Can K Heads Explore Better Than One in Online Reinforcement Learning?

Abhishek Jha [1] **Satyapragnya Kar** [1] **Kishlay Kumar** [1] **Stephanie Milani** [1] **Rajesh Ranganath** [1]

Abstract

Ensemble methods provide principled exploration for Gaussian policies in reinforcement learning, but their extension to the increasingly popular family of generative policies based on diffusion and flow matching is non-trivial. The standard ensemble template, namely K value heads trained on bootstrapped data subsets and selected by exploration strategies, fails for Q-weighted generative losses. When the K policy heads share a common Q-network, the Q-weighted importance weights are identical across heads by construction, so the denoising loss produces head invariant gradients in expectation and the heads collapse to a single function. Per-sample bootstrap masking cannot prevent this collapse because the mask only modulates a signal that is already head invariant before it enters the loss. This failure mode is specific to importance weighted generative losses and does not arise for standard Gaussian policy ensembles. To resolve this, we introduce Ensemble Soft Actor-Critic (`ESAC`), an ensemble framework that pairs each generative head with its own independent Q-network, so that per-head Q-weighted objectives drive the heads toward distinct optima. Within the same architecture we propose Trajectory Integrated Disagreement Exploration (`TIDE`), an action-selection rule that scores candidates by ensemble disagreement integrated across the full denoising trajectory and adaptively calibrates the bonus by the running correlation between velocity disagreement and Q-uncertainty. We formalize the collapse failure mode and the diversity restoration mechanism as theorems. On MuJoCo locomotion benchmarks with both diffusion and conditional flow matching backbones, ESAC delivers consistent multi-seed gains over its soft diffusion actor-critic backbone with the largest gains on tasks where directed exploration matters most.

[1]New York University, New York, USA. Correspondence to: Abhishek Jha <aj4718@nyu.edu>.

Accepted at ICML 2026 Workshop on Decision-Making from Offline Datasets to Online Adaptation: Black-Box Optimization to Reinforcement Learning.

1. Introduction

The choice of policy representation fundamentally shapes what an RL agent can learn. Gaussian policies, as used by SAC (Haarnoja et al., 2018) and its variants, parametrize actions as a diagonal Gaussian with network-predicted mean and variance. This is inherently unimodal, the policy expresses only one strategy per state, even when multiple qualitatively different actions are equally viable. The limitation bites in manipulation, where distinct grasp strategies may all succeed, and in locomotion, where multiple stable gaits coexist.

Generative policies based on diffusion (Ho et al., 2020; Song et al., 2021) and flow matching (Lipman et al., 2023; Liu et al., 2023) close this expressiveness gap by replacing the parametric Gaussian with an iterative sampling process, and recent work shows substantial empirical gains (Ma et al., 2025; Ding et al., 2024; McAllister et al., 2025; Wang et al., 2023; Hansen-Estruch et al., 2024; Ada et al., 2024). Yet *exploration* in these methods remains primitive. Every published flow or diffusion RL algorithm explores by adding isotropic Gaussian noise to the selected action or by inheriting SAC's entropy regularization, perturbing actions uniformly regardless of where the agent's uncertainty actually lies. On smooth reward landscapes like HalfCheetah this suffices, but on tasks with coordinated multi-joint control and early termination, such as Ant-v4 or Humanoid-v4, it is catastrophically sample inefficient. The Bayesian RL literature offers a principled alternative: Bootstrapped DQN (Osband et al., 2016) maintains K value heads trained on bootstrapped experience and selects one per episode via Thompson Sampling (Thompson, 1933), producing temporally coherent exploration. SUNRISE (Lee et al., 2021) extended this to continuous control using ensemble disagreement as a UCB-style (Auer et al., 2002) uncertainty proxy, and REDQ (Chen et al., 2021) showed that Q-ensembles support stable high-UTD training. But these methods operate on Gaussian policies (Fujimoto et al., 2018; Lillicrap

et al., 2016), and their extension to generative policy classes is non-trivial.

The naive extension fails. We show theoretically and empirically that any ensemble of generative policy heads trained with a Q-weighted denoising loss against a *shared* Q-network collapses to a single function regardless of bootstrap masking, independent network initialization, or architectural separation. The mechanism is structural: the Q-weighted denoising gradient depends on the policy heads only through the candidate actions they sample, and a shared Q-landscape produces head-invariant gradients in expectation, so independent initialization shifts only the early dynamics rather than the long-run optimum. Bootstrap masking adds per-sample noise but this variance is dwarfed by the consistent mean signal from the shared Q, and Thompson Sampling degenerates to random selection among K near-identical policies. The collapse is a property of the Q-weighted objective itself and applies to any generative policy class trained by importance-weighted regression on a shared critic.

ESAC resolves this by training a small team of K independent (policy, Q) pairs. Each head k optimizes a Q-weighted objective using its *own* Q_{ϕ_k} for the importance weights, so different Q-landscapes drive different policies even though all heads are trained on the same data. What independence buys is not a different asymptote but a different *trajectory* through policy space, and the heads explore distinct regions early and concentrate the team's coverage of the action manifold. At every environment step each policy proposes a small batch of candidate actions, the team scores every candidate by its ensemble-mean return plus a disagreement bonus measured along the action-generation trajectory, and the highest-scoring action is executed. The disagreement bonus, which we call *TIDE* (Trajectory Integrated Disagreement Exploration), measures inter-head velocity-field disagreement at multiple points along the denoising trajectory rather than only at the final action. Two heads can produce nearly identical final actions via very different denoising paths; conversely, two heads can produce different final actions by chance even when their underlying posteriors agree. Trajectory-level disagreement separates these cases, and we calibrate the bonus strength online by the running Pearson correlation between trajectory disagreement and value uncertainty so that the bonus is active only when the two signals agree. The same recipe works with diffusion or flow-matching backbones with no further change.

ESAC is the first ensemble exploration framework for expressive generative policies, and the first to identify and resolve the shared-critic collapse mode. This combination unlocks directed, posterior-driven exploration over multimodal action distributions in continuous control, with consistent gains on hard locomotion tasks where undirected exploration plateaus and unimodal policies cannot represent the multiple viable gaits. The key contributions for this work are as follows:

1. **Collapse theorem and architectural fix for generative-policy ensembles.** We prove that any ensemble of generative heads trained with a Q-weighted loss against a shared critic collapses to a single function regardless of initialization or bootstrap masking, and identify per-head independent critics with circular twin-Q pessimism as the minimal architectural change that breaks this collapse. The result generalizes to any generative policy class trained by importance-weighted regression.

2. **Trajectory-integrated disagreement exploration (TIDE).** We introduce an action-selection rule that exploits the iterative structure of generative sampling by measuring inter-head disagreement at multiple denoising timesteps rather than only at the output. Bonus strength is gated online by the correlation between disagreement and Q-uncertainty, preventing over-exploration on unimodal landscapes while preserving gains where genuine posterior diversity exists.

3. **Backbone comparison and modality-aware analysis.** We provide the systematic comparison of diffusion and conditional flow-matching backbones inside a unified ensemble framework across five MuJoCo locomotion benchmarks and four additional tasks, showing flow matching maintains substantially higher inter-head diversity throughout training, and characterize when each backbone wins via Q-landscape modality analysis.

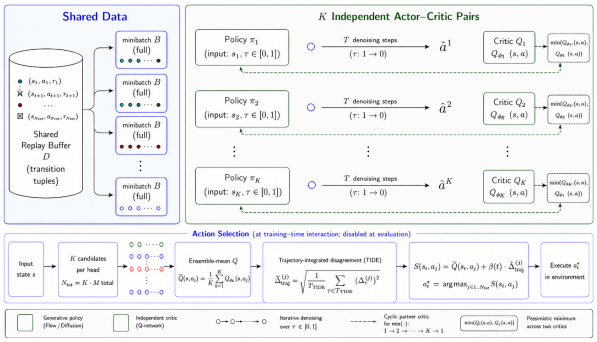

Figure 1. **ESAC training and inference pipeline.** A shared replay-buffer minibatch is fed to K independent actor-critic pairs. Each pair $(\pi_{\theta_k}, Q_{\phi_k})$ is trained jointly: the Q-loss uses a REDQ-style random subset minimum across the K Q-networks as a shared TD target, while the policy loss uses per-head importance weights computed via circular twin-Q pairing, so that each head's policy update sees a different pair of Q-networks. At interaction time (bottom), every head proposes candidate actions and TIDE selects one by combining the ensemble Q-mean with a trajectory-integrated inter-head disagreement bonus.

2. Preliminaries

Entropy Regularized Reinforcement Learning. We consider the standard infinite-horizon discounted MDP $(\mathcal{S}, \mathcal{A}, P, r, \gamma)$ with continuous state space $\mathcal{S} \subseteq \mathbb{R}^{d_s}$, continuous action space $\mathcal{A} \subseteq \mathbb{R}^{d_a}$, transition dynamics $P(s'|s,a)$, reward function $r(s,a)$, and discount factor $\gamma \in (0,1)$. Following the maximum entropy RL framework (Haarnoja et al., 2018), we seek a policy π maximizing the entropy-augmented objective

$$J(\pi) = \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t \left(r(s_t, a_t) + \tau_{\text{ent}} \, \mathcal{H}(\pi(\cdot|s_t)) \right) \right] \quad (1)$$

where $\tau_{\text{ent}} > 0$ is a learned entropy temperature coefficient and $\mathcal{H}(\pi(\cdot|s)) = -\mathbb{E}_{a\sim\pi}[\log \pi(a|s)]$ is the policy entropy. SAC and its derivatives (Haarnoja et al., 2018; Fujimoto et al., 2018) form the foundation for modern continuous control. The soft Q-function satisfies the soft Bellman equation

$$\begin{aligned} Q^\pi(s,a) &= r(s,a) + \gamma \mathbb{E}_{s'\sim P}\left[V^\pi(s')\right], \\ V^\pi(s) &= \mathbb{E}_{a\sim\pi}\left[Q^\pi(s,a) - \tau_{\text{ent}} \log \pi(a|s)\right] \end{aligned} \quad (2)$$

In practice, SAC (Haarnoja et al., 2018) maintains two Q-networks Q_{ϕ_1}, Q_{ϕ_2} (twin-Q) and uses the minimum for both policy optimization and Bellman targets to suppress overestimation bias. The target Q-networks $Q_{\bar{\phi}_i}$ are updated via exponential moving average with coefficient ρ, namely $\bar{\phi}_i \leftarrow (1-\rho)\bar{\phi}_i + \rho\,\phi_i$. For numerical stability, implementations parametrize $\log \tau_{\text{ent}}$ rather than τ_{ent} directly and recover the temperature via exponentiation.

Diffusion based policies SDAC (Ma et al., 2025) parametrizes the policy as a denoising diffusion probabilistic model (DDPM) (Ho et al., 2020). Given state s, action candidates are generated by starting from Gaussian noise $x_T \sim \mathcal{N}(0, I)$ and iteratively applying a learned denoiser ϵ_θ over T diffusion steps,

$$x_{k-1} = \frac{1}{\sqrt{\alpha_k}} \left(x_k - \frac{1-\alpha_k}{\sqrt{1-\bar{\alpha}_k}} \epsilon_\theta(x_k, k, s) \right) + \sigma_k z, \quad (3)$$

with $z \sim \mathcal{N}(0, I)$, where $\bar{\alpha}_k = \prod_{i=1}^{k} \alpha_i$ follows a predetermined noise schedule and σ_k controls the stochasticity of the reverse process. The forward (noising) process adds Gaussian noise, $x_k = \sqrt{\bar{\alpha}_k}\, a_0 + \sqrt{1-\bar{\alpha}_k}\, \epsilon$, where a_0 is a clean action and $\epsilon \sim \mathcal{N}(0, I)$.

