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**onfounded **O**ffline-to-**O**nline **T**ransfer), a reliability-weighted framework inspired by causal reasoning about confounded logs 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**onfounded **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 role of causal reasoning in COOT is deliberately modest. COOT does not perform causal identification and does not estimate the interventional value $\mathbb{E}[G \mid o, \mathrm{do}(a)]$; its bridge regressors fit observational, proxy-conditioned returns by standard supervised learning. What the causal model of Section 2 contributes is a reason to distrust certain observational correlations: it identifies the residual gap between the proxy-conditioned signal and the online effect needed for improvement as the quantity that transfer must respect. 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. COOT is therefore best described as a reliability-weighted offline-to-online framework inspired by causal reasoning, contributing a practical mechanism for deciding which offline samples should influence online learning rather than a causal estimation procedure.

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 reliability views of hidden execution regimes rather than as ordinary auxiliary features or as causal adjustment variables. 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. Here o_t is the fully observed state, playing the role of s_t in a standard MDP, while u_t is an exogenous execution condition rather than a POMDP hidden state, and g is an execution map COOT never estimates. The logged action $a_t^{\log}$ and executed action a_t^{exec} coincide only in the fully observed ideal case, where g ignores u_t. The latent regime u_t may represent actuator wear, environment drift, operator intent, demand state, or market liquidity. The learner observes

$$\mathcal{D}_{\text{off}} = \{(o_t, a_t^{\log}, r_t, o_{t+1}, z_t, w_t)\}_{t=1}^{N},$$

where z_t, w_t are proxy variables that provide partial information about u_t 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 $u_t \to a_t^{\text{exec}} \to (r_t, o_{t+1})$ 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:

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

Here do(·) is Pearl's intervention operator (Pearl, 2009), and both sides average the same regime-conditioned reward over u_t, the left under its posterior given that the action was logged and the right under its state-conditional prior (Appendix B.1, which also fixes the reading of every symbol in Eq. 1). This gap is invisible from (o, a, r, o') alone, so a supported action may look good because it was logged under an easy regime, or 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: Confounded Offline-to-Online Transfer for Diffusion Policy Reinforcement Learning

3.1 From Confounded Logs to Reliability-Weighted 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 $\mathbb{E}[G \mid o, \text{do}(a)]$ is identified, and the bridge regressors estimate observational conditional expectations rather than causal effects. The causal structure is nevertheless useful because it specifies what must be estimated for cautious transfer: the reliability of the proxy-conditioned evidence relative to the hidden-regime bias in Eq. 2. Cross-proxy disagreement is accordingly interpreted as a reliability signal rather than as a direct estimate or proof of hidden confounding, since it can also reflect finite-sample, model, and optimization error. The confounding model explains why a systematic component of that disagreement is relevant to transfer, while bootstrap variation and support estimation prevent every disagreement from being attributed to confounding, and Appendix D.4 separates these components empirically. This still yields 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 reliability questions of benefit, stability, and support: a pessimistic proxy-stability value surrogate, a state potential, a deployable risk field, and a behavioral support penalty, where benefit is read from the surrogate relative to the potential. We write a_i for the logged action $a_i^{\log}$ and G_i for its empirical discounted return. 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 $\mathcal{I}_b$,

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

with ridge strength $\alpha_{\mathrm{br}} > 0$, each bridge approximating the proxy-conditioned return $\mathbb{E}[G \mid o, a, \nu]$. The predictor is called a *bridge* because it connects the unobserved regime u_t to observable predictions: conditioning on a proxy integrates over the regime through its proxy-conditioned posterior, so COOT compares the two proxies instead of recovering u_t (Appendix B.2). These predictions are not interpreted as causal effects. The two proxies are kept separate rather than concatenated, so that their *disagreement* becomes a usable warning signal: a return estimate that shifts with the proxy read is fragile evidence, while one that is stable across both proxies is more dependable, though still not certified unbiased, since two proxies 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, so agreement is measured as low pooled dispersion rather than pairwise equality of the two ensembles (Appendix B.2). 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. Appendix C.6 quantifies what this caution buys for transfer: the gap between $\hat{Q}_{\mathrm{rob}}$ and the interventional value splits into the measurable

width $2\beta_{\mathrm{br}}\hat{\sigma}_{\mathrm{br}}$ and an irreducible residual-confounding radius, so the surrogate never over-trusts what the proxies cannot certify. 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 over a ridge model class $\mathcal{F}_R$, and behavioral plausibility is scored by the positive-class probability $\hat{p}_{\mathrm{sup}}$ of a support classifier trained against three negative families of state-action pairs, namely permuted, uniform, and locally perturbed actions (defined in Appendices B.2 and B.3),

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

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 $\mathcal{D}_{\mathrm{off}}$ and fuses them into a reliability-adjusted desirability whose exponential, symmetrically clipped tilt becomes the transfer weight,

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

with empirical means and floored standard deviations taken over $\mathcal{D}_{\mathrm{off}}$. 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 $s_i, s_j \in [-c_{\mathrm{off}}, c_{\mathrm{off}}]$, the relative emphasis is $\omega_{\mathrm{off},i}/\omega_{\mathrm{off},j} = \exp(s_i - s_j)$, and every weight obeys $e^{-c_{\mathrm{off}}} \le \omega_{\mathrm{off}} \le e^{c_{\mathrm{off}}}$.*

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, proxy-informed 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 $\ell_{\mathrm{diff}}(\theta; o, a^{\mathrm{g}}) = \mathbb{E}_{t,\epsilon}\left[\left\|\epsilon - \epsilon_\theta(o, a_t, t)\right\|_2^2\right]$ 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 first projected onto a fixed leverage interval $[\omega_-, \omega_+]$ and then batch-normalized to fix the gradient scale,

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

Distributionally, Eq. 7 distills not the behavior policy $\mu(a \mid o)$ but a reweighted target $\propto \omega_{\mathrm{off}}(o, a, z, w)\, \mu(a \mid o)$, shifting mass toward logged actions that are robustly advantageous, low-risk, and well supported, and

Algorithm 1 COOT: Confounded Offline-to-Online Transfer

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

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).
Selective offline pretraining.
for each offline update **do**
  Sample $(o, a, r, o', d, z, w) \sim \mathcal{D}_{\text{off}}$. Set guide $a^{\text{g}} \leftarrow a$ and weight $\omega^{\text{g}} \leftarrow \omega_{\text{off}}(o, a, z, w)$.
  Shape reward and update critics (Eq. 8). Update π_θ by weighted denoising toward a^{g} (Eq. 7). Polyak-update targets.
end for
Freeze the target policy as the persistent prior π_{prior}.
Prior-guided online learning.
for each online step **do**
  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.
  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.
end for

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 projection and normalization preserve the bundle's ranking while keeping the influence of any sample provably bounded (Appendix B.7).

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, projected, and normalized as in Eqs. 6 and 7, then* $0 \le \bar{\omega}_{\text{off},i}\, \ell_{\text{diff}}(\theta; o_i, a_i) \le \kappa_\omega L_{\max}$ *with* $\kappa_\omega = \min\{e^{2c_{\text{off}}}, \omega_+/\omega_-\}$.

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, reward-scale-consistent shaping term, and forms bounded double-Q targets (Fujimoto et al., 2018; Haarnoja et al., 2018) with $a' \sim \pi_{\bar{\theta}}(\cdot \mid o')$,

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

where $\ell_H^{\delta_Q}$ is the Huber loss with threshold δ_Q and targets are additionally confined to $[-c_y, c_y]$. The clips and the Huber loss bound per-step shaping, bootstrapped targets, and the influence of any single transition, extending the bounded-influence principle of Proposition 2 to value learning (Appendices B.5 and B.7). Potential-based shaping leaves the optimal policy invariant for any fixed bounded $\hat{\Phi}$ (Ng et al., 1999) wherever the clip is inactive, and the anneal of η to zero makes the shaped and unshaped objectives coincide beyond the shaping horizon in all cases (Appendix C.5), so densification injects early signal without altering the solution. 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 standardized online value penalized by the surviving offline signals,

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

where K and M are the numbers of candidates *drawn per online step* from the current actor π_θ and the frozen prior π_{prior}, not history lengths or the replay-buffer size (defaults K=6, M=2; the offline experience is held in the frozen weights of π_{prior}), $\hat{Q}^{\,\text{std}}(o,a) = \text{clip}\left((\min_j Q_{\phi_j}(o,a) - \hat{\mu}_Q)/\hat{\sigma}_Q, -c_Q, c_Q\right)$ standardizes the clipped double-Q estimate by running critic statistics, and $\hat{R}_c, \hat{S}_c$ are the risk and support fields clipped to $[0, c_P]$, so the fixed prices $(\lambda_R^{\text{on}}, \lambda_S^{\text{on}})$ keep their meaning as the raw critic scale drifts over training (Appendix B.6). 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. Appendix C.7 formalizes this arbitration: whenever the candidate set retains a reliable action whose advantage clears the reliability price, the critic error, and the score-clipping slack, the rule yields a quantified policy-improvement guarantee, the precise sense in which the persistent prior is more than a warm start. The same rule supplies guide actions and their weights for the actor,

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

feeding the batch-normalized online objective

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

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.6). 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. All methods receive the same clean MuJoCo observation o_t with no added sensor-noise model, and none observes the hidden regime u_t; Appendix D.3 gives the full mapping from simulator quantities to the variables of Section 2. 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

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

the strongest evaluated policy found by that seed, and report $R^{500\text{k}}_{\text{best}}$ 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),

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

We report $\text{AUC}^{500\text{k}}$ as the mean $\pm$ standard deviation of $\text{AUC}^{500\text{k}}_s$ 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: $R^{500\text{k}}_{\text{best}}$ measures the strongest policy found within the budget, while $\text{AUC}^{500\text{k}}$ 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 $R^{500\text{k}}_{\text{best}}$ and $\text{AUC}^{500\text{k}}$ under the seed-first protocol of Section 4.1, where the five runs are five independent full-budget training seeds following standard practice in diffusion-policy RL (Appendix D.3). 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 control is a full online learner rather than a frozen clone, running COOT's online diffusion actor-critic on the same budget with only the reliability machinery disabled, and each

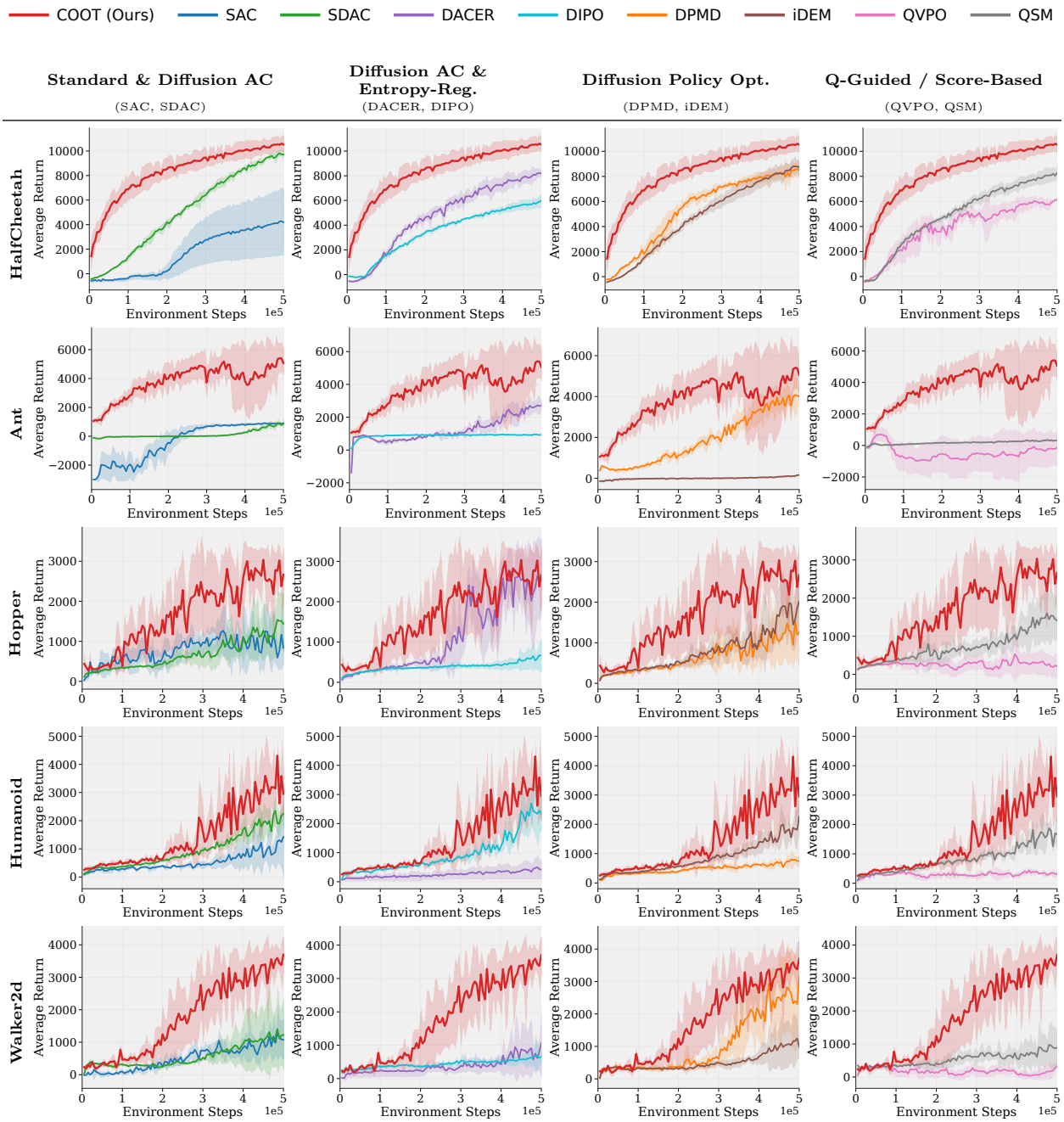

Figure 1: **Early online learning curves up to** 500k **environment steps** across five MuJoCo locomotion tasks. Rows are environments, and columns group COOT against one baseline family, namely standard and diffusion actor-critic methods, entropy-regularized diffusion actor-critic methods, diffusion-policy 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**

of the five seeds draws its own confounded dataset that both methods share within that seed (Appendices D.1 and D.3).

