COOT Accelerates Online Diffusion Policy Optimization from Confounded Offline Data

Anonymous authors
Paper under double-blind review

Abstract

Diffusion policies are powerful multimodal action generators, but online reinforcement learning with diffusion policies remains sample intensive when useful behaviors must be discovered through interaction alone. Offline-to-online learning can reduce this cost, yet practical logs are often biased by hidden execution regimes: unobserved factors may affect the logged action, the action actually executed by the environment, and the realized outcome. As a result, naive behavior cloning or offline value initialization can transfer spurious action-outcome correlations rather than reliable control knowledge. We propose COOT (**C**ausally Motivated **O**ffline-to-**O**nline **T**ransfer), a framework for accelerating online diffusion-policy optimization from biased but informative offline data. COOT uses observed proxy variables to build a proxy-informed reliability transfer bundle that estimates robust proxy-conditioned transfer value, bridge uncertainty, state potential, and behavior-support reliability. These signals selectively weight diffusion-policy pretraining, shape critic initialization, and regularize online candidate selection from both the current policy and a persistent offline prior. Under a same-data control with the identical offline dataset and pretraining budget, COOT improves over a uniform behavior-cloning warm start, showing that its gains are not explained by generic offline pretraining alone. Across five MuJoCo locomotion tasks and five random seeds, COOT also outperforms eight strong actor-critic, diffusion-policy, energy-based, Q-guided, and score-based online baselines in both peak return and sample efficiency. Ablations, proxy stress tests, and covariate-shift diagnostics further show that COOT's advantage comes from proxy-informed, uncertainty-aware, and support-aware transfer.

1 Introduction

Offline-to-online reinforcement learning aims to reduce online exploration cost by reusing pre-collected experience before interacting with the target environment. This is especially valuable in continuous control, where exploration can be expensive, unstable, or unsafe, and where deployed systems often accumulate large logs of feasible behavior (Levine et al., 2020; Nair et al., 2020; Kumar et al., 2022; Lee et al., 2022; Ball et al., 2023). Diffusion policies make this setting particularly attractive. As expressive multimodal action generators (Ho et al., 2020; Song et al., 2021b; Karras et al., 2022), they can represent many plausible actions at the same state and have become a strong policy class for imitation, planning, offline RL, and online RL (Janner et al., 2022; Chi et al., 2025; Wang et al., 2023; Hansen-Estruch et al., 2023; Ma et al., 2025; Wang et al., 2024b; Ding et al., 2024; Psenka et al., 2024). The key opportunity is to use offline logs as a behavioral prior that accelerates online diffusion-policy learning.

The difficulty is that practical logs are rarely clean evidence of action quality. They are often collected under hidden execution regimes, where unobserved factors influence the logged action, the action effectively executed by the environment, and the realized outcome. In robotics, such factors include friction, battery level, actuator wear, payload variation, and calibration drift (Wang et al., 2024a; Liu et al., 2025; Fey et al., 2025; Lin et al., 2026; Liu et al., 2026; Wen et al., 2025). In strategy optimization, they include user intent, demand state, competitive pressure, and operational context (Hou et al., 2025; Shao et al., 2025; Huang et al., 2025; Shen et al., 2025). In finance, they include liquidity, volatility, hidden order flow, execution pressure, and transaction costs (Nagy et al., 2025; Li et al., 2025a). Under such confounding, an action may

look useful because it co-occurred with favorable hidden conditions, not because it will transfer well online. This is more subtle than standard extrapolation error in offline RL (Fujimoto et al., 2019; Kumar et al., 2020; Levine et al., 2020; Kostrikov et al., 2022; Agarwal et al., 2020), because misleading actions may still lie inside the offline support.

A central observation is that biased logs often come with auxiliary telemetry. Robotic systems record motor currents, torque residuals, contacts, inertial signals, temperature, and battery diagnostics (Liu et al., 2025; Wen et al., 2025; Hou et al., 2025; Nagy et al., 2025; Jin et al., 2024; Yu et al., 2025; Li et al., 2025a; Fey et al., 2025). Recommendation and strategy systems record user activity, dwell time, inventory signals, calendar effects, and demand indicators (Hou et al., 2025; Hoo et al., 2024; Silva et al., 2024; Jin et al., 2024). Financial systems record bid-ask spreads, order-book depth, volume imbalance, fill ratios, message timing, and transaction-cost statistics (Nagy et al., 2025; Li et al., 2025a). These variables usually do not reveal the hidden regime or identify the exact interventional value. Instead, they provide imperfect proxy views of the conditions under which actions were selected and outcomes were realized. Motivated by proxy-based and proximal causal inference (Miao et al., 2018; Tchetgen Tchetgen et al., 2020; Deaner, 2021; Liu et al., 2023; Xu et al., 2021), we ask a practical question: which offline actions are reliable enough to guide online diffusion-policy learning?

We propose COOT, short for **C**ausally Motivated **O**ffline-to-**O**nline **T**ransfer. COOT treats offline data as biased but useful evidence. Its core is a proxy-informed reliability transfer bundle that estimates robust proxy-conditioned transfer value, bridge uncertainty, state potential, and behavior-support reliability. These signals ask whether an action appears beneficial relative to the state baseline, whether this benefit is stable across proxy views and resampled bridge models, and whether the action remains close to reliable offline behavior. COOT therefore does not clone the offline policy uniformly and does not treat logged returns as clean causal labels. It transfers offline experience in proportion to estimated reliability.

The causal role of COOT is deliberately modest but important. We do not claim that proxies identify the exact interventional value $\mathbb{E}[G \mid o, \mathrm{do}(a)]$ in general. Instead, the causal structure tells us what to distrust: the residual gap between the proxy-conditioned observational signal and the online effect needed for improvement. This makes the proxies more than extra conditioning features. They define a testable reliability criterion. If proxy information preserves useful views of the hidden execution regime, transfer should improve. If proxies are weakened, removed, shuffled, or made spurious, the benefit should decrease. Our proxy stress tests directly verify this behavior. Thus, COOT contributes a practical mechanism for using imperfect causal information in offline-to-online transfer without requiring exact identification.

COOT integrates this reliability principle across the full learning pipeline. During offline pretraining, reliability weights make the diffusion policy imitate robust, stable, and supported actions rather than the full behavior distribution. The state potential initializes the critic through potential-shaped targets. During online learning, the pretrained diffusion policy remains as a persistent prior. Candidate actions are sampled from both the current policy and the prior, then ranked by online value with proxy-derived risk and support penalties. This allows reliable offline structure to guide early exploration when the online critic is immature, while allowing online evidence to override biased offline guidance. COOT is complementary to existing diffusion-policy RL methods (Wang et al., 2023; Hansen-Estruch et al., 2023; Ma et al., 2025; Wang et al., 2024b; Ding et al., 2024; Psenka et al., 2024). It addresses not only how to optimize diffusion policies, but how they should reuse biased offline logs.

We evaluate COOT on five MuJoCo locomotion tasks under a matched same-data protocol. A Uniform Behavior Cloning warm-start receives the same offline transitions, pretraining budget, diffusion actor, and online budget as COOT, differing only in the reliability machinery. COOT improves over this generic-transfer control on every task in both peak return and sample efficiency, isolating reliability-weighted transfer from mere offline-data access. It also achieves the best performance under both metrics against eight strong actor-critic, diffusion-policy, energy-based, Q-guided, and score-based baselines: SAC (Haarnoja et al., 2018), DIPO (Yang et al., 2023), SDAC (Ma et al., 2025), DACER (Wang et al., 2024b), DPMD (Ma et al., 2025), iDEM (Akhound-Sadegh et al., 2024), QVPO (Ding et al., 2024), and QSM (Psenka et al., 2024). The largest gains occur early in training, showing that COOT provides more than a better initialization.

Component ablations, proxy stress tests, and covariate-shift diagnostics further show that the gains arise from proxy-informed, uncertainty-aware, and support-aware transfer rather than generic offline pretraining.

Our contributions are summarized as follows:

- **Confounded offline-to-online diffusion RL.** We identify hidden execution confounding as a practical failure mode for offline-to-online diffusion-policy RL, where unobserved regimes jointly affect logged actions, executed behavior, and outcomes. This explains why even supported offline actions can be unreliable for transfer.
- **Proxy-informed reliability transfer.** We introduce a proxy-informed reliability transfer bundle that converts biased offline logs into robust proxy-conditioned transfer value, bridge uncertainty, state potential, and behavior-support reliability. Proxies are used as imperfect views of hidden execution regimes, not as ordinary auxiliary features. The resulting disagreement and support signals estimate when offline actions should or should not guide online learning.
- **Unified COOT algorithm.** We integrate these reliability signals into diffusion actor-critic learning through selective policy pretraining, potential-shaped critic initialization, and prior-guided online candidate selection. COOT preserves useful offline structure while filtering fragile, uncertain, or unsupported logged correlations.
- **Strong empirical validation.** Across five MuJoCo tasks and five seeds, COOT improves over a matched same-data Uniform Behavior Cloning warm-start on every task in both peak return and sample efficiency, and further outperforms eight strong baselines under both metrics. Ablations, proxy stress tests, and covariate-shift diagnostics show that the gains come from reliability-weighted transfer rather than generic offline pretraining.

2 Problem Formulation: Offline-to-Online RL with Hidden Execution Confounding

We study offline-to-online reinforcement learning in which an agent first receives a fixed offline dataset and then improves through online interaction. Let $o_t \in \mathcal{O}$, $a_t \in \mathcal{A}$, $r_t \in \mathbb{R}$, and o_{t+1} denote observation, action, reward, and next observation. Most offline-to-online pipelines assume that the logged action is the action effectively executed by the environment, so that $(r_t, o_{t+1}) \sim P(\cdot \mid o_t, a_t)$. Under this premise, logged action-outcome associations can be reused as reliable signals for online improvement.

Our setting departs from this premise. In many deployed systems, logs are collected under *hidden execution regimes*: unobserved factors that affect which action is logged, how that action is realized, and what outcome follows. We model this with a latent variable $u_t \in \mathcal{U}$, unobserved by the learner, through the offline generative process

$$u_t \sim p(u_t \mid o_t), \quad a_t^{\log} \sim \mu(\cdot \mid o_t, u_t), \quad a_t^{\text{exec}} = g(o_t, a_t^{\log}, u_t), \quad (r_t, o_{t+1}) \sim P_u(\cdot \mid o_t, a_t^{\text{exec}}, u_t). \tag{1}$$

Eq. 1 is the formal distinction of our problem. The logged action $a_t^{\log}$ and executed action a_t^{exec} coincide only in the fully observed ideal case. The latent regime u_t may represent actuator wear, environment drift, operator intent, demand state, or market liquidity. The learner observes

$$\mathcal{D}_{\text{off}} = \{(o_i, a_i^{\log}, r_i, o_i', z_i, w_i)\}_{i=1}^N,$$

where z_i, w_i are proxy variables that provide partial information about u_i without revealing it.

The goal is to use $\mathcal{D}_{\text{off}}$ to accelerate online learning without treating logged returns as unbiased causal effects. Eq. 1 creates two sources of bias. First, the backdoor path $a_t^{\log} \leftarrow u_t \rightarrow (r_t, o_{t+1})$ makes logged action values regime dependent. Second, the execution path $u_t \rightarrow a_t^{\text{exec}} \rightarrow (r_t, o_{t+1})$ means that the logged action may differ from the action whose effect produced the outcome. Thus the observational quantity exposed by the log can differ from the interventional quantity needed online:

$$\underbrace{\mathbb{E}\Big[r_t \mid o_t, a_t^{\log}\Big]}_{\text{observational, logged}} \neq \underbrace{\mathbb{E}\Big[r_t \mid o_t, \text{do}(a_t^{\log})\Big]}_{\text{interventional, online}}. \tag{2}$$

This gap is invisible from (o, a, r, o') alone. A supported action may look good because it was logged under an easy regime, or look poor because it was executed under a difficult one.

We do not assume that Eq. 2 is fully identifiable from proxies. Instead, the causal model gives proxies a precise role. If z, w preserve information about the hidden regime, then the proxy-conditioned signal can remove part of the regime variation that contaminates raw logged returns. If the proxies are incomplete, their failure should appear as instability across proxy views, bootstrap bridge models, and support estimates. COOT exploits this partial regime information. It does not ask whether a logged action has high observed return. It asks whether the evidence for that action remains stable after conditioning on proxy views of the hidden execution regime.

3 COOT: Causally Motivated Offline-to-Online Transfer for Diffusion Policy Reinforcement Learning

3.1 From Confounded Logs to Causally Motivated Reliable Transfer

Eq. 1 and Eq. 2 show why standard offline reuse can fail even when logged actions are well supported. Behavior cloning imitates every logged action, and value initialization fits critics to observational returns. Both can therefore reinforce actions whose apparent value comes from favorable hidden regimes rather than from effects that remain useful online. The key transfer question is not only whether (o, a) is familiar. It is whether the offline evidence for that action is stable, proxy-consistent, and useful for online improvement.

COOT addresses the practical case where offline logs are biased but informative. We do not claim that $\mathbb{E}[G \mid o, \mathrm{do}(a)]$ is exactly identified. The causal structure is nevertheless essential because it specifies what must be estimated for safe transfer: the reliability of the proxy-conditioned evidence relative to the hidden-regime bias in Eq. 2. This separates COOT from a generic ensemble penalty. Without the confounding model, cross-proxy disagreement is only prediction noise. Under Eq. 1, it is evidence that the proxies have failed to reveal part of the hidden regime that made the logged action look useful. This gives a direct empirical prediction. Transfer should improve when proxies carry regime information, and should degrade when proxies are weakened, removed, shuffled, or made spurious. The stress tests in Appendix D.4 evaluate this prediction directly.

Assumptions for reliable COOT transfer

A1. Hidden bias	A latent regime u_t can influence logged actions, executed transitions, or rewards, so raw logged returns may be observationally biased.
A2. Informative proxies	Auxiliary variables z_t, w_t provide partial information about u_t or its outcome effects, but need not identify u_t or exact interventional values.
A3. Useful support	The offline data contain task-relevant state-action regions, so some logged actions remain useful as pretraining targets or online proposals.
A4. Reliable uncertainty	Proxy disagreement, bootstrap variation, and support scores are informative about when offline evidence is fragile or extrapolative.
A5. Online correction	The learner can collect online transitions, allowing reliable offline guidance to accelerate early learning while online evidence corrects biased priors.

Under these assumptions, COOT replaces uniform offline reuse with reliability-weighted transfer. Its proxy-informed reliability transfer bundle estimates three control-relevant properties: **benefit**, the proxy-conditioned advantage of a logged action over its state baseline, **stability**, the robustness of this advantage across proxy views and bootstrap resampling, and **support**, the credibility of evaluating this action within the logged behavior manifold. These properties determine how strongly each offline action should influence diffusion pretraining, critic initialization, and online candidate selection.

The remainder of this section implements this principle. Section 3.2 constructs the proxy-informed reliability transfer bundle. Section 3.3 uses it for selective diffusion-policy pretraining and potential-shaped critic initialization. Section 3.4 keeps the pretrained policy as a persistent prior whose proposals are used only when supported by online value, risk, and support scores. Algorithm 1 summarizes the full pipeline.

3.2 The Proxy-Informed Reliability Transfer Bundle

The bundle is the inferential core of COOT. It compiles the confounded dataset into four reusable functionals that answer the three reliability questions: a pessimistic proxy-stability value surrogate, a state potential, a deployable risk field, and a behavioral support penalty. We write a_i for the logged action $a_i^{\log}$ and G_i for its empirical discounted return.

Proxy views and a pessimistic stability surrogate. The regime u_i is unobserved, so value estimation cannot condition on it directly, and the interventional action value is in general not identifiable from proxies alone. COOT does not use z and w as adjustment variables and claims no proximal identification guarantee. In the hidden-regime setting of Eq. 1, conditioning on measurements of the latent regime can improve return prediction yet need not close every backdoor path and can mislead when the two proxies share a common bias. Proximal causal inference (Miao et al., 2018; Tchetgen Tchetgen et al., 2024; Xu et al., 2021; Deshpande et al., 2022) instead supplies only the guiding intuition that auxiliary measurements carrying regime information make residual sensitivity *detectable*. COOT therefore treats the proxies operationally, as auxiliary *reliability views* rather than causal adjustments, and fits one *bridge* predictor per view as a bootstrap ridge ensemble over resamples $\mathcal{I}_b$,

$$\hat{h}_\nu^{(b)} \in \arg\min_{h\in\mathcal{H}} \sum_{i\in\mathcal{I}_b} \big[G_i - h(o_i, a_i, \nu_i)\big]^2 + \alpha_{\mathrm{br}}\|h\|_2^2, \qquad \nu \in \{z, w\}, \;\; b = 1, \ldots, B, \tag{3}$$

each view approximating the proxy-conditioned return $\mathbb{E}[G \mid o, a, \nu]$. These predictions are not interpreted as causal effects. The two views are kept separate rather than concatenated, so that their *disagreement* becomes a usable warning signal: a return estimate that shifts with the slice of telemetry read is fragile evidence, while one that is stable across both windows is more dependable, though still not certified unbiased, since two views sharing a common bias could agree while both err. We therefore treat agreement as one cautious reliability cue, not as proof that the estimate reflects the action rather than the regime. Pooling the $2B$ predictions into $\mathcal{B}(o, a, z, w)$, we summarize them by their mean and dispersion and discount the latter,

$$\hat{Q}_{\mathrm{rob}} = \hat{\mu}_{\mathrm{br}} - \beta_{\mathrm{br}}\,\hat{\sigma}_{\mathrm{br}}, \qquad \hat{\mu}_{\mathrm{br}} = \frac{1}{2B}\sum_{v\in\mathcal{B}} v, \qquad \hat{\sigma}_{\mathrm{br}} = \Big[\tfrac{1}{2B}\sum_{v\in\mathcal{B}}\big(v - \hat{\mu}_{\mathrm{br}}\big)^2\Big]^{1/2}, \quad \beta_{\mathrm{br}} \geq 0, \tag{4}$$

where $\hat{\sigma}_{\mathrm{br}}$ fuses two unreliability sources into one scalar, the *epistemic* variance exposed by resampling and the *cross-view disagreement* between the z- and w-conditioned regressors. The surrogate $\hat{Q}_{\mathrm{rob}}$ thus rewards an action only when its proxy-conditioned return is *both* high *and* agreed upon. Under a mild calibration condition it lower-bounds the proxy-conditioned target value, not the interventional value (Appendix C.1), so the discount is a principled conservatism rather than an arbitrary penalty, the cautious departure from naive transfer that would propagate $\hat{\mu}_{\mathrm{br}}$ at face value. We make the residual limitation explicit. Shared proxy bias, where both views err in the same direction, can still pass the agreement check, so the proxy stress tests of Appendix D.4 measure directly how COOT behaves when proxy information is noisy, incomplete, spurious, or misaligned.