The key innovation of SDAC is a *Q-weighted* denoising loss that addresses a fundamental mismatch in applying diffusion models to RL standard denoising score matching fits the policy to whatever actions appear in the replay buffer, but in RL we want the policy to concentrate mass on *high-value* actions, not arbitrary ones. SDAC resolves this by reweighting the denoising loss by Q-values, so the generative capacity is steered toward high-reward modes of the action distribution. For each state s in the batch, SDAC (i) draws N_{mc} Monte Carlo action candidates$\{\hat{a}_i\}_{i=1}^{N_{\text{mc}}}$ from the current policy, (ii) scores each with the Q-function and converts the scores into softmax importance weights w_i, (iii) samples a diffusion step $k \sim \mathcal{U}\{1, \ldots, T\}$ and i.i.d. noise $\epsilon^{(i)} \sim \mathcal{N}(0, I)$, and (iv) constructs the noisy targets $x_k^{(i)}$ via the forward diffusion process. Concretely,

$$\begin{aligned} w_i &= \frac{\exp(c\,Q(s,\hat{a}_i)\,/\,\tau_{\text{ent}})}{\sum_{j=1}^{M} \exp(c\,Q(s,\hat{a}_j)\,/\,\tau_{\text{ent}})}, \\ x_k^{(i)} &= \sqrt{\bar{\alpha}_k}\,\hat{a}_i + \sqrt{1-\bar{\alpha}_k}\,\epsilon^{(i)}, \end{aligned} \quad (4)$$

$$\mathcal{L}_{\text{SDAC}}(\theta) = \sum_{i=1}^{N_{\text{mc}}} w_i \left\| \epsilon_\theta\left(x_k^{(i)}, k, s\right) - \epsilon^{(i)} \right\|^2 \quad (5)$$

where $c = 5$ is the temperature scaling constant from the original work and $\bar{\alpha}_k$ is the cumulative noise schedule of Eq. 3. Each candidate $\hat{a}_i$ thus serves as a clean-action target a_0 in the standard DDPM forward process, and the score-matching loss is evaluated on the candidate-specific noisy state $x_k^{(i)}$ with its own injected noise $\epsilon^{(i)}$. The denominator in w_i is the entropy temperature itself; the original SDAC code expresses it as $\exp(\log \tau_{\text{ent}})$ because the log-temperature is the learned parameter, but the two forms are equivalent. At inference, N candidates are sampled and the one with the highest $\min(Q_{\phi_1}, Q_{\phi_2})$ is selected (best-of-N selection).

Flow matching policies Conditional flow matching (Lipman et al., 2023) offers an alternative generative framework that replaces the complex noise schedule of diffusion with straight-line interpolation between noise and data. The forward process defines a linear path

$$x_t = (1-t)\,a_0 + t\,\epsilon, \quad \epsilon \sim \mathcal{N}(0, I), \quad t \in [0,1] \quad (6)$$

where $t = 0$ corresponds to the clean action and $t = 1$ to pure noise. A velocity network v_θ is trained to predict the conditional velocity field $v^* = \epsilon - a_0$ that transports points along these linear paths,

$$\mathcal{L}_{\text{FM}}(\theta) = \mathbb{E}_{t\sim\mathcal{U}[0,1],\,\epsilon\sim\mathcal{N}(0,I)} \left[\left\| v_\theta(x_t, t, s) - (\epsilon - a_0) \right\|^2 \right] \quad (7)$$

Sampling proceeds via Euler ODE integration from $t = 1$ (noise) to $t = 0$ (action), with the update $x_{t-\Delta t} = x_t - v_\theta(x_t, t, s) \cdot \Delta t$. Flow matching produces simpler velocity fields and more consistent candidates via deterministic ODE integration.

Ensemble Exploration. Bootstrapped DQN (Osband et al., 2016) maintains K Q-network heads $\{Q_{\phi_k}\}_{k=1}^{K}$, each trained on a bootstrapped data subset obtained by independently including each transition with probability p (Bernoulli masking). Thompson Sampling selects one

head at the start of each episode and follows the greedy policy of throughout, producing *temporally coherent* exploration, the agent commits to one head's strategy for the full episode rather than switching strategies every step, which enables deep exploration of long-horizon plans that single-step methods cannot. SUNRISE (Lee et al., 2021) replaces per-episode commitment with per-step UCB selection The shift to per-step selection trades temporal coherence for finer-grained uncertainty exploitation at every state, which works better in continuous-control settings where episodes are long and value uncertainty varies meaningfully across nearby states. REDQ (Chen et al., 2021) introduces random subset minimization for Bellman targets, with the target using $\min_{i\in\mathcal{M}} Q_{\bar{\phi}_i}(s', a')$ where $\mathcal{M}$ is a random subset of size M from the K Q-networks, providing a tunable pessimism-optimism knob ($M = 2$ approximates twin-Q, $M = K$ is maximally pessimistic).

3. Methodology

We saw in Section 1 that the naive extension of ensemble exploration to generative policies fails: K flow or diffusion heads trained against a shared Q-network collapse to functionally identical policies regardless of bootstrap masking or initialization, because the Q-weighted denoising gradient is head-invariant in expectation when the same Q drives every head's importance weights. Prior work on ensemble exploration in deep RL (Osband et al., 2016; Lee et al., 2021; Chen et al., 2021) sidesteps this collapse only because Gaussian policies have a fundamentally different gradient structure: their parameters enter the loss directly through a reparametrized log-density, not indirectly through Q-weighted importance sampling. The transition to generative policies therefore requires a non-obvious architectural change.

ESAC makes this change. Each policy head is paired with its *own* independently initialized Q-network and trained with Q-weighted importance weights computed from that head's Q rather than a shared one, so different Q-landscapes drive different policies even on shared data. A circular twin-Q pairing recovers SAC-style pessimism without doubling critic parameter count, and a trajectory-integrated disagreement bonus (TIDE) drives the resulting ensemble to act on its uncertainty during environment interaction. Figure 1 summarizes the full pipeline; the rest of this section describes the architecture (§3.1) and TIDE (§3.2) in detail.

3.1. Independent Actor-Critic Pairs

The core architectural decision in ESAC is to maintain K independently-initialized flow policy heads $\{\pi_{\theta_k}\}_{k=1}^{K}$ paired with K independently-initialized Q-networks $\{Q_{\phi_k}\}_{k=1}^{K}$. Twin-Q pessimism is recovered through circular pairing where head k uses the index set $S_k = \{k,\ (k{+}1) \bmod K\}$ in its policy loss, so each Q_{ϕ_k} participates in exactly two pairings while keeping critic parameter count linear in K.

Each Q_{ϕ_k} is trained via Bellman backup with REDQ-style random subset minimization (Chen et al., 2021). At every gradient step we redraw a random size-M subset $\mathcal{M}_t \subseteq \{1, \ldots, K\}$ of the target Q-networks and use

$$y = r + \gamma \min_{i\in\mathcal{M}_t} Q_{\bar{\phi}_i}(s', a'), \quad a' \sim \pi_{\theta_0}(\cdot \mid s') \tag{8}$$

with the same scalar target y training all K Q-networks at that step. Sampling a' from a single anchor head decouples policy diversity that is carried by the per-head importance weights below from value-target stability. Entropy regularization is realized implicitly in the actor through the Q-weighted softmax loss in Eq. 9, equivalent to maximizing $\mathbb{E}_{a\sim\pi}[Q] + \tau_{\text{ent}}\,\mathcal{H}[\pi]$ under the softmax distribution over the candidate pool. This avoids evaluating $\log \pi_\theta$ on the generative policy. Auto-tuning of τ_{ent} uses the standard SAC (Haarnoja et al., 2018) dual objective applied to the analytic Gaussian entropy of the action-time exploration noise $\mathcal{N}(0, (0.1\,\tau_{\text{ent}})^2 I)$.

Per-head pairing is applied at the policy loss. Each head π_{θ_k} is trained with a Q-weighted denoising / flow-matching loss (Eq. 5, Eq. 7) using importance weights computed via its circular twin-Q pair,

$$w_{k,i} \;=\; \mathrm{softmax}_i\left(\frac{c \cdot \min_{l\in S_k} Q_{\phi_l}(s, \hat{a}_i)}{\tau_{\text{ent}}}\right) \tag{9}$$

where the softmax is over the N_{mc} Monte Carlo candidates $\{\hat{a}_i\}_{i=1}^{N_{\text{mc}}}$ drawn from π_{θ_k}. For any $j \neq k$ the index sets S_j and S_k overlap in at most one element, so each head's gradient depends on at least one Q-network the other's does not and the structural property formalized in Theorem 2. Since each Q_{ϕ_l} converges differently, the pairwise minima differ across heads, the weights $w_{k,i}$ differ across heads, and each policy is driven toward different regions of action space. The total policy loss is averaged across heads, $\mathcal{L}(\theta) = \frac{1}{K}\sum_{k=1}^{K} \mathcal{L}_k(\theta_k)$.

3.2. Trajectory Integrated Disagreement Exploration

UCB measures disagreement only at the final action, ignoring the iterative denoising structure. TIDE exploits this additional structure by measuring disagreement at multiple points along the denoising trajectory and using the integrated disagreement as the exploration bonus. Intuitively, when the K heads disagree throughout the generation process rather than only at the output, they represent genuinely distinct posterior samples and the associated state is a more informative one to visit. Concretely, for each candidate action a_j produced by head k and any denoising timestep $\tau \in \mathcal{T}_{\text{TIDE}} := \{0, T/4, T/2, 3T/4\}$, we evaluate all K velocity (or noise) networks on the noised intermediate $x_\tau^{(j)}$

and collect the prediction disagreement,

$$\Delta_\tau^{(j)} = \frac{1}{d_a}\sum_{d=1}^{d_a}\left[\frac{1}{K-1}\sum_{k=1}^{K}\Big(v_{\theta_k}(x_\tau^{(j)},\tau,s)_d - \bar{v}_\tau(x_\tau^{(j)},s)_d\Big)^2\right]^{1/2} \tag{10}$$

where $\bar{v}_\tau$ is the mean velocity across heads (for diffusion, v_{θ_k} is replaced by the noise prediction ϵ_{θ_k}). The trajectory-integrated disagreement is

$$\widehat{\Delta}_{\text{traj}}^{(j)} = \sqrt{\frac{1}{|\mathcal{T}_{\text{TIDE}}|}\sum_{\tau\in\mathcal{T}_{\text{TIDE}}}\left(\Delta_\tau^{(j)}\right)^2}, \tag{11}$$

i.e. the root-mean-square disagreement over the sampled denoising timesteps for candidate j, used as the exploration bonus in action selection. This is the empirical estimator of the population quantity used in Proposition 1,

$$\Delta_{\text{traj}}^2 := \mathbb{E}_{t\sim\mathcal{T}}\Big[\operatorname{Var}_k f_{\theta_k}\Big(x_t^{(k)},t,s\Big)\Big], \tag{12}$$

where $x_t^{(k)}$ denotes the head-k-specific intermediate state at denoising time t. Since $(\Delta_\tau^{(j)})^2$ in Eq. (10) is already the per-timestep empirical variance across heads, $(\widehat{\Delta}_{\text{traj}}^{(j)})^2 = \frac{1}{|\mathcal{T}_{\text{TIDE}}|}\sum_\tau \operatorname{Var}_k f_{\theta_k}(x_\tau^{(j)},\tau,s)$ is an unbiased estimator of Δ_{traj}^2 when $x_\tau^{(j)}$ is drawn from the marginal of the reverse process. Candidates are then scored as

$$a_t^* = \arg\max_{a_j}\ \bar{Q}(s_t,a_j) + \beta(t)\cdot\widehat{\Delta}_{\text{traj}}^{(j)}, \tag{13}$$

where t denotes the environment step and a_t^* is the action executed at s_t. We use τ exclusively for the denoising timestep (Eq. 10) and t exclusively for the environment step (Eq. 13, Eq. 14) throughout this section.