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 $R^{500\text{k}}_{\text{best}}$, and 9.4%, 32.2%, 64.2%, 56.5%, and 49.7% in $\text{AUC}^{500\text{k}}$ 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 $\text{AUC}^{500\text{k}}$ 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 vary-

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)	$\mathbf{8370.5 \pm 657.5}$	$\mathbf{3797.6 \pm 906.9}$	$\mathbf{1661.8 \pm 466.1}$	$\mathbf{1442.9 \pm 360.1}$	$\mathbf{1748.4 \pm 413.0}$

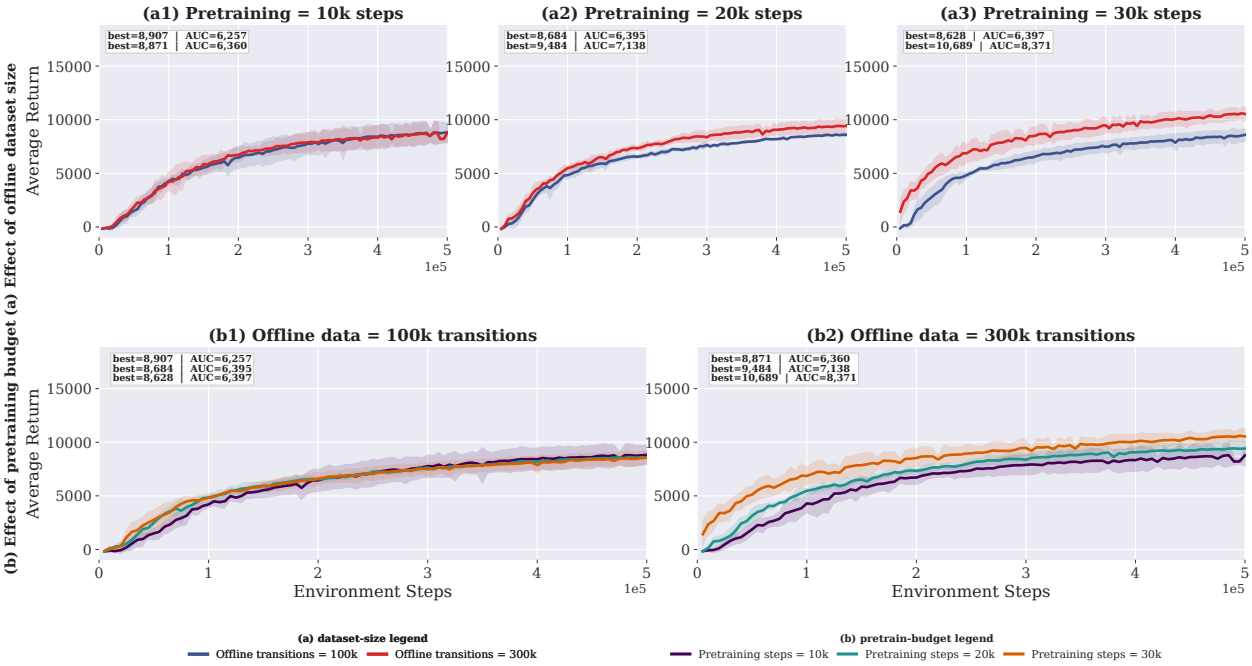

Figure 2: **COOT sensitivity to offline coverage and pretraining budget** on HalfCheetah up to 500k online steps, separating offline dataset size (top) from offline pretraining budget (bottom). Larger datasets help only when paired with sufficient pretraining, while longer pretraining barely helps under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

ing 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 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 one would collapse, whereas COOT tracks the amount of usable proxy information, exactly as a mechanism that reads proxy-conditioned reliability should. The same appendix isolates this mechanism from conventional uncertainty-aware transfer through two same-data controls, a no-proxy bootstrap ensemble and a proxy-concat ensemble, neither of which recovers full COOT performance (Appendix D.4), and its held-out variance decomposition states what bridge disagreement measures in practice: a cross-view instability cue for transfer reliability, not a calibrated estimate of hidden confounding (Appendix D.4). 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. Appendix D.6 extends this necessity check beyond synthetic confounding to realistic industrial telemetry, verifying out of sample that such telemetry is incrementally outcome-relevant, correctly event-aligned, and yields disagreement that tracks predictive unreliability, the empirical prerequisites of proxy-informed transfer.

5 Conclusion

We introduced COOT, a reliability-weighted framework for offline-to-online diffusion-policy reinforcement learning from confounded logs, inspired by causal reasoning about hidden execution regimes rather than performing causal identification. 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.

Zhihong Deng, Jing Jiang, Guodong Long, and Chengqi Zhang. Causal reinforcement learning: A survey. *arXiv preprint arXiv:2307.01452*, 2023.

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.

Shutong Ding, Ke Hu, Shan Zhong, Haoyang Luo, Weinan Zhang, Jingya Wang, Jun Wang, and Ye Shi. GenPO: Generative diffusion models meet on-policy reinforcement learning. *Advances in Neural Information Processing Systems*, 38:130443–130474, 2026.

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.

Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *Proceedings of the Nineteenth International Conference on Machine Learning (ICML)*, pp. 267–274. Morgan Kaufmann, 2002.

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.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York, NY, USA, second edition, 2009. ISBN 9780521895606.

Judea Pearl. Causal diagrams for empirical research. In *Probabilistic and Causal Inference: The Works of Judea Pearl*, pp. 255–316. Association for Computing Machinery, 2022.

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.

Yunpeng Qing, Yixiao Chi, Shuo Chen, Shunyu Liu, Kexuan Zhou, Sixu Lin, Litao Liu, and Changqing Zou. BiTrajDiff: Bidirectional trajectory generation with diffusion models for offline reinforcement learning. *arXiv preprint arXiv:2506.05762*, 2025.

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 inspired 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 Reading of Eq. 1: Relation to MDP and POMDP Formulations

Eq. 1 reuses symbols that carry established meanings in MDP and POMDP formulations, so this subsection fixes their reading precisely.

Observation, latent regime, and the POMDP contrast. COOT is a fully observed MDP augmented with an exogenous confounder, not a POMDP. The variable o_t is the fully observed state and plays exactly the role of s_t in a standard MDP; the difference from the usual notation is purely symbolic, and we write o_t because the offline logs also record proxy channels alongside it. The variable u_t is *not* a POMDP hidden state. It is an unobserved exogenous execution condition, such as actuator wear, calibration drift, or operator intent, that influences how actions are selected and realized. The distinction is structural: in a POMDP the agent's uncertainty concerns the state itself and $o_t \sim p(o_t \mid s_t)$ is a noisy measurement of it, whereas here the state is known exactly and the unobserved quantity is a separate factor acting on selection, execution, and dynamics. Consequently u_t is a hidden confounder in the sense of Eq. 2, not a state to be estimated, and COOT never attempts to infer it.

Direction of the factorization $u_t \sim p(u_t \mid o_t)$. Because o_t is observed and u_t is not, the generative direction is the reverse of the POMDP emission model. We do not write $o_t \sim p(o_t \mid u_t)$, which would make the state a measurement of the regime. Instead $p(u_t \mid o_t)$ states that which execution conditions are plausible may itself depend on the current state, for example when high-load configurations make actuator saturation more likely. The factorization is therefore a statement about where the regime is likely to arise, not about how the state is generated.

Why P_u conditions on both o_t and u_t. The two variables are distinct, non-nested causes that jointly determine the transition. The state o_t carries the ordinary dynamical information of an MDP, such as position and velocity, while u_t shifts the dynamics on top of it. Neither determines the other, so conditioning on u_t alone would discard the state dependence that any control problem requires, and conditioning on o_t alone would be the unconfounded model we depart from. Writing $P_u(\cdot \mid o_t, a_t^{\text{exec}}, u_t)$ makes the regime an additional source of dynamical variation rather than a substitute for the state.

The execution map g and the fully observed ideal case. The function g in Eq. 1 is an unspecified execution map sending the logged action to the action the environment actually realizes, as when a commanded wheel speed of 1 m/s is attenuated to 0.7 m/s by unobserved slip. COOT neither assumes a parametric form for g nor attempts to estimate it; the proxies are used to assess how reliable a logged action is under the prevailing execution conditions, which sidesteps g entirely. The fully observed ideal case referred to in Section 2 is exactly the condition that g ignores the regime,

$$g(o_t, a_t^{\log}, u_t) = a_t^{\log} \quad \text{for all } u_t, \qquad \text{hence} \qquad a_t^{\text{exec}} = a_t^{\log}, \tag{13}$$

under which the execution path of Eq. 1 vanishes and only the backdoor path through action selection remains.

The do operator and the expectations in Eq. 2. We use $\text{do}(\cdot)$ in the sense of Pearl's do-calculus (Pearl, 2009; 2022), denoting an action that is *forced* rather than one that is *observed to have occurred*; the notation is standard in causal reinforcement learning under unobserved confounding (Li et al., 2025c; Deng et al., 2023). Both sides of Eq. 2 average the same regime-conditioned reward function

$$m(o, a, u) = \mathbb{E}[r_t \mid o_t = o,\ a_t^{\text{exec}} = g(o, a, u),\ u_t = u] \tag{14}$$

over the hidden regime u, and differ only in the distribution used for that average. The observational quantity weights u by its posterior given that action a was in fact logged, whereas the interventional quantity weights u by its state-conditional prior, since forcing the action severs the dependence of selection on the regime,

$$\begin{aligned} \mathbb{E}\Big[r_t \mid o_t = o,\ a_t^{\log} = a\Big] &= \int m(o, a, u)\, p_{\text{off}}(u \mid o, a)\, du, \qquad p_{\text{off}}(u \mid o, a) = \frac{\mu(a \mid o, u)\, p(u \mid o)}{\int \mu(a \mid o, \tilde{u})\, p(\tilde{u} \mid o)\, d\tilde{u}}, \\ \mathbb{E}\Big[r_t \mid o_t = o,\ \text{do}(a_t^{\log} = a)\Big] &= \int m(o, a, u)\, p(u \mid o)\, du. \end{aligned} \tag{15}$$

The inequality in Eq. 2 is thus a statement about these two averaging measures rather than about the reward function itself: they coincide precisely when the logging policy μ does not depend on u, so that $p_{\text{off}}(u \mid o, a) = p(u \mid o)$. Concretely, if a 1 m/s command is logged on dry floor 90% of the time but dry and slippery floors are equally likely under intervention, the logged reward exceeds the interventional one

because the command was recorded predominantly under favorable conditions, not because it causes better outcomes. This is the bias COOT is built to discount.

B.2 The Bridge Model: Meaning, Regularization, and Agreement

What the bridge is and why it is called a bridge. A *bridge* in COOT is a lightweight regression model that predicts the expected return from the observed state, action, and a single proxy view. With the regime-conditioned reward $m(o, a, u) = \mathbb{E}[G \mid O = o, A^{\log} = a, U = u]$ of Eq. 14, each view predictor targets

$$\hat{h}_z(o, a, z) \approx \mathbb{E}[G \mid o, a, z] = \int m(o, a, u)\, p_{\text{off}}(u \mid o, a, z)\, du,$$
$$\hat{h}_w(o, a, w) \approx \mathbb{E}[G \mid o, a, w] = \int m(o, a, u)\, p_{\text{off}}(u \mid o, a, w)\, du, \tag{16}$$

with all expectations taken under the offline-data distribution. The name reflects the role of these models rather than any particular functional form. The hidden regime u_t is never observed, but it leaves a measurable imprint on the proxies z_t and w_t; each bridge *connects* that unobserved regime to observable predictions by conditioning on a proxy view, so $\mathbb{E}[G \mid o, a, \nu]$ implicitly integrates over the regime through its proxy-conditioned posterior $p_{\text{off}}(u \mid o, a, \nu)$. COOT never recovers u_t directly. It fits one bridge per view and compares their return predictions: when the regime barely affects the return, the two views predict similarly and the evidence is treated as reliable, whereas when the regime strongly affects the return, the two proxies can capture different parts of that effect and the bridges disagree, so COOT treats the action as uncertain and uses it cautiously.

