ATTACKING AND DEFENDING ONE-STEP FLOW POLICIES IN OFFLINE REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

One-step flow policies enable efficient action generation in offline reinforcement learning, but remain vulnerable to observation perturbations. We introduce Behavior-Decoupled Adversarial Perturbation (BDAP), which trains a generator against a frozen surrogate to lower estimated action value at the true state while increasing action deviation from a behavior-only reference. At deployment, BDAP generates perturbations in one forward pass without victim queries or gradients. The existing Diffusion Model-Based Predictor (DMBP) partially mitigates these attacks but leaves reconstruction errors. We therefore propose the Residual-Corrected Filter (RCF), which predicts these residuals from the perturbed observation and the DMBP-reconstructed observation and corrects the reconstruction before policy execution. RCF is trained only on simulator rollouts with uniform random observation perturbations, keeping the policy and DMBP frozen, and requires no attack-type information at deployment. Our analysis bounds reconstruction error through prediction loss, target clipping, and correction constraints, and connects the remaining error to a return-gap bound relative to the clean policy under Lipschitz assumptions. Experiments on OGBench and D4RL across four policy families demonstrate effective cross-policy transfer of BDAP and aggregate performance gains from RCF over DMBP under adversarial attacks and Gaussian noise.

1 INTRODUCTION

Offline reinforcement learning (RL) learns policies from fixed datasets, making it useful for control tasks where online exploration is costly or unsafe (Levine et al., 2020; Kumar et al., 2022). These methods balance value improvement with constraints or regularization based on dataset behavior (Fujimoto et al., 2019; Kumar et al., 2019; 2020; Fujimoto & Gu, 2021; Kostrikov et al., 2021). In continuous-action control, policies must capture multimodal action distributions while generating actions efficiently at deployment. Diffusion policies provide expressive action models but typically require iterative sampling (Wang et al., 2022; Chi et al., 2025). Consistency policies reduce the number of sampling steps (Ding & Jin, 2024), while recent *one-step flow policies* combine expressive policy learning with single-step candidate-action generation (Park et al., 2025b; Wang et al., 2026; Nguimatsia Tiofack et al., 2026; Koirala & Fleming, 2026). Their benchmark performance makes them promising for control, but efficient action generation alone does not establish robustness to measurement noise or adversarial observation perturbations.

Observation attacks can change a policy's actions and reduce its return without directly altering the physical state (Lin et al., 2017; Zhang et al., 2020; 2021). For offline RL, perturbed observations may differ from those represented in the fixed training dataset (Levine et al., 2020; Yang & Xu, 2024). Studies of multistep diffusion policies also show that tolerance to random noise does not ensure robustness to adversarial perturbations (Chen et al., 2024). For one-step policies, fitting actions on training observations likewise leaves their response to perturbed observations untested. We therefore investigate the vulnerability of one-step flow policies to observation perturbations. In particular, can an attacker degrade their performance with zero queries to the victim model at deployment?

We show that such query-free attacks are possible with Behavior-Decoupled Adversarial Perturbation (BDAP). Unlike gradient-based optimization and critic-scored sampling, which repeatedly evaluate the victim at each state (Madry et al., 2017; Zhang et al., 2020; Yang & Xu, 2024), BDAP

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

learns a perturbation generator offline using a frozen Flow Q-Learning (FQL) surrogate (Park et al., 2025b) and its critic, together with a separately trained behavior-only reference. Its objective seeks to lower the critic's value estimate for the induced action at the true state while increasing its displacement from the reference. BDAP decouples the learning of this behavior reference from critic-driven value optimization, providing a separate behavioral anchor for the attack. Once trained for a task and budget, the generator is reused across victim policies through surrogate transfer (Papernot et al., 2017; Dong et al., 2018), producing each perturbation in one forward pass without victim queries or gradients.

The vulnerability exposed by these attacks motivates protection at the same observation interface. The Diffusion Model-Based Predictor (DMBP) (Yang & Xu, 2024) reconstructs states from perturbed observations and temporal context. Applied to the attacked flow policies, it partially restores performance. Although DMBP uses a non-Markovian multistep objective to mitigate error accumulation through reconstructed histories, residual reconstruction errors may still enter the frozen policy and alter its actions, affecting subsequent control. We therefore introduce the Residual-Corrected Filter (RCF) to correct these remaining errors before action generation. RCF predicts the residual from the perturbed observation and the DMBP reconstruction, then applies a bounded, scaled correction. It learns only from uniformly perturbed simulator rollouts, keeping both the policy and DMBP frozen. At deployment, the predictor uses these two inputs without requiring knowledge of the attack type.

Experiments on OGBench (Park et al., 2025a) and D4RL (Fu et al., 2020) across FQL (Park et al., 2025b), MeanFlowQL (Wang et al., 2026), SSCP (Koirala & Fleming, 2026), and GFP (Nguimatsia Tiofack et al., 2026) show that FQL-trained BDAP generators expose cross-policy vulnerabilities without target-model queries, matching or exceeding the victim-dependent APGD (Croce & Hein, 2020) and min-Q (Yang & Xu, 2024) baselines in most reported task-group comparisons. DMBP (Yang & Xu, 2024) provides partial protection, and RCF further improves its performance across policy families under adversarial perturbations and Gaussian noise, beyond the uniform perturbations used for fitting.

Our contributions are threefold. (1) BDAP decouples behavior-reference learning from value optimization to construct transferable attacks on one-step flow policies without victim access at deployment. (2) RCF strengthens DMBP through residual prediction, using only uniformly perturbed simulator rollouts for training and no attack-type information at deployment. (3) We derive a reconstruction-error bound for RCF in terms of residual prediction loss and the correction mechanism, and relate the remaining observation error to the return gap between the defended and clean systems under Lipschitz assumptions.

2 Preliminaries and Related Work

2.1 One-step flow policies

One-step flow policies generate each candidate action from an observation and a random latent in a single forward pass. We detail FQL (Park et al., 2025b), which provides BDAP's surrogate and learns a one-step actor through value optimization and behavior-flow distillation.

Let $\mathcal{D}$ denote the offline dataset of observations $s \in \mathcal{S} \subseteq \mathbb{R}^d$ and actions $a \in [-1,1]^{d_a}$, where d and d_a are the observation and action dimensions. The behavior flow models the conditional action distribution using a velocity field $v_\theta : [0,1] \times \mathcal{S} \times \mathbb{R}^{d_a} \to \mathbb{R}^{d_a}$ (Lipman et al., 2022; Liu et al., 2022). Given $z \sim \mathcal{N}(0, I_{d_a})$ and $t \sim \mathrm{Unif}[0,1]$, define $x_t = (1-t)z + ta$. Here t is flow time, distinct from the environment timestep. We train the behavior flow only by minimizing

$$\mathcal{L}_{\mathrm{FM}}(\theta) = \mathbb{E}_{\substack{(s,a)\sim\mathcal{D},\ z\sim\mathcal{N}(0,I) \\ t\sim\mathrm{Unif}([0,1])}} \left[\frac{1}{d_a} \left\| v_\theta(t, s, x_t) - (a - z) \right\|_2^2 \right]. \tag{1}$$

Integrating $\dot{x}_t = v_\theta(t, s, x_t)$ from $x_0 = z$ gives the endpoint x_1 and the behavior action $\mu_\theta(s, z) = \mathrm{clip}(x_1, -1, 1)$. Sampling z induces a stochastic policy.

Since the behavior flow requires numerical integration, FQL uses a separate one-step actor for deployment. Let $u_\eta(s, z)$ denote its raw output, with executed action $\pi_\eta(s, z) = \mathrm{clip}(u_\eta(s, z), -1, 1)$. The critic $Q_\psi(s, a)$ averages the critic heads to estimate discounted return. Holding the flow and

critic fixed during the actor update, FQL minimizes

$$\mathcal{L}_{\mathrm{FQL}}(\eta) = \mathbb{E}_{s\sim\mathcal{D},\, z\sim\mathcal{N}(0,I)}\left[-Q_\psi\big(s, \pi_\eta(s,z)\big) + \frac{\alpha}{d_a}\left\|u_\eta(s,z) - \mu_\theta(s,z)\right\|_2^2\right]. \tag{2}$$

The first term maximizes estimated value, while the second distills the behavior-flow endpoint with weight α. Both squared losses average over action coordinates. This actor avoids backpropagation through the ODE solver during optimization and iterative integration at deployment.

Beyond FQL, we consider a broader class of flow-based policies that generate candidate actions in a single forward pass, including GFP, MeanFlowQL, and SSCP (Park et al., 2025b; Nguimatsia Tiofack et al., 2026; Wang et al., 2026; Koirala & Fleming, 2026). Throughout the paper, we use $\pi(s,z)$ to denote the deployed action map of such a policy, where z collects all sampling randomness. This notation includes any action clipping and candidate-selection procedure used at deployment.

2.2 Observation corruption in SA-MDPs

Let $M = (\mathcal{S}, \mathcal{A}, P, R, \gamma)$ denote the underlying Markov decision process (MDP), with state space $\mathcal{S} \subseteq \mathbb{R}^d$ and action space $\mathcal{A} = [-1,1]^{d_a}$. The transition kernel P, expected reward function R, and discount factor $\gamma \in [0,1)$ specify the environment dynamics and return criterion. In the state-adversarial Markov decision process (SA-MDP) formulation (Zhang et al., 2020), corruption changes the policy input without directly modifying the true state. At state s, the policy receives $\tilde{s} = s + \delta$ and produces $a = \pi(\tilde{s}, z)$, while the next state is drawn as $s' \sim P(\cdot \mid s, a)$. Thus, the perturbation affects transitions through the selected action.

Let $\sigma \in \mathbb{R}^d_{>0}$ denote the numerically stabilized coordinatewise standard deviation of offline observations. For a perturbation budget $\epsilon \geq 0$, define the admissible observation set as

$$\mathcal{B}(s) := \left\{\tilde{s} \in \mathbb{R}^d : \left\|\mathrm{diag}(\sigma)^{-1}(\tilde{s} - s)\right\|_\infty \leq \epsilon\right\}. \tag{3}$$

The ℓ_∞ constraint limits the perturbation in each standardized coordinate. Gaussian noise is a separate random-corruption control, with its scale controlling noise dispersion rather than imposing a hard bound.

2.3 DMBP-based state reconstruction

Observation robustness can be addressed during policy training or through input reconstruction before inference. RADIAL-RL, WocaR-RL, and RORL follow the training-based approach (Oikarinen et al., 2021; Liang et al., 2022; Yang et al., 2022), while Defense-GAN and DiffPure use generative models to reconstruct inputs in vision (Samangouei et al., 2018; Nie et al., 2022). For offline RL, DMBP (Yang & Xu, 2024) uses conditional diffusion to reconstruct the underlying clean state from a corrupted observation and temporal context before policy execution. Let $h_t := [\hat{s}_{t-N}, \ldots, \hat{s}_{t-1}]$ denote a length-N window of previous reconstructions. Given $\tilde{s}_t$, the preceding action a_{t-1}, and this history, its reverse-denoising procedure produces

$$\hat{s}_t = F(\tilde{s}_t, a_{t-1}, h_t, \xi_t), \tag{4}$$

where F is the DMBP reconstruction map, held fixed in our framework, and ξ_t collects its sampling randomness. Under DMBP, $\hat{s}_t$ is supplied to the policy and also enters the next reconstruction history, so errors can accumulate through temporal conditioning. DMBP uses a non-Markovian multistep objective to mitigate this accumulation (Yang & Xu, 2024), but residual reconstruction errors may still remain. We denote the remaining residual by

$$r_t := s_t - \hat{s}_t. \tag{5}$$

3 Methodology

BDAP learns transferable observation perturbations offline to expose the vulnerability of one-step policies without querying the victim at deployment. To strengthen DMBP's protection against such perturbations, RCF predicts and corrects its remaining reconstruction residuals. Figure 1 summarizes the training and deployment of both methods.

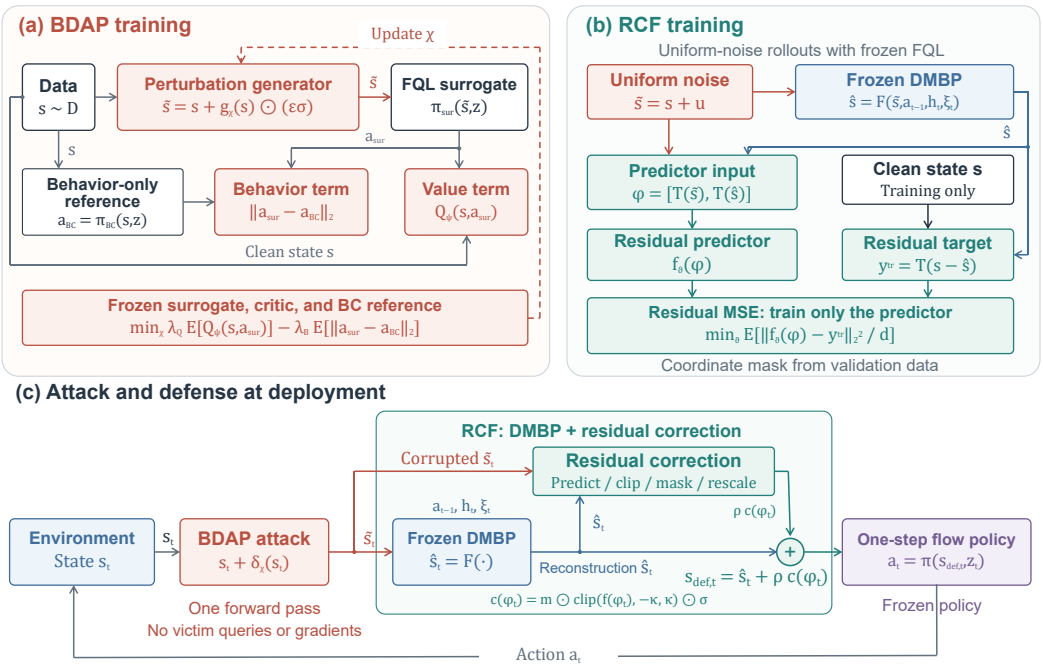


Figure 1: BDAP and RCF overview. (a) BDAP learns perturbations using value and behavior signals. (b) RCF learns residual prediction under uniform random perturbations. (c) At deployment, BDAP requires no victim queries, and RCF corrects DMBP reconstructions before the policy acts.

3.1 Attack Side: The Behavior-Decoupled Adversarial Perturbation

Attack construction and bounded perturbation. BDAP trains a perturbation generator on $\mathcal{D}$ using a frozen FQL surrogate π_{sur} with critic Q_ψ and a separately trained behavior-only reference π_{BC}. The deterministic generator g_χ parameterizes a bounded observation perturbation as

$$\delta_\chi(s) = g_\chi(s) \odot (\epsilon\sigma), \qquad \tilde{s} = s + \delta_\chi(s). \tag{6}$$

Here $\odot$ denotes elementwise multiplication. The bound $|g_{\chi,i}(s)| \le 1$ ensures $\tilde{s} \in \mathcal{B}(s)$ under the coordinate scaling σ and budget ϵ. Training evaluates the induced action by its value at the true state and its displacement from the behavior-only reference, which we prepare below.

Behavior-only reference preparation. To decouple the behavior reference from value optimization, we train a separate one-step policy to model dataset behavior without the critic objective in (2). We first learn a behavior velocity field v_ζ by minimizing (1) with parameters ζ. Numerical integration from z, followed by clipping to $[-1, 1]$, defines its action output $\mu_\zeta(s, z)$. A separate one-step actor with raw output $u_\omega(s, z)$ learns to match this output through the distillation loss

$$\mathcal{L}_{\text{dist}}(\omega; \zeta) = \mathbb{E}_{s\sim\mathcal{D},\, z\sim\mathcal{N}(0,I_{d_a})} \frac{1}{d_a} \left\| u_\omega(s, z) - \mu_\zeta(s, z) \right\|_2^2 . \tag{7}$$

Each actor update minimizes $\mathcal{L}_{\text{dist}}$ over ω, with ζ held fixed. Neither model uses a critic objective. After training, we freeze $\pi_{\text{BC}}(s, z) = \text{clip}(u_\omega(s, z), -1, 1)$ as the action reference for generator training.

Behavior-decoupled attack objective. For a shared latent $z \sim \mathcal{N}(0, I_{d_a})$, define $a_{\text{sur}} = \pi_{\text{sur}}(\tilde{s}, z)$, $a_{\text{BC}} = \pi_{\text{BC}}(s, z)$, and $\Delta_{\text{B}}(s, z) = a_{\text{sur}} - a_{\text{BC}}$, using the clipped action maps defined in the preliminaries. The value response $q_{\text{sur}} = Q_\psi(s, a_{\text{sur}})$ evaluates the induced action at the true state. Both q_{sur} and Δ_{B} depend on χ through the perturbed observation $\tilde{s}$. The two channels serve different roles. The critic directs the attack toward low estimated value, while the behavior-only reference anchors the action comparison to behavior learned from the offline dataset. Keeping this reference free of value optimization provides an additional attack signal without directly relying on critic estimates.

Since displacement alone does not imply low value, BDAP combines the channels by minimizing

$$\mathcal{L}(\chi) = \underbrace{\lambda_Q \mathbb{E}_{s\sim\mathcal{D},z}[q_{\mathrm{sur}}]}_{\text{value channel}} - \underbrace{\lambda_{\mathrm{B}} \mathbb{E}_{s\sim\mathcal{D},z}\|\Delta_{\mathrm{B}}(s,z)\|_2}_{\text{behavior channel}}, \tag{8}$$

where $\lambda_Q, \lambda_{\mathrm{B}} \geq 0$ weight the value and behavior channels, respectively.

Offline gradient-based generator optimization. We train the generator using stochastic gradients computed from mini-batches of observations and policy latents. At optimization step $k \geq 1$, let χ_k denote the current generator parameters. We sample B observations $s_i \overset{\text{i.i.d.}}{\sim} \mathcal{D}$ and independently draw n_z latents $z_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_{d_a})$ per observation, with $i = 1, \ldots, B$ and $j = 1, \ldots, n_z$. For any χ, define $\tilde{s}_i(\chi) = s_i + \delta_\chi(s_i)$, $a_{ij}(\chi) = \pi_{\mathrm{sur}}(\tilde{s}_i(\chi), z_{ij})$, and $a_{ij}^{\mathrm{BC}} = \pi_{\mathrm{BC}}(s_i, z_{ij})$. The batch responses are

$$\big(\widehat{q}_k(\chi), \widehat{d}_k(\chi)\big) := \frac{1}{Bn_z} \sum_{i=1}^{B} \sum_{j=1}^{n_z} \big(Q_\psi(s_i, a_{ij}(\chi)), \|a_{ij}(\chi) - a_{ij}^{\mathrm{BC}}\|_2\big). \tag{9}$$

These batch averages estimate the two expectations in (8). At step k, we minimize $\widehat{\mathcal{L}}_k(\chi) = \lambda_{Q,k}\widehat{q}_k(\chi) - \lambda_{\mathrm{B},k}\widehat{d}_k(\chi)$. The weights normalize the batch value and gradually introduce the behavior channel, respectively, and are held fixed during differentiation. Appendix B specifies their computation.

Let $\Delta_{ij} = a_{ij}(\chi_k) - a_{ij}^{\mathrm{BC}}$ and let $J_u f$ denote the Jacobian of f with respect to u. Define $D_i = J_\chi \delta_\chi(s_i)|_{\chi=\chi_k}$ and $A_{ij} = J_u \pi_{\mathrm{sur}}(u, z_{ij})|_{u=\tilde{s}_i(\chi_k)}$. The action-space gradient combines the critic gradient with the gradient of displacement from the fixed behavior reference. At differentiable points, holding the sampled observations and latents fixed, the chain rule gives (Appendix A.3)

$$G_k := \nabla_\chi \widehat{\mathcal{L}}_k(\chi_k) = \frac{1}{Bn_z} \sum_{i=1}^{B} \sum_{j=1}^{n_z} D_i^\top A_{ij}^\top \left[\lambda_{Q,k} \nabla_a Q_\psi\big(s_i, a_{ij}(\chi_k)\big) - \lambda_{\mathrm{B},k} \frac{\Delta_{ij}}{\|\Delta_{ij}\|_2}\right]. \tag{10}$$

The value channel backpropagates through the critic and surrogate actor, while the behavior channel backpropagates through the surrogate actor with the reference action held fixed. We update only the generator parameters χ, keeping the surrogate and reference parameters frozen.

At deployment, BDAP supplies $s + \delta_\chi(s)$ to the victim using one generator evaluation and no victim queries or gradients. The surrogate models are discarded. Applying the generator to a different victim gives a surrogate-transfer attack (Papernot et al., 2017), evaluated in Section 4.2. Algorithm 1 summarizes offline generator training.

3.2 Defense Side: The Residual-Corrected Filter

RCF learns to correct the residual left by the DMBP reconstruction in (4). It predicts the residual from the perturbed observation and the DMBP-reconstructed observation, then applies a gated, bounded correction to the policy input. DMBP and the policy remain frozen throughout.

