SAMPLE-ANCHORED ADVANTAGE WEIGHTING FOR FLOW POLICY LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Offline-to-online reinforcement learning combines offline pretraining with online fine-tuning, resulting in replay buffers that mix offline data from diverse behavior policies with online experience from the evolving policy. Flow policies are well suited to capturing multimodal action distributions through iterative generation, but their implicit action distributions make conventional RL objectives difficult to apply, leaving behavior cloning as a natural yet limited training strategy. To improve performance, advantage-weighted regression (AWR) retains the imitation objective while weighting it according to estimated advantages. However, this weighting compares actions across modes on a shared value scale, which can suppress valid behavior modes whose estimated values are temporarily lower and thereby reduce behavioral diversity. To address this issue, we propose **S**ample-anchored **A**dvantage-**W**eighted **R**egression (SAWR), a variant of AWR for flow policies that replaces conventional global value comparison with a sample-specific local comparison against a reference action constructed from each replay action. Specifically, SAWR constructs the reference action by partially noising the replay action and completing it with the flow policy, then weights the flow matching loss by the value difference between the replay action and its reference. This sample-specific comparison uses local value differences to guide policy learning while mitigating the suppression of valid behavior modes. Integrated into Flow Q-Learning, SAWR improves average online success rates over conventional AWR on OGBench and D4RL Adroit. Navigation and manipulation diagnostics further demonstrate better preservation of diverse successful behaviors, while a route-blockage test shows improved robustness relative to conventional AWR when previously preferred behaviors become unavailable. Code is provided in the supplementary material.

1 INTRODUCTION

Reinforcement learning (RL) has demonstrated strong performance in sequential decision-making tasks across various domains (Hafner et al., 2025; Kaufmann et al., 2023), but its reliance on extensive online interaction limits practical deployment. Offline RL reduces this cost by learning from previously collected experience, yet its performance is constrained by dataset coverage and quality. Offline-to-online (O2O) RL combines offline pretraining with online adaptation, enabling efficient policy improvement while retaining pretrained knowledge (Nair et al., 2020).

Most existing O2O RL methods rely on unimodal Gaussian policies, which struggle to capture multimodal action distributions in replay buffers combining diverse offline behaviors and online interactions. This mismatch can produce averaged actions or prevent the policy from representing distinct behavior modes, discarding behaviors that may be useful during adaptation. Generative policies, including diffusion and flow policies, capture such distributions through iterative generation (Chi et al., 2023; Wang et al., 2023; Park et al., 2025b). However, their common reliance on behavior cloning limits improvement beyond the data, while their implicit action distributions complicate conventional RL methods requiring tractable policy densities or action optimization. Existing approaches therefore develop tailored objectives and optimization procedures for generative policies (Wang et al., 2023; Park et al., 2025b; Lyu et al., 2025; McAllister et al., 2026; Zhang et al., 2025).

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.

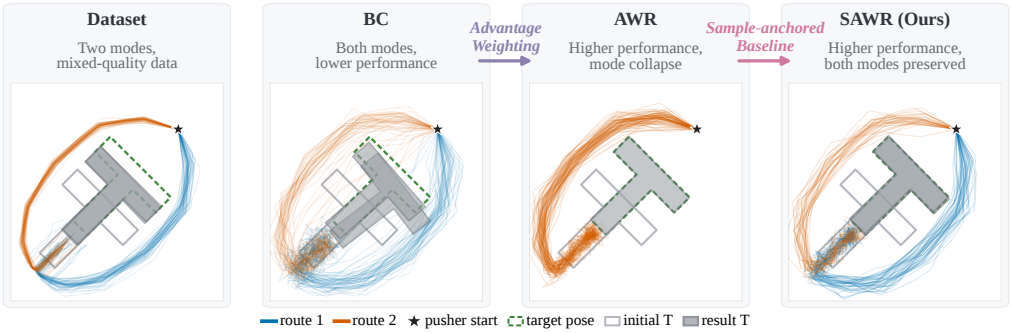


Figure 1: **AWR improves task performance at the cost of behavioral diversity.** In a controlled PushT experiment, the dataset contains two approach routes with trajectories of differing quality. BC preserves both modes but achieves lower performance, while AWR improves performance but discards one route. SAWR uses a sample-anchored baseline to mitigate this bias, improving performance while maintaining both modes.

Meanwhile, advantage-weighted regression computes an estimated advantage as $A_V(s,a) = Q(s,a) - V(s)$ and weights the behavior cloning loss for each replay action by the exponential of this advantage (Peng et al., 2019; Nair et al., 2020), improving performance over unweighted behavior cloning. For flow policies, it can be applied directly to the conditional flow matching loss without backpropagating through the iterative action generation process. However, conventional AWR places replay actions from different behavior modes on a common exponential weighting scale. Modes with temporarily higher estimated action values therefore receive systematically larger training weights, which can prematurely suppress other valid modes before their relative values are reliably established, limiting subsequent exploration and gradually eroding the intrinsic multimodality of the flow policy. Figure 1 illustrates this effect in a controlled PushT setting: with mixed-quality data covering two approach routes, AWR improves performance over behavior cloning but concentrates the policy's behavior on a single route.

We address this problem with **S**ample-anchored **A**dvantage-**W**eighted **R**egression (SAWR), a variant of advantage-weighted regression for flow policies. SAWR replaces the state value baseline used by conventional AWR, which is shared across actions at a given state, with the value of a local reference action tailored to each replay action. It constructs the reference by partially noising the replay action and completing it with the flow policy, then uses the resulting value difference between the replay action and reference action to weight the conditional flow matching loss. This sample-specific comparison favors local actions with higher value while preserving diverse behavior modes. As shown in the right panel of Figure 1, SAWR improves performance while retaining both action modes.

Our contributions are threefold: (i) We analyze how conventional AWR reweights behavior modes according to their estimated advantage and empirically demonstrate that this weighting can suppress valid modes in flow policies. (ii) We introduce SAWR, which constructs a reference action through partial noising and flow completion and uses a sample-specific value difference to weight the conditional flow matching loss, improving policy performance while mitigating mode suppression. (iii) We integrate SAWR into FQL and demonstrate improved performance on OGBench and D4RL Adroit. Navigation and manipulation diagnostics show better preservation of diverse successful behaviors, while a route-blockage test shows improved robustness compared with conventional AWR.

2 Background

2.1 Preliminaries

Off-policy reinforcement learning. We consider a Markov decision process (MDP) $(\mathcal{S}, \mathcal{A}, P, r, \gamma)$ with continuous action space $\mathcal{A} = [-1,1]^{d_a}$. The goal is to maximize the discounted return $J(\pi) = \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k r_k\right]$. We denote the Euclidean projection onto $\mathcal{A}$ by $\Pi_{\mathcal{A}}$. The policy is trained off-policy

using a replay buffer $\mathcal{D} = \{(s_i, a_i, r_i, s'_i)\}_{i=1}^N$, which may be fixed or continually updated during training. Because the buffer can aggregate transitions collected by heterogeneous behavior policies, its conditional action distribution $\pi_{\text{data}}(a|s)$ may be highly multimodal.

Flow policies. A flow policy models a vector field $u_\theta(s, x, \tau)$ and induces a conditional action distribution $\pi_\theta(a|s)$ through the ordinary differential equation $\mathrm{d}x_\tau/\mathrm{d}\tau = u_\theta(s, x_\tau, \tau)$, which transports samples from a base distribution $x_0 \sim \mathcal{N}(0, I)$ to actions $x_1 = a$. Given a state–action pair $(s, a) \sim \mathcal{D}$, a base sample $\varepsilon \sim \mathcal{N}(0, I)$, and a flow time $\tau \sim U[0, 1]$, conditional flow matching considers the linear interpolant $x_\tau = (1 - \tau)\varepsilon + \tau a$ and optimizes

$$\mathcal{L}_{\text{FM}}(\theta) = \mathbb{E}\left[\|u_\theta(s, x_\tau, \tau) - (a - \varepsilon)\|_2^2\right]. \tag{1}$$

Throughout, we use τ to denote flow time and reserve t for the reference construction introduced in Section 3.

Advantage-weighted regression. AWR incorporates value information into behavior cloning by weighting each replay action by $w = \exp(\beta A(s, a))$ with inverse temperature $\beta > 0$ (Peng et al., 2019; Nair et al., 2020). With the advantage $A_V(s, a) = Q(s, a) - V(s)$, where $V(s) = \mathbb{E}_{a\sim\pi(\cdot|s)}[Q(s, a)]$, a weighted regression model fits

$$\pi(a|s) \propto \pi_{\text{data}}(a|s) \exp(\beta A_V(s, a)). \tag{2}$$

For flow policies, the same principle applies by multiplying each flow matching loss in Equation 1 by $w(s, a)$, thereby incorporating value information without differentiating through the flow generation trajectory.

2.2 Problem: Global weighting across behavior modes

Consider a state s whose action distribution in the replay buffer comprises multiple behavior modes:

$$\pi_{\text{data}}(a|s) = \sum_{m=1}^{M} \omega_m(s)\, p_m(a|s), \qquad \sum_{m=1}^{M} \omega_m(s) = 1, \tag{3}$$

where $\omega_m(s)$ is the probability mass of mode m, and $p_m(\cdot|s)$ is its conditional action distribution, supported on a measurable set $\mathcal{M}_m(s) \subseteq \mathcal{A}$. We assume that the mode sets $\{\mathcal{M}_m(s)\}_{m=1}^M$ are pairwise disjoint. For $\beta > 0$, define the entropic value of mode m as

$$\mathcal{Q}_m^{(\beta)}(s) := \frac{1}{\beta} \log \mathbb{E}_{a\sim p_m(\cdot|s)}[\exp(\beta Q(s, a))]. \tag{4}$$

Let $\omega_m^{\text{AWR}}(s)$ denote the mass assigned to mode m by Equation 2. Because the state value baseline $V(s)$ is shared by all actions at state s, it cancels when comparing the relative masses of any two modes i and j with positive probability:

$$\frac{\omega_i^{\text{AWR}}(s)/\omega_j^{\text{AWR}}(s)}{\omega_i(s)/\omega_j(s)} = \frac{\mathbb{E}_{a\sim p_i(\cdot|s)}[\exp(\beta Q(s, a))]}{\mathbb{E}_{a\sim p_j(\cdot|s)}[\exp(\beta Q(s, a))]} = \exp\Big(\beta\left[\mathcal{Q}_i^{(\beta)}(s) - \mathcal{Q}_j^{(\beta)}(s)\right]\Big). \tag{5}$$

Equation 5 shows that the state value does not affect relative mode mass at a fixed state. This global comparison is less problematic for unimodal Gaussian policies, which represent a single behavior mode, but can be harmful for multimodal policies. In practice, even behavior modes of comparable quality can receive different value estimates due to variations in data quality and training error. Conventional AWR compares all modes at a state on a common scale through exponentiated action values, which can amplify such estimation errors, causing modest underestimation to translate into disproportionately smaller weights. As a result, valid behavior modes may be progressively suppressed, reducing behavioral diversity and robustness. Section 4.2.2 evaluates these effects.

3 Method

To enable effective reinforcement learning for flow policies while maintaining their inherent multimodal behavior, SAWR replaces the state value baseline of AWR with a reference action value baseline constructed for each replay action, turning the global comparison across modes into a comparison based primarily on local value differences. Figure 2 illustrates the construction and its intended effect on mode preservation. SAWR modifies the weighting of the flow matching loss while leaving the surrounding RL objectives unchanged.

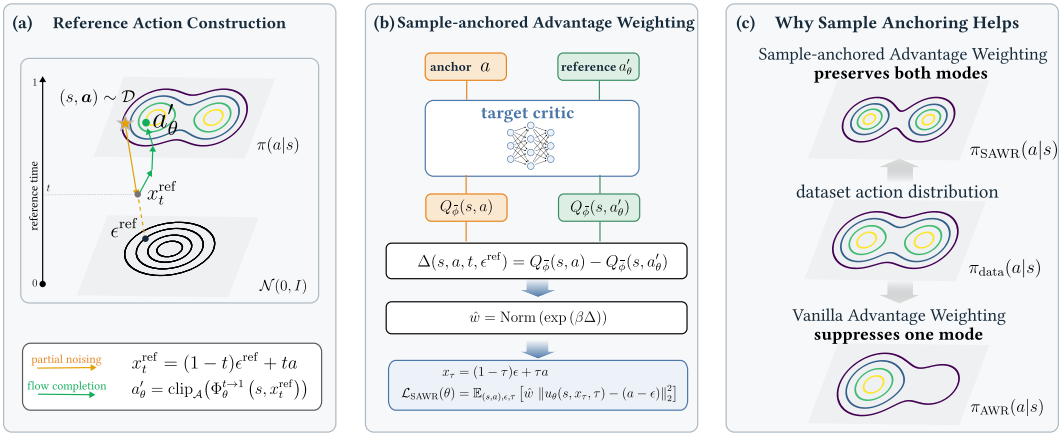


Figure 2: **Overview of sample-anchored advantage-weighted regression (SAWR). (a)** A reference action is constructed through partial noising and flow completion. **(b)** The resulting sample-anchored advantage produces a sample-specific weight for the flow matching loss. **(c)** When the reference remains in the anchor mode, the advantage removes the offset shared by that mode and avoids a direct comparison with other modes.

3.1 Reference Action Construction

To evaluate each replay action against a more local baseline and thereby reduce cross-mode competition, SAWR constructs a reference action coupled to the replay action, as illustrated in Figure 2(a). During training, given a state–action pair $(s, a) \sim \mathcal{D}$ drawn from the replay buffer, sample a reference time $t \sim P_t$ and independent Gaussian noise $\varepsilon^{\text{ref}} \sim \mathcal{N}(0, I)$, then partially noise the replay action to obtain the intermediate flow state:

$$x_t^{\text{ref}} = (1-t)\varepsilon^{\text{ref}} + ta. \tag{6}$$

Starting from (x_t^{ref}, t), we integrate the current flow policy to $t = 1$ conditioned on the same state s:

$$a'_\theta(s, a, t, \varepsilon^{\text{ref}}) = \Pi_{\mathcal{A}}\left(\Phi_\theta^{t\to 1}\left(s, x_t^{\text{ref}}\right)\right), \tag{7}$$

where $\Phi_\theta^{t\to 1}$ denotes the flow map induced by the vector field u_θ, approximated using K Euler steps. We refer to a as the **anchor action** and to a'_θ as its **reference action**. Analogous to SDEdit (Meng et al., 2022), partial noising followed by flow completion leverages the learned generative structure while retaining information from the anchor, thereby biasing a'_θ toward the anchor's local action neighborhood rather than producing an independent policy sample.

3.2 Sample-anchored Advantage Weighting

Sample-anchored advantage. As shown in Figure 2(b), with the reference action in place, its action value serves as the local baseline for evaluating the anchor. We define the resulting **sample-anchored advantage** as

$$\Delta(s, a; t, \varepsilon^{\text{ref}}) = Q_{\bar{\phi}}(s, a) - Q_{\bar{\phi}}\left(s, a'_\theta(s, a, t, \varepsilon^{\text{ref}})\right). \tag{8}$$

Here $Q_{\bar{\phi}}$ denotes the target critic, and Δ measures the advantage of the anchor a over the reference a'_θ. Thus, the action value of the reference action serves as a sample-specific baseline for weighting each replay action. Proposition 1 formalizes when the reference remains in the anchor's mode and shows that the resulting advantage cancels value offsets shared by that mode.

Proposition 1 (Sufficient condition for the reference to remain in the same mode)**.** *Let $U_{\max}$ uniformly bound the flow velocity norm along reference rollouts for all reference times, and let L_Q be a Lipschitz constant of $Q_{\bar{\phi}}(s, \cdot)$ on $\mathcal{A}$. Define*

$$R_t := (1-t)\left(\|\varepsilon^{\text{ref}} - a\|_2 + U_{\max}\right). \tag{9}$$

Then

$$\left\|a'_\theta(s, a, t, \varepsilon^{\text{ref}}) - a\right\|_2 \le R_t, \qquad \left|\Delta(s, a; t, \varepsilon^{\text{ref}})\right| \le L_Q R_t. \tag{10}$$

For fixed $(s, a, \varepsilon^{\mathrm{ref}})$, both bounds decay linearly to zero as $t \to 1$. If the certified radius R_t is smaller than the anchor margin $\rho_m(a) := \mathrm{dist}(a, \mathcal{A} \setminus \mathcal{M}_m(s))$, where $\mathcal{M}_m(s)$ is the mode containing a, then the reference remains in the anchor's mode. Writing $Q_{\bar{\phi}}(s, b) = \mu_m(s) + q_m(s, b)$ within this mode, where $\mu_m(s)$ is a shared offset, we have

$$\Delta(s, a; t, \varepsilon^{\mathrm{ref}}) = q_m(s, a) - q_m\big(s, a'_\theta(s, a, t, \varepsilon^{\mathrm{ref}})\big)\,, \tag{11}$$

so the critic offset shared by the mode cancels.