For each environment step, scoring the M_c candidates per head requires evaluating all K velocity networks at each of the $|\mathcal{T}_{\text{TIDE}}| = 4$ timesteps on every noised intermediate, giving $|\mathcal{T}_{\text{TIDE}}|\cdot K\cdot(K\cdot M) = 4K^2M$ velocity forward passes for exploration scoring alone. Empirically, because each velocity network is a small MLP and the $4K^2M$ evaluations are batched into a single forward call, TIDE inference adds $\approx 35\%$ to per step wall clock time over best-of-N Q-selection at training time and is disabled at evaluation.

A key failure mode of any disagreement-based bonus is that in relatively unimodal environments (e.g. HalfCheetah), the K velocity networks may disagree simply due to parameter noise, even when the Q-landscape is flat. We provide empirical evidence that Ant and Humanoid possess relatively large multimodal Q-landscapes supporting multiple qualitatively distinct high-reward action strategies, while HalfCheetah exhibits a comparatively unimodal structure, via Q-value distribution analysis and behavioural clustering of independently trained agents across all three environments; We provide empirical evidence in Appendix E that Ant and Humanoid possess multimodal Q-landscapes supporting multiple qualitatively distinct high-reward action strategies, while HalfCheetah exhibits a comparatively unimodal structure.

$$\beta(t) = \beta_0\cdot\max(0,\rho_t), \quad \rho_t = \frac{\operatorname{Cov}(\Delta_{\text{traj}},\sigma_Q)}{\sigma(\Delta_{\text{traj}})\,\sigma(\sigma_Q)}. \tag{14}$$

When velocity disagreement and Q-uncertainty agree on where the agent is uncertain ($\rho_t \approx 1$), the full bonus is applied, and when they disagree ($\rho_t \le 0$) the bonus is suppressed and TIDE falls back to pure exploitation at that step.

4. Theoretical Analysis

We provide formal justification for three claims.

Theorem 1 (Collapse under Shared Backbone and Shared Twin-Q)**.** *Let K policy heads share a backbone ϕ and a single twin-Q pair, with per-head Bernoulli masks $m_{k,i}\sim$ Bernoulli(p). Then the expected gradient is head-invariant:*

$$\mathbb{E}_{m_k}[\nabla_{\eta_k}\mathcal{L}_k] = p\cdot\nabla_{\eta_k}\mathcal{L}^{(0)}(\theta_k;Q_\phi) \quad \forall\, k, \tag{15}$$

and the pairwise function-space distance $\|f_{\theta_i}(\cdot)-f_{\theta_j}(\cdot)\|^2$ contracts to zero at rate $O(\eta_t)$ under standard SGD. Bootstrap masking cannot prevent collapse since $\operatorname{Var}[\bar{m}_k] = p(1-p)/B$ is a factor $1/B$ smaller than the per-sample variant of Bootstrapped DQN (Osband et al., 2016).

Figure 4 in Appendix D confirms collapse empirically the off-diagonal action-cosine mean reaches exactly 1.000 by 200k steps under shared Q, while under independent Q it stabilizes near 0.68.

Lemma 1 (Non-identifiability of Heads under Shared Q)**.** *$\mathcal{L}^{(0)}$ is invariant to permutations of the head index, so the K-head model is non-identifiable up to head reparameterization. Function-space collapse therefore does not require parameter-space collapse; heads can retain distinct parameters that all implement the same function.*

Figure 2 confirms this that under shared Q function-space distance crashes to near zero by 200k steps while parameter-space distance remains substantially nonzero throughout. More detailed explanation on this is in Appendix D

Theorem 2 (Initial-Step Gradient Non-Collapse under Circular Twin-Q Pairing)**.** *Replace the shared twin-Q with K independently-initialized Q-networks $Q_{\phi_k}\sim\mathcal{N}(0,\sigma_{\text{init}}^2 I)$ and use circular twin-Q pairing $S_k = \{k,(k+1)\bmod K\}$ for per-head importance weights. For any $j\neq k$ the index sets S_k and S_j overlap in at most one element, so each*

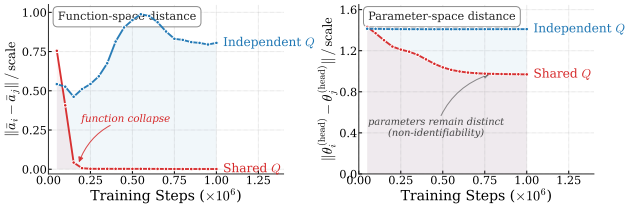

Figure 2. Empirical confirmation of Lemma 1 on Ant-v4 with K=5. Each panel reports a mean pairwise inter-head L2 distance, normalized by the mean L2 magnitude of a single head's output .Curves show shared Q (red) and independent Q (blue); annotations highlight the collapse and non-identifiability regimes.

head's gradient depends on at least one Q-network the other's does not. Under Assumptions 1–2,

$$\mathbb{E}\left\|\nabla_{\eta_k}\mathcal{L}_k - \nabla_{\eta_j}\mathcal{L}_j\right\|^2 \;\geq\; c\cdot\frac{\sigma_{\text{init}}^2}{K} \quad \textit{for all } j\neq k, \tag{16}$$

breaking the head-invariance that drives collapse in Theorem 1.

Proposition 1 (TIDE as a Tighter Epistemic Uncertainty Proxy)**.** *Let* $\Delta_0 := \mathrm{std}_k\, f_{\theta_k}(x_0, 0, s)$ *be terminal disagreement and* $\Delta_{\text{traj}}^2 := \mathbb{E}_{t\sim\mathcal{T}}[\mathrm{Var}_k\, f_{\theta_k}(x_t^{(k)}, t, s)]$ *be trajectory-integrated disagreement (Eq. 12). Under Assumption 1,*

$$\mathrm{Var}_k[a_k(s)] \;\leq\; C_{L_x}^2\cdot\Big(\sum_{t\in\mathcal{T}} L_g(t)\Big)^2\cdot\Delta_{\text{traj}}^2 + O(\Delta_0^2), \tag{17}$$

so inter-head action variance is bounded by trajectory-integrated disagreement alone, while Δ_0 *captures only the final denoising step. On Ant-v4,* Δ_{traj} *is up to* $3\times$ *larger than* Δ_0 *during mid-training, providing a richer signal exactly when the policy has begun to specialize but not yet committed to a single strategy.*

Full proofs, figures, and the Observation on finite-horizon diversity preservation are in Appendix D.

5. Results and Discussions

We evaluate ESAC on standard MuJoCo continuous control benchmarks, comparing against the SDAC baseline (Ma et al., 2025) and other baselines. Our central question is whether directed, posterior-driven exploration meaningfully outperforms the noise-based exploration used by every prior generative-policy RL method on tasks where exploration is the actual bottleneck. We therefore expect ESAC's gains to be largest on tasks like Ant-v4 and Humanoid-v4 with multimodal action distributions and tight termination conditions, and smallest on smooth landscapes like HalfCheetah where any reasonable exploration strategy suffices.

5.1. Setup

We evaluate on Mujoco environments HalfCheetah, Hopper, Walker2d, Ant, Pusher, Reacher, and Humanoid running all experiments for 1M environment steps except Humanoid which runs for 5M, with data collected through 5 vectorized environments. All experiments and training is carried out on NVIDIA RTX A6000 48GB VRAM and methods share the same hyperparameters, including a learning rate of 3×10^{-4} with linear decay to 3×10^{-5}, an α learning rate of 7×10^{-3}, $T = 20$ denoising or flow steps, batch size 256, a replay buffer of 10^6 transitions, $\tau = 0.005$, $\gamma = 0.99$, loss scaling probability $p = 0.8$, and REDQ $M = 2$. Full details are provided in Appendix F.

5.2. Results

Figure 3 shows the training learning curves across the eight MuJoCo locomotion tasks of our benchmark. We compare ESAC with the flow-matching and diffusion backbones (red and green) against the SDAC baseline (orange), DPMD (blue), and the SAC baseline (gray). ESAC with the flow-matching backbone matches or exceeds the strongest baseline on Ant-v4, Walker2d-v4, and Humanoid-v4, the three tasks where directed exploration matters most, and stays competitive on the comparatively unimodal tasks Hopper-v4 and HalfCheetah-v4. Table 1 reports final evaluation returns across all environments. ESAC (FM) achieves the best result on three of the five locomotion tasks shown in the figure, with Ant-v4 (5646 ± 1001), Walker2d-v4 (4305 ± 529), and Humanoid-v4 (6773 ± 510), and remains within $\sim15\%$ of the leader on the remaining two. Relative to the SDAC backbone (Ma et al., 2025), ESAC (FM) yields a $7.08\times$ improvement on Ant-v4, $1.36\times$ on Walker2d-v4, $1.18\times$ on Humanoid-v4, and $1.12\times$ on Hopper-v4.

The improvement scales monotonically with the exploration difficulty of the task. Ensemble diversity helps most when the Q-landscape is multimodal and high-dimensional, and yields diminishing returns when a single mode suffices. The signature is visible in the saturation pattern on HalfCheetah-v4, where the four SDAC-family methods all converge to within $\sim14\%$ of each other (10,328–11,949), while the gap between the best and worst diffusion-policy method on Ant exceeds $58\times$ (95 for DPPO vs 5646 for ESAC FM). Against the strongest single competing baseline per environment, ESAC (FM) outperforms DPMD (Ma et al., 2025) on Ant (5646 vs 5520), Walker2d (4305 vs 4280), and Humanoid (6773 vs 6710); DPMD edges out on Hopper-v4 (3220 vs 2759) and the original SDAC retains a slim HalfCheetah-v4 lead. On Humanoid-v4, ESAC (FM) also matches the variance profile of the strongest stable diffusion-policy baselines its seed standard deviation (±510) is an order of magnitude tighter than the deeper diffusion baselines that exhibit seed-level training collapses on 17-DoF control (DACER ±3210, SAC ±2415) and the parameter matched comparison in Appendix H shows SDAC and SAC scaled to ESAC's budget still trail ESAC (FM) by roughly 1000 return points on Ant-v4, confirming that the gains stem from ensemble structure

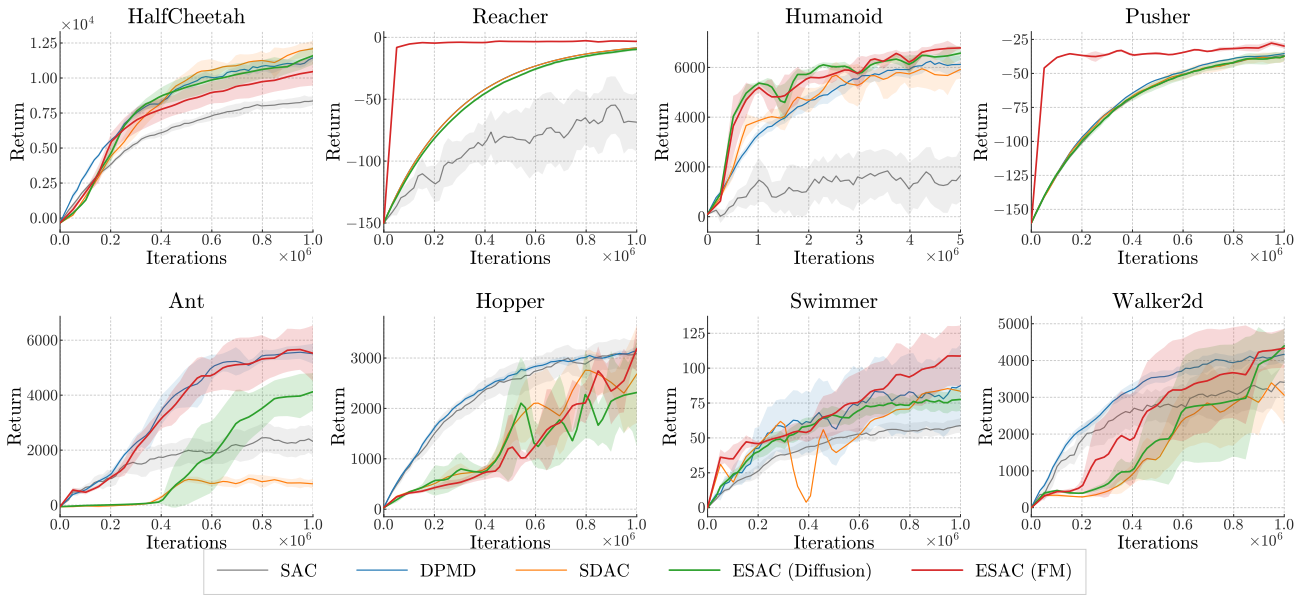

Figure 3. Training learning curves on five MuJoCo locomotion tasks trained for 1M steps (200k iterations across 5 vectorized environments) for top 5 baselines

rather than added capacity or candidate count.