The ridge coefficient is strictly positive. The regularization strength $\alpha_{\text{br}} > 0$ in Eqs. 3 and 21 is a genuine ridge penalty. A positive value shrinks the bridge coefficients, stabilizing the fit and reducing overfitting of the proxy-conditioned return, which matters here because the bootstrap dispersion of these fits is itself used downstream as a reliability signal. The limit $\alpha_{\text{br}} = 0$ removes regularization, and $\alpha_{\text{br}} < 0$ is inadmissible because it would reward large coefficients and destabilize the fit. All experiments use $\alpha_{\text{br}} = 5.0$ (Table 3).

Agreement is low pooled dispersion, not pairwise equality. The word *agreement* refers to the spread of the pooled prediction set, not to index-by-index equality of the two view ensembles. For a query (o, a, z, w), the $2B$ predictions are pooled into $\mathcal{B}(o, a, z, w) = \{\hat{h}_z^{(1)}, \ldots, \hat{h}_z^{(B)}, \hat{h}_w^{(1)}, \ldots, \hat{h}_w^{(B)}\}$, and agreement is quantified by the pooled dispersion $\hat{\sigma}_{\text{br}}$ of Eq. 23. Writing $\bar{h}_z, \bar{h}_w$ for the two view means and s_z^2, s_w^2 for their within-view bootstrap variances, the pooled variance decomposes as

$$\hat{\sigma}_{\text{br}}^2 = \tfrac{1}{2} s_z^2 + \tfrac{1}{2} s_w^2 + \tfrac{1}{4}\big(\bar{h}_z - \bar{h}_w\big)^2, \tag{17}$$

so $\hat{\sigma}_{\text{br}}$ fuses within-view bootstrap variability with between-view disagreement in a single scalar. Small $\hat{\sigma}_{\text{br}}$ means strong agreement and large $\hat{\sigma}_{\text{br}}$ means weak agreement. Two consequences make the definition precise. Exact equality of all $2B$ predictions is the only case giving $\hat{\sigma}_{\text{br}} = 0$, and it is sufficient but not necessary for high reliability, since COOT rewards low spread rather than exact equality. Conversely, the predictions are not compared pairwise at matched bootstrap index: even if $\hat{h}_z^{(b)} = \hat{h}_w^{(b)}$ for every b, the pooled dispersion can remain positive through the within-view terms s_z^2, s_w^2 when predictions vary across resamples. The empirical size of the between-view term relative to the within-view terms is reported in Appendix D.4.

The risk model class $\mathcal{F}_R$. The proxy-free risk field of Eqs. 5 and 27 regresses the bridge dispersion $\hat{\sigma}_{\text{br}}(o, a, z, w)$, computable offline where proxies are present, onto the proxy-free pair (o, a) that is all the online selector observes. The class $\mathcal{F}_R$ is the hypothesis set for this regression: in our implementation, ridge-regularized linear models over standardized state-action features,

$$\mathcal{F}_R = \big\{\, R(o, a) = \theta^\top \phi(o, a) + b \;\big|\; \theta \in \mathbb{R}^{d_\phi},\, b \in \mathbb{R} \,\big\}, \qquad \phi(o, a) = \big[\, \tilde{o}^\top \;\; \tilde{a}^\top \,\big]^\top, \tag{18}$$

where $\tilde{o}, \tilde{a}$ are standardized state and action features. Fitting $R \in \mathcal{F}_R$ distills proxy-derived uncertainty into a score deployable from (o, a) alone: a logged action that the proxies flagged as high-dispersion trains R to return a high risk score for similar state-action inputs online, where the proxies are unavailable. The support model class is fixed analogously as class-balanced logistic regression over the same features.

B.3 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

$$h^\star_\nu(o, a, \nu) = \mathbb{E}[\, G \mid o, a, \nu \,], \qquad Q^\star_{\text{prox}}(o, a, z, w) = \mathbb{E}[\, G \mid o, a, z, w \,], \tag{19}$$

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. 19 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

$$x_i^\nu = \left[\, o_i^\top \;\; a_i^\top \;\; \nu_i^\top \,\right]^\top, \qquad \nu \in \{z, w\}, \tag{20}$$

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

$$\hat{h}_\nu^{(b)} \in \operatorname*{arg\,min}_{h \in \mathcal{H}} \sum_{i \in \mathcal{I}_b} \big(G_i - h(x_i^\nu)\big)^2 + \alpha_{\text{br}} \|h\|_2^2, \qquad b = 1, \ldots, B. \tag{21}$$

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),

$$H(o, a, z, w) = \big(\hat{h}_z^{(1)}(o, a, z), \ldots, \hat{h}_z^{(B)}(o, a, z),\; \hat{h}_w^{(1)}(o, a, w), \ldots, \hat{h}_w^{(B)}(o, a, w)\big)^\top \in \mathbb{R}^{2B}, \tag{22}$$

its first two moments give the ensemble summary

$$\hat{\mu}_{\text{br}} = \tfrac{1}{2B}\, \mathbf{1}^\top H, \qquad \hat{\sigma}_{\text{br}} = \left(\tfrac{1}{2B} \left\| H - \hat{\mu}_{\text{br}} \mathbf{1} \right\|_2^2\right)^{1/2}, \tag{23}$$

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

$$\hat{Q}_{\text{rob}}(o, a, z, w) = \hat{\mu}_{\text{br}}(o, a, z, w) - \beta_{\text{br}}\, \hat{\sigma}_{\text{br}}(o, a, z, w), \qquad \beta_{\text{br}} \geq 0. \tag{24}$$

A single scalar $\hat{\sigma}_{\text{br}}$ thus fuses two distinct uncertainties, bootstrap (epistemic) variability and z-versus-w disagreement, so the discount in Eq. 24 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. 19.

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

$$\hat{\Phi} \in \operatorname*{arg\,min}_{\Phi \in \mathcal{F}_\Phi} \sum_{i=1}^{N} \big(G_i - \Phi(o_i)\big)^2 + \alpha_\Phi \|\Phi\|_2^2, \qquad \hat{\Phi}(o) \approx \mathbb{E}[\, G \mid o \,], \tag{25}$$

and subtracts it to obtain the state-normalized robust advantage

$$\hat{A}_{\text{rob}}(o, a, z, w) = \hat{Q}_{\text{rob}}(o, a, z, w) - \hat{\Phi}(o). \tag{26}$$

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,

$$\tilde{R} \in \operatorname*{arg\,min}_{R \in \mathcal{F}_R} \sum_{i=1}^{N} \big(\hat{\sigma}_{\mathrm{br}}(o_i, a_i, z_i, w_i) - R(o_i, a_i)\big)^2 + \alpha_R \|R\|_2^2, \qquad \hat{R}(o,a) = \big[\tilde{R}(o,a)\big]_+, \tag{27}$$

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,

$$\mathcal{N} = \Big\{ (o_i, a_{\pi(i)}),\ (o_i, a_i^{\mathrm{unif}}),\ \big(o_i, \mathrm{clip}(a_i + \epsilon_i, -1, 1)\big) \Big\}_{i=1}^{N}, \quad a_i^{\mathrm{unif}} \sim \mathcal{U}[-1,1]^{d_a},\ \ \epsilon_i \sim \mathcal{N}(0, \sigma_{\mathrm{neg}}^2 I), \tag{28}$$

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

$$\hat{S}(o,a) = -\log \mathrm{clip}\big(\hat{p}_{\mathrm{sup}}(o,a),\, \varepsilon_{\mathrm{sup}},\, 1 - \varepsilon_{\mathrm{sup}}\big). \tag{29}$$

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 $\mathcal{D}_{\mathrm{off}}$ with a variance floor before they are combined. Writing $T \in \{A, R, S\}$ for the streams $\hat{A}_{\mathrm{rob}}, \hat{R}, \hat{S}$,

$$\mu_T = \tfrac{1}{N} \sum_{i=1}^{N} T_i, \qquad \sigma_T = \max\Big\{ \big(\tfrac{1}{N} \sum_{i=1}^{N} (T_i - \mu_T)^2\big)^{1/2},\ \varepsilon_{\mathrm{std}} \Big\}. \tag{30}$$

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

$$s_{\mathrm{off}}(o,a,z,w) = \frac{1}{\tau_{\mathrm{off}}} \underbrace{\frac{\hat{A}_{\mathrm{rob}} - \mu_A}{\sigma_A}}_{\text{benefit}} - \lambda_R \underbrace{\frac{\hat{R} - \mu_R}{\sigma_R}}_{\text{instability}} - \lambda_S \underbrace{\frac{\hat{S} - \mu_S}{\sigma_S}}_{\text{implausibility}}, \tag{31}$$

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

$$\omega_{\mathrm{off}}(o,a,z,w) = \exp\Big[\mathrm{clip}\big(s_{\mathrm{off}}(o,a,z,w),\, -c_{\mathrm{off}},\, c_{\mathrm{off}}\big) \Big]. \tag{32}$$

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 $\varepsilon_{\mathrm{std}}$ keeps the standardization well-posed even when a stream is nearly constant on a given dataset. Appendix C.2 shows that Eq. 32 is not posited but *derived*, as the unique Gibbs tilt of a KL-anchored reliability program.

B.4 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,

$$\ell_{\mathrm{diff}}(\theta; o, a^{\mathrm{g}}) = \mathbb{E}_{t,\epsilon} \big\| \epsilon - \epsilon_\theta(o, a_t, t) \big\|_2^2. \tag{33}$$

Given a batch B with nonnegative guide weights $\{\omega_i\}_{i \in B}$, COOT first projects each weight onto a fixed leverage interval and then batch-normalizes to fix the gradient scale,

$$\tilde{\omega}_i = \mathrm{clip}\big(\omega_i, \omega_-, \omega_+\big), \qquad \bar{\omega}_i = \frac{\tilde{\omega}_i}{|B|^{-1} \sum_{j \in B} \tilde{\omega}_j + \varepsilon_\omega}, \tag{34}$$

with $0 < \omega_- \leq 1 \leq \omega_+$. The projection guards against a failure mode the exponential tilt alone cannot rule out: when a few samples receive weights orders of magnitude above the rest, the effective sample size of the weighted objective collapses and pretraining is driven by a handful of transitions. Bounding every weight into $[\omega_-, \omega_+]$ before normalization caps that concentration directly, while the mean normalization keeps the loss scale-free, and non-finite weights are mapped into the same interval so no single corrupted score can enter the update. COOT then minimizes the per-coordinate weighted objective

$$\mathcal{L}_{\text{diff}}(\theta) = \frac{1}{|B|\, d_a} \sum_{i \in B} \bar{\omega}_i \left\| \epsilon_i - \epsilon_\theta(o_i, a_{t_i,i}, t_i) \right\|_2^2, \tag{35}$$

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

$$\left(a_i^{\text{g}},\, \omega_i\right) = \begin{cases} \left(a_i,\ \omega_{\text{off}}(o_i, a_i, z_i, w_i)\right), & \text{offline (selective imitation)}, \\ \left(\arg\max\limits_{a \in \mathcal{C}(o_i)} \mathcal{J}_{\text{on}}(o_i, a),\ \ \omega_{\text{on}}(o_i)\right), & \text{online (value-guided improvement)}. \end{cases} \tag{36}$$

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.5 Potential-Shaped Critic Targets

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

$$\begin{aligned} \tilde{r}_t &= \rho\, r_t + \eta_t \operatorname{clip}\big(\rho\big(\gamma \hat{\Phi}(o_{t+1}) - \hat{\Phi}(o_t)\big), -c_\Phi, c_\Phi\big), \\ y_t &= \operatorname{clip}\Big(\tilde{r}_t + \gamma(1 - d_t) \min_{j \in \{1,2\}} Q_{\bar{\phi}_j}(o_{t+1}, a_{t+1}), -c_y, c_y\Big), \qquad a_{t+1} \sim \pi_{\bar{\theta}}(\cdot \mid o_{t+1}), \end{aligned} \tag{37}$$

and fits both critics by robust regression onto the shared target,

$$\mathcal{L}_Q(\phi_1, \phi_2) = \sum_{j=1}^{2} \mathbb{E}\big[\ell_H^{\delta_Q}\big(Q_{\phi_j}(o_t, a_t) - y_t\big)\big], \qquad \ell_H^{\delta}(u) = \begin{cases} \frac{1}{2} u^2, & |u| \leq \delta, \\ \delta\big(|u| - \frac{1}{2}\delta\big), & |u| > \delta. \end{cases} \tag{38}$$

Each element of Eq. 37 plays a distinct stabilizing role. Multiplying the potential difference by ρ keeps the shaping increment on the same scale as the scaled reward it augments, so a single coefficient η_t has a consistent meaning across environments with different return magnitudes. The clip c_Φ bounds the per-step shaping magnitude so a poorly extrapolated potential at a rarely visited state cannot inject an outsized pseudo-reward. The target clip c_y confines Bellman targets to a plausible return interval, preventing transient critic spikes from compounding through bootstrapping. The Huber loss makes the critic gradient bounded in the residual, which is the value-learning analogue of the bounded actor influence in Proposition 2: no single transition, however surprising its target, can dominate a critic update. The squared-error objective is recovered in the limit $\delta_Q \to \infty$, $c_y \to \infty$, $c_\Phi \to \infty$. All critic, policy, and temperature updates are additionally clipped by global gradient norm, and the entropy temperature is confined to a bounded interval $\log \alpha \in [\alpha_{\min}, \alpha_{\max}]$, so every update channel of the learner operates under an explicit influence budget (Appendix B.7).