Advantage, risk, and support. Because a large $\hat{Q}_{\mathrm{rob}}$ may reflect a good *state* rather than a good *action*, COOT subtracts a state potential $\hat{\Phi}(o) \approx \mathbb{E}[G \mid o]$, a ridge fit on observations alone, to form the state-normalized robust advantage $\hat{A}_{\mathrm{rob}} = \hat{Q}_{\mathrm{rob}} - \hat{\Phi}(o)$, and the same $\hat{\Phi}$ is reused online as a dense shaping signal. Since $\hat{\sigma}_{\mathrm{br}}$ requires proxies but online candidates lack them, the dispersion is distilled across the offline-online boundary into a proxy-free risk field, and behavioral plausibility is scored by a discriminative support model $\hat{p}_{\mathrm{sup}}$ trained against three negative families (permuted, uniform, and locally perturbed actions),

$$\tilde{R} \in \arg\min_{R\in\mathcal{F}_R} \sum_i \big[\hat{\sigma}_{\mathrm{br},i} - R(o_i, a_i)\big]^2 + \alpha_R\|R\|_2^2, \quad \hat{R} = \max\{\tilde{R}, 0\}, \quad \hat{S} = -\log \mathrm{clip}\big(\hat{p}_{\mathrm{sup}}, \varepsilon_{\mathrm{sup}}, 1 - \varepsilon_{\mathrm{sup}}\big). \tag{5}$$

Being discriminative rather than density-based, $\hat{R}$ and $\hat{S}$ are cheap to evaluate on arbitrary online candidates, which is exactly where they are needed.

Composing reliability into a transfer weight. The three signals live on incomparable scales, so COOT standardizes each over $\mathcal{D}_{\text{off}}$ and fuses them into a reliability-adjusted desirability whose exponential, symmetrically clipped tilt becomes the transfer weight,

$$\omega_{\text{off}} = \exp\Big[\,\text{clip}\,\big(s_{\text{off}}, -c_{\text{off}}, c_{\text{off}}\big)\Big], \qquad s_{\text{off}} = \underbrace{\tfrac{\hat{A}_{\text{rob}}-\mu_A}{\sigma_A}}_{\text{benefit}} - \lambda_R \underbrace{\tfrac{\hat{R}-\mu_R}{\sigma_R}}_{\text{instability}} - \lambda_S \underbrace{\tfrac{\hat{S}-\mu_S}{\sigma_S}}_{\text{implausibility}}, \tag{6}$$

with empirical means and floored standard deviations taken over $\mathcal{D}_{\text{off}}$. The exponential turns an additive reliability budget into a multiplicative emphasis that integrates cleanly into a weighted likelihood (Sec. 3.3), while the clip bounds the leverage of any single sample. The construction is monotone in the intended directions, increasing in advantage and decreasing in instability and implausibility, and uniformly bounded.

Proposition 1 (Selective offline transfer)**.** *For samples i, j with unclipped scores $s_i, s_j \in [-c_{\text{off}}, c_{\text{off}}]$, the relative emphasis is $\omega_{\text{off},i}/\omega_{\text{off},j} = \exp(s_i - s_j)$, and every weight obeys $e^{-c_{\text{off}}} \le \omega_{\text{off}} \le e^{c_{\text{off}}}$.*

Appendix C.2 derives ω_{off} as the optimal exponential tilt of a KL-anchored selective-transfer program, and Appendix C.3 proves the proposition. The bundle is not discarded after pretraining: ω_{off} governs imitation, $\hat{\Phi}$ shapes the critic, and the proxy-free pair $(\hat{R}, \hat{S})$ survives into online selection. It is the single channel through which cautious, causally motivated reliability persists across the pipeline.

3.3 Weighted Diffusion Pretraining as Selective Imitation

COOT represents the policy as a conditional diffusion model π_θ (Ho et al., 2020; Song et al., 2021b;a; Chi et al., 2025; Wang et al., 2023; Kang et al., 2023; Ren et al., 2025) for a reason intrinsic to the problem: confounded logs are unavoidably multimodal, since different hidden regimes induce different behavior modes at the same state, and a unimodal actor would be forced to average across them. With the standard denoising loss $\ell_{\text{diff}}(\theta; o, a^{\text{g}}) = \mathbb{E}_{t,\epsilon}\big[\big\|\epsilon - \epsilon_\theta(o, a_t, t)\big\|_2^2\big]$ for guide action a^{g}, the conceptual move is in *how* that loss is aggregated. Ordinary cloning minimizes the unweighted expectation and thereby reproduces the behavior distribution together with its bias. COOT instead minimizes a reliability-weighted objective whose per-sample weight is batch-normalized to fix the gradient scale,

$$\mathcal{L}_{\text{diff}}(\theta) = \mathbb{E}_{(o,a,z,w)\sim\mathcal{D}_{\text{off}}}\big[\bar{\omega}_{\text{off}}\,\ell_{\text{diff}}(\theta; o, a)\big], \qquad \bar{\omega}_{\text{off},i} = \frac{\omega_{\text{off},i}}{|B|^{-1}\sum_{j\in B}\omega_{\text{off},j} + \varepsilon_\omega}. \tag{7}$$

Distributionally, Eq. 7 distills not the behavior policy $\mu(a \mid o)$ but a reweighted target $\propto \omega_{\text{off}}(o, a, z, w)\,\mu(a \mid o)$, shifting mass toward logged actions that are robustly advantageous, low-risk, and well supported, and away from those whose apparent quality is fragile or regime-dependent. The multimodal capacity of the diffusion model makes this reweighting expressive rather than mode-collapsing, retaining several reliable modes at once. The normalization preserves the bundle's ranking while keeping the influence of any sample provably bounded.

Proposition 2 (Bounded influence of weighted denoising)**.** *If $0 \le \ell_{\text{diff}}(\theta; o_i, a_i) \le L_{\max}$ for all $i \in B$, with ω_{off} clipped and normalized as in Eqs. 6 and 7, then $0 \le \bar{\omega}_{\text{off},i}\,\ell_{\text{diff}}(\theta; o_i, a_i) \le e^{2c_{\text{off}}} L_{\max}$.*

Appendix C.4 gives the proof, reconciling selectivity with stability: COOT can emphasize trustworthy samples aggressively without letting a handful dominate the update.

Potential-shaped critic initialization. The critics Q_{ϕ_1}, Q_{ϕ_2} are pretrained on offline transitions, but the same regimes that bias actions also bias raw rewards. COOT therefore injects the state structure already captured by $\hat{\Phi}$ as a dense, policy-order-preserving shaping term, and forms clipped double-Q targets (Fujimoto et al., 2018; Haarnoja et al., 2018) with $a' \sim \pi_{\bar{\theta}}(\cdot \mid o')$,

$$\tilde{r} = \rho\, r + \eta\big(\gamma\hat{\Phi}(o') - \hat{\Phi}(o)\big), \qquad y = \tilde{r} + \gamma(1-d)\min_{j\in\{1,2\}} Q_{\bar{\phi}_j}(o', a'), \qquad \mathcal{L}_Q = \sum_{j=1}^{2}\mathbb{E}\big[(Q_{\phi_j}(o,a) - y)^2\big]. \tag{8}$$

Potential-based shaping leaves the optimal policy invariant for any fixed bounded $\hat{\Phi}$ (Ng et al., 1999) (Appendix C.5), so densification injects early signal without altering the solution. The weight η is annealed

Algorithm 1 COOT: Causally Motivated Offline-to-Online Transfer

Require: Confounded offline dataset $\mathcal{D}_{\text{off}}$, online environment, diffusion actor critic $(\pi_\theta, Q_{\phi_1}, Q_{\phi_2})$.

1: **Fit proxy-informed reliability transfer bundle:** bridge ensembles $\{\hat{h}_z^{(b)}, \hat{h}_w^{(b)}\}$ (Eq. 3), pessimistic surrogate $\hat{Q}_{\text{rob}}$ (Eq. 4), potential $\hat{\Phi}$, risk $\hat{R}$, support $\hat{S}$ (Eq. 5), and weights ω_{off} (Eq. 6).
2: **Selective offline pretraining.**
3: **for** each offline update **do**
4: Sample $(o, a, r, o', d, z, w) \sim \mathcal{D}_{\text{off}}$. Set guide $a^{\text{g}} \leftarrow a$ and weight $\omega^{\text{g}} \leftarrow \omega_{\text{off}}(o, a, z, w)$.
5: Shape reward and update critics (Eq. 8). Update π_θ by weighted denoising toward a^{g} (Eq. 7). Polyak-update targets.
6: **end for**
7: Freeze the target policy as the persistent prior π_{prior}.
8: **Prior-guided online learning.**
9: **for** each online step **do**
10: Observe o_t, build $\mathcal{C}(o_t) = \mathcal{C}_\theta(o_t) \cup \mathcal{C}_{\text{prior}}(o_t)$, and execute $a_t = \arg\max_{a \in \mathcal{C}(o_t)} \mathcal{J}_{\text{on}}(o_t, a)$ (Eq. 9), then store in replay.
11: Sample replay and recompute guides $a^{\text{g}}(o)$ and weights $\omega_{\text{on}}(o)$ (Eq. 10). Update critics on *executed* transitions and π_θ toward *selected* guides (Eq. 11), then anneal shaping and update targets.
12: **end for**

online (Appendix B.3) so the learned potential bootstraps early value estimation and then yields to environment reward. The two channels are complementary by design: the actor is distilled only toward reliable logged actions, while the critic learns a shaped landscape over all transitions. After pretraining, the target actor is frozen as the persistent prior $\pi_{\text{prior}} \leftarrow \pi_{\bar{\theta}}$.

3.4 Prior-Guided Online Learning: Retaining Offline Structure Under Correction

Many offline-to-online methods use the offline policy only as an initialization and then overwrite it as updates accumulate, discarding reliable offline structure precisely when early online estimates are noisiest. COOT instead keeps the pretrained policy alive as a *persistent proposal distribution* and arbitrates between it and the adapting policy through a reliability-regularized value. Candidates are drawn from both sources and scored by online value penalized by the surviving offline signals,

$$\mathcal{C}(o) = \{a_k^\theta \sim \pi_\theta(\cdot \mid o)\}_{k=1}^K \cup \{a_m^{\text{prior}} \sim \pi_{\text{prior}}(\cdot \mid o)\}_{m=1}^M, \qquad \mathcal{J}_{\text{on}}(o, a) = \underbrace{\min_{j \in \{1,2\}} Q_{\phi_j}(o, a)}_{\text{online evidence}} - \underbrace{\lambda_R^{\text{on}} \hat{R}(o, a)}_{\text{risk}} - \underbrace{\lambda_S^{\text{on}} \hat{S}(o, a)}_{\text{support}}. \tag{9}$$

The interplay of the three terms is the crux of the mechanism. The critic supplies evidence that sharpens with experience, while $\hat{R}$ and $\hat{S}$, the proxy-free survivors of the bundle, act as priors over reliability that matter most while the critic is poorly calibrated: early on they suppress speculative candidates in regions the offline data never vouched for, and as the critic sharpens its term dominates, freeing online learning to override offline guidance that proves suboptimal. The same rule supplies guide actions and their weights for the actor,

$$a^{\text{g}}(o) \in \arg\max_{a \in \mathcal{C}(o)} \mathcal{J}_{\text{on}}(o, a), \qquad \omega_{\text{on}}(o) = \exp\Big[\operatorname{clip}\big(\tau_{\text{act}}^{-1} \mathcal{J}_{\text{on}}(o, a^{\text{g}}(o)), -c_{\text{on}}, c_{\text{on}}\big)\Big], \tag{10}$$

feeding the batch-normalized online objective

$$\mathcal{L}_{\text{diff,on}}(\theta) = \mathbb{E}_{o \sim \mathcal{B}_{\text{replay}}}\big[\bar{\omega}_{\text{on}}(o)\, \ell_{\text{diff}}(\theta; o, a^{\text{g}}(o))\big]. \tag{11}$$

A deliberate asymmetry governs the two updates: the critic is trained only on *executed* transitions, anchoring value estimates to real online data, while the actor is distilled toward the *selected* guides, which may originate from the prior (Appendix B.4). This lets COOT exploit offline structure for proposal diversity while ensuring value learning never drifts from on-policy reality, so the single denoising interface serves both stages, unifying selective imitation and value-guided improvement under one expressive policy class.

4 Experiments

Our experiments test a single thesis, that COOT's gains come from *proxy-informed reliability transfer* rather than from the generic ingredients it shares with prior work. We therefore make the same-data Uniform Behavior Cloning warm-start control a central main-table comparison, so the offline data, pretraining budget, and diffusion actor are matched before measuring the benefit of reliability-weighted transfer. We do not claim to recover exact interventional action values, only that proxy-conditioned reliability signals, motivated by the confounding structure of Section 2, are what drive the improvement. We frame the study as four falsifiable claims, each tied to a component of Section 3 and each established either below or, where noted, in the appendix. (i) *Sufficiency*: against strong actor-critic and diffusion-policy baselines, the reliability-weighted prior should accelerate online learning, with the sharpest gains in the early-budget regime where the offline signal is most valuable (below). (ii) *Generality*: the effect should persist across locomotion tasks spanning a wide range of state and action dimensions, since the transfer principle is not tuned to any one geometry (below). (iii) *Attribution*: ablating the bridge, prior, shaping, or support pathway should each remove a distinct portion of the gain, and degrading the proxies should degrade performance *gracefully and monotonically*, the behavioral fingerprint of a method genuinely reading proxy information rather than exploiting an incidental regularizer (Appendix D.4). (iv) *Necessity*: the offline data must be measurably confounded relative to the target environment, so that face-value reuse would indeed be biased and the machinery of Section 3 is warranted (Appendix D.5).

4.1 Experimental Setup

Tasks and offline data. We use five MuJoCo locomotion benchmarks, HalfCheetah, Ant, Hopper, Humanoid, and Walker2d, chosen because their state and action dimensions span a wide range and thus test claim (ii) (Todorov et al., 2012; Duan et al., 2016). Each follows the protocol of Section 2: COOT is pretrained on a confounded offline dataset, generated by the latent-regime process of Eq. 1 and described in Appendix D.3, then improves online. Unless noted, we use 300k offline transitions and 30k pretraining updates.

Protocol and metrics. We focus on the first 500k online environment steps, the regime in which transfer is most valuable, and evaluate every 5k steps over five random seeds. We summarize each run by two complementary scalars computed under a *seed-first* protocol (Zeng et al., 2023; Ma et al., 2025; Adkins et al., 2024; Agarwal et al., 2021; Vincent et al., 2025), meaning the per-seed statistic is formed first and only then aggregated across seeds. This is more robust than reading a peak off a single across-seed mean curve and yields tighter, less optimistic estimates, and it follows standard learning-curve-based evaluation practice in reinforcement learning (Machado et al., 2018; Agarwal et al., 2021; Jordan et al., 2024). The first metric is the best return attained within the budget. For each seed s we take

$$R^{500\mathrm{k}}_{\mathrm{best},s} = \max_{t_i \le 500\mathrm{k}} R_s(t_i),$$

the strongest evaluated policy found by that seed, and report $R^{500\mathrm{k}}_{\mathrm{best}}$ as the mean $\pm$ standard deviation across seeds. Reporting peak training performance is a long-established evaluation summary (Machado et al., 2018; Yang et al., 2023; Ding et al., 2024; Ma et al., 2025). The second metric is the time-normalized area under the learning curve, computed separately for each seed by the trapezoidal rule (Stadie et al., 2015; Machado et al., 2018; He et al., 2022),

$$\mathrm{AUC}^{500\mathrm{k}}_s = \frac{1}{t_n - t_1} \sum_{i=1}^{n-1} \frac{R_s(t_i) + R_s(t_{i+1})}{2} (t_{i+1} - t_i), \qquad t_n \le 500\mathrm{k}. \tag{12}$$

We report $\mathrm{AUC}^{500\mathrm{k}}$ as the mean $\pm$ standard deviation of $\mathrm{AUC}^{500\mathrm{k}}_s$ across seeds. Dividing by $t_n - t_1$ converts the area into the average return sustained over the evaluated online horizon within each task. This pairing separates two forms of early online progress: $R^{500\mathrm{k}}_{\mathrm{best}}$ measures the strongest policy found within the budget, while $\mathrm{AUC}^{500\mathrm{k}}$ measures how consistently the method converts offline evidence into sustained online improvement.

Baselines. We include two types of comparisons. The first is a matched offline-transfer control, *Uniform Behavior Cloning warm-start*, which uses the same 300k confounded offline transitions, the same 30k pretraining updates, and the same diffusion actor architecture as COOT, but assigns unit weight to every offline action and removes the proxy bridge, bridge uncertainty, risk penalty, support penalty, potential shaping, and persistent prior candidates. This is the direct test of whether COOT improves beyond generic offline pretraining. The second comparison set contains eight representative online optimizers spanning the relevant design space: the off-policy actor-critic SAC (Haarnoja et al., 2018), and the diffusion, energy-based, Q-guided, and score-based policy optimizers DIPO (Yang et al., 2023), SDAC and DPMD (Ma et al., 2025), DACER (Wang et al., 2024b), iDEM (Akhound-Sadegh et al., 2024), QVPO (Ding et al., 2024), and QSM (Psenka et al., 2024). All methods share the same online evaluation protocol and metrics.

4.2 Experimental Results

Same-data evidence for reliability-weighted transfer. Tables 1 and 2 report R^{500k}_{best} and AUC^{500k} under the seed-first protocol of Section 4.1. The most important fairness comparison is the Uniform Behavior Cloning warm-start control. It receives exactly the same 300k confounded offline transitions, the same 30k offline pretraining updates, and the same diffusion actor architecture as COOT. The only difference is the transfer rule: Uniform Behavior Cloning imitates all offline actions equally and then continues online learning without the proxy bridge, uncertainty risk, support penalty, potential shaping, or persistent offline prior candidates.