6. Related Work

Generative policies for RL. Diffusion Policy (Chi et al., 2023) pioneered diffusion models for behavioral cloning in robotics. For online RL, SDAC (Ma et al., 2025) introduced Q-weighted score matching for diffusion actors, QVPO (Ding et al., 2024) combined Q-weighted diffusion with variational policy optimization, and Q-score matching (Psenka et al., 2024) connected the policy score to the Q-function gradient. FPO (McAllister et al., 2025) introduced flow matching (Lipman et al., 2023; Liu et al., 2023) for policy gradients. In the offline setting, diffusion policies have been applied via implicit Q-learning (Hansen-Estruch et al., 2024), conditional generation (Wang et al., 2023), and out-of-distribution generalization (Ada et al., 2024), while policy-guided diffusion (Jackson et al., 2024) fine-tunes pre-trained diffusion policies via RL objectives. More recently, LSAC (Ishfaq et al., 2025) approximates Thompson sampling in continuous control through Langevin Monte Carlo updates on a distributional critic, side-stepping the cost of an explicit Q-ensemble while still relying on a Gaussian actor. All of these methods, both online and offline, use a single generative policy with noise-based exploration. ESAC is the first to extend ensemble exploration to generative policy classes, and the first to identify and resolve the shared-Q collapse mode that breaks the naive ensemble extension.

Ensemble methods in RL. The use of ensembles for exploration traces back to Thompson Sampling (Thompson, 1933; Russo et al., 2018). Bootstrapped DQN (Osband et al., 2016) and randomized prior functions (Osband et al., 2018) introduced multi-head Q-networks with Thompson Sampling for deep exploration in discrete domains. REDQ (Chen et al., 2021) scaled Q-ensembles to continuous control with high update-to-data ratios. SUNRISE (Lee et al., 2021) combined ensemble disagreement with UCB-style optimism (Auer et al., 2002). OAC (Ciosek et al., 2019) used approximate upper confidence bounds for Gaussian policies. EDAC (An et al., 2021) penalized Q-function uncertainty for offline RL, and Dropout Q-functions (Hiraoka et al., 2022) provided a lightweight alternative to full ensembles. TEEN (Li et al., 2023) explicitly promoted trajectory diversity across ensemble members. Theoretical foundations include tighter Bayesian regret bounds (Moradipari et al., 2023) and approximate posterior sampling (Ishfaq et al., 2024). All of these methods operate on Gaussian or otherwise unimodal policy classes, and our results (Theorem 1) show that their default shared-Q designs do not transfer to generative policies without the architectural change ESAC introduces. We also note a surface resemblance between ESAC's ensemble-mean aggregation and value-decomposition methods such as VDN (Sunehag et al., 2018) and QMIX (Rashid et al., 2020), where the total Q value factorizes a joint Q-function across cooperative agents with private observations. The two are structurally distinct, ESAC's K Q-networks are independent estimates of the *same* single-agent Q-function on the *same* observation, summarized by an ensemble mean, whereas VDN's Q_i are pieces of a joint Q-function across *different* agents. The shared shape (a linear combination of K Q-values) addresses orthogonal problems.

Exploration in continuous control. Beyond ensembles,

Table 1. Final evaluation return on MuJoCo locomotion tasks, mean ± std over 5 seeds. **Bold** marks methods within one standard deviation of the highest mean per environment.

	Method	HALFCHEETAH	REACHER	HUMANOID	PUSHER
Classic Model-Free RL	PPO (Schulman et al., 2017)	4715 ± 810	-7.56 ± 10.68	1015 ± 280	-27.52 ± 3.40
	TD3 (Fujimoto et al., 2018)	8345 ± 645	$\mathbf{-3.19 \pm 0.04}$	5705 ± 405	$\mathbf{-25.23 \pm 0.96}$
	SAC (Haarnoja et al., 2018)	8745 ± 415	-61.58 ± 58.11	2945 ± 2415	-34.28 ± 1.16
Diffusion Policy RL	QSM (Psenka et al., 2024)	10495 ± 510	-4.45 ± 0.45	5520 ± 480	-83.18 ± 2.20
	DIPO (Yang et al., 2023)	9285 ± 720	$\mathbf{-3.28 \pm 0.03}$	4790 ± 985	-32.11 ± 0.44
	DACER (Wang et al., 2024)	11055 ± 285	$\mathbf{-3.12 \pm 0.07}$	2880 ± 3210	-31.02 ± 0.18
	QVPO (Ding et al., 2024)	7195 ± 1015	-35.89 ± 13.25	445 ± 95	-134.21 ± 1.08
	DPPO (Ren et al., 2024)	1245 ± 415	-5.62 ± 1.67	510 ± 75	-95.45 ± 11.87
	DPMD (Ma et al., 2025)	$\mathbf{11750 \pm 670}$	$\mathbf{-3.18 \pm 0.08}$	$\mathbf{6710 \pm 525}$	-30.18 ± 0.33
	SDAC (Ma et al., 2025)	$\mathbf{11949 \pm 659}$	-3.43 ± 0.41	5766 ± 386	-32.28 ± 4.98
Ours	ESAC (Diff)	$\mathbf{11310 \pm 728}$	-3.84 ± 1.96	$\mathbf{6493 \pm 672}$	-31.52 ± 5.88
	ESAC (FM)	10328 ± 1119	$\mathbf{-3.03 \pm 0.87}$	$\mathbf{6773 \pm 510}$	-28.90 ± 4.53

	Method	ANT	HOPPER	SWIMMER	WALKER2D
Classic Model-Free RL	PPO (Schulman et al., 2017)	3320 ± 945	$\mathbf{3155 \pm 220}$	$\mathbf{84.5 \pm 12.4}$	$\mathbf{4220 \pm 740}$
	TD3 (Fujimoto et al., 2018)	3690 ± 1180	1855 ± 985	71.9 ± 15.3	2640 ± 1190
	SAC (Haarnoja et al., 2018)	2480 ± 820	$\mathbf{3110 \pm 345}$	63.5 ± 10.2	3110 ± 950
Diffusion Policy RL	QSM (Psenka et al., 2024)	870 ± 195	2715 ± 510	57.0 ± 7.7	2480 ± 805
	DIPO (Yang et al., 2023)	1020 ± 145	1245 ± 690	46.7 ± 2.9	2050 ± 1380
	DACER (Wang et al., 2024)	4080 ± 615	$\mathbf{3145 \pm 145}$	$\mathbf{103.0 \pm 45.8}$	3275 ± 1640
	QVPO (Ding et al., 2024)	745 ± 290	2810 ± 555	53.4 ± 5.0	2415 ± 1095
	DPPO (Ren et al., 2024)	95 ± 35	2245 ± 580	$\mathbf{106.1 \pm 6.5}$	1085 ± 720
	DPMD (Ma et al., 2025)	$\mathbf{5520 \pm 245}$	$\mathbf{3220 \pm 85}$	84.33 ± 34.56	$\mathbf{4280 \pm 320}$
	SDAC (Ma et al., 2025)	798 ± 190	2472 ± 993	83.84 ± 16.9	3176 ± 782
Ours	ESAC (Diff)	4043 ± 617	2284 ± 1147	77.05 ± 16.04	$\mathbf{4172 \pm 483}$
	ESAC (FM)	$\mathbf{5646 \pm 1001}$	2759 ± 437	$\mathbf{107.83 \pm 23.82}$	$\mathbf{4305 \pm 529}$

exploration has been addressed via curiosity-driven intrinsic motivation (Pathak et al., 2017), random network distillation (Burda et al., 2019), count-based bonuses (Bellemare et al., 2016), and maximum entropy objectives (Haarnoja et al., 2018). These methods target a different exploration regime they augment the reward function with a state-novelty signal that depends on visitation history, encoding-prediction error, or policy entropy none of which use the structure of an ensemble's posterior disagreement to direct exploration. They are complementary to ESAC rather than competing alternatives, and could in principle be added on top of our ensemble framework as additional reward shaping. We do not benchmark against them here because they address state-level novelty (where in $\mathcal{S}$ the agent has not been) rather than action-level posterior diversity (which actions in $\mathcal{A}$ the agent is uncertain about), which is the regime ESAC is designed for. We treat their integration as future work.

7. Conclusion

We explored whether ensemble exploration, a proven technique for simple Gaussian policies, transfers to the increasingly popular family of generative RL policies. Our central finding is that it does not, at least not without independent Q-targets. The Q-weighted denoising loss that trains generative policies creates a gradient alignment failure mode where shared Q-functions drive all ensemble heads to identical policies, regardless of bootstrap masking or architectural independence. With independent actor-critic pairs and our new TIDE exploration rule, ESAC delivers substantial gains that scale with task difficulty, especially on tasks that demand exploration the most. The gains are regime-dependent: Appendix I shows ESAC plateaus on unimodal manipulation skills where coverage helps less than sharpness, pointing to multi-solution manipulation and adaptive K as natural future work, alongside extending ESAC to offline pretraining and offline-to-online finetuning where Q-uncertainty behaves differently. One limitation is that ESAC trades a $3.4\times$ parameter and $\sim$35% wall-clock overhead for these gains.

References

Ada, S. E., Oztop, E., and Ugur, E. Diffusion policies for out-of-distribution generalization in offline reinforcement learning. *IEEE Robotics and Automation Letters*, 9(4):3116–3123, 2024.

An, G., Moon, S., Kim, J.-H., and Song, H. O. Uncertainty-based offline reinforcement learning with diversified Q-ensemble. In *Advances in Neural Information Processing Systems*, 2021.

Auer, P., Cesa-Bianchi, N., and Fischer, P. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2-3): 235–256, 2002.

Bellemare, M., Srinivasan, S., Ostrovski, G., Schaul, T., Saxton, D., and Munos, R. Unifying count-based exploration and intrinsic motivation. In *Advances in Neural Information Processing Systems*, 2016.

Bradbury, J., Frostig, R., Hawkins, P., Johnson, M. J., Leary, C., Maclaurin, D., Necula, G., Paszke, A., VanderPlas, J., Wanderman-Milne, S., and Zhang, Q. JAX: composable transformations of Python+NumPy programs, 2018. URL http://github.com/google/jax.

Burda, Y., Edwards, H., Storkey, A., and Klimov, O. Exploration by random network distillation. In *International Conference on Learning Representations*, 2019.