Online, the shaping coefficient is annealed linearly to zero,

$$\eta_t = \eta_0 \big[\, 1 - t/T_{\text{shape}} \,\big]_+, \qquad [x]_+ = \max\{x, 0\}. \tag{39}$$

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. Potential shaping is policy-order preserving for any fixed bounded $\hat{\Phi}$ whenever the shaping clip is inactive, and the anneal makes the shaped and unshaped objectives coincide for all $t \geq T_{\text{shape}}$ even where the clip binds (Appendix C.5), so the schedule alters only the optimization path, never the fixed point it converges to.

B.6 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,

$$\mathcal{C}(o) = \underbrace{\{a_k^\theta \sim \pi_\theta(\cdot \mid o)\}_{k=1}^K}_{\mathcal{C}_\theta(o)} \cup \underbrace{\{a_m^{\text{prior}} \sim \pi_{\text{prior}}(\cdot \mid o)\}_{m=1}^M}_{\mathcal{C}_{\text{prior}}(o)}, \tag{40}$$

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 standardized by running statistics of recent critic values before the penalties are applied,

$$\hat{Q}^{\,\text{std}}(o,a) = \text{clip}\left(\frac{\min_j Q_{\phi_j}(o,a) - \hat{\mu}_Q}{\hat{\sigma}_Q}, -c_Q,\, c_Q\right), \qquad \mathcal{J}_{\text{on}}(o,a) = \hat{Q}^{\,\text{std}}(o,a) - \lambda_R^{\text{on}} \hat{R}_c(o,a) - \lambda_S^{\text{on}} \hat{S}_c(o,a), \tag{41}$$

where $\hat{R}_c = \text{clip}(\hat{R}, 0, c_P)$ and $\hat{S}_c = \text{clip}(\hat{S}, 0, c_P)$ bound the reliability penalties, non-finite predictions are mapped into the same intervals, and the standardization statistics are exponential moving averages of the batch mean and dispersion of critic values,

$$\hat{\mu}_Q \leftarrow (1-\kappa_Q)\,\hat{\mu}_Q + \kappa_Q\, \widehat{\text{mean}}_B(Q), \qquad \hat{\sigma}_Q \leftarrow (1-\kappa_Q)\,\hat{\sigma}_Q + \kappa_Q \max\{\widehat{\text{std}}_B(Q), \varepsilon_\sigma\}, \tag{42}$$

updated once per learner step with rate κ_Q and dispersion floor $\varepsilon_\sigma > 0$. Three design intentions are encoded here. First, standardization expresses online evidence in units of the critic's own dispersion, so that $(\lambda_R^{\text{on}}, \lambda_S^{\text{on}})$ keep a fixed meaning as the raw critic scale drifts over training, rather than being silently outgrown by it. Second, the statistics are slow global averages rather than per-state quantities, so at any fixed observation the map from $\min_j Q_{\phi_j}$ to $\hat{Q}^{\,\text{std}}$ is a positive affine transformation whose scale is common to all candidates, which is exactly the property the improvement guarantee of Appendix C.7 exploits. Third, the clips c_Q and c_P cap the leverage of any single candidate's score, extending the bounded-influence principle from parameter updates to action selection: no overconfident critic spike and no runaway penalty prediction can single-handedly decide which action is executed. Selection then occurs at two points that must not be conflated. The agent *executes*

$$a_t \in \arg\max_{a \in \mathcal{C}(o_t)} \mathcal{J}_{\text{on}}(o_t, a), \tag{43}$$

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

$$a_i^{\text{g}} \in \arg\max_{a \in \mathcal{C}(o_i)} \mathcal{J}_{\text{on}}(o_i, a), \qquad \omega_{\text{on}}(o_i) = \exp\Big[\text{clip}\Big(\tfrac{1}{\tau_{\text{act}}}\, \mathcal{J}_{\text{on}}(o_i, a_i^{\text{g}}),\, -c_{\text{on}},\, c_{\text{on}}\Big)\Big]. \tag{44}$$

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.

B.7 Bounded-Update Robustification

The safeguards introduced across Appendices B.4 to B.6 are instances of a single design principle that this subsection states once, in one place. COOT transfers from biased data, so every channel through which the offline signal can act on the learner carries an explicit influence budget: no sample, transition, candidate, or update step may exert unbounded leverage. Proposition 2 establishes this budget for the actor's imitation weights, and the same budget governs the remaining channels, each safeguard closing a specific pathway through which a single erroneous signal could otherwise dominate a stochastic training trajectory.

The complete set of safeguards, with their default constants, is as follows. *Actor weights*: offline and online guide weights are projected onto $[\omega_-, \omega_+] = [0.25, 4.0]$ and mean-normalized (Eq. 34), bounding relative

leverage by $\omega_+/\omega_- = 16$ regardless of how extreme the raw reliability scores are. *Critic regression*: targets are clipped to $[-c_y, c_y]$ with $c_y = 250$ and fit with the Huber loss at threshold $\delta_Q = 10$ (Eq. 38), so the critic gradient is bounded in the residual. *Shaping*: the potential increment is reward-scale-consistent and clipped to $[-c_\Phi, c_\Phi]$ with $c_\Phi = 2$ (Eq. 37), so a poorly extrapolated potential cannot inject an outsized pseudo-reward at any single step. *Candidate scoring*: the standardized critic score is clipped to $[-c_Q, c_Q]$ with $c_Q = 8$, and the risk and support fields to $[0, c_P]$ with $c_P = 20$ (Eq. 41), so no critic spike or runaway penalty prediction decides an action alone. *Optimization*: critic, policy, and temperature gradients are clipped by global norm at 10, 5, and 1 respectively, and the entropy temperature is confined to $\log \alpha \in [-8, 1]$. *Numerical hygiene*: sampled and selected actions are projected onto the action box, and non-finite weights or penalty predictions are mapped into their admissible intervals before use. Environment-specific constants follow the same template.

Two properties of this layer deserve emphasis. First, it is conservative by construction: every safeguard is a projection or a robust loss, each of which acts only where the corresponding bound binds and coincides with the unguarded operation everywhere else, so on well-behaved batches the safeguards are inert and the objectives of Appendices B.4 to B.6 are optimized exactly. Second, the theory tracks the safeguards explicitly rather than idealizing them away: the bounded-influence guarantee of Appendix C.4 incorporates the weight projection, the shaping analysis of Appendix C.5 bounds the distortion of the clipped increment, and the improvement guarantee of Appendix C.7 is stated directly for the standardized, clipped selection rule as implemented, carrying its clipping slack as an explicit term. The safeguards are therefore the completion of the bounded-influence principle that motivates COOT's transfer weights in the first place: a method built on distrusting individual offline samples should not, at any other point in its pipeline, allow an individual sample, target, or candidate to act without bound.

C Proofs and Derivations

This appendix formalizes the 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. Four properties govern the construction. *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*: weight projection and batch normalization cap the gradient influence of any single sample (Sec. C.4). *Invariance*: bounded potential shaping preserves policy ordering exactly wherever its truncation is inactive and distorts return differences by at most an explicitly bounded, annealed amount (Sec. C.5). Two further results state what these properties buy in performance terms. *Transfer*: the gap between the transferred value and the interventional value splits into a measurable dispersion term and an irreducible residual-confounding term (Sec. C.6). *Improvement*: reliability-penalized candidate selection yields conditional policy improvement whenever a reliable candidate's advantage clears its penalty, the critic error, and the score-clipping slack (Sec. C.7). Conservatism makes the transferred value trustworthy, selectivity turns it into a principled imitation emphasis, stability keeps the optimization well-behaved, invariance ensures the dense critic signal accelerates rather than redirects online learning, and the final two results connect the pipeline to transfer accuracy and online improvement.

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

$$Q^\star_{\text{prox}}(o, a, z, w) := \mathbb{E}[\, G \mid o, a, z, w \,], \tag{45}$$

the best return predictor measurable from the full logged tuple. Eq. 45 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

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

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

$$\hat{\mu}_{\text{br}} = \tfrac{1}{2B}\mathbf{1}^\top H, \qquad \hat{\sigma}_{\text{br}} = \Big(\tfrac{1}{2B}\big\|H - \hat{\mu}_{\text{br}}\mathbf{1}\big\|_2^2\Big)^{1/2}. \tag{46}$$

COOT transfers the lower-confidence summary

$$\hat{Q}_{\text{rob}}(o, a, z, w) = \hat{\mu}_{\text{br}}(o, a, z, w) - \beta_{\text{br}}\,\hat{\sigma}_{\text{br}}(o, a, z, w), \qquad \beta_{\text{br}} \geq 0, \tag{47}$$

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,*

$$\big|\, Q^\star_{\text{prox}}(o, a, z, w) - \hat{\mu}_{\text{br}}(o, a, z, w) \,\big| \leq \beta_{\text{br}}\,\hat{\sigma}_{\text{br}}(o, a, z, w). \tag{48}$$

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. 48 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*

$$\hat{Q}_{\text{rob}}(o, a, z, w) \leq Q^\star_{\text{prox}}(o, a, z, w), \tag{49}$$

and the conservatism gap is controlled by the same dispersion,

$$0 \leq Q^\star_{\text{prox}}(o, a, z, w) - \hat{Q}_{\text{rob}}(o, a, z, w) \leq 2\beta_{\text{br}}\,\hat{\sigma}_{\text{br}}(o, a, z, w). \tag{50}$$

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

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

which is Eq. equation 49 and the lower half of Eq. equation 50. For the upper half, the other side of calibration gives $Q^\star_{\text{prox}} - \hat{\mu}_{\text{br}} \leq \beta_{\text{br}}\hat{\sigma}_{\text{br}}$. Adding the identity $\hat{\mu}_{\text{br}} - \hat{Q}_{\text{rob}} = \beta_{\text{br}}\hat{\sigma}_{\text{br}}$ yields

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

Two consequences explain design choices in the main text rather than merely restating them. First, the gap bound $2\beta_{\text{br}}\hat{\sigma}_{\text{br}}$ makes conservatism *self-limiting*: where the proxies pin down the regime, dispersion is small and $\hat{Q}_{\text{rob}}$ 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 ($\beta_{\text{br}} = 0$) 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,

$$\bar{A}_{\text{rob}} = \frac{\hat{A}_{\text{rob}} - \mu_A}{\sigma_A}, \qquad \bar{R} = \frac{\hat{R} - \mu_R}{\sigma_R}, \qquad \bar{S} = \frac{\hat{S} - \mu_S}{\sigma_S}, \qquad \hat{A}_{\text{rob}} = \hat{Q}_{\text{rob}} - \hat{\Phi}, \tag{51}$$

and define the per-sample reliability utility

$$u(o, a, z, w) = \tau_{\text{off}}^{-1} \bar{A}_{\text{rob}}(o, a, z, w) - \lambda_R \bar{R}(o, a) - \lambda_S \bar{S}(o, a). \tag{52}$$

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:

$$q^\star \in \operatorname*{arg\,max}_{q \ll p_0} \Big\{ \mathbb{E}_q[u(o, a, z, w)] - \mathrm{KL}(q \,\|\, p_0) \Big\}. \tag{53}$$

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

$$q^\star(o, a, z, w) = \frac{p_0(o, a, z, w) \, \exp\big(u(o, a, z, w)\big)}{\mathbb{E}_{p_0}\big[\exp(u(o, a, z, w))\big]}. \tag{54}$$

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

$$s_{\text{off}}(o, a, z, w) = \tau_{\text{off}}^{-1} \bar{A}_{\text{rob}}(o, a, z, w) - \lambda_R \bar{R}(o, a) - \lambda_S \bar{S}(o, a), \tag{55}$$

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

$$\omega_{\text{off}}(o, a, z, w) = \exp\Big[\operatorname{clip}\big(s_{\text{off}}(o, a, z, w), \, -c_{\text{off}}, \, c_{\text{off}}\big)\Big]. \tag{56}$$

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 56 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*,

$$\frac{\omega_{\text{off},i}}{\omega_{\text{off},j}} = \exp(s_i - s_j), \tag{57}$$

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

$$e^{-c_{\text{off}}} \le \omega_{\text{off}}(o, a, z, w) \le e^{c_{\text{off}}}. \tag{58}$$

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

$$\frac{\partial s_{\text{off}}}{\partial \bar{A}_{\text{rob}}} = \tau_{\text{off}}^{-1} > 0, \qquad \frac{\partial s_{\text{off}}}{\partial \bar{R}} = -\lambda_R \le 0, \qquad \frac{\partial s_{\text{off}}}{\partial \bar{S}} = -\lambda_S \le 0, \tag{59}$$

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 projection and batch normalization

$$\tilde{\omega}_{\text{off},i} = \text{clip}\big(\omega_{\text{off},i}, \omega_-, \omega_+\big), \qquad \bar{\omega}_{\text{off},i} = \frac{\tilde{\omega}_{\text{off},i}}{|B|^{-1}\sum_{j\in B}\tilde{\omega}_{\text{off},j} + \varepsilon_\omega} \tag{60}$$

preserve within-batch ordering while making the loss scale-free, and the next result shows they also cap every sample's contribution.