This offset cancellation is the mechanism illustrated in Figure 2(c): when the reference remains in the anchor mode, weighting is determined by value differences within that mode rather than shared offsets across modes, thereby reducing the tendency to suppress an entire mode. The reference time therefore controls a trade-off between comparison locality and the scale of raw value differences:

- **As $t \to 0$:** the comparison becomes more global and more likely to cross modes, while the certified bounds on the anchor–reference distance and raw advantage become larger, allowing potentially larger raw value differences.
- **As $t \to 1$:** the comparison becomes more local and more likely to remain within the anchor's mode, while the certified bounds on the anchor–reference distance and raw advantage contract toward zero, so the raw value difference vanishes.

This trade-off motivates the proposed reference time distribution introduced below; Appendix C.1 gives the full setup and proof.

Weight normalization. To improve weighting stability, we standardize the minibatch advantages $\{\Delta_i\}_{i=1}^B$ using the median and median absolute deviation (MAD), select the temperature used for exponential weighting to match a target effective sample size (ESS), and normalize the weights to have mean one:

$$z_i = \frac{\Delta_i - \mathrm{median}(\Delta_{1:B})}{\max\{1.4826\,\mathrm{MAD}(\Delta_{1:B}), \varepsilon_{\mathrm{num}}\}}, \quad \hat{w}_i = \mathrm{sg}\left[1 + g\left(\frac{\exp(z_i/\eta^\star)}{B^{-1}\sum_j \exp(z_j/\eta^\star)} - 1\right)\right]. \tag{12}$$

Here $\eta^\star$ is the temperature selected to meet the target ESS, $g \in [0, 1]$ gates reweighting according to critic reliability, and sg denotes stop-gradient. The resulting weights stabilize policy reweighting while preserving the overall scale of the policy loss; complete definitions are provided in Appendix C.2.

Weighted conditional flow matching loss. We then use $\hat{w}_i$ to modulate the corresponding conditional flow matching loss for each sample. Specifically, for each anchor action, we draw $\tau_i \sim U[0, 1]$ and $\varepsilon_i^{\mathrm{fm}} \sim \mathcal{N}(0, I)$ independently of the reference variables $(t_i, \varepsilon_i^{\mathrm{ref}})$ used to compute $\hat{w}_i$. The resulting weighted objective is

$$\mathcal{L}_{\mathrm{FM}}^{\mathrm{SAWR}}(\theta) = \frac{1}{B}\sum_{i=1}^{B} \hat{w}_i \left\| u_\theta\big(s_i, (1-\tau_i)\varepsilon_i^{\mathrm{fm}} + \tau_i a_i, \tau_i\big) - \big(a_i - \varepsilon_i^{\mathrm{fm}}\big)\right\|_2^2. \tag{13}$$

By prioritizing replay actions through sample-anchored advantages, SAWR improves performance while reducing mode suppression. Operationally, it only reweights the conditional flow matching loss for each sample, and the stop-gradient weights prevent backpropagation through the target critic or reference rollout. This lightweight design allows SAWR to serve as a modular weighting method for other algorithms with flow policies.

3.3 Practical implementation

Reference time distribution. To balance local comparisons that favor mode preservation with broader value comparisons, we use the following reference time distribution, whose density is

$$p_t(t) \propto \begin{cases} 0.1 + 1.8t, & 0 \le t \le 0.5 \\ 2(1-t), & 0.5 < t \le 1 \end{cases} \tag{14}$$

The density peaks at intermediate reference times, retaining information from the anchor while allowing flow completion to modify the action. Nonzero density near $t = 0$ also permits occasional broader comparisons to filter infeasible or clearly suboptimal behavior modes. Given the resulting intermediate state, we use three Euler steps by default to integrate the flow to the reference action at $t = 1$.

FQL instantiation and deployment. SAWR can be integrated into other flow policies by replacing only their flow matching objective. We instantiate it with FQL, which combines an expressive multi-step flow policy with efficient one-step action generation at deployment. Specifically, we apply SAWR to the learning of FQL's multi-step flow policy while leaving its distillation, actor, critic, and target network updates unchanged. This integration introduces no additional trainable parameters; Appendix A.1 provides the complete algorithm.

4 EXPERIMENTS

We evaluate SAWR through its FQL instantiation and organize our experiments around three questions. (1) Can SAWR retain the performance benefits of conventional AWR? (2) Does SAWR mitigate mode suppression in multimodal action distributions? (3) Which aspects of reference construction and locality govern the balance between task performance and mode preservation?

4.1 EXPERIMENTAL SETUP

Baselines. We compare SAWR with (i) FQL (Park et al., 2025b), the unweighted baseline, and (ii) FQL-V, the conventional AWR baseline, which replaces the value of the reference action $Q_{\bar{\phi}}(s, a')$ in Equation 8 with a learned state value $V(s)$. The value function is parameterized by a network with the same architecture as the critic and trained to match the Monte Carlo estimate $\widehat{V}(s) = \frac{1}{M}\sum_{m=1}^{M} Q_{\bar{\phi}}(s, a^{(m)})$, where $a^{(m)} \sim \pi(\cdot|s)$. Appendix A.2 gives the complete FQL-V objectives and implementation details. All other components are left unchanged. This controlled comparison isolates the effect of baseline construction: a conventional state value baseline versus the value of a reference action constructed from the same anchor.

Benchmarks and protocol. We evaluate a subset of OGBench navigation and manipulation tasks (Park et al., 2025a), together with D4RL Adroit (Fu et al., 2020). These benchmarks cover long-horizon navigation, compositional manipulation, and dexterous control under heterogeneous offline data. We adopt an offline-to-online protocol consisting of 1M offline updates followed by 1M online updates, and report evaluation success rate at the end of each phase. For a controlled comparison, all methods use the same FQL architecture and task-specific settings. Appendix A.3 provides detailed information about the evaluated tasks, while Appendix A.4 reports the corresponding parameter settings and complete experimental configurations. We report the mean and population standard deviation across five independent seeds.

4.2 EXPERIMENTAL RESULTS

4.2.1 OFFLINE-TO-ONLINE BENCHMARK PERFORMANCE

Table 1 presents the main offline and online results on OGBench and D4RL Adroit. The gains from advantage weighting are modest during offline pretraining and more pronounced after online fine-tuning. On OGBench, adding conventional AWR increases the average online success rate from 46.4 for FQL to 54.4 for FQL-V, while SAWR further raises it to 56.7 and achieves the highest mean online success rate. This trend is even more pronounced on Adroit, where the average online success rate increases from 80.0 for FQL to 85.9 for FQL-V and 91.0 for SAWR. Moreover, SAWR attains or ties for the highest online success rate on 10 of the 12 Adroit tasks.

These results show that SAWR retains the performance benefits of advantage weighting while further improving average online performance on the evaluated OGBench and D4RL Adroit tasks. Several of the largest gains occur in settings with suboptimal offline data, including `antsoccer-arena-stitch` and Adroit Human tasks. Detailed per-task results and learning curves are provided in Appendix B.

Because SAWR only reweights the flow matching loss, it can be readily combined with other flow policy algorithms. Applied to Q-chunking (Li et al., 2025), SAWR increases the average online success rate across four OGBench manipulation families from 67.5% for QC-V to 72.0%. These results support its applicability beyond FQL; Appendix B.5 provides the full comparison.

Table 1: **Main offline-to-online benchmark results on OGBench and D4RL Adroit.** Each entry reports offline → online success rate (%). Bold indicates the best online success rate.

Task	FQL	FQL-V	SAWR (Ours)
antmaze-giant-navigate-singletask (5 tasks)	$17.3_{\pm 9.8} \rightarrow 75.0_{\pm 8.7}$	$8.9_{\pm 6.8} \rightarrow 69.1_{\pm 7.7}$	$\mathbf{7.4}_{\pm 3.7} \rightarrow \mathbf{80.5}_{\pm 2.8}$
antmaze-large-stitch-singletask (5 tasks)	$\mathbf{62.7}_{\pm 10.3} \rightarrow \mathbf{99.8}_{\pm 0.3}$	$55.7_{\pm 5.0} \rightarrow 99.6_{\pm 0.4}$	$\mathbf{63.4}_{\pm 5.3} \rightarrow \mathbf{99.8}_{\pm 0.3}$
humanoidmaze-medium-navigate-singletask (5 tasks)	$52.7_{\pm 6.3} \rightarrow 50.5_{\pm 11.6}$	$48.2_{\pm 8.3} \rightarrow 44.8_{\pm 7.7}$	$\mathbf{57.9}_{\pm 1.9} \rightarrow \mathbf{56.2}_{\pm 7.5}$
humanoidmaze-medium-stitch-singletask (5 tasks)	$\mathbf{50.6}_{\pm 7.8} \rightarrow \mathbf{56.2}_{\pm 7.8}$	$48.2_{\pm 5.7} \rightarrow 53.1_{\pm 15.5}$	$46.9_{\pm 8.3} \rightarrow 56.1_{\pm 8.3}$
humanoidmaze-large-navigate-singletask (5 tasks)	$\mathbf{3.8}_{\pm 2.2} \rightarrow \mathbf{11.0}_{\pm 3.4}$	$5.3_{\pm 3.6} \rightarrow 8.2_{\pm 2.5}$	$2.8_{\pm 0.8} \rightarrow 6.7_{\pm 4.6}$
antsoccer-arena-navigate-singletask (5 tasks)	$57.8_{\pm 4.0} \rightarrow 84.3_{\pm 6.8}$	$\mathbf{58.7}_{\pm 9.2} \rightarrow \mathbf{89.6}_{\pm 1.7}$	$53.4_{\pm 4.1} \rightarrow 88.1_{\pm 5.1}$
antsoccer-arena-stitch-singletask (5 tasks)	$6.0_{\pm 1.0} \rightarrow 21.3_{\pm 2.8}$	$7.8_{\pm 2.3} \rightarrow 38.8_{\pm 4.2}$	$\mathbf{10.6}_{\pm 1.3} \rightarrow \mathbf{55.8}_{\pm 5.7}$
cube-double-noisy-singletask (5 tasks)	$2.5_{\pm 1.3} \rightarrow 19.8_{\pm 0.3}$	$\mathbf{5.8}_{\pm 0.4} \rightarrow \mathbf{80.5}_{\pm 1.2}$	$6.7_{\pm 1.2} \rightarrow 75.0_{\pm 13.4}$
cube-double-play-singletask (5 tasks)	$25.6_{\pm 2.8} \rightarrow 77.1_{\pm 3.9}$	$\mathbf{23.8}_{\pm 4.3} \rightarrow \mathbf{79.7}_{\pm 4.8}$	$23.8_{\pm 7.3} \rightarrow 77.6_{\pm 1.6}$
cube-triple-play-singletask (5 tasks)	$2.4_{\pm 1.1} \rightarrow 19.4_{\pm 1.8}$	$1.4_{\pm 0.4} \rightarrow 13.7_{\pm 4.9}$	$\mathbf{2.9}_{\pm 1.2} \rightarrow \mathbf{19.6}_{\pm 3.5}$
scene-play-singletask (5 tasks)	$55.5_{\pm 2.4} \rightarrow 59.9_{\pm 0.2}$	$\mathbf{58.3}_{\pm 0.9} \rightarrow \mathbf{60.0}_{\pm 0.0}$	$\mathbf{58.8}_{\pm 1.6} \rightarrow \mathbf{60.0}_{\pm 0.0}$
scene-noisy-singletask (5 tasks)	$23.0_{\pm 1.9} \rightarrow 39.8_{\pm 0.2}$	$\mathbf{43.2}_{\pm 3.3} \rightarrow \mathbf{60.0}_{\pm 0.0}$	$\mathbf{39.4}_{\pm 1.6} \rightarrow \mathbf{60.0}_{\pm 0.0}$
puzzle-3x3-play-singletask (5 tasks)	$\mathbf{30.7}_{\pm 2.4} \rightarrow \mathbf{20.0}_{\pm 0.0}$	$\mathbf{24.6}_{\pm 1.5} \rightarrow \mathbf{20.0}_{\pm 0.0}$	$\mathbf{28.2}_{\pm 1.6} \rightarrow \mathbf{20.0}_{\pm 0.0}$
puzzle-4x4-play-singletask (5 tasks)	$15.2_{\pm 3.8} \rightarrow 47.9_{\pm 9.9}$	$\mathbf{25.0}_{\pm 3.3} \rightarrow \mathbf{60.0}_{\pm 0.0}$	$24.4_{\pm 3.5} \rightarrow 56.0_{\pm 8.0}$
puzzle-4x4-noisy-singletask (5 tasks)	$3.3_{\pm 1.1} \rightarrow 13.8_{\pm 3.1}$	$14.1_{\pm 2.8} \rightarrow 38.6_{\pm 4.9}$	$\mathbf{11.2}_{\pm 2.4} \rightarrow \mathbf{39.8}_{\pm 4.3}$
OGBench average	$27.3_{\pm 2.3} \rightarrow 46.4_{\pm 1.6}$	$28.6_{\pm 0.7} \rightarrow 54.4_{\pm 1.6}$	$\mathbf{29.2}_{\pm 1.0} \rightarrow \mathbf{56.7}_{\pm 1.4}$
pen-cloned-v1	$46.8_{\pm 5.2} \rightarrow 90.0_{\pm 6.1}$	$48.8_{\pm 2.7} \rightarrow 94.0_{\pm 1.8}$	$\mathbf{47.6}_{\pm 6.6} \rightarrow \mathbf{96.4}_{\pm 2.3}$
door-cloned-v1	$6.0_{\pm 4.2} \rightarrow 99.2_{\pm 1.0}$	$7.6_{\pm 1.5} \rightarrow 96.8_{\pm 4.5}$	$\mathbf{8.8}_{\pm 4.3} \rightarrow \mathbf{99.6}_{\pm 0.8}$
hammer-cloned-v1	$9.6_{\pm 10.6} \rightarrow 96.0_{\pm 2.8}$	$\mathbf{4.8}_{\pm 3.7} \rightarrow \mathbf{97.2}_{\pm 3.0}$	$6.0_{\pm 4.0} \rightarrow 96.0_{\pm 2.2}$
relocate-cloned-v1	$1.2_{\pm 1.6} \rightarrow 38.4_{\pm 32.4}$	$1.6_{\pm 2.3} \rightarrow 55.2_{\pm 31.9}$	$\mathbf{2.0}_{\pm 1.3} \rightarrow \mathbf{80.0}_{\pm 12.8}$
pen-human-v1	$22.0_{\pm 5.2} \rightarrow 92.0_{\pm 5.5}$	$31.2_{\pm 6.4} \rightarrow 93.6_{\pm 1.5}$	$\mathbf{24.8}_{\pm 6.0} \rightarrow \mathbf{95.6}_{\pm 2.0}$
door-human-v1	$\mathbf{0.4}_{\pm 0.8} \rightarrow \mathbf{96.4}_{\pm 3.2}$	$0.0_{\pm 0.0} \rightarrow 95.2_{\pm 5.9}$	$1.2_{\pm 2.4} \rightarrow 94.8_{\pm 9.4}$
hammer-human-v1	$1.2_{\pm 1.6} \rightarrow 40.0_{\pm 48.2}$	$1.2_{\pm 2.4} \rightarrow 92.0_{\pm 8.3}$	$\mathbf{1.6}_{\pm 0.8} \rightarrow \mathbf{94.4}_{\pm 2.9}$
relocate-human-v1	$0.0_{\pm 0.0} \rightarrow 8.0_{\pm 5.8}$	$0.0_{\pm 0.0} \rightarrow 7.6_{\pm 5.6}$	$\mathbf{0.0}_{\pm 0.0} \rightarrow \mathbf{35.6}_{\pm 19.5}$
pen-expert-v1	$96.0_{\pm 4.6} \rightarrow 99.6_{\pm 0.8}$	$98.4_{\pm 0.8} \rightarrow 99.2_{\pm 1.0}$	$\mathbf{98.0}_{\pm 1.3} \rightarrow \mathbf{99.6}_{\pm 0.8}$
door-expert-v1	$\mathbf{99.6}_{\pm 0.8} \rightarrow \mathbf{100.0}_{\pm 0.0}$	$\mathbf{100.0}_{\pm 0.0} \rightarrow \mathbf{100.0}_{\pm 0.0}$	$\mathbf{100.0}_{\pm 0.0} \rightarrow \mathbf{100.0}_{\pm 0.0}$
hammer-expert-v1	$\mathbf{99.6}_{\pm 0.8} \rightarrow \mathbf{100.0}_{\pm 0.0}$	$99.2_{\pm 1.0} \rightarrow 99.6_{\pm 0.8}$	$\mathbf{98.0}_{\pm 1.3} \rightarrow \mathbf{100.0}_{\pm 0.0}$
relocate-expert-v1	$\mathbf{98.8}_{\pm 1.6} \rightarrow \mathbf{100.0}_{\pm 0.0}$	$\mathbf{100.0}_{\pm 0.0} \rightarrow \mathbf{100.0}_{\pm 0.0}$	$\mathbf{99.6}_{\pm 0.8} \rightarrow \mathbf{100.0}_{\pm 0.0}$
D4RL Adroit average	$40.1_{\pm 1.3} \rightarrow 80.0_{\pm 6.0}$	$41.1_{\pm 0.7} \rightarrow 85.9_{\pm 2.9}$	$\mathbf{40.6}_{\pm 1.6} \rightarrow \mathbf{91.0}_{\pm 1.7}$

4.2.2 Mechanism Diagnostics

We first examine whether reference actions remain in the same behavioral mode as their paired anchor actions. In three OGBench tasks, we measure the fraction of anchor–reference pairs that share the same behavioral mode. Across all three tasks, this same-mode rate generally increases with reference time t and approaches one as $t \rightarrow 1$ (Figure 3), as predicted by Proposition 1. Appendix B.1 details the setting and protocol.