Residual prediction and gated correction. Let $f_\vartheta : \mathbb{R}^{2d} \to \mathbb{R}^d$ denote the residual predictor. Its input concatenates the perturbed observation and the DMBP-reconstructed observation after coordinatewise scaling and clipping, giving $\varphi_t = \mathrm{clip}([\tilde{s}_t/\sigma, \hat{s}_t/\sigma], -10, 10)$. For fitting, we collect simulator rollouts under uniform random observation perturbations with RCF disabled. The clean observations provide the targets $y_t^{\mathrm{tr}} = \mathrm{clip}(r_t/\sigma, -10, 10)$. Let $\mathcal{P}_{\mathrm{tr}}$ denote the resulting distribution of training pairs. We train f_ϑ by minimizing the mean-squared error

$$\mathcal{L}_{\mathrm{RCF}}(\vartheta) := \mathbb{E}_{\mathcal{P}_{\mathrm{tr}}}\left[\frac{1}{d}\|f_\vartheta(\varphi) - y^{\mathrm{tr}}\|_2^2\right]. \tag{11}$$

We minimize this objective with mini-batch stochastic gradients, backpropagating the prediction error through f_ϑ while keeping the collected inputs and targets fixed. Only the residual predictor is updated. After fitting, write $f = f_{\vartheta^\star}$. To select the coordinates to correct, we compare each output's validation MSE with that of a zero-residual predictor. The score and fixed coordinate mask are

$$\mathrm{fe}_i^{\mathrm{val}} = 1 - \frac{\mathbb{E}[(f_i(\varphi) - y_i^{\mathrm{tr}})^2]}{\mathbb{E}[(y_i^{\mathrm{tr}})^2] + \eta_{\mathrm{fe}}}, \qquad m_i = \mathbf{1}\{\mathrm{fe}_i^{\mathrm{val}} > \tau\}. \tag{12}$$

Here f_i is the ith output, expectations use validation data, and $\eta_{\text{fe}} > 0$ stabilizes the denominator. The threshold τ determines which coordinates are retained. At deployment, we use this mask to form the correction and policy input

$$c(\varphi_t) = m \odot \text{clip}(f(\varphi_t), -\kappa, \kappa) \odot \sigma, \qquad s_{\text{def},t} = \hat{s}_t + \rho c(\varphi_t). \tag{13}$$

Masking removes unselected coordinates, clipping limits each retained prediction to $[-\kappa, \kappa]$, and multiplication by σ restores observation units. The coefficient $\rho \in [0, 1]$ scales the correction added to $\hat{s}_t$. The fitted predictor and mask are reused without attack-type information. Algorithm 2 summarizes predictor fitting and coordinate selection.

Reconstruction error controlled by residual prediction. The predictor is fitted to a clipped target, while the policy receives a gated and scaled correction. We account for these operations when relating prediction loss to the remaining reconstruction error. Fix f, $m \in \{0, 1\}^d$, and $\kappa > 0$. Write the standardized residual as $y^{\text{raw}} = r/\sigma$ and correction as $c_\sigma(\varphi) = c(\varphi)/\sigma$. Then $(s_{\text{def}} - s)/\sigma = \rho c_\sigma(\varphi) - y^{\text{raw}}$. Under a fixed joint distribution $\mathcal{Q}$ of $(y^{\text{raw}}, \varphi)$ with finite second moments for y^{raw} and $f(\varphi)$, the risks before and after correction are

$$\mathcal{R}_0 := \frac{1}{d}\mathbb{E}_{\mathcal{Q}}\|y^{\text{raw}}\|_2^2, \qquad \mathcal{R}_\rho(f) := \frac{1}{d}\mathbb{E}_{\mathcal{Q}}\|y^{\text{raw}} - \rho c_\sigma(\varphi)\|_2^2. \tag{14}$$

Both risks use the same underlying DMBP reconstructions and distribution $\mathcal{Q}$. The bound accounts for three sources of error. The first source is residual prediction error, which measures how well f fits the clipped residual. The second source is target-clipping error, which measures the difference between the raw and clipped residuals. They are defined as

$$\mathcal{L}_{\mathcal{Q}}(f) := \frac{1}{d}\mathbb{E}_{\mathcal{Q}}\|f(\varphi) - y^{\text{tr}}\|_2^2, \qquad \mathcal{L}_{\text{clip}} := \frac{1}{d}\mathbb{E}_{\mathcal{Q}}\|y^{\text{raw}} - y^{\text{tr}}\|_2^2. \tag{15}$$

On the training-pair distribution, $\mathcal{L}_{\mathcal{Q}}(f_\vartheta)$ is exactly the fitting objective $\mathcal{L}_{\text{RCF}}(\vartheta)$.

The third source is error imposed by the correction constraints. Masking and output clipping project the prediction onto the closed convex set $\mathcal{C} = \{v \in \mathbb{R}^d : v_i = 0 \text{ if } m_i = 0,\ |v_i| \le \kappa \text{ if } m_i = 1\}$. Let $\Pi_{\mathcal{C}}(w) := \arg\min_{v \in \mathcal{C}} \|w - v\|_2^2$ denote the Euclidean projection onto $\mathcal{C}$, so $c_\sigma(\varphi) = \Pi_{\mathcal{C}}(f(\varphi))$. Even with the raw residual known, the closest correction in $\mathcal{C}$ can leave error, measured by

$$\mathcal{L}_{\mathcal{C}} := \frac{1}{d}\mathbb{E}_{\mathcal{Q}}\|y^{\text{raw}} - \Pi_{\mathcal{C}}(y^{\text{raw}})\|_2^2. \tag{16}$$

Theorem 1 (Reconstruction Error Bound for Residual Correction)**.** *Under the preceding definitions and assumptions, for every fixed* $\rho \in [0, 1]$,

$$\mathcal{R}_\rho(f) \le (1 - \rho)\mathcal{R}_0 + \rho\left[\mathcal{L}_{\mathcal{C}} + \left(\sqrt{\mathcal{L}_{\mathcal{Q}}(f)} + \sqrt{\mathcal{L}_{\text{clip}}}\right)^2\right]. \tag{17}$$

For fixed $\mathcal{Q}$, m, κ, and $\rho > 0$, reducing prediction loss tightens the bound, while $\mathcal{L}_{\text{clip}}$ and $\mathcal{L}_{\mathcal{C}}$ account for target clipping and correction constraints. Appendix A.1 gives the proof and the coordinate expansion of $\mathcal{L}_{\mathcal{C}}$.

Connection to policy performance. We next relate the reconstruction error bounded in Theorem 1 to the defended policy's return. Let π be the clean policy and $\bar{\pi}$ the same policy receiving the defended input $s_{\text{def}}(x, Z)$, where x is the augmented state of the environment and filter and Z is policy sampling randomness drawn independently of x. Let $J(\pi)$ and $J(\bar{\pi})$ denote the expected discounted returns of the clean and defended systems, respectively, and $d^{\bar{\pi}}$ the normalized discounted state distribution under $\bar{\pi}$. Appendix A.2 specifies this process and the distribution definitions.

Proposition 1 (Return Gap under Reconstruction Error)**.** *Assume that both systems share the same initial physical-state distribution, rewards are bounded, and* $0 \le \gamma < 1$. *Let* Q^π *be the exact clean-policy action-value function. Take* $\mathcal{Q}$ *in* $\mathcal{R}_\rho(f)$ *to be the joint distribution of* $(y^{\text{raw}}, \varphi)$ *induced by* $x \sim d^{\bar{\pi}}$ *and policy randomness* Z. *Suppose uniform constants* L_Q, L_π *satisfy*

$$|Q^\pi(s, a) - Q^\pi(s, a')| \le L_Q\|a - a'\|_2,$$
$$\|\pi(u, z) - \pi(u', z)\|_2 \le L_\pi \left\|\text{diag}(\sigma)^{-1}(u - u')\right\|_2.$$

The first condition holds for every $s \in \mathcal{S}$ *and* $a, a' \in \mathcal{A}$. *The second holds for almost every* z, *uniformly over* $u, u' \in \mathbb{R}^d$. *Then*

$$|J(\bar{\pi}) - J(\pi)| \le \frac{L_Q L_\pi}{1 - \gamma}\sqrt{d\mathcal{R}_\rho(f)}. \tag{18}$$

Theorem 1 bounds $\mathcal{R}_\rho(f)$, which in turn controls the return-gap bound in Proposition 1. Applying both results under the same trajectory-induced distribution $\mathcal{Q}$ relates residual prediction error, target-clipping error, and correction-constraint error to the return gap.

4 Experimental Results

4.1 Experimental Setup

Benchmarks. Following FQL (Park et al., 2025b), we evaluate state-based policies on OGBench and D4RL. For OGBench (Park et al., 2025a), we use 50 reward-based single-task variants from five locomotion and five manipulation environments, each providing five distinct evaluation goals. We also include the six AntMaze and twelve Adroit tasks from D4RL (Fu et al., 2020) used by FQL.

Methods. We evaluate four one-step flow-policy families as victims: FQL (Park et al., 2025b), MeanFlowQL (Wang et al., 2026), SSCP (Koirala & Fleming, 2026), and GFP (Nguimatsia Tiofack et al., 2026), with frozen checkpoints. The main attack comparison evaluates BDAP against a 10-step Auto-PGD-style attack, denoted APGD (Croce & Hein, 2020), and min-Q sampling (Yang & Xu, 2024). BDAP, APGD, and min-Q share the coordinate-normalized ℓ_∞ constraint in (3). Our APGD baseline maximizes squared Euclidean distance between perturbed and clean actions. Min-Q selects among 20 uniformly sampled perturbations using the induced actions' estimated values at the true state. Appendix B.1 gives both implementations. For defense, we compare frozen DMBP (Yang & Xu, 2024) with DMBP+RCF. BDAP generators are trained against FQL for each task and budget, then transferred to the other three policy families without adaptation. Each task-specific RCF predictor is trained only on uniformly perturbed frozen-FQL rollouts at $\epsilon_{\text{tr}} = 0.4$ and reused across victim families and perturbations without attack-type information. The main attack and defense comparisons use $\epsilon = 0.1$, with $\rho = 0.3$ for RCF. Gaussian noise serves as a random-perturbation control.

Evaluation. The performances are averaged over 5 seeds, with 50 evaluation episodes per seed. We report success percentages for OGBench and D4RL AntMaze and normalized returns for Adroit. Raw clean uses unmodified observations without DMBP or RCF. Main tables average task-level means, including all five tasks per OGBench group. Appendices C.1 and C.3 report per-task means and standard deviations.

4.2 Attack Effectiveness and Transfer

Table 1 examines the vulnerability of undefended one-step flow policies to FQL-trained BDAP generators, including transfer to other policy families. BDAP constructs each perturbation with one generator evaluation and no target-model queries. APGD and min-Q instead access the victim policy or critic repeatedly at each state.

Transfer without target-model queries. On OGBench cube-single, transferring the FQL-trained generator reduces success from 49.6% to 4.8% on MeanFlowQL and from 54.4% to 8.4% on SSCP. APGD obtains 7.6% and 25.6% on these victims, while min-Q obtains 15.6% and 40.4%. On GFP scene, BDAP reduces success from 49.2% to 8.8%, compared with 23.2% under APGD and 20.8% under min-Q. Thus, in these task groups, the transferred generator reduces performance more than the methods that search against the deployed victim.

On D4RL, BDAP reduces GFP AntMaze success from 94.7% to 53.7%, while APGD and min-Q yield 91.3% and 81.7%. On GFP Adroit, BDAP yields a normalized return of 48.5, compared with 51.7 and 51.9 for the two baselines. The ranking depends on the victim. APGD is stronger on SSCP AntMaze, yielding 23.3% versus BDAP's 27.3%. BDAP matches or exceeds both victim-dependent online attacks in most reported task-group comparisons, despite requiring no access to the victim model at deployment. Appendix C.1 reports the full per-task results for all four victim families, including attacks at $\epsilon = 0.05$, 0.1, and 0.2 and Gaussian-noise controls, all without defense. BDAP remains effective across transfer victims at the larger budget without target-model queries.

Contribution of the two attack objectives. Appendix C.2 compares Full BDAP with Q-only and Behavior-only generators on five tasks. On the three transfer victim families, Full BDAP yields

Table 1: Attack performance at $\epsilon = 0.1$ without defense. Scores are success percentages for OGBench and D4RL AntMaze, and normalized returns for D4RL Adroit. Lower attack scores are better. Task-level results appear in Appendix C.1.

Task group	FQL				MeanFlowQL				SSCP				GFP			
	Raw clean	APGD	min-Q	BDAP	Raw clean	APGD	min-Q	BDAP	Raw clean	APGD	min-Q	BDAP	Raw clean	APGD	min-Q	BDAP
OGBench																
antmaze-large	80.8	34.8	18.8	11.2	0.0	0.0	0.0	0.0	27.6	0.0	0.0	0.4	94.4	63.2	48.8	41.6
antmaze-giant	14.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	24.8	0.0	3.6	0.0
antsoccer-arena	56.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.6	0.0	0.0	0.0	64.4	0.0	0.0	0.0
humanoidmaze-medium	63.6	5.2	46.0	0.4	57.2	0.4	54.0	0.4	0.4	0.0	0.4	0.0	74.4	1.6	48.0	0.4
humanoidmaze-large	3.2	0.0	0.8	0.0	18.0	0.0	11.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
cube-single	96.8	15.2	0.0	0.0	49.6	7.6	15.6	4.8	54.4	25.6	40.4	8.4	98.0	7.6	0.0	0.4
cube-double	25.2	0.4	0.0	0.0	0.8	0.4	0.4	0.4	1.2	0.0	2.0	0.0	48.8	1.6	2.4	3.2
scene	56.0	19.6	2.8	4.4	2.4	3.6	4.8	0.4	36.0	3.2	1.6	4.8	49.2	23.2	20.8	8.8
puzzle-3x3	27.2	14.0	2.0	2.0	14.0	5.2	10.8	2.0	26.0	12.0	14.0	17.2	21.2	20.0	17.2	9.2
puzzle-4x4	13.2	2.0	0.0	0.0	17.2	14.0	14.8	2.0	2.0	0.0	0.0	0.0	37.2	6.4	4.0	0.8
D4RL																
AntMaze	83.7	78.0	58.7	55.7	62.7	53.3	31.7	18.0	36.0	23.3	26.3	27.3	94.7	91.3	81.7	53.7
Adroit	52.3	43.3	51.4	42.7	1.4	0.2	0.4	0.8	55.6	47.3	51.7	45.6	59.4	51.7	51.9	48.5

Table 2: Defense against BDAP at $\epsilon = 0.1$ with $\rho = 0.3$. +RCF denotes DMBP+RCF. Scores are success percentages for OGBench and D4RL AntMaze, and normalized returns for D4RL Adroit. Higher scores are better. Bold highlights +RCF.

Task group	FQL			MeanFlowQL			SSCP			GFP		
	Raw clean	DMBP	+RCF	Raw clean	DMBP	+RCF	Raw clean	DMBP	+RCF	Raw clean	DMBP	+RCF
OGBench												
antmaze-large	80.8	42.8	**54.4**	0.0	0.0	**1.6**	27.6	2.0	**8.0**	94.4	72.4	**82.8**
antmaze-giant	14.8	0.0	**0.0**	0.0	0.0	**0.0**	0.0	0.0	**0.0**	24.8	0.0	**0.8**
antsoccer-arena	56.4	0.0	**0.0**	0.0	0.0	**0.0**	1.6	0.0	**0.0**	64.4	0.0	**0.0**
humanoidmaze-medium	63.6	0.0	**0.0**	57.2	0.0	**0.8**	0.4	0.0	**0.0**	74.4	0.0	**0.4**
humanoidmaze-large	3.2	0.0	**0.0**	18.0	0.0	**0.0**	0.0	0.0	**0.0**	0.0	0.0	**0.0**
cube-single	96.8	82.0	**85.2**	49.6	35.6	**39.6**	54.4	47.6	**49.6**	98.0	88.0	**92.4**
cube-double	25.2	12.0	**14.4**	0.8	1.6	**2.4**	1.2	0.0	**0.0**	48.8	24.8	**28.8**
scene	56.0	48.0	**49.2**	2.4	1.2	**1.2**	36.0	35.6	**38.4**	49.2	39.2	**42.0**
puzzle-3x3	27.2	22.8	**25.2**	14.0	9.6	**12.4**	26.0	21.6	**23.6**	21.2	17.2	**19.2**
puzzle-4x4	13.2	0.0	**0.0**	17.2	0.4	**1.2**	2.0	0.0	**0.0**	37.2	1.6	**8.8**
D4RL												
AntMaze	83.7	62.7	**69.0**	62.7	19.7	**21.3**	36.0	25.7	**27.0**	94.7	63.7	**67.7**
Adroit	52.3	43.1	**47.4**	1.4	0.1	**0.5**	55.6	41.5	**44.7**	59.4	44.8	**48.0**

lower victim scores than both variants in 12 of 15 comparisons and ties them in the remaining three. All five GFP tasks benefit from combining the two signals.

4.3 Defense Effectiveness

We examine whether corrections learned from uniform random perturbations improve DMBP under the evaluated attacks and across victim policies.

Improvement beyond DMBP. Tables 1 and 2 show successive recovery under BDAP. On FQL antmaze-large, DMBP raises success from 11.2% to 42.8%, still below the raw-clean 80.8%. RCF with $\rho = 0.3$ further raises success to 54.4%. On GFP antmaze-large, it improves DMBP's 72.4% to 82.8%. Across all 200 OGBench victim–task entries, RCF raises pooled success from 15.2% to 17.1%, with gains also appearing on cube and puzzle tasks.

Transfer across policy families. On OGBench cube-single, the FQL-trained corrector increases success from 35.6% to 39.6% for MeanFlowQL and from 47.6% to 49.6% for SSCP. On D4RL, both AntMaze success and Adroit normalized return increase for all four policy families. For example,

Adroit scores rise from 41.5 to 44.7 for SSCP and from 44.8 to 48.0 for GFP, showing gains beyond the collection policy.

Generalization across perturbations. Correctors trained with uniform random perturbations are evaluated under all four perturbation types: BDAP, APGD, min-Q, and Gaussian noise (Appendix C.3). Pooled OGBench success changes from 19.5% to 22.4% under APGD, from 19.3% to 21.7% under min-Q, and from 23.1% to 25.4% under Gaussian noise. These gains demonstrate that residual correction learned from random perturbations generalizes to the evaluated attacks and Gaussian noise.

4.4 Robustness Across Attack Budgets

We next examine whether RCF's benefit persists as the attack budget increases. Figure 2 compares DMBP and DMBP+RCF with fixed $\rho = 0.3$ at $\epsilon \in \{0.1, 0.2, 0.3\}$. Success rates use the $[0, 1]$ scale.

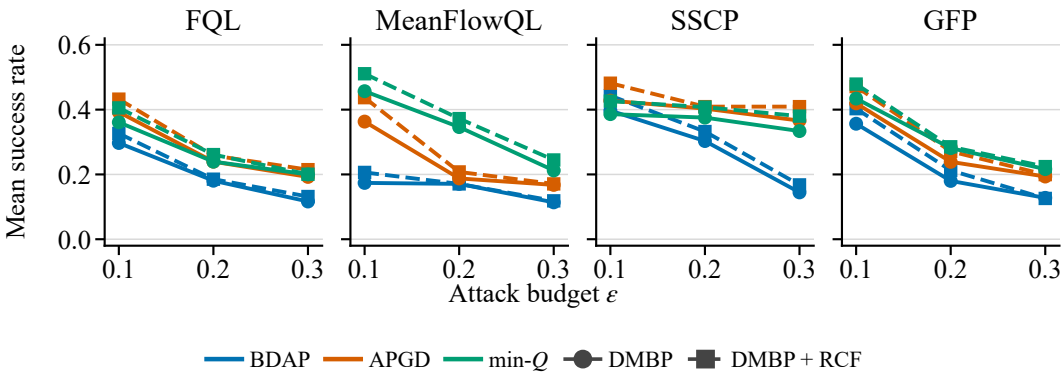


Figure 2: OGBench success rates across attack budgets on fixed task sets. Solid and dashed lines denote DMBP and DMBP+RCF ($\rho = 0.3$), respectively. Numerical results appear in Table 20.

RCF generally retains its benefit as the attack budget increases, with the dashed curves remaining above the corresponding DMBP curves in most settings. For FQL and SSCP, RCF improves success under all three attacks at every tested budget, including $\epsilon = 0.3$. The improvements under APGD and min-Q also persist across budgets for MeanFlowQL and GFP. Thus, the benefit of residual correction extends beyond the lowest attack budget, with the same fixed coefficient $\rho = 0.3$ supporting improved defense across the tested budget range.

4.5 Additional Analyses

Effect of correction strength. Appendix C.4 compares $\rho \in \{0.3, 0.5, 1.0\}$. Among these settings, $\rho = 0.3$ gives the highest pooled success under all three attacks at $\epsilon = 0.1$ and under Gaussian noise. Stronger correction can help at larger budgets. Under BDAP at $\epsilon = 0.3$, increasing ρ from 0.3 to 1.0 raises pooled success from 0.133 to 0.153. The same change under Gaussian noise reduces success from 0.520 to 0.401. The sweep therefore favors stronger correction for some larger-budget attacks and weaker correction for Gaussian noise.

5 Conclusion

BDAP reveals transferable vulnerabilities in one-step flow policies without victim queries or gradients at deployment. DMBP partially restores performance under observation perturbations, and RCF strengthens this defense by predicting and correcting its reconstruction residuals. RCF learns only from uniformly perturbed simulator rollouts, keeping the policy and DMBP frozen and requiring no attack-type information at deployment. Our analysis bounds reconstruction error using prediction loss, target clipping, and correction constraints, then relates this error to a return-gap bound under Lipschitz conditions. Experiments on OGBench and D4RL demonstrate effective query-free transfer attacks and further performance recovery over DMBP under adversarial and random perturbations.

AI USE STATEMENT

Generative AI tools were used to improve the clarity and readability of the manuscript. They were not used to formulate the research problem, develop the proposed methodology, derive theoretical results, conduct experiments, or interpret the experimental results. All AI-assisted edits were reviewed and verified by the authors. The authors take full responsibility for the final content of the paper.

REFERENCES

Yipu Chen, Haotian Xue, and Yongxin Chen. Diffusion policy attacker: Crafting adversarial attacks for diffusion-based policies. *Advances in Neural Information Processing Systems*, 37:119614–119637, 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.

Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International conference on machine learning*, pp. 2206–2216. PMLR, 2020.

Zihan Ding and Chi Jin. Consistency models as a rich and efficient policy class for reinforcement learning. In *International Conference on Learning Representations*, volume 2024, pp. 53047–53066, 2024.

Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9185–9193. IEEE, 2018.

Justin Fu, Aviral Kumar, Ofir Nachum, George Tucker, and Sergey Levine. D4rl: Datasets for deep data-driven reinforcement learning. *arXiv preprint arXiv:2004.07219*, 2020.