Proof. By Eq. equation 58 every raw weight obeys $e^{-c_{\text{off}}} \le \omega_{\text{off},i} \le e^{c_{\text{off}}}$. Assume the clip intervals intersect, $\omega_- \le e^{c_{\text{off}}}$ and $e^{-c_{\text{off}}} \le \omega_+$, which holds in every practical configuration since the projection is meant to tighten the exponential envelope rather than exclude it. Then the projection confines every weight to the intersection,

$$\underline{\omega} \le \tilde{\omega}_{\text{off},i} \le \overline{\omega}, \qquad \underline{\omega} = \max\{e^{-c_{\text{off}}}, \omega_-\}, \quad \overline{\omega} = \min\{e^{c_{\text{off}}}, \omega_+\}, \qquad i \in B, \tag{61}$$

so the batch mean satisfies $|B|^{-1}\sum_{j\in B}\tilde{\omega}_{\text{off},j} \ge \underline{\omega}$. Since $\varepsilon_\omega \ge 0$, the denominator of Eq. equation 60 is at least $\underline{\omega}$, and combined with the numerator bound $\tilde{\omega}_{\text{off},i} \le \overline{\omega}$ this gives the composition-free envelope

$$\bar{\omega}_{\text{off},i} \le \frac{\overline{\omega}}{\underline{\omega}} \le \min\left\{e^{2c_{\text{off}}}, \frac{\omega_+}{\omega_-}\right\} = \kappa_\omega. \tag{62}$$

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

$$0 \le \bar{\omega}_{\text{off},i}\, \ell_{\text{diff}}(\theta; o_i, a_i) \le \kappa_\omega L_{\max}. \tag{63}$$

□

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 $\kappa_\omega L_{\max}$ to the update. The two clips play complementary roles made explicit by Propositions 1 and 2: the log-space clip c_{off} caps relative emphasis on the tilt scale, the weight-space projection $[\omega_-, \omega_+]$ caps absolute leverage directly, and whichever is tighter governs the envelope, so selectivity and optimization stability are controlled by two interpretable constants rather than traded off against each other. The projection also confers robustness the exponential bound alone cannot: any non-finite weight produced by a corrupted score is mapped into $[\omega_-, \omega_+]$ before entering the update, so the envelope holds unconditionally.

C.5 Exact and Approximate Policy-Order Preservation of Bounded 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 implementation uses a reward-scale-consistent potential difference and clips it before adding it to the critic target (Eq. 37). The following proposition therefore separates the exact potential-shaping identity from the bounded perturbation introduced by that clip, so that what is preserved exactly and what is distorted by how much are both stated rather than assumed.

Proposition 4 (Potential shaping with bounded truncation)**.** *Let $\Phi : \mathcal{O} \to \mathbb{R}$ be fixed and bounded, $\gamma \in [0, 1)$, reward scale $\rho > 0$, and fixed shaping coefficient $\eta \ge 0$. Define*

$$F_t = \rho\big[\gamma\Phi(o_{t+1}) - \Phi(o_t)\big], \qquad \bar{F}_t = \text{clip}(F_t, -c_\Phi, c_\Phi), \qquad e_t = \bar{F}_t - F_t. \tag{64}$$

For the untruncated shaped reward $r_t^F = \rho\, r_t + \eta F_t$, every trajectory satisfies

$$G_0^F = \rho\, G_0 - \eta\rho\, \Phi(o_0). \tag{65}$$

For the implemented reward $\bar{r}_t^F = \rho\, r_t + \eta \bar{F}_t$,

$$\left|\bar{G}_0^F - G_0^F\right| \le \frac{\eta\, \epsilon_\Phi}{1-\gamma}, \qquad \epsilon_\Phi := \sup_t |e_t|. \tag{66}$$

Consequently, for any two policies π_1, π_2 *from the same initial observation, the clipping-induced error in their shaped return difference is at most* $2\eta\epsilon_\Phi/(1-\gamma)$. *If clipping is inactive along the trajectories,* $\epsilon_\Phi = 0$ *and policy ordering is exactly preserved.*

Proof. For the untruncated return,

$$\begin{aligned} G_0^F &= \sum_{t=0}^{\infty} \gamma^t \big[\rho\, r_t + \eta\rho\big(\gamma\Phi(o_{t+1}) - \Phi(o_t)\big)\big] \\ &= \rho\, G_0 + \eta\rho \sum_{t=0}^{\infty} \left[\gamma^{t+1}\Phi(o_{t+1}) - \gamma^t \Phi(o_t)\right] = \rho\, G_0 - \eta\rho\, \Phi(o_0), \end{aligned} \tag{67}$$

where boundedness of Φ and $\gamma < 1$ give absolute convergence of both potential series, so the telescoping rearrangement is valid and all terms cancel except the initial one. For a fixed o_0 the offset $-\eta\rho\,\Phi(o_0)$ is independent of the action sequence, hence of the policy, and cannot change the argmax over policies. For the clipped version,

$$\bar{G}_0^F - G_0^F = \eta \sum_{t=0}^{\infty} \gamma^t e_t, \tag{68}$$

so the triangle inequality and $|e_t| \le \epsilon_\Phi$ give

$$\left|\bar{G}_0^F - G_0^F\right| \le \eta \sum_{t=0}^{\infty} \gamma^t \epsilon_\Phi = \frac{\eta\, \epsilon_\Phi}{1-\gamma}. \tag{69}$$

Applying this bound once to each policy proves the return-difference statement. □

The proposition licenses the annealing schedule of Appendix B.5 in two regimes at once. Wherever the potential proposes increments within c_Φ, which is its intended operating regime, shaping changes returns only by a policy-independent constant and the fixed point is untouched, so COOT can inject a strong dense signal from $\hat{\Phi}$ when on-policy returns are scarce. Where the clip binds, the distortion of any policy comparison is at most $2\eta_t\epsilon_\Phi/(1-\gamma)$, a controlled trust region that exists precisely to stop an unusually large increment from a poorly extrapolated potential, and the schedule of Eq. 39 drives this perturbation to zero since $\eta_t = 0$ for all $t \ge T_{\text{shape}}$. Target clipping is a separate numerical safeguard and is absorbed into the critic-calibration error of Theorem 2, and no claim is made that a persistently saturated TD target has the same fixed point as the unbounded Bellman operator.

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. The remaining two subsections state what this well-behaved pipeline buys in performance terms: how far residual confounding can distort the value it transfers, and when its selection rule converts what was transferred into policy improvement.

C.6 Transfer Error Under Residual Confounding

The first question any transferred value must answer is how far it can sit from the interventional value that online improvement actually requires. Recall the proxy-conditioned target of Eq. 45,

$$Q^\star_{\text{prox}}(o, a, z, w) = \mathbb{E}[G \mid o, a, z, w]\,, \tag{70}$$

and define the interventional value

$$Q_{\text{int}}(o, a) = \mathbb{E}[G \mid o, \text{do}(a)]\,. \tag{71}$$

Under the confounding model of Eq. 1, $Q^\star_{\text{prox}}$ absorbs the component of the latent regime revealed by (z, w) but in general differs from Q_{int} by the component the proxies do not reveal. We quantify that residual with a single scalar, introduced as a property of the data-generating process rather than of any estimator.

Definition 2 (Residual confounding radius)**.** *The proxies are ϵ_{conf}-informative at (o, a, z, w) if*

$$\left| Q_{\text{int}}(o, a) - Q^\star_{\text{prox}}(o, a, z, w) \right| \;\le\; \epsilon_{\text{conf}}. \tag{72}$$

The radius ϵ_{conf} is small when the proxies carry most of the regime information relevant to the return, and large when they are weak, spurious, or misaligned. It is exactly the quantity manipulated by the proxy stress tests of Appendix D.4, which makes the following bound empirically testable rather than purely formal.

Theorem 1 (Transfer error decomposes into a measurable part and an irreducible part)**.** *Fix a query (o, a, z, w). If the bridge ensemble is β_{br}-calibrated at the query in the sense of Definition 1 and the proxies are ϵ_{conf}-informative there in the sense of Definition 2, then*

$$\hat{Q}_{\text{rob}} - \epsilon_{\text{conf}} \;\le\; Q_{\text{int}} \;\le\; \hat{Q}_{\text{rob}} + 2\beta_{\text{br}}\hat{\sigma}_{\text{br}} + \epsilon_{\text{conf}}, \tag{73}$$

and in particular

$$\left| Q_{\text{int}} - \hat{Q}_{\text{rob}} \right| \;\le\; 2\beta_{\text{br}}\hat{\sigma}_{\text{br}} + \epsilon_{\text{conf}}, \tag{74}$$

where all quantities are evaluated at the query.

Proof. Proposition 3 gives the two-sided conservatism bound

$$0 \;\le\; Q^\star_{\text{prox}} - \hat{Q}_{\text{rob}} \;\le\; 2\beta_{\text{br}}\hat{\sigma}_{\text{br}}. \tag{75}$$

For the lower bound in Eq. 73, Definition 2 gives $Q_{\text{int}} \ge Q^\star_{\text{prox}} - \epsilon_{\text{conf}}$, and the left inequality of Eq. 75 gives $Q^\star_{\text{prox}} \ge \hat{Q}_{\text{rob}}$, so

$$Q_{\text{int}} \;\ge\; Q^\star_{\text{prox}} - \epsilon_{\text{conf}} \;\ge\; \hat{Q}_{\text{rob}} - \epsilon_{\text{conf}}. \tag{76}$$

For the upper bound, Definition 2 gives $Q_{\text{int}} \le Q^\star_{\text{prox}} + \epsilon_{\text{conf}}$, and the right inequality of Eq. 75 gives $Q^\star_{\text{prox}} \le \hat{Q}_{\text{rob}} + 2\beta_{\text{br}}\hat{\sigma}_{\text{br}}$, so

$$Q_{\text{int}} \;\le\; Q^\star_{\text{prox}} + \epsilon_{\text{conf}} \;\le\; \hat{Q}_{\text{rob}} + 2\beta_{\text{br}}\hat{\sigma}_{\text{br}} + \epsilon_{\text{conf}}. \tag{77}$$

Combining the two chains yields Eq. 73, and Eq. 74 follows since both deviations of Q_{int} from $\hat{Q}_{\text{rob}}$ are bounded by $2\beta_{\text{br}}\hat{\sigma}_{\text{br}} + \epsilon_{\text{conf}}$. □

Theorem 1 assigns each error term an owner. The width $2\beta_{\text{br}}\hat{\sigma}_{\text{br}}$ is observable at fit time, is controlled by the bridge dispersion, and shrinks with more data and more agreeing proxy views. The radius ϵ_{conf} is irreducible from the logs alone and shrinks only with more informative proxies. This is the precise sense in which COOT reads what to distrust: it cannot erase ϵ_{conf}, but its pessimism is anchored to the one term it can measure, so it never over-trusts the transferred value. The decomposition also predicts the empirical pattern of Appendix D.4: attenuating, dropping, or misaligning the proxies inflates ϵ_{conf}, so performance should degrade in an ordered, monotone way rather than collapse, which is exactly the behavior reported in Tables 5 and 7.