We next test whether AWR with a sample-anchored baseline better preserves behavioral multimodality. Using fixed checkpoints from benchmark experiments, we compare the diversity of successful trajectories. For each fixed policy and initial state, we measure the entropy of successful trajectory modes as $H = -\sum_m p_m \log p_m$, where p_m is the fraction of successful trajectories assigned to mode m. Higher entropy indicates greater diversity among successful behaviors, whereas zero indicates that all successful trajectories belong to a single mode.

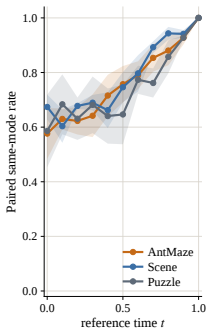


Figure 3: **Reference locality analysis.**

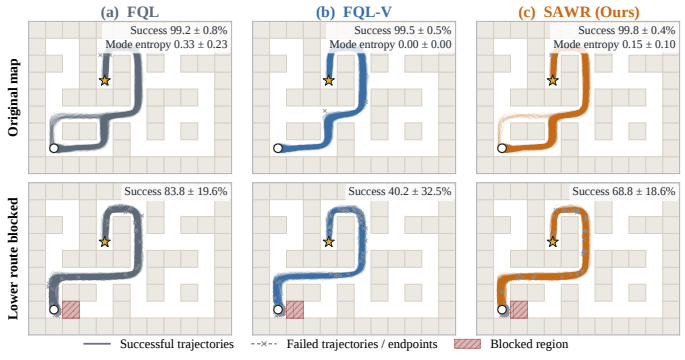


Figure 4: **Mechanism diagnostic on OGBench `antmaze-large-stitch-singletask-task5-v0`.**

In `antmaze-large-stitch-singletask-task5-v0`, all methods achieve near-perfect success on the original map, but FQL-V collapses to one route, whereas SAWR retains both, increasing mode entropy from zero to 0.154 ± 0.099 (Figure 4). To test whether behavioral multimodality affects robustness, we block the lower route and evaluate the policies without further training. Under this intervention, SAWR achieves 68.8% success versus 40.2% for FQL-V, consistent with improved robustness from retaining alternative routes. FQL remains more diverse and robust in this diagnostic.

The manipulation tasks also show the same tendency (Figure 5): SAWR preserves alternative manipulation orders in `scene-noisy-singletask-task1-v0` and a broader set of valid solutions in `puzzle-4x4-noisy-singletask-task1-v0`, while FQL-V concentrates on a single or nearly single mode. Appendix B.2 provides the settings and quantitative results.

4.3 Ablations

We ablate two design axes of the sample-anchored advantage, evaluating performance on five `antsoccer-arena-stitch` tasks and examining the mechanism in `antmaze-large-stitch-singletask-task5-v0`.

- **Reference construction**: *noise-perturbed action* applies a local perturbation to the action without flow completion, whereas *flow-integrated noise* generates the reference by integrating from pure noise at $t = 0$, removing the anchor connection. Our method combines partial noising that retains information from the anchor with flow completion, isolating the contribution of each component.
- **Reference time distribution**: we replace the proposed reference time distribution with uniform *early* ($t \in [0, 0.3]$), *mid* ($t \in [0.4, 0.7]$), *late* ($t \in [0.7, 1]$), or *full* ($t \in [0, 1]$) sampling. This ablation separates the effects of temporal support and the allocation of probability mass on reference locality.

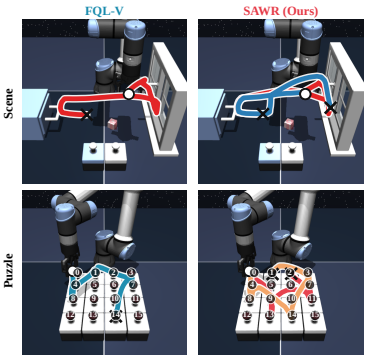


Figure 5: **Manipulation mechanism diagnostics.**

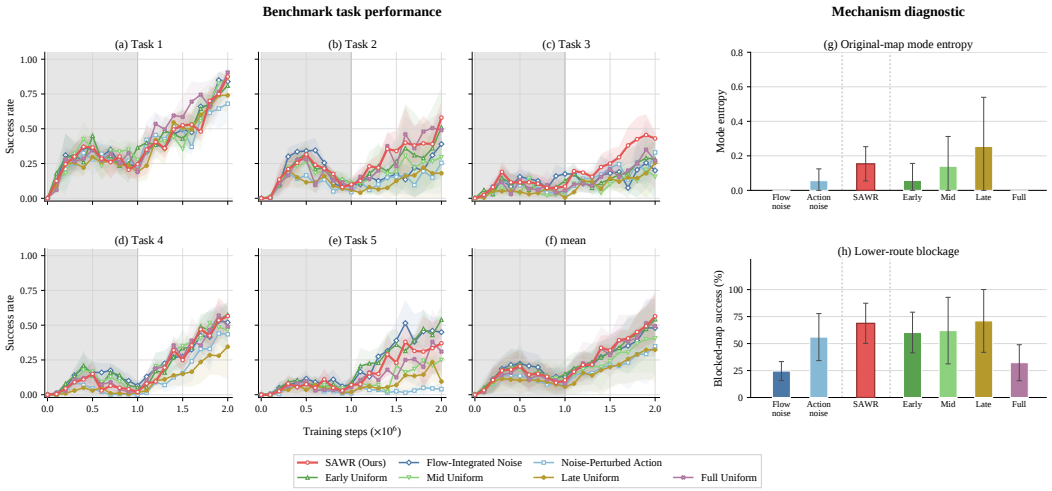


Figure 6: **Reference construction and reference time distribution ablations.** Left (a–f): offline-to-online success rate on five `antsoccer-arena-stitch` tasks. Right: mode entropy of successful routes on the original `antmaze-large-stitch-singletask-task5-v0` map (g) and success after blocking the lower route (h).

Across both the benchmark evaluation and the mechanism diagnostic (Figure 6), the ablations reveal a trade-off controlled by the contribution of the replay action to the reference. As this contribution

increases, references better preserve the behavior mode of the replay action, whereas reducing it allows broader comparisons through flow completion. The reference-time ablations are consistent with the locality analysis in Proposition 1, while their performance differences empirically support balancing local and broader comparisons. Together, SAWR's sample-anchored reference construction and the proposed reference time distribution balance the task performance provided by flow completion with the multimodal behavior preserved by anchoring each reference to a replay action.

5 Related Work

Reinforcement learning for generative policies. Diffusion and flow models provide expressive policies for multimodal, high-dimensional action sequences (Ho et al., 2020; Lipman et al., 2023; Chi et al., 2023), but their iterative sampling procedures and typically intractable action likelihoods complicate reinforcement learning. To enable value-based optimization, existing methods distill a generative policy into a separate actor, incorporate value information into policy regularization, or adapt value optimization to the generative process through stabilized differentiation, distribution constraints, and intermediate-state value guidance (Park et al., 2025b; Lyu et al., 2025; Zhang et al., 2026; Doo et al., 2026). To apply policy-gradient optimization, other methods construct likelihood-ratio surrogates for flow matching, introduce stochastic flow transitions, or approximate policy updates along interpolation paths (McAllister et al., 2026; Zhang et al., 2025; Yang et al., 2026). Additional approaches optimize the initial noise distribution, select critic-ranked action chunks, or reduce sampling cost through one-step generation (Wagenmaker et al., 2025; Li et al., 2025; Geng et al., 2025; Zhan et al., 2026). SAWR follows the AWR framework, using weighted flow matching to improve the policy without evaluating policy likelihoods or differentiating through the sampling process.

Advantage-weighted regression. AWR and AWAC formulate policy improvement as supervised regression on exponentially weighted replay actions (Peng et al., 2019; Nair et al., 2020), while CRR, IQL, and value-aligned behavior cloning develop related policy extraction objectives (Wang et al., 2020; Kostrikov et al., 2022; Jiang et al., 2025). Related methods extend policy extraction and guidance to expressive generative actors (Hansen-Estruch et al., 2023; Tagle et al., 2025; NGUIMATSIA TIOFACK et al., 2026). Within this framework, SAWR focuses on modifying the baseline used to compute advantages, an aspect that remains relatively underexplored in advantage-weighted learning for generative policies.

6 Conclusion

We study how conventional AWR can suppress valid behavior modes when training expressive flow policies. To mitigate this effect, we introduce SAWR, which replaces the shared state value baseline with a local reference action value baseline constructed through partial noising and flow completion. The resulting sample-specific value difference weights the conditional flow matching loss. Integrated into FQL, SAWR improves average online success rates on OGBench and D4RL Adroit. Navigation and manipulation diagnostics show better preservation of diverse successful behaviors, while a route-blockage test shows improved robustness.

Future work should develop more adaptive reference time distributions that adjust locality according to both the stage of training and the policy's observed performance, rather than using a fixed distribution throughout learning. Another promising direction is to extend our method to large generative decision models, including vision language action models (VLAs) and world action models (WAMs), where heterogeneous data and multimodal action distributions make preserving diverse, high-value behaviors especially important.

AI use statement

In this work, we used generative AI tools to aid and polish writing, assist with research ideation and execution, draft portions of the paper, and assist with mathematical claims. For research execution, LLMs were used to assist with code implementation. We did not use generative AI tools for retrieval and discovery or for generating synthetic datasets, and other listed purposes are not applicable to this work. All AI-assisted outputs were reviewed by the authors. We take full responsibility for the final

content of this work, including all text, claims, code, and other artifacts produced with the aid of generative AI.

Reproducibility statement

We include anonymized source code in the supplementary material. Detailed experimental configurations are provided in the appendix. The full codebase will be released after publication.

References

Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin Burchfiel, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. In Kostas E. Bekris, Kris Hauser, Sylvia L. Herbert, and Jingjin Yu (eds.), *Robotics: Science and Systems XIX, Daegu, Republic of Korea, July 10-14, 2023*, 2023. doi: 10.15607/RSS.2023.XIX.026. URL https://doi.org/10.15607/RSS.2023.XIX.026.

JaeHyeok Doo, Byeongguk Jeon, Seonghyeon Ye, Kimin Lee, and Minjoon Seo. Q-flow: Stable and expressive reinforcement learning with flow-based policy. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=oZqOS1N6Ag.

Justin Fu, Aviral Kumar, Ofir Nachum, George Tucker, and Sergey Levine. D4RL: datasets for deep data-driven reinforcement learning. *CoRR*, abs/2004.07219, 2020. URL https://arxiv.org/abs/2004.07219.

Zhengyang Geng, Mingyang Deng, Xingjian Bai, Zico Kolter, and Kaiming He. Mean flows for one-step generative modeling. In Danielle Belgrave, Cheng Zhang, Laura N. Montoya, Hsuan-Tien Lin, Razvan Pascanu, Piotr Koniusz, Marzyeh Ghassemi, Nancy Chen, Iván Vladimir Meza Ruíz, and Arturo Loaiza-Bonilla (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025. URL http://papers.nips.cc/paper_files/paper/2025/hash/6d13e085b79d454da5910e4ca82a3d9d-Abstract-Conference.html.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy P. Lillicrap. Mastering diverse control tasks through world models. *Nat.*, 640(8059):647–653, 2025. doi: 10.1038/S41586-025-08744-2. URL https://doi.org/10.1038/s41586-025-08744-2.

Philippe Hansen-Estruch, Ilya Kostrikov, Michael Janner, Jakub Grudzien Kuba, and Sergey Levine. IDQL: implicit q-learning as an actor-critic method with diffusion policies. *CoRR*, abs/2304.10573, 2023. doi: 10.48550/ARXIV.2304.10573. URL https://doi.org/10.48550/arXiv.2304.10573.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html.

Xingyu Jiang, Ning Gao, Xiuhui Zhang, Hongkun Dou, and Yue Deng. Value-aligned behavior cloning for offline reinforcement learning via bi-level optimization. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=elTJBP7Fbv.

Elia Kaufmann, Leonard Bauersfeld, Antonio Loquercio, Matthias Müller, Vladlen Koltun, and Davide Scaramuzza. Champion-level drone racing using deep reinforcement learning. *Nat.*, 620 (7976):982–987, 2023. doi: 10.1038/S41586-023-06419-4. URL https://doi.org/10.1038/s41586-023-06419-4.

Ilya Kostrikov, Ashvin Nair, and Sergey Levine. Offline reinforcement learning with implicit q-learning. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=68n2s9ZJWF8.

Qiyang Li, Zhiyuan Zhou, and Sergey Levine. Reinforcement learning with action chunking. In Danielle Belgrave, Cheng Zhang, Laura N. Montoya, Hsuan-Tien Lin, Razvan Pascanu, Piotr Koniusz, Marzyeh Ghassemi, Nancy Chen, Iván Vladimir Meza Ruíz, and Arturo Loaiza-Bonilla (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025. URL http://papers.nips.cc/paper_files/paper/2025/hash/50348e8f9aef984abe0ea1ec2a326f78-Abstract-Conference.html.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Lei Lyu, Yunfei Li, Yu Luo, Fuchun Sun, Tao Kong, Jiafeng Xu, and Xiao Ma. Flow-based policy for online reinforcement learning. In Danielle Belgrave, Cheng Zhang, Laura N. Montoya, Hsuan-Tien Lin, Razvan Pascanu, Piotr Koniusz, Marzyeh Ghassemi, Nancy Chen, Iván Vladimir Meza Ruíz, and Arturo Loaiza-Bonilla (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025. URL http://papers.nips.cc/paper_files/paper/2025/hash/871a06b60cf087bbdb854ebfcdf5349a-Abstract-Conference.html.

David McAllister, Songwei Ge, Brent Yi, Chung Min Kim, Ethan Weber, Hongsuk Choi, Haiwen Feng, and Angjoo Kanazawa. Flow matching policy gradients. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 36352–36372, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/3d43cc5692bf68944ee7cd31b97d0c11-Paper-Conference.pdf.

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=aBsCjcPu_tE.

Ashvin Nair, Murtaza Dalal, Abhishek Gupta, and Sergey Levine. Accelerating online reinforcement learning with offline datasets. *CoRR*, abs/2006.09359, 2020. URL https://arxiv.org/abs/2006.09359.

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 C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 30601–30625, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/33f131b806a93d376cf0ce1a456464bb-Paper-Conference.pdf.

Seohong Park, Kevin Frans, Benjamin Eysenbach, and Sergey Levine. Ogbench: Benchmarking offline goal-conditioned rl. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 94937–94982, 2025a. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/ecd92623ac899357312aaa8915853699-Paper-Conference.pdf.

Seohong Park, Qiyang Li, and Sergey Levine. Flow q-learning. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, volume 267 of *Proceedings of Machine Learning Research*. PMLR

/ OpenReview.net, 2025b. URL https://proceedings.mlr.press/v267/park25f.html.

Xue Bin Peng, Aviral Kumar, Grace Zhang, and Sergey Levine. Advantage-weighted regression: Simple and scalable off-policy reinforcement learning. *CoRR*, abs/1910.00177, 2019. URL http://arxiv.org/abs/1910.00177.

Augusto Tagle, Javier Ruiz-del-Solar, and Felipe Tobar. Diffusion self-weighted guidance for offline reinforcement learning. *Trans. Mach. Learn. Res.*, 2025, 2025. URL https://openreview.net/forum?id=jmXBnpmznv.

Andrew Wagenmaker, Yunchu Zhang, Mitsuhiko Nakamoto, Seohong Park, Waleed Yagoub, Anusha Nagabandi, Abhishek Gupta, and Sergey Levine. Steering your diffusion policy with latent space reinforcement learning. In Joseph Lim, Shuran Song, and Hae-Won Park (eds.), *Proceedings of The 9th Conference on Robot Learning*, volume 305 of *Proceedings of Machine Learning Research*, pp. 258–282. PMLR, 27–30 Sep 2025. URL https://proceedings.mlr.press/v305/wagenmaker25a.html.