This matched control directly tests whether the improvement comes from generic offline pretraining or from COOT's reliability-transfer mechanism. COOT outperforms Uniform Behavior Cloning warm-start on every task under both metrics. The gains are 16.0%, 6.5%, 38.4%, 86.1%, and 16.7% in R^{500k}_{best}, and 9.4%, 32.2%, 64.2%, 56.5%, and 49.7% in AUC^{500k} on HalfCheetah, Ant, Hopper, Humanoid, and Walker2d. Averaging these per-task relative gains over the matched control gives 32.7% higher peak early return and 42.4% higher time-normalized AUC. Because this comparison fixes offline data access, pretraining compute, policy class, online budget, and evaluation protocol, the remaining advantage isolates the value of selective, proxy-informed reliability transfer.

The broader online baselines provide complementary context. COOT is the top method on every one of the five domains under both metrics, which supports the generality claim that the same transfer principle works across distinct state-action geometries. The matched Behavior Cloning control is often already much stronger than from-scratch online learning, confirming that the offline data are useful. COOT's further improvement shows that the key question is not whether to use offline data, but how to decide which offline actions should remain influential after biased logging. This is the central role of ω_{off}, $\hat{R}$, $\hat{S}$, and the persistent prior in Section 3.

Learning-curve interpretation. Figure 1 consolidates early learning across all environments and baseline families, and the curves make the mechanism visible. COOT shifts the trajectories both upward and leftward within the budget, which is the graphical content of its AUC^{500k} advantage in Table 2. The deeper point is that COOT changes the *search problem* the diffusion policy faces rather than merely its starting point. Baseline diffusion policies must locate high-return action modes through online critic learning alone, paying the full exploration cost, whereas COOT instead enters online training with a proxy-informed prior and, through Eq. 9, evaluates every candidate as online value discounted by the surviving offline risk $\hat{R}$ and support $\hat{S}$. Early on, when $\min_j Q_{\phi_j}$ is noisy, these proxy-free penalties dominate and steer interaction away from regions the offline data never vouched for, and as the critic sharpens its term takes over while the offline penalties recede, so the prior is exploited but never blindly trusted. The curves thus corroborate the central design claim that confounded offline experience accelerates online RL when it is converted into risk- and support-aware guidance, not when it is imitated as clean demonstration.

Coverage versus distillation. Figure 2 disentangles the two knobs that control the offline signal, namely how much evidence the dataset contains and how thoroughly the actor distills it. Fixing pretraining and varying offline size, then fixing size and varying pretraining, exposes an interaction rather than two independent effects. With 300k transitions, longer pretraining produces a clean ordering of curves and the 30k-update setting is strongest, whereas with only 100k transitions the pretraining curves collapse together. The reading

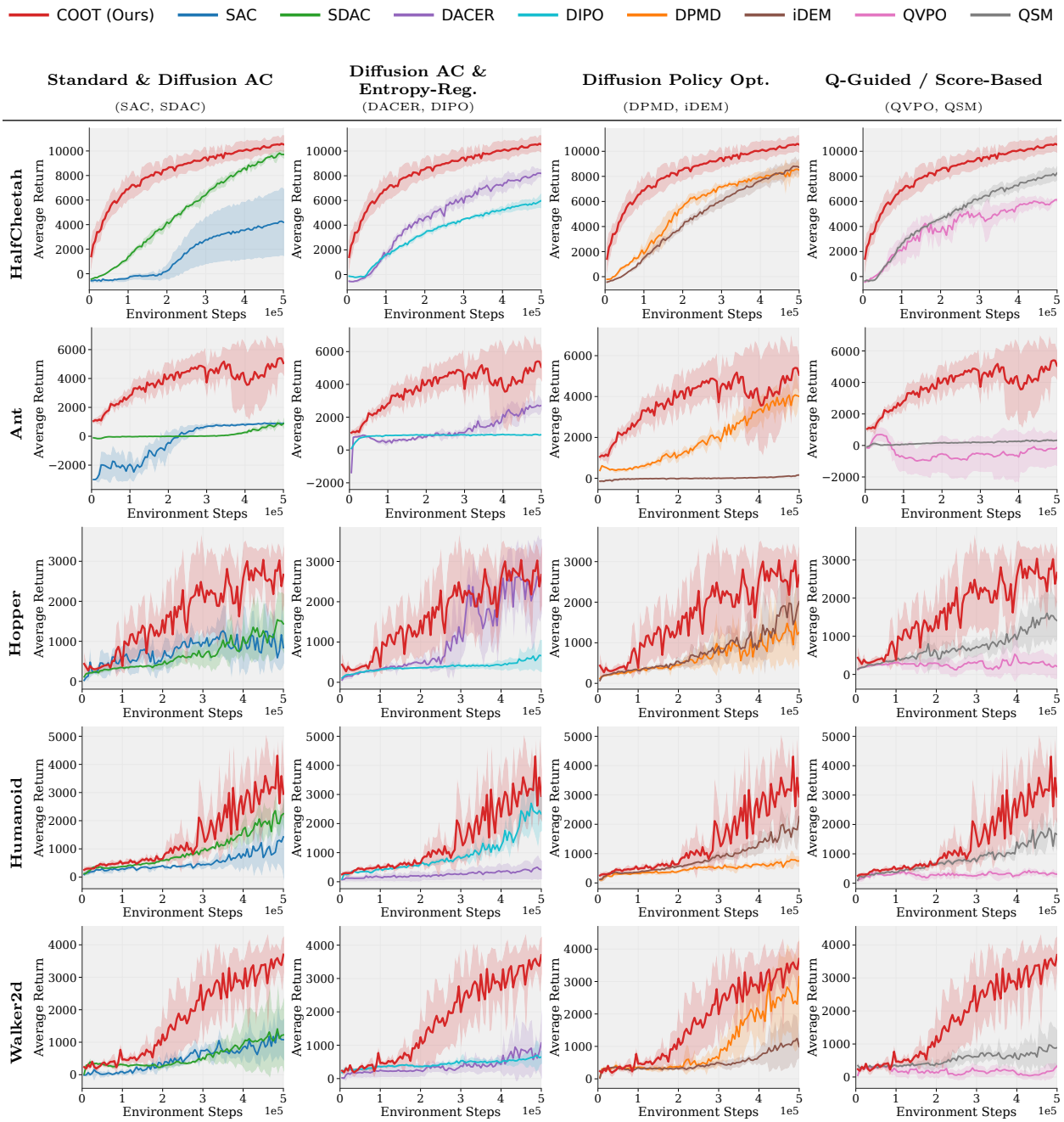

Figure 1: **Early online learning curves up to** 500k **environment steps** across five MuJoCo locomotion tasks. Rows are environments, and columns group COOT against one baseline family, namely standard and diffusion actor-critic methods, entropy-regularized diffusion actor-critic methods, diffusion-policy optimization methods, and Q-guided or score-based diffusion methods. Curves show across-seed mean evaluation return over five random seeds, with shaded bands denoting ± 1 standard deviation. The consolidated view shows that COOT shifts trajectories both upward and leftward, consistently accelerating early online learning across diverse dynamics and structurally different diffusion-policy baselines.

Table 1: **Best return within** 500k **online environment steps** on five MuJoCo continuous-control tasks (higher is better). Under the seed-first protocol, we first compute each seed's peak evaluation return over steps $\leq$ 500k, then report the across-seed mean $\pm$ standard deviation over five seeds. Uniform Behavior Cloning warm-start is the same-data control: it uses the same 300k offline transitions, 30k pretraining updates, and diffusion actor as COOT, but transfers all offline actions uniformly and removes COOT's reliability-transfer mechanisms. Best per column in **bold**.

Method	**HalfCheetah** ↑	**Ant** ↑	**Hopper** ↑	**Humanoid** ↑	**Walker2d** ↑
Conventional model-free RL					
SAC (Haarnoja et al., 2018)	4334.2 ± 2792.5	934.9 ± 35.0	1835.4 ± 461.8	1809.1 ± 1439.4	1743.2 ± 490.1
Diffusion and energy-based RL					
DIPO (Yang et al., 2023)	6013.7 ± 582.2	964.8 ± 5.9	680.5 ± 394.6	3342.4 ± 664.4	777.4 ± 396.7
QVPO (Ding et al., 2024)	6386.7 ± 212.6	1092.1 ± 203.7	851.0 ± 564.8	567.5 ± 68.2	482.6 ± 65.5
DACER (Wang et al., 2024b)	8413.0 ± 463.3	3115.0 ± 341.8	3020.2 ± 467.1	541.0 ± 416.7	1106.2 ± 1092.5
QSM (Psenka et al., 2024)	8298.0 ± 596.4	418.5 ± 51.9	1842.3 ± 706.7	2233.2 ± 410.2	1184.8 ± 690.2
DPMD (Ma et al., 2025)	8669.4 ± 690.8	4429.5 ± 620.7	1971.5 ± 1040.3	930.1 ± 177.5	3694.7 ± 270.3
iDEM (Akhound-Sadegh et al., 2024)	9009.3 ± 619.0	177.5 ± 28.5	2415.1 ± 841.0	2364.8 ± 569.7	1367.0 ± 743.7
SDAC (Ma et al., 2025)	9922.8 ± 279.4	1078.9 ± 254.8	2001.4 ± 735.3	2669.8 ± 441.3	1452.8 ± 1375.9
Same-data offline-transfer control					
Uniform Behavior Cloning warm-start	9215.3 ± 75.6	5199.3 ± 238.5	2335.7 ± 875.0	2564.4 ± 607.7	3421.3 ± 543.8
COOT (Ours)	**10688.9 ± 619.0**	**5538.6 ± 853.8**	**3231.6 ± 91.1**	**4771.2 ± 517.7**	**3991.1 ± 357.6**

Table 2: **Time-normalized AUC up to** 500k **online environment steps** on five MuJoCo continuous-control tasks (higher is better). Under the seed-first protocol, we first compute each seed's trapezoidal AUC of evaluation return over steps $\leq$ 500k, normalized by the elapsed horizon to the average return sustained over the early online budget, then report the across-seed mean $\pm$ standard deviation over five seeds. Uniform Behavior Cloning warm-start is the same-data control and isolates generic offline pretraining from COOT's reliability-weighted transfer. Best per column in **bold**.

Method	**HalfCheetah** ↑	**Ant** ↑	**Hopper** ↑	**Humanoid** ↑	**Walker2d** ↑
Conventional model-free RL					
SAC (Haarnoja et al., 2018)	1647.3 ± 1274.7	−340.2 ± 306.6	760.7 ± 191.7	505.8 ± 202.2	493.5 ± 115.6
Diffusion and energy-based RL					
DIPO (Yang et al., 2023)	3486.2 ± 106.7	884.2 ± 12.6	386.2 ± 68.6	922.3 ± 91.4	437.6 ± 62.7
QVPO (Ding et al., 2024)	3928.0 ± 359.1	−467.8 ± 919.4	262.9 ± 147.5	299.1 ± 62.0	174.5 ± 67.4
DACER (Wang et al., 2024b)	4759.9 ± 280.2	1231.8 ± 77.9	1114.7 ± 305.9	247.9 ± 165.2	373.3 ± 226.2
QSM (Psenka et al., 2024)	4958.8 ± 229.5	175.3 ± 17.2	717.9 ± 186.6	815.6 ± 45.3	530.1 ± 178.2
DPMD (Ma et al., 2025)	5399.6 ± 316.1	1880.8 ± 211.4	653.1 ± 198.9	475.6 ± 46.6	1023.7 ± 289.8
iDEM (Akhound-Sadegh et al., 2024)	4726.1 ± 343.7	1.7 ± 2.7	779.2 ± 168.2	844.1 ± 84.9	501.3 ± 123.7
SDAC (Ma et al., 2025)	5082.9 ± 84.2	123.4 ± 35.3	656.1 ± 170.7	925.1 ± 94.1	559.8 ± 273.8
Same-data offline-transfer control					
Uniform Behavior Cloning warm-start	7652.7 ± 106.9	2871.7 ± 940.5	1011.8 ± 380.7	922.1 ± 155.5	1167.7 ± 348.8
COOT (Ours)	**8370.5 ± 657.5**	**3797.6 ± 906.9**	**1661.8 ± 466.1**	**1442.9 ± 360.1**	**1748.4 ± 413.0**

is precise and consistent with the framework. Dataset size controls how informative the proxy-conditioned bundle can be, because the calibration underlying $\hat{Q}_{\text{rob}}$ (Appendix C.1) and the support boundary $\hat{S}$ both sharpen with coverage, whereas pretraining controls how much of that information is actually transferred into the actor. Additional optimization cannot manufacture evidence absent from the data's support, which is why longer pretraining helps only once coverage is sufficient. The benefit is thus not more behavior cloning but more reliable cloning, exactly the distinction the reliability weight is built to make.

The gains track proxy reliability, not incidental regularization (claims iii-iv). Two appendix studies isolate the source of the improvement. The **component ablations and proxy stress tests** (Appendix D.4) show that each pathway, the bridge, persistent prior, potential shaping, and support penalty, carries a distinct and non-recoverable share of the gain, with the proxy bridge dominant, and that performance falls *monotonically* as proxies are attenuated, dropped, or misaligned. This ordered decay is itself the diagnostic: a method riding an incidental regularizer would be indifferent to proxy quality, and a brittle

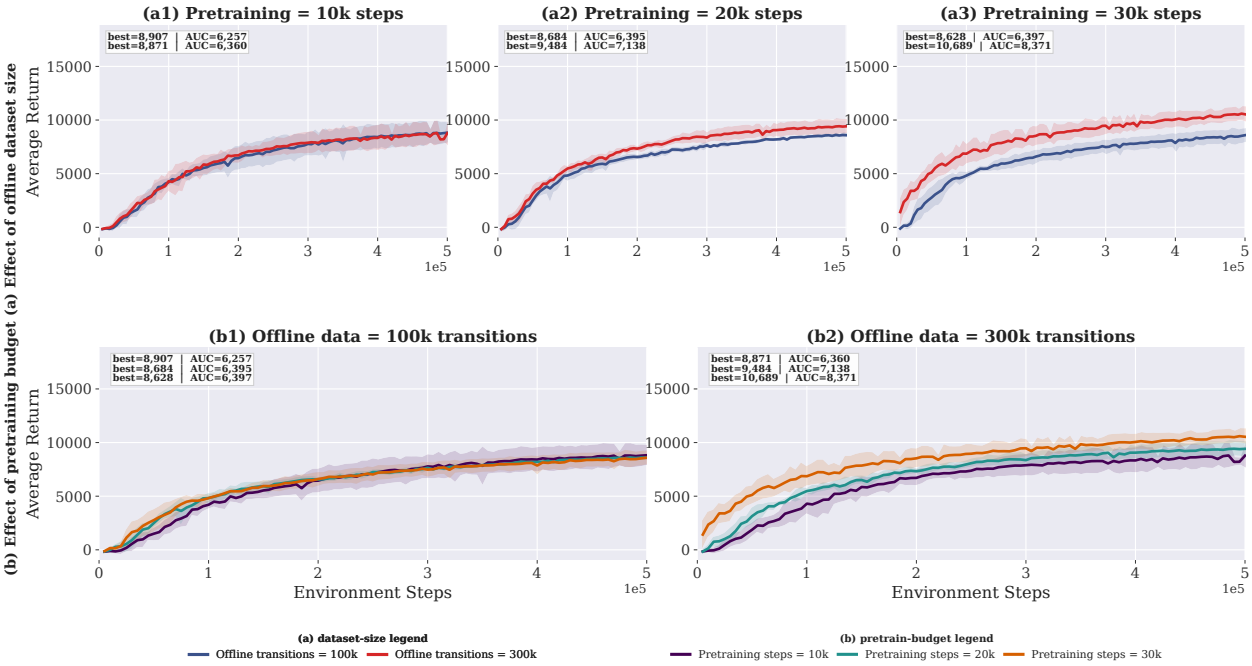

Figure 2: **COOT sensitivity to offline coverage and pretraining budget** on HalfCheetah up to 500k online steps, separating offline dataset size (top) from offline pretraining budget (bottom). Larger datasets help only when paired with sufficient pretraining, while longer pretraining barely helps under thin coverage. The interaction indicates that COOT's benefit comes from distilling reliable proxy-conditioned signal, not from more offline imitation.

one would collapse, whereas COOT tracks the amount of usable proxy information, exactly as a mechanism that reads proxy-conditioned reliability should. The **covariate-shift diagnostics** (Appendix D.5) supply the necessity check, showing the offline state-action distribution is strongly separable from ordinary interaction (domain-classifier AUC 0.87-0.997), so face-value reuse would import real bias and the cautious, proxy-conditioned design of Section 3 is warranted rather than ornamental.

5 Conclusion

We introduced COOT, a causally motivated framework for offline-to-online diffusion-policy reinforcement learning from confounded logs. Its central contribution is a reliability principle made operational: offline experience should be transferred in proportion to estimated proxy-conditioned reliability rather than raw observed return. A single proxy-informed bundle estimates robust transfer value, bridge uncertainty, state potential, and behavior support, and these signals jointly drive weighted diffusion pretraining, potential-shaped critic learning, and prior-guided online exploration. Against a same-data Uniform Behavior Cloning warm-start with the identical dataset, pretraining budget, and actor, COOT improves on every task, isolating the gain to how biased data are reused rather than to data or compute. It also leads eight strong baselines in peak return and sample efficiency, most sharply in early online learning, and ablations and proxy stress tests attribute the advantage to the reliability mechanism itself, since performance tracks proxy quality and degrades gracefully as the proxies weaken. COOT thus reframes biased logs not as data to copy, discard, or pessimistically constrain, but as imperfect evidence whose transfer reliability can be estimated and corrected online, a practical path toward sample-efficient learning in the telemetry-rich systems where offline data are abundant but rarely clean.

References

Jacob Adkins, Michael Bowling, and Adam White. A method for evaluating hyperparameter sensitivity in reinforcement learning. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and

C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37. Curran Associates, Inc., 2024. URL https://openreview.net/forum?id=4OJdZhcwBb.

