/MR/E denote medium/medium-replay/expert.