Zhendong Wang, Jonathan J. Hunt, and Mingyuan Zhou. Diffusion policies as an expressive policy class for offline reinforcement learning. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=AHvFDPi-FA.

Ziyu Wang, Alexander Novikov, Konrad Zolna, Josh Merel, Jost Tobias Springenberg, Scott E. Reed, Bobak Shahriari, Noah Y. Siegel, Çaglar Gülçehre, Nicolas Heess, and Nando de Freitas. Critic regularized regression. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/588cb956d6bbe67078f29f8de420a13d-Abstract.html.

Shunpeng Yang, Ben Liu, and Hua Chen. Policyflow: Policy optimization with continuous normalizing flow in reinforcement learning. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 12553–12573, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/1529a05fdff470f4e7c239c80d85e28e-Paper-Conference.pdf.

Guojian Zhan, Letian Tao, Pengcheng Wang, Yixiao Wang, Yuxin Chen, Yiheng Li, Hongyang Li, Masayoshi Tomizuka, and Shengbo Li. Mean flow policy with instantaneous velocity constraint for one-step action generation. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 157132–157154, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/fe7f375ef01e43f17d2c32b28a176577-Paper-Conference.pdf.

Tonghe Zhang, Chao Yu, Sichang Su, and Yu Wang. Reinflow: Fine-tuning flow matching policy with online reinforcement learning. In Danielle Belgrave, Cheng Zhang, Laura N. Montoya, Hsuan-Tien Lin, Razvan Pascanu, Piotr Koniusz, Marzyeh Ghassemi, Nancy Chen, Iván Vladimir Meza Ruíz, and Arturo Loaiza-Bonilla (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025. URL http://papers.nips.cc/paper_files/paper/2025/hash/98d6f928497d9eaad9f320fb2db3040d-Abstract-Conference.html.

Yixian Zhang, Shu-ang Yu, Tonghe Zhang, Mo Guang, Haojia Hui, Kaiwen Long, Yu Wang, Chao Yu, and Wenbo Ding. Sac flow: Sample-efficient reinforcement learning of flow-based policies via velocity-reparameterized sequential modeling. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 121983–122016, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/c66a3f3d3317da8d4b9c1579e2bc97d5-Paper-Conference.pdf.

A Implementation and Experimental Details

A.1 Complete training algorithm

Algorithm 1 gives the specific FQL instantiation used in our experiments. It leaves the original critic and one-step policy updates unchanged and multiplies only the flow matching loss by the sample-anchored weight $\hat{w}$.

Algorithm 1 FQL with sample-anchored weighting

Input: Replay buffer $\mathcal{D}$, BC flow policy π_θ with vector field u_θ, one-step policy π_ω with generator μ_ω, critic Q_ϕ, target critic $Q_{\bar{\phi}}$, batch size B, discount factor γ, regularization coefficient α, reference time distribution P_t, target update rate ρ
function $\mu_\theta(s, x, t_0)$ ▷ BC flow policy
  $x_1 \leftarrow \mathrm{ODESolve}(\dot{x}_\tau = u_\theta(s, x_\tau, \tau),\ x_{t_0} = x,\ \tau \in [t_0, 1])$
  return $\Pi_{\mathcal{A}}(x_1)$
end function
while not converged **do**
  Sample batch $\{(s, a, r, s')\} \sim \mathcal{D}$
  ▷ Train critic Q_ϕ
  $z' \sim \mathcal{N}(0, I)$; $a' \leftarrow \mu_\omega(s', z')$
  Update ϕ to minimize $\mathbb{E}[(Q_\phi(s, a) - r - \gamma Q_{\bar{\phi}}(s', a'))^2]$
  ▷ Sample-anchored advantage-weighted regression
  $t \sim P_t$; $\varepsilon^{\mathrm{ref}} \sim \mathcal{N}(0, I)$
  $x_t^{\mathrm{ref}} \leftarrow (1 - t)\varepsilon^{\mathrm{ref}} + ta$ ▷ Partial noising
  $a'_\theta \leftarrow \mathrm{sg}[\mu_\theta(s, x_t^{\mathrm{ref}}, t)]$ ▷ Flow completion
  $\Delta \leftarrow \mathrm{sg}[Q_{\bar{\phi}}(s, a) - Q_{\bar{\phi}}(s, a'_\theta)]$ ▷ Sample-anchored advantage
  $\hat{w} \leftarrow \mathrm{sg}[\textsc{Normalize}(\Delta)]$, where $\frac{1}{B}\sum_{i=1}^{B} \hat{w}_i = 1$
  $\varepsilon^{\mathrm{fm}} \sim \mathcal{N}(0, I)$; $\tau \sim \mathrm{Unif}([0, 1])$
  $x_\tau \leftarrow (1 - \tau)\varepsilon^{\mathrm{fm}} + \tau a$
  Update θ to minimize $\mathbb{E}[\hat{w}\|u_\theta(s, x_\tau, \tau) - (a - \varepsilon^{\mathrm{fm}})\|_2^2]$ ▷ Weighted conditional flow matching loss
  ▷ Train one-step policy π_ω
  $z \sim \mathcal{N}(0, I)$; $a^\pi \leftarrow \mu_\omega(s, z)$
  Update ω to minimize $\mathbb{E}[-Q_\phi(s, a^\pi) + \alpha\|a^\pi - \mathrm{sg}[\mu_\theta(s, z, 0)]\|_2^2]$
  Update target critic: $\bar{\phi} \leftarrow \rho\phi + (1 - \rho)\bar{\phi}$
end while
return One-step policy π_ω

A.2 FQL-V baseline details

State-value target. The FQL-V baseline is an implementation of conventional AWR using a state value baseline with the same FQL backbone. In addition to the ensemble of two critics used by FQL, it introduces a value network $V_\psi(s)$ with a scalar output. The value network uses the same architecture as each critic: four hidden layers of width 512 and layer normalization. For every replay state s_i in a minibatch, we draw M independent Gaussian inputs $z_i^{(m)} \sim \mathcal{N}(0, I)$ and generate actions with the distilled one-step policy,

$$a_i^{(m)} = \Pi_{\mathcal{A}}\left(\mu_\omega(s_i, z_i^{(m)})\right), \qquad \widehat{V}_i = \mathrm{sg}\left[\frac{1}{M}\sum_{m=1}^{M} Q_{\bar{\phi}}\left(s_i, a_i^{(m)}\right)\right]. \tag{15}$$

The sampled actions and the critic evaluations in the target are stop-gradient. We train the value network by minimizing the mean squared error,

$$\mathcal{L}_V(\psi) = \frac{1}{B}\sum_{i=1}^{B}\left(V_\psi(s_i) - \widehat{V}_i\right)^2. \tag{16}$$

Thus V_ψ estimates the expected value from the target critic under the deployed one-step policy rather than fitting returns from the replay action. We use $M = 8$ policy samples per state and give $\mathcal{L}_V$ a unit coefficient in the joint training loss.

Conventional advantage and flow policy update. For each replay pair (s_i, a_i), FQL-V forms the conventional advantage using a state value baseline,

$$A_i^V = \mathrm{sg}\big[Q_{\bar{\phi}}(s_i, a_i) - V_\psi(s_i)\big]\,. \tag{17}$$

We apply exactly the same robust minibatch standardization, temperature selection for the target ESS, normalization to mean one, and critic reliability gate as in Appendix C.2, replacing Δ_i there by A_i^V. We denote the resulting stop-gradient weight by $\hat{w}_i^V$, and optimize the multi-step flow policy with

$$\mathcal{L}_{\mathrm{FM}}^{\text{FQL-V}}(\theta) = \frac{1}{B}\sum_{i=1}^{B} \hat{w}_i^V \, \|u_\theta(s_i, (1-\tau_i)\varepsilon_i + \tau_i a_i, \tau_i) - (a_i - \varepsilon_i)\|_2^2\,, \tag{18}$$

with independently sampled $\tau_i \sim U[0,1]$ and $\varepsilon_i \sim \mathcal{N}(0, I)$. Weighting is active in both offline and online updates.

Updates shared with FQL. The critic Bellman loss, one-step policy distillation and actor losses based on critic values, target critic exponential moving average, optimizer, and deployed policy are otherwise unchanged from FQL and Algorithm 1. Concretely, each minibatch update minimizes the sum of the critic loss, the FQL actor loss with Equation 18 in place of unweighted flow matching, and Equation 16. The value network is used only to construct training weights and is not required for action generation at evaluation time. This design makes the comparison with SAWR differ primarily in the baseline: a Monte Carlo state value shared by all actions for FQL-V versus the value of a reference action constructed from the same anchor for SAWR.

A.3 Environments and datasets

We evaluate on OGBench (Park et al., 2025a) and D4RL (Fu et al., 2020), which cover long-horizon navigation, whole-body control, and dexterous manipulation. We describe the two benchmarks separately because similarly named dataset suffixes need not share the same collection procedure. In particular, `play` denotes natural expert interaction data in OGBench manipulation but a hand-specified set of navigation queries in D4RL AntMaze.

A.3.1 OGBench

OGBench pairs each physical environment with several offline behavior distributions. Each task family comprises five fixed evaluation subtasks, corresponding to five predefined goals. Each dataset identifier specifies the environment (e.g., `antsoccer-arena`), behavior distribution (`stitch`), evaluation protocol (`singletask`), subtask index (e.g., `task4`), and release version (`v0`). We introduce the six task families used in our evaluation and highlight how their dataset variants differ in behavior coverage, trajectory coherence, and compositional difficulty.

AntMaze. AntMaze controls an 8-DoF quadruped and couples low-level locomotion with long-horizon navigation to a specified maze location. The `navigate` dataset is collected by a noisy expert that repeatedly travels to randomly selected locations, forming the standard, relatively coherent data regime. The `stitch` dataset instead contains short and disconnected trajectory segments; solving an evaluation task generally requires composing transitions that do not occur in one complete demonstration. OGBench also provides an `explore` dataset of random, low-quality trajectories with high state coverage, although it is not used in Table 1.

HumanoidMaze. HumanoidMaze replaces the Ant with a 21-DoF humanoid, substantially increasing the difficulty of low-level control while retaining the long-horizon maze objective. Its `navigate` dataset is collected by a noisy expert moving between randomly selected locations. The `stitch` variant contains short, disconnected motion segments and therefore tests whether the learner can compose locally valid humanoid behaviors into a complete route. As in AntMaze, the optional `explore` data provide broad state coverage through low-quality random exploration but are not used in our benchmark table.

AntSoccer. AntSoccer augments Ant locomotion with ball interaction: the agent must first approach the ball and then dribble it to a target location. The `navigate` dataset contains trajectories that

repeatedly approach the ball and move it to random locations, and therefore includes both locomotion and ball-control behavior within individual rollouts. The `stitch` dataset deliberately separates these skills, combining maze-navigation trajectories without the ball with local dribbling trajectories in which the ball is already near the agent. Completing the full task thus requires stitching across the two data sources.

Cube. Cube evaluates robotic pick-and-place, with `single`, `double`, and `triple` indicating the number of objects. The five goals include moving, stacking, swapping, and permuting cubes. The `play` dataset is generated by a non-Markovian scripted expert that repeatedly selects a cube and places it at a random location or on another cube, with temporally correlated action noise producing natural interaction trajectories. The `noisy` dataset uses a Markovian expert with larger, temporally uncorrelated Gaussian action noise. It provides broader state coverage but contains more suboptimal transitions.

Scene. Scene combines a cube, drawer, window, and buttons that control object locks. Its evaluation goals require multi-stage behaviors such as unlocking an object, manipulating it, and restoring the lock state. The `play` dataset records a non-Markovian scripted expert randomly interacting with the objects under temporally correlated noise, yielding coherent but unstructured behavior sequences. The `noisy` variant is collected with a Markovian expert and larger independent Gaussian perturbations, increasing coverage and behavioral heterogeneity at the cost of lower data quality.

Puzzle. Puzzle is a continuous-control realization of Lights Out: pressing a button toggles that button and its neighbors, and the robot must reach a target board configuration. Grid sizes such as `3x3` and `4x4` determine the combinatorial state space. The `play` dataset consists of natural button-press sequences produced by a non-Markovian scripted expert with temporally correlated noise. The `noisy` dataset uses the corresponding Markovian expert with larger independent Gaussian perturbations, providing greater coverage of both continuous robot states and discrete board configurations.

A.3.2 D4RL

D4RL uses a different naming convention: the suffix directly identifies the offline dataset associated with an environment, and `v1` or `v2` indicates a dataset release rather than a difficulty level.

AntMaze. D4RL AntMaze requires a quadruped to reach a fixed target through one of three maze layouts of increasing size. The `play` datasets are collected from a fixed, hand-selected set of initial and goal locations and therefore contain structured but incomplete navigation experience. The `diverse` datasets command the Ant toward randomly sampled goal locations, producing broader maze coverage. Neither dataset is tailored to the evaluation start–goal pair, so successful offline learning can require stitching subtrajectories across demonstrations. This use of `play` is distinct from the OGBench manipulation datasets above.

Adroit. Adroit controls a 24-DoF anthropomorphic hand in four tasks: orienting a pen, opening a door, hammering a nail, and relocating an object. The `human` datasets contain the 25 motion-captured demonstrations released with DAPG. The `cloned` datasets form an equal mixture of human demonstrations and trajectories from a behavior-cloned policy trained on those demonstrations; they consequently combine expert-like actions with systematic imitation errors. The `expert` datasets contain trajectories sampled from a task-solving expert policy and represent the highest-quality collection regime.

A.4 Benchmark configurations

We follow the experimental protocol and hyperparameter settings of FQL. Each experiment consists of 1M offline updates followed by 1M online updates. The flow policy uses 10 Euler steps to generate actions, while SAWR uses 3 Euler steps to integrate a partially noised action from the sampled reference time to $t = 1$, thereby producing the reference action used to compute the sample-anchored advantage. For all tasks and methods, both the actor and critic use 4 hidden layers of width 512, with a batch size of 256, a learning rate of 3×10^{-4}, and a target update rate of 0.005. Layer normalization is enabled for the critic but disabled for the actor. Both weighted methods use an ESS target of 0.7,

the same critic reliability gate, and weighting during both offline and online updates. Task-specific hyperparameters, including α, the discount factor, and the critic aggregation method, are reported in Table 2.

Table 2: Task-specific hyperparameters.

Task	α	γ	Q agg.
OGBench			
antmaze-giant-navigate-singletask-v0	10	0.995	min
antmaze-large-stitch-singletask-v0	10	0.995	min
humanoidmaze-medium-navigate-singletask-v0	30	0.995	mean
humanoidmaze-medium-stitch-singletask-v0	30	0.995	mean
humanoidmaze-large-navigate-singletask-v0	30	0.995	mean
antsoccer-arena-navigate-singletask-v0	10	0.995	mean
antsoccer-arena-stitch-singletask-v0	30	0.995	mean
cube-double-noisy-singletask-v0	300	0.990	mean
cube-double-play-singletask-v0	300	0.990	mean
cube-triple-play-singletask-v0	300	0.990	mean
scene-play-singletask-v0	300	0.990	mean
scene-noisy-singletask-v0	300	0.990	mean
puzzle-3x3-play-singletask-v0	1000	0.990	mean
puzzle-4x4-play-singletask-v0	1000	0.990	mean
puzzle-4x4-noisy-singletask-v0	1000	0.990	mean
D4RL AntMaze			
antmaze-umaze-v2	10	0.990	mean
antmaze-umaze-diverse-v2	10	0.990	mean
antmaze-medium-play-v2	10	0.990	mean
antmaze-medium-diverse-v2	10	0.990	mean
antmaze-large-play-v2	3	0.990	mean
antmaze-large-diverse-v2	3	0.990	mean
D4RL Adroit			
pen-cloned-v1	10000	0.990	min
door-cloned-v1	30000	0.990	min
hammer-cloned-v1	10000	0.990	min
relocate-cloned-v1	30000	0.990	min
pen-human-v1	10000	0.990	min
door-human-v1	30000	0.990	min
hammer-human-v1	30000	0.990	min
relocate-human-v1	10000	0.990	min
pen-expert-v1	3000	0.990	min
door-expert-v1	30000	0.990	min
hammer-expert-v1	30000	0.990	min
relocate-expert-v1	30000	0.990	min

B Detailed Benchmark Results

B.1 Reference Locality

Reference construction and paired evaluation. Figure 3 evaluates SAWR at the final checkpoints on `antmaze-large-stitch-singletask-task5-v0`, `scene-noisy-singletask-task1-v0`, and `puzzle-4x4-noisy-singletask-task1-v0`. We select initial states at which policy rollouts can exhibit two alternative modes and use the first action from an alternative mode as the anchor a. For each anchor, we construct reference actions $a'_r(t)$ at $t \in \{0, 0.1, \ldots, 1\}$ by partial noising followed by three Euler flow completion steps and action clipping. We execute either the anchor or a reference action from the same initial state, then continue with the frozen policy using shared policy and environment randomness. This diagnostic measures locality on selected states with alternative behaviors.