Rishabh Agarwal, Dale Schuurmans, and Mohammad Norouzi. An optimistic perspective on offline reinforcement learning. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 104–114. PMLR, 2020.

Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron C. Courville, and Marc G. Bellemare. Deep reinforcement learning at the edge of the statistical precipice. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 29304–29320. Curran Associates, Inc., 2021.

Tara Akhound-Sadegh, Jarrid Rector-Brooks, Avishek Joey Bose, Sarthak Mittal, Pablo Lemos, Cheng-Hao Liu, Marcin Sendera, Siamak Ravanbakhsh, Gauthier Gidel, Yoshua Bengio, Nikolay Malkin, and Alexander Tong. Iterated denoising energy matching for sampling from boltzmann densities. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 760–786. PMLR, 2024.

Peter C. Austin. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Statistics in Medicine*, 28(25):3083–3107, 2009. doi: 10.1002/sim.3697.

Philip J. Ball, Laura Smith, Ilya Kostrikov, and Sergey Levine. Efficient online reinforcement learning with offline data. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 1577–1594. PMLR, 2023.

Tianyu Chen, Zhendong Wang, and Mingyuan Zhou. Diffusion policies creating a trust region for offline reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, 44(10–11):1684–1704, 2025. doi: 10.1177/02783649241273668.

Ben Deaner. Proxy controls and panel data. *arXiv preprint arXiv:1810.00283*, 2021.

Shachi Deshpande, Kaiwen Wang, Dhruv Sreenivas, Zheng Li, and Volodymyr Kuleshov. Deep multi-modal structural equations for causal effect estimation with unstructured proxies. In S. Koyejo, S. Mohamed, A. Agarwal, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35. Curran Associates, Inc., 2022.

Shutong Ding, Ke Hu, Zhenhao Zhang, Kan Ren, Weinan Zhang, Jingyi Yu, Jingya Wang, and Ye Shi. Diffusion-based reinforcement learning via q-weighted variational policy optimization. *Advances in Neural Information Processing Systems*, 37:53945–53968, 2024.

Yan Duan, Xi Chen, Rein Houthooft, John Schulman, and Pieter Abbeel. Benchmarking deep reinforcement learning for continuous control. In Maria Florina Balcan and Kilian Q. Weinberger (eds.), *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1329–1338. PMLR, 2016.

Nolan Fey, Gabriel B. Margolis, Martin Peticco, and Pulkit Agrawal. Bridging the sim-to-real gap for athletic loco-manipulation. In *Proceedings of Robotics: Science and Systems (RSS)*, Los Angeles, CA, USA, 2025. doi: 10.15607/RSS.2025.XXI.125.

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

Scott Fujimoto, David Meger, and Doina Precup. Off-policy deep reinforcement learning without exploration. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2052–2062. PMLR, 2019. URL https://proceedings.mlr.press/v97/fujimoto19a.html.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012.

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

James A. Hanley and Barbara J. McNeil. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1):29–36, 1982. doi: 10.1148/radiology.143.1.7063747.

Philippe Hansen-Estruch, Ilya Kostrikov, Michael Janner, Jakub Grudzien Kuba, and Sergey Levine. IDQL: Implicit Q-learning as an actor-critic method with diffusion policies. *arXiv preprint arXiv:2304.10573*, 2023.

Tairan He, Yuge Zhang, Kan Ren, Minghuan Liu, Che Wang, Weinan Zhang, Yuqing Yang, and Dongsheng Li. Reinforcement learning with automated auxiliary loss search. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 1820–1834. Curran Associates, Inc., 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html.

Shi Bin Hoo, Samuel Müller, David Salinas, and Frank Hutter. The tabular foundation model TabPFN outperforms specialized time series forecasting models based on simple features. In *NeurIPS 2024 Workshop on Table Representation Learning*, 2024. URL https://openreview.net/forum?id=H02X7RO3OC.

Yupeng Hou, Jianmo Ni, Zhankui He, Noveen Sachdeva, Wang-Cheng Kang, Ed H. Chi, Julian McAuley, and Derek Zhiyuan Cheng. ActionPiece: Contextually tokenizing action sequences for generative recommendation. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*. PMLR, 2025. Spotlight.

Hao Hu, Yanchao Yang, Qianchuan Zhao, and Chongjie Zhang. Bayesian design principles for offline-to-online reinforcement learning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 20731–20755. PMLR, 2024.

Kexin Huang, Ziqian Chen, Xue Wang, Chongming Gao, Jinyang Gao, Bolin Ding, and Xiang Wang. Learning Bayesian Nash equilibrium in auction games via approximate best response. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 25159–25184. PMLR, 2025.

Michael Janner, Yilun Du, Joshua B. Tenenbaum, and Sergey Levine. Planning with diffusion for flexible behavior synthesis. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 9902–9915. PMLR, 2022.

Tianying Ji, Yongyuan Liang, Yan Zeng, Yu Luo, Guowei Xu, Jiawei Guo, Ruijie Zheng, Furong Huang, Fuchun Sun, and Huazhe Xu. ACE: Off-policy actor-critic with causality-aware entropy regularization.

In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 21620–21647. PMLR, 2024.

Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Y. Zhang, Xiaoming Shi, Pin-Yu Chen, Yuxuan Liang, Yuan-Fang Li, Shirui Pan, and Qingsong Wen. Time-LLM: Time series forecasting by reprogramming large language models. In *International Conference on Learning Representations*, 2024.

Scott M. Jordan, Adam White, Bruno Castro Da Silva, Martha White, and Philip S. Thomas. Position: Benchmarking is limited in reinforcement learning research. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 22551–22569. PMLR, 2024.

Mateo Juliani, Mingxuan Li, and Elias Bareinboim. Confounding robust continuous control via automatic reward shaping. *arXiv preprint arXiv:2602.10305*, 2026.

Bingyi Kang, Xiao Ma, Chao Du, Tianyu Pang, and Shuicheng Yan. Efficient diffusion policies for offline reinforcement learning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 67195–67212. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/d45e0bfb5a39477d56b55c0824200008-Abstract-Conference.html.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In S. Koyejo, S. Mohamed, A. Agarwal, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35. Curran Associates, Inc., 2022.

Ilya Kostrikov, Ashvin Nair, and Sergey Levine. Offline reinforcement learning with implicit Q-learning. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=68n2s9ZJWF8.

Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative q-learning for offline reinforcement learning. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1179–1191. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/0d2b2061826a5df3221116a5085a6052-Abstract.html.

Aviral Kumar, Joey Hong, Anikait Singh, and Sergey Levine. Should i run offline reinforcement learning or behavioral cloning? In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=AP1MKT37rJ.

Erich Kummerfeld, Jaewon Lim, and Xu Shi. Data-driven automated negative control estimation (DANCE): Search for, validation of, and causal inference with negative controls. *Journal of Machine Learning Research*, 25(106):1–35, 2024.

Seunghyun Lee, Younggyo Seo, Kimin Lee, Pieter Abbeel, and Jinwoo Shin. Offline-to-online reinforcement learning via balanced replay and pessimistic Q-ensemble. In Aleksandra Faust, David Hsu, and Gerhard Neumann (eds.), *Proceedings of the 5th Conference on Robot Learning*, volume 164 of *Proceedings of Machine Learning Research*, pp. 1702–1712. PMLR, 2022. URL https://proceedings.mlr.press/v164/lee22d.html.

Seunghyun Lee, Younggyo Seo, Kimin Lee, Pieter Abbeel, and Jinwoo Shin. Offline-to-online reinforcement learning via balanced replay and pessimistic q-ensembles. In *Proceedings of the 40th International Conference on Machine Learning*, Proceedings of Machine Learning Research. PMLR, 2023.

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.

Junjie Li, Yang Liu, Weiqing Liu, Shikai Fang, Lewen Wang, Chang Xu, and Jiang Bian. MarS: A financial market simulation engine powered by generative foundation model. In *The Thirteenth International Conference on Learning Representations*, 2025a. URL https://openreview.net/forum?id=Yqk7EyT52H.

Mingxuan Li, Junzhe Zhang, and Elias Bareinboim. Causally aligned curriculum learning. In *International Conference on Learning Representations*, 2024.

Mingxuan Li, Junzhe Zhang, and Elias Bareinboim. Automatic reward shaping from confounded offline data. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 36765–36793. PMLR, 2025b.

Mingxuan Li, Junzhe Zhang, and Elias Bareinboim. Confounding robust deep reinforcement learning: A causal approach. In *Advances in Neural Information Processing Systems*, volume 38, pp. 138292–138325, 2025c.

Mingxuan Li, Junzhe Zhang, and Elias Bareinboim. Causal flow q-learning for robust offline reinforcement learning. *arXiv preprint arXiv:2602.02847*, 2026.

Xiu-Chuan Li and Tongliang Liu. Causal effect identification in the presence of latent confounding with a single imperfect proxy variable. In *International Conference on Learning Representations*, 2026.

Sixu Lin, Guanren Qiao, Yunxin Tai, Ang Li, Kui Jia, and Guiliang Liu. HWC-Loco: A hierarchical whole-body control approach to robust humanoid locomotion. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=3UE3Aatcjy.

Jizhou Liu, Eric J. Tchetgen Tchetgen, and Carlos Varjão. Proximal causal inference for synthetic control with surrogates. *arXiv preprint arXiv:2308.09527*, 2023.

Songming Liu, Lingxuan Wu, Bangguo Li, Hengkai Tan, Huayu Chen, Zhengyi Wang, Ke Xu, Hang Su, and Jun Zhu. RDT-1B: A diffusion foundation model for bimanual manipulation. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=yAzN4tz7oI.

Xu-Hui Liu, Tian-Shuo Liu, Shengyi Jiang, Ruifeng Chen, Zhilong Zhang, Xinwei Chen, and Yang Yu. Energy-guided diffusion sampling for offline-to-online reinforcement learning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 31541–31565. PMLR, 2024.

Xueyi Liu, He Wang, and Li Yi. DexNDM: Closing the reality gap for dexterous in-hand rotation via joint-wise neural dynamics model. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=80vjyj5o7l.

David Lopez-Paz and Maxime Oquab. Revisiting classifier two-sample tests. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=SJkXfE5xx.

Haitong Ma, Tianyi Chen, Kai Wang, Na Li, and Bo Dai. Efficient online reinforcement learning for diffusion policy. In *Proceedings of the 42nd International Conference on Machine Learning*, Proceedings of Machine Learning Research. PMLR, 2025.

Marlos C Machado, Marc G Bellemare, Erik Talvitie, Joel Veness, Matthew Hausknecht, and Michael Bowling. Revisiting the arcade learning environment: Evaluation protocols and open problems for general agents. *Journal of Artificial Intelligence Research*, 61:523–562, 2018.

Liyuan Mao, Haoran Xu, Xianyuan Zhan, Weinan Zhang, and Amy Zhang. Diffusion-DICE: In-sample diffusion guidance for offline reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Wang Miao, Zhi Geng, and Eric J. Tchetgen Tchetgen. Identifying causal effects with proxy variables of an unmeasured confounder. *Biometrika*, 105(4):987–993, 12 2018. doi: 10.1093/biomet/asy038.

Peer Nagy, Sascha Yves Frey, Kang Li, Bidipta Sarkar, Svitlana Vyetrenko, Stefan Zohren, Ani Calinescu, and Jakob Nicolaus Foerster. LOB-Bench: Benchmarking generative AI for finance — an application to limit order book data. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 45437–45460. PMLR, 2025. URL https://proceedings.mlr.press/v267/nagy25a.html.

Ashvin Nair, Abhishek Gupta, Murtaza Dalal, and Sergey Levine. AWAC: Accelerating online reinforcement learning with offline datasets. *arXiv preprint arXiv:2006.09359*, 2020.

Mitsuhiko Nakamoto, Yifeng Zhai, Anikait Singh, Max Sobol Mark, Yi Ma, Chelsea Finn, Aviral Kumar, and Sergey Levine. Cal-QL: Calibrated offline RL pre-training for efficient online fine-tuning. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Andrew Y. Ng, Daishi Harada, and Stuart J. Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In *Proceedings of the Sixteenth International Conference on Machine Learning (ICML)*, pp. 278–287, San Francisco, CA, USA, 1999. Morgan Kaufmann. ISBN 1-55860-612-2.

Michael Psenka, Alejandro Escontrela, Pieter Abbeel, and Yi Ma. Learning a diffusion model policy from rewards via q-score matching. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 41163–41182. PMLR, 2024.

Allen Z. Ren, Justin Lidard, Lars L. Ankile, Anthony Simeonov, Pulkit Agrawal, Anirudha Majumdar, Benjamin Burchfiel, Hongkai Dai, and Max Simchowitz. Diffusion policy policy optimization. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=mEpqHvbD2h.

Donald B. Rubin. Using propensity scores to help design observational studies: Application to the tobacco litigation. *Health Services and Outcomes Research Methodology*, 2(3–4):169–188, 2001. doi: 10.1023/A:1020363010465.

Han Shao, Shuo Xie, and Kunhe Yang. Should decision-makers reveal classifiers in online strategic classification? In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 54254–54274. PMLR, 2025.

Yishan Shen, Yuyang Ye, Hui Xiong, and Yong Chen. SAFER: A calibrated risk-aware multimodal recommendation model for dynamic treatment regimes. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 54611–54631. PMLR, 2025.

Yongjae Shin, Jeonghye Kim, Whiyoung Jung, Sunghoon Hong, Deunsol Yoon, Youngsoo Jang, Geonhyeong Kim, Jongseong Chae, Youngchul Sung, Kanghoon Lee, and Woohyung Lim. Online pre-training for offline-to-online reinforcement learning. In *Proceedings of the 42nd International Conference on Machine Learning*, Proceedings of Machine Learning Research. PMLR, 2025.

Álvaro Labarca Silva, Denis Parra, and Rodrigo Toro Icarte. On the unexpected effectiveness of reinforcement learning for sequential recommendation. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 45432–45450. PMLR, 2024.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations (ICLR)*, 2021a. URL https://openreview.net/forum?id=St1giarCHLP.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations (ICLR)*, 2021b. URL https://openreview.net/forum?id=PxTIG12RRHS.

Yuda Song, Yifei Zhang, Xinyue Zhang, Jianye Li, Zhendong Wang, and Xianyuan Zhan. Uni-O4: Unifying online and offline deep reinforcement learning with multi-step on-policy optimization. In *International Conference on Learning Representations*, 2024.

Bradly C Stadie, Sergey Levine, and Pieter Abbeel. Incentivizing exploration in reinforcement learning with deep predictive models. *arXiv preprint arXiv:1507.00814*, 2015.

Masashi Sugiyama, Matthias Krauledat, and Klaus-Robert Müller. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8(35):985–1005, 2007.

Eric J. Tchetgen Tchetgen, Andrew Ying, Yifan Cui, Xu Shi, and Wang Miao. An introduction to proximal causal learning. *arXiv preprint arXiv:2009.10982*, 2020.

Eric J. Tchetgen Tchetgen, Andrew Ying, Yifan Cui, Xu Shi, and Wang Miao. An introduction to proximal causal inference. *Statistical Science*, 39(3):375–390, 8 2024. doi: 10.1214/23-STS911.

Emanuel Todorov, Tom Erez, and Yuval Tassa. MuJoCo: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 5026–5033. IEEE, 2012. doi: 10.1109/IROS.2012.6386109.

Théo Vincent, Fabian Wahren, Jan Peters, Boris Belousov, and Carlo D'Eramo. Adaptive q-network: On-the-fly target selection for deep reinforcement learning. In *International Conference on Learning Representations*, volume 2025, pp. 88542–88569, 2025.

Lirui Wang, Xinlei Chen, Jialiang Zhao, and Kaiming He. Scaling proprioceptive-visual learning with heterogeneous pre-trained transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024a.

Yinuo Wang, Likun Wang, Yuxuan Jiang, Wenjun Zou, Tong Liu, Xujie Song, Wenxuan Wang, Liming Xiao, Jiang Wu, Jingliang Duan, et al. Diffusion actor-critic with entropy regulator. *Advances in Neural Information Processing Systems*, 37:54183–54204, 2024b.

Zhendong Wang, Jonathan J. Hunt, and Mingyuan Zhou. Diffusion policies as an expressive policy class for offline reinforcement learning. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=AHvFDPi-FA.

Junjie Wen, Yichen Zhu, Minjie Zhu, Zhibin Tang, Jinming Li, Zhongyi Zhou, Xiaoyu Liu, Chaomin Shen, Yaxin Peng, and Feifei Feng. DiffusionVLA: Scaling robot foundation models via unified diffusion and autoregression. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 66558–66574. PMLR, 2025. URL https://proceedings.mlr.press/v267/wen25g.html.

Liyuan Xu, Heishiro Kanagawa, and Arthur Gretton. Deep proxy causal learning and its application to confounded bandit policy evaluation. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 26264–26275. Curran Associates, Inc., 2021.

Long Yang, Zhixiong Huang, Fenghao Lei, Yucun Zhong, Yiming Yang, Cong Fang, Shiting Wen, Binbin Zhou, and Zhouchen Lin. Policy representation via diffusion probability model for reinforcement learning. *arXiv preprint arXiv:2305.13122*, 2023.

Jiawen Yu, Hairuo Liu, Qiaojun Yu, Jieji Ren, Ce Hao, Haitong Ding, Guangyu Huang, Guofan Huang, Yan Song, Panpan Cai, Wenqiang Zhang, and Cewu Lu. ForceVLA: Enhancing VLA models with a force-aware MoE for contact-rich manipulation. In *Advances in Neural Information Processing Systems*, 2025.

Zilai Zeng, Ce Zhang, Shijie Wang, and Chen Sun. Goal-conditioned predictive coding for offline reinforcement learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Ruoqi Zhang, Ziwei Luo, Jens Sjölund, Thomas B. Schön, and Per Mattsson. Entropy-regularized diffusion policy with q-ensembles for offline reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Zihan Zhou, Aviral Kumar, Abhishek Gupta, and Sergey Levine. Efficient online reinforcement learning fine-tuning need not retain offline data. In *International Conference on Learning Representations*, 2025.