Scott Fujimoto and Shixiang Shane Gu. A minimalist approach to offline reinforcement learning. *Advances in neural information processing systems*, 34:20132–20145, 2021.

Scott Fujimoto, David Meger, and Doina Precup. Off-policy deep reinforcement learning without exploration. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2052–2062. PMLR, 09–15 Jun 2019.

Peyré Gabriel and Cuturi Marco. Computational optimal transport with applications to data sciences. *Foundations and trends® in machine learning*, 11(5-6):355–607, 2019.

Zhengyang Geng, Mingyang Deng, Xingjian Bai, Zico Kolter, and Kaiming He. Mean flows for one-step generative modeling. *Advances in Neural Information Processing Systems*, 38:75460–75482, 2026.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *Proceedings of the nineteenth international conference on machine learning*, pp. 267–274, 2002.

Bingyi Kang, Xiao Ma, Chao Du, Tianyu Pang, and Shuicheng Yan. Efficient diffusion policies for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 36:67195–67212, 2023.

Prajwal Koirala and Cody Fleming. Flow-based single-step completion for efficient and expressive policy learning. In *International Conference on Learning Representations*, volume 2026, pp. 145668–145701, 2026.

Ilya Kostrikov, Ashvin Nair, and Sergey Levine. Offline reinforcement learning with implicit q-learning. *arXiv preprint arXiv:2110.06169*, 2021.

Aviral Kumar, Justin Fu, Matthew Soh, George Tucker, and Sergey Levine. Stabilizing off-policy q-learning via bootstrapping error reduction. *Advances in neural information processing systems*, 32, 2019.

Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative q-learning for offline reinforcement learning. *Advances in neural information processing systems*, 33:1179–1191, 2020.

Aviral Kumar, Anikait Singh, Frederik Ebert, Mitsuhiko Nakamoto, Yanlai Yang, Chelsea Finn, and Sergey Levine. Pre-training for robots: Offline rl enables learning new tasks from a handful of trials. *arXiv preprint arXiv:2210.05178*, 2022.

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.

Yongyuan Liang, Yanchao Sun, Ruijie Zheng, and Furong Huang. Efficient adversarial training without attacking: Worst-case-aware robust reinforcement learning. *Advances in neural information processing systems*, 35:22547–22561, 2022.

Yen-Chen Lin, Zhang-Wei Hong, Yuan-Hong Liao, Meng-Li Shih, Ming-Yu Liu, and Min Sun. Tactics of adversarial attack on deep reinforcement learning agents. *arXiv preprint arXiv:1703.06748*, 2017.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Franki Nguimatsia Tiofack, Théotime Le Hellard, Fabian Schramm, Nicolas Perrin-Gilbert, and Justin Carpentier. Guided flow policy: Learning from high-value actions in offline reinforcement learning. In *International Conference on Learning Representations*, volume 2026, pp. 30601–30625, 2026.

Weili Nie, Brandon Guo, Yujia Huang, Chaowei Xiao, Arash Vahdat, and Anima Anandkumar. Diffusion models for adversarial purification. *arXiv preprint arXiv:2205.07460*, 2022.

Tuomas Oikarinen, Wang Zhang, Alexandre Megretski, Luca Daniel, and Tsui-Wei Weng. Robust deep reinforcement learning through adversarial loss. *Advances in Neural Information Processing Systems*, 34:26156–26167, 2021.

Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia conference on computer and communications security*, pp. 506–519, 2017.

Seohong Park, Kevin Frans, Benjamin Eysenbach, and Sergey Levine. Ogbench: Benchmarking offline goal-conditioned rl. In *International Conference on Learning Representations*, volume 2025, pp. 94937–94982, 2025a.

Seohong Park, Qiyang Li, and Sergey Levine. Flow q-learning. *arXiv preprint arXiv:2502.02538*, 2025b.

Pouya Samangouei, Maya Kabkab, and Rama Chellappa. Defense-gan: Protecting classifiers against adversarial attacks using generative models. *arXiv preprint arXiv:1805.06605*, 2018.

Zeyuan Wang, Da Li, Yulin Chen, Ye Shi, Liang Bai, Tianyuan Yu, and Yanwei Fu. One-step generative policies with q-learning: A reformulation of meanflow. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 26751–26759, 2026.

Zhendong Wang, Jonathan J Hunt, and Mingyuan Zhou. Diffusion policies as an expressive policy class for offline reinforcement learning. *arXiv preprint arXiv:2208.06193*, 2022.

Rui Yang, Chenjia Bai, Xiaoteng Ma, Zhaoran Wang, Chongjie Zhang, and Lei Han. Rorl: Robust offline reinforcement learning via conservative smoothing. *Advances in neural information processing systems*, 35:23851–23866, 2022.

Zhihe Yang and Yunjian Xu. Dmbp: Diffusion model-based predictor for robust offline reinforcement learning against state observation perturbations. In *International Conference on Learning Representations*, volume 2024, pp. 38549–38578, 2024.

Huan Zhang, Hongge Chen, Chaowei Xiao, Bo Li, Mingyan Liu, Duane Boning, and Cho-Jui Hsieh. Robust deep reinforcement learning against adversarial perturbations on state observations. *Advances in neural information processing systems*, 33:21024–21037, 2020.

Huan Zhang, Hongge Chen, Duane Boning, and Cho-Jui Hsieh. Robust reinforcement learning on state observations with learned optimal adversary. *arXiv preprint arXiv:2101.08452*, 2021.

A PROOFS

A.1 PROOF OF THEOREM 1 (RECONSTRUCTION ERROR BOUND)

Proof. All expectations are under $\mathcal{Q}$. We use the standardized residual y^{raw} and correction $c_\sigma(\varphi) = \Pi_{\mathcal{C}}(f(\varphi))$ defined in the main text. Projection onto $\mathcal{C}$ sets masked coordinates to zero and clips the others to $[-\kappa, \kappa]$. Consequently, the constraint residual in (16) has the coordinate expansion

$$\mathcal{L}_{\mathcal{C}} = \frac{1}{d}\mathbb{E}_{\mathcal{Q}}\left[\sum_{i:m_i=0}(y_i^{\mathrm{raw}})^2 + \sum_{i:m_i=1}(|y_i^{\mathrm{raw}}| - \kappa)_+^2\right],$$

where $(q)_+ = \max(q, 0)$. The first sum is the residual energy in uncorrected coordinates, and the second is the excess beyond the correction limit in selected coordinates.

Decomposing $f(\varphi) - y^{\mathrm{raw}} = (f(\varphi) - y^{\mathrm{tr}}) + (y^{\mathrm{tr}} - y^{\mathrm{raw}})$ separates the prediction error from the target-clipping error. Expanding the squared norm and taking expectations gives

$$\begin{aligned}\frac{1}{d}\mathbb{E}\|f(\varphi) - y^{\mathrm{raw}}\|_2^2 &= \frac{1}{d}\mathbb{E}\|f(\varphi) - y^{\mathrm{tr}}\|_2^2 + \frac{1}{d}\mathbb{E}\|y^{\mathrm{tr}} - y^{\mathrm{raw}}\|_2^2 + \frac{2}{d}\mathbb{E}\langle f(\varphi) - y^{\mathrm{tr}}, y^{\mathrm{tr}} - y^{\mathrm{raw}}\rangle \\ &= \mathcal{L}_{\mathcal{Q}}(f) + \mathcal{L}_{\mathrm{clip}} + \frac{2}{d}\mathbb{E}\langle f(\varphi) - y^{\mathrm{tr}}, y^{\mathrm{tr}} - y^{\mathrm{raw}}\rangle.\end{aligned}$$

The Cauchy–Schwarz inequality bounds the cross term by

$$\frac{1}{d}\mathbb{E}\langle f(\varphi) - y^{\mathrm{tr}}, y^{\mathrm{tr}} - y^{\mathrm{raw}}\rangle \le \left(\frac{1}{d}\mathbb{E}\|f(\varphi) - y^{\mathrm{tr}}\|_2^2\right)^{1/2}\left(\frac{1}{d}\mathbb{E}\|y^{\mathrm{tr}} - y^{\mathrm{raw}}\|_2^2\right)^{1/2} = \sqrt{\mathcal{L}_{\mathcal{Q}}(f)\mathcal{L}_{\mathrm{clip}}}.$$

Substitution yields $d^{-1}\mathbb{E}\|f(\varphi) - y^{\mathrm{raw}}\|_2^2 \le \mathcal{L}_{\mathcal{Q}}(f) + \mathcal{L}_{\mathrm{clip}} + 2\sqrt{\mathcal{L}_{\mathcal{Q}}(f)\mathcal{L}_{\mathrm{clip}}} = (\sqrt{\mathcal{L}_{\mathcal{Q}}(f)} + \sqrt{\mathcal{L}_{\mathrm{clip}}})^2$. Taking nonnegative square roots gives

$$\left(\frac{1}{d}\mathbb{E}\|f(\varphi) - y^{\mathrm{raw}}\|_2^2\right)^{1/2} \le \sqrt{\mathcal{L}_{\mathcal{Q}}(f)} + \sqrt{\mathcal{L}_{\mathrm{clip}}}. \tag{19}$$

For the projected correction, let $y_{\mathcal{C}} = \Pi_{\mathcal{C}}(y^{\mathrm{raw}})$ and $\Delta_{\mathrm{proj}} = c_\sigma(\varphi) - y_{\mathcal{C}}$ for each realization of $(y^{\mathrm{raw}}, \varphi)$. The optimality condition for projection onto the closed convex set $\mathcal{C}$ gives $\langle f(\varphi) - c_\sigma(\varphi), y_{\mathcal{C}} - c_\sigma(\varphi)\rangle \le 0$, or equivalently $\langle c_\sigma(\varphi) - f(\varphi), \Delta_{\mathrm{proj}}\rangle \le 0$. Hence

$$\begin{aligned}\|y^{\mathrm{raw}} - c_\sigma(\varphi)\|_2^2 - \|y^{\mathrm{raw}} - y_{\mathcal{C}}\|_2^2 &= 2\langle c_\sigma(\varphi) - y^{\mathrm{raw}}, \Delta_{\mathrm{proj}}\rangle - \|\Delta_{\mathrm{proj}}\|_2^2 \\ &\le 2\langle f(\varphi) - y^{\mathrm{raw}}, \Delta_{\mathrm{proj}}\rangle - \|\Delta_{\mathrm{proj}}\|_2^2 \\ &= \|f(\varphi) - y^{\mathrm{raw}}\|_2^2 - \|f(\varphi) - y^{\mathrm{raw}} - \Delta_{\mathrm{proj}}\|_2^2 \\ &\le \|f(\varphi) - y^{\mathrm{raw}}\|_2^2.\end{aligned} \tag{20}$$

Rearranging (20), taking expectations, and dividing by d yields

$$\frac{1}{d}\mathbb{E}\|y^{\mathrm{raw}} - c_\sigma(\varphi)\|_2^2 \le \frac{1}{d}\mathbb{E}\|y^{\mathrm{raw}} - y_{\mathcal{C}}\|_2^2 + \frac{1}{d}\mathbb{E}\|f(\varphi) - y^{\mathrm{raw}}\|_2^2.$$

The left side is $\mathcal{R}_1(f)$ by (14). Since $y_{\mathcal{C}} = \Pi_{\mathcal{C}}(y^{\mathrm{raw}})$, the first term on the right is $d^{-1}\mathbb{E}\|y^{\mathrm{raw}} - y_{\mathcal{C}}\|_2^2 = \mathcal{L}_{\mathcal{C}}$ by (16). Squaring the nonnegative sides of (19) bounds the remaining prediction term, giving

$$\mathcal{R}_1(f) \le \mathcal{L}_{\mathcal{C}} + \frac{1}{d}\mathbb{E}\|f(\varphi) - y^{\mathrm{raw}}\|_2^2 \le \mathcal{L}_{\mathcal{C}} + \left(\sqrt{\mathcal{L}_{\mathcal{Q}}(f)} + \sqrt{\mathcal{L}_{\mathrm{clip}}}\right)^2. \tag{21}$$

Finally, $y^{\mathrm{raw}} - \rho c_\sigma(\varphi) = (1-\rho)y^{\mathrm{raw}} + \rho(y^{\mathrm{raw}} - c_\sigma(\varphi))$ implies

$$\begin{aligned}\mathcal{R}_\rho(f) &= (1-\rho)\mathcal{R}_0 + \rho\mathcal{R}_1(f) - \frac{\rho(1-\rho)}{d}\mathbb{E}\|c_\sigma(\varphi)\|_2^2 \\ &\le (1-\rho)\mathcal{R}_0 + \rho\mathcal{R}_1(f),\end{aligned} \tag{22}$$

where the inequality uses $0 \le \rho \le 1$. Substituting (21) proves (17). □

A.2 Proof of Proposition 1 (return gap under reconstruction error)

Let $x = (s, h, a_-, t, \xi)$ denote the augmented state, containing the true state, filter history, previous action, timestep, and exogenous attack and denoiser randomness. The policy sampling randomness Z is drawn independently of x. The attack may use this draw, so we allow the defended input $s_{\mathrm{def}}(x, Z)$ to depend on both. We define expected discounted return and the normalized discounted state occupancy measure (Kakade & Langford, 2002) by

$$J(\pi) := \mathbb{E}_\pi\left[\sum_{t=0}^{\infty}\gamma^t R_t\right], \qquad d^{\bar\pi} := (1-\gamma)\sum_{t=0}^{\infty}\gamma^t \,\mathrm{Law}_{\bar\pi}(x_t). \tag{23}$$

Here R_t is the timestep reward and $\mathrm{Law}_{\bar\pi}(x_t)$ is the distribution of x_t under the defended system. The weights $(1-\gamma)\gamma^t$ sum to one. The return $J(\bar\pi)$ uses the same definition under defended trajectories.

Proof. Let P_s and $\bar P_x$ denote the action distributions induced by $\pi(s, Z)$ and $\pi(s_{\mathrm{def}}(x, Z), Z)$, respectively. Consider a trajectory generated by the defended system, with $a_t \sim \bar P_{x_t}$ and $s_{t+1} \sim P(\cdot \mid s_t, a_t)$. Observation corruption and correction leave the physical transition kernel and reward function unchanged. Define the clean-policy advantage as

$$A^\pi(s, a) := Q^\pi(s, a) - \mathbb{E}_{a'\sim P_s} Q^\pi(s, a').$$

This compares the value of action a with the mean value of actions $a' \sim P_s$ selected by the clean policy. Since $R(s, a)$ is the expected immediate reward, Q^π satisfies the Bellman equation

$$Q^\pi(s, a) = R(s, a) + \gamma \mathbb{E}_{s'\sim P(\cdot|s,a)} \mathbb{E}_{a'\sim P_{s'}} Q^\pi(s', a').$$

Along the defended trajectory, $\mathbb{E}_{\bar\pi}[R_t \mid x_t, a_t] = R(s_t, a_t)$. Applying the Bellman equation at (s_t, a_t) expresses the advantage in terms of the immediate reward and the clean-policy mean action values at consecutive states. Taking expectations over defended trajectories and summing with weights γ^t from $t = 0$ to T gives

$$\begin{aligned}
&\mathbb{E}_{\bar\pi}\sum_{t=0}^{T}\gamma^t A^\pi(s_t, a_t) \\
&= \mathbb{E}_{\bar\pi}\sum_{t=0}^{T}\gamma^t\left[R_t + \gamma \mathbb{E}_{a'\sim P_{s_{t+1}}} Q^\pi(s_{t+1}, a') - \mathbb{E}_{a'\sim P_{s_t}} Q^\pi(s_t, a')\right] \\
&= \mathbb{E}_{\bar\pi}\sum_{t=0}^{T}\gamma^t R_t - \mathbb{E}_{\bar\pi}\mathbb{E}_{a'\sim P_{s_0}} Q^\pi(s_0, a') + \gamma^{T+1}\mathbb{E}_{\bar\pi}\mathbb{E}_{a'\sim P_{s_{T+1}}} Q^\pi(s_{T+1}, a').
\end{aligned}$$

The mean action values at timesteps $1, \ldots, T$ cancel, since each appears once with a positive weight and once with the same negative weight. By the return definition in (23),

$$\lim_{T\to\infty} \mathbb{E}_{\bar\pi}\sum_{t=0}^{T}\gamma^t R_t = J(\bar\pi).$$

At the initial state, averaging $Q^\pi(s_0, a')$ over the clean action distribution gives the expected return obtained by following π from s_0. Since both systems share the same initial physical-state distribution,

$$\mathbb{E}_{\bar\pi}\mathbb{E}_{a'\sim P_{s_0}} Q^\pi(s_0, a') = \mathbb{E}_\pi \mathbb{E}_{a'\sim P_{s_0}} Q^\pi(s_0, a') = J(\pi).$$

Bounded rewards and $\gamma < 1$ imply that Q^π and its clean-policy action average are bounded, so the final discounted value term vanishes as $T \to \infty$. Taking the limit gives the performance-difference identity (Kakade & Langford, 2002),

$$J(\bar\pi) - J(\pi) = \mathbb{E}_{\bar\pi}\sum_{t=0}^{\infty}\gamma^t A^\pi(s_t, a_t).$$

Conditioning on x_t replaces each expected advantage along the trajectory by $\mathbb{E}_{a\sim\bar P_{x_t}} A^\pi(s_t, a)$. The normalized discounted state occupancy measure in (23) then combines the state distributions across timesteps, giving

$$J(\bar\pi) - J(\pi) = \frac{1}{1-\gamma}\mathbb{E}_{x\sim d^{\bar\pi}}\mathbb{E}_{a\sim\bar P_x} A^\pi(s, a).$$

To bound the expected advantage, fix x and let $\Pi(\bar{P}_x, P_s)$ denote the set of joint action distributions with marginals $\bar{P}_x$ and P_s, using the optimal-transport coupling definition (Gabriel & Marco, 2019). For $p \in \{1, 2\}$, the Wasserstein distance with the Euclidean action metric is $W_p(\bar{P}_x, P_s) = \inf_{\Gamma \in \Pi(\bar{P}_x, P_s)} (\mathbb{E}_\Gamma \|a - a'\|_2^p)^{1/p}$. Since $\mathbb{E}_{a' \sim P_s} A^\pi(s, a') = 0$, any coupling $\Gamma \in \Pi(\bar{P}_x, P_s)$ satisfies

$$\begin{aligned} \left|\mathbb{E}_{a \sim \bar{P}_x} A^\pi(s, a)\right| &= \left|\mathbb{E}_{(a,a') \sim \Gamma}\left[A^\pi(s, a) - A^\pi(s, a')\right]\right| \\ &= \left|\mathbb{E}_{(a,a') \sim \Gamma}\left[Q^\pi(s, a) - Q^\pi(s, a')\right]\right| \\ &\le \mathbb{E}_{(a,a') \sim \Gamma} \left|Q^\pi(s, a) - Q^\pi(s, a')\right| \\ &\le L_Q \mathbb{E}_{(a,a') \sim \Gamma} \|a - a'\|_2. \end{aligned}$$

The last inequality uses the action-value Lipschitz condition. Taking the infimum over Γ gives $\left|\mathbb{E}_{a \sim \bar{P}_x} A^\pi(s, a)\right| \le L_Q W_1(\bar{P}_x, P_s)$. Applying the triangle inequality to the preceding return identity therefore yields

$$|J(\bar{\pi}) - J(\pi)| \le \frac{1}{1-\gamma} \mathbb{E}_{x \sim d^{\bar{\pi}}} \left|\mathbb{E}_{a \sim \bar{P}_x} A^\pi(s, a)\right| \le \frac{L_Q}{1-\gamma} \mathbb{E}_{x \sim d^{\bar{\pi}}} W_1(\bar{P}_x, P_s).$$

We bound this action-distribution distance using shared policy sampling randomness. Define the root-mean-square action difference

$$\mathfrak{R}_\pi(x) := \left(\mathbb{E}_Z \|\pi(s_{\text{def}}(x, Z), Z) - \pi(s, Z)\|_2^2\right)^{1/2}. \tag{24}$$

The bounded action space ensures finite second moments. Cauchy–Schwarz gives, for every $\Gamma \in \Pi(\bar{P}_x, P_s)$,

$$W_1(\bar{P}_x, P_s) \le \mathbb{E}_\Gamma \|a - a'\|_2 \le \left(\mathbb{E}_\Gamma \|a - a'\|_2^2\right)^{1/2}.$$

Taking the infimum of the last expression over Γ yields $W_1(\bar{P}_x, P_s) \le W_2(\bar{P}_x, P_s)$. Using the same Z for both actions gives the coupling induced by $(\pi(s_{\text{def}}(x, Z), Z), \pi(s, Z))$. By the definition of W_2 and (24),

$$W_2(\bar{P}_x, P_s) \le \left(\mathbb{E}_Z \|\pi(s_{\text{def}}(x, Z), Z) - \pi(s, Z)\|_2^2\right)^{1/2} = \mathfrak{R}_\pi(x).$$

Combining these bounds gives

$$|J(\bar{\pi}) - J(\pi)| \le \frac{L_Q}{1-\gamma} \mathbb{E}_{x \sim d^{\bar{\pi}}} W_1(\bar{P}_x, P_s) \le \frac{L_Q}{1-\gamma} \mathbb{E}_{x \sim d^{\bar{\pi}}} \mathfrak{R}_\pi(x). \tag{25}$$

The policy Lipschitz condition now relates this action difference to the remaining reconstruction error. Applying it to the defended observation and the true state gives, for almost every z,

$$\|\pi(s_{\text{def}}(x, z), z) - \pi(s, z)\|_2 \le L_\pi \left\|\text{diag}(\sigma)^{-1}\left(s_{\text{def}}(x, z) - s\right)\right\|_2.$$

Squaring, averaging over Z, and taking square roots gives $\mathfrak{R}_\pi(x) \le L_\pi \left(\mathbb{E}_Z \|\,\text{diag}(\sigma)^{-1}(s_{\text{def}}(x, Z) - s)\|_2^2\right)^{1/2}$. Substituting into (25) and applying Jensen's inequality to the square root under the probability measure $d^{\bar{\pi}}$ yields

$$\begin{aligned} |J(\bar{\pi}) - J(\pi)| &\le \frac{L_Q L_\pi}{1-\gamma} \mathbb{E}_{x \sim d^{\bar{\pi}}} \left(\mathbb{E}_Z \left\|\text{diag}(\sigma)^{-1}\left(s_{\text{def}}(x, Z) - s\right)\right\|_2^2\right)^{1/2} \\ &\le \frac{L_Q L_\pi}{1-\gamma} \left(\mathbb{E}_{x \sim d^{\bar{\pi}},\, Z} \left\|\text{diag}(\sigma)^{-1}\left(s_{\text{def}}(x, Z) - s\right)\right\|_2^2\right)^{1/2}, \end{aligned}$$

Since $(s_{\text{def}} - s)/\sigma = \rho c_\sigma(\varphi) - y^{\text{raw}}$, the joint distribution $\mathcal{Q}$ specified in the proposition gives

$$\mathbb{E}_{x \sim d^{\bar{\pi}},\, Z} \left\|\text{diag}(\sigma)^{-1}\left(s_{\text{def}}(x, Z) - s\right)\right\|_2^2 = \mathbb{E}_{\mathcal{Q}} \|y^{\text{raw}} - \rho c_\sigma(\varphi)\|_2^2 = d\mathcal{R}_\rho(f).$$

The factor d follows from the per-coordinate averaging in (14). Substitution proves (18). □

A.3 Derivation of the BDAP generator gradient

Fix update k and the sampled observations s_i and latents z_{ij}. The weights $\lambda_{Q,k}$ and $\lambda_{\mathrm{B},k}$ are constants for this differentiation. In particular, the stop-gradient operation in (28) prevents differentiation through the batch-dependent value weight. The reference actions a_{ij}^{BC} are independent of χ because their inputs and network parameters are fixed. We use the definitions of D_i, A_{ij}, and Δ_{ij} from Section 3.1, with all derivatives below evaluated at $\chi = \chi_k$ wherever the displayed derivatives exist and $\Delta_{ij} \neq 0$.