C.7 Conditional Policy Improvement of Reliability-Penalized Selection

The second question is why the online rule of Appendix B.6 yields improvement rather than merely a different initialization. We analyze the standardized-and-clipped rule exactly as implemented, carrying its clipping slack as an explicit term rather than assuming it away. We condition on the diffusion-sampling randomness, so each realized candidate set is fixed, and the statement for the stochastic policy follows by averaging when the stated conditions hold conditionally. Fix a learner step and observation o, write π for the current policy, and define

$$\hat{q}(o, a) = \min_j Q_{\phi_j}(o, a), \qquad \hat{z}(o, a) = \text{clip}\left(\frac{\hat{q}(o, a) - \hat{\mu}_Q}{\hat{\sigma}_Q}, -c_Q, c_Q \right), \tag{78}$$

with the running statistics of Eq. 42 satisfying $\hat{\sigma}_Q \geq \varepsilon_\sigma > 0$. Two derived objects translate the dimensionless score back into the units in which critic error and policy advantage are measured, which is the deliberate step that lets the analysis speak about returns rather than scores,

$$\tilde{q}(o,a) = \hat{\mu}_Q + \hat{\sigma}_Q\, \hat{z}(o,a), \qquad P(o,a) = \hat{\sigma}_Q \big(\lambda_R^{\text{on}} \hat{R}_c(o,a) + \lambda_S^{\text{on}} \hat{S}_c(o,a)\big) \;\geq\; 0, \tag{79}$$

the reconstructed clipped critic and the reliability price in critic units. The clipping distortion at o is

$$\epsilon_{\text{clip}}(o) = \sup_{a \in \mathcal{C}(o)} \big|\tilde{q}(o,a) - \hat{q}(o,a)\big|, \tag{80}$$

which is zero whenever all candidate critic values lie within c_Q running standard deviations of the running mean, the intended operating regime of the score clip.

Lemma 1 (Exact reduction of the standardized selection rule)**.** *At the fixed step and observation, the executed rule of Eqs. 41 and 43 satisfies*

$$\hat{a}(o) \in \arg\max_{a \in \mathcal{C}(o)} \big[\tilde{q}(o,a) - P(o,a)\big]. \tag{81}$$

Proof. The statistics $\hat{\mu}_Q$ and $\hat{\sigma}_Q$ are computed once per learner step from replay batches, so over $a \in \mathcal{C}(o)$ they are constants with $\hat{\sigma}_Q > 0$. Multiplying $\mathcal{J}_{\text{on}} = \hat{z} - \lambda_R^{\text{on}} \hat{R}_c - \lambda_S^{\text{on}} \hat{S}_c$ by the positive constant $\hat{\sigma}_Q$ and adding the action-independent constant $\hat{\mu}_Q$ gives exactly $\tilde{q} - P$, so the maximizer is unchanged. Nonnegativity of P follows from $\hat{\sigma}_Q > 0$ and $\hat{R}_c, \hat{S}_c \geq 0$ by the clipping in Eq. 41. □

The lemma is unconditional because it maximizes the reconstructed clipped critic, and the price of that exactness is the distortion term ϵ_{clip}, which reenters the guarantee below instead of being hidden inside an assumption. One further bookkeeping choice keeps the statement faithful to the implementation. The critics are trained on the scaled reward $r^\rho = \rho\, r$ of Eq. 37, so let Q_ρ^π, V_ρ^π, and A_ρ^π denote values under r^ρ. Positive reward scaling preserves policy ordering, and unscaled return differences are recovered by dividing by ρ. Critic calibration below is with respect to Q_ρ^π and therefore includes function-approximation error, residual early shaping, and any target-clipping bias, so the theorem does not pretend the critic tracks an idealized unscaled value.

Theorem 2 (Conditional policy improvement of reliability-penalized selection)**.** *Let π_{new} denote the policy that executes $\hat{a}(o)$ of Lemma 1 at every observation. Assume $|r^\rho| \leq R_{\max}^\rho$, so that $|A_\rho^\pi| \leq A_{\max} = \frac{2R_{\max}^\rho}{1-\gamma}$, and let $\mathcal{G} \subseteq \mathcal{O}$ be the set of observations at which the following two conditions hold:*

(i) Calibrated critic on candidates. $\big|\hat{q}(o,a) - Q_\rho^\pi(o,a)\big| \leq \epsilon_{\text{crit}}$ *for all* $a \in \mathcal{C}(o)$*, and* $\epsilon_{\text{clip}}(o) \leq \epsilon_{\text{clip}}$.

(ii) Reliable advantageous candidate. *There exists* $a^+ \in \mathcal{C}(o)$ *with* $A_\rho^\pi(o,a^+) \geq \Delta$ *and* $P(o,a^+) \leq p$.

Then for every $o \in \mathcal{G}$,

$$A_\rho^\pi\big(o, \hat{a}(o)\big) \;\geq\; L_{\mathcal{G}} := \Delta - p - 2\epsilon_{\text{crit}} - 2\epsilon_{\text{clip}}, \tag{82}$$

and if $L_{\mathcal{G}} \geq -A_{\max}$ and the normalized discounted state distribution of π_{new} satisfies $d^{\pi_{\text{new}}}(\mathcal{G}) \geq 1 - \delta$, then

$$J_\rho(\pi_{\text{new}}) - J_\rho(\pi) \;\geq\; \frac{(1-\delta)\, L_{\mathcal{G}} - \delta A_{\max}}{1-\gamma}. \tag{83}$$

In particular, the right side is positive whenever the reliable-candidate margin Δ exceeds the combined reliability price, critic error, clipping slack, and bad-state mass.

Proof. Fix $o \in \mathcal{G}$, abbreviate $\hat{a} = \hat{a}(o)$, and take any $a \in \mathcal{C}(o)$. Lemma 1 gives

$$\tilde{q}(o,\hat{a}) - P(o,\hat{a}) \;\geq\; \tilde{q}(o,a) - P(o,a). \tag{84}$$

Applying critic calibration to $\hat{a}$, the distortion bound, $P(o, \hat{a}) \geq 0$, Eq. 84, the distortion bound again, and critic calibration to a, in sequence,

$$\begin{aligned} Q^{\pi}_{\rho}(o, \hat{a}) &\geq \hat{q}(o, \hat{a}) - \epsilon_{\text{crit}} \geq \tilde{q}(o, \hat{a}) - \epsilon_{\text{clip}} - \epsilon_{\text{crit}} \geq \tilde{q}(o, \hat{a}) - P(o, \hat{a}) - \epsilon_{\text{clip}} - \epsilon_{\text{crit}} \\ &\geq \tilde{q}(o, a) - P(o, a) - \epsilon_{\text{clip}} - \epsilon_{\text{crit}} \geq \hat{q}(o, a) - P(o, a) - 2\epsilon_{\text{clip}} - \epsilon_{\text{crit}} \\ &\geq Q^{\pi}_{\rho}(o, a) - P(o, a) - 2\epsilon_{\text{clip}} - 2\epsilon_{\text{crit}}. \end{aligned} \tag{85}$$

Instantiating $a = a^{+}$ from condition (ii), using $P(o, a^{+}) \leq p$, and subtracting $V^{\pi}_{\rho}(o)$ from both sides yields Eq. 82. For Eq. 83, the performance-difference lemma (Kakade & Langford, 2002) gives

$$J_{\rho}(\pi_{\text{new}}) - J_{\rho}(\pi) = \frac{1}{1-\gamma} \mathbb{E}_{o \sim d^{\pi_{\text{new}}}} \Big[A^{\pi}_{\rho}\big(o, \pi_{\text{new}}(o)\big) \Big]. \tag{86}$$

On $\mathcal{G}$ the integrand is at least $L_{\mathcal{G}}$ by Eq. 82, and on the complement it is at least $-A_{\max}$ by bounded rewards. Writing $g = d^{\pi_{\text{new}}}(\mathcal{G})$, the split expectation is at least $g\, L_{\mathcal{G}} - (1-g) A_{\max}$, which is non-decreasing in g because $L_{\mathcal{G}} + A_{\max} \geq 0$, so $g \geq 1 - \delta$ proves Eq. 83. □

The theorem names the quantities each algorithmic component is designed to control, and this mapping is the result that separates COOT from a warm start. A warm start only moves the initialization, whereas condition (ii) must hold throughout training, which is exactly what the persistent prior maintains by keeping reliable candidates inside $\mathcal{C}(o)$ at every online step through Eq. 40, raising the margin Δ. Risk and support screening keep the price p small for those candidates by construction of $\hat{R}$ and $\hat{S}$. Huber regression, bounded targets, and training on executed transitions contain ϵ_{crit} and drive it down as experience accumulates. The running standardization makes the price commensurate with critic scale, while the score clip contributes the explicit ϵ_{clip} term rather than being hidden inside the proof. The bound then reads as a timeline. Early in training ϵ_{crit} is large, so improvement requires a wide advantage margin and the penalty is what prevents acting on critic noise. Later ϵ_{crit} shrinks and genuine advantages clear the threshold $\Delta > p + 2\epsilon_{\text{crit}} + 2\epsilon_{\text{clip}}$. This matches the empirical pattern that gains concentrate early and that removing the persistent prior produces the largest AUC degradation in Table 5. We emphasize the conditional scope. The theorem does not assert convergence of the full nonlinear diffusion actor-critic loop, and conditions (i) and (ii) are assumptions whose plausibility is supported empirically rather than proved. Its role is to name the quantities that govern online improvement, Δ, p, ϵ_{crit}, ϵ_{clip}, and δ, and to tie each to the component of the algorithm responsible for it.

Read together with Theorem 1, the two results complete the pipeline view of this appendix. The transfer theorem bounds how far residual confounding can distort what the bundle carries across the offline-online boundary, and the improvement theorem states when the online selection rule converts what was carried into policy gains. The earlier results now serve these two as supporting structure: conservatism supplies the calibrated value entering the decomposition, the exponential tilt and bounded influence keep pretraining faithful to it, and shaping invariance guarantees the dense critic signal accelerates learning without altering the target that Theorem 2 evaluates.

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 hyper-parameter 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 $N = 300$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. Crucially, it is not a frozen behavior-cloning policy: cloning is used only for offline initialization, after which the control runs the identical online diffusion actor-critic algorithm as COOT, on the same online budget, differing solely in that the reliability weights, risk and support penalties, potential shaping, and persistent-prior candidates are disabled (equivalently $\omega_i \equiv 1$, $\lambda_R^{\text{on}} = \lambda_S^{\text{on}} = 0$, and $M = 0$). 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

Tables 3 and 4 report 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 tables separate 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 (Table 3). 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 (Table 4). 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.

Table 3: Default COOT hyperparameters, grouped by algorithmic role: proxy-informed reliability transfer bundle and diffusion actor-critic architecture.

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.

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

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

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

$$\mu_k(o_t, u_t) = \tanh(f_k(o_t) + v_k u_t + b_k)\,, \qquad a_t = \text{clip}\,(\mu_k(o_t, u_t) + \epsilon_t, -1, 1)\,, \qquad \epsilon_t \sim \mathcal{N}(0, \sigma_k^2 I), \tag{87}$$

where f_k is a task-scaled linear observation map. Here $\mu_k(o_t, u_t)$ is the deterministic mean action of the k-th logging policy, the center of its action distribution before exploration noise. The term $v_k u_t$ makes this mean depend on the hidden regime, which is precisely the confounding path $u_t \to a_t^{\log}$ of Eq. 1. The Gaussian perturbation ϵ_t then makes the logging policy stochastic rather than deterministic: it gives the offline data local action coverage around μ_k, induces a well-defined behavior-policy density and log-probability ℓ_i, and matches realistic noisy collection, while the clip keeps actions in the valid range $[-1, 1]$. 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

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

Table 4: Default COOT hyperparameters, grouped by algorithmic role (continued): offline pretraining and online prior-guided learning.

Hyperparameter	**Value**	**Role in COOT**
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	Rescales rewards and the potential difference so shaping stays commensurate with critic targets.
Guide-weight interval $[\omega_-, \omega_+]$	$[0.25, 4.0]$	Projects offline and online guide weights before mean normalization, capping any sample's leverage (Eq. 34).
Potential-increment cap c_Φ	2.0	Bounds the reward-scale-consistent shaping increment before the annealed coefficient is applied (Eq. 37).
Huber threshold δ_Q	10.0	Makes critic regression quadratic near the target and linear for large residuals (Eq. 38).
TD-target cap c_y	250	Confines Bellman targets to a plausible return interval so transient spikes cannot compound through bootstrapping.
Gradient-norm caps (critic, policy, temperature)	10, 5, 1	Bounds each optimizer step by global norm under heavy-tailed targets.
Log-temperature interval	$[-8, 1]$	Confines the entropy temperature, preventing collapse or explosive diffusion noise.
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.
Running-statistic update rate κ_Q	0.01	Updates the running mean and dispersion that place critic evidence on a stable dimensionless scale (Eq. 42).
Dispersion floor ε_σ	10^{-3}	Keeps the standardization scale strictly positive.
Standardized-score cap c_Q	8.0	Limits the influence of transient critic-scale errors in candidate ranking (Eq. 41).
Risk/support cap c_P	20.0	Bounds the two reliability penalties and replaces non-finite predictions before scoring.
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.

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

$$G_i = r_i + \gamma(1 - d_i)G_{i+1}, \tag{88}$$

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.