Chen, X., Wang, C., Zhou, Z., and Ross, K. Randomized ensembled double Q-learning: Learning fast without a model. In *International Conference on Learning Representations*, 2021.

Chi, C., Feng, S., Du, Y., Xu, Z., Cousineau, E., Burchfiel, B., and Song, S. Diffusion policy: Visuomotor policy learning via action diffusion. In *Robotics: Science and Systems*, 2023.

Ciosek, K., Vuong, Q., Loftin, R., and Hofmann, K. Better exploration with optimistic actor-critic. In *Advances in Neural Information Processing Systems*, 2019.

Ding, S., Zheng, K., Zhang, C., Zhao, J., Shi, Y., and Chen, C. Diffusion-based reinforcement learning via Q-weighted variational policy optimization. In *Advances in Neural Information Processing Systems*, 2024.

Fujimoto, S., van Hoof, H., and Meger, D. Addressing function approximation error in actor-critic methods. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, 2018.

Haarnoja, T., Zhou, A., Abbeel, P., and Levine, S. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International Conference on Machine Learning*, 2018.

Hansen-Estruch, P., Kostrikov, I., Janner, M., Kuba, J. G., and Levine, S. IDQL: Implicit Q-learning as an actor-critic method with diffusion policies. In *International Conference on Machine Learning*, 2024.

Hiraoka, T., Imagawa, T., Hashimoto, T., Onishi, T., and Tsuruoka, Y. Dropout Q-functions for doubly efficient reinforcement learning. In *International Conference on Learning Representations*, 2022.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, 2020.

Ishfaq, H., Tan, Y., Yang, Y., Lan, Q., Lu, J., Mahmood, A. R., Precup, D., and Xu, P. More efficient randomized exploration for reinforcement learning via approximate sampling. *Reinforcement Learning Journal*, 1:1558–1595, 2024. URL https://rlj.cs.umass.edu/2024/papers/Paper148.html.

Ishfaq, H., Wang, G., Islam, S., and Precup, D. Langevin soft actor-critic: Efficient exploration through uncertainty-driven critic learning. In *International Conference on Learning Representations*, volume 2025, pp. 13758–13784, 2025.

Jackson, M. T., Matthews, M. T., Lu, C., Ellis, B., Whiteson, S., and Foerster, J. Policy-guided diffusion. *Reinforcement Learning Journal*, 2024.

Lee, K., Laskin, M., Srinivas, A., and Abbeel, P. SUNRISE: A simple unified framework for ensemble learning in deep reinforcement learning. In *International Conference on Machine Learning*, 2021.

Li, C., Gong, C., He, Q., and Hou, X. Keep various trajectories: Promoting exploration of ensemble policies in continuous control. In *Advances in Neural Information Processing Systems*, volume 36, pp. 5223–5235, 2023.

Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., and Wierstra, D. Continuous control with deep reinforcement learning. *International Conference on Learning Representations*, 2016.

Lipman, Y., Chen, R. T., Ben-Hamu, H., Nickel, M., and Le, M. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023.

Liu, X., Gong, C., and Liu, Q. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations*, 2023.

Ma, H., Yan, H., Hao, J., and Xu, Z. Efficient online reinforcement learning for diffusion policy. In *International Conference on Machine Learning*, 2025.

McAllister, D., Ge, S., Yi, B., Kim, C. M., Weber, E., Choi, H., Feng, H., and Kanazawa, A. Flow matching policy gradients. *arXiv preprint arXiv:2507.21053*, 2025.

Moradipari, A., Pedramfar, M., Lykouris, T., and Combes, R. Improved Bayesian regret bounds for Thompson sampling in reinforcement learning. In *Advances in Neural Information Processing Systems*, 2023.

Osband, I., Blundell, C., Pritzel, A., and Van Roy, B. Deep exploration via bootstrapped DQN. In *Advances in Neural Information Processing Systems*, 2016.

Osband, I., Aslanides, J., and Cassirer, A. Randomized prior functions for deep reinforcement learning. In *Advances in Neural Information Processing Systems*, 2018.

Pathak, D., Agrawal, P., Efros, A. A., and Darrell, T. Curiosity-driven exploration by self-supervised prediction. In *International Conference on Machine Learning*, 2017.

Psenka, M., Escontrela, A., Abbeel, P., and Ma, Y. Learning a diffusion model policy from rewards via Q-score matching. In *International Conference on Machine Learning*, 2024.

Rashid, T., Samvelyan, M., de Witt, C. S., Farquhar, G., Foerster, J., and Whiteson, S. Monotonic value function factorisation for deep multi-agent reinforcement learning. *Journal of Machine Learning Research*, 21(178):1–51, 2020.

Ren, A. Z., Lidard, J., Ankile, L. L., Simeonov, A., Agrawal, P., Majumdar, A., Burchfiel, B., Dai, H., and Simchowitz, M. Diffusion policy policy optimization. *arXiv preprint arXiv:2409.00588*, 2024.

Russo, D. J., Van Roy, B., Kazerouni, A., Osband, I., and Wen, Z. A tutorial on Thompson sampling. *Foundations and Trends in Machine Learning*, 11(1):1–96, 2018.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. *International Conference on Learning Representations*, 2021.

Sunehag, P., Lever, G., Gruslys, A., Czarnecki, W. M., Zambaldi, V., Jaderberg, M., Lanctot, M., Sonnerat, N., Leibo, J. Z., Tuyls, K., and Graepel, T. Value-decomposition networks for cooperative multi-agent learning based on team reward. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*, pp. 2085–2087, 2018.

Thompson, W. R. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3-4):285–294, 1933.

Wang, Y., Wang, L., Jiang, Y., Zou, W., Liu, T., Song, X., Wang, W., Xiao, L., Wu, J., Duan, J., and Li, S. E. Diffusion actor-critic with entropy regulator. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

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

Yang, L., Huang, Z., Lei, F., Zhong, Y., Yang, Y., Fang, C., Wen, S., Zhou, B., and Lin, Z. Policy representation via diffusion probability model for reinforcement learning. *arXiv preprint arXiv:2305.13122*, 2023.

Yu, T., Quillen, D., He, Z., Julian, R., Hausman, K., Finn, C., and Levine, S. Meta-World: A benchmark and evaluation for multi-task and meta reinforcement learning. In *Conference on Robot Learning*, 2020.

A. Proof of Theorem 1 (Collapse under Shared Backbone and Shared Twin-Q)

Proof. Recall the deployed Q-weighted loss for head k under the shared-backbone, shared-twin-Q regime,

$$\mathcal{L}_k(\theta_k; Q_\phi) \;=\; \bar{m}_k \cdot \sum_{i=1}^{B} w_i(Q_\phi) \cdot \ell_i(\theta_k), \qquad (18)$$

where $\bar{m}_k = \frac{1}{B}\sum_i m_{k,i}$ is the per-head Bernoulli-mask mean and $w_i(Q_\phi) = \mathrm{softmax}(\min(Q_{\phi_1}, Q_{\phi_2})(s_i, \hat{a}_i) \cdot c/\exp(\alpha))$ depends only on the shared twin-Q pair, hence is head-invariant. The per-sample loss is $\ell_i(\theta_k) = \|f_{\theta_k}(x_t^{(i)}, t, s_i) - \mathrm{target}_i\|^2$ with $f_{\theta_k} = h_k \circ \phi$ (shared backbone ϕ, head-specific linear h_k).

Part (i): Expected gradient direction is head-invariant. The mask is independent of the data and parameters, so

$$\mathbb{E}_{m_k}\big[\nabla_{\eta_k}\mathcal{L}_k\big] = \mathbb{E}[\bar{m}_k] \cdot \sum_i w_i(Q_\phi) \cdot \nabla_{\eta_k}\ell_i(\theta_k) \qquad (19)$$

$$= p \cdot \nabla_{\eta_k}\mathcal{L}^{(0)}(\theta_k; Q_\phi), \qquad (20)$$

where $\mathcal{L}^{(0)} = \sum_i w_i \ell_i$ is the unmasked Q-weighted loss. The right-hand side depends on θ_k but not on the head index k itself, and for any two heads i, j at the same parameter point the expected gradient is identical. The same reasoning applies to the shared-backbone gradient $\nabla_\xi \mathcal{L}_k$, where contributions from all K heads sum into the same trunk parameters with head-invariant magnitude.

Part (ii): Function-space collapse. Assume ℓ_i is L-smooth as a function of the network output $f_{\theta_k}(x, t, s)$. Define the per-head output function $F_k(\cdot) := f_{\theta_k}(\cdot)$ and the function-space discrepancy $d_t := \|F_i^{(t)}(\cdot) - F_j^{(t)}(\cdot)\|_\mu^2$, where $\|\cdot\|_\mu$ denotes the L^2 norm over the candidate-action distribution μ. Because the expected gradient $\mathbb{E}_{m_k}[\nabla_{\eta_k}\mathcal{L}_k] = p\,\nabla_{\eta_k}L^{(0)}(\theta_k; Q_\phi)$ is identical for all heads k (from Part (i)), and $L^{(0)}$ depends on θ_k only through the output F_k, the SGD iterates for F_i and F_j satisfy the same expected update rule. By L-smoothness applied in output space and the Polyak–Łojasiewicz (PL) condition on $L^{(0)}$ viewed as a functional of F_k (which holds without requiring a unique minimiser in parameter space, in contrast to strong convexity), we obtain

$$d_{t+1} \;\le\; (1 - 2\eta_t\, p\, \mu_{\mathrm{PL}})\, d_t + O(\eta_t^2), \qquad (21)$$

where $\mu_{\mathrm{PL}} > 0$ is the PL constant of $L^{(0)}$ in function space. Under standard step-size schedules $\sum_t \eta_t = \infty$, $\sum_t \eta_t^2 < \infty$, this gives $d_t \to 0$ at rate $O(\eta_t)$, i.e. $\|F_i^{(t)}(\cdot) - F_j^{(t)}(\cdot)\|_\mu^2 \to 0$.

Note: $L^{(0)}$ is permutation-invariant (Lemma 1), so its minimisers form a manifold rather than a unique point; the parameter-space distance $\|\eta_i - \eta_j\|$ need not vanish and empirically does not. The collapse is purely in function space: all heads converge to the same input–output map while potentially retaining distinct parameter vectors on the permutation-symmetry orbit.

Part (iii): Variance bound. The Bernoulli-mask variance of the deployed scalar mask is $\mathrm{Var}[\bar{m}_k] = \mathrm{Var}[\frac{1}{B}\sum_i m_{k,i}] = p(1-p)/B$. The per-sample mask used by Bootstrapped DQN (Osband et al., 2016) contributes $p(1-p)\sum_i w_i^2(\nabla\ell_i)^2$, which is a factor of B larger but still vanishes in expectation. Either form is dominated by the head-invariant mean signal of part (i) and cannot break the collapse.

Remark. Even with randomized prior functions (Osband et al., 2018) added to each head's output, the gradient of the *trainable* parameters still follows the shared Q-landscape, so priors alone cannot resolve the collapse. □

B. Proof of Theorem 2 (Initial-Step Gradient Non-Collapse under Circular Twin-Q Pairing)

Assumption 1 (Regularity)**.** *The MDP $(\mathcal{S}, \mathcal{A}, P, r, \gamma)$ has bounded rewards $|r(s,a)| \le R_{\max}$ and the Q-function class $\mathcal{F} = \{Q_\phi : \phi \in \Phi\}$ has bounded complexity (e.g., bounded Rademacher complexity).*

Assumption 2 (Approximation Diversity)**.** *There exist states (s,a) with positive function approximation uncertainty $\sigma^2_{approx}(s,a) := Var_{\mathcal{D}\sim P}[\hat{Q}(s,a;\mathcal{D})] > 0$, where $\hat{Q}(s,a;\mathcal{D})$ is the Q-function estimate trained on dataset $\mathcal{D}$.*

Proof. The argument is purely structural and holds at $t = 0$ under independent initialization. Write $S_k = \{k, (k+1) \bmod K\}$ for head k's pessimism index set, so the importance weights take the form $w_{k,i} \propto \exp(\beta\, q_{k,i})$ with $q_{k,i} := \min_{\ell \in S_k} Q_{\phi_\ell}(s, \hat{a}_i)$, and the per-head policy gradient is $\nabla_{\eta_k}\mathcal{L}_k = \sum_{i=1}^{M} w_{k,i}\, \nabla_{\eta_k}\ell_i(\eta_k)$ where ℓ_i is the per-sample denoising or velocity loss on candidate $\hat{a}_i$.