Since $\tilde{s}_i(\chi) = s_i + \delta_\chi(s_i)$ and $a_{ij}(\chi) = \pi_{\mathrm{sur}}(\tilde{s}_i(\chi), z_{ij})$, the action Jacobian is

$$J_\chi a_{ij}(\chi)\big|_{\chi=\chi_k} = A_{ij} D_i.$$

The critic receives the fixed true state s_i, so its dependence on χ is entirely through the induced action. Applying the chain rule to the batch value gives

$$\nabla_\chi \widehat{q}_k(\chi)\big|_{\chi=\chi_k} = \frac{1}{Bn_z} \sum_{i=1}^{B} \sum_{j=1}^{n_z} D_i^\top A_{ij}^\top \nabla_a Q_\psi\big(s_i, a_{ij}(\chi_k)\big).$$

For the behavior term, differentiating the Euclidean norm with respect to its action argument yields

$$\nabla_a \|a - a_{ij}^{\mathrm{BC}}\|_2\big|_{a=a_{ij}(\chi_k)} = \left.\frac{a - a_{ij}^{\mathrm{BC}}}{\sqrt{(a - a_{ij}^{\mathrm{BC}})^\top (a - a_{ij}^{\mathrm{BC}})}}\right|_{a=a_{ij}(\chi_k)} = \frac{\Delta_{ij}}{\|\Delta_{ij}\|_2}.$$

The reference contributes no parameter derivative. Propagating this action gradient through the surrogate actor and generator therefore gives

$$\nabla_\chi \widehat{d}_k(\chi)\Big|_{\chi=\chi_k} = \frac{1}{Bn_z} \sum_{i=1}^{B} \sum_{j=1}^{n_z} D_i^\top A_{ij}^\top \frac{\Delta_{ij}}{\|\Delta_{ij}\|_2}.$$

Combining the two derivatives with weights $\lambda_{Q,k}$ and $-\lambda_{\mathrm{B},k}$ gives G_k in (10). The implementation computes these products by reverse-mode automatic differentiation of the batch loss with respect to χ, without constructing the full Jacobian matrices. At $\Delta_{ij} = 0$, the Euclidean norm is nondifferentiable. A zero-subgradient extension would set this sample's behavior-channel contribution to zero while retaining its critic contribution.

B Experimental Details

B.1 Attack Baselines

Both baselines search over $\mathcal{B}(s)$ in (3) at each attacked timestep, using the victim's action interface. Their perturbations are constructed before filtering in defense evaluations, so the search objectives use the unfiltered policy response. DMBP and RCF process the selected observation after the search.

APGD. We implement an Auto-PGD-style projected gradient search (Croce & Hein, 2020) with the action-deviation objective

$$\max_{\tilde{s} \in \mathcal{B}(s)} \|\pi(\tilde{s}, z) - \pi(s, z)\|_2^2 . \tag{26}$$

The clean action is computed once and held fixed. Policy sampling randomness is also fixed throughout the search, and gradients are taken with respect to the perturbed observation through the clipped action interface. We initialize uniformly within the perturbation box and perform 10 updates. The initial coordinatewise step size is $2\epsilon\sigma$. Each update takes a projected gradient-sign step, blends its displacement with the previous iterate displacement using weights 0.75 and 0.25, and projects again onto $\mathcal{B}(s)$. Every two updates, the step size is halved and the next iterate is reset to the best candidate if fewer than 75% of the updates improved the best objective value. The search returns the highest-scoring candidate, including the initialization.

Min-Q. Following the sampling attack evaluated by Yang & Xu (2024), we draw 20 independent candidate perturbations with coordinatewise samples $\delta_i^{(k)} \sim \mathcal{U}[-\epsilon\sigma_i, \epsilon\sigma_i]$. Each candidate observation $\tilde{s}^{(k)} = s + \delta^{(k)}$ is passed through the victim sampler to produce an action $a^{(k)}$. Let Q_{vic} denote the mean of the victim's frozen target-critic heads, including its observation preprocessing. We select

$$k^\star \in \arg\min_{1 \le k \le 20} Q_{\text{vic}}(s, a^{(k)}), \qquad \tilde{s} = s + \delta^{(k^\star)}. \tag{27}$$

The critic receives the true state because the environment executes the induced action from that state. This search uses victim action samples and critic scores, but no gradients.

Gaussian noise. We use independent zero-mean Gaussian perturbations with coordinatewise standard deviation $\epsilon\sigma_i/\sqrt{3}$ and scale $\epsilon = 0.2$. The factor $1/\sqrt{3}$ matches the variance of uniform noise on $[-\epsilon\sigma_i, \epsilon\sigma_i]$ at the same scale. Unlike the bounded attacks, Gaussian noise has no hard ℓ_∞ constraint. Task-level results appear in Appendices C.1 and C.3.

B.2 Training and Evaluation Protocol

Hardware. The experiments use a compute instance with one NVIDIA RTX PRO 6000 GPU with 96 GB of GPU memory, 22 vCPUs on an Intel Xeon Platinum 8470Q processor, and 110 GB of RAM.

BDAP generator training. The bounded generator g_χ in (6) has three hidden layers of width 256, with GELU activations, layer normalization, and an elementwise $\tanh$ output. For each task and perturbation budget, we train a separate generator against the frozen FQL surrogate for 200,000 updates. Each batch contains $B = 256$ observations with $n_z = 8$ Gaussian latent samples per observation, shared by the surrogate and behavior-only actors. The batch loss uses

$$\lambda_{Q,k} := \frac{1}{|\operatorname{sg}[\widehat{q}_k(\chi_k)]| + c_Q}, \qquad \lambda_{\text{B},k} := \min(1, k/K_{\text{w}}), \tag{28}$$

where sg denotes stop-gradient, $c_Q = 10^{-3}$ stabilizes value normalization, and $K_{\text{w}} = 50{,}000$ sets the behavior-weight warm-up period. Adam uses learning rate 3×10^{-4} with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and numerical stability constant 10^{-8}. Only the generator is updated. The surrogate actor, its two critic heads, and the behavior-only reference remain fixed. The resulting generator is reused on MeanFlowQL, SSCP, and GFP without further training.

RCF data collection and fitting. For each task, we collect $(\varphi_t, y_t^{\text{tr}})$ from frozen-FQL rollouts with DMBP enabled and residual correction disabled. All training perturbations are sampled uniformly from the coordinate-normalized ℓ_∞ ball at $\epsilon_{\text{tr}} = 0.4$, without generating BDAP, APGD, or min-Q attacks for predictor training. The predictor has two hidden layers of width 256 with ReLU activations and is trained with Adam using learning rate 10^{-3} and batch size 256. Both benchmarks use the $2d$-dimensional input defined in the main text. On OGBench, targets are clipped to $[-10, 10]$. Training runs for 400 epochs, retaining the checkpoint with the lowest validation MSE. The validation pairs provide the coordinate scores in (12), with $\eta_{\text{fe}} = 10^{-9}$ and threshold $\tau = 0.3$. At deployment, predictions are clipped at $\kappa = 5$. Each task-specific predictor is reused across victim families and perturbations, taking only φ_t as input and requiring no attack-type information.

Evaluation and aggregation. Each evaluation run uses 50 episodes, following Park et al. (2025b), and ends each rollout at environment termination or the task horizon. Raw-clean evaluation feeds unmodified observations directly to the frozen policy. The main OGBench group means average all five tasks. The budget and correction-strength comparisons in Appendix C.4 use the same victim–task combinations. D4RL averages include all tasks.

B.3 BDAP Training and Deployment

We first fit the behavior flow with $\mathcal{L}_{\mathrm{FM}}(\zeta)$ and its one-step reference with $\alpha_{\mathrm{dist}}\mathcal{L}_{\mathrm{dist}}(\omega;\zeta)$ for a positive loss weight α_{dist}, holding the distillation target fixed during each reference-actor update. The resulting π_{BC} is frozen for generator training in Algorithm 1.

Algorithm 1 BDAP generator training.

Require: Offline dataset $\mathcal{D}$, frozen FQL surrogate π_{sur} and critic Q_ψ, frozen behavior reference π_{BC}, initialized generator g_{χ_1}, observation scales σ, budget ϵ, batch size B, latent count n_z, training step count K, and weight settings c_Q, K_{w}.

Initialize the Adam optimizer state at χ_1.
for $k = 1, \dots, K$ **do**
  Sample observations $s_i \sim \mathcal{D}$, $i = 1, \dots, B$.
  Sample independent latents $z_{ij} \sim \mathcal{N}(0, I_{d_a})$, $j = 1, \dots, n_z$, for each s_i.
  $\tilde{s}_i(\chi) = s_i + \epsilon\sigma \odot g_\chi(s_i)$.
  $a_{ij}(\chi) = \pi_{\mathrm{sur}}(\tilde{s}_i(\chi), z_{ij})$, $a_{ij}^{\mathrm{BC}} = \pi_{\mathrm{BC}}(s_i, z_{ij})$.
  $\widehat{q}_k(\chi) = \frac{1}{Bn_z}\sum_{i=1}^{B}\sum_{j=1}^{n_z} Q_\psi(s_i, a_{ij}(\chi))$.
  $\widehat{d}_k(\chi) = \frac{1}{Bn_z}\sum_{i=1}^{B}\sum_{j=1}^{n_z} \|a_{ij}(\chi) - a_{ij}^{\mathrm{BC}}\|_2$.
  $\lambda_{Q,k} = \left(|\operatorname{sg}[\widehat{q}_k(\chi_k)]| + c_Q\right)^{-1}$, $\lambda_{\mathrm{B},k} = \min(1, k/K_{\mathrm{w}})$.
  $\widehat{\mathcal{L}}_k(\chi) = \lambda_{Q,k}\widehat{q}_k(\chi) - \lambda_{\mathrm{B},k}\widehat{d}_k(\chi)$.
  Compute $G_k = \nabla_\chi \widehat{\mathcal{L}}_k(\chi)\Big|_{\chi=\chi_k}$ by automatic differentiation.
  Update χ_k to χ_{k+1} with Adam using G_k.
end for
return trained generator $g_{\chi_{K+1}}$.

At deployment, set $\chi = \chi_{K+1}$. The frozen generator supplies $\tilde{s}_t = s_t + \epsilon\sigma \odot g_\chi(s_t)$ in one forward pass per observation. The surrogate, critic, and behavior reference are no longer needed. The same generator transfers to other victim policies for the same task and budget without further training or victim queries.

B.4 RCF Training and Deployment

We collect frozen-FQL rollouts with DMBP enabled and RCF disabled, using uniform random observation perturbations at budget ϵ_{tr}. Each record provides $\varphi_t = \mathrm{clip}([\tilde{s}_t/\sigma, \hat{s}_t/\sigma], -10, 10)$ and $y_t^{\mathrm{tr}} = \mathrm{clip}((s_t - \hat{s}_t)/\sigma, -10, 10)$. We split these pairs into training and validation data for Algorithm 2.

Algorithm 2 RCF predictor fitting.

Require: Training and validation pairs $(\varphi, y^{\mathrm{tr}})$, initialized predictor f_ϑ, epoch count $E \geq 1$, batch size B, gate threshold τ, and denominator stabilizer $\eta_{\mathrm{fe}} > 0$.

Initialize the Adam optimizer state and set the best validation MSE to $+\infty$.
for $e = 1, \ldots, E$ **do**
  Shuffle the training pairs and divide them into mini-batches of at most B pairs.
  for each mini-batch $\mathcal{B}$ **do**
    $\widehat{\mathcal{L}}_{\mathrm{RCF}}(\vartheta) = \frac{1}{d|\mathcal{B}|} \sum_{(\varphi, y^{\mathrm{tr}}) \in \mathcal{B}} \| f_\vartheta(\varphi) - y^{\mathrm{tr}} \|_2^2$.
    Update ϑ with Adam using $\nabla_\vartheta \widehat{\mathcal{L}}_{\mathrm{RCF}}(\vartheta)$.
  end for
  Compute the mean-squared error of f_ϑ on the validation pairs.
  if the validation MSE is lower than the best recorded value **then**
    Store $\vartheta^\star \leftarrow \vartheta$ and update the best validation MSE.
  end if
end for
Set $f = f_{\vartheta^\star}$ and evaluate its outputs on the validation pairs.
$\mathrm{fe}_i^{\mathrm{val}} = 1 - \frac{\mathbb{E}_{\mathrm{val}}[(f_i(\varphi) - y_i^{\mathrm{tr}})^2]}{\mathbb{E}_{\mathrm{val}}[(y_i^{\mathrm{tr}})^2] + \eta_{\mathrm{fe}}}$, $i = 1, \ldots, d$, using validation averages.
$m_i = \mathbf{1}\{\mathrm{fe}_i^{\mathrm{val}} > \tau\}$, $i = 1, \ldots, d$.
return frozen predictor f and coordinate mask m.

At deployment, DMBP produces $\hat{s}_t = F(\tilde{s}_t, a_{t-1}, h_t, \xi_t)$. Using the same input transformation, RCF computes $c(\varphi_t) = m \odot \mathrm{clip}(f(\varphi_t), -\kappa, \kappa) \odot \sigma$ and supplies $s_{\mathrm{def},t} = \hat{s}_t + \rho c(\varphi_t)$ to the frozen policy. The policy executes $a_t = \pi(s_{\mathrm{def},t}, z_t)$. DMBP appends $\hat{s}_t$ to its history, retaining the most recent N reconstructed observations, and uses a_t as the preceding action at the next timestep.

C Additional Experimental Results

We report per-task results underlying the main tables. Scores are success percentages for OGBench and D4RL AntMaze, and normalized returns for D4RL Adroit. Evaluations follow Section 4, with 50 episodes per seed and task-specific horizons. Entries with $\pm$ report the mean and standard deviation over 5 seeds.

C.1 Per-Task Attack Results

The following tables report results without defense under APGD, min-Q, and BDAP at $\epsilon \in \{0.05, 0.1, 0.2\}$, together with raw-clean results and Gaussian noise at scale $\epsilon = 0.2$. Each OGBench environment includes five distinct evaluation goals, denoted task1 through task5. D4RL tasks use their full environment IDs. The group means in Table 1 average five tasks per OGBench environment, six D4RL AntMaze tasks, or twelve D4RL Adroit tasks.

Table 3: FQL performance on individual OGBench tasks without defense. Scores are success percentages.

Task	Raw clean	Gaussian	APGD			min-Q			BDAP		
		0.2	0.05	0.1	0.2	0.05	0.1	0.2	0.05	0.1	0.2
antmaze-large											
task1	82.0±7.7	80.0±8.4	70.0±10.5	24.0±10.2	0.0±0.0	46.0±12.0	16.0±9.2	0.0±0.0	56.0±12.4	12.0±8.4	0.0±0.0
task2	64.0±9.7	58.0±10.5	60.0±11.4	0.0±0.0	0.0±0.0	38.0±11.8	0.0±0.0	0.0±0.0	42.0±12.5	2.0±3.5	0.0±0.0
task3	96.0±2.4	96.0±2.4	90.0±4.0	82.0±5.5	0.0±0.0	94.0±3.5	56.0±7.3	0.0±0.0	88.0±4.7	42.0±7.5	0.0±0.0
task4	90.0±2.8	88.0±3.5	76.0±4.9	32.0±5.5	0.0±0.0	66.0±5.7	16.0±4.5	0.0±0.0	62.0±6.0	0.0±0.0	0.0±0.0
task5	72.0±4.7	76.0±4.7	80.0±4.9	36.0±6.0	0.0±0.0	38.0±6.2	6.0±3.2	0.0±0.0	76.0±5.7	0.0±0.0	0.0±0.0
antmaze-giant											
task1	24.0±10.9	6.0±6.5	6.0±6.9	0.0±0.0	0.0±0.0	12.0±9.9	0.0±0.0	0.0±0.0	10.0±9.6	0.0±0.0	0.0±0.0
task2	50.0±12.2	14.0±8.9	14.0±9.8	0.0±0.0	0.0±0.0	20.0±11.7	0.0±0.0	0.0±0.0	12.0±9.9	0.0±0.0	0.0±0.0
task3	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task4	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task5	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
antsoccer-arena											
task1	78.0±4.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	6.0±2.8	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task2	88.0±2.8	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task3	56.0±6.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task4	42.0±6.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task5	18.0±7.2	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
humanoidmaze-medium											
task1	24.0±13.1	24.0±13.7	10.0±10.6	0.0±0.0	0.0±0.0	18.0±14.1	2.0±4.5	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task2	98.0±2.0	96.0±2.8	92.0±4.0	0.0±0.0	0.0±0.0	90.0±4.5	60.0±7.7	2.0±2.0	42.0±7.7	0.0±0.0	0.0±0.0
task3	98.0±4.5	96.0±7.9	92.0±12.7	14.0±17.1	0.0±0.0	92.0±13.4	82.0±19.7	10.0±13.5	42.0±25.3	2.0±4.5	0.0±0.0
task4	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	2.0±3.5	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task5	98.0±1.4	100.0±0.0	100.0±0.0	12.0±4.5	0.0±0.0	100.0±0.0	86.0±5.1	4.0±2.4	86.0±5.1	0.0±0.0	0.0±0.0
humanoidmaze-large											
task1	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	4.0±5.5	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task2	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task3	16.0±8.2	20.0±9.4	12.0±8.4	0.0±0.0	0.0±0.0	18.0±10.3	4.0±5.5	0.0±0.0	2.0±3.5	0.0±0.0	0.0±0.0
task4	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	4.0±4.9	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task5	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	4.0±6.9	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
cube-single											
task1	98.0±1.4	68.0±5.7	84.0±4.9	16.0±5.3	4.0±2.4	10.0±4.2	0.0±0.0	0.0±0.0	26.0±6.5	0.0±0.0	0.0±0.0
task2	98.0±1.4	66.0±5.8	86.0±4.7	26.0±6.2	0.0±0.0	20.0±5.5	0.0±0.0	0.0±0.0	28.0±6.6	0.0±0.0	0.0±0.0
task3	100.0±0.0	84.0±5.5	96.0±3.2	8.0±4.7	4.0±3.2	4.0±3.5	0.0±0.0	0.0±0.0	12.0±5.8	0.0±0.0	0.0±0.0
task4	94.0±2.8	72.0±6.0	80.0±5.8	10.0±4.5	0.0±0.0	14.0±5.3	0.0±0.0	0.0±0.0	14.0±5.5	0.0±0.0	0.0±0.0
task5	94.0±2.8	78.0±5.1	88.0±4.5	16.0±5.3	0.0±0.0	28.0±6.3	0.0±0.0	0.0±0.0	20.0±6.0	0.0±0.0	0.0±0.0
cube-double											
task1	52.0±9.2	18.0±7.5	22.0±8.8	2.0±2.8	0.0±0.0	10.0±6.6	0.0±0.0	0.0±0.0	2.0±3.5	0.0±0.0	0.0±0.0
task2	28.0±5.5	10.0±4.0	4.0±2.8	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task3	22.0±4.9	6.0±2.8	8.0±3.7	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task4	12.0±2.8	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task5	12.0±8.4	2.0±3.5	2.0±4.5	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
scene											
task1	100.0±0.0	100.0±0.0	100.0±0.0	86.0±2.8	0.0±0.0	88.0±2.4	14.0±2.8	0.0±0.0	96.0±1.4	22.0±3.5	0.0±0.0
task2	92.0±5.7	80.0±8.9	80.0±9.7	2.0±3.5	0.0±0.0	26.0±11.1	0.0±0.0	0.0±0.0	62.0±12.8	0.0±0.0	0.0±0.0
task3	86.0±2.4	84.0±2.8	90.0±2.4	10.0±2.8	0.0±0.0	10.0±2.8	0.0±0.0	0.0±0.0	24.0±3.7	0.0±0.0	0.0±0.0
task4	2.0±1.4	2.0±1.4	4.0±1.4	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task5	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
puzzle-3x3											
task1	78.0±5.5	92.0±3.7	86.0±5.3	68.0±7.5	4.0±2.8	28.0±7.2	4.0±3.2	0.0±0.0	66.0±7.9	10.0±5.3	0.0±0.0
task2	22.0±5.5	22.0±6.0	28.0±6.9	2.0±2.4	0.0±0.0	12.0±5.3	2.0±2.4	0.0±0.0	4.0±3.2	0.0±0.0	0.0±0.0
task3	4.0±2.0	2.0±1.4	4.0±2.4	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task4	4.0±2.8	6.0±3.5	10.0±4.7	0.0±0.0	0.0±0.0	4.0±3.2	0.0±0.0	0.0±0.0	4.0±3.2	0.0±0.0	0.0±0.0
task5	28.0±3.7	8.0±2.4	22.0±4.0	0.0±0.0	0.0±0.0	10.0±2.8	4.0±2.0	0.0±0.0	2.0±1.4	0.0±0.0	0.0±0.0
puzzle-4x4											
task1	32.0±7.9	38.0±8.6	36.0±9.3	10.0±6.0	0.0±0.0	12.0±6.6	0.0±0.0	0.0±0.0	26.0±9.3	0.0±0.0	0.0±0.0
task2	6.0±3.2	12.0±4.7	8.0±4.2	0.0±0.0	0.0±0.0	2.0±2.4	0.0±0.0	0.0±0.0	8.0±4.5	0.0±0.0	0.0±0.0
task3	18.0±4.9	26.0±6.0	22.0±6.2	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	14.0±5.5	0.0±0.0	0.0±0.0
task4	4.0±2.0	2.0±1.4	6.0±2.8	0.0±0.0	0.0±0.0	6.0±2.8	0.0±0.0	0.0±0.0	4.0±2.4	0.0±0.0	0.0±0.0
task5	6.0±2.8	6.0±2.8	2.0±2.0	0.0±0.0	0.0±0.0	2.0±2.0	0.0±0.0	0.0±0.0	2.0±2.0	0.0±0.0	0.0±0.0

Table 4: MeanFlowQL performance on individual OGBench tasks without defense. Scores are success percentages.