Mapping MuJoCo quantities to the Section 2 variables. The correspondence between simulator quantities and the notation of Eq. 1 is direct, and clarifies in particular that measured signals map to observables while u_t stays hidden. The proprioceptive simulator state, joint angles, joint velocities, and body pose, is the observation o_t. The commanded control written to the log is $a_t^{\log}$, the realized reward and next state are r_t and o_{t+1}, and the auxiliary telemetry generated by the confounded environment, playing the role of quantities such as motor current, torque residuals, or temperature, is the proxy pair (z_t, w_t). The latent regime u_t is the *unmeasured* physical condition behind these signals, standing in for factors such as actuator wear, friction change, payload variation, or calibration drift. A worn motor illustrates the distinction: the wear level itself is never a sensor reading, so it is u_t, whereas its measurable symptoms, elevated current or torque error, are the proxies (z_t, w_t), and the ordinary robot state remains o_t. Real measurements therefore populate o_t, z_t, w_t, and u_t is the hidden cause behind them, never observed by any method.

Observations are clean simulator states, identical across all methods. In every MuJoCo experiment o_t is the standard observation returned by the simulator, and we add no sensor model and no observation-noise model. The uncertainty COOT reasons about does not originate from noisy sensing of o_t. It originates from the hidden execution regime u_t, which can shift the executed action and the resulting reward and next state relative to the logged action, exactly as in Eq. 1. For example, with o_t holding joint angles and velocities, a commanded torque $a_t^{\log} = 0.8$ may be realized as $a_t^{\text{exec}} \approx 0.6$ under a weak-actuator regime, and the proxies carry only partial clues about it. All eight online baselines and COOT receive the *same* clean o_t: no method is given a cleaner or noisier observation, and none is given u_t. The only additional information COOT consumes is the proxy telemetry (z_t, w_t), which is the proposed mechanism rather than a privileged observation channel. The eight baselines do not use proxies because their published formulations condition only on (o, a); supplying proxies to them would require nonstandard modifications, whereas the same-data Uniform Behavior Cloning warm-start of Appendix D.1 is precisely the control that shares COOT's data access while omitting the reliability machinery.

Seed protocol and per-seed datasets. The reported five runs are five independent random seeds, not five short rollouts. Each seed is a complete training run over the full 500k-step online budget, evaluated every 5k steps, and every method is run over the same five seeds on each of the five tasks. Five-seed reporting under a full online budget is standard in computationally intensive continuous-control and diffusion-policy RL, including closely related diffusion-policy work (Ren et al., 2025; Ma et al., 2025; Qing et al., 2025; Ding et al., 2026), and we pair it with the seed-first aggregation of Section 4.1 and the seed-paired intervals of Appendix D.4 to report variability honestly rather than relying on seed count alone. Each seed draws its *own* independently generated confounded offline dataset $\mathcal{D}^s_{\text{off}}$, so variability across seeds reflects both data collection and online training, testing robustness to different logs. Within a seed, however, every method that consumes offline data uses the identical $\mathcal{D}^s_{\text{off}}$: COOT and the Uniform Behavior Cloning warm-start share the same seed-specific transitions, pretraining budget, and actor, so the same-seed comparison is exactly matched and only the transfer rule differs.

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, $R^{500\text{k}}_{\text{best}}$ 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 5 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

Table 5: **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. $R_{\text{best}}^{500\text{k}}$ 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	$R_{\text{best}}^{500\text{k}} \uparrow$	$\Delta R\,(\%)$	$\text{AUC}^{500\text{k}} \uparrow$	$\Delta\text{AUC}\,(\%)$
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

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 5. 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 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.

Comparison with conventional uncertainty-aware transfer. The preceding component ablations establish that COOT benefits from its bridge, prior, shaping, and support pathways, but they do not by themselves determine whether the bridge advantage can be explained by conventional ensemble uncertainty or by simply adding proxy variables as ordinary predictive features. We therefore introduce two additional same-data controls whose purpose is to isolate the specific value of COOT's separated proxy-view reliability construction. The *bootstrap ensemble without proxies* fits an ensemble using only observations and actions,

Table 6: **Comparison with conventional uncertainty-aware transfer on HalfCheetah.** All methods use five matched seeds, the same 300k confounded offline transitions, the same 30k offline-pretraining updates, the same diffusion actor-critic, and the same 500k-step online budget. The no-proxy bootstrap ensemble tests generic resampling uncertainty, while the proxy-concat ensemble tests ordinary proxy-feature conditioning without separated proxy views. The Δ columns report COOT minus the corresponding control.

Method	$R^{500\text{k}}_{\text{best}}$ ↑	$\text{AUC}^{500\text{k}}$ ↑	ΔR	ΔAUC
Bootstrap ensemble, no proxies	9833.4 ± 904.6	7528.1 ± 507.1	855.5	842.4
Proxy-concat ensemble	10194.8 ± 790.1	7788.6 ± 425.0	494.2	581.9
COOT (separated proxy views)	$\mathbf{10688.9 \pm 619.0}$	$\mathbf{8370.5 \pm 657.5}$	–	–

so its dispersion captures conventional resampling-based predictive uncertainty but contains no proxy information. The *proxy-concat ensemble* supplies both proxies to the predictor as one concatenated feature vector, preserving their predictive information while removing COOT's separate proxy views and the resulting cross-view disagreement signal. All three methods use the same five random seeds, the same 300k confounded offline transitions, the same 30k offline-pretraining updates, the same diffusion actor and critic, and the same 500k-step online budget. The comparison therefore changes only how uncertainty and proxy information are represented and used for transfer.

Table 6 shows that full COOT achieves the highest mean performance on both metrics. Relative to the no-proxy bootstrap ensemble, COOT improves $R^{500\text{k}}_{\text{best}}$ from 9833.4 to 10688.9, an increase of 855.5, or 8.7%, and improves $\text{AUC}^{500\text{k}}$ from 7528.1 to 8370.5, an increase of 842.4, or 11.2%. This result shows that generic ensemble uncertainty is useful but does not recover the performance of the proxy-informed mechanism. Relative to the proxy-concat ensemble, COOT improves peak return by 494.2, or 4.8%, and AUC by 581.9, or 7.5%. Thus, providing the same proxy variables as ordinary input features also does not recover the benefit of preserving them as separate reliability views. COOT outperforms each uncertainty-aware control on both metrics in four of the five paired seeds, showing that the aggregate advantage is not driven by a single favorable run.

The seed-paired differences are also directionally consistent. Against the no-proxy bootstrap ensemble, the paired gains are 855.5 in best return with a bootstrap 95% confidence interval of $[-29.7, 1508.8]$ and 842.4 in AUC with an interval of $[-12.6, 1568.0]$. Against the proxy-concat ensemble, the corresponding gains are 494.2 with an interval of $[-206.7, 1195.0]$ and 581.9 with an interval of $[-18.6, 1081.5]$. The paired p-values are 0.132 and 0.136 for the no-proxy control and 0.310 and 0.145 for the proxy-concat control. With five seeds, these intervals still include zero, so we do not claim conventional statistical significance for these two control comparisons. Nevertheless, the same ordering appears across both aggregate metrics and four of five paired seeds. The pattern is mechanistically informative: conventional bootstrap uncertainty recovers part of the benefit, ordinary proxy conditioning recovers an additional part, and the best performance is obtained when COOT keeps the proxies separate and treats their disagreement as a transfer-reliability warning. This supports the narrower interpretation of bridge disagreement as a useful proxy-view stability signal rather than as proof of hidden confounding.

What bridge disagreement measures in practice. The held-out bridge diagnostic clarifies both the value and the limitation of disagreement. For an equally weighted pair of proxy-conditioned ensembles, the total bridge variance decomposes as

$$\widehat{\sigma}^2_{\text{br}} = \underbrace{\frac{1}{2}\widehat{s}^2_Z + \frac{1}{2}\widehat{s}^2_W}_{\text{within-view bootstrap uncertainty}} + \underbrace{\frac{1}{4}\left(\widehat{\mu}_Z - \widehat{\mu}_W\right)^2}_{\text{cross-view disagreement}} . \tag{89}$$

Across the five matched seeds, the cross-view term accounts for 56.4% to 60.3% of the per-sample bridge variance, with a mean share of 58.5%, and exceeds the within-view bootstrap contribution for every seed. The bridge signal is therefore not merely conventional resampling uncertainty. It is dominated by a systematic difference between what the two proxy views imply about the same transition. At the same time, cross-view disagreement is only a weak standalone detector of the synthetic oracle gap. Its mean Spearman

correlation is 0.018 and its mean high-gap AUROC is 0.514, compared with 0.012 and 0.507 for the no-proxy bootstrap ensemble and 0.033 and 0.523 for within-view uncertainty. Thus, disagreement is slightly more informative than generic no-proxy dispersion in this diagnostic, but it is not uniformly more predictive than every conventional uncertainty component and should not be interpreted as a calibrated estimator of hidden confounding. The decomposition and the online controls answer complementary questions. The decomposition shows that proxy-view conflict is a distinct source of uncertainty, while the controls test whether preserving and using that distinction improves transfer decisions. Full COOT exceeds the no-proxy bootstrap ensemble by 855.5 in best return and 842.4 in AUC, and exceeds the proxy-concat ensemble by 494.2 and 581.9, respectively, with the same ordering in four of five paired seeds. Although the five-seed confidence intervals include zero, this consistent direction indicates that the useful role of disagreement is not to act as a general-purpose error detector in isolation. Rather, it provides a view-specific instability warning that becomes useful when combined with COOT's support checks, reliability weighting, and conservative transfer mechanism. We therefore interpret bridge disagreement as a confounding-sensitive transfer-reliability signal, not as proof of hidden confounding or a direct estimate of causal bias.

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 7 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 5, 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 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.

Table 7: **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

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. The uncertainty-aware controls additionally show that these gains are not reproduced by generic bootstrap uncertainty or by treating the proxies as ordinary concatenated features. Instead, the strongest performance is obtained when the proxy views remain separate and their disagreement is used as a reliability signal. 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,

$$\tilde{x} = \frac{x - \hat{\mu}_{\text{pool}}}{\hat{\sigma}_{\text{pool}} + \varepsilon}, \tag{90}$$

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

$$y = 1\left[\tilde{x} \sim \mathcal{D}_{\text{conf}}\right]. \tag{91}$$

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.

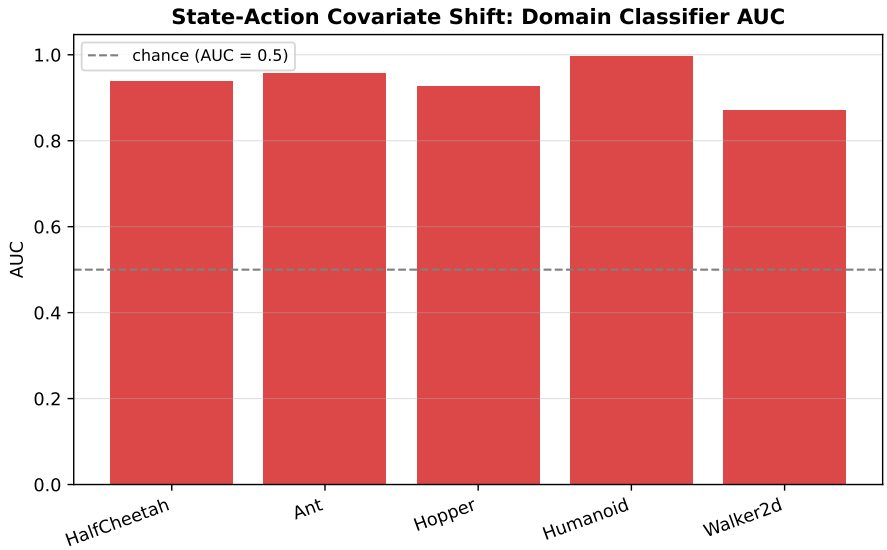

Figure 3: State-action domain-classifier AUC between confounded and plain reference datasets. Values far above chance show that the confounded data-generation process induces state-action distributions that are systematically distinguishable from ordinary interaction data.