The geometry of the circular pairing is what drives the result. For any two distinct heads $j \ne k$ (with $K \ge 3$), the index sets S_k and S_j each consist of two consecutive integers modulo K, so they coincide only when $j \equiv k \pmod K$ and otherwise overlap in at most one element, as can be checked case by case for $j = k \pm 1, k \pm 2$. Consequently, at least one critic index $\ell^\star \in S_k \setminus S_j$ always exists, and head k's importance weights depend on $\phi_{\ell^\star}$ while head j's do not.

This single non-overlapping critic is sufficient to break gradient symmetry. Holding all ϕ_ℓ with $\ell \ne \ell^\star$ fixed and viewing $q_{k,i}$ as a function of $\phi_{\ell^\star}$ alone, independent Gaussian initialization $\phi_\ell \sim \mathcal{N}(0, \sigma^2_{\mathrm{init}} I)$ makes $Q_{\phi_{\ell^\star}}(s, \hat{a}_i)$ a ran-

dom variable with positive variance $\Theta(\sigma_{\text{init}}^2)$ (the implicit constant absorbing the layerwise gain). With probability strictly greater than $1/2$, this random Q-value is the active argument of the $\min$ over S_k, so $\partial q_{k,i}/\partial \phi_{\ell^\star}$ is non-zero on a set of positive measure. By contrast, $q_{j,i}$ has no dependence on $\phi_{\ell^\star}$ at all because $\ell^\star \notin S_j$, so head j's weights are deterministically constant in this direction.

It remains to translate this asymmetry on the importance weights into a lower bound on the gradient gap. Let $u_i := \nabla_{\eta_k} \ell_i(\eta_k)$ and $v_i := \nabla_{\eta_j} \ell_i(\eta_j)$ denote the head-k and head-j gradients of the same per-sample loss evaluated at their respective initial parameters. Assumption 2 guarantees $\|u_i\|, \|v_i\| \geq c_\ell > 0$ in expectation over the candidate distribution, and the independent initialization of η_k and η_j implies $\mathbb{E}[\langle u_i, v_i \rangle] = 0$ since the directions $u_i/\|u_i\|$ and $v_i/\|v_i\|$ are independent on the unit sphere. Expanding the squared gradient difference and dropping the cross term yields

$$\mathbb{E}\left\|\nabla_{\eta_k}\mathcal{L}_k - \nabla_{\eta_j}\mathcal{L}_j\right\|^2 \geq \mathbb{E}\left[\sum_i (w_{k,i} - w_{j,i})^2 \|u_i\|^2\right] + \mathbb{E}\left[\sum_i w_{j,i}^2 \|u_i - v_i\|^2\right]. \tag{22}$$

The first term inherits a lower bound from the asymmetry above. Along the direction $\phi_{\ell^\star}$, the weights $w_{k,i}$ vary through the softmax composed with $\min$ while $w_{j,i}$ does not, and the candidate Q-values are bounded by Assumption 1, so the softmax does not saturate. Hence there exists $c_w > 0$ depending only on the temperature β and on $R_{\max}$ such that $\mathbb{E}[\sum_i (w_{k,i} - w_{j,i})^2] \geq c_w \cdot \sigma_{\text{init}}^2/K$, the $1/K$ factor arising from the normalization over M candidates and the averaging over the random pair (j, k). Combined with the lower bound on $\|u_i\|$, this gives

$$\mathbb{E}\left\|\nabla_{\eta_k}\mathcal{L}_k - \nabla_{\eta_j}\mathcal{L}_j\right\|^2 \;\geq\; c \cdot \frac{\sigma_{\text{init}}^2}{K}, \qquad c = c_w\, c_\ell^2, \tag{23}$$

which is Eq. 16. The second term in the decomposition is non-negative and only tightens the bound, but is not needed to establish the claim. □

Remark 1 (Parameter count: K Q-networks vs. $2K$ independent pairs)**.** *Theorem 2 uses circular twin-Q pairing on K Q-networks rather than $2K$ fully independent twin-Q pairs. This keeps the parameter count linear in K at the cost of coupling adjacent heads through one shared Q-network. The structural lower bound is essentially unchanged: for any $j \neq k$ the pessimism index sets $S_k = \{k, k{+}1\}$ and $S_j = \{j, j{+}1\}$ overlap in at most one element, so each head's gradient depends on at least one Q-network the other's does not. The constant in Eq. 16 is only a factor $\sqrt{2}$ tighter than the $2K$-independent-pair case (where index sets do not overlap at all), which we confirm empirically: at $K = 5$ we observe ℓ_2-distance between the K Q-networks remaining at 3–5× the initialization noise floor throughout training (Figure 7, bottom row).*

Observation 1 (Finite-Horizon Preservation of Q-Network Diversity)**.** *The shared scalar TD target y_t in Eq. 8 acts as a homogenizing regularizer rather than a source of inter-network diversity: when all K critics regress toward the same target, the difference dynamics $(Q_{\phi_k}^{(t+1)} - Q_{\phi_j}^{(t+1)})$ contract toward zero in the limit of infinite training, with the unique fixed point being the Bellman optimum Q^* on the data-induced state-action distribution. The mechanism that preserves a non-vanishing inter-network gap over the training horizons used in our experiments is, in order of importance, (i) independent Gaussian initialization placing the K networks in distinct basins of attraction, (ii) independent Adam optimizer state (per-parameter momentum and second-moment buffers) which amplifies small initial differences over the first few thousand updates, and (iii) the circular twin-Q pairing of Theorem 2, which routes each head's policy gradient through a different pair of Q-networks. Empirically, the inter-network ℓ_2 distance starts at the initialization noise floor, expands to roughly 5–8× that floor in the early phase of training, and stabilizes at 3–5× the floor by 10^6 steps (Fig. 7, bottom row). We make this finite-horizon claim empirically rather than as a theorem; a rigorous time-uniform bound would require additional assumptions that we do not impose, and we leave a formal dynamical analysis to future work.*

C. Proof of Proposition 1 (TIDE Epistemic Uncertainty Bound)

Proof. We treat the reverse process as a discrete-time recurrence on the discretization grid $\mathcal{T} = \{t_0 = 1, t_1, \ldots, t_T = 0\}$ with $|\mathcal{T}| = T + 1$ points. For head k, let $x_{t_i}^{(k)}$ denote the state at step t_i. Both the flow matching ODE solver and the DDIM reverse step can be written as

$$x_{t_{i+1}}^{(k)} \;=\; g_{t_i}\big(x_{t_i}^{(k)},\; f_{\theta_k}(x_{t_i}^{(k)}, t_i, s)\big), \tag{24}$$

where g_{t_i} is the one-step reverse map and f_{θ_k} is the denoising-direction network of head k. For flow matching, $g_{t_i}(x, u) = x - (t_i - t_{i+1})\, u$ and $L_g(t_i) = |t_i - t_{i+1}|$ is simply the step size. For diffusion with the DDIM predictor the one-step reverse map is a smooth affine transformation in u whose Lipschitz constant in u depends on the noise schedule coefficients and is bounded on the unit interval.

By the Lipschitz assumption and the triangle inequality, for any two heads k, j and any step i,

$$D_{i+1}^{k,j} \;\leq L_x(t_i)\, D_i^{k,j} + L_g(t_i)\, F_i^{k,j},$$

where $D_i^{k,j} := \|x_{t_i}^{(k)} - x_{t_i}^{(j)}\|$, $F_i^{k,j} := \|f_{\theta_k}(x_{t_i}^{(k)}, t_i, s) -$

$f_{\theta_j}(x_{t_i}^{(j)}, t_i, s)\|$, and $L_x(t_i)$ is the Lipschitz constant of g_{t_i} in its first argument. Unrolling this recurrence from $i = 0$ (both heads share the same initial noise x_1, so the first term on the right vanishes) gives

$$\|x_0^{(k)} - x_0^{(j)}\| \;\le\; \sum_{i=0}^{T-1} \Big(\prod_{\ell=i+1}^{T-1} L_x(t_\ell)\Big) L_g(t_i) \cdot \|f_{\theta_k}(x_{t_i}^{(k)}, t_i, s) - f_{\theta_j}(x_{t_i}^{(j)}, t_i, s)\|. \tag{25}$$

Taking variance across the K heads on both sides, applying Cauchy–Schwarz, and using the triangle inequality to bound the cross-head disagreement at the *head-specific* intermediate states $x_{t_i}^{(k)}$ via the spatial Lipschitz constant $L_x(t_i)$ of g_{t_i} in its first argument yields a further error term proportional to $\|x_{t_i}^{(k)} - x_{t_i}^{(j)}\|$. Applying the discrete Grönwall inequality to control the accumulated state divergence $\|x_{t_i}^{(k)} - x_1\|$ (both heads share the same initial noise x_1, so the divergence starts at zero and grows through the per-step velocity disagreement), and using the conditional independence assumption, we obtain

$$\mathrm{Var}_k[a_k(s)] \;\le C_{L_x}^2 \Big(\sum_{t\in\mathcal{T}} L_g(t)\Big)^2 \cdot \mathbb{E}_{t\sim\mathcal{T}}\Big[\mathrm{Var}_k\, f_{\theta_k}\big(x_t^{(k)}, t, s\big)\Big] + O(\Delta_0^2), \tag{26}$$

where $C_{L_x} = \prod_i (1 + L_x(t_i))$ is the Grönwall amplification constant that absorbs the accumulated spatial Lipschitz factors, and Δ_0 captures the residual contribution of the final denoising step $t = 0$. The right-hand side of Eq. (26) contains the term $\mathbb{E}_{t\sim\mathcal{T}}\big[\mathrm{Var}_k\, f_{\theta_k}(x_t^{(k)}, t, s)\big]$, which equals Δ_{traj}^2 by definition (Eq. (12)). This substitution is an *equality*, not an application of Jensen's inequality: Δ_{traj}^2 is defined directly as the expected per-timestep variance, so no Jensen step is involved and the inequality direction is preserved throughout. The empirical estimator $(\widehat{\Delta}_{\text{traj}}^{(j)})^2$ used in action selection (Eq. (11)) concentrates around Δ_{traj}^2 by the law of large numbers over candidates and timesteps, so the bound of Eq. (17) applies to the population quantity that $\widehat{\Delta}_{\text{traj}}^{(j)}$ estimates. This yields Eq. (17) directly, with the Lipschitz constant $L_g(t)$ absorbing the Grönwall amplification constant C_{L_x}. The terminal-only bound is recovered by taking $\mathcal{T} = \{0\}$, which coincides with what a UCB-style terminal disagreement captures, and the trajectory-integrated bound of TIDE uses $|\mathcal{T}| = 4$ equally spaced points and is therefore tighter whenever the heads disagree more along the trajectory than at its endpoint. □

D. Why Shared-Q Networks Collapse?

The proof of Theorem 1 is given in Appendix A. It follows from the head-invariance of w_i under a shared Q together with a Bottou-style L-smooth SGD descent argument. Under shared Q the K heads are functionally one policy, so Thompson Sampling reduces to picking among K identical policies. Even with randomized prior functions (Osband et al., 2018) added to each head's output, the trainable-parameter gradient still follows the shared Q-landscape, so priors alone cannot resolve the collapse.