Task	Raw clean	Gaussian	APGD			min-Q			BDAP		
		0.2	0.05	0.1	0.2	0.05	0.1	0.2	0.05	0.1	0.2
antmaze-large											
task1	0.0±0.0	0.0±0.0	2.0±3.5	0.0±0.0	0.0±0.0	2.0±3.5	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task2	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task3	0.0±0.0	0.0±0.0	2.0±2.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task4	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task5	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
antmaze-giant											
task1	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task2	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task3	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task4	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task5	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
antsoccer-arena											
task1	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task2	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task3	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task4	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task5	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
humanoidmaze-medium											
task1	90.0±9.2	98.0±4.5	80.0±14.1	0.0±0.0	0.0±0.0	94.0±8.5	80.0±15.3	10.0±12.4	56.0±18.2	0.0±0.0	0.0±0.0
task2	96.0±2.4	92.0±3.5	94.0±3.5	0.0±0.0	0.0±0.0	96.0±2.8	96.0±3.2	8.0±4.5	76.0±6.5	0.0±0.0	0.0±0.0
task3	0.0±0.0	0.0±0.0	2.0±4.5	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task4	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task5	100.0±0.0	96.0±2.4	96.0±2.8	2.0±2.0	0.0±0.0	98.0±2.0	94.0±3.5	8.0±4.2	84.0±5.3	2.0±2.0	0.0±0.0
humanoidmaze-large											
task1	54.0±11.7	24.0±10.5	14.0±9.4	0.0±0.0	0.0±0.0	42.0±13.3	22.0±12.2	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task2	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task3	6.0±5.3	0.0±0.0	2.0±3.5	0.0±0.0	0.0±0.0	2.0±3.5	6.0±6.6	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task4	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task5	30.0±13.8	20.0±12.7	4.0±6.8	0.0±0.0	0.0±0.0	22.0±14.4	28.0±16.9	2.0±4.5	0.0±0.0	0.0±0.0	0.0±0.0
cube-single											
task1	16.0±4.2	12.0±4.0	14.0±4.7	6.0±3.5	2.0±2.4	14.0±4.7	12.0±4.7	6.0±3.7	16.0±5.3	14.0±5.3	6.0±4.0
task2	86.0±4.0	62.0±6.0	68.0±6.3	14.0±5.1	0.0±0.0	82.0±5.3	6.0±3.5	0.0±0.0	60.0±6.9	6.0±3.5	0.0±0.0
task3	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task4	74.0±5.5	46.0±6.6	56.0±7.2	10.0±4.7	0.0±0.0	74.0±6.5	36.0±7.6	0.0±0.0	52.0±7.6	0.0±0.0	0.0±0.0
task5	72.0±5.3	38.0±6.0	44.0±6.8	8.0±4.0	0.0±0.0	80.0±5.5	24.0±6.3	2.0±2.4	36.0±6.8	4.0±2.8	0.0±0.0
cube-double											
task1	4.0±3.5	0.0±0.0	2.0±2.8	2.0±3.5	0.0±0.0	6.0±4.9	2.0±3.5	0.0±0.0	6.0±5.3	2.0±3.5	0.0±0.0
task2	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task3	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task4	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task5	0.0±0.0	0.0±0.0	2.0±4.5	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
scene											
task1	2.0±1.4	0.0±0.0	4.0±2.4	4.0±2.4	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	8.0±3.2	2.0±2.0	2.0±2.0
task2	10.0±6.3	26.0±9.7	26.0±10.6	14.0±9.1	2.0±3.5	18.0±9.3	24.0±11.2	16.0±10.4	6.0±6.0	0.0±0.0	0.0±0.0
task3	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task4	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task5	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
puzzle-3x3											
task1	22.0±5.5	22.0±5.8	16.0±5.5	26.0±7.3	14.0±6.2	24.0±6.5	14.0±5.8	10.0±5.5	12.0±5.3	10.0±5.3	14.0±6.5
task2	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task3	8.0±2.8	2.0±1.4	0.0±0.0	0.0±0.0	0.0±0.0	20.0±4.5	28.0±5.5	0.0±0.0	2.0±1.4	0.0±0.0	0.0±0.0
task4	40.0±6.8	42.0±7.1	36.0±7.5	0.0±0.0	0.0±0.0	38.0±7.6	12.0±5.5	0.0±0.0	20.0±6.5	0.0±0.0	0.0±0.0
task5	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	2.0±1.4	0.0±0.0	0.0±0.0
puzzle-4x4											
task1	72.0±7.6	90.0±5.3	88.0±6.3	70.0±9.7	0.0±0.0	86.0±6.8	68.0±9.8	12.0±7.5	72.0±9.1	10.0±6.6	0.0±0.0
task2	0.0±0.0	0.0±0.0	2.0±2.4	0.0±0.0	0.0±0.0	2.0±2.4	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task3	8.0±3.5	0.0±0.0	2.0±2.0	0.0±0.0	0.0±0.0	0.0±0.0	6.0±4.0	0.0±0.0	2.0±2.4	0.0±0.0	0.0±0.0
task4	6.0±2.4	4.0±2.0	8.0±3.2	0.0±0.0	0.0±0.0	4.0±2.4	0.0±0.0	0.0±0.0	2.0±1.4	0.0±0.0	0.0±0.0
task5	0.0±0.0	0.0±0.0	2.0±2.0	0.0±0.0	0.0±0.0	2.0±2.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0

Table 5: SSCP performance on individual OGBench tasks without defense. Scores are success percentages.

Task	Raw clean	Gaussian	APGD			min-Q			BDAP		
		0.2	0.05	0.1	0.2	0.05	0.1	0.2	0.05	0.1	0.2
antmaze-large											
task1	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	4.0±4.5	0.0±0.0	0.0±0.0	10.0±7.2	0.0±0.0	0.0±0.0
task2	28.0±9.1	4.0±4.2	0.0±0.0	0.0±0.0	0.0±0.0	6.0±5.5	0.0±0.0	0.0±0.0	14.0±8.4	0.0±0.0	0.0±0.0
task3	78.0±4.9	26.0±5.5	14.0±4.7	0.0±0.0	0.0±0.0	12.0±4.5	0.0±0.0	0.0±0.0	70.0±6.5	2.0±2.0	0.0±0.0
task4	20.0±4.0	8.0±2.8	8.0±3.2	0.0±0.0	0.0±0.0	6.0±2.8	0.0±0.0	0.0±0.0	14.0±4.2	0.0±0.0	0.0±0.0
task5	12.0±3.5	8.0±2.8	4.0±2.4	0.0±0.0	0.0±0.0	14.0±4.2	0.0±0.0	0.0±0.0	10.0±3.7	0.0±0.0	0.0±0.0
antmaze-giant											
task1	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task2	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task3	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task4	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task5	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	4.0±7.9	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
antsoccer-arena											
task1	2.0±1.4	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task2	6.0±2.4	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task3	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task4	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task5	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
humanoidmaze-medium											
task1	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task2	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task3	2.0±4.5	2.0±4.5	0.0±0.0	0.0±0.0	0.0±0.0	4.0±7.9	2.0±4.5	16.0±16.8	2.0±4.5	0.0±0.0	0.0±0.0
task4	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task5	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	6.0±3.2	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
humanoidmaze-large											
task1	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task2	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task3	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task4	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task5	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
cube-single											
task1	78.0±4.9	56.0±6.0	68.0±6.3	40.0±6.9	0.0±0.0	72.0±6.0	70.0±6.6	2.0±2.0	38.0±6.8	14.0±5.3	0.0±0.0
task2	50.0±6.0	40.0±6.0	64.0±6.5	32.0±6.8	10.0±4.5	72.0±6.0	42.0±7.1	0.0±0.0	44.0±6.9	16.0±5.5	0.0±0.0
task3	74.0±6.3	52.0±7.5	62.0±8.0	42.0±8.6	4.0±3.5	84.0±6.0	76.0±7.5	24.0±7.9	50.0±8.6	10.0±5.5	0.0±0.0
task4	2.0±2.0	4.0±2.8	2.0±2.0	0.0±0.0	2.0±2.4	4.0±2.8	2.0±2.4	14.0±5.7	4.0±2.8	0.0±0.0	0.0±0.0
task5	68.0±5.5	38.0±6.0	68.0±6.3	14.0±4.9	0.0±0.0	88.0±4.5	12.0±4.7	0.0±0.0	44.0±6.9	2.0±2.0	0.0±0.0
cube-double											
task1	6.0±4.5	0.0±0.0	2.0±2.8	0.0±0.0	0.0±0.0	6.0±4.9	10.0±6.8	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task2	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task3	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task4	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task5	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
scene											
task1	94.0±2.4	96.0±2.0	78.0±4.7	16.0±4.5	0.0±0.0	36.0±5.5	8.0±3.2	0.0±0.0	92.0±3.2	24.0±5.5	0.0±0.0
task2	26.0±9.2	22.0±9.2	20.0±9.7	0.0±0.0	0.0±0.0	4.0±4.7	0.0±0.0	0.0±0.0	34.0±12.0	0.0±0.0	0.0±0.0
task3	54.0±4.9	72.0±4.7	46.0±5.7	0.0±0.0	0.0±0.0	18.0±4.5	0.0±0.0	0.0±0.0	6.0±2.8	0.0±0.0	0.0±0.0
task4	6.0±2.4	2.0±1.4	14.0±4.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task5	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	2.0±1.4	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
puzzle-3x3											
task1	100.0±0.0	100.0±0.0	98.0±2.0	58.0±8.0	8.0±4.7	90.0±4.5	52.0±8.1	0.0±0.0	100.0±0.0	86.0±6.0	0.0±0.0
task2	4.0±2.8	2.0±2.0	8.0±4.2	2.0±2.4	0.0±0.0	0.0±0.0	10.0±4.9	0.0±0.0	2.0±2.4	0.0±0.0	0.0±0.0
task3	8.0±2.8	2.0±1.4	2.0±1.4	0.0±0.0	0.0±0.0	18.0±4.5	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task4	6.0±3.2	6.0±3.5	16.0±5.8	0.0±0.0	0.0±0.0	6.0±3.7	4.0±3.2	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task5	12.0±3.2	34.0±4.9	32.0±5.3	0.0±0.0	0.0±0.0	18.0±4.5	4.0±2.4	0.0±0.0	2.0±1.4	0.0±0.0	0.0±0.0
puzzle-4x4											
task1	4.0±3.2	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	6.0±4.7	0.0±0.0	0.0±0.0
task2	2.0±2.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task3	2.0±2.0	0.0±0.0	6.0±3.5	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task4	2.0±1.4	2.0±1.4	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	2.0±1.4	0.0±0.0	0.0±0.0
task5	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0

Table 6: GFP performance on individual OGBench tasks without defense. Scores are success percentages.

Task	Raw clean	Gaussian	APGD			min-Q			BDAP		
		0.2	0.05	0.1	0.2	0.05	0.1	0.2	0.05	0.1	0.2
antmaze-large											
task1	98.0±3.5	92.0±4.9	96.0±4.5	56.0±8.6	0.0±0.0	86.0±6.5	50.0±8.8	0.0±0.0	94.0±5.3	22.0±8.0	0.0±0.0
task2	92.0±4.9	90.0±5.8	82.0±7.3	40.0±9.1	0.0±0.0	92.0±6.0	14.0±7.5	0.0±0.0	74.0±8.2	42.0±9.7	0.0±0.0
task3	96.0±2.8	94.0±3.5	92.0±4.0	84.0±5.5	0.0±0.0	92.0±4.0	92.0±4.5	0.0±0.0	96.0±3.2	68.0±7.2	0.0±0.0
task4	90.0±4.0	80.0±5.5	86.0±5.3	60.0±7.3	0.0±0.0	84.0±5.5	56.0±7.6	0.0±0.0	82.0±6.0	24.0±6.9	0.0±0.0
task5	96.0±2.8	84.0±4.9	94.0±4.0	76.0±6.6	0.0±0.0	92.0±4.5	32.0±7.2	0.0±0.0	88.0±5.1	52.0±7.9	0.0±0.0
antmaze-giant											
task1	4.0±3.2	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	6.0±4.2	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task2	84.0±5.3	32.0±6.9	66.0±7.5	0.0±0.0	0.0±0.0	22.0±6.9	0.0±0.0	0.0±0.0	80.0±6.9	0.0±0.0	0.0±0.0
task3	28.0±4.7	0.0±0.0	2.0±2.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	4.0±2.8	0.0±0.0	0.0±0.0
task4	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task5	8.0±8.9	12.0±10.6	18.0±12.1	0.0±0.0	0.0±0.0	28.0±13.1	18.0±13.2	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
antsoccer-arena											
task1	88.0±4.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	16.0±5.5	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task2	96.0±2.8	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	2.0±2.4	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task3	44.0±6.3	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task4	50.0±6.3	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task5	44.0±6.9	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
humanoidmaze-medium											
task1	82.0±6.5	70.0±8.0	60.0±9.1	0.0±0.0	0.0±0.0	74.0±8.6	38.0±9.8	0.0±0.0	34.0±9.2	0.0±0.0	0.0±0.0
task2	94.0±2.8	82.0±4.9	74.0±6.0	0.0±0.0	0.0±0.0	82.0±5.5	50.0±7.5	0.0±0.0	66.0±6.8	0.0±0.0	0.0±0.0
task3	98.0±4.5	90.0±7.7	90.0±8.4	4.0±7.9	0.0±0.0	94.0±7.9	76.0±10.9	8.0±9.2	50.0±11.2	2.0±4.5	0.0±0.0
task4	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task5	98.0±2.0	94.0±3.2	98.0±2.4	4.0±3.2	0.0±0.0	90.0±4.2	76.0±6.0	0.0±0.0	88.0±4.5	0.0±0.0	0.0±0.0
humanoidmaze-large											
task1	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	2.0±3.5	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task2	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task3	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task4	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task5	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
cube-single											
task1	100.0±0.0	86.0±4.2	78.0±5.5	6.0±3.5	0.0±0.0	54.0±6.6	0.0±0.0	0.0±0.0	56.0±6.8	0.0±0.0	0.0±0.0
task2	96.0±2.4	64.0±5.8	82.0±5.1	8.0±4.0	0.0±0.0	72.0±6.0	0.0±0.0	0.0±0.0	50.0±6.8	0.0±0.0	0.0±0.0
task3	98.0±2.0	56.0±6.0	72.0±6.0	2.0±2.4	0.0±0.0	26.0±6.0	0.0±0.0	0.0±0.0	28.0±6.0	0.0±0.0	0.0±0.0
task4	100.0±0.0	68.0±6.0	64.0±6.5	10.0±4.7	0.0±0.0	12.0±4.7	0.0±0.0	0.0±0.0	20.0±5.8	2.0±2.8	0.0±0.0
task5	96.0±2.8	80.0±5.3	68.0±6.5	12.0±4.9	0.0±0.0	50.0±6.9	0.0±0.0	0.0±0.0	34.0±6.8	0.0±0.0	0.0±0.0
cube-double											
task1	78.0±6.0	42.0±7.7	24.0±7.5	6.0±5.5	2.0±4.5	34.0±8.2	8.0±6.2	0.0±0.0	16.0±6.9	16.0±7.7	0.0±0.0
task2	54.0±6.3	14.0±5.3	10.0±5.3	2.0±3.5	0.0±0.0	42.0±7.6	2.0±3.5	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task3	56.0±6.0	18.0±5.5	12.0±5.3	0.0±0.0	0.0±0.0	18.0±6.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task4	6.0±2.8	0.0±0.0	4.0±2.8	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task5	50.0±7.2	2.0±3.5	12.0±6.6	0.0±0.0	0.0±0.0	22.0±7.7	2.0±4.5	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
scene											
task1	98.0±2.0	100.0±0.0	100.0±0.0	84.0±4.5	0.0±0.0	96.0±2.4	76.0±5.5	0.0±0.0	96.0±2.4	44.0±6.8	0.0±0.0
task2	64.0±6.8	44.0±7.7	66.0±8.1	2.0±4.5	0.0±0.0	24.0±7.6	0.0±0.0	0.0±0.0	34.0±8.4	0.0±0.0	0.0±0.0
task3	84.0±4.0	84.0±4.2	78.0±5.3	30.0±6.0	0.0±0.0	74.0±5.5	28.0±6.0	0.0±0.0	80.0±5.3	0.0±0.0	0.0±0.0
task4	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	4.0±2.8	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task5	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
puzzle-3x3											
task1	98.0±2.4	98.0±2.4	94.0±4.0	92.0±4.5	50.0±7.7	96.0±3.5	64.0±7.5	4.0±4.0	98.0±2.8	46.0±7.9	0.0±0.0
task2	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task3	4.0±2.8	4.0±2.8	0.0±0.0	2.0±2.8	0.0±0.0	6.0±3.7	10.0±4.7	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task4	2.0±2.4	4.0±3.5	18.0±6.0	2.0±3.5	0.0±0.0	14.0±5.5	8.0±4.9	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task5	2.0±2.0	24.0±5.5	46.0±6.8	4.0±3.2	0.0±0.0	12.0±4.7	4.0±3.5	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
puzzle-4x4											
task1	62.0±6.6	50.0±7.5	46.0±8.1	26.0±7.7	0.0±0.0	36.0±8.0	6.0±5.5	0.0±0.0	42.0±8.4	4.0±5.1	0.0±0.0
task2	26.0±5.5	14.0±5.1	16.0±5.7	4.0±4.0	0.0±0.0	18.0±6.0	4.0±4.0	0.0±0.0	12.0±5.5	0.0±0.0	0.0±0.0
task3	48.0±6.0	36.0±6.5	22.0±6.3	2.0±3.5	0.0±0.0	22.0±6.5	4.0±4.0	0.0±0.0	28.0±6.9	0.0±0.0	0.0±0.0
task4	30.0±5.3	20.0±5.3	18.0±5.5	0.0±0.0	0.0±0.0	12.0±4.7	4.0±3.5	0.0±0.0	8.0±4.2	0.0±0.0	0.0±0.0
task5	20.0±4.5	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	8.0±4.0	2.0±2.8	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0

Table 7: FQL performance on individual D4RL tasks without defense. Scores are success percentages for AntMaze and normalized returns for Adroit.