A Related Work

Offline-to-online reinforcement learning. Offline-to-online RL uses previously collected experience to accelerate subsequent environment interaction. Recent methods improve this transfer by calibrating conservative value estimates, balancing pessimism with online plasticity, selectively reusing offline data, or extracting priors for exploration (Nakamoto et al., 2023; Lee et al., 2023; Hu et al., 2024; Liu et al., 2024; Song et al., 2024; Zhou et al., 2025; Shin et al., 2025). Cal-QL stabilizes online fine-tuning through calibrated offline pretraining, Bayesian offline-to-online methods formalize the tradeoff between pessimistic safety and optimistic improvement, Uni-O4 unifies offline and online updates, and EDIS transfers offline knowledge through energy-guided diffusion sampling (Nakamoto et al., 2023; Hu et al., 2024; Song et al., 2024; Liu et al., 2024). These works address important statistical obstacles such as distribution shift, extrapolation error, replay imbalance, and value miscalibration. However, they generally assume that the offline action–outcome relationship provides a reliable signal for the online environment. COOT targets the harder setting where offline logs are informative but biased by hidden execution regimes that may influence both logged behavior and realized rewards. Thus, rather than directly reusing offline data as demonstrations or replay, COOT learns proxy-conditioned reliability signals that determine how biased offline experience should guide diffusion pretraining, critic shaping, and online improvement.

Diffusion policies for control. Diffusion policies have become a strong policy class for continuous control because they can represent multimodal and non-Gaussian action distributions. In offline RL, diffusion policies have been used to model behavior distributions, constrain policy improvement, and reduce out-of-distribution value error (Wang et al., 2023; Chen et al., 2024; Zhang et al., 2024; Mao et al., 2024). Diffusion-QL performs Q-guided diffusion policy improvement, DTQL introduces a diffusion trust region, entropy-regularized diffusion policies use Q-ensembles to reduce overestimation, and Diffusion-DICE guides the behavior distribution toward an optimal stationary distribution (Wang et al., 2023; Chen et al., 2024; Zhang et al., 2024; Mao et al., 2024). In online RL, DACER, QVPO, DPPO, DPMD, and SDAC further make diffusion policies trainable through actor-critic, variational, policy-gradient, and reweighted score-matching objectives (Wang et al., 2024b; Ding et al., 2024; Ren et al., 2025; Ma et al., 2025). These methods primarily advance the optimization of the diffusion policy itself. COOT is complementary: it asks not only how to optimize a diffusion policy, but how such a policy should use offline data when the logged action–outcome associations may be biased by hidden regimes. Its main innovation is a proxy-informed reliability transfer layer that modulates weighted diffusion pretraining, potential-based critic shaping, and risk/support-aware candidate selection.

Causality-aware and confounding-robust reinforcement learning. Causal structure has been used to improve exploration, robustness, and transfer in RL. Causality-aware actor-critic methods exploit relationships between action dimensions and rewards to guide exploration (Ji et al., 2024), while causally aligned curriculum learning uses causal abstractions to improve transfer (Li et al., 2024). More closely related work studies RL under unobserved confounding. Confounding-robust deep RL optimizes policies against worst-case environments compatible with biased observations (Li et al., 2025c); automatic reward shaping derives potentials from confounded offline data, with extensions to continuous-control reward shaping (Li et al., 2025b; Juliani et al., 2026); and Causal Flow Q-Learning studies robust offline RL with expressive flow-based policies under confounded demonstrations (Li et al., 2026). These methods show that hidden confounding can invalidate standard offline RL objectives. COOT advances this direction by moving from robust correction alone to proxy-conditioned offline-to-online transfer. Instead of claiming full identification of interventional action values from confounded logs, COOT uses proxy information to estimate transfer reliability and converts this reliability into modular guidance for diffusion pretraining, uncertainty-aware transfer, support-aware deployment, and online candidate ranking.

Proxy variables and proximal causal inference. Proxy-variable and negative-control methods provide tools for reasoning about latent confounding when the confounder is not observed but auxiliary measurements contain information about it. Recent work develops automated negative-control estimation under unmeasured confounding (Kummerfeld et al., 2024) and studies identification with imperfect proxies (Li & Liu, 2026). This perspective is well matched to modern decision logs, which often contain telemetry beyond the primary state, action, and reward, such as actuator diagnostics, demand indicators, market signals, user-intent measurements, or execution metadata. COOT is motivated by this proximal-causal perspective, but does not require or claim exact proximal identification of interventional action values. Importantly, it does not simply concatenate proxies to the policy input. Instead, proxies are used to infer the reliability of offline transfer, producing potential, uncertainty, support, and risk signals that are consumed by the online diffusion-policy learner while preserving the standard state-action deployment interface.

Positioning of COOT. COOT connects three previously separate threads. Offline-to-online RL shows how offline data can accelerate online learning, but usually assumes sufficiently reliable logged experience (Nakamoto et al., 2023; Hu et al., 2024; Liu et al., 2024; Song et al., 2024). Diffusion-policy RL provides expressive action generators and strong online optimization objectives, but generally studies clean offline data or standard online interaction rather than biased transfer under hidden execution regimes (Wang et al., 2024b; Ding et al., 2024; Ma et al., 2025; Ren et al., 2025). Confounding-robust RL identifies the causal failure mode of biased logs, but has not provided a practical proxy-conditioned reliability mechanism for scalable online diffusion-policy optimization (Li et al., 2025b;c; Juliani et al., 2026; Li et al., 2026). COOT fills this gap by combining causally motivated proxy reasoning with offline-to-online diffusion-policy RL. The resulting proxy-informed reliability transfer bundle transforms biased but informative offline experience into structured guidance for diffusion pretraining, critic shaping, and online candidate selection, providing a practical path for reusing imperfect real-world logs without assuming that they are causally clean or that exact interventional values are identifiable.

B Additional Method Details

This appendix gives the implementation-level realization of the framework in Section 3. The main text states the reliability principle behind COOT. Here we make precise the estimands each component targets, the estimators that approximate them, and the normalizations and online interfaces that turn them into a single coherent transfer mechanism. Two conventions fix the scope of every quantity below. First, all bundle estimators target *proxy-conditioned* population objects, expectations that condition on the observed tuple (o, a, z, w). None is asserted to equal an interventional value, and the dispersion of the ensemble is used precisely to quantify how far that observational target may sit from a transferable one. Second, the bundle is fit once, offline, and frozen, so the online learner alone is responsible for validating and correcting the induced policy through environment interaction. Throughout, a_i is the logged action, G_i its empirical discounted return, and $\mathcal{B}_{\text{replay}}$ the online replay buffer.

B.1 Proxy-Informed Reliability Transfer Bundle: Estimators and Normalization

Estimand. Each proxy view $\nu \in \{z, w\}$ carries partial information about the latent regime u of Eq. 1. The associated population target is the view-conditioned return surface

$$h^\star_\nu(o, a, \nu) = \mathbb{E}[\, G \mid o, a, \nu \,], \qquad Q^\star_{\text{prox}}(o, a, z, w) = \mathbb{E}[\, G \mid o, a, z, w \,], \tag{13}$$

the best regime-aware predictors measurable from the logged covariates. Conditioning on (z, w) captures only the component of the regime that the proxies happen to reveal in prediction, and does not by itself close the backdoor paths of Eq. 1. The residual, unrevealed component is what separates Eq. 13 from any interventional quantity, and the ensemble dispersion introduced below is the statistic COOT uses to react to it. The bundle therefore estimates a reliability-weighted version of an *observational* target, and never presumes identification of $\mathbb{E}[G \mid o, \mathrm{do}(a)]$.

Proxy-conditioned bridge ensembles. For each view we form the augmented feature

$$x_i^\nu = \left[\, o_i^\top \;\; a_i^\top \;\; \nu_i^\top \,\right]^\top, \qquad \nu \in \{z, w\}, \tag{14}$$

and fit a standardized ridge ensemble over B bootstrap resamples. On resample $\mathcal{I}_b$, the b-th view predictor solves

$$\hat{h}_\nu^{(b)} \in \arg\min_{h\in\mathcal{H}} \sum_{i\in\mathcal{I}_b} \big(G_i - h(x_i^\nu)\big)^2 + \alpha_{\mathrm{br}}\|h\|_2^2, \qquad b = 1,\dots,B. \tag{15}$$

The two views are fit *separately*, never on a concatenated $[z, w]$ feature, so that disagreement between them survives as a measurable signal rather than being averaged into a single regressor. Stacking the full pool of $2B$ predictions at a query (o, a, z, w),

$$H(o,a,z,w) = \big(\hat{h}_z^{(1)}(o,a,z),\dots,\hat{h}_z^{(B)}(o,a,z),\ \hat{h}_w^{(1)}(o,a,w),\dots,\hat{h}_w^{(B)}(o,a,w)\big)^\top \in \mathbb{R}^{2B}, \tag{16}$$

its first two moments give the ensemble summary

$$\hat{\mu}_{\mathrm{br}} = \tfrac{1}{2B}\,\mathbf{1}^\top H, \qquad \hat{\sigma}_{\mathrm{br}} = \Big(\tfrac{1}{2B}\,\big\|H - \hat{\mu}_{\mathrm{br}}\mathbf{1}\big\|_2^2\Big)^{1/2}, \tag{17}$$

and the *robust* proxy-conditioned transfer value is the lower-confidence summary

$$\hat{Q}_{\mathrm{rob}}(o,a,z,w) = \hat{\mu}_{\mathrm{br}}(o,a,z,w) - \beta_{\mathrm{br}}\,\hat{\sigma}_{\mathrm{br}}(o,a,z,w), \qquad \beta_{\mathrm{br}} \geq 0. \tag{18}$$

A single scalar $\hat{\sigma}_{\mathrm{br}}$ thus fuses two distinct uncertainties, bootstrap (epistemic) variability and z-versus-w disagreement, so the discount in Eq. 18 responds to both with no second hyperparameter. A logged action earns a high transfer value only when its proxy-conditioned return is simultaneously large and stable across views and resamples. Stability across views is a reliability cue rather than a certificate of unbiasedness, since two views sharing a common bias can agree while both err, a failure mode probed directly by the proxy stress tests of Appendix D.4. Appendix C.1 shows this summary is conservative with respect to the target in Eq. 13.

State potential and robust advantage. A large $\hat{Q}_{\mathrm{rob}}$ may reflect a favorable *state* rather than a reliable *action*. COOT therefore fits a state-only potential

$$\hat{\Phi} \in \arg\min_{\Phi\in\mathcal{F}_\Phi} \sum_{i=1}^{N} \big(G_i - \Phi(o_i)\big)^2 + \alpha_\Phi\|\Phi\|_2^2, \qquad \hat{\Phi}(o) \approx \mathbb{E}[\,G \mid o\,], \tag{19}$$

and subtracts it to obtain the state-normalized robust advantage

$$\hat{A}_{\mathrm{rob}}(o,a,z,w) = \hat{Q}_{\mathrm{rob}}(o,a,z,w) - \hat{\Phi}(o). \tag{20}$$

This isolates action-level transfer desirability from state-level return structure, and the same $\hat{\Phi}$ is reused as the dense potential-shaping signal for the critic.

Proxy-free risk field. The dispersion $\hat{\sigma}_{\mathrm{br}}$ is defined only where proxies are present, yet online candidates are scored through the proxy-free state-action interface. COOT bridges this gap by distilling the proxy-conditioned dispersion onto proxy-free inputs,

$$\tilde{R} \in \arg\min_{R\in\mathcal{F}_R} \sum_{i=1}^{N} \big(\hat{\sigma}_{\mathrm{br}}(o_i,a_i,z_i,w_i) - R(o_i,a_i)\big)^2 + \alpha_R\|R\|_2^2, \qquad \hat{R}(o,a) = \big[\tilde{R}(o,a)\big]_+, \tag{21}$$

with $[x]_+ = \max\{x, 0\}$. The field $\hat{R}$ renders proxy-derived uncertainty queryable online, exactly where the proxies themselves are unavailable.

Behavior-support model. COOT also scores whether a state-action pair lies within the reliable logged manifold. Against the positive set $\{(o_i, a_i)\}$ it contrasts three complementary negative families, action permutation, uniform resampling, and local Gaussian perturbation,

$$\mathcal{N} = \Big\{\,(o_i, a_{\pi(i)}),\ (o_i, a_i^{\mathrm{unif}}),\ \big(o_i, \mathrm{clip}(a_i + \epsilon_i, -1, 1)\big)\,\Big\}_{i=1}^{N}, \quad a_i^{\mathrm{unif}} \sim \mathcal{U}[-1,1]^{d_a},\ \ \epsilon_i \sim \mathcal{N}(0, \sigma_{\mathrm{neg}}^2 I), \tag{22}$$

which jointly pin the decision boundary from three directions, global mismatch, coverage gaps, and local off-support drift. A class-balanced logistic model yields $\hat{p}_{\text{sup}}(o, a)$, and the support penalty is its clipped negative log-probability,

$$\hat{S}(o, a) = -\log \text{clip}\big(\hat{p}_{\text{sup}}(o, a),\, \varepsilon_{\text{sup}},\, 1 - \varepsilon_{\text{sup}}\big). \tag{23}$$

Large $\hat{S}$ marks weak behavioral support, hence lower transfer reliability. Being discriminative rather than density-based, both $\hat{R}$ and $\hat{S}$ evaluate in closed form on arbitrary online candidates.

Standardized reliability composition. The three signals carry incomparable units, a return-scale advantage, a return-scale dispersion, and a log-probability penalty, so each is standardized over $\mathcal{D}_{\text{off}}$ with a variance floor before they are combined. Writing $T \in \{A, R, S\}$ for the streams $\hat{A}_{\text{rob}}, \hat{R}, \hat{S}$,

$$\mu_T = \tfrac{1}{N}\sum_{i=1}^{N} T_i, \qquad \sigma_T = \max\Big\{\big(\tfrac{1}{N}\sum_{i=1}^{N}(T_i - \mu_T)^2\big)^{1/2},\, \varepsilon_{\text{std}}\Big\}. \tag{24}$$

The standardized reliability score linearly trades benefit against instability and implausibility on this common scale,

$$s_{\text{off}}(o, a, z, w) = \frac{1}{\tau_{\text{off}}}\, \underbrace{\frac{\hat{A}_{\text{rob}} - \mu_A}{\sigma_A}}_{\text{benefit}} - \lambda_R \underbrace{\frac{\hat{R} - \mu_R}{\sigma_R}}_{\text{instability}} - \lambda_S \underbrace{\frac{\hat{S} - \mu_S}{\sigma_S}}_{\text{implausibility}}, \tag{25}$$

and the offline transfer weight is its symmetrically clipped exponential tilt,

$$\omega_{\text{off}}(o, a, z, w) = \exp\Big[\text{clip}\big(s_{\text{off}}(o, a, z, w),\, -c_{\text{off}},\, c_{\text{off}}\big)\Big]. \tag{26}$$

The map is monotone in the intended directions, increasing in robust advantage and decreasing in instability and implausibility, and the log-space clip bounds the leverage of any single sample. The variance floor ε_{std} keeps the standardization well-posed even when a stream is nearly constant on a given dataset. Appendix C.2 shows that Eq. 26 is not posited but *derived*, as the unique Gibbs tilt of a KL-anchored reliability program.

B.2 Weighted Denoising: A Shared Interface for Offline and Online Policy Updates

COOT uses one weighted denoising interface for both offline pretraining and online actor improvement, and only the guide-weight pair (a^{g}, ω) changes between stages. For guide action a^{g}, diffusion time t, noised action a_t, and injected noise ϵ,

$$\ell_{\text{diff}}(\theta; o, a^{\text{g}}) = \mathbb{E}_{t,\epsilon}\big\|\epsilon - \epsilon_\theta(o, a_t, t)\big\|_2^2. \tag{27}$$

Given a batch B with nonnegative guide weights $\{\omega_i\}_{i\in B}$, COOT batch-normalizes them to fix the gradient scale,

$$\bar{\omega}_i = \frac{\omega_i}{|B|^{-1}\sum_{j\in B}\omega_j + \varepsilon_\omega}, \tag{28}$$

and minimizes the per-coordinate weighted objective

$$\mathcal{L}_{\text{diff}}(\theta) = \frac{1}{|B|\, d_a}\sum_{i\in B}\bar{\omega}_i\, \big\|\epsilon_i - \epsilon_\theta(o_i, a_{t_i,i}, t_i)\big\|_2^2, \tag{29}$$

where d_a is the action dimension. The two stages differ only in their target distribution,

$$\big(a_i^{\text{g}},\, \omega_i\big) = \begin{cases} \big(a_i,\ \omega_{\text{off}}(o_i, a_i, z_i, w_i)\big), & \text{offline (selective imitation)}, \\ \Big(\arg\max\limits_{a\in\mathcal{C}(o_i)} \mathcal{J}_{\text{on}}(o_i, a),\ \ \omega_{\text{on}}(o_i)\Big), & \text{online (value-guided improvement)}. \end{cases} \tag{30}$$

Sharing one interface guarantees that the offline and online policy objectives are mechanically identical and that selective imitation and value-guided improvement are expressed under a single expressive policy class rather than two bespoke losses.

B.3 Potential-Shaped Critic Targets

With reward scale $\rho > 0$ and shaping coefficient $\eta_t \geq 0$, COOT forms the potential-shaped reward and clipped double-Q target

$$\tilde{r}_t = \rho\, r_t + \eta_t\big(\gamma\hat{\Phi}(o_{t+1}) - \hat{\Phi}(o_t)\big), \qquad y_t = \tilde{r}_t + \gamma(1-d_t)\min_{j\in\{1,2\}} Q_{\bar{\phi}_j}(o_{t+1}, a_{t+1}), \quad a_{t+1} \sim \pi_{\bar{\theta}}(\cdot \mid o_{t+1}), \tag{31}$$

and fits both critics by regression onto the shared target,

$$\mathcal{L}_Q(\phi_1, \phi_2) = \sum_{j=1}^{2} \mathbb{E}\big[(Q_{\phi_j}(o_t, a_t) - y_t)^2\big]. \tag{32}$$

Online, the shaping coefficient is annealed linearly to zero,

$$\eta_t = \eta_0\big[\,1 - t/T_{\text{shape}}\,\big]_+, \qquad [x]_+ = \max\{x, 0\}. \tag{33}$$

The schedule encodes a deliberate handover: while on-policy transitions are scarce, the offline potential supplies dense state-level value structure, and as $\mathcal{B}_{\text{replay}}$ fills, the shaping term vanishes and the critic is governed entirely by environment reward. Because potential shaping is policy-order preserving for any fixed bounded $\hat{\Phi}$ (Appendix C.5), the anneal alters only the optimization path, never the fixed point it converges to.