Mode labels and agreement. AntMaze modes correspond to the upper and lower routes, identified by passage through the route-discriminating corridor. Scene modes indicate whether the drawer or window target is reached first; trajectories that do not reach both targets or reach them simultaneously are invalid. Puzzle modes are the sets of buttons pressed an odd number of times, validated against task success and button parity.

Let $M(a, \xi)$ denote the mode obtained by executing action a and continuing with random inputs ξ, and let $\mathcal{V}$ denote the valid labels. We estimate the same-mode rate as

$$\widehat{P}_t = \frac{\sum_{(a,\xi,r)} \mathbf{1}\{M(a,\xi) \in \mathcal{V},\ M(a'_r(t),\xi) = M(a,\xi)\}}{\sum_{(a,\xi,r)} \mathbf{1}\{M(a,\xi) \in \mathcal{V}\}}. \tag{19}$$

The valid-anchor panel is fixed across reference times. Invalid anchor continuations are excluded, while invalid reference continuations count as mismatches. We compute the rate separately for each training run and report the mean and standard deviation across runs. At $t = 0$, reference completion starts from pure Gaussian noise; at $t = 1$, the reference equals the anchor, so agreement is one by construction. Agreement away from this endpoint serves as an empirical measure of behavioral locality.

B.2 Mechanistic Diagnostic on Manipulation Tasks

Experimental objective. We conduct a manipulation diagnostic to test whether SAWR preserves successful high-level manipulation orders while retaining the success rate improvement of conventional advantage weighting.

Environment and protocol. We use `scene-noisy-singletask-task1-v0` and `puzzle-4x4-noisy-singletask-task1-v0` from OGBench for complementary mechanism diagnostics. In Scene, the robot must open an initially unlocked drawer and window while preserving the target cube position and button states. In Puzzle, the robot must transform a 4×4 button grid from all zeros to all ones, where each press toggles the selected button and its orthogonal neighbors. These tasks admit interpretable behavioral alternatives—different manipulation orders in Scene and different button-press solutions in Puzzle—making them suitable for examining whether a policy preserves multiple successful behaviors when learning from noisy offline data. For each method, we reuse the five final checkpoints from the benchmark experiments. Given a fixed initial state, we sample 100 stochastic rollouts from each checkpoint, using identical initial states across methods and training seeds. We measure multimodality through the diversity of successful behavioral modes under a fixed policy and initial state. In Scene, modes are defined by the order in which the drawer and window first reach their target configurations: drawer-first or window-first. In Puzzle, modes are defined by the set of buttons pressed an odd number of times.

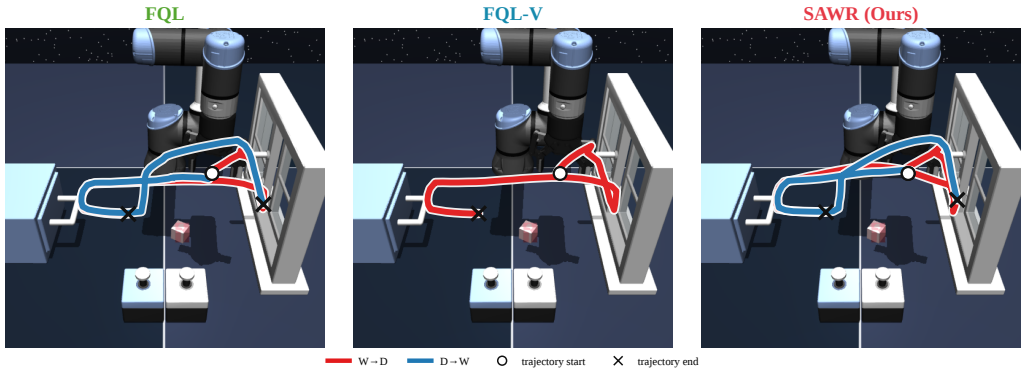


Figure 7: **Manipulation mechanism diagnostic on OGBench Scene.**

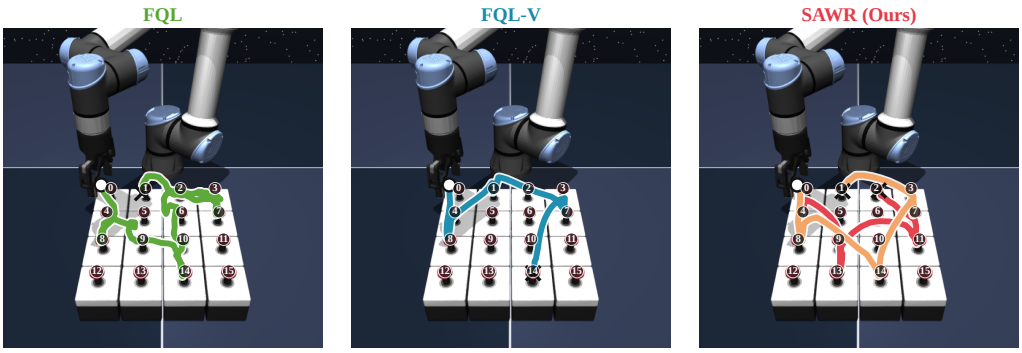


Figure 8: **Manipulation mechanism diagnostic on OGBench Puzzle.**

Table 3: Success rates and mode entropy from a fixed initial state in each task.

Task	Method	Success (%)	Mode entropy
Scene	FQL	100.0 ± 0.0	0.609 ± 0.093
	FQL-V	100.0 ± 0.0	0.000 ± 0.000
	SAWR	100.0 ± 0.0	0.213 ± 0.140
Puzzle	FQL	76.8 ± 14.4	0.625 ± 0.174
	FQL-V	100.0 ± 0.0	0.027 ± 0.060
	SAWR	100.0 ± 0.0	0.282 ± 0.332

Results. Table 3 shows that SAWR retains more diverse successful behaviors than FQL-V while matching its success rate on both tasks. In Scene, all methods achieve 100% success, but FQL-V consistently selects a single manipulation order, yielding zero mode entropy, whereas SAWR preserves both drawer-first and window-first behaviors (mode entropy 0.213 ± 0.140). In Puzzle, both FQL-V and SAWR achieve 100% success, exceeding the $76.8\% \pm 14.4\%$ of FQL, while SAWR increases mode entropy from 0.027 ± 0.060 to 0.282 ± 0.332. Although FQL exhibits greater mode diversity in both tasks, SAWR retains more of this diversity than FQL-V without sacrificing success. Overall, these diagnostics suggest that SAWR mitigates the concentration of successful behavior associated with conventional AWR.

B.3 D4RL AntMaze performance

Table 4 reports the complete D4RL AntMaze results. SAWR achieves comparable final online performance on D4RL AntMaze, where all methods approach ceiling success rates.

Table 4: D4RL AntMaze offline-to-online success rate (%).

Task	FQL	FQL-V	SAWR
antmaze-umaze-v2	$97.3_{\pm 2.5} \rightarrow 99.3_{\pm 0.9}$	$\mathbf{96.7}_{\pm 2.5} \rightarrow \mathbf{100.0}_{\pm 0.0}$	$99.3_{\pm 0.9} \rightarrow 99.3_{\pm 0.9}$
antmaze-umaze-diverse-v2	$\mathbf{92.7}_{\pm 1.9} \rightarrow \mathbf{100.0}_{\pm 0.0}$	$84.0_{\pm 4.3} \rightarrow 98.7_{\pm 0.9}$	$83.3_{\pm 3.4} \rightarrow 99.3_{\pm 0.9}$
antmaze-medium-play-v2	$80.0_{\pm 5.9} \rightarrow 96.0_{\pm 0.0}$	$\mathbf{77.3}_{\pm 6.8} \rightarrow \mathbf{97.3}_{\pm 0.9}$	$76.7_{\pm 5.0} \rightarrow 94.7_{\pm 2.5}$
antmaze-medium-diverse-v2	$68.0_{\pm 2.8} \rightarrow 98.0_{\pm 1.6}$	$70.0_{\pm 7.1} \rightarrow 97.3_{\pm 1.9}$	$\mathbf{73.3}_{\pm 2.5} \rightarrow \mathbf{99.3}_{\pm 0.9}$
antmaze-large-play-v2	$86.7_{\pm 3.4} \rightarrow 94.0_{\pm 1.6}$	$87.3_{\pm 0.9} \rightarrow 94.7_{\pm 4.1}$	$\mathbf{73.3}_{\pm 20.8} \rightarrow \mathbf{95.3}_{\pm 0.9}$
antmaze-large-diverse-v2	$\mathbf{90.0}_{\pm 4.3} \rightarrow \mathbf{94.7}_{\pm 4.7}$	$84.7_{\pm 5.0} \rightarrow 94.0_{\pm 1.6}$	$85.3_{\pm 6.8} \rightarrow 92.0_{\pm 1.6}$
D4RL AntMaze average	$\mathbf{85.8}_{\pm 1.1} \rightarrow \mathbf{97.0}_{\pm 0.9}$	$\mathbf{83.3}_{\pm 2.9} \rightarrow \mathbf{97.0}_{\pm 0.5}$	$81.9_{\pm 3.7} \rightarrow 96.7_{\pm 0.3}$

B.4 OGBench per-subtask endpoints

Table 5 presents detailed results for each of the five subtasks across all 15 OGBench benchmark settings, using the same selected runs as Table 1.