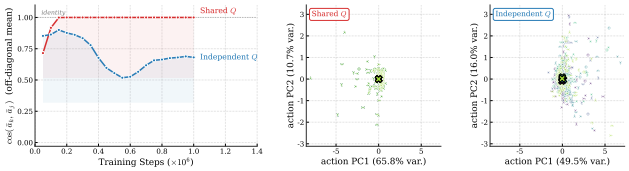

Figure 4. Empirical confirmation of Theorem 1 on Ant-v4 with K=5. *Left*, off-diagonal mean of the per-head action-cosine matrix $C_{kj} = \cos(\bar{a}_k, \bar{a}_j)$ over training; values approaching 1 indicate that all heads produce essentially the same action distribution (collapse), while values bounded away from 1 indicate preserved diversity. *Middle and right*, action-space PCA of $K \times N$ candidate actions sampled at the end of training, one panel per condition, with crosses marking per-head centroids; collapsed heads produce overlapping point clouds with overlapping centroids, while diverse heads occupy spatially separated regions of PC1–PC2 space.

The proof of Theorem 2 is in Appendix B. The formal claim is deliberately restricted to the initial step $t = 0$, which is what the structural argument supports without further assumptions on non-convex SGD dynamics. Whether the inter-head gap is preserved over the full training horizon is addressed empirically below.

Figure 4 provides the empirical confirmation alongside the formal claim. The cosine off-diagonal mean in the left panel is computed from the matrix $C_{kj} = \cos(\bar{a}_k, \bar{a}_j)$, where $\bar{a}_k$ is head k's mean action over a batch of probe states; the off-diagonal mean equals 1 exactly when all heads are functionally identical and 0 when their action distributions are orthogonal in expectation. Under shared Q this quantity rises from ≈ 0.7 at initialization to essentially 1.0 within 200k steps and stays there for the remainder of training, confirming the collapse predicted by Theorem 1. With independent twin-Q pairs the same quantity dips to ≈ 0.5 during the early-training phase and stabilizes around 0.7, indicating that the heads continue to disagree in expectation throughout training. The action-space PCA scatters in the middle and right panels make the geometric distinction visible: under shared Q all K head clouds collapse onto a single dense cluster with overlapping centroids, while under independent twin-Q each head occupies its own region of action space and the centroids spread out along PC1.

Figure 2 sharpens the analysis. The function-space panel (left) shows that under shared Q the policies the heads represent become indistinguishable within 200k training steps. The parameter-space panel (right) shows that the underlying parameter vectors remain $\Omega(1)$ apart in ℓ_2 throughout

training. Together the two panels demonstrate that shared-Q collapse is not a parameter-tying phenomenon but a non-identifiability of the function class under the shared critic: distinct parameter vectors representing identical policies. This is the empirical content of Lemma 1, and it explains why naive ensembling techniques such as randomized prior functions (Osband et al., 2018) cannot rescue the diversity of a shared-Q ensemble. Section 2 establishes that replacing the shared critic with K independent twin-Q pairs eliminates the non-identifiability and provides a structural lower bound on inter-head disagreement.

Remark 2 (Parameter count: K Q-networks vs. $2K$ independent pairs)**.** *Circular twin-Q pairing on K Q-networks keeps parameter count linear in K while preserving the structural lower bound: for any $j \neq k$ the pessimism index sets S_k and S_j overlap in at most one element, so the constant in Eq. 16 is only a factor $\sqrt{2}$ tighter than the $2K$-independent-pair case. Empirically, at $K = 5$ the ℓ_2-distance between Q-networks remains at 3–5$\times$ the initialization noise floor throughout training which can be analyzed through Fig. 7*

Observation 2 (Finite-Horizon Preservation of Q-Network Diversity)**.** *The shared scalar TD target y_t in Eq. 8 is a homogenizing regularizer whose unique fixed point is the Bellman optimum Q^*. The mechanisms that preserve a non-vanishing inter-network gap over the training horizons used in our experiments are, in order of importance: (i) independent Gaussian initialization placing the K networks in distinct basins of attraction, (ii) independent Adam optimizer state amplifying small initial differences over the first few thousand updates, and (iii) the circular twin-Q pairing of Theorem 2. Empirically, the inter-network ℓ_2 distance expands to 5–8$\times$ the initialization noise floor in the early phase and stabilizes at 3–5$\times$ that floor by 10^6 steps (Fig. 7, bottom row). We make this finite-horizon claim empirically rather than as a theorem; a rigorous time-uniform bound would require additional assumptions we do not impose.*

The proof of Proposition 1 is in Appendix C and uses a discrete Grönwall argument. The terminal-only UCB bound is recovered by taking $\mathcal{T} = \{0\}$; the TIDE bound with $|\mathcal{T}| = 4$ timesteps is strictly tighter whenever heads disagree more along the trajectory than at the endpoint.

E. Empirical Investigation of Policy Expressiveness

A central concern with any ensemble policy method is whether the K heads actually learn distinct behaviors, or whether they collapse to K functionally identical copies of a single policy. Theorem 1 predicts exactly such a collapse for shared-Q ensembles, and although the architectural fix of independent Q-networks (Theorem 2) breaks this collapse in theory, the empirical question of whether ESAC expresses genuinely multimodal behavior remains a separate test. We address it with two complementary experiments. The first directly observes the K heads converging to different gait modes on Swimmer, where the multimodal structure of the optimal policy is well documented and admits a clean frequency-domain analysis. The second examines the Q-landscape itself on the main locomotion benchmarks to characterize which environments admit multimodal value structure that ESAC can exploit.

E.1. Gait Spectrum Analysis on Swimmer

Swimmer-v4 is the canonical hard-exploration benchmark for off-policy actor-critic methods because its return surface admits multiple high-value gait modes that correspond to qualitatively different oscillation frequencies of the two body joints. A unimodal Gaussian or single-policy diffusion class can only commit to one such mode. To test whether our K=5 ensemble actually discovers the underlying multimodality rather than producing K near-copies of one policy, we trained both an SDAC single-policy baseline and ESAC with a flow-matching backbone for 1M environment steps at the same seed, deterministically rolled out each policy for 1,000 steps, and computed the discrete Fourier transform of the resulting action time series. The dominant peak of the power spectrum identifies the gait frequency that each policy has converged to.

Figure 5 reports the outcome. SDAC concentrates almost all of its action power around 0.65 Hz in panel (a), with a single dominant peak and an episode return of 106.58. The five ESAC heads in panel (b) fan out into two well-separated clusters. Heads 0 and 4 converge to a slow gait near 0.60 to 0.70 Hz that resembles the SDAC mode, while heads 1, 2, and 3 converge to a faster gait centred between 1.45 and 1.90 Hz. Panel (c) maps dominant gait frequency to episode return and shows that every ESAC head outperforms the SDAC baseline by 23 to 32 reward points, with the highest-return head sitting at the 1.9 Hz fast-gait mode that the single-policy baseline never reaches.

This experiment substantiates the central methodological claim of our work, namely that a K-headed actor paired with K independent critics expresses a genuinely multimodal exploration distribution rather than collapsing to a single policy. A single-policy class is empirically insufficient on Swimmer because its training trajectory commits to one gait mode and forgets the other, while ESAC recovers the missing fast-gait mode and improves returns at every individual head. The diagnostic generalizes beyond Swimmer.

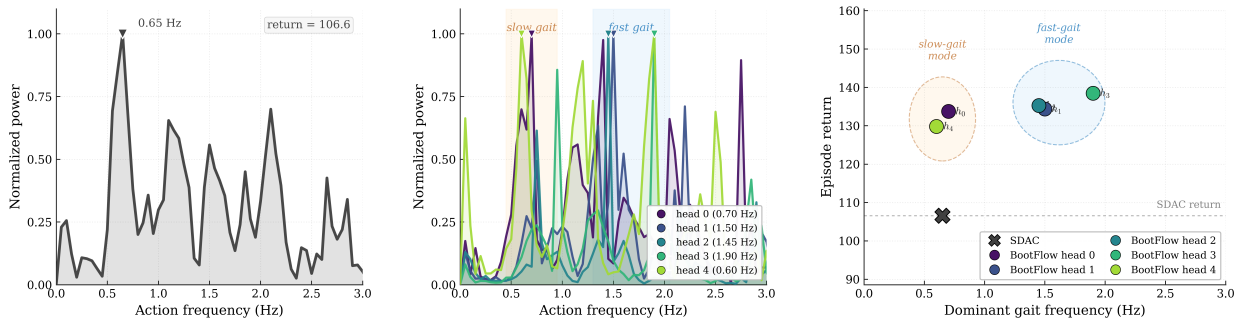

Figure 5. Gait spectrum analysis on Swimmer-v4 at 1M training steps with seed 100. Panel (a) shows the action power spectrum of an SDAC single-policy baseline, dominated by one slow oscillation near 0.65 Hz with episode return 106.58. Panel (b) overlays the action power spectra of the K=5 ESAC heads, which split into a slow cluster around 0.60 to 0.70 Hz and a fast cluster around 1.45 to 1.90 Hz. Panel (c) plots dominant gait frequency against episode return, where every ESAC head reaches 129.83 to 138.50 and the strongest head occupies the fast-gait mode that the single-policy baseline does not reach.

E.2. Q-Landscape Modality on Locomotion Benchmarks

The Swimmer gait analysis shows that ESAC expresses multimodal behavior where the value surface clearly admits it. A natural follow-up question is which of the main locomotion benchmarks used in our experiments also have multimodal Q-landscapes that ESAC can exploit. We examine this for HalfCheetah, Ant, and Humanoid by training N=5 independent SDAC agents, equivalent to ESAC with K=1, from distinct random seeds for 1,000,000 environment steps (Humanoid for 5,000,000) using the same hyperparameters as the main experiments. We then characterize the Q-landscape from two angles. First we evaluate the distribution of Q-values over 3,000 uniformly sampled random actions at ten representative reference states drawn from a trained agent's on-policy rollout. Second we project these actions to two dimensions with UMAP and color each point by its normalized Q-value to visualize spatial concentration of high-value action regions.

Figure 6 reveals an ordering of Q-landscape multimodality. The HalfCheetah Q-distribution concentrates around a single dominant peak with mild secondary structure and a long left tail, consistent with a comparatively unimodal value surface. Ant exhibits two clearly separated peaks of similar prominence, indicating two distinct high-value action regions. Humanoid is the most strongly bimodal of the three, with two prominent peaks and a deeper trough between them. The action-space UMAPs in the bottom row corroborate this ordering. HalfCheetah produces a near-uniform Q-coloring, with high-Q points scattered across the embedding rather than concentrated. Ant shows mild but visible concentration of high-Q points, and Humanoid shows the cleanest spatial concentration of high-Q regions into distinct pockets, the signature of a multimodal value surface.

F. Implementation Details

All experiments are implemented in JAX (Bradbury et al., 2018) using the dm-haiku library for neural network definitions. Q-networks are trained with the Adam optimizer. The entire training step (Q-loss, policy loss for all K heads, α-loss, target updates) is compiled as a single JIT-traced function for maximum throughput. More details on hyperparameter is given in Table 2.

G. Ablation Study

Table 3 isolates the contribution of each exploration strategy across all five environments and both backbones. TIDE matches or improves on both UCB and Thompson Sampling on Ant, Walker2d, and Hopper across both backbones; on Humanoid and HalfCheetah, UCB modestly edges out TIDE, where the Q-landscape is comparatively unimodal and Q-filtering alone suffices. The improvement is largest on Ant-v4, and the improvement of TIDE over UCB is largest with the flow-matching backbone on the higher-dimensional locomotion tasks (Ant and Walker2d), where the velocity-field disagreement signal is richest.