B.4 Prior-Guided Online Candidate Selection

After pretraining, COOT freezes the target diffusion policy as a persistent prior π_{prior}. At online observation o, the candidate set unions samples from the current actor and the prior,

$$\mathcal{C}(o) = \underbrace{\{a_k^\theta \sim \pi_\theta(\cdot \mid o)\}_{k=1}^{K}}_{\mathcal{C}_\theta(o)} \cup \underbrace{\{a_m^{\text{prior}} \sim \pi_{\text{prior}}(\cdot \mid o)\}_{m=1}^{M}}_{\mathcal{C}_{\text{prior}}(o)}, \tag{34}$$

and each candidate is scored by online value regularized by the proxy-free reliability fields. To make the penalty prices scale-invariant, the critic term is range-normalized across the candidate set before the penalties are applied,

$$\hat{Q}^{\,\text{nrm}}(o,a) = \frac{\min_j Q_{\phi_j}(o,a) - \min_{a'\in\mathcal{C}(o)} \min_j Q_{\phi_j}(o,a')}{\text{range}_{a'\in\mathcal{C}(o)} \min_j Q_{\phi_j}(o,a') + \varepsilon_Q}, \qquad \mathcal{J}_{\text{on}}(o,a) = \hat{Q}^{\,\text{nrm}}(o,a) - \lambda_R^{\text{on}}\hat{R}(o,a) - \lambda_S^{\text{on}}\hat{S}(o,a), \tag{35}$$

so that $(\lambda_R^{\text{on}}, \lambda_S^{\text{on}})$ keep a fixed meaning regardless of the absolute magnitude of Q. Selection then occurs at two points that must not be conflated. The agent *executes*

$$a_t \in \arg\max_{a\in\mathcal{C}(o_t)} \mathcal{J}_{\text{on}}(o_t, a), \tag{36}$$

storing the realized transition. For actor improvement it *independently* recomputes guides on replay observations and converts the selected score into the online guide weight,

$$a_i^{\text{g}} \in \arg\max_{a\in\mathcal{C}(o_i)} \mathcal{J}_{\text{on}}(o_i, a), \qquad \omega_{\text{on}}(o_i) = \exp\Big[\,\text{clip}\Big(\tfrac{1}{\tau_{\text{act}}}\,\mathcal{J}_{\text{on}}(o_i, a_i^{\text{g}}),\, -c_{\text{on}},\, c_{\text{on}}\Big)\Big]. \tag{37}$$

The asymmetry is the design's safeguard: the Bellman update is anchored to the *executed* action the environment actually evaluated, so value learning never drifts from on-policy reality, while the actor is free to be distilled toward the best *reliable* candidate, which may originate from the prior. COOT thereby exploits offline structure for proposal diversity while structurally preventing unverified offline actions from corrupting the value estimates that drive learning.

C Proofs and Derivations

This appendix formalizes the four properties that make COOT a principled transfer mechanism from biased but informative logs. We are explicit about what is proved and what is not. We do not claim identification of interventional action values from proxies, which is not possible without further assumptions. Instead, each result certifies a guarantee that is both attainable from observational data and sufficient for reliable offline-to-online transfer. The four guarantees compose into a single pipeline. *Conservatism*: the robust bridge value lower-bounds a proxy-conditioned return target whenever the ensemble is locally calibrated (Sec. C.1). *Selectivity*: the offline weight is the unique exponential tilt of a KL-anchored reliability program and is monotone in the three reliability criteria (Sec. C.2, C.3). *Stability*: batch normalization caps the gradient influence of any single sample (Sec. C.4). *Invariance*: potential shaping preserves the optimal-policy set for any fixed bounded potential (Sec. C.5). Conservatism makes the transferred value trustworthy, selectivity turns it into a principled imitation emphasis, stability keeps the optimization well-behaved, and invariance ensures the dense critic signal accelerates rather than redirects online learning.

C.1 The Robust Bridge Value as a Conservative Proxy-Conditioned Surrogate

The first question any transfer signal must answer is what it is conservative *with respect to*. We fix this target as the proxy-conditioned return surface

$$Q^{\star}_{\text{prox}}(o, a, z, w) := \mathbb{E}[\, G \mid o, a, z, w \,], \tag{38}$$

the best return predictor measurable from the full logged tuple. Eq. 38 is an observational object: under the confounding model of Eq. 1 it absorbs the regime information carried by (z, w) but, in general, still differs from the interventional value by the residual that the proxies do not reveal. COOT does not attempt to close that residual. It treats $Q^{\star}_{\text{prox}}$ as the strongest target the proxies can support and converts the ensemble's two-sided estimation error into a one-sided, transferable guarantee with respect to it. The pessimistic principle is deliberately asymmetric: a logged action may be trusted *below* what the proxies certify, never above.

For a query (o, a, z, w), let

$$H(o, a, z, w) = \big(\hat{h}^{(1)}_z, \ldots, \hat{h}^{(B)}_z,\ \hat{h}^{(1)}_w, \ldots, \hat{h}^{(B)}_w\big)^{\top} \in \mathbb{R}^{2B}$$

be the pool of bridge predictions at the corresponding proxy view, with first two moments

$$\hat{\mu}_{\text{br}} = \tfrac{1}{2B}\mathbf{1}^{\top} H, \qquad \hat{\sigma}_{\text{br}} = \Big(\tfrac{1}{2B}\big\|H - \hat{\mu}_{\text{br}}\mathbf{1}\big\|_2^2\Big)^{1/2}. \tag{39}$$

COOT transfers the lower-confidence summary

$$\hat{Q}_{\text{rob}}(o, a, z, w) = \hat{\mu}_{\text{br}}(o, a, z, w) - \beta_{\text{br}}\, \hat{\sigma}_{\text{br}}(o, a, z, w), \qquad \beta_{\text{br}} \geq 0, \tag{40}$$

in which the single dispersion $\hat{\sigma}_{\text{br}}$ pools two qualitatively distinct error sources, variation across bootstrap fits and disagreement between the two proxy views. The surrogate is therefore high only when the proxy-conditioned predictions are simultaneously large and stable.

Definition 1 (Pointwise bridge calibration)**.** *The bridge ensemble is β_{br}-calibrated at (o, a, z, w) if its dispersion brackets the residual to the proxy-conditioned target,*

$$\big|\, Q^{\star}_{\text{prox}}(o, a, z, w) - \hat{\mu}_{\text{br}}(o, a, z, w) \,\big| \leq \beta_{\text{br}}\, \hat{\sigma}_{\text{br}}(o, a, z, w). \tag{41}$$

The condition asks only that $\hat{\sigma}_{\text{br}}$ be an honest *local* uncertainty radius for $Q^{\star}_{\text{prox}}$ at the query of interest. It is deliberately pointwise: COOT uses the surrogate to rank and weight heterogeneous logged actions, so the relevant guarantee is one that holds where the ensemble is calibrated and degrades gracefully where it is not, rather than a global parametric assumption. Because $\hat{\sigma}_{\text{br}}$ already fuses epistemic and cross-view uncertainty, Eq. 41 is exactly the radius that must dominate the residual for one-sided conservatism to hold.

Proposition 3 (Lower-confidence proxy-conditioned surrogate)**.** *If the bridge ensemble is β_{br}-calibrated at (o, a, z, w), then*

$$\hat{Q}_{\mathrm{rob}}(o, a, z, w) \leq Q^{\star}_{\mathrm{prox}}(o, a, z, w), \tag{42}$$

and the conservatism gap is controlled by the same dispersion,

$$0 \leq Q^{\star}_{\mathrm{prox}}(o, a, z, w) - \hat{Q}_{\mathrm{rob}}(o, a, z, w) \leq 2\beta_{\mathrm{br}}\, \hat{\sigma}_{\mathrm{br}}(o, a, z, w). \tag{43}$$

Proof. Calibration gives $Q^{\star}_{\mathrm{prox}} - \hat{\mu}_{\mathrm{br}} \geq -\beta_{\mathrm{br}}\hat{\sigma}_{\mathrm{br}}$, hence

$$Q^{\star}_{\mathrm{prox}} \geq \hat{\mu}_{\mathrm{br}} - \beta_{\mathrm{br}}\hat{\sigma}_{\mathrm{br}} = \hat{Q}_{\mathrm{rob}},$$

which is Eq. equation 42 and the lower half of Eq. equation 43. For the upper half, the other side of calibration gives $Q^{\star}_{\mathrm{prox}} - \hat{\mu}_{\mathrm{br}} \leq \beta_{\mathrm{br}}\hat{\sigma}_{\mathrm{br}}$. Adding the identity $\hat{\mu}_{\mathrm{br}} - \hat{Q}_{\mathrm{rob}} = \beta_{\mathrm{br}}\hat{\sigma}_{\mathrm{br}}$ yields

$$Q^{\star}_{\mathrm{prox}} - \hat{Q}_{\mathrm{rob}} = \big(Q^{\star}_{\mathrm{prox}} - \hat{\mu}_{\mathrm{br}}\big) + \big(\hat{\mu}_{\mathrm{br}} - \hat{Q}_{\mathrm{rob}}\big) \leq 2\beta_{\mathrm{br}}\hat{\sigma}_{\mathrm{br}}. \qquad \square$$

Two consequences explain design choices in the main text rather than merely restating them. First, the gap bound $2\beta_{\mathrm{br}}\hat{\sigma}_{\mathrm{br}}$ makes conservatism *self-limiting*: where the proxies pin down the regime, dispersion is small and $\hat{Q}_{\mathrm{rob}}$ is nearly tight, so well-identified actions are barely penalized, whereas where the views disagree or bootstraps scatter, the surrogate retreats in exact proportion to the measured uncertainty. The single coefficient β_{br} thus interpolates between optimistic mean transfer ($\beta_{\mathrm{br}} = 0$) and strict worst-case transfer, with no separate mechanism per uncertainty source. Second, the guarantee is the right *shape* for the task, local and dispersion-controlled, which is precisely the reliability behavior selective transfer requires.

C.2 The Offline Weight as an Exponential Reliability Tilt

The conservative value of Sec. C.1 answers *benefit*; reliable transfer additionally demands *stability* and *support*. Collect the three criteria in their standardized, dimensionless forms,

$$\bar{A}_{\mathrm{rob}} = \frac{\hat{A}_{\mathrm{rob}} - \mu_A}{\sigma_A}, \qquad \bar{R} = \frac{\hat{R} - \mu_R}{\sigma_R}, \qquad \bar{S} = \frac{\hat{S} - \mu_S}{\sigma_S}, \qquad \hat{A}_{\mathrm{rob}} = \hat{Q}_{\mathrm{rob}} - \hat{\Phi}, \tag{44}$$

and define the per-sample reliability utility

$$u(o, a, z, w) = \tau_{\mathrm{off}}^{-1}\bar{A}_{\mathrm{rob}}(o, a, z, w) - \lambda_R\bar{R}(o, a) - \lambda_S\bar{S}(o, a). \tag{45}$$

We now derive, rather than posit, the offline weight. Let p_0 be the empirical offline distribution placing uniform mass on logged samples, the imitation target of ordinary cloning. COOT seeks a reweighted distribution q that maximizes expected reliability while remaining anchored to p_0, the anchor encoded as a Kullback-Leibler penalty:

$$q^{\star} \in \operatorname*{arg\,max}_{q \ll p_0} \Big\{ \mathbb{E}_q[u(o, a, z, w)] - \mathrm{KL}(q \,\|\, p_0) \Big\}. \tag{46}$$

The program is strictly concave in q over the simplex, and its unique maximizer is the Gibbs tilt

$$q^{\star}(o, a, z, w) = \frac{p_0(o, a, z, w)\, \exp\big(u(o, a, z, w)\big)}{\mathbb{E}_{p_0}\big[\exp(u(o, a, z, w))\big]}. \tag{47}$$

Hence the per-sample emphasis relative to uniform cloning is exactly $\exp(u)$. Identifying the normalized score

$$s_{\mathrm{off}}(o, a, z, w) = \tau_{\mathrm{off}}^{-1}\bar{A}_{\mathrm{rob}}(o, a, z, w) - \lambda_R\bar{R}(o, a) - \lambda_S\bar{S}(o, a), \tag{48}$$

and bounding per-sample leverage by a symmetric log-space clip, yields the offline transfer weight

$$\omega_{\mathrm{off}}(o, a, z, w) = \exp\Big[\,\mathrm{clip}\big(s_{\mathrm{off}}(o, a, z, w),\, -c_{\mathrm{off}},\, c_{\mathrm{off}}\big)\Big]. \tag{49}$$

The derivation gives every element an operational meaning. The exponential is the unique form turning an additive reliability budget into a multiplicative likelihood reweighting, τ_{off} sets how aggressively benefit is exploited against the uniform anchor, the pair (λ_R, λ_S) are the Lagrange prices of the stability and support constraints, and c_{off} caps the influence a single sample can purchase. The offline weight is thus the optimal solution of a constrained transfer program, not an ad hoc scoring rule.

C.3 Monotone Selectivity: Proof of Proposition 1

Proposition 1 states the operational content of the tilt: it preserves the reliability ordering of samples and uniformly bounds their leverage. We prove both and record the monotonicity that makes "reliability ordering" precise.

Proof. Write $s_i = s_{\text{off}}(o_i, a_i, z_i, w_i)$. If $s_i, s_j \in [-c_{\text{off}}, c_{\text{off}}]$ the clip in Eq. equation 49 is inactive, so $\omega_{\text{off},i} = e^{s_i}$, $\omega_{\text{off},j} = e^{s_j}$, and the relative emphasis depends only on the score *gap*,

$$\frac{\omega_{\text{off},i}}{\omega_{\text{off},j}} = \exp(s_i - s_j), \tag{50}$$

which is what makes the subsequent batch normalization (Sec. C.4) ranking-preserving. For arbitrary samples the clipped argument lies in $[-c_{\text{off}}, c_{\text{off}}]$, so monotonicity of exp gives the uniform envelope

$$e^{-c_{\text{off}}} \le \omega_{\text{off}}(o, a, z, w) \le e^{c_{\text{off}}}. \tag{51}$$

Finally, on the unclipped region the score has constant-sign partial derivatives

$$\frac{\partial s_{\text{off}}}{\partial \bar{A}_{\text{rob}}} = \tau_{\text{off}}^{-1} > 0, \qquad \frac{\partial s_{\text{off}}}{\partial \bar{R}} = -\lambda_R \le 0, \qquad \frac{\partial s_{\text{off}}}{\partial \bar{S}} = -\lambda_S \le 0, \tag{52}$$

so $\omega_{\text{off}} = e^{s_{\text{off}}}$ is strictly increasing in robust proxy-conditioned advantage and non-increasing in instability and implausibility. This establishes the ratio identity, the uniform bound, and the claimed monotone selectivity. □

The proposition separates COOT from uniform cloning at the level of the objective, not the architecture: cloning is the degenerate tilt $\omega \equiv 1$, recovered as $\tau_{\text{off}}^{-1} \to 0$ and $\lambda_R, \lambda_S \to 0$, and any positive setting of these prices strictly reorders imitation emphasis toward the reliable region of the dataset while keeping every sample's leverage explicitly bounded.

C.4 Bounded Gradient Influence: Proof of Proposition 2

Selectivity would be self-defeating if a few high-reliability samples could dominate the actor update and collapse the learned modes. The batch normalization

$$\bar{\omega}_{\text{off},i} = \frac{\omega_{\text{off},i}}{|B|^{-1} \sum_{j \in B} \omega_{\text{off},j} + \varepsilon_\omega} \tag{53}$$

preserves within-batch ordering while making the loss scale-free, and the next result shows it also caps every sample's contribution.

Proof. By Eq. equation 51, every weight obeys

$$e^{-c_{\text{off}}} \le \omega_{\text{off},i} \le e^{c_{\text{off}}}, \qquad i \in B, \tag{54}$$

so the batch mean satisfies $|B|^{-1} \sum_{j \in B} \omega_{\text{off},j} \ge e^{-c_{\text{off}}}$. Since $\varepsilon_\omega \ge 0$, the denominator of Eq. equation 53 is at least $e^{-c_{\text{off}}}$, and combined with the numerator bound $\omega_{\text{off},i} \le e^{c_{\text{off}}}$ this gives the composition-free envelope

$$\bar{\omega}_{\text{off},i} \le \frac{e^{c_{\text{off}}}}{e^{-c_{\text{off}}}} = e^{2c_{\text{off}}}. \tag{55}$$

If the denoising loss satisfies $0 \le \ell_{\text{diff}}(\theta; o_i, a_i) \le L_{\max}$ on the batch, then because $\bar{\omega}_{\text{off},i} \ge 0$,

$$0 \le \bar{\omega}_{\text{off},i}\, \ell_{\text{diff}}(\theta; o_i, a_i) \le e^{2c_{\text{off}}} L_{\max}. \tag{56}$$

□

The bound is informative precisely because it is independent of batch composition and of how extreme the raw scores are: no matter how the reliability mass is distributed, no single sample contributes more than $e^{2c_{\text{off}}} L_{\max}$ to the update. The clip c_{off} thus plays a dual role made explicit by Propositions 1 and 2: it caps both relative emphasis and absolute gradient leverage through the same constant, so selectivity and optimization stability are governed by one interpretable hyperparameter rather than traded off against each other.

C.5 Policy-Order Invariance of Potential Shaping

COOT injects the learned potential $\hat{\Phi}$ to densify critic targets early, but a transfer method must not let this densification silently redefine the task. The following identity shows that, for any fixed bounded potential, shaping changes discounted returns only by a state-dependent constant, leaving the optimal-policy set invariant.

Proposition 4 (Return reparameterization under potential shaping)**.** *Let $\Phi : \mathcal{O} \to \mathbb{R}$ be fixed and bounded and $\gamma \in [0, 1)$, and define the shaped reward*

$$\tilde{r}(o_t, a_t, o_{t+1}) = r(o_t, a_t, o_{t+1}) + \gamma\Phi(o_{t+1}) - \Phi(o_t). \tag{57}$$

Then for every trajectory $\tau = (o_0, a_0, o_1, \ldots)$,

$$\tilde{G}_0(\tau) = G_0(\tau) - \Phi(o_0), \tag{58}$$

so potential shaping preserves the policy ordering at each fixed initial observation.