Table 5: **OGBench per-subtask results.**

Task	FQL	FQL-V	SAWR
antmaze-giant-navigate-singletask-task1-v0	$7.2_{\pm 6.8} \rightarrow 70.0_{\pm 37.1}$	$11.2_{\pm 12.2} \rightarrow 77.2_{\pm 29.5}$	$5.2_{\pm 6.1} \rightarrow 96.4_{\pm 2.0}$
antmaze-giant-navigate-singletask-task2-v0	$11.2_{\pm 12.9} \rightarrow 98.8_{\pm 1.6}$	$1.2_{\pm 2.4} \rightarrow 99.2_{\pm 1.6}$	$9.6_{\pm 12.8} \rightarrow 99.6_{\pm 0.8}$
antmaze-giant-navigate-singletask-task3-v0	$0.8_{\pm 1.6} \rightarrow 7.6_{\pm 15.2}$	$0.0_{\pm 0.0} \rightarrow 8.8_{\pm 16.6}$	$0.0_{\pm 0.0} \rightarrow 7.2_{\pm 12.5}$
antmaze-giant-navigate-singletask-task4-v0	$13.2_{\pm 22.5} \rightarrow 98.8_{\pm 1.6}$	$1.2_{\pm 2.4} \rightarrow 60.4_{\pm 48.5}$	$0.0_{\pm 0.0} \rightarrow 99.2_{\pm 1.0}$
antmaze-giant-navigate-singletask-task5-v0	$54.0_{\pm 27.6} \rightarrow 100.0_{\pm 0.0}$	$30.8_{\pm 32.3} \rightarrow 100.0_{\pm 0.0}$	$22.4_{\pm 27.6} \rightarrow 100.0_{\pm 0.0}$
antmaze-large-stitch-singletask-task1-v0	$81.6_{\pm 17.5} \rightarrow 100.0_{\pm 0.0}$	$78.0_{\pm 29.2} \rightarrow 100.0_{\pm 0.0}$	$93.2_{\pm 4.1} \rightarrow 100.0_{\pm 0.0}$
antmaze-large-stitch-singletask-task2-v0	$55.6_{\pm 28.4} \rightarrow 100.0_{\pm 0.0}$	$0.0_{\pm 0.0} \rightarrow 98.8_{\pm 1.6}$	$43.2_{\pm 32.7} \rightarrow 99.2_{\pm 1.6}$
antmaze-large-stitch-singletask-task3-v0	$88.0_{\pm 10.8} \rightarrow 100.0_{\pm 0.0}$	$88.4_{\pm 3.4} \rightarrow 99.6_{\pm 0.8}$	$88.8_{\pm 6.3} \rightarrow 100.0_{\pm 0.0}$
antmaze-large-stitch-singletask-task4-v0	$0.0_{\pm 0.0} \rightarrow 100.0_{\pm 0.0}$	$22.4_{\pm 27.2} \rightarrow 99.6_{\pm 0.8}$	$3.2_{\pm 6.4} \rightarrow 100.0_{\pm 0.0}$
antmaze-large-stitch-singletask-task5-v0	$88.4_{\pm 7.1} \rightarrow 99.2_{\pm 1.6}$	$89.6_{\pm 5.6} \rightarrow 100.0_{\pm 0.0}$	$88.4_{\pm 4.6} \rightarrow 100.0_{\pm 0.0}$
humanoidmaze-medium-navigate-singletask-task1-v0	$12.8_{\pm 7.8} \rightarrow 8.8_{\pm 9.5}$	$13.6_{\pm 15.7} \rightarrow 4.8_{\pm 6.4}$	$20.4_{\pm 22.9} \rightarrow 2.8_{\pm 3.7}$
humanoidmaze-medium-navigate-singletask-task2-v0	$92.0_{\pm 10.2} \rightarrow 92.0_{\pm 14.0}$	$67.2_{\pm 35.9} \rightarrow 98.4_{\pm 2.3}$	$96.8_{\pm 3.2} \rightarrow 96.0_{\pm 8.0}$
humanoidmaze-medium-navigate-singletask-task3-v0	$59.2_{\pm 39.3} \rightarrow 51.2_{\pm 44.8}$	$80.0_{\pm 10.0} \rightarrow 20.0_{\pm 40.0}$	$73.6_{\pm 15.4} \rightarrow 62.0_{\pm 45.1}$
humanoidmaze-medium-navigate-singletask-task4-v0	$0.4_{\pm 0.8} \rightarrow 0.4_{\pm 0.8}$	$0.8_{\pm 1.0} \rightarrow 0.8_{\pm 1.0}$	$0.4_{\pm 0.8} \rightarrow 20.4_{\pm 38.8}$
humanoidmaze-medium-navigate-singletask-task5-v0	$99.2_{\pm 1.0} \rightarrow 100.0_{\pm 0.0}$	$79.2_{\pm 39.6} \rightarrow 100.0_{\pm 0.0}$	$98.4_{\pm 3.2} \rightarrow 100.0_{\pm 0.0}$
humanoidmaze-medium-stitch-singletask-task1-v0	$12.4_{\pm 5.9} \rightarrow 6.0_{\pm 9.3}$	$17.6_{\pm 12.5} \rightarrow 25.6_{\pm 29.3}$	$23.6_{\pm 19.2} \rightarrow 3.2_{\pm 4.7}$
humanoidmaze-medium-stitch-singletask-task2-v0	$91.2_{\pm 1.0} \rightarrow 99.6_{\pm 0.8}$	$92.8_{\pm 4.3} \rightarrow 99.6_{\pm 0.8}$	$95.2_{\pm 2.0} \rightarrow 100.0_{\pm 0.0}$
humanoidmaze-medium-stitch-singletask-task3-v0	$51.2_{\pm 34.7} \rightarrow 0.0_{\pm 0.0}$	$32.8_{\pm 22.2} \rightarrow 0.8_{\pm 1.6}$	$17.2_{\pm 25.9} \rightarrow 0.4_{\pm 0.8}$
humanoidmaze-medium-stitch-singletask-task4-v0	$0.0_{\pm 0.0} \rightarrow 76.0_{\pm 38.1}$	$0.4_{\pm 0.8} \rightarrow 59.6_{\pm 48.7}$	$0.4_{\pm 0.8} \rightarrow 77.6_{\pm 38.8}$
humanoidmaze-medium-stitch-singletask-task5-v0	$98.4_{\pm 1.5} \rightarrow 99.6_{\pm 0.8}$	$97.2_{\pm 3.7} \rightarrow 80.0_{\pm 39.0}$	$98.0_{\pm 3.1} \rightarrow 99.2_{\pm 1.0}$
humanoidmaze-large-navigate-singletask-task1-v0	$8.4_{\pm 7.8} \rightarrow 0.4_{\pm 0.8}$	$10.0_{\pm 12.4} \rightarrow 0.0_{\pm 0.0}$	$1.2_{\pm 1.6} \rightarrow 0.0_{\pm 0.0}$
humanoidmaze-large-navigate-singletask-task2-v0	$0.0_{\pm 0.0} \rightarrow 0.0_{\pm 0.0}$	$0.0_{\pm 0.0} \rightarrow 0.0_{\pm 0.0}$	$0.0_{\pm 0.0} \rightarrow 0.0_{\pm 0.0}$
humanoidmaze-large-navigate-singletask-task3-v0	$8.8_{\pm 6.4} \rightarrow 41.2_{\pm 20.7}$	$16.0_{\pm 7.6} \rightarrow 31.6_{\pm 16.8}$	$9.6_{\pm 5.3} \rightarrow 33.6_{\pm 22.9}$
humanoidmaze-large-navigate-singletask-task4-v0	$0.0_{\pm 0.0} \rightarrow 13.2_{\pm 16.8}$	$0.4_{\pm 0.8} \rightarrow 1.2_{\pm 1.6}$	$0.0_{\pm 0.0} \rightarrow 0.0_{\pm 0.0}$
humanoidmaze-large-navigate-singletask-task5-v0	$2.0_{\pm 1.8} \rightarrow 0.0_{\pm 0.0}$	$0.0_{\pm 0.0} \rightarrow 8.0_{\pm 16.0}$	$3.2_{\pm 6.4} \rightarrow 0.0_{\pm 0.0}$
antsoccer-arena-navigate-singletask-task1-v0	$75.6_{\pm 7.0} \rightarrow 97.6_{\pm 1.5}$	$77.2_{\pm 10.7} \rightarrow 98.8_{\pm 2.4}$	$81.2_{\pm 5.0} \rightarrow 97.2_{\pm 2.0}$
antsoccer-arena-navigate-singletask-task2-v0	$91.6_{\pm 3.2} \rightarrow 98.8_{\pm 1.0}$	$90.8_{\pm 7.9} \rightarrow 97.2_{\pm 1.0}$	$94.4_{\pm 2.9} \rightarrow 98.4_{\pm 1.5}$
antsoccer-arena-navigate-singletask-task3-v0	$50.0_{\pm 2.2} \rightarrow 94.0_{\pm 3.3}$	$54.8_{\pm 13.6} \rightarrow 94.0_{\pm 2.2}$	$42.8_{\pm 6.5} \rightarrow 84.4_{\pm 9.2}$
antsoccer-arena-navigate-singletask-task4-v0	$35.6_{\pm 7.1} \rightarrow 50.4_{\pm 33.2}$	$45.2_{\pm 3.7} \rightarrow 71.6_{\pm 7.4}$	$30.0_{\pm 9.8} \rightarrow 70.4_{\pm 23.4}$
antsoccer-arena-navigate-singletask-task5-v0	$36.4_{\pm 12.0} \rightarrow 80.8_{\pm 5.7}$	$25.6_{\pm 17.7} \rightarrow 86.4_{\pm 8.7}$	$18.8_{\pm 12.8} \rightarrow 90.0_{\pm 7.5}$
antsoccer-arena-stitch-singletask-task1-v0	$20.0_{\pm 4.9} \rightarrow 69.2_{\pm 6.6}$	$22.8_{\pm 6.0} \rightarrow 76.8_{\pm 8.8}$	$25.6_{\pm 9.0} \rightarrow 87.2_{\pm 6.0}$
antsoccer-arena-stitch-singletask-task2-v0	$6.6_{\pm 2.2} \rightarrow 1.2_{\pm 2.4}$	$7.2_{\pm 3.0} \rightarrow 37.8_{\pm 20.4}$	$11.2_{\pm 5.2} \rightarrow 54.4_{\pm 12.6}$
antsoccer-arena-stitch-singletask-task3-v0	$1.6_{\pm 1.5} \rightarrow 21.2_{\pm 13.4}$	$2.4_{\pm 2.3} \rightarrow 24.8_{\pm 9.8}$	$9.2_{\pm 3.5} \rightarrow 43.2_{\pm 4.1}$
antsoccer-arena-stitch-singletask-task4-v0	$0.0_{\pm 0.0} \rightarrow 12.0_{\pm 13.9}$	$2.8_{\pm 3.0} \rightarrow 43.6_{\pm 6.2}$	$1.6_{\pm 1.5} \rightarrow 56.4_{\pm 6.2}$
antsoccer-arena-stitch-singletask-task5-v0	$2.0_{\pm 1.8} \rightarrow 2.8_{\pm 2.4}$	$4.0_{\pm 3.1} \rightarrow 10.8_{\pm 7.5}$	$5.6_{\pm 6.6} \rightarrow 38.0_{\pm 17.1}$
cube-double-noisy-singletask-task1-v0	$10.4_{\pm 5.0} \rightarrow 98.4_{\pm 1.5}$	$25.6_{\pm 3.4} \rightarrow 99.2_{\pm 1.0}$	$23.2_{\pm 1.6} \rightarrow 99.6_{\pm 0.8}$
cube-double-noisy-singletask-task2-v0	$0.0_{\pm 0.0} \rightarrow 0.0_{\pm 0.0}$	$0.4_{\pm 0.8} \rightarrow 98.8_{\pm 1.6}$	$4.0_{\pm 3.3} \rightarrow 99.6_{\pm 0.8}$
cube-double-noisy-singletask-task3-v0	$1.6_{\pm 2.0} \rightarrow 0.4_{\pm 0.8}$	$0.8_{\pm 1.0} \rightarrow 98.4_{\pm 2.0}$	$0.8_{\pm 1.0} \rightarrow 99.2_{\pm 1.0}$
cube-double-noisy-singletask-task4-v0	$0.0_{\pm 0.0} \rightarrow 0.4_{\pm 0.8}$	$0.8_{\pm 1.0} \rightarrow 6.0_{\pm 4.6}$	$2.4_{\pm 0.8} \rightarrow 16.4_{\pm 30.8}$
cube-double-noisy-singletask-task5-v0	$0.4_{\pm 0.8} \rightarrow 0.0_{\pm 0.0}$	$1.6_{\pm 1.5} \rightarrow 100.0_{\pm 0.0}$	$3.2_{\pm 1.6} \rightarrow 60.0_{\pm 49.0}$
cube-double-play-singletask-task1-v0	$60.0_{\pm 8.1} \rightarrow 95.6_{\pm 4.5}$	$63.2_{\pm 12.1} \rightarrow 99.2_{\pm 1.0}$	$56.8_{\pm 17.3} \rightarrow 98.4_{\pm 3.2}$
cube-double-play-singletask-task2-v0	$28.8_{\pm 5.0} \rightarrow 94.0_{\pm 4.2}$	$29.2_{\pm 13.5} \rightarrow 93.6_{\pm 4.3}$	$28.8_{\pm 12.7} \rightarrow 91.2_{\pm 6.8}$
cube-double-play-singletask-task3-v0	$20.0_{\pm 11.3} \rightarrow 94.0_{\pm 3.3}$	$12.4_{\pm 4.6} \rightarrow 96.4_{\pm 4.3}$	$19.2_{\pm 6.0} \rightarrow 96.8_{\pm 4.1}$
cube-double-play-singletask-task4-v0	$4.0_{\pm 1.3} \rightarrow 7.2_{\pm 8.9}$	$5.2_{\pm 4.3} \rightarrow 16.4_{\pm 21.4}$	$3.2_{\pm 2.0} \rightarrow 4.8_{\pm 4.7}$
cube-double-play-singletask-task5-v0	$15.2_{\pm 7.2} \rightarrow 94.8_{\pm 3.2}$	$9.2_{\pm 5.9} \rightarrow 92.8_{\pm 4.1}$	$10.8_{\pm 8.4} \rightarrow 96.8_{\pm 2.0}$
cube-triple-play-singletask-task1-v0	$12.0_{\pm 5.5} \rightarrow 88.8_{\pm 11.9}$	$7.2_{\pm 2.0} \rightarrow 65.6_{\pm 22.5}$	$13.6_{\pm 5.6} \rightarrow 91.2_{\pm 9.6}$
cube-triple-play-singletask-task2-v0	$0.0_{\pm 0.0} \rightarrow 0.0_{\pm 0.0}$	$0.0_{\pm 0.0} \rightarrow 0.0_{\pm 0.0}$	$0.0_{\pm 0.0} \rightarrow 0.0_{\pm 0.0}$
cube-triple-play-singletask-task3-v0	$0.0_{\pm 0.0} \rightarrow 8.0_{\pm 6.3}$	$0.0_{\pm 0.0} \rightarrow 2.8_{\pm 3.7}$	$0.8_{\pm 1.0} \rightarrow 6.8_{\pm 10.1}$
cube-triple-play-singletask-task4-v0	$0.0_{\pm 0.0} \rightarrow 0.0_{\pm 0.0}$	$0.0_{\pm 0.0} \rightarrow 0.0_{\pm 0.0}$	$0.0_{\pm 0.0} \rightarrow 0.0_{\pm 0.0}$
cube-triple-play-singletask-task5-v0	$0.0_{\pm 0.0} \rightarrow 0.0_{\pm 0.0}$	$0.0_{\pm 0.0} \rightarrow 0.0_{\pm 0.0}$	$0.0_{\pm 0.0} \rightarrow 0.0_{\pm 0.0}$
scene-play-singletask-task1-v0	$100.0_{\pm 0.0} \rightarrow 100.0_{\pm 0.0}$	$100.0_{\pm 0.0} \rightarrow 100.0_{\pm 0.0}$	$100.0_{\pm 0.0} \rightarrow 100.0_{\pm 0.0}$
scene-play-singletask-task2-v0	$75.2_{\pm 11.0} \rightarrow 100.0_{\pm 0.0}$	$92.0_{\pm 4.7} \rightarrow 100.0_{\pm 0.0}$	$90.0_{\pm 6.6} \rightarrow 100.0_{\pm 0.0}$
scene-play-singletask-task3-v0	$94.0_{\pm 3.3} \rightarrow 99.6_{\pm 0.8}$	$98.4_{\pm 1.5} \rightarrow 100.0_{\pm 0.0}$	$99.2_{\pm 1.0} \rightarrow 100.0_{\pm 0.0}$
scene-play-singletask-task4-v0	$8.4_{\pm 3.4} \rightarrow 0.0_{\pm 0.0}$	$1.2_{\pm 1.6} \rightarrow 0.0_{\pm 0.0}$	$4.8_{\pm 3.5} \rightarrow 0.0_{\pm 0.0}$
scene-play-singletask-task5-v0	$0.0_{\pm 0.0} \rightarrow 0.0_{\pm 0.0}$	$0.0_{\pm 0.0} \rightarrow 0.0_{\pm 0.0}$	$0.0_{\pm 0.0} \rightarrow 0.0_{\pm 0.0}$
scene-noisy-singletask-task1-v0	$83.6_{\pm 6.1} \rightarrow 100.0_{\pm 0.0}$	$100.0_{\pm 0.0} \rightarrow 100.0_{\pm 0.0}$	$99.6_{\pm 0.8} \rightarrow 100.0_{\pm 0.0}$
scene-noisy-singletask-task2-v0	$0.8_{\pm 1.0} \rightarrow 0.0_{\pm 0.0}$	$28.6_{\pm 16.3} \rightarrow 100.0_{\pm 0.0}$	$11.6_{\pm 1.5} \rightarrow 100.0_{\pm 0.0}$
scene-noisy-singletask-task3-v0	$28.0_{\pm 5.8} \rightarrow 99.2_{\pm 1.0}$	$86.0_{\pm 6.6} \rightarrow 100.0_{\pm 0.0}$	$83.2_{\pm 7.0} \rightarrow 100.0_{\pm 0.0}$
scene-noisy-singletask-task4-v0	$2.4_{\pm 2.3} \rightarrow 0.0_{\pm 0.0}$	$1.6_{\pm 2.3} \rightarrow 0.0_{\pm 0.0}$	$2.4_{\pm 2.0} \rightarrow 0.0_{\pm 0.0}$
scene-noisy-singletask-task5-v0	$0.0_{\pm 0.0} \rightarrow 0.0_{\pm 0.0}$	$0.0_{\pm 0.0} \rightarrow 0.0_{\pm 0.0}$	$0.0_{\pm 0.0} \rightarrow 0.0_{\pm 0.0}$
puzzle-3x3-play-singletask-task1-v0	$92.8_{\pm 3.0} \rightarrow 100.0_{\pm 0.0}$	$90.4_{\pm 2.9} \rightarrow 100.0_{\pm 0.0}$	$93.2_{\pm 2.0} \rightarrow 100.0_{\pm 0.0}$
puzzle-3x3-play-singletask-task2-v0	$17.2_{\pm 4.1} \rightarrow 0.0_{\pm 0.0}$	$4.4_{\pm 3.9} \rightarrow 0.0_{\pm 0.0}$	$8.4_{\pm 3.4} \rightarrow 0.0_{\pm 0.0}$
puzzle-3x3-play-singletask-task3-v0	$9.2_{\pm 6.0} \rightarrow 0.0_{\pm 0.0}$	$7.6_{\pm 4.3} \rightarrow 0.0_{\pm 0.0}$	$8.0_{\pm 2.2} \rightarrow 0.0_{\pm 0.0}$
puzzle-3x3-play-singletask-task4-v0	$12.8_{\pm 6.3} \rightarrow 0.0_{\pm 0.0}$	$8.8_{\pm 2.7} \rightarrow 0.0_{\pm 0.0}$	$13.2_{\pm 2.7} \rightarrow 0.0_{\pm 0.0}$
puzzle-3x3-play-singletask-task5-v0	$21.6_{\pm 6.4} \rightarrow 0.0_{\pm 0.0}$	$11.6_{\pm 4.6} \rightarrow 0.0_{\pm 0.0}$	$18.4_{\pm 5.9} \rightarrow 0.0_{\pm 0.0}$
puzzle-4x4-play-singletask-task1-v0	$28.0_{\pm 7.0} \rightarrow 100.0_{\pm 0.0}$	$41.2_{\pm 6.5} \rightarrow 100.0_{\pm 0.0}$	$41.2_{\pm 9.2} \rightarrow 100.0_{\pm 0.0}$
puzzle-4x4-play-singletask-task2-v0	$14.0_{\pm 4.6} \rightarrow 0.0_{\pm 0.0}$	$16.0_{\pm 6.8} \rightarrow 0.0_{\pm 0.0}$	$21.6_{\pm 3.7} \rightarrow 0.0_{\pm 0.0}$
puzzle-4x4-play-singletask-task3-v0	$22.0_{\pm 8.7} \rightarrow 79.6_{\pm 38.8}$	$34.8_{\pm 7.2} \rightarrow 100.0_{\pm 0.0}$	$32.0_{\pm 6.1} \rightarrow 100.0_{\pm 0.0}$
puzzle-4x4-play-singletask-task4-v0	$6.8_{\pm 3.0} \rightarrow 60.0_{\pm 49.0}$	$19.2_{\pm 7.9} \rightarrow 100.0_{\pm 0.0}$	$13.6_{\pm 4.1} \rightarrow 80.0_{\pm 40.0}$
puzzle-4x4-play-singletask-task5-v0	$5.2_{\pm 3.9} \rightarrow 0.0_{\pm 0.0}$	$13.6_{\pm 6.5} \rightarrow 0.0_{\pm 0.0}$	$13.6_{\pm 4.1} \rightarrow 0.0_{\pm 0.0}$

Task	FQL	FQL-V	SAWR
puzzle-4x4-noisy-singletask-task1-v0	$7.2_{\pm 2.7} \rightarrow 65.2_{\pm 13.1}$	$36.8_{\pm 5.3} \rightarrow 100.0_{\pm 0.0}$	$22.4_{\pm 9.8} \rightarrow 100.0_{\pm 0.0}$
puzzle-4x4-noisy-singletask-task2-v0	$3.2_{\pm 2.7} \rightarrow 0.0_{\pm 0.0}$	$4.0_{\pm 2.5} \rightarrow 0.0_{\pm 0.0}$	$10.4_{\pm 3.9} \rightarrow 0.0_{\pm 0.0}$
puzzle-4x4-noisy-singletask-task3-v0	$3.2_{\pm 2.7} \rightarrow 3.2_{\pm 4.1}$	$22.4_{\pm 5.0} \rightarrow 88.8_{\pm 21.4}$	$14.0_{\pm 5.2} \rightarrow 12.8_{\pm 14.8}$
puzzle-4x4-noisy-singletask-task4-v0	$2.0_{\pm 1.3} \rightarrow 0.4_{\pm 0.8}$	$3.6_{\pm 2.9} \rightarrow 4.0_{\pm 8.0}$	$4.8_{\pm 3.2} \rightarrow 86.0_{\pm 13.9}$
puzzle-4x4-noisy-singletask-task5-v0	$0.8_{\pm 1.0} \rightarrow 0.0_{\pm 0.0}$	$3.6_{\pm 2.3} \rightarrow 0.0_{\pm 0.0}$	$4.4_{\pm 3.2} \rightarrow 0.0_{\pm 0.0}$

B.5 Instantiation on Q-chunking

Because SAWR only reweights the conditional flow matching loss, the same weighting can be applied to other flow-policy algorithms. We instantiate it on Q-chunking (Li et al., 2025), which uses a flow policy to generate action chunks and best-of-n sampling to select among candidates. The comparison uses QC, QC-V, and SAWR, matching the roles of FQL, FQL-V, and SAWR in Table 1.