Task	Raw clean	Gaussian	APGD			min-Q			BDAP		
		0.2	0.05	0.1	0.2	0.05	0.1	0.2	0.05	0.1	0.2
AntMaze											
antmaze-umaze-v2	98.0±2.0	98.0±2.0	94.0±2.8	94.0±3.2	90.0±4.0	96.0±2.8	92.0±3.7	36.0±7.7	94.0±3.2	94.0±3.5	94.0±3.7
antmaze-umaze-diverse-v2	96.0±4.9	96.0±5.1	92.0±6.5	96.0±5.8	88.0±7.2	90.0±6.9	94.0±6.0	48.0±10.5	84.0±8.5	38.0±11.2	4.0±4.7
antmaze-medium-play-v2	70.0±6.9	74.0±7.6	78.0±7.7	78.0±8.1	34.0±10.3	46.0±10.2	22.0±8.8	0.0±0.0	66.0±9.3	0.0±0.0	0.0±0.0
antmaze-medium-diverse-v2	70.0±13.0	68.0±13.4	74.0±13.1	62.0±14.8	8.0±8.5	58.0±15.2	8.0±8.9	0.0±0.0	66.0±15.0	66.0±15.8	0.0±0.0
antmaze-large-play-v2	86.0±6.9	92.0±5.8	84.0±7.7	84.0±8.2	0.0±0.0	88.0±6.9	62.0±10.1	0.0±0.0	76.0±8.9	68.0±10.2	0.0±0.0
antmaze-large-diverse-v2	82.0±4.0	70.0±5.5	84.0±4.9	54.0±7.1	0.0±0.0	88.0±4.7	74.0±6.5	0.0±0.0	68.0±6.9	68.0±7.5	0.0±0.0
Pen											
pen-human-v1	55.5±6.0	50.1±8.3	38.6±9.0	10.5±6.8	-2.1±3.7	59.9±9.2	60.5±10.1	48.3±11.5	52.4±9.7	25.2±9.4	-1.3±4.0
pen-cloned-v1	76.2±11.0	87.2±10.0	79.4±11.8	73.4±12.5	42.6±14.2	72.8±12.1	73.8±13.0	55.8±14.6	75.7±12.7	57.7±14.0	3.7±6.2
pen-expert-v1	148.0±6.0	145.6±6.5	144.2±7.0	140.4±7.8	98.3±11.3	145.5±6.8	143.1±7.5	130.1±9.7	147.7±6.6	132.7±9.5	68.6±14.0
Hammer											
hammer-human-v1	0.8±1.0	0.8±1.1	0.5±0.9	0.3±0.7	0.0±0.5	0.7±1.1	1.6±1.5	1.0±1.3	0.4±0.8	0.2±0.6	0.2±0.6
hammer-cloned-v1	17.9±9.0	11.2±8.1	13.4±9.0	10.8±8.8	0.1±1.3	22.2±10.6	12.7±9.8	2.9±4.3	22.9±11.3	4.9±5.6	0.1±1.5
hammer-expert-v1	115.6±3.0	116.0±3.8	120.7±4.2	73.4±10.5	10.4±8.0	119.9±4.1	115.5±5.3	112.7±6.2	115.2±5.0	92.7±8.9	10.0±8.5
Door											
door-human-v1	0.1±0.3	0.2±0.4	0.0±0.3	0.0±0.3	0.0±0.3	0.2±0.4	0.4±0.6	0.5±0.7	0.2±0.4	0.4±0.6	0.0±0.3
door-cloned-v1	3.4±1.0	2.2±1.2	0.7±0.8	0.9±1.0	-0.1±0.3	-0.1±0.3	-0.1±0.3	-0.2±0.4	-0.1±0.3	-0.1±0.3	-0.1±0.3
door-expert-v1	102.8±1.0	100.8±1.8	102.8±1.5	102.7±1.7	99.8±2.5	99.6±2.4	102.3±1.9	101.7±2.2	102.5±1.8	99.1±2.8	61.2±10.5
Relocate											
relocate-human-v1	0.1±0.2	0.1±0.2	0.1±0.2	0.1±0.2	0.0±0.2	0.2±0.3	0.0±0.2	0.1±0.2	0.1±0.2	0.0±0.2	0.1±0.2
relocate-cloned-v1	-0.1±0.2	-0.1±0.2	-0.1±0.2	0.0±0.2	-0.2±0.3	0.0±0.2	0.0±0.2	0.0±0.2	-0.1±0.2	-0.1±0.2	-0.1±0.2
relocate-expert-v1	107.3±1.0	107.2±1.2	107.9±1.2	107.6±1.4	101.0±3.5	107.0±1.5	106.4±1.8	105.0±2.5	106.7±1.7	99.8±3.8	32.0±12.0

Table 8: MeanFlowQL performance on individual D4RL tasks without defense. Scores are success percentages for AntMaze and normalized returns for Adroit.

Task	Raw clean	Gaussian	APGD			min-Q			BDAP		
		0.2	0.05	0.1	0.2	0.05	0.1	0.2	0.05	0.1	0.2
AntMaze											
antmaze-umaze-v2	92.0±2.8	80.0±4.0	72.0±4.7	92.0±3.2	90.0±3.6	68.0±5.3	74.0±4.9	84.0±4.2	82.0±4.6	80.0±4.7	72.0±5.2
antmaze-umaze-diverse-v2	96.0±4.5	54.0±7.0	48.0±7.6	86.0±5.8	80.0±6.4	20.0±6.8	24.0±7.4	72.0±6.7	38.0±8.5	12.0±6.0	0.0±0.0
antmaze-medium-play-v2	94.0±5.2	80.0±7.5	74.0±8.0	92.0±6.0	0.0±0.0	32.0±9.0	70.0±8.2	50.0±10.0	30.0±9.2	0.0±0.0	0.0±0.0
antmaze-medium-diverse-v2	94.0±7.0	30.0±13.8	24.0±12.7	50.0±15.0	0.0±0.0	14.0±10.5	22.0±12.6	2.0±4.5	16.0±11.5	16.0±12.0	0.0±0.0
antmaze-large-play-v2	0.0±0.0	2.0±3.5	6.0±4.5	0.0±0.0	0.0±0.0	4.0±4.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
antmaze-large-diverse-v2	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
Pen											
pen-human-v1	-1.4±0.8	-2.6±1.0	-1.3±0.9	-2.1±1.1	-1.6±1.2	-0.1±0.7	-1.8±1.0	-1.7±1.1	-2.6±1.2	-2.1±1.2	-0.2±0.8
pen-cloned-v1	-1.3±0.9	-1.0±0.9	-2.2±1.1	-2.0±1.2	-2.6±1.4	-2.0±1.2	0.1±0.5	0.1±0.5	-1.8±1.2	-2.4±1.4	-1.8±1.3
pen-expert-v1	20.3±5.5	15.4±5.8	4.5±4.6	7.8±5.0	3.4±4.1	10.4±5.3	7.8±5.2	13.3±6.0	7.9±5.4	14.3±6.1	3.3±4.3
Hammer											
hammer-human-v1	-0.1±0.1	-0.2±0.2	-0.2±0.2	-0.2±0.2	-0.1±0.2	-0.2±0.2	-0.2±0.2	-0.2±0.2	-0.2±0.2	-0.2±0.2	-0.1±0.2
hammer-cloned-v1	0.1±0.2	0.0±0.2	0.0±0.2	0.0±0.2	-0.1±0.2	0.1±0.2	0.1±0.2	0.0±0.2	0.1±0.3	0.1±0.3	0.0±0.2
hammer-expert-v1	0.1±0.2	0.1±0.2	0.1±0.2	0.0±0.2	0.0±0.2	0.1±0.2	0.1±0.2	0.1±0.2	0.1±0.2	0.1±0.2	0.0±0.2
Door											
door-human-v1	-0.2±0.1	-0.2±0.1	-0.2±0.1	-0.2±0.1	-0.1±0.1	-0.2±0.1	-0.1±0.1	-0.2±0.1	-0.2±0.1	-0.1±0.1	-0.2±0.1
door-cloned-v1	-0.2±0.1	-0.2±0.1	-0.2±0.1	-0.2±0.1	-0.3±0.1	-0.2±0.1	-0.2±0.1	-0.2±0.1	-0.2±0.1	-0.2±0.1	-0.2±0.1
door-expert-v1	-0.2±0.1	-0.3±0.1	-0.2±0.1	-0.3±0.1	-0.3±0.1	-0.2±0.1	-0.2±0.1	-0.2±0.1	-0.2±0.1	-0.2±0.1	-0.2±0.1
Relocate											
relocate-human-v1	-0.1±0.1	0.0±0.1	-0.1±0.1	-0.1±0.1	-0.1±0.1	-0.1±0.1	-0.1±0.1	-0.1±0.1	-0.1±0.1	0.0±0.1	0.1±0.1
relocate-cloned-v1	-0.1±0.1	-0.1±0.1	-0.1±0.1	-0.3±0.1	0.0±0.1	-0.1±0.1	-0.1±0.1	-0.1±0.1	-0.1±0.1	0.0±0.1	0.0±0.1
relocate-expert-v1	-0.2±0.1	-0.2±0.1	-0.2±0.1	-0.2±0.1	-0.2±0.1	-0.2±0.1	-0.2±0.1	-0.2±0.1	-0.2±0.1	-0.2±0.1	-0.2±0.1

Table 9: SSCP performance on individual D4RL tasks without defense. Scores are success percentages for AntMaze and normalized returns for Adroit.

Task	Raw clean	Gaussian	APGD			min-Q			BDAP		
		0.2	0.05	0.1	0.2	0.05	0.1	0.2	0.05	0.1	0.2
AntMaze											
antmaze-umaze-v2	96.0±2.4	70.0±4.9	56.0±6.0	72.0±4.9	20.0±5.3	64.0±5.5	70.0±4.9	58.0±6.3	80.0±4.5	86.0±4.0	20.0±5.5
antmaze-umaze-diverse-v2	86.0±5.5	76.0±6.2	64.0±6.9	62.0±7.5	24.0±6.9	66.0±7.2	62.0±7.7	80.0±6.5	92.0±5.1	74.0±6.9	32.0±8.2
antmaze-medium-play-v2	4.0±4.2	2.0±3.5	0.0±0.0	0.0±0.0	0.0±0.0	4.0±4.5	8.0±5.5	0.0±0.0	2.0±3.5	0.0±0.0	0.0±0.0
antmaze-medium-diverse-v2	30.0±12.5	0.0±0.0	4.0±4.9	6.0±6.0	0.0±0.0	8.0±6.9	18.0±10.0	28.0±12.0	4.0±5.5	2.0±3.5	0.0±0.0
antmaze-large-play-v2	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	2.0±3.5	0.0±0.0
antmaze-large-diverse-v2	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
Pen											
pen-human-v1	73.2±7.0	62.0±8.5	58.7±9.0	52.7±9.8	30.0±11.0	68.0±8.8	64.1±9.5	70.4±10.2	62.8±9.2	46.8±10.5	16.2±8.0
pen-cloned-v1	76.1±10.5	72.8±11.0	69.6±11.8	48.4±13.0	22.9±13.5	71.8±11.5	70.1±12.2	55.0±13.8	71.2±12.0	40.7±14.0	16.6±10.0
pen-expert-v1	147.3±6.0	148.6±6.2	147.1±6.5	136.5±8.0	56.3±13.0	147.6±6.3	145.3±7.0	130.9±9.5	148.6±6.4	134.3±9.0	53.1±13.5
Hammer											
hammer-human-v1	0.0±0.2	-0.2±0.2	-0.1±0.2	-0.2±0.2	-0.3±0.2	-0.1±0.2	-0.2±0.2	-0.1±0.2	-0.3±0.2	-0.1±0.2	-0.2±0.2
hammer-cloned-v1	4.5±4.5	3.1±4.0	5.5±4.8	2.3±3.5	0.3±1.2	0.6±1.8	2.3±3.4	0.6±1.8	0.5±1.6	0.2±1.0	1.2±2.5
hammer-expert-v1	147.0±4.0	130.9±7.5	127.2±8.0	130.8±7.8	104.3±11.0	126.6±8.2	128.1±7.5	129.1±7.0	129.8±7.6	128.2±8.0	102.4±11.5
Door											
door-human-v1	-0.2±0.1	-0.2±0.1	-0.1±0.1	-0.2±0.1	-0.2±0.1	-0.2±0.1	-0.2±0.1	-0.2±0.1	-0.2±0.1	-0.2±0.1	-0.2±0.1
door-cloned-v1	0.0±0.2	0.2±0.3	0.1±0.2	0.0±0.2	0.0±0.2	0.0±0.2	0.0±0.2	0.0±0.2	0.0±0.2	0.3±0.4	0.0±0.2
door-expert-v1	105.3±1.1	104.4±1.7	105.6±1.5	103.3±2.0	51.8±12.0	105.3±1.8	104.9±2.0	102.3±2.8	103.6±2.2	98.3±3.5	96.1±4.0
Relocate											
relocate-human-v1	-0.2±0.1	-0.2±0.1	-0.2±0.1	-0.2±0.1	-0.2±0.1	-0.2±0.1	-0.2±0.1	-0.2±0.1	-0.2±0.1	-0.2±0.1	-0.2±0.1
relocate-cloned-v1	-0.1±0.2	0.2±0.3	-0.1±0.2	0.0±0.2	-0.2±0.3	0.0±0.2	-0.1±0.2	-0.1±0.2	-0.1±0.2	-0.1±0.2	-0.1±0.2
relocate-expert-v1	113.7±1.5	107.2±2.0	102.8±2.5	94.1±4.5	13.3±10.5	104.1±2.8	106.1±2.5	105.8±3.0	103.1±3.2	98.6±4.2	61.3±9.0

Table 10: GFP performance on individual D4RL tasks without defense. Scores are success percentages for AntMaze and normalized returns for Adroit.

Task	Raw clean	Gaussian	APGD			min-Q			BDAP		
		0.2	0.05	0.1	0.2	0.05	0.1	0.2	0.05	0.1	0.2
AntMaze											
antmaze-umaze-v2	98.0±2.0	94.0±2.8	92.0±3.5	94.0±2.8	92.0±3.5	90.0±3.5	98.0±2.4	98.0±2.4	94.0±3.2	94.0±3.2	96.0±2.8
antmaze-umaze-diverse-v2	98.0±4.5	92.0±6.0	86.0±6.9	98.0±3.5	82.0±7.2	98.0±3.5	98.0±4.5	96.0±4.7	84.0±6.9	26.0±10.5	24.0±10.8
antmaze-medium-play-v2	98.0±4.5	98.0±4.5	90.0±6.5	96.0±4.9	66.0±9.5	94.0±6.2	84.0±7.5	84.0±7.6	92.0±6.8	0.0±0.0	0.0±0.0
antmaze-medium-diverse-v2	90.0±13.0	86.0±13.0	82.0±13.5	88.0±13.0	58.0±15.0	78.0±14.0	64.0±15.0	46.0±16.0	76.0±14.5	78.0±14.5	86.0±13.0
antmaze-large-play-v2	92.0±6.9	88.0±7.2	84.0±7.7	92.0±6.5	0.0±0.0	92.0±6.8	74.0±8.9	0.0±0.0	80.0±8.5	80.0±8.7	0.0±0.0
antmaze-large-diverse-v2	92.0±4.0	92.0±4.2	88.0±4.9	80.0±6.0	0.0±0.0	84.0±5.5	72.0±6.8	0.0±0.0	60.0±7.5	44.0±8.0	0.0±0.0
Pen											
pen-human-v1	79.2±6.0	70.1±8.0	69.4±8.5	54.5±9.5	20.0±9.0	67.2±8.8	66.0±9.2	61.0±10.0	66.4±9.0	51.6±10.2	26.4±9.5
pen-cloned-v1	98.0±11.0	87.0±10.5	88.8±11.5	95.3±10.0	63.9±12.8	80.8±12.0	77.3±12.5	74.4±13.0	76.6±12.2	73.7±13.5	51.5±14.0
pen-expert-v1	158.4±6.0	149.9±6.5	146.9±7.0	139.5±8.0	103.9±11.0	143.0±7.2	141.1±7.8	125.0±9.5	150.8±6.5	137.8±9.0	60.6±14.0
Hammer											
hammer-human-v1	2.3±1.0	1.6±1.1	1.5±1.0	0.9±0.9	0.2±0.6	1.7±1.1	2.1±1.2	1.9±1.1	1.9±1.1	2.1±1.2	0.4±0.7
hammer-cloned-v1	6.0±9.0	2.6±7.0	3.8±8.0	1.5±6.0	0.2±1.2	4.1±8.0	2.0±6.5	0.6±2.5	4.8±8.5	2.1±6.5	0.1±1.0
hammer-expert-v1	129.6±3.0	124.7±4.5	121.5±5.0	119.6±5.5	106.2±7.0	121.0±5.0	119.8±5.5	110.3±7.5	123.2±5.0	118.3±6.0	31.3±12.0
Door											
door-human-v1	5.5±3.0	0.9±1.3	1.3±1.5	0.0±0.4	0.0±0.4	1.3±1.5	3.8±2.2	0.4±0.8	0.1±0.5	1.3±1.5	0.0±0.4
door-cloned-v1	0.1±0.3	0.1±0.3	0.1±0.3	0.0±0.2	-0.1±0.3	0.0±0.2	0.0±0.2	-0.1±0.3	0.0±0.2	-0.1±0.3	0.0±0.2
door-expert-v1	112.0±1.0	103.2±2.0	101.9±2.0	102.6±2.0	100.6±2.5	102.8±2.2	103.0±2.2	102.0±2.8	101.8±2.5	87.1±5.0	63.6±10.0
Relocate											
relocate-human-v1	1.8±1.2	0.6±0.7	0.8±0.8	0.3±0.5	0.1±0.3	0.9±0.8	0.6±0.7	0.3±0.5	0.6±0.7	0.8±0.8	0.1±0.3
relocate-cloned-v1	2.5±1.5	1.6±1.2	0.3±0.6	0.2±0.5	-0.1±0.3	0.1±0.4	0.1±0.4	0.3±0.6	0.7±0.8	0.1±0.4	-0.1±0.3
relocate-expert-v1	117.7±1.0	107.7±1.5	107.6±1.6	106.2±2.0	99.2±3.8	108.3±1.8	107.5±1.8	106.0±2.5	108.2±1.9	106.7±2.1	64.6±9.0

C.2 BDAP Objective Ablation

We compare the two terms in (8) on five tasks and four victim families. Q-only removes the behavior term by setting $\lambda_{\mathrm{B},k} = 0$, while Behavior-only removes the value term by setting $\lambda_{Q,k} = 0$. Full BDAP retains both terms. The attacks use the FQL surrogate and are evaluated without a defense at $\epsilon = 0.1$.

Table 11 shows that Full BDAP scores below both single-objective variants in 17 of 20 comparisons and ties them in the remaining three. On GFP, the combined objective outperforms both variants on all five tasks, supporting the contribution of both channels to transfer in these evaluations. The three ties occur on antmaze-large task1 for MeanFlowQL and SSCP, and on scene task1 for MeanFlowQL. Their raw-clean success rates are 0%, 0%, and 2%, respectively, leaving little room to distinguish attack strength.

Table 11: BDAP objective ablation at $\epsilon = 0.1$ without defense. Scores are success percentages for OGBench and normalized returns for D4RL Adroit. Lower attack scores are better. Bold marks the best attack in each row, including ties.

Victim	Raw clean	Q-only	Behavior-only	Full BDAP
		OGBench		
antmaze-large / task1				
FQL	82.0	18.0	36.0	**12.0**
MeanFlowQL	0.0	**0.0**	**0.0**	**0.0**
SSCP	0.0	**0.0**	**0.0**	**0.0**
GFP	98.0	48.0	34.0	**22.0**
scene / task1				
FQL	100.0	28.0	46.0	**22.0**
MeanFlowQL	2.0	**2.0**	**2.0**	**2.0**
SSCP	94.0	30.0	40.0	**24.0**
GFP	98.0	68.0	54.0	**44.0**
cube-single / task2				
FQL	98.0	6.0	18.0	**0.0**
MeanFlowQL	86.0	14.0	22.0	**6.0**
SSCP	50.0	22.0	28.0	**16.0**
GFP	96.0	12.0	8.0	**0.0**
puzzle-3x3 / task1				
FQL	78.0	18.0	30.0	**10.0**
MeanFlowQL	22.0	14.0	18.0	**10.0**
SSCP	100.0	88.0	92.0	**86.0**
GFP	98.0	60.0	52.0	**46.0**
		D4RL		
pen-human-v1				
FQL	55.5	29.0	40.0	**25.2**
MeanFlowQL	-1.4	-1.8	-1.2	**-2.1**
SSCP	73.2	50.0	54.0	**46.8**
GFP	79.2	61.0	58.0	**51.6**

C.3 PER-TASK DEFENSE RESULTS

The following tables compare DMBP with DMBP+RCF for each victim at $\rho = 0.3$. APGD, min-Q, and BDAP use $\epsilon = 0.1$, matching Table 2, while Gaussian noise uses scale $\epsilon = 0.2$. For each task, the same corrector trained with uniform random perturbations is reused across all four perturbation types without attack-type information. Raw-clean results evaluate the victim on clean observations without DMBP or RCF.

Table 12: FQL: OGBench success percentages under attacks ($\epsilon = 0.1$) and Gaussian noise (scale $\epsilon = 0.2$). +RCF denotes DMBP+RCF with $\rho = 0.3$.