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):

$$D_\mu = \left\| \frac{1}{n_{\text{conf}}} \sum_{i=1}^{n_{\text{conf}}} \tilde{x}_i^{\text{conf}} - \frac{1}{n_{\text{plain}}} \sum_{j=1}^{n_{\text{plain}}} \tilde{x}_j^{\text{plain}} \right\|_2 . \tag{92}$$

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):

$$\begin{aligned} \text{MMD}^2\left(\mathcal{D}_{\text{conf}}, \mathcal{D}_{\text{plain}}\right) = & \frac{1}{n_{\text{conf}}^2} \sum_{i,i'} k\left(\tilde{x}_i^{\text{conf}}, \tilde{x}_{i'}^{\text{conf}}\right) + \frac{1}{n_{\text{plain}}^2} \sum_{j,j'} k\left(\tilde{x}_j^{\text{plain}}, \tilde{x}_{j'}^{\text{plain}}\right) \\ & - \frac{2}{n_{\text{conf}} n_{\text{plain}}} \sum_{i,j} k\left(\tilde{x}_i^{\text{conf}}, \tilde{x}_j^{\text{plain}}\right), \end{aligned} \tag{93}$$

$$k(x, y) = \exp\left(-\frac{\|x - y\|_2^2}{2\sigma^2}\right), \tag{94}$$

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 8 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. 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

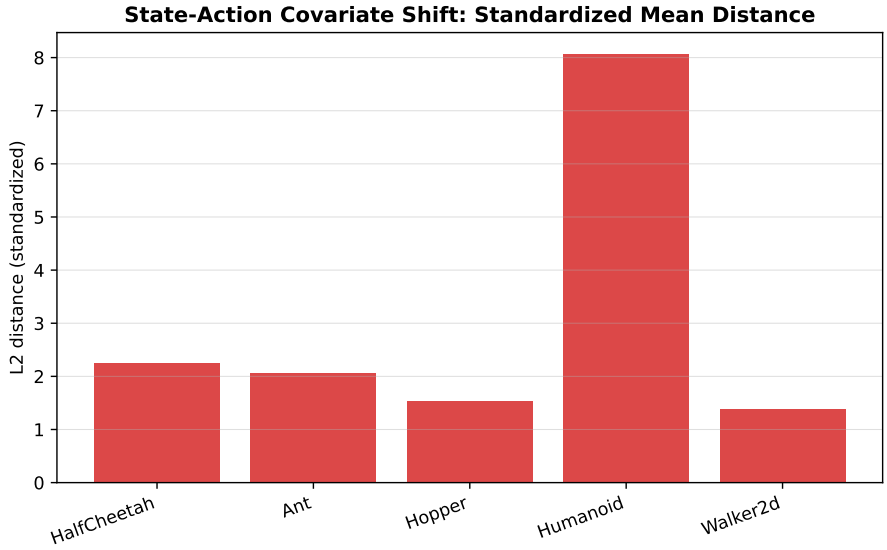

Figure 4: Standardized state-action mean distance. The large Humanoid value reflects accumulated coordinate-wise shifts in a high-dimensional state-action representation.

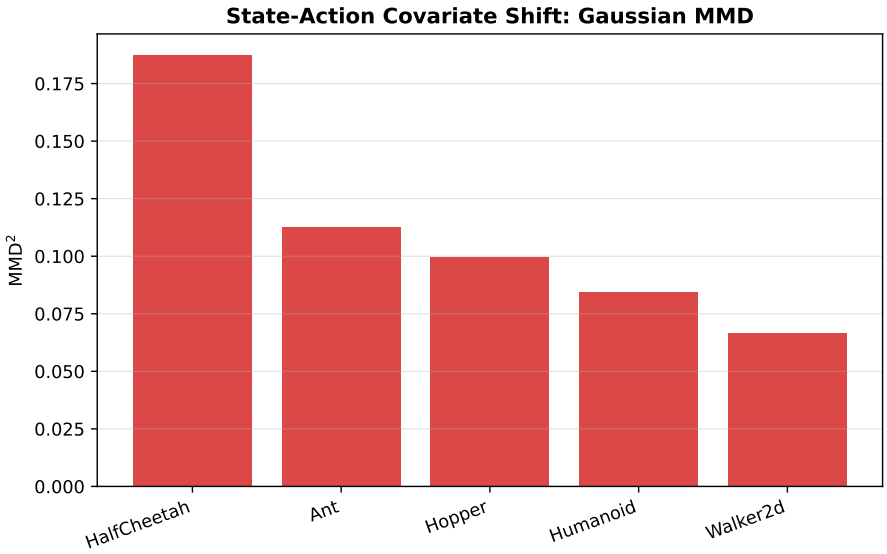

Figure 5: Gaussian-kernel MMD between confounded and plain state-action distributions. Positive values across all environments show distributional discrepancy beyond a simple mean shift.

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.

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

Table 8: 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

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.

D.6 Offline Validation on Realistic Industrial Telemetry: HAI 22.04

Purpose and experimental setting. The synthetic MuJoCo experiments provide a controlled test because the hidden execution regime is generated by construction and is therefore known. They do not, however, establish that realistic telemetry contains the observable structure needed by COOT. We therefore conduct a complementary offline validation on HAI 22.04,[1] a hardware-in-the-loop industrial-control dataset containing physical measurements, controller commands, actuator feedback, machine diagnostics, and documented disturbance intervals. HAI is not treated as an offline-RL benchmark, and this experiment does not claim online industrial policy improvement. Its narrower purpose is to test whether realistic telemetry satisfies the empirical prerequisites for proxy-informed transfer.

For each valid transition, let

$$X_t = (O_t, A_t), \qquad Z_t \in \mathbb{R}^9, \qquad W_t \in \mathbb{R}^{19}, \tag{95}$$

where $O_t \in \mathbb{R}^{13}$ contains ordinary process measurements, $A_t \in \mathbb{R}^{19}$ contains logged valve, pump, solenoid, speed-control, and set-point commands, Z_t contains local actuator and normalized feedback channels, and W_t contains rotating-machine and hardware-in-the-loop diagnostic channels. The two views are separated by physical role rather than selected to maximize the reported scores. Z_t describes local execution, whereas W_t describes broader plant condition. This gives disagreement a concrete interpretation because two physically distinct measurements of the same operation may support compatible or conflicting predictions.

Because HAI is a security dataset rather than an RL benchmark, it does not provide an operator-defined reward. We therefore construct a transparent, label-free future process-health outcome. Let m_j and $s_j = Q_{0.75}(O_j) - Q_{0.25}(O_j) + \epsilon$ be the median and robust scale of physical coordinate j, estimated only from normal training files. We define

$$M_t = \frac{1}{d_O} \sum_{j=1}^{d_O} \mathrm{clip}\left(\frac{|O_{t+1,j} - O_{t,j}|}{s_j}, 0, 10 \right), \tag{96}$$

$$L_t = \frac{1}{d_O} \sum_{j=1}^{d_O} \mathrm{clip}\left(\frac{|O_{t+1,j} - m_j|}{s_j}, 0, 10 \right), \tag{97}$$

$$Y_t = -\left(0.60 M_t + 0.20 L_t + 2.00 S_t\right), \tag{98}$$

where M_t measures abrupt physical motion, L_t measures departure from the normal operating region, and $S_t \in [0, 1]$ is the fraction of active safety indicators. The negative sign makes larger values correspond to healthier operation. The fixed coefficients keep rapid transients prominent, retain sensitivity to sustained drift, and prevent sparse safety indicators from being diluted by many continuous sensors. The official HAI disturbance label is excluded from this outcome, all predictors, the proxy definitions, and model selection. It is used only for a separate post-hoc check.

[1] https://github.com/icsdataset/hai

Validation design. Industrial streams are strongly autocorrelated, so random row-wise splitting can place nearly identical neighboring measurements in both training and test sets. We instead use five blocked cross-validation folds and remove a purge region of $g = 100$ retained transitions around each test block. All reported predictions are therefore out of block and temporally separated from their training samples.

The first object measures whether telemetry adds future-outcome information beyond the record available to an ordinary state and action learner:

$$\Delta_Y = R^2(Y_t \mid X_t, Z_t, W_t) - R^2(Y_t \mid X_t)\,. \tag{99}$$

A positive Δ_Y means that the telemetry contains predictive content not recoverable from (O_t, A_t) alone. To ensure that this content is tied to the correct physical event rather than to generic sensor ranges or long-term trends, we circularly shift both views in time. If π_s denotes shift s, the alignment gain is

$$\Delta_{\text{align}} = R^2(Y_t \mid X_t, Z_t, W_t) - \frac{1}{S}\sum_{s=1}^{S} R^2\big(Y_t \mid X_t, Z_{\pi_s(t)}, W_{\pi_s(t)}\big)\,. \tag{100}$$

The shift preserves the values and much of their serial structure while deliberately breaking event-level correspondence. Hence $\Delta_{\text{align}} > 0$ isolates correctly aligned telemetry information.

We next test the two observable pathways that motivate hidden-execution analysis. The dynamics lift

$$\Delta_{\text{dyn}} = R^2(O_{t+1} - O_t \mid X_t, Z_t, W_t) - R^2(O_{t+1} - O_t \mid X_t) \tag{101}$$

asks whether telemetry helps explain how the plant physically responds after conditioning on state and command. The control-selection lift

$$\Delta_A = R^2(A_t \mid O_t, Z_t, W_t) - R^2(A_t \mid O_t) \tag{102}$$

asks whether the logged command depends on information beyond the ordinary physical state. Positive values do not prove a hidden common cause, but they establish the observable dependencies that a proxy-based transfer mechanism must exploit.

Finally, we test whether separated views provide a useful reliability warning. Let

$$\widehat{Y}_t^Z = f_Z(X_t, Z_t), \qquad \widehat{Y}_t^W = f_W(X_t, W_t), \qquad D_t = \left|\widehat{Y}_t^Z - \widehat{Y}_t^W\right|, \tag{103}$$

and let $E_t = |Y_t - \widehat{Y}_t^{ZW}|$ be the held-out residual of the predictor using both views. We evaluate the Spearman association $\rho_S(D, E)$ and the AUROC of D_t for identifying the highest-residual quintile. These quantities give disagreement an operational role. It need not recover the hidden regime exactly, but it should increase when transferring the offline prediction is less reliable. We additionally train a domain classifier between normal training files and later HAI test files. Its AUROC measures whether the historical and later state and control populations occupy measurably different regions.

Results and interpretation. Table 9 reports results on 50,000 uniformly retained transitions, five blocked folds, a purge gap of 100, and 20 circular-shift controls. The retained rows and blocked folds are fixed. The five audit repetitions vary only the null-shift and matched-domain constructions, so the reported standard deviation is informative for those randomized controls rather than an independent-training-seed uncertainty estimate.

The strongest result is incremental outcome relevance. Adding (Z_t, W_t) raises held-out R^2 from 0.4548 to 0.6475, giving $\Delta_Y = 0.1927$. Equivalently, the unexplained fraction falls from 0.5452 to 0.3525, a 35.3% reduction. The alignment gain is even larger, 0.2145 ± 0.0016. Correctly matched telemetry is substantially more informative than the same streams attached to the wrong transitions. This rules out the simple explanation that the gain arises only from adding high-dimensional, autocorrelated sensor channels.

The pathway diagnostics are also positive. Telemetry improves prediction of the next physical change by 0.0424 in R^2 and improves prediction of logged commands by 0.1009 beyond the ordinary physical state.

Table 9: Offline validation on HAI 22.04 hardware-in-the-loop industrial telemetry. Lifts compare blocked cross-fitted models with and without the two telemetry views. Official disturbance labels are excluded from predictors, proxy construction, outcome construction, and model selection.

Diagnostic	**Held-out result**
Outcome relevance	$R^2 : 0.4548 \to 0.6475\,(\Delta_Y = 0.1927)$
Correct event alignment	$\Delta_{\text{align}} = 0.2145 \pm 0.0016$, $p = 0.0476$
Physical evolution	$R^2 : 0.0943 \to 0.1368\,(\Delta_{\text{dyn}} = 0.0424)$
Control selection	$R^2 : 0.3741 \to 0.4750\,(\Delta_A = 0.1009)$
Disagreement and residual association	$\rho_S(D, E) = 0.1844$
High-residual identification	AUROC $= 0.5886$
Historical/test shift	domain AUROC $= 0.9536$
Withheld disturbance regime	proxy-only AUROC $= 0.6309$

Thus, realistic auxiliary channels carry information about both how the plant evolves and which controls are issued. These findings do not identify a unique latent confounder, but they demonstrate that the observable structure motivating COOT is present in a realistic control log rather than only in the synthetic generator.

Cross-view disagreement provides a weaker but useful reliability signal. Its rank correlation with held-out residual magnitude is 0.1844, and it reaches 0.5886 AUROC for identifying the highest-error samples. Hence disagreement is not a precise fault detector, but larger disagreement systematically marks less reliable outcome predictions, which is the graded warning required for cautious transfer. The domain AUROC of 0.9536 further shows that normal historical files and later HAI operation occupy strongly different state and control distributions, motivating explicit support and reliability checks. In a label-withheld post-hoc check, the proxies alone obtain 0.6309 AUROC for the documented disturbance regime, confirming that they contain measurable regime information independently of the constructed outcome.

Overall, HAI 22.04 positively supports the empirical plausibility of COOT's assumptions beyond synthetic MuJoCo. The telemetry is incrementally outcome-relevant, correctly aligned, related to physical evolution and control selection, informative under strong source to target shift, and yields disagreement that modestly tracks predictive unreliability. The scope remains deliberately limited. HAI is a hardware-in-the-loop testbed rather than a deployed production plant. Its commands are not actions selected by an RL policy, and Y_t is an interpretable process-health surrogate rather than an operator-recorded return. Moreover, disagreement is near chance for directly classifying the official attack label (AUROC 0.5098). COOT's disagreement should therefore be interpreted as a transfer-reliability score, not as a general cyberattack detector. The same pre-deployment logic extends to robotics, recommender systems, and industrial control. Candidate proxies should be accepted only when they demonstrate incremental outcome relevance, correct event alignment, action and transition relevance, measurable source to target shift, and calibrated disagreement. The HAI study complements the controlled MuJoCo experiments by showing that these requirements are quantitatively testable and substantially supported in realistic industrial telemetry.