We evaluate four OGBench manipulation families—`cube-triple-play`, `cube-triple-noisy`, `scene-noisy`, and `puzzle-4x4-noisy`—with five subtasks each. Training follows the same 1M-offline then 1M-online protocol, with 50 evaluation episodes per checkpoint. All methods use a horizon of 5 and a best-of-n actor with 32 samples. We report the mean and population standard deviation across three independent seeds.

Table 6 reports family-level offline $\rightarrow$ online success rate. SAWR attains the highest average online success rate (72.0), compared with 64.0 for QC and 67.5 for QC-V. The largest improvement is on `puzzle-4x4-noisy`, where online success rate rises from 73.5 for QC and 66.7 for QC-V to 86.7 for SAWR. Table 7 gives the corresponding per-subtask results. These results further demonstrate that SAWR generalizes beyond a specific flow policy algorithm and can improve performance across different flow policy backbones.

Table 6: Q-chunking offline-to-online success rate (%).

Task	QC	QC-V	SAWR
cube-triple-play-singletask (5 tasks)	$3.3_{\pm 1.1} \rightarrow 61.2_{\pm 1.4}$	$\mathbf{5.9}_{\pm 2.5} \rightarrow \mathbf{66.8}_{\pm 1.4}$	$4.5_{\pm 2.0} \rightarrow 64.7_{\pm 3.1}$
cube-triple-noisy-singletask (5 tasks)	$1.2_{\pm 0.3} \rightarrow 24.0_{\pm 2.4}$	$\mathbf{0.8}_{\pm 0.6} \rightarrow \mathbf{38.7}_{\pm 1.6}$	$0.9_{\pm 0.7} \rightarrow 38.5_{\pm 4.3}$
scene-noisy-singletask (5 tasks)	$75.2_{\pm 9.9} \rightarrow 97.3_{\pm 0.7}$	$74.3_{\pm 5.1} \rightarrow 97.6_{\pm 0.6}$	$\mathbf{72.0}_{\pm 5.4} \rightarrow \mathbf{97.9}_{\pm 0.2}$
puzzle-4x4-noisy-singletask (5 tasks)	$0.3_{\pm 0.2} \rightarrow 73.5_{\pm 9.2}$	$0.4_{\pm 0.3} \rightarrow 66.7_{\pm 18.9}$	$\mathbf{0.8}_{\pm 0.3} \rightarrow \mathbf{86.7}_{\pm 9.4}$
Q-chunking average	20.0$\rightarrow$ 64.0	20.4$\rightarrow$ 67.5	**19.6$\rightarrow$ 72.0**

Table 7: **Q-chunking per-subtask results.**

Task	QC	QC-V	SAWR
cube-triple-play-singletask-task1-v0	$16.0_{\pm 5.9} \rightarrow 100.0_{\pm 0.0}$	$25.3_{\pm 13.9} \rightarrow 100.0_{\pm 0.0}$	$21.3_{\pm 10.0} \rightarrow 100.0_{\pm 0.0}$
cube-triple-play-singletask-task2-v0	$0.0_{\pm 0.0} \rightarrow 94.0_{\pm 4.3}$	$0.7_{\pm 0.9} \rightarrow 90.7_{\pm 5.2}$	$0.7_{\pm 0.9} \rightarrow 92.7_{\pm 1.9}$
cube-triple-play-singletask-task3-v0	$0.7_{\pm 0.9} \rightarrow 71.3_{\pm 11.1}$	$3.3_{\pm 3.4} \rightarrow 78.0_{\pm 0.0}$	$0.7_{\pm 0.9} \rightarrow 79.3_{\pm 4.7}$
cube-triple-play-singletask-task4-v0	$0.0_{\pm 0.0} \rightarrow 40.7_{\pm 6.2}$	$0.0_{\pm 0.0} \rightarrow 65.3_{\pm 9.8}$	$0.0_{\pm 0.0} \rightarrow 51.3_{\pm 19.7}$
cube-triple-play-singletask-task5-v0	$0.0_{\pm 0.0} \rightarrow 0.0_{\pm 0.0}$	$0.0_{\pm 0.0} \rightarrow 0.0_{\pm 0.0}$	$0.0_{\pm 0.0} \rightarrow 0.0_{\pm 0.0}$
cube-triple-noisy-singletask-task1-v0	$6.0_{\pm 1.6} \rightarrow 100.0_{\pm 0.0}$	$4.0_{\pm 2.8} \rightarrow 100.0_{\pm 0.0}$	$4.0_{\pm 2.8} \rightarrow 100.0_{\pm 0.0}$
cube-triple-noisy-singletask-task2-v0	$0.0_{\pm 0.0} \rightarrow 15.3_{\pm 11.6}$	$0.0_{\pm 0.0} \rightarrow 81.3_{\pm 5.2}$	$0.0_{\pm 0.0} \rightarrow 82.7_{\pm 14.7}$
cube-triple-noisy-singletask-task3-v0	$0.0_{\pm 0.0} \rightarrow 1.3_{\pm 0.9}$	$0.0_{\pm 0.0} \rightarrow 2.0_{\pm 1.6}$	$0.7_{\pm 0.9} \rightarrow 2.0_{\pm 2.8}$
cube-triple-noisy-singletask-task4-v0	$0.0_{\pm 0.0} \rightarrow 3.3_{\pm 0.9}$	$0.0_{\pm 0.0} \rightarrow 10.0_{\pm 5.7}$	$0.0_{\pm 0.0} \rightarrow 8.0_{\pm 5.9}$
cube-triple-noisy-singletask-task5-v0	$0.0_{\pm 0.0} \rightarrow 0.0_{\pm 0.0}$	$0.0_{\pm 0.0} \rightarrow 0.0_{\pm 0.0}$	$0.0_{\pm 0.0} \rightarrow 0.0_{\pm 0.0}$
scene-noisy-singletask-task1-v0	$100.0_{\pm 0.0} \rightarrow 100.0_{\pm 0.0}$	$100.0_{\pm 0.0} \rightarrow 100.0_{\pm 0.0}$	$100.0_{\pm 0.0} \rightarrow 100.0_{\pm 0.0}$
scene-noisy-singletask-task2-v0	$99.3_{\pm 0.9} \rightarrow 100.0_{\pm 0.0}$	$98.0_{\pm 1.6} \rightarrow 100.0_{\pm 0.0}$	$97.3_{\pm 0.9} \rightarrow 100.0_{\pm 0.0}$
scene-noisy-singletask-task3-v0	$92.7_{\pm 2.5} \rightarrow 100.0_{\pm 0.0}$	$89.3_{\pm 3.4} \rightarrow 100.0_{\pm 0.0}$	$93.3_{\pm 2.5} \rightarrow 100.0_{\pm 0.0}$
scene-noisy-singletask-task4-v0	$48.7_{\pm 29.0} \rightarrow 97.3_{\pm 2.5}$	$46.0_{\pm 25.7} \rightarrow 97.3_{\pm 0.9}$	$47.3_{\pm 12.3} \rightarrow 98.0_{\pm 0.0}$
scene-noisy-singletask-task5-v0	$35.3_{\pm 18.0} \rightarrow 89.3_{\pm 4.1}$	$38.0_{\pm 1.6} \rightarrow 90.7_{\pm 1.9}$	$22.0_{\pm 11.8} \rightarrow 91.3_{\pm 0.9}$
puzzle-4x4-noisy-singletask-task1-v0	$0.0_{\pm 0.0} \rightarrow 67.3_{\pm 46.2}$	$0.7_{\pm 0.9} \rightarrow 33.3_{\pm 47.1}$	$0.7_{\pm 0.9} \rightarrow 100.0_{\pm 0.0}$
puzzle-4x4-noisy-singletask-task2-v0	$0.0_{\pm 0.0} \rightarrow 100.0_{\pm 0.0}$	$0.0_{\pm 0.0} \rightarrow 33.3_{\pm 47.1}$	$0.0_{\pm 0.0} \rightarrow 66.7_{\pm 47.1}$
puzzle-4x4-noisy-singletask-task3-v0	$0.0_{\pm 0.0} \rightarrow 0.0_{\pm 0.0}$	$0.0_{\pm 0.0} \rightarrow 66.7_{\pm 47.1}$	$0.0_{\pm 0.0} \rightarrow 66.7_{\pm 47.1}$
puzzle-4x4-noisy-singletask-task4-v0	$0.7_{\pm 0.9} \rightarrow 100.0_{\pm 0.0}$	$0.0_{\pm 0.0} \rightarrow 100.0_{\pm 0.0}$	$2.7_{\pm 0.9} \rightarrow 100.0_{\pm 0.0}$
puzzle-4x4-noisy-singletask-task5-v0	$0.7_{\pm 0.9} \rightarrow 100.0_{\pm 0.0}$	$1.3_{\pm 0.9} \rightarrow 100.0_{\pm 0.0}$	$0.7_{\pm 0.9} \rightarrow 100.0_{\pm 0.0}$

B.6 Complete benchmark learning curves

We present the corresponding training curves for Table 1 in Figures 9, 10, and 11.