Task	Raw clean	Gaussian (0.2)		APGD (0.1)		min-Q (0.1)		BDAP (0.1)	
		DMBP	+RCF	DMBP	+RCF	DMBP	+RCF	DMBP	+RCF
antmaze-large									
task1	82.0±7.7	68.0±9.8	72.0±8.5	76.0±9.4	86.0±6.9	28.0±10.3	34.0±9.9	50.0±11.5	56.0±10.4
task2	64.0±9.7	56.0±10.5	66.0±9.1	42.0±11.0	50.0±10.1	16.0±8.5	24.0±9.1	12.0±7.5	42.0±10.5
task3	96.0±2.4	90.0±3.7	94.0±2.8	80.0±5.3	82.0±4.5	82.0±5.3	84.0±4.5	70.0±6.2	82.0±4.7
task4	90.0±2.8	82.0±4.0	90.0±2.8	72.0±4.9	76.0±4.2	56.0±5.7	62.0±5.1	40.0±5.5	42.0±5.3
task5	72.0±4.7	76.0±4.7	84.0±3.7	70.0±5.5	84.0±4.0	16.0±4.5	56.0±5.5	42.0±6.0	50.0±5.5
antmaze-giant									
task1	24.0±10.9	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task2	50.0±12.2	8.0±6.9	10.0±6.9	4.0±5.1	4.0±4.7	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task3	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task4	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task5	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
antsoccer-arena									
task1	78.0±4.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task2	88.0±2.8	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task3	56.0±6.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task4	42.0±6.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task5	18.0±7.2	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
humanoidmaze-medium									
task1	24.0±13.1	10.0±9.6	12.0±9.4	0.0±0.0	0.0±0.0	4.0±6.9	8.0±8.7	0.0±0.0	0.0±0.0
task2	98.0±2.0	82.0±5.1	92.0±3.2	26.0±6.0	50.0±6.3	54.0±7.2	64.0±6.5	0.0±0.0	0.0±0.0
task3	98.0±4.5	88.0±14.0	84.0±14.3	44.0±22.4	44.0±20.4	90.0±14.2	84.0±15.8	0.0±0.0	0.0±0.0
task4	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task5	98.0±1.4	90.0±3.7	98.0±1.4	60.0±6.3	66.0±5.5	88.0±4.5	90.0±3.7	0.0±0.0	0.0±0.0
humanoidmaze-large									
task1	0.0±0.0	2.0±3.5	2.0±2.8	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task2	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task3	16.0±8.2	10.0±6.9	14.0±7.5	0.0±0.0	0.0±0.0	2.0±3.5	4.0±4.5	0.0±0.0	0.0±0.0
task4	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task5	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
cube-single									
task1	98.0±1.4	94.0±2.8	96.0±2.4	90.0±4.0	94.0±2.8	96.0±2.8	98.0±1.4	86.0±4.7	90.0±3.7
task2	98.0±1.4	92.0±3.2	96.0±2.4	96.0±2.4	96.0±2.4	98.0±2.0	98.0±1.4	78.0±5.5	84.0±4.5
task3	100.0±0.0	98.0±2.0	98.0±2.0	88.0±5.1	96.0±2.8	90.0±4.9	98.0±2.0	94.0±4.0	94.0±3.5
task4	94.0±2.8	88.0±4.2	96.0±2.4	84.0±5.1	86.0±4.5	90.0±4.5	94.0±3.2	74.0±6.5	74.0±5.8
task5	94.0±2.8	86.0±4.2	92.0±2.8	92.0±3.5	92.0±3.2	94.0±3.2	96.0±2.4	78.0±5.5	84.0±4.5
cube-double									
task1	52.0±9.2	34.0±9.2	34.0±8.4	22.0±8.4	32.0±8.6	18.0±8.2	26.0±8.5	34.0±10.1	36.0±9.3
task2	28.0±5.5	22.0±5.5	26.0±5.3	22.0±5.7	28.0±5.5	18.0±5.5	26.0±5.8	0.0±0.0	8.0±3.5
task3	22.0±4.9	18.0±4.9	26.0±4.9	16.0±4.9	18.0±4.5	22.0±5.8	26.0±5.5	24.0±6.0	24.0±5.5
task4	12.0±2.8	2.0±1.4	4.0±1.4	0.0±0.0	4.0±2.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task5	12.0±8.4	4.0±5.1	4.0±4.7	4.0±5.5	8.0±6.9	4.0±5.7	6.0±6.5	2.0±4.5	4.0±5.1
scene									
task1	100.0±0.0	100.0±0.0	100.0±0.0	100.0±0.0	100.0±0.0	100.0±0.0	100.0±0.0	100.0±0.0	100.0±0.0
task2	92.0±5.7	86.0±7.7	86.0±6.9	74.0±10.2	78.0±8.7	62.0±11.8	66.0±10.5	68.0±11.3	68.0±10.3
task3	86.0±2.4	76.0±3.2	86.0±2.4	80.0±3.2	84.0±2.8	64.0±4.0	78.0±3.2	72.0±3.7	76.0±3.2
task4	2.0±1.4	0.0±0.0	2.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	2.0±0.0	0.0±0.0	2.0±0.0
task5	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
puzzle-3x3									
task1	78.0±5.5	82.0±5.5	90.0±3.7	74.0±6.5	76.0±5.7	22.0±6.5	38.0±6.8	76.0±6.5	86.0±4.9
task2	22.0±5.5	2.0±2.0	4.0±2.4	10.0±4.5	12.0±4.5	10.0±4.7	16.0±5.3	10.0±4.7	10.0±4.2
task3	4.0±2.0	4.0±2.0	6.0±2.4	0.0±0.0	4.0±2.0	2.0±1.4	6.0±2.4	8.0±3.2	8.0±2.8
task4	4.0±2.8	4.0±2.8	4.0±2.4	6.0±3.5	10.0±4.0	2.0±2.4	8.0±4.0	2.0±2.4	4.0±2.8
task5	28.0±3.7	16.0±3.2	16.0±2.8	10.0±2.8	30.0±3.7	4.0±2.0	8.0±2.4	18.0±3.5	18.0±3.2
puzzle-4x4									
task1	32.0±7.9	26.0±7.7	32.0±7.5	6.0±4.5	14.0±6.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task2	6.0±3.2	6.0±3.5	6.0±3.2	0.0±0.0	0.0±0.0	2.0±2.4	2.0±2.0	0.0±0.0	0.0±0.0
task3	18.0±4.9	8.0±3.7	20.0±4.9	2.0±2.0	10.0±4.0	2.0±2.0	2.0±2.0	0.0±0.0	0.0±0.0
task4	4.0±2.0	0.0±0.0	2.0±1.4	0.0±0.0	2.0±1.4	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task5	6.0±2.8	2.0±1.4	2.0±1.4	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0

Table 13: MeanFlowQL: OGBench success percentages under attacks ($\epsilon = 0.1$) and Gaussian noise (scale $\epsilon = 0.2$). +RCF denotes DMBP+RCF with $\rho = 0.3$.

Task	Raw clean	Gaussian (0.2)		APGD (0.1)		min-Q (0.1)		BDAP (0.1)	
		DMBP	+RCF	DMBP	+RCF	DMBP	+RCF	DMBP	+RCF
antmaze-large									
task1	0.0±0.0	12.0±6.8	18.0±7.2	26.0±10.0	26.0±9.1	2.0±3.5	2.0±2.8	0.0±0.0	4.0±4.2
task2	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task3	0.0±0.0	2.0±2.0	50.0±6.0	4.0±2.8	50.0±6.8	0.0±0.0	12.0±4.2	0.0±0.0	4.0±2.8
task4	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task5	0.0±0.0	0.0±0.0	2.0±1.4	2.0±2.0	2.0±1.4	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
antmaze-giant									
task1	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task2	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task3	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task4	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task5	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
antsoccer-arena									
task1	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task2	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task3	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task4	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task5	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
humanoidmaze-medium									
task1	90.0±9.2	88.0±10.5	96.0±5.7	18.0±13.6	52.0±16.1	84.0±12.9	88.0±10.5	0.0±0.0	0.0±0.0
task2	96.0±2.4	80.0±5.3	88.0±4.0	46.0±7.3	72.0±6.0	84.0±5.5	92.0±3.5	0.0±0.0	0.0±0.0
task3	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task4	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task5	100.0±0.0	86.0±4.0	96.0±2.0	80.0±5.3	82.0±4.5	92.0±3.5	96.0±2.4	0.0±0.0	4.0±2.4
humanoidmaze-large									
task1	54.0±11.7	16.0±8.9	20.0±8.9	0.0±0.0	0.0±0.0	12.0±8.8	18.0±9.5	0.0±0.0	0.0±0.0
task2	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task3	6.0±5.3	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	2.0±3.5	0.0±0.0	0.0±0.0
task4	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task5	30.0±13.8	4.0±6.2	8.0±7.7	0.0±0.0	0.0±0.0	2.0±4.5	8.0±8.6	0.0±0.0	0.0±0.0
cube-single									
task1	16.0±4.2	4.0±2.4	10.0±3.2	4.0±2.8	6.0±2.8	2.0±2.0	10.0±3.7	8.0±3.7	8.0±3.5
task2	86.0±4.0	86.0±4.2	92.0±3.2	88.0±4.5	88.0±4.0	86.0±4.7	90.0±3.7	80.0±5.7	86.0±4.5
task3	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task4	74.0±5.5	52.0±6.6	54.0±6.0	60.0±7.1	60.0±6.5	60.0±7.1	64.0±6.3	46.0±7.5	52.0±6.9
task5	72.0±5.3	60.0±6.0	68.0±5.3	64.0±6.5	64.0±6.0	60.0±6.6	64.0±6.0	44.0±6.9	52.0±6.5
cube-double									
task1	4.0±3.5	4.0±3.7	18.0±6.8	6.0±4.9	8.0±5.3	2.0±2.8	6.0±4.5	8.0±6.0	12.0±6.6
task2	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task3	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task4	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task5	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
scene									
task1	2.0±1.4	0.0±0.0	0.0±0.0	0.0±0.0	2.0±1.4	0.0±0.0	2.0±1.4	6.0±2.8	6.0±2.8
task2	10.0±6.3	10.0±6.6	10.0±6.0	6.0±5.8	8.0±6.0	2.0±3.5	2.0±2.8	0.0±0.0	0.0±0.0
task3	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task4	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task5	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
puzzle-3x3									
task1	22.0±5.5	20.0±5.5	20.0±5.1	12.0±4.9	14.0±4.7	16.0±5.5	18.0±5.5	6.0±3.7	14.0±5.1
task2	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task3	8.0±2.8	2.0±1.4	4.0±2.0	4.0±2.4	6.0±2.4	8.0±3.2	8.0±2.8	2.0±1.4	2.0±1.4
task4	40.0±6.8	38.0±6.9	40.0±6.5	44.0±7.7	46.0±7.2	36.0±7.6	46.0±7.2	40.0±8.1	46.0±7.5
task5	0.0±0.0	2.0±1.4	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
puzzle-4x4									
task1	72.0±7.6	78.0±7.5	84.0±6.0	50.0±9.8	76.0±7.6	58.0±9.6	68.0±8.4	2.0±2.8	6.0±4.5
task2	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task3	8.0±3.5	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task4	6.0±2.4	2.0±1.4	2.0±1.4	0.0±0.0	0.0±0.0	0.0±0.0	2.0±1.4	0.0±0.0	0.0±0.0
task5	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0

Table 14: SSCP: OGBench success percentages under attacks ($\epsilon = 0.1$) and Gaussian noise (scale $\epsilon = 0.2$). +RCF denotes DMBP+RCF with $\rho = 0.3$.

Task	Raw clean	Gaussian (0.2)		APGD (0.1)		min-Q (0.1)		BDAP (0.1)	
		DMBP	+RCF	DMBP	+RCF	DMBP	+RCF	DMBP	+RCF
antmaze-large									
task1	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task2	28.0±9.1	2.0±2.8	8.0±5.3	2.0±3.5	2.0±2.8	0.0±0.0	8.0±5.5	2.0±3.5	22.0±8.8
task3	78.0±4.9	54.0±6.5	46.0±5.5	10.0±4.0	22.0±4.9	0.0±0.0	0.0±0.0	4.0±2.8	8.0±3.5
task4	20.0±4.0	2.0±1.4	4.0±2.0	2.0±1.4	2.0±1.4	0.0±0.0	0.0±0.0	2.0±1.4	2.0±1.4
task5	12.0±3.5	20.0±4.7	20.0±4.0	0.0±0.0	8.0±2.8	6.0±2.8	10.0±3.2	2.0±2.0	8.0±3.2
antmaze-giant									
task1	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task2	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task3	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task4	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task5	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
antsoccer-arena									
task1	2.0±1.4	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task2	6.0±2.4	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task3	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task4	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task5	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
humanoidmaze-medium									
task1	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task2	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task3	2.0±4.5	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	2.0±4.5	2.0±4.5	0.0±0.0	0.0±0.0
task4	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task5	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	2.0±2.8	2.0±2.4	0.0±0.0	0.0±0.0
humanoidmaze-large									
task1	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task2	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task3	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task4	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	2.0±2.4	0.0±0.0	0.0±0.0
task5	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
cube-single									
task1	78.0±4.9	66.0±6.0	70.0±5.1	64.0±6.5	68.0±5.5	68.0±6.3	70.0±5.5	72.0±6.5	76.0±5.3
task2	50.0±6.0	36.0±6.0	42.0±5.5	40.0±6.6	42.0±5.8	48.0±6.8	48.0±6.0	42.0±6.9	44.0±6.2
task3	74.0±6.3	62.0±7.5	70.0±6.3	66.0±7.7	74.0±6.3	68.0±7.7	76.0±6.0	56.0±8.6	60.0±7.5
task4	2.0±2.0	2.0±2.0	2.0±1.4	2.0±2.0	2.0±2.0	2.0±2.0	2.0±2.0	2.0±2.4	2.0±2.0
task5	68.0±5.5	64.0±6.0	64.0±5.5	70.0±6.2	72.0±5.3	66.0±6.5	68.0±5.5	66.0±6.8	66.0±6.0
cube-double									
task1	6.0±4.5	0.0±0.0	2.0±2.4	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task2	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task3	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task4	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task5	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
scene									
task1	94.0±2.4	88.0±3.5	96.0±2.0	100.0±0.0	100.0±0.0	90.0±3.5	94.0±2.4	96.0±2.4	98.0±1.4
task2	26.0±9.2	36.0±10.9	42.0±9.8	34.0±11.4	42.0±10.4	20.0±9.6	26.0±9.2	22.0±10.4	28.0±9.9
task3	54.0±4.9	56.0±5.3	66.0±4.5	70.0±5.3	72.0±4.5	48.0±5.5	60.0±4.7	52.0±6.0	56.0±5.1
task4	6.0±2.4	16.0±4.0	16.0±3.5	8.0±3.2	10.0±2.8	10.0±3.5	18.0±4.0	8.0±3.2	10.0±3.2
task5	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
puzzle-3x3									
task1	100.0±0.0	96.0±1.4	96.0±1.4	86.0±3.2	88.0±2.8	86.0±3.2	86.0±2.8	90.0±2.8	96.0±1.4
task2	4.0±2.8	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task3	8.0±2.8	4.0±2.0	4.0±2.0	4.0±2.4	4.0±2.0	6.0±2.8	6.0±2.4	6.0±2.8	6.0±2.4
task4	6.0±3.2	4.0±2.8	4.0±2.4	0.0±0.0	10.0±4.0	8.0±4.2	10.0±4.0	2.0±2.4	4.0±2.8
task5	12.0±3.2	16.0±4.0	24.0±4.0	12.0±3.7	34.0±4.7	2.0±1.4	8.0±2.8	10.0±3.5	12.0±3.5
puzzle-4x4									
task1	4.0±3.2	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task2	2.0±2.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task3	2.0±2.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task4	2.0±1.4	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task5	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0

Table 15: GFP: OGBench success percentages under attacks ($\epsilon = 0.1$) and Gaussian noise (scale $\epsilon = 0.2$). +RCF denotes DMBP+RCF with $\rho = 0.3$.

Task	Raw clean	Gaussian (0.2)		APGD (0.1)		min-Q (0.1)		BDAP (0.1)	
		DMBP	+RCF	DMBP	+RCF	DMBP	+RCF	DMBP	+RCF
antmaze-large									
task1	98.0±3.5	90.0±3.7	98.0±2.4	98.0±2.8	98.0±2.4	76.0±5.7	92.0±3.7	78.0±5.7	92.0±4.0
task2	92.0±4.9	82.0±5.3	94.0±3.7	66.0±6.6	74.0±5.7	30.0±6.8	38.0±6.3	56.0±7.3	70.0±6.5
task3	96.0±2.8	88.0±4.0	96.0±2.4	94.0±3.5	98.0±2.4	88.0±4.5	92.0±3.7	90.0±4.5	92.0±4.0
task4	90.0±4.0	80.0±5.1	78.0±4.7	64.0±6.2	80.0±5.1	74.0±6.2	78.0±5.3	62.0±6.8	72.0±5.8
task5	96.0±2.8	84.0±4.2	86.0±3.7	84.0±4.7	84.0±4.2	64.0±6.0	90.0±4.0	76.0±5.7	88.0±4.2
antmaze-giant									
task1	4.0±3.2	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task2	84.0±5.3	22.0±6.5	24.0±6.0	2.0±4.5	4.0±4.5	0.0±0.0	2.0±4.5	0.0±0.0	4.0±4.9
task3	28.0±4.7	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task4	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task5	8.0±8.9	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
antsoccer-arena									
task1	88.0±4.0	0.0±0.0	2.0±2.8	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task2	96.0±2.8	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task3	44.0±6.3	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task4	50.0±6.3	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task5	44.0±6.9	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
humanoidmaze-medium									
task1	82.0±6.5	60.0±6.8	70.0±5.8	0.0±0.0	0.0±0.0	32.0±7.7	40.0±7.2	0.0±0.0	0.0±0.0
task2	94.0±2.8	50.0±6.2	60.0±5.7	6.0±4.7	20.0±5.5	34.0±6.9	44.0±6.5	0.0±0.0	0.0±0.0
task3	98.0±4.5	80.0±7.1	80.0±6.3	26.0±8.2	34.0±7.7	70.0±8.8	76.0±7.7	0.0±0.0	2.0±4.5
task4	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task5	98.0±2.0	82.0±5.3	84.0±4.7	46.0±6.9	68.0±6.0	72.0±6.6	78.0±5.8	0.0±0.0	0.0±0.0
humanoidmaze-large									
task1	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task2	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task3	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task4	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task5	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
cube-single									
task1	100.0±0.0	88.0±3.7	96.0±2.4	100.0±0.0	100.0±0.0	98.0±2.4	100.0±0.0	96.0±3.2	100.0±0.0
task2	96.0±2.4	96.0±2.8	98.0±2.0	94.0±3.2	96.0±2.8	88.0±4.2	96.0±2.8	94.0±3.5	96.0±3.2
task3	98.0±2.0	98.0±2.0	98.0±2.0	98.0±2.4	98.0±2.0	92.0±3.7	94.0±3.2	92.0±4.0	98.0±2.4
task4	100.0±0.0	84.0±4.0	90.0±3.2	82.0±4.7	90.0±3.5	98.0±2.4	98.0±2.4	82.0±5.1	88.0±4.0
task5	96.0±2.8	86.0±4.0	94.0±2.8	86.0±4.2	92.0±3.2	88.0±4.2	92.0±3.5	76.0±5.5	80.0±4.9
cube-double									
task1	78.0±6.0	62.0±6.6	62.0±6.0	52.0±7.3	72.0±6.3	66.0±7.5	66.0±6.9	54.0±8.1	60.0±7.3
task2	54.0±6.3	32.0±6.6	38.0±6.0	36.0±7.3	40.0±6.8	50.0±7.9	56.0±7.1	16.0±6.9	22.0±6.8
task3	56.0±6.0	32.0±6.3	42.0±6.0	30.0±7.1	40.0±6.6	34.0±7.5	40.0±7.1	38.0±7.9	40.0±7.3
task4	6.0±2.8	6.0±4.0	8.0±4.0	2.0±3.5	2.0±3.5	2.0±3.5	4.0±4.0	0.0±0.0	0.0±0.0
task5	50.0±7.2	20.0±6.3	26.0±6.2	12.0±6.3	20.0±6.3	20.0±7.3	28.0±7.2	16.0±7.3	22.0±7.2
scene									
task1	98.0±2.0	100.0±0.0	100.0±0.0	100.0±0.0	100.0±0.0	100.0±0.0	100.0±0.0	100.0±0.0	100.0±0.0
task2	64.0±6.8	38.0±6.9	42.0±6.3	48.0±7.7	58.0±7.1	24.0±7.3	32.0±7.1	20.0±7.5	26.0±7.2
task3	84.0±4.0	72.0±5.5	74.0±4.9	72.0±5.8	80.0±5.1	76.0±6.0	76.0±5.5	76.0±6.3	84.0±5.3
task4	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task5	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
puzzle-3x3									
task1	98.0±2.4	90.0±4.0	94.0±3.2	96.0±3.5	96.0±3.2	82.0±5.3	86.0±4.7	82.0±5.7	90.0±4.5
task2	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task3	4.0±2.8	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task4	2.0±2.4	0.0±0.0	0.0±0.0	12.0±4.7	14.0±4.5	2.0±2.8	6.0±3.5	2.0±3.5	2.0±2.8
task5	2.0±2.0	14.0±4.5	14.0±4.0	18.0±5.1	24.0±5.1	2.0±2.8	4.0±3.2	2.0±3.5	4.0±3.5
puzzle-4x4									
task1	62.0±6.6	44.0±6.9	48.0±6.3	28.0±7.2	30.0±6.6	8.0±5.8	14.0±6.0	8.0±6.2	40.0±7.6
task2	26.0±5.5	2.0±3.5	8.0±4.2	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
task3	48.0±6.0	38.0±6.6	48.0±6.2	0.0±0.0	24.0±6.2	6.0±5.3	8.0±5.3	0.0±0.0	4.0±4.7
task4	30.0±5.3	8.0±4.5	10.0±4.2	6.0±4.7	6.0±4.2	6.0±4.9	8.0±4.7	0.0±0.0	0.0±0.0
task5	20.0±4.5	4.0±3.7	6.0±3.7	0.0±0.0	0.0±0.0	0.0±0.0	4.0±4.0	0.0±0.0	0.0±0.0

Table 16: FQL: defense results on individual D4RL tasks under attacks ($\epsilon = 0.1$) and Gaussian noise (scale $\epsilon = 0.2$). +RCF denotes DMBP+RCF with $\rho = 0.3$. Scores are success percentages for D4RL AntMaze and normalized returns for D4RL Adroit.