Proof. Expanding the shaped return and separating the potential terms,

$$\begin{aligned}\tilde{G}_0(\tau) &= \sum_{t=0}^{\infty} \gamma^t \big[r(o_t, a_t, o_{t+1}) + \gamma\Phi(o_{t+1}) - \Phi(o_t)\big] \\ &= G_0(\tau) + \sum_{t=0}^{\infty} \gamma^{t+1}\Phi(o_{t+1}) - \sum_{t=0}^{\infty} \gamma^t \Phi(o_t).\end{aligned} \tag{59}$$

Reindexing the first potential sum by $t' = t + 1$ gives $\sum_{t'=1}^{\infty} \gamma^{t'}\Phi(o_{t'})$. Since Φ is bounded and $\gamma < 1$ both series converge absolutely, so the rearrangement is valid and all terms cancel except the initial one,

$$\tilde{G}_0(\tau) = G_0(\tau) - \Phi(o_0). \tag{60}$$

For a fixed o_0 the offset $-\Phi(o_0)$ is independent of the action sequence, hence of the policy, and cannot change the argmax over policies. The optimal-policy set is invariant. □

This invariance is exactly what licenses the annealing schedule of Appendix B.3. Because the fixed point is independent of the shaping magnitude, COOT can inject a strong dense signal from $\hat{\Phi}$ when on-policy returns are scarce and withdraw it as the replay buffer fills with genuine environment reward, so the schedule changes how fast the critic reaches its target, never which target it reaches. Read together, the four results close the loop: bridge conservatism makes the transferred value trustworthy, the exponential tilt turns it into a principled imitation emphasis, bounded influence keeps the optimization stable, and shaping invariance ensures the offline signal accelerates rather than redirects online learning.

D Experiment Details

This appendix provides additional experimental details that are not repeated in the main paper. We focus on the default COOT setting used for the reported results: the same-data offline-transfer control, the hyperparameter configuration, the construction of biased offline datasets, and the covariate-shift diagnostics used to assess whether the offline data exhibit the distributional structure targeted by COOT.

D.1 Same-data Uniform Behavior Cloning Warm-start Control

The Uniform Behavior Cloning warm-start control is designed as a controlled intervention on the offline-transfer channel. Its purpose is to hold fixed the offline data, pretraining compute, policy class, and online budget, while replacing COOT's selective reliability transfer with the identity transfer rule. This makes the control a direct test of whether COOT's gains come from generic offline pretraining or from the proposed proxy-informed reliability mechanism.

Let $\mathcal{D}_{\text{off}} = \{(o_i, a_i, r_i, o'_i, d_i, z_i, w_i)\}_{i=1}^{N}$ denote the same confounded dataset used by COOT, with $N = 300\text{k}$ transitions. COOT first maps $\mathcal{D}_{\text{off}}$ to a reliability bundle

$$\mathcal{T}_{\text{COOT}}(\mathcal{D}_{\text{off}}) = \big(\omega_{\text{off}}, \hat{\Phi}, \hat{R}, \hat{S}, \pi_{\text{prior}}\big),$$

where ω_{off} selects reliable imitation targets, $\hat{\Phi}$ shapes critic initialization, $\hat{R}$ and $\hat{S}$ regularize online candidate selection, and π_{prior} preserves the pretrained offline actor as a proposal source. The Uniform Behavior Cloning control applies the identity transfer map

$$\mathcal{T}_{\text{UBC}}(\mathcal{D}_{\text{off}}) = \big(\mathbf{1}, 0, 0, 0, \big).$$

Thus every logged action receives the same offline weight, the critic receives no potential-shaped reward, online candidates receive no proxy-derived risk or support penalty, and the frozen offline policy is not reused as a persistent proposal distribution.

The offline actor pretraining objective is the standard diffusion behavior-cloning loss with unit weights:

$$\theta_{\text{UBC}} \in \arg\min_{\theta} \mathbb{E}_{(o,a)\sim\mathcal{D}_{\text{off}}} \mathbb{E}_{t,\epsilon} \left[\|\epsilon - \epsilon_\theta(o, a_t, t)\|_2^2 \right], \qquad a_t = \sqrt{\bar{\alpha}_t} a + \sqrt{1 - \bar{\alpha}_t}\epsilon.$$

Online learning is then initialized from θ_{UBC} and proceeds with the same online budget and the same diffusion actor architecture as COOT, but candidate selection uses only the current online actor and online critic:

$$\mathcal{C}_{\text{UBC}}(o) = \{a_k \sim \pi_\theta(\cdot \mid o)\}_{k=1}^{K}, \qquad J_{\text{UBC}}(o, a) = \min_{j\in\{1,2\}} Q_{\phi_j}(o, a).$$

In contrast, COOT evaluates candidates using $J_{\text{on}}(o, a) = \min_j Q_{\phi_j}(o, a) - \lambda_R^{\text{on}} \hat{R}(o, a) - \lambda_S^{\text{on}} \hat{S}(o, a)$ over candidates drawn from both the current actor and π_{prior}. The comparison therefore isolates the incremental contribution of reliability-weighted transfer after controlling for the benefits of seeing the same offline data.

This control is intentionally strong. Uniform Behavior Cloning can already benefit from offline coverage, multimodal diffusion pretraining, and a better starting policy. If COOT still improves over this matched control, the advantage cannot be explained by offline data access, pretraining compute, actor capacity, or warm-start initialization. The remaining difference is precisely the mechanism proposed in this work: biased offline actions should not all be copied with equal strength. They should be transferred according to proxy-conditioned benefit, stability, and support, then revised through online evidence.

D.2 COOT Hyperparameters

Table 3 reports the default COOT hyperparameters used by the implementation. We group the parameters by the role they play in the algorithm rather than by implementation module. This organization highlights the main design principle of COOT: offline data are not treated as uniformly reliable demonstrations, but are first converted into calibrated transfer signals and then used as a persistent, filtered prior during online learning.

The table separates four coupled sources of behavior. The proxy-informed reliability transfer bundle determines *which* logged actions are trusted by combining proxy-conditioned value, uncertainty, and support information. The diffusion actor-critic determines *how* action proposals are represented. Offline pretraining determines *how strongly* reliable offline actions shape the initial policy and critics. Online candidate selection determines *when* the offline prior should influence interaction after online evidence becomes available. This separation is central to COOT's interpretability: the method transfers useful structure from biased offline data while retaining online correction through critic-guided and support-aware action selection.

D.3 Confounded Offline Datasets

The offline datasets are designed to test whether an offline-to-online learner can use biased experience without mistaking logged correlations for reliable action effects. For each environment and seed, we collect a dataset

$$\mathcal{D}_{\text{off}} = \{(o_i, a_i, r_i, o'_i, d_i, z_i, w_i, \ell_i, G_i)\}_{i=1}^{N},$$

Table 3: Default COOT hyperparameters, grouped by algorithmic role.

Hyperparameter	**Value**	**Role in COOT**
Proxy-informed reliability transfer bundle		
Discount factor γ	0.99	Defines empirical discounted returns and Bellman targets.
Bridge ridge regularization α_{br}	5.0	Stabilizes proxy-conditioned return regressions.
Potential ridge regularization α_Φ	2.0	Regularizes the state-only potential fit $\hat{\Phi}(o)$.
Risk ridge regularization α_R	2.0	Regularizes the proxy-free risk regression $\hat{R}(o, a)$.
Support classifier strength C_{sup}	0.5	Inverse-regularization of the behavior-support logistic model.
Bridge bootstrap repeats B	5	Estimates uncertainty by resampling proxy-conditioned bridge models.
Pessimism coefficient β_{br}	1.0	Forms a lower-confidence proxy-conditioned value estimate.
Benefit temperature τ_{off}	2.0	Scales the standardized robust advantage in the offline reliability score.
Offline risk coefficient λ_R	0.75	Penalizes logged actions whose bridge value is unstable.
Offline support coefficient λ_S	0.25	Penalizes logged actions outside reliable behavior support.
Diffusion actor-critic architecture		
Policy and critic hidden width	256	Network capacity for actor and critic approximation.
Number of hidden layers	3	Depth of the policy, critic, and denoising networks.
Diffusion denoising steps	20	Discretization level of the diffusion action sampler.
Diffusion hidden width	256	Capacity of the denoising model used by the actor.
Number of action particles	16	Diversity of stochastic diffusion action proposals.
Diffusion noise scale	0.08	Sampling noise scale for the diffusion policy.
Beta schedule scale	0.8	Controls the diffusion noise schedule.
Target entropy scale	0.9	Sets the entropy target for actor-critic temperature adaptation.
Offline pretraining		
Offline transitions	300k	Default confounded-dataset size used in all main experiments.
Offline pretraining updates	30k	Weighted diffusion and critic updates before online learning, used by default in all main experiments and varied over $10\text{k}, 20\text{k}, 30\text{k}$ in the sensitivity study.
Batch size	256	Mini-batch size for critic and weighted denoising updates.
Learning rate	3×10^{-4}	Optimizer step size for policy and critic networks.
Temperature learning rate	3×10^{-3}	Optimizer step size for the entropy-temperature parameter.
Target-network update rate τ	0.005	Polyak averaging rate for target networks.
Reward scale	0.2	Reward rescaling factor for stable critic learning.
Online prior-guided learning		
Current-policy candidates	$K_\theta = 6$	Candidate actions drawn from the current diffusion actor and scored by Eq. 9.
Offline-prior candidates	$K_{\text{prior}} = 2$	Candidate actions drawn from the persistent pretrained prior and scored by Eq. 9.
Actor guide temperature τ_{act}	5.0	Controls the sharpness of online guide weights.
Online risk coefficient λ_R^{on}	0.30	Penalizes high-risk candidate actions during online selection.
Online support coefficient λ_S^{on}	0.10	Penalizes unsupported candidate actions during online selection.
Initial shaping scale	1.0	Strength of dense early critic guidance from the learned state potential.
Shaping decay horizon	100k	Steps over which potential shaping is annealed to zero.

where o_i and o'_i are consecutive observations, a_i is the logged behavior action, r_i is the reward observed under the confounded environment, d_i is the terminal indicator, z_i and w_i are observed proxy variables, ℓ_i is the behavior-policy log probability, and G_i is the empirical discounted return. The generator also records the latent regime u_i for diagnostics, but COOT never uses u_i during training. This separation is important: the benchmark provides proxy information, not privileged access to the hidden regime.

All five environments use the same data-generation principle. At each step, the environment maintains an unobserved latent regime u_t. The behavior policy observes o_t only through the data-collection procedure, but its action distribution is also conditioned on u_t. A behavior policy k generates

$$\mu_k(o_t, u_t) = \tanh(f_k(o_t) + v_k u_t + b_k)\,, \qquad a_t = \text{clip}\,(\mu_k(o_t, u_t) + \epsilon_t, -1, 1)\,, \qquad \epsilon_t \sim \mathcal{N}(0, \sigma_k^2 I), \tag{61}$$

where f_k is a task-scaled linear observation map. A small ensemble of such policies is randomly initialized, evaluated in the confounded environment, and the strongest candidates are retained for data collection. Sampling from this ensemble produces heterogeneous but structured behavior, avoiding the degenerate case of a single narrow logging policy.

The key confounding structure is

$$u_t \to a_t, \qquad u_t \to (r_t, o_{t+1}), \qquad u_t \to (z_t, w_t),$$

while the learner observes only $(o_t, a_t, r_t, o_{t+1}, z_t, w_t)$. Thus, the same hidden regime can influence which action is logged, how the transition unfolds, and what reward is observed. A logged action with high return may therefore be genuinely useful, or it may only look useful because it was collected under a favorable hidden regime. This is precisely the ambiguity COOT is designed to address.

The proxy variables z_t and w_t make the setting informative rather than hopeless. They are noisy observable summaries produced by the confounded environment, and they carry partial information about the latent regime and transition context without revealing u_t. COOT uses these proxies to construct proxy-conditioned bridge estimates, uncertainty scores, state potentials, and support penalties. The dataset therefore represents a realistic middle ground: it is not a clean expert dataset, but it is also not arbitrary corrupted data. It contains useful behavioral evidence whose reliability must be estimated before transfer.

After collection, returns are computed within trajectories by the backward recursion

$$G_i = r_i + \gamma(1 - d_i)G_{i+1}, \tag{62}$$

with the recursion reset at terminal transitions. The stored dataset contains observations, logged actions, rewards, next observations, terminal indicators, two proxy views, behavior log probabilities, empirical returns, and the latent diagnostic value. During COOT training, only the observed transition fields and proxies are used.

This construction deliberately separates *offline association* from *deployable action value*. Direct behavior cloning treats every logged action as equally trustworthy and can therefore amplify confounded action-outcome correlations. COOT instead treats the offline data as informative but biased: logged actions receive high influence only when they have robust proxy-conditioned advantage, low uncertainty, and sufficient support. The confounded datasets therefore evaluate the central claim of COOT: biased offline experience can accelerate online reinforcement learning when it is converted into calibrated, risk-aware, and support-aware transfer guidance rather than used as clean supervision.

D.4 Ablation Studies and Failure-Mode Analysis

The results in Section 4.2 demonstrate that COOT improves both peak policy quality and online sample efficiency under biased offline data. This section attributes these gains to the reliability mechanisms introduced by COOT and examines their robustness under imperfect proxy information. Unlike generic offline pretraining, COOT relies on the premise that proxy-conditioned reliability structure can be converted into actionable online inductive bias. We therefore isolate each transfer pathway in the proxy-informed reliability transfer bundle while keeping the diffusion actor-critic, offline data, online budget, and all remaining hyperparameters fixed. This allows the performance change of each variant to be interpreted as the contribution

of a specific reliability pathway rather than as an artifact of model capacity or optimization. We further stress-test the proxy premise by corrupting, dropping, and misaligning proxy variables, reflecting practical deployment settings where auxiliary telemetry may be noisy, incomplete, or spuriously correlated with outcomes. These studies show not only that COOT improves performance, but also why the improvement occurs: the proxy-conditioned bridge provides the dominant reliability-transfer signal, the prior, shaping, and support terms play complementary roles in converting this signal into efficient online improvement, and the uncertainty- and support-aware design causes COOT to degrade conservatively rather than amplify unreliable proxy information.

All ablations are run on HalfCheetah under the default protocol of Section 4.1: 300k confounded offline transitions, 30k offline pretraining updates, and a 500k-step online budget. We report the same two learning-curve metrics used in the main experiments, $R_{\text{best}}^{500\text{k}}$ and $\text{AUC}^{500\text{k}}$. The former measures the best policy discovered within the online budget, while the latter measures how quickly and stably performance is accumulated throughout training. This distinction is central for offline-to-online transfer: a useful transfer mechanism should not only improve the final policy, but should also reduce the online exploration cost required to reach strong policies.

Component ablations. Table 4 isolates the role of each transfer pathway by disabling or weakening one mechanism at a time while keeping the diffusion actor-critic, offline dataset, online budget, and all other hyperparameters fixed. This controlled design makes the ablation diagnostic of the proposed reliability-transfer mechanism rather than of model capacity, training budget, or optimization differences. The *no proxy bridge* variant removes the proxy-conditioned bridge pathway, including bridge-induced offline weights, bridge risk, and bridge potential. The *no support penalty* variant sets the offline and online support coefficients to zero, testing whether explicit support-awareness is needed to avoid selecting actions outside reliable offline coverage. The *no persistent prior* variant removes offline-prior candidates from online action selection, testing whether the pretrained policy continues to provide useful candidate actions after online updates begin. The *no potential shaping* variant sets the shaping coefficient to zero, testing whether the bridge-derived potential provides useful dense feedback during online critic learning. The two *weak/noisy proxies* variants attenuate both proxy views by a retained signal scale κ and add Gaussian noise with scale σ, testing how COOT behaves as the proxy information used to infer hidden execution regimes becomes progressively less reliable.

The results show that every pathway contributes positively and that no ablated or weakened variant recovers full COOT performance. This supports the view that COOT's gains arise from an integrated reliability-transfer mechanism rather than from a single heuristic or from diffusion-policy pretraining alone. The proxy bridge is the dominant component. Removing it reduces $R_{\text{best}}^{500\text{k}}$ by 20.6% and $\text{AUC}^{500\text{k}}$ by 15.2%, producing the largest peak-return degradation in Table 4. This identifies the bridge as the load-bearing reliability module of COOT. Its role is to convert proxy information into a robust estimate of which logged actions are likely to remain useful under the target online regime. Without this module, the learner loses the mechanism that separates actions with reliable transfer value from actions that only appear favorable because they were collected or executed under hidden regimes in the offline data.

The non-bridge ablations reveal a finer decomposition of how COOT converts the bridge signal into online improvement. Removing the persistent prior causes the largest AUC degradation in the table, -17.4%, while retaining a higher peak return than the no-bridge variant. This is the expected signature of an early-learning accelerator. The persistent prior keeps offline-informed actions available in the online candidate set while the critic is still poorly calibrated, thereby reducing the exploration burden at the beginning of online training. Potential shaping shows a different profile. It reduces AUC by only 5.3%, but reduces peak return by 11.2%, indicating that shaping is not primarily an early-speed mechanism. Instead, it helps the critic and policy reach a stronger solution within the fixed online budget by translating the bridge potential into denser reward feedback. Support regularization occupies a stabilizing role. Its removal reduces AUC by 7.4% and peak return by 4.3%, suggesting that support-awareness mainly improves learning efficiency by discouraging optimistic candidate selection in regions where the offline data provide weak evidence.