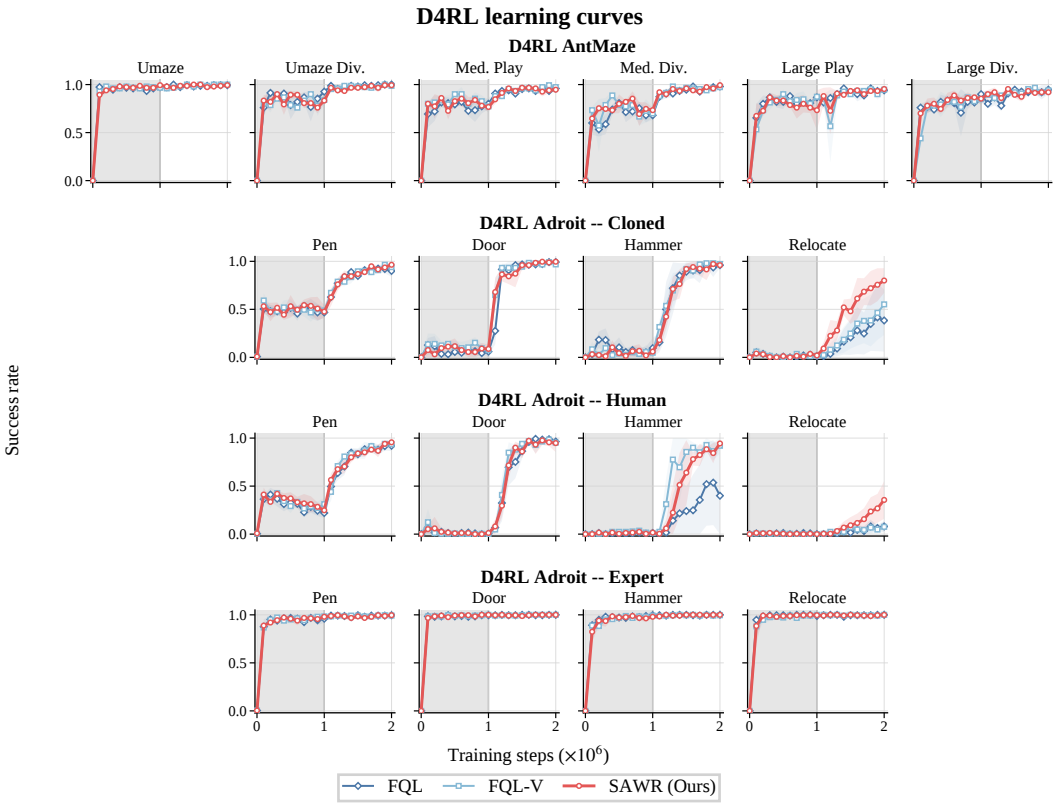


Figure 9: **D4RL learning curves.**

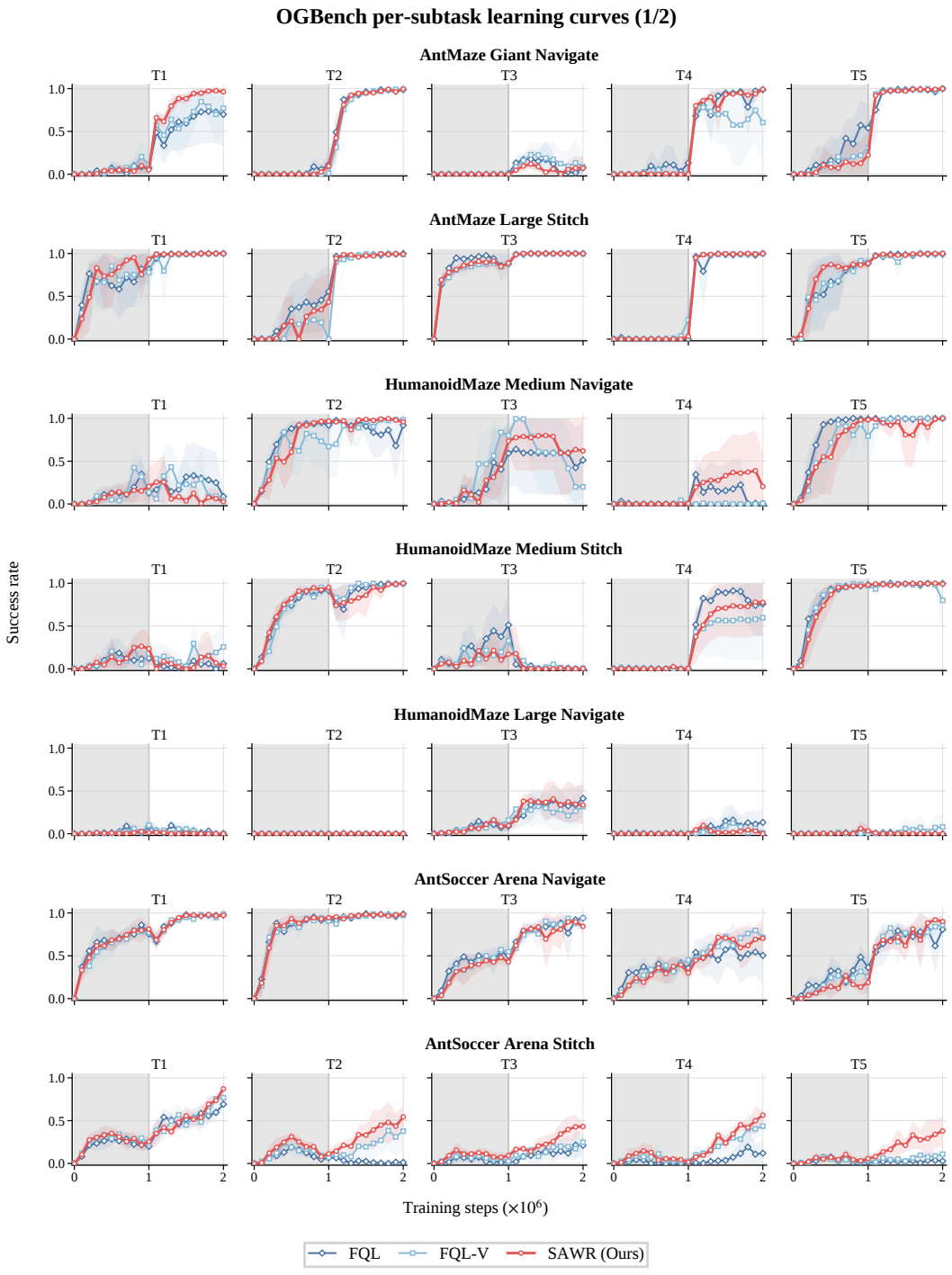


Figure 10: **OGBench learning curves (1/2).**

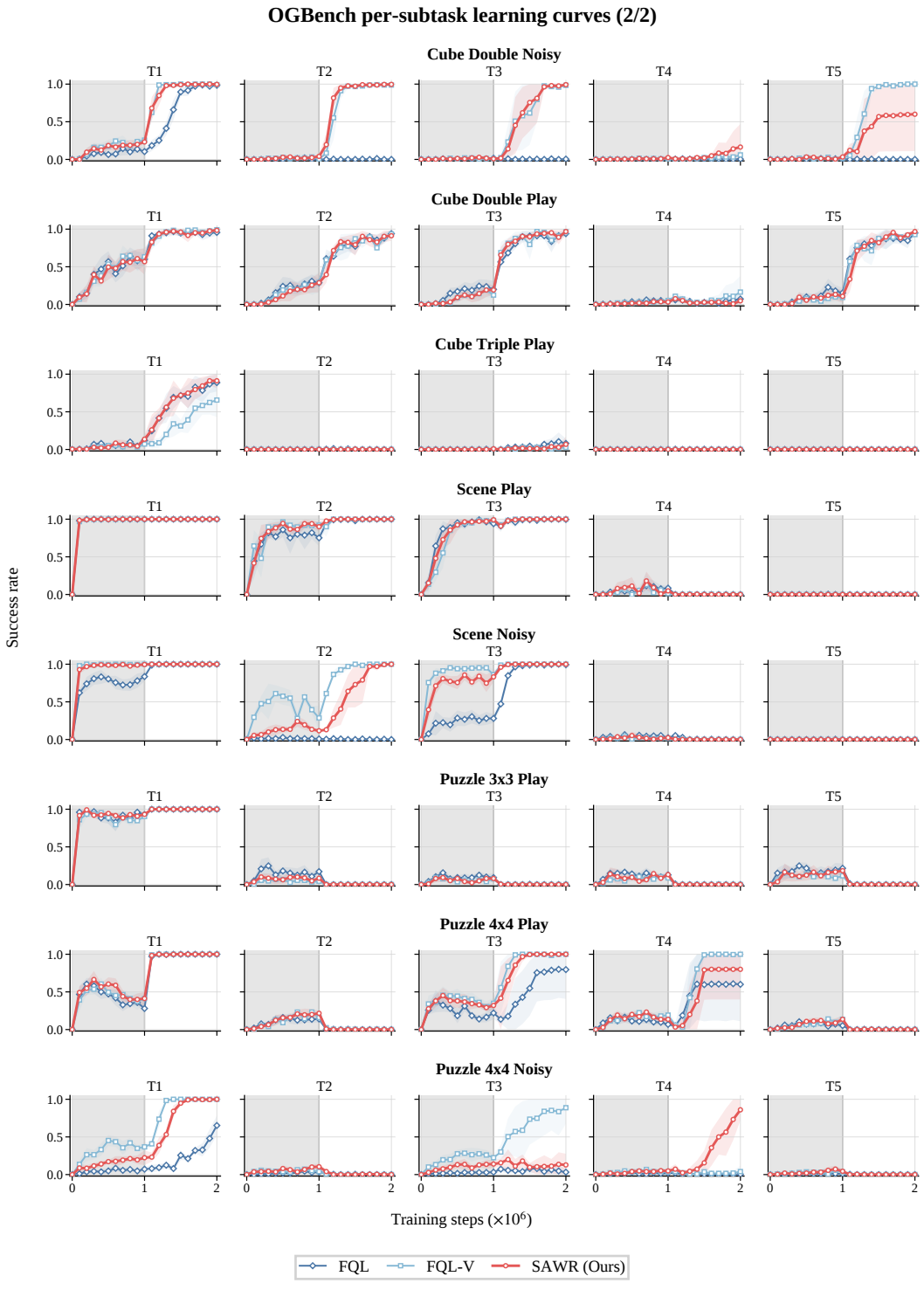


Figure 11: **OGBench learning curves (2/2).**

C Theoretical Details

C.1 Technical setup and proof of the condition ensuring that the reference remains in the same mode

Mode region setup. Fix a state s and let $\mathcal{M}_m(s) \subseteq \mathcal{A}$ denote a mode region containing the anchor a. Define the anchor margin

$$\rho_m(a) := \operatorname{dist}(a, \mathcal{A} \setminus \mathcal{M}_m(s)). \tag{20}$$

Thus, every $b \in \mathcal{A}$ satisfying $\|b - a\|_2 < \rho_m(a)$ belongs to the same mode as a. The condition $\rho_m(a) > 0$ restricts the result to anchors in the interior of a mode region. Within mode m, write

$$Q_{\bar{\phi}}(s, b) = \mu_m(s) + q_m(s, b), \qquad b \in \mathcal{M}_m(s), \tag{21}$$

where $\mu_m(s)$ is an offset shared by mode m and $q_m(s, b)$ captures variation within that mode.

Regularity assumptions. For every reference time and noise realization under consideration, assume that the flow policy produces a valid rollout from x_t^{ref} to time 1. Suppose a deterministic constant $U_{\max} < \infty$ uniformly bounds $\|u_\theta(s, x, \tau)\|_2$ along every such trajectory and at every Euler evaluation point, and that $Q_{\bar{\phi}}(s, \cdot)$ is L_Q-Lipschitz on $\mathcal{A}$. For the Euler implementation, the step sizes are assumed nonnegative and to sum to $1 - t$.

Proof of Proposition 1. Write $x_t = x_t^{\mathrm{ref}}$ and let $y_\tau = \Phi_\theta^{t \to \tau}(s, x_t)$ denote the reference trajectory before endpoint clipping. By the integral form of the flow ODE and the assumed velocity bound,

$$\|y_1 - x_t\|_2 = \left\| \int_t^1 u_\theta(s, y_\tau, \tau)\, \mathrm{d}\tau \right\|_2 \le \int_t^1 \|u_\theta(s, y_\tau, \tau)\|_2\, \mathrm{d}\tau \le U_{\max}(1 - t). \tag{22}$$

The same inequality holds for the K-step Euler implementation when its nonnegative step sizes sum to $1 - t$ and the velocity bound holds at every Euler evaluation point. From the partial noising construction,

$$\|x_t - a\|_2 = (1 - t)\|\varepsilon^{\mathrm{ref}} - a\|_2. \tag{23}$$

Euclidean projection onto a closed convex set is non-expansive and fixes every $a \in \mathcal{A}$. Therefore, the triangle inequality together with Equation 22 and Equation 23 gives

$$\begin{aligned} \|a'_\theta - a\|_2 &= \|\Pi_{\mathcal{A}}(y_1) - a\|_2 \le \|y_1 - a\|_2 && (24)\\ &\le \|y_1 - x_t\|_2 + \|x_t - a\|_2 && (25)\\ &\le (1 - t)\big(U_{\max} + \|\varepsilon^{\mathrm{ref}} - a\|_2\big) = R_t. && (26) \end{aligned}$$

For fixed $(s, a, \varepsilon^{\mathrm{ref}})$, the factor multiplying $1 - t$ in R_t is constant, so the certified distance bound decays linearly as $t \to 1$. The L_Q-Lipschitz property of the target critic then gives

$$|\Delta(s, a; t, \varepsilon^{\mathrm{ref}})| = |Q_{\bar{\phi}}(s, a) - Q_{\bar{\phi}}(s, a'_\theta)| \le L_Q \|a - a'_\theta\|_2 \le L_Q R_t.$$

Finally, if $R_t < \rho_m(a)$, then $\|a'_\theta - a\|_2 < \rho_m(a)$. By the definition of the anchor margin, every action outside $\mathcal{M}_m(s)$ is at least $\rho_m(a)$ away from a; hence $a'_\theta \in \mathcal{M}_m(s)$. Both actions therefore share the decomposition within mode m, and

$$\begin{aligned} \Delta &= Q_{\bar{\phi}}(s, a) - Q_{\bar{\phi}}(s, a'_\theta)\\ &= \mu_m(s) + q_m(s, a) - \mu_m(s) - q_m(s, a'_\theta)\\ &= q_m(s, a) - q_m(s, a'_\theta), \end{aligned}$$

which establishes cancellation of the shared offset within the mode. □

C.2 Weight normalization and reliability gate

Robust batch standardization. For a minibatch of sample-anchored advantages, define

$$\begin{aligned} m_\Delta &:= \operatorname{median}_i \Delta_i, \qquad & d_\Delta &:= \operatorname{median}_i |\Delta_i - m_\Delta|,\\ \varepsilon_{\mathrm{num}} &:= 10^{-6}, \qquad & s_\Delta &:= \max\{1.4826 d_\Delta, \varepsilon_{\mathrm{num}}\}. \end{aligned} \tag{27}$$

We use $z_i = (\Delta_i - m_\Delta)/s_\Delta$. The factor 1.4826 makes the MAD a consistent scale estimate under Gaussian data, while the numerical floor handles batches with nearly identical advantages. In particular, if all Δ_i are equal, then every $z_i = 0$.

Temperature selection for the target ESS. For a candidate temperature $\eta > 0$, we compute numerically stable unnormalized weights

$$v_i(\eta) := \exp\left(\frac{z_i}{\eta} - \max_j \frac{z_j}{\eta}\right) \tag{28}$$

and their normalized effective sample size fraction

$$\frac{\mathrm{ESS}(\eta)}{B} := \frac{\left(\sum_i v_i(\eta)\right)^2}{B \sum_i v_i(\eta)^2}. \tag{29}$$

We target $\mathrm{ESS}/B = 0.7$. Since this quantity is nondecreasing in η, we perform 14 bisection iterations on a logarithmic temperature scale over $\eta \in [10^{-3}, 10^3]$ and use the nearest endpoint when the target is outside the attainable range. With the selected $\eta^\star$, the weights normalized to have mean one are

$$\widetilde{w}_i := \frac{v_i(\eta^\star)}{B^{-1} \sum_j v_j(\eta^\star)}. \tag{30}$$

The maximum subtraction above affects neither $\widetilde{w}_i$ nor its ESS and prevents overflow in the exponential. The implementation additionally adds 10^{-12} to the ESS and mean weight denominators; we omit these terms above for readability.

Critic reliability gate. Let $\mathcal{L}_{Q,k}$ be the critic squared Bellman error and $V_{y,k}$ the minibatch variance of its target at update k. We maintain exponential moving averages with decay $\lambda_g = 0.999$,

$$\bar{\mathcal{L}}_{Q,k} = \lambda_g \bar{\mathcal{L}}_{Q,k-1} + (1 - \lambda_g)\mathcal{L}_{Q,k}, \qquad \bar{V}_{y,k} = \lambda_g \bar{V}_{y,k-1} + (1 - \lambda_g)V_{y,k}, \tag{31}$$

initialized from the first valid update. The reliability statistic and gate are

$$R^2_{Q,k} := 1 - \frac{\bar{\mathcal{L}}_{Q,k}}{\max\{\bar{V}_{y,k}, 10^{-8}\}}, \qquad g_k := \sigma\left(\frac{R^2_{Q,k} - 0.5}{0.05}\right), \tag{32}$$

where an invalid or uninitialized statistic sets $g_k = 0$. The final weights are

$$\hat{w}_i = (1 - g_k) + g_k \widetilde{w}_i, \tag{33}$$

and are treated as stop-gradient. Thus $g_k = 0$ recovers uniform flow matching, $g_k = 1$ applies the full weights selected for the target ESS, and $B^{-1} \sum_i \hat{w}_i = 1$. The reported experiments enable this weighting in both offline and online updates.

The ESS target applies before gating. If $e := \mathrm{ESS}(\widetilde{w})/B$, then the final ESS fraction is

$$\frac{\mathrm{ESS}(\hat{w})}{B} = \frac{1}{1 + g_k^2(e^{-1} - 1)}, \tag{34}$$

so the reliability gate can only make the weights more uniform. The entire normalization is unchanged whenever $\Delta_{1:B}$ is unchanged; hence the cancellation of the shared offset within a mode in Proposition 1 carries through to the implemented weights when the gate statistic is held fixed.

Batch scale invariance. The distinction between raw advantage contraction and the strength of the normalized weights is important. Let $\Delta_i^{c,d} = c\Delta_i + d$ for constants $c > 0$ and $d \in \mathbb{R}$. The transformed median and MAD are $m_\Delta^{c,d} = c m_\Delta + d$ and $d_\Delta^{c,d} = c d_\Delta$. Therefore, whenever the MAD term rather than the numerical floor determines both scales,

$$\frac{\Delta_i^{c,d} - m_\Delta^{c,d}}{1.4826 d_\Delta^{c,d}} = \frac{\Delta_i - m_\Delta}{1.4826 d_\Delta}. \tag{35}$$

The standardized scores, selected temperature, and final weights are then exactly unchanged for a fixed gate. Thus a common positive rescaling of the raw anchor–reference gaps does not, by itself, weaken the implemented weights. In particular, the raw advantage bound in Proposition 1 does not establish monotonic weakening of normalized policy updates as reference time increases.

Reference time and relative weighting. This batch-level invariance does not imply invariance to the reference times. In the implementation, we independently draw $t_i \sim p_t$ for each sample, whereas the median, MAD, and ESS temperature are computed jointly across the minibatch. Changing the reference times can alter whether comparisons remain within a mode and can change the relative magnitudes and ordering of the advantages; these changes need not constitute a common affine transformation of $\Delta_{1:B}$. Even when a local approximation takes the form $\Delta_i \approx (1 - t_i)d_i$, with d_i denoting a sample-specific local coefficient, the factors $1 - t_i$ generally differ across samples and are not separately canceled by a single batch scale. Normalization therefore preserves the cancellation of shared within-mode offsets established above, while retaining dependence on the relative advantage structure induced by the references.

The ESS target controls weight concentration before gating, but does not determine which samples receive greater weight. For per-sample flow matching losses $\ell_i(\theta)$, the stop-gradient weights give

$$\nabla_\theta \mathcal{L}_{\mathrm{FM}}^{\mathrm{SAWR}} = \frac{1}{B}\sum_i \nabla_\theta \ell_i + \frac{1}{B}\sum_i (\hat{w}_i - 1)\nabla_\theta \ell_i. \tag{36}$$

Consequently, even at the same ESS, different weight assignments can produce different policy updates. The trade-off discussed in the main text concerns comparison locality and raw value differences; its effect after normalization depends on the resulting relative weights and their association with sample gradients.

Mode mass under the normalized update. Offset cancellation alone does not imply exact preservation of mode probabilities. To make the connection precise, fix a realized minibatch and partition its anchors into mode groups I_m (at a fixed state, these are the regions $\mathcal{M}_m(s)$). Let $p_m = |I_m|/B > 0$ denote the unweighted empirical mass, and define

$$A_m := \frac{1}{|I_m|}\sum_{i \in I_m} v_i(\eta^\star), \qquad A := \frac{1}{B}\sum_i v_i(\eta^\star) = \sum_m p_m A_m. \tag{37}$$

All quantities are evaluated after the implemented MAD standardization and ESS temperature selection. Retaining the implementation's mean-denominator stabilizer $\delta_w = 10^{-12}$, the normalized mass assigned to group m by the weighted imitation objective is

$$p_m^+ := \frac{\sum_{i \in I_m} \hat{w}_i}{\sum_i \hat{w}_i} = p_m \frac{(1-g)(A+\delta_w) + gA_m}{(1-g)(A+\delta_w) + gA}. \tag{38}$$

This identity follows by substituting $\hat{w}_i = (1-g) + gv_i/(A+\delta_w)$ and summing over the group and batch. It yields the following precise preservation statements.

Proposition 2 (Conditional mode-mass preservation)**.** *For the normalized weighted target in Equation 38, if $g = 0$ or $A_m = A$ for every mode, then $p_m^+ = p_m$. In particular, identical empirical distributions of standardized advantages across modes suffice for exact preservation, even when the weights vary within each mode. More generally,*

$$p_m^+ \geq (1-g)p_m. \tag{39}$$

Thus, if $g \leq 1 - \gamma$ for some $\gamma > 0$, every represented mode retains at least a fraction γ of its replay mass in this weighted target. If all anchor–reference pairs remain within their respective modes, changes to the shared mode offsets $\mu_m(s)$ leave these masses unchanged for fixed references and gate.

Proof. The equality statements follow directly from Equation 38. For the lower bound, $v_i \geq 0$ implies $\hat{w}_i \geq 1 - g$, while $B^{-1}\sum_i \hat{w}_i = (1-g) + gA/(A+\delta_w) \leq 1$. Dividing the group weight sum by the total therefore gives Equation 39. Finally, same-mode references cancel each $\mu_m(s)$ before standardization. The entire advantage vector, and hence its MAD, selected temperature, weights, and normalized group masses, are unchanged when the gate and references are held fixed. □

These statements concern the reweighted replay target of the flow matching term. They transfer to the generated distribution only under idealized exact fitting of that target with sufficient model capacity; they do not guarantee the mode probabilities of a finite neural-policy update or of the distilled execution policy. In particular, locality does not force the within-mode factors A_m to be

equal, and an ESS target controls overall concentration rather than the mass of each mode. At $g = 1$, unequal factors can make a mode's mass arbitrarily small despite positive weights in exact arithmetic. Even with $g < 1$, repeated reweighting of a changing replay distribution provides no time-uniform floor without a lower bound on replay mode mass and on $1 - g$. Thus, the implemented update removes suppression caused solely by shared mode offsets and admits the conditional preservation and mass-floor guarantees above, rather than an unconditional claim that mode collapse is impossible.

Endpoint behavior. Uniform weighting is recovered at the exact endpoint. If every $t_i = 1$, Proposition 1 gives $\Delta_i = 0$, so the numerical scale floor yields $z_i = 0$ and hence $\widetilde{w}_i = \hat{w}_i = 1$. More generally, the positive numerical floor and the lower bound on the temperature imply $\hat{w}_i \to 1$ as every reference time in a fixed batch approaches one. In this limit the value-based reweighting correction vanishes, while the ordinary flow matching gradient remains. This endpoint continuity does not imply that the normalized weights weaken monotonically over the interior of $[0, 1]$.