Task	Raw clean	Gaussian (0.2)		APGD (0.1)		min-Q (0.1)		BDAP (0.1)	
		DMBP	+RCF	DMBP	+RCF	DMBP	+RCF	DMBP	+RCF
AntMaze									
antmaze-umaze-v2	98.0±2.0	96.0±2.4	98.0±2.0	88.0±4.2	96.0±2.4	96.0±2.8	100.0±0.0	94.0±3.2	98.0±2.0
antmaze-umaze-diverse-v2	96.0±4.9	86.0±6.2	94.0±4.2	100.0±0.0	100.0±0.0	96.0±4.5	98.0±2.8	76.0±8.2	88.0±6.0
antmaze-medium-play-v2	70.0±6.9	74.0±7.7	78.0±6.6	82.0±7.1	84.0±6.3	48.0±10.0	64.0±8.5	0.0±0.0	0.0±0.0
antmaze-medium-diverse-v2	70.0±13.0	82.0±12.0	96.0±6.8	76.0±13.0	78.0±11.5	36.0±15.2	54.0±13.0	76.0±13.8	80.0±11.7
antmaze-large-play-v2	86.0±6.9	84.0±7.5	90.0±5.8	76.0±9.2	80.0±7.7	62.0±10.8	64.0±9.6	72.0±9.5	78.0±8.4
antmaze-large-diverse-v2	82.0±4.0	82.0±5.1	86.0±4.2	84.0±5.1	88.0±4.5	68.0±6.8	74.0±6.0	58.0±7.5	70.0±6.2
Pen									
pen-human-v1	55.5±6.0	63.3±8.8	67.8±7.9	53.0±9.6	55.1±8.7	56.5±9.4	57.7±8.6	54.7±10.0	58.2±9.1
pen-cloned-v1	76.2±11.0	78.3±11.8	85.4±10.2	76.8±12.0	80.2±10.8	73.7±12.8	79.5±11.0	76.4±13.0	78.4±11.5
pen-expert-v1	148.0±6.0	149.9±6.8	149.3±6.3	145.6±7.8	146.1±7.0	142.0±8.2	144.1±7.4	131.7±10.2	138.9±8.8
Hammer									
hammer-human-v1	0.8±1.0	0.4±0.8	1.0±1.1	0.9±1.0	1.1±1.1	0.4±0.8	1.0±1.1	1.2±1.2	1.8±1.4
hammer-cloned-v1	17.9±9.0	8.0±8.2	12.2±8.4	7.5±8.0	8.7±7.5	7.1±8.3	11.3±8.0	0.6±3.8	3.9±5.5
hammer-expert-v1	115.6±3.0	106.5±5.0	115.5±4.0	106.6±5.4	115.9±4.0	116.7±4.5	118.1±4.0	106.9±5.8	114.7±4.8
Door									
door-human-v1	0.1±0.3	0.8±0.9	3.7±2.0	1.2±1.2	1.7±1.4	0.6±0.8	3.7±2.0	0.9±1.0	3.2±1.8
door-cloned-v1	3.4±1.0	2.6±1.2	5.0±1.8	3.0±1.3	3.0±1.2	2.1±1.2	2.3±1.2	1.1±0.9	1.2±0.9
door-expert-v1	102.8±1.0	103.2±1.6	103.3±1.4	103.2±1.7	103.3±1.5	102.5±2.0	103.0±1.7	100.8±2.4	101.2±2.1
Relocate									
relocate-human-v1	0.1±0.2	0.1±0.2	0.1±0.2	0.1±0.2	0.1±0.2	0.1±0.2	0.1±0.2	0.1±0.2	0.1±0.2
relocate-cloned-v1	-0.1±0.2	-0.1±0.2	-0.1±0.2	-0.1±0.2	-0.1±0.2	-0.1±0.2	-0.1±0.2	-0.1±0.2	-0.1±0.2
relocate-expert-v1	107.3±1.0	44.4±14.0	64.0±12.0	42.9±14.5	67.2±11.5	46.2±13.8	68.9±11.0	42.3±15.0	66.7±11.8

Table 17: MeanFlowQL: defense results on individual D4RL tasks under attacks ($\epsilon = 0.1$) and Gaussian noise (scale $\epsilon = 0.2$). +RCF denotes DMBP+RCF with $\rho = 0.3$. Scores are success percentages for D4RL AntMaze and normalized returns for D4RL Adroit.

Task	Raw clean	Gaussian (0.2)		APGD (0.1)		min-Q (0.1)		BDAP (0.1)	
		DMBP	+RCF	DMBP	+RCF	DMBP	+RCF	DMBP	+RCF
AntMaze									
antmaze-umaze-v2	92.0±2.8	74.0±4.9	76.0±4.5	58.0±6.0	66.0±5.3	66.0±5.5	76.0±4.7	76.0±4.9	80.0±4.5
antmaze-umaze-diverse-v2	96.0±4.5	10.0±5.5	10.0±4.9	28.0±7.2	30.0±6.6	18.0±6.5	18.0±6.0	6.0±4.5	8.0±4.5
antmaze-medium-play-v2	94.0±5.2	74.0±7.5	82.0±6.5	78.0±7.2	84.0±6.0	70.0±7.7	78.0±6.8	0.0±0.0	0.0±0.0
antmaze-medium-diverse-v2	94.0±7.0	24.0±12.8	40.0±11.0	20.0±11.5	32.0±10.5	28.0±12.8	32.0±11.2	36.0±13.5	40.0±12.0
antmaze-large-play-v2	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	2.0±3.5	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
antmaze-large-diverse-v2	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
Pen									
pen-human-v1	-1.4±0.8	-0.2±0.7	1.0±0.9	-1.3±1.0	-0.7±0.9	-1.1±1.0	0.5±0.8	-0.7±1.0	-0.7±0.9
pen-cloned-v1	-1.3±0.9	-3.2±1.2	-2.4±1.1	-2.6±1.3	-2.6±1.2	-3.3±1.4	-2.8±1.2	-2.6±1.3	-2.4±1.2
pen-expert-v1	20.3±5.5	11.7±5.4	12.8±5.1	10.4±5.5	11.1±5.1	10.9±5.6	11.5±5.2	5.4±4.6	9.1±5.0
Hammer									
hammer-human-v1	-0.1±0.1	-0.1±0.1	-0.2±0.2	-0.1±0.1	-0.1±0.1	-0.1±0.1	-0.1±0.1	-0.1±0.1	-0.1±0.1
hammer-cloned-v1	0.1±0.2	0.1±0.2	0.1±0.2	0.1±0.2	0.1±0.2	0.1±0.2	0.1±0.2	0.1±0.2	0.1±0.2
hammer-expert-v1	0.1±0.2	0.1±0.2	0.1±0.2	0.1±0.2	0.1±0.2	0.1±0.2	0.1±0.2	0.1±0.2	0.1±0.2
Door									
door-human-v1	-0.2±0.1	0.0±0.2	0.0±0.2	0.0±0.2	0.0±0.2	0.0±0.2	0.0±0.2	0.0±0.2	0.1±0.2
door-cloned-v1	-0.2±0.1	0.1±0.2	0.1±0.2	0.0±0.2	0.1±0.2	0.0±0.2	0.1±0.2	0.1±0.2	0.1±0.2
door-expert-v1	-0.2±0.1	-0.2±0.1	-0.2±0.1	-0.2±0.1	-0.2±0.1	-0.2±0.1	-0.2±0.1	-0.2±0.1	-0.2±0.1
Relocate									
relocate-human-v1	-0.1±0.1	-0.1±0.1	-0.1±0.1	-0.1±0.1	-0.1±0.1	-0.1±0.1	-0.1±0.1	-0.1±0.1	-0.1±0.1
relocate-cloned-v1	-0.1±0.1	-0.1±0.1	-0.1±0.1	-0.2±0.1	-0.1±0.1	-0.1±0.1	-0.1±0.1	-0.1±0.1	-0.1±0.1
relocate-expert-v1	-0.2±0.1	-0.2±0.1	-0.2±0.1	-0.2±0.1	-0.2±0.1	-0.2±0.1	-0.2±0.1	-0.2±0.1	-0.2±0.1

Table 18: SSCP: defense results on individual D4RL tasks under attacks ($\epsilon = 0.1$) and Gaussian noise (scale $\epsilon = 0.2$). +RCF denotes DMBP+RCF with $\rho = 0.3$. Scores are success percentages for D4RL AntMaze and normalized returns for D4RL Adroit.

Task	Raw clean	Gaussian (0.2)		APGD (0.1)		min-Q (0.1)		BDAP (0.1)	
		DMBP	+RCF	DMBP	+RCF	DMBP	+RCF	DMBP	+RCF
AntMaze									
antmaze-umaze-v2	96.0±2.4	56.0±6.0	60.0±5.5	50.0±6.5	52.0±5.8	54.0±6.2	64.0±5.5	62.0±6.5	68.0±5.7
antmaze-umaze-diverse-v2	86.0±5.5	46.0±7.5	48.0±6.8	56.0±7.2	58.0±6.6	48.0±7.6	52.0±6.9	90.0±5.3	92.0±4.5
antmaze-medium-play-v2	4.0±4.2	4.0±4.5	2.0±3.5	2.0±3.5	2.0±2.8	2.0±3.5	10.0±4.9	0.0±0.0	0.0±0.0
antmaze-medium-diverse-v2	30.0±12.5	0.0±0.0	2.0±4.5	2.0±4.5	2.0±3.5	2.0±4.5	4.0±5.1	2.0±4.5	2.0±4.5
antmaze-large-play-v2	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
antmaze-large-diverse-v2	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0
Pen									
pen-human-v1	73.2±7.0	58.6±9.0	61.1±8.2	52.7±9.8	60.0±8.8	58.9±9.5	63.6±8.6	53.5±10.0	54.5±9.2
pen-cloned-v1	76.1±10.5	76.7±11.3	78.1±10.2	63.7±12.8	64.7±11.7	68.4±12.2	73.3±11.0	68.8±12.5	73.3±11.2
pen-expert-v1	147.3±6.0	146.0±6.6	149.4±6.1	136.6±8.4	144.5±7.2	137.8±8.1	144.2±7.0	138.8±8.6	146.6±6.8
Hammer									
hammer-human-v1	0.0±0.2	0.0±0.2	0.0±0.2	0.0±0.2	0.1±0.2	0.1±0.2	0.2±0.3	0.1±0.2	0.3±0.4
hammer-cloned-v1	4.5±4.5	0.7±1.9	1.7±2.7	0.5±1.6	1.2±2.2	1.6±2.8	1.9±2.8	0.3±1.3	0.4±1.4
hammer-expert-v1	147.0±4.0	121.0±6.5	128.5±5.6	120.8±6.8	126.6±5.9	120.6±7.0	128.1±5.8	120.9±7.2	130.7±5.7
Door									
door-human-v1	-0.2±0.1	-0.1±0.2	0.0±0.2	-0.1±0.2	0.0±0.2	-0.1±0.2	0.0±0.2	0.0±0.2	0.0±0.2
door-cloned-v1	0.0±0.2	0.0±0.2	1.2±0.8	0.0±0.2	0.2±0.3	0.0±0.2	0.0±0.2	0.0±0.2	0.0±0.2
door-expert-v1	105.3±1.1	93.4±4.5	101.6±2.5	105.1±1.9	105.4±1.7	93.4±4.7	102.5±2.3	101.7±2.7	103.2±2.1
Relocate									
relocate-human-v1	-0.2±0.1	-0.1±0.1	-0.1±0.1	-0.1±0.1	-0.1±0.1	-0.1±0.1	-0.1±0.1	-0.1±0.1	-0.1±0.1
relocate-cloned-v1	-0.1±0.2	-0.2±0.3	0.0±0.2	-0.1±0.2	-0.1±0.2	-0.1±0.2	-0.1±0.2	-0.1±0.2	0.0±0.2
relocate-expert-v1	113.7±1.5	16.7±10.5	28.7±10.0	14.0±9.8	25.2±9.3	12.5±9.5	27.3±9.6	14.1±10.0	27.0±9.4

Table 19: GFP: defense results on individual D4RL tasks under attacks ($\epsilon = 0.1$) and Gaussian noise (scale $\epsilon = 0.2$). +RCF denotes DMBP+RCF with $\rho = 0.3$. Scores are success percentages for D4RL AntMaze and normalized returns for D4RL Adroit.

Task	Raw clean	Gaussian (0.2)		APGD (0.1)		min-Q (0.1)		BDAP (0.1)	
		DMBP	+RCF	DMBP	+RCF	DMBP	+RCF	DMBP	+RCF
AntMaze									
antmaze-umaze-v2	98.0±2.0	88.0±4.0	92.0±3.5	92.0±3.7	92.0±3.5	90.0±4.2	98.0±2.4	92.0±4.0	92.0±3.5
antmaze-umaze-diverse-v2	98.0±4.5	86.0±6.5	86.0±5.8	86.0±6.8	90.0±5.5	86.0±6.6	94.0±4.7	62.0±8.5	66.0±7.7
antmaze-medium-play-v2	98.0±4.5	88.0±6.9	94.0±5.8	84.0±7.7	86.0±6.8	88.0±7.2	94.0±5.8	0.0±0.0	0.0±0.0
antmaze-medium-diverse-v2	90.0±13.0	76.0±13.5	88.0±10.5	82.0±12.0	88.0±10.0	80.0±12.8	86.0±10.3	88.0±11.0	90.0±9.5
antmaze-large-play-v2	92.0±6.9	78.0±8.5	82.0±7.5	80.0±8.4	84.0±7.2	68.0±9.7	70.0±8.8	74.0±8.9	80.0±7.7
antmaze-large-diverse-v2	92.0±4.0	80.0±5.7	82.0±4.9	84.0±5.3	86.0±4.7	76.0±6.5	82.0±5.5	66.0±7.2	78.0±6.0
Pen									
pen-human-v1	79.2±6.0	74.9±8.2	82.0±7.3	63.7±9.3	68.4±8.4	72.8±8.8	75.6±8.0	56.4±10.0	59.6±9.2
pen-cloned-v1	98.0±11.0	70.8±12.2	79.7±10.8	90.8±10.0	92.8±9.0	71.3±12.5	80.4±10.8	77.1±12.8	81.0±11.2
pen-expert-v1	158.4±6.0	137.0±8.0	145.1±7.0	136.1±8.4	143.0±7.3	132.2±8.8	142.3±7.4	142.5±8.0	143.8±7.2
Hammer									
hammer-human-v1	2.3±1.0	0.6±0.8	1.2±1.0	1.1±1.0	1.3±1.1	0.8±0.9	1.3±1.1	1.2±1.1	2.1±1.4
hammer-cloned-v1	6.0±9.0	3.9±7.5	4.3±6.8	2.7±6.3	3.1±5.8	5.9±8.0	6.1±7.1	3.6±7.0	8.2±8.0
hammer-expert-v1	129.6±3.0	121.2±4.8	125.0±4.2	125.3±4.8	125.3±4.3	124.5±5.0	124.5±4.4	123.8±5.4	124.1±4.7
Door									
door-human-v1	5.5±3.0	1.9±1.6	3.4±2.0	0.8±1.1	2.9±1.8	1.3±1.4	3.7±2.1	1.9±1.7	2.0±1.6
door-cloned-v1	0.1±0.3	0.4±0.5	0.0±0.2	0.1±0.3	0.0±0.2	0.0±0.2	0.0±0.2	0.0±0.2	0.0±0.2
door-expert-v1	112.0±1.0	103.6±1.8	103.6±1.6	103.4±1.9	103.5±1.7	103.3±2.0	103.3±1.8	100.2±2.5	100.8±2.2
Relocate									
relocate-human-v1	1.8±1.2	0.4±0.6	0.5±0.6	0.3±0.5	0.6±0.7	0.4±0.6	0.7±0.7	0.3±0.5	0.6±0.7
relocate-cloned-v1	2.5±1.5	0.8±0.9	0.8±0.8	0.2±0.5	0.7±0.8	0.1±0.4	0.2±0.5	0.1±0.4	0.6±0.7
relocate-expert-v1	117.7±1.0	39.4±14.0	63.1±12.0	35.0±14.5	54.6±12.4	33.9±14.2	54.5±12.0	30.9±14.8	52.9±12.3

C.4 Detailed Budget and Correction-Strength Results

Table 20 provides the numerical results for Figure 2 and extends the comparison to $\rho \in \{0.3, 0.5, 1.0\}$ and Gaussian noise. All settings use the same OGBench victim–task combinations, with success rates reported on the $[0, 1]$ scale. These analyses use a fixed subset of 94 combinations with raw-clean success of at least 10% in the initial evaluation

Among the tested coefficients, $\rho = 0.3$ yields the highest pooled success under all three attacks at $\epsilon = 0.1$. At $\epsilon = 0.3$, stronger correction performs better: $\rho = 1.0$ yields 0.244 under APGD and 0.153 under BDAP, compared with DMBP's 0.213 and 0.124, respectively. For min-Q, $\rho = 0.5$ yields the highest pooled success of 0.250, compared with 0.226 for DMBP. Under Gaussian noise, $\rho = 0.3$ improves pooled success from 0.480 to 0.520, while $\rho = 0.5$ and $\rho = 1.0$ yield 0.492 and 0.401. These results favor stronger correction for some larger-budget attacks and a smaller correction coefficient for Gaussian noise.

Table 20: OGBench success rates across correction strengths. $\rho = 0$ denotes DMBP alone, and $\rho > 0$ denotes DMBP+RCF. Bold marks the highest score within each perturbation setting and column, including ties.

Perturbation	ϵ	ρ	FQL	MeanFlowQL	SSCP	GFP	Pooled
Raw clean			0.627 ±0.026	0.586 ±0.031	0.534 ±0.028	0.746 ±0.022	0.651 ±0.014
Gaussian noise	0.2	0	0.464 ±0.028	0.478 ±0.030	0.460 ±0.029	0.504 ±0.026	0.480 ±0.014
		0.3	**0.502 ±0.024**	**0.528 ±0.026**	**0.498 ±0.025**	**0.543 ±0.023**	**0.520 ±0.012**
		0.5	0.462 ±0.026	0.517 ±0.028	0.486 ±0.027	0.515 ±0.025	0.492 ±0.013
		1.0	0.374 ±0.030	0.434 ±0.032	0.454 ±0.029	0.396 ±0.028	0.401 ±0.015
APGD	0.1	0	0.392 ±0.031	0.363 ±0.034	0.428 ±0.029	0.419 ±0.032	0.403 ±0.016
		0.3	**0.432 ±0.027**	**0.437 ±0.030**	**0.482 ±0.026**	**0.471 ±0.028**	**0.454 ±0.014**
		0.5	0.408 ±0.029	0.426 ±0.031	0.454 ±0.027	0.438 ±0.030	0.427 ±0.015
		1.0	0.335 ±0.033	0.371 ±0.035	0.442 ±0.028	0.344 ±0.033	0.358 ±0.017
	0.2	0	0.239 ±0.027	0.188 ±0.030	0.403 ±0.025	0.239 ±0.028	0.255 ±0.014
		0.3	0.258 ±0.025	0.208 ±0.028	0.409 ±0.024	0.271 ±0.026	0.277 ±0.013
		0.5	**0.277 ±0.024**	0.215 ±0.027	**0.417 ±0.023**	**0.278 ±0.025**	**0.288 ±0.012**
		1.0	0.257 ±0.027	**0.258 ±0.030**	0.400 ±0.025	0.261 ±0.027	0.279 ±0.014
	0.3	0	0.192 ±0.024	0.168 ±0.026	0.366 ±0.023	0.194 ±0.025	0.213 ±0.012
		0.3	0.215 ±0.023	0.171 ±0.025	**0.409 ±0.021**	0.199 ±0.024	0.230 ±0.012
		0.5	0.216 ±0.022	0.169 ±0.024	0.405 ±0.021	0.208 ±0.023	0.233 ±0.011
		1.0	**0.230 ±0.024**	**0.191 ±0.026**	0.405 ±0.022	**0.216 ±0.024**	**0.244 ±0.012**
min-Q	0.1	0	0.361 ±0.030	0.457 ±0.031	0.386 ±0.029	0.434 ±0.031	0.404 ±0.015
		0.3	**0.405 ±0.027**	**0.511 ±0.027**	0.426 ±0.026	**0.479 ±0.027**	**0.449 ±0.013**
		0.5	0.383 ±0.028	0.485 ±0.029	**0.442 ±0.025**	0.464 ±0.028	0.434 ±0.014
		1.0	0.316 ±0.032	0.411 ±0.033	0.432 ±0.028	0.341 ±0.032	0.354 ±0.016
	0.2	0	0.239 ±0.026	0.346 ±0.029	0.375 ±0.025	0.281 ±0.027	0.288 ±0.013
		0.3	**0.261 ±0.024**	0.372 ±0.027	0.408 ±0.024	0.285 ±0.025	0.306 ±0.012
		0.5	0.255 ±0.024	**0.392 ±0.026**	**0.422 ±0.023**	**0.304 ±0.024**	**0.315 ±0.012**
		1.0	0.241 ±0.026	0.338 ±0.029	**0.422 ±0.024**	0.274 ±0.026	0.291 ±0.013
	0.3	0	0.200 ±0.023	0.212 ±0.026	0.334 ±0.023	0.216 ±0.024	0.226 ±0.012
		0.3	0.201 ±0.022	0.245 ±0.024	0.380 ±0.022	0.224 ±0.023	0.240 ±0.011
		0.5	**0.209 ±0.022**	**0.249 ±0.024**	0.382 ±0.021	**0.241 ±0.022**	**0.250 ±0.011**
		1.0	0.205 ±0.023	0.237 ±0.025	**0.414 ±0.022**	0.232 ±0.023	0.248 ±0.012
BDAP	0.1	0	0.297 ±0.030	0.174 ±0.032	0.397 ±0.029	0.356 ±0.031	0.315 ±0.015
		0.3	**0.326 ±0.027**	0.206 ±0.030	0.443 ±0.026	**0.403 ±0.027**	**0.354 ±0.013**
		0.5	**0.326 ±0.027**	0.232 ±0.029	**0.449 ±0.025**	0.389 ±0.028	0.353 ±0.014
		1.0	0.289 ±0.031	**0.268 ±0.032**	0.429 ±0.028	0.314 ±0.031	0.315 ±0.016
	0.2	0	0.181 ±0.025	0.171 ±0.027	0.303 ±0.024	0.180 ±0.026	0.196 ±0.013
		0.3	0.185 ±0.024	0.171 ±0.026	0.332 ±0.023	0.212 ±0.025	0.213 ±0.012
		0.5	0.195 ±0.023	0.155 ±0.025	**0.352 ±0.022**	0.223 ±0.024	0.221 ±0.012
		1.0	**0.223 ±0.025**	**0.197 ±0.027**	0.351 ±0.023	**0.235 ±0.025**	**0.241 ±0.013**
	0.3	0	0.116 ±0.021	0.114 ±0.023	0.145 ±0.021	0.128 ±0.022	0.124 ±0.011
		0.3	0.132 ±0.020	0.118 ±0.022	0.168 ±0.020	0.126 ±0.021	0.133 ±0.010
		0.5	0.135 ±0.020	**0.140 ±0.021**	0.186 ±0.020	0.135 ±0.021	0.143 ±0.010
		1.0	**0.142 ±0.021**	0.123 ±0.022	**0.237 ±0.021**	**0.144 ±0.022**	**0.153 ±0.011**