The weak/noisy-proxy variants further clarify the role of proxy quality. With $\kappa = 0.75$ and $\sigma = 0.25$, COOT loses 11.7% in peak return and 13.1% in AUC. When the proxy signal is further weakened to $\kappa = 0.50$ and $\sigma = 0.50$, the degradation increases to 15.6% in peak return and 15.2% in AUC. This ordered degradation

Table 4: **Component ablations of COOT on HalfCheetah** within the 500k-step online budget. Each row disables or weakens one transfer signal from the proxy-informed reliability transfer bundle while holding the diffusion actor-critic, offline dataset, online budget, and all remaining hyperparameters fixed. For the weak/noisy-proxy variants κ is the retained proxy-signal scale and σ the additive Gaussian noise scale. $R^{500\text{k}}_{\text{best}}$ measures the best policy found within the budget and $\text{AUC}^{500\text{k}}$ summarizes sample efficiency over the full learning curve, while the Δ columns give the relative change versus full COOT. Every pathway contributes positively and no variant recovers full performance, identifying the proxy bridge as the load-bearing reliability module.

Variant	$R^{500\text{k}}_{\text{best}} \uparrow$	$\Delta R\,(\%)$	$\text{AUC}^{500\text{k}} \uparrow$	$\Delta\text{AUC}\,(\%)$
No proxy bridge	8489.3	−20.6	7095.2	−15.2
No persistent prior	9109.7	−14.8	6914.0	−17.4
No potential shaping	9492.0	−11.2	7929.3	−5.3
Weak / noisy proxies ($\kappa = 0.75, \sigma = 0.25$)	9437.2	−11.7	7277.9	−13.1
Weak / noisy proxies ($\kappa = 0.50, \sigma = 0.50$)	9022.2	−15.6	7095.9	−15.2
No support penalty	10231.7	−4.3	7751.9	−7.4
COOT (full)	**10688.9**	0.0	**8370.5**	0.0

shows that the proxy bridge is not merely acting as a generic regularizer. Its benefit depends on the amount of usable proxy information available for capturing hidden execution-regime information. At the same time, the weakened-proxy variants remain comparable to the no-bridge condition in sample efficiency and retain higher peak returns than the no-bridge variant. This indicates that COOT can still extract value from partially informative telemetry rather than requiring perfectly measured proxies. Overall, the ablations support the intended design of COOT: the bridge supplies the reliability signal, the persistent prior improves early candidate generation, potential shaping densifies online feedback, and support regularization prevents unsafe extrapolation from confounded offline data.

Failure modes: spurious, incomplete, and misaligned proxies. COOT assumes that the observed proxies contain partial information about the hidden execution regime. This assumption is substantially weaker than observing the regime itself, but it is still testable: if proxies are unrelated to the latent regime, missing for many transitions, or misaligned with the corresponding outcomes, the bridge should become less informative and the offline-to-online transfer benefit should decrease. Table 5 therefore stress-tests the proxy pathway directly. In the *spurious* setting, proxies are replaced by independent Gaussian noise, so the bridge receives variables with no sample-level regime content. In the *incomplete* setting, a fraction of proxy entries is randomly zeroed, emulating telemetry dropout or missing auxiliary logs. In the *shuffled* setting, proxy values are permuted across samples, preserving the marginal proxy distribution while breaking the alignment between proxies, actions, rewards, and next states. The shuffled test is especially diagnostic: it rules out the possibility that COOT benefits only from marginal proxy statistics, since useful transfer requires the proxies to remain aligned with the transition on which the bridge is trained.

The results show that proxy violations primarily reduce sample efficiency, while the online learner continues to improve under all tested corruptions. The strongest degradation occurs under spurious proxies with $\sigma_{\text{px}} = 1.0$, where $R^{500\text{k}}_{\text{best}}$ drops by 18.8% and $\text{AUC}^{500\text{k}}$ drops by 17.5%. This is the expected behavior: when the bridge is fitted to uninformative auxiliary variables, it can no longer reliably identify which logged actions provide useful transfer signal rather than hidden-regime bias. Importantly, this degradation is not a catastrophic policy failure. The peak return remains 8681.7, which is above the no-bridge peak return in Table 4, although its AUC is lower than the no-bridge AUC. This distinction is informative. It indicates that misleading proxies mainly damage the early transfer signal that accelerates online learning, while the online reinforcement learning component can still recover useful behavior later in training.

The incomplete and shuffled proxy settings provide a more favorable but still revealing picture. With 80% proxy dropout, COOT obtains 9594.7 peak return and 7229.0 AUC, corresponding to 10.2% and 13.6% degradation relative to clean proxies. With 50% dropout, the peak return nearly recovers the clean-proxy result, reaching 10423.5, only 2.5% below full COOT, while AUC remains 7.3% lower. Thus, partial telemetry can still be sufficient to identify strong actions within the online budget, but missing proxy information

Table 5: **Proxy failure-mode analysis on HalfCheetah** within the 500k-step online budget. Proxies are degraded during bundle fitting and deployment by additive noise (*spurious*), random entry dropout (*incomplete*), or batch-level permutation (*shuffled*), and the Δ columns report the relative change versus clean-proxy COOT. Unreliable proxies primarily reduce sample efficiency, while partially informative and sample-aligned proxies retain substantial offline-to-online transfer benefit, showing that COOT degrades gracefully rather than depending blindly on corrupted telemetry.

Deployment proxy condition	$R_{\text{best}}^{500\text{k}} \uparrow$	$\Delta R\,(\%)$	$\text{AUC}^{500\text{k}} \uparrow$	$\Delta\text{AUC}\,(\%)$
Spurious ($\sigma_{\text{px}} = 1.0$)	8681.7	−18.8	6904.9	−17.5
Incomplete (dropout 0.8)	9594.7	−10.2	7229.0	−13.6
Shuffled (misaligned)	9622.3	−10.0	7225.6	−13.7
Incomplete (dropout 0.5)	10423.5	−2.5	7760.8	−7.3
COOT (clean proxies)	**10688.9**	0.0	**8370.5**	0.0

slows down the conversion of offline experience into early online improvement. The shuffled condition shows a similar sample-efficiency loss, with 9622.3 peak return and 7225.6 AUC (10.0% and 13.7% below clean proxies). Because shuffling preserves the marginal proxy distribution but destroys sample-level correspondence, this result confirms that COOT is exploiting transition-aligned proxy information rather than generic proxy statistics.

These failure modes clarify the practical boundary of COOT. Clean and aligned proxies provide the strongest transfer. Incomplete proxies retain much of the peak-performance benefit but reduce early learning efficiency. Misaligned or spurious proxies remove a substantial fraction of the bridge advantage, especially in AUC. This behavior is desirable for an offline-to-online method under hidden confounding: COOT benefits from auxiliary telemetry when it carries usable regime information, but when that information becomes unreliable, the method degrades toward conservative online improvement rather than depending blindly on corrupted proxy signals.

Overall, the ablation and failure-mode studies support COOT's main mechanism claim. The component ablations show that the gains are not explained by diffusion pretraining alone: removing the bridge, prior, shaping, or support pathway consistently reduces peak policy quality, online sample efficiency, or both. The proxy failure-mode analysis further shows that COOT behaves predictably when its proxy informativeness is weakened. Informative and aligned proxies produce the strongest transfer, incomplete proxies preserve much of the peak-performance benefit but slow down learning, and misaligned or spurious proxies remove a substantial portion of the bridge advantage. This is the desired behavior for a practical offline-to-online method under hidden confounding: COOT exploits auxiliary telemetry when it contains useful regime information, and falls back toward more conservative online improvement when that telemetry becomes less reliable.

D.5 Offline Covariate Shift Diagnostics

We diagnose whether the confounded offline datasets induce a genuine shift relative to ordinary target-environment interaction. This check follows the standard covariate-shift view, where the input distribution changes across domains while the learning algorithm must still transfer useful information from the source data (Sugiyama et al., 2007). For each environment, we compare the confounded offline dataset against a plain reference dataset generated in the corresponding unconfounded MuJoCo environment. The plain reference data follow the same storage schema, with proxy and latent fields filled only for compatibility, and are used only for diagnostics rather than COOT training.

Let $x = [o, a]$ denote a concatenated state-action feature. For each environment, samples from the confounded and plain datasets are standardized with shared pooled statistics,

$$\tilde{x} = \frac{x - \hat{\mu}_{\text{pool}}}{\hat{\sigma}_{\text{pool}} + \varepsilon}, \tag{63}$$

Table 6: State-action covariate-shift diagnostics for seed 100 with 300k transitions. Domain AUC measures separability between confounded and plain samples. Random AUC is a random-label sanity check. D_μ is the standardized mean distance, and MMD^2 is the Gaussian-kernel two-sample distance.

Environment	**Domain AUC**	**Random AUC**	D_μ	MMD^2	**Δ Reward**	**Δ Return**
HalfCheetah	0.940	0.494	2.243	0.187	−0.137	−13.094
Ant	0.959	0.485	2.056	0.113	0.732	66.929
Hopper	0.927	0.492	1.525	0.100	0.351	2.640
Humanoid	0.997	0.498	8.068	0.084	0.098	24.788
Walker2d	0.871	0.494	1.392	0.067	0.159	8.303

where $\hat{\mu}_{\text{pool}}$ and $\hat{\sigma}_{\text{pool}}$ are computed from the union of both domains. Shared standardization makes the diagnostics comparable across coordinates and ensures that the measured discrepancy reflects distributional structure rather than raw feature scale.

Domain-classifier AUC. We train a balanced logistic domain classifier to distinguish confounded samples from plain samples, following classifier two-sample testing (Lopez-Paz & Oquab, 2017). The binary domain label is

$$y = 1\left[\tilde{x} \sim \mathcal{D}_{\text{conf}}\right]. \tag{64}$$

The resulting area under the ROC curve measures how easily the two state-action distributions can be separated (Hanley & McNeil, 1982). An AUC near 0.5 indicates that the two domains are indistinguishable to the classifier, while an AUC near 1.0 indicates strong covariate shift. We also compute a random-label AUC sanity check, which remains close to chance and verifies that high domain AUC is not an artifact of sample size or classifier calibration.

Standardized mean distance. We compute the Euclidean distance between standardized state-action means. This statistic is a multivariate analogue of standardized mean-difference diagnostics used to assess covariate imbalance across groups (Rubin, 2001; Austin, 2009):

$$D_\mu = \left\| \frac{1}{n_{\text{conf}}} \sum_{i=1}^{n_{\text{conf}}} \tilde{x}_i^{\text{conf}} - \frac{1}{n_{\text{plain}}} \sum_{j=1}^{n_{\text{plain}}} \tilde{x}_j^{\text{plain}} \right\|_2 . \tag{65}$$

This diagnostic captures first-order shift in the average state-action representation. Because it aggregates across all coordinates, high-dimensional environments can have larger values when many coordinates shift moderately at the same time.

Gaussian MMD. We also compute the squared maximum mean discrepancy with a Gaussian kernel, a standard kernel two-sample statistic for measuring distributional discrepancy (Gretton et al., 2012):

$$\begin{aligned} \mathrm{MMD}^2\left(\mathcal{D}_{\text{conf}}, \mathcal{D}_{\text{plain}}\right) = \frac{1}{n_{\text{conf}}^2} \sum_{i,i'} k\left(\tilde{x}_i^{\text{conf}}, \tilde{x}_{i'}^{\text{conf}}\right) + \frac{1}{n_{\text{plain}}^2} \sum_{j,j'} k\left(\tilde{x}_j^{\text{plain}}, \tilde{x}_{j'}^{\text{plain}}\right) \\ - \frac{2}{n_{\text{conf}} n_{\text{plain}}} \sum_{i,j} k\left(\tilde{x}_i^{\text{conf}}, \tilde{x}_j^{\text{plain}}\right), \end{aligned} \tag{66}$$

$$k(x, y) = \exp\left(-\frac{\|x - y\|_2^2}{2\sigma^2}\right), \tag{67}$$

where σ is chosen by the median heuristic. Unlike D_μ, MMD compares kernel mean embeddings and can detect distributional differences beyond a shift in the mean, including changes in local geometry and higher-order structure.

The diagnostics in Figures 3 to 5 and Table 6 show that the confounded datasets are not merely noisy versions of plain MuJoCo interaction data. The domain-classifier AUC is far above chance for every task, ranging from 0.871 on Walker2d to 0.997 on Humanoid, while the random-label AUC remains close to 0.5.

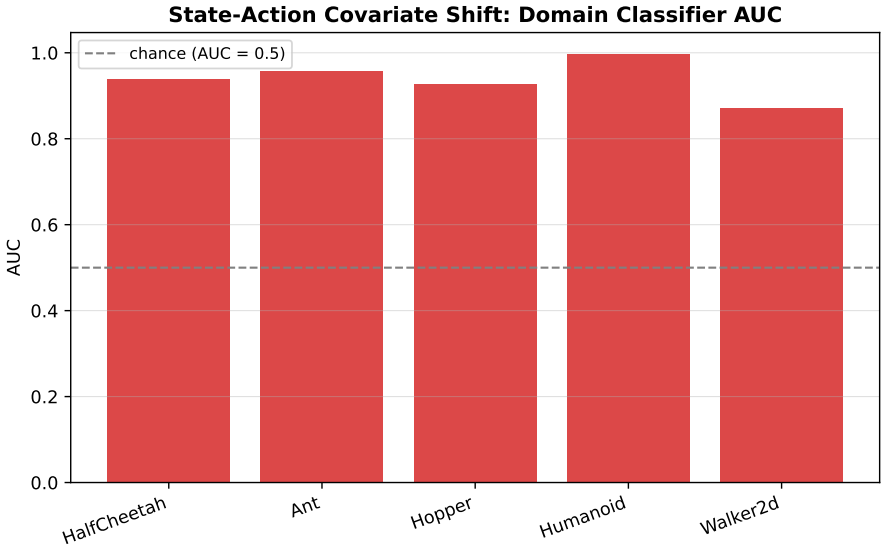

Figure 3: State-action domain-classifier AUC between confounded and plain reference datasets. Values far above chance show that the confounded data-generation process induces state-action distributions that are systematically distinguishable from ordinary interaction data.

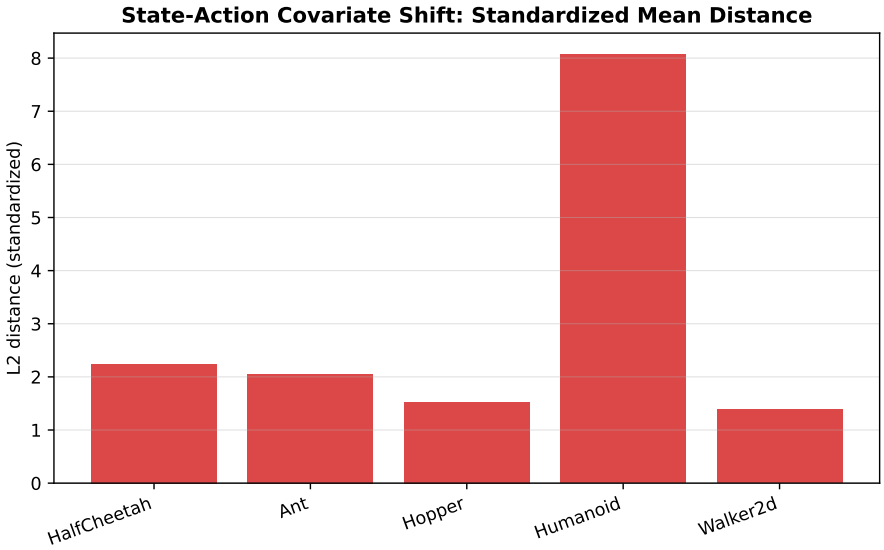

Figure 4: Standardized state-action mean distance. The large Humanoid value reflects accumulated coordinate-wise shifts in a high-dimensional state-action representation.

This confirms that the confounded data generator changes the state-action distribution in a systematic and measurable way.

The three diagnostics expose complementary aspects of this shift. Domain AUC shows separability between source and target state-action populations. Standardized mean distance shows how far the average confounded representation moves from the plain reference distribution. Humanoid obtains the largest D_μ because its observation-action representation has many dimensions, so moderate coordinate-level changes accumulate under the ℓ_2 norm. This explains why Humanoid can have the largest domain AUC and mean distance while not having the largest MMD. MMD depends on the Gaussian kernel scale and measures distributional discrepancy through kernel embeddings, whereas D_μ grows directly with accumulated mean displacement.

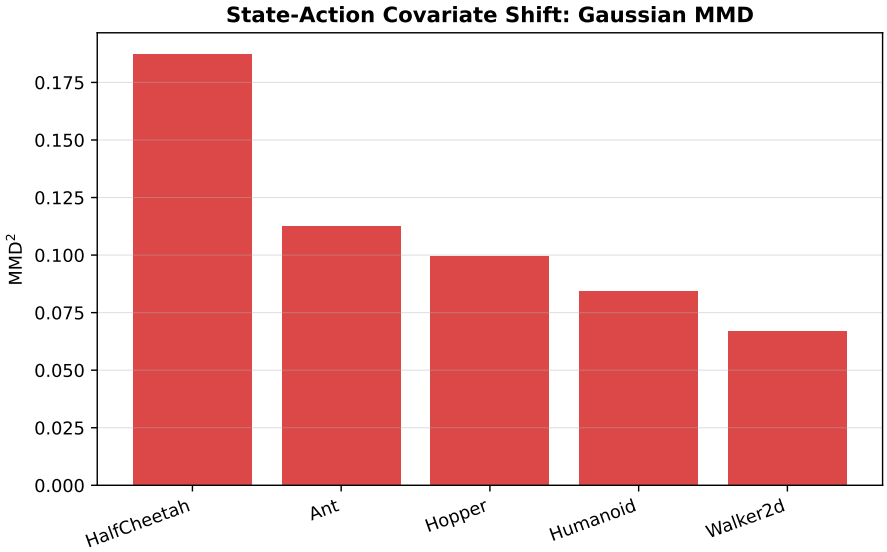

Figure 5: Gaussian-kernel MMD between confounded and plain state-action distributions. Positive values across all environments show distributional discrepancy beyond a simple mean shift.

These results strengthen the experimental motivation for COOT. The offline data contain useful trajectories, rewards, actions, and proxy variables, but the logged state-action-outcome associations are shifted relative to the target online environment. This is precisely the regime where direct behavior cloning or unfiltered offline value transfer can be misleading. COOT is designed to use such data cautiously: proxy-conditioned bridge estimates identify robust advantage, bootstrap disagreement estimates risk, and the support model discourages unsupported action transfer. The covariate-shift diagnostics therefore support the central claim of the paper: confounded offline data can accelerate online reinforcement learning when it is transformed into risk-aware and support-aware guidance rather than treated as clean expert supervision.