Two ablations isolate the source of the gain. First, swapping the diffusion backbone for flow matching while holding the rest of ESAC fixed yields a 1.40× improvement on Ant-v4 (5646 vs 4043) and consistent gains on Hopper and Humanoid, confirming that velocity-field heads support more diverse trajectory rollouts than denoising-step heads. The velocity field is integrated through ODE rollout into a fully formed action, exposing more inter-head disagreement than a single denoising step where heads only differ in the residual noise prediction. Second, candidate count alone is not the explanation. The $K = 1$, $N = 160$ ablation (a single-policy backbone with 5× more action proposals than vanilla SDAC) underperforms SDAC's 32-particle baseline, since best-of-N from a single mode collapses into overexploita-

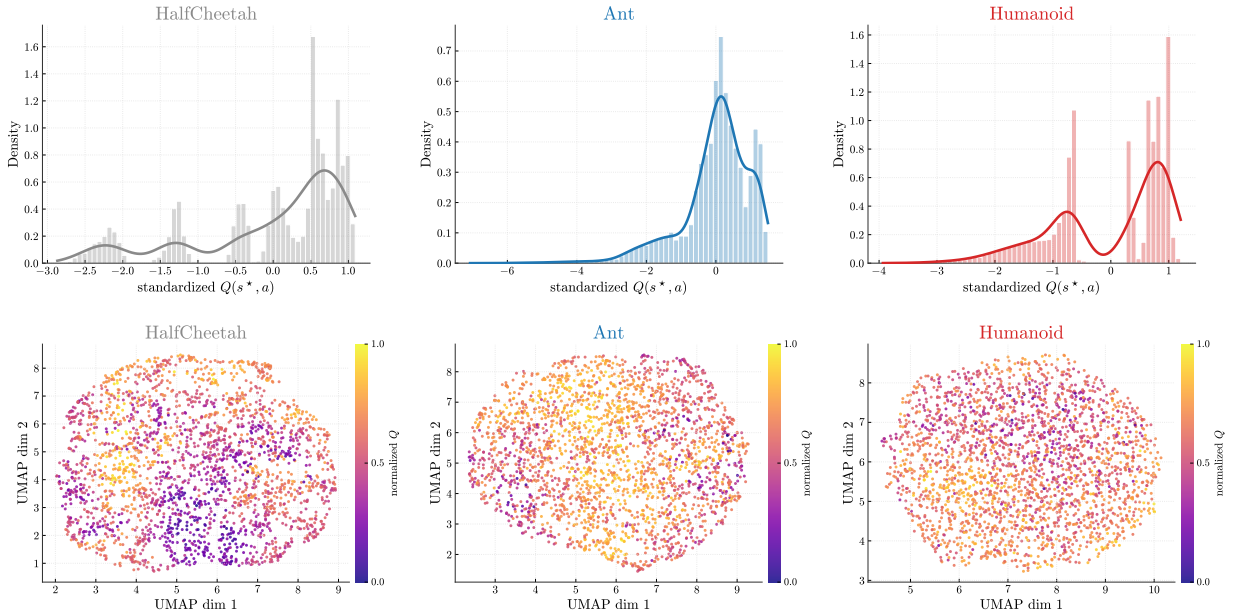

Figure 6. Q-landscape modality at 1M training steps (5M for Humanoid). Top row, Q-value distributions over uniformly random actions. Bottom row, UMAP of action space coloured by normalized Q-value. Columns correspond to HalfCheetah, Ant, and Humanoid environments.

tion of one local Q-maximum. Diversity must come from *population* structure, not from sampling resolution.

With shared twin-Q, Thompson ($K = 5$) matches SDAC ($K = 1$) exactly, confirming Theorem 1, and independent Q-targets raise Thompson to ~ 2400 and UCB to ~ 3300 on Ant-v4 (full numbers in Table 3). Across all configurations, UCB matches or outperforms Thompson on Humanoid and HalfCheetah, while TIDE matches or outperforms UCB on tasks with non-trivial exploration structure (Ant, Walker2d, Hopper). UCB uses all K heads at every step ($K \times N$ candidates with Q-based selection), while Thompson commits to one head per episode, leaving $K - 1$ heads unused. UCB's advantage grows when head diversity is limited (high p or shared Q), as it compensates via Q-filtering. TIDE adds the trajectory-integrated velocity disagreement on top of UCB, which on Ant-v4 moves the final return from 4357 (single seed, UCB with FM) to 5646 ± 1001 (five seeds, TIDE with FM). Thompson's advantage of temporally coherent multi-step strategies is less impactful on locomotion tasks where single-step action quality dominates.

Figure 7 reports two metrics over training. The top row shows per-head action disagreement, defined as the per-dimension standard deviation across K policy heads evaluated at denoising time $t = 0$. ESAC (FM) consistently maintains this disagreement at 3 to 5 times the level of ESAC (Diffusion) on every environment, confirming that flow matching supports a more diverse ensemble. The bottom row shows the inter-network ℓ_2 distance across the K independently initialized Q-networks. Consistent with Observation 2, both backbones maintain a non-vanishing inter-network gap at 3–5$\times$ the initialization noise floor throughout the 10^6-step training horizon, despite the shared scalar TD target, preventing the head-collapse failure mode that a shared Q-network would induce.

H. Capacity vs. Population Structure

ESAC with K=5 has roughly 3.4$\times$ the trainable parameters of the SDAC backbone at the default hidden width h=256, raising the natural question of whether the locomotion gains in Table 1 reflect ensemble structure or simply added capacity. The headline finding of this section is that ESAC's gain is structural rather than capacity-driven. A properly capacity-matched SDAC at 0.84M parameters trails ESAC by an order of magnitude on Ant-v4, and an SAC baseline scaled to 1.64M parameters, which is 13 percent *larger* than ESAC's 1.45M parameter budget, still trails ESAC (FM) by roughly 1000 return points. The population structure of K independent actor-critic pairs converts parameters into directed exploration in a way that no single-policy method at a comparable parameter budget can replicate.

The Q-network width axis is fragile in SDAC. Naive capacity matching by widening both the actor and critic of SDAC to h=512, including diffusion_hidden_dim, produces a catastrophic regression rather than a fair comparison. Widened SDAC drops from 798 ± 190 to 149 ± 48 on Ant-v4. To isolate the failure mechanism, we ran multi-seed ablations on Ant-v4 holding one component at h=256

Table 2. Complete hyperparameter specification for all experiments.

Hyperparameter	Value	Notes
Q-network architecture	MLP [256, 256, 256]	Mish activation
Policy network architecture	MLP [256, 256, 256]	Mish activation, sinusoidal time emb.
Learning rate (Q, policy init)	3×10^{-4}	Linear decay to 3×10^{-5}
Entropy coefficient lr (α)	7×10^{-3}	
Delayed update interval	250 steps	Policy and α update every 250 Q-updates
Discount factor γ	0.99	
Soft target update τ	0.005	
Replay buffer capacity	10^6 transitions	
Batch size B	256	
Number of heads K	5	$K = 1$ reduces to SDAC
REDQ subset size M	2	Deterministic pairing: head k with $(k + 1) \bmod K$
Scaling probability p	0.8	
Particles per head N	32	UCB pools $K \times N = 160$ total
Denoising / flow steps T	20	
UCB exploration bonus β	1.0	
TIDE base bonus β_0	2.0	Modulated online by $\max(0, \rho_t)$
TIDE timesteps $\mathcal{T}$	$\{0, T/4, T/2, 3T/4\}$	Four equally spaced denoising points
TIDE correlation window	4096 decisions	Sliding window for ρ_t
Exploration noise scale	0.1	Multiplied by $\exp(\alpha)$
Beta schedule (diffusion only)	Linear, scale 0.8	
Warmup steps (random policy)	30000	
Vectorized environments	5	Per-env Thompson head tracking
Monte Carlo samples (Q-weights)	64	For computing w_i in Eq. 5

while widening the other. Widening only the Q-networks to h=512 reproduces the collapse, while widening only the actor to h=512 preserves performance (≈750 return, within noise of the 798 baseline). The failure axis is the critic width. A wider Q-network exhibits higher output variance at initialisation, which sharpens the $\mathrm{softmax}(Q/\alpha)$ importance weights over the 64 Monte Carlo action candidates and concentrates policy gradients onto miscalibrated modes early in training. Self-distillation then amplifies this bias, locking the actor into a narrow suboptimal distribution. This finding is itself an architectural observation about the SDAC objective that is independent of ESAC, and it dictates the protocol for a fair scaled comparison: any capacity matching of SDAC must widen the actor only.

Capacity-matched comparison. Based on the diagnosis above, we scale SDAC by widening only the actor to h_π=512 while retaining h_Q=256, yielding ≈0.84M trainable parameters on Ant-v4 against ESAC's 1.45M at K=5. We also re-train SAC at full h=512 for both actor and critic (≈1.64M parameters), which over-scales SAC slightly past ESAC's parameter budget. Both scaled baselines use identical training budgets and hyperparameters to the main experiments, with 5 seeds per condition at 1M environment steps.

Table 4 reports the final evaluation returns. SDAC scaled to 0.84M parameters recovers substantially on Ant-v4 (719 ± 168) relative to the collapsed full-width variant (149±48), yet remains far below ESAC (FM) at 5646±1001 despite consuming the full budget that the SDAC objective can absorb without collapse. SAC scaled to 1.64M parameters nearly matches ESAC (Diff.) on Ant-v4 (4658 ± 940 vs 4043 ± 617) but trails ESAC (FM) by roughly 1000 return points while consuming 13 percent more parameters. The same ordering holds across Walker2d-v4 and Hopper-v4, and on HalfCheetah-v4 all four scaled and unscaled diffusion-family methods cluster within 14 percent of each other, consistent with the unimodality of that environment as discussed in Appendix E. ESAC's gain on the multimodal locomotion tasks is therefore not a parameter-count artefact. The ensemble structure converts parameters into a different functional resource (directed exploration over distinct policy modes) that single-policy methods cannot replicate at any tested scale.

I. Experiments on Manipulation Tasks

We evaluate ESAC on five MetaWorld v3 (Yu et al., 2020) manipulation tasks to characterize the regime in which ensemble-based exploration provides a measurable benefit. The headline finding aligns with the central thesis of the paper. ESAC's gains on locomotion arise specifically from multimodal Q-landscapes that reward coverage of distinct strategies, and tasks that admit only a single contact-rich solution sequence offer no such headroom. ESAC matches its SDAC backbone on three of five MetaWorld tasks, ties on one within seed noise, and incurs a measurable regression on `peg-insert-side-v3`. We present the experimental setup and results, then analyze the task-structural

Table 3. Ablation of exploration strategy across all five environments and both backbones. Thompson, UCB, and TIDE values are single-seed (seed 100) final returns at 1M training steps (5M for Humanoid). Bold indicates the best exploration strategy per env-backbone row.[†]

Environment	Backbone	Thompson	UCB	TIDE
Ant-v4	Diffusion	2412	3308	**4186**
	Flow Matching	4391	4357	**5733**
Walker2d-v4	Diffusion	3987	3697	**4555**
	Flow Matching	4120	2722	**4489**
Hopper-v4	Diffusion	2197	2696	**3263**
	Flow Matching	2200	2554	**2817**
Humanoid-v4	Diffusion	6044	**6631**	6493
	Flow Matching	5578	**6674**	6597
HalfCheetah-v4	Diffusion	12054	**12431**	12186
	Flow Matching	11816	**11717**	10328

Table 4. Parameter-scaled ablation on MuJoCo locomotion. SDAC scaled uses h_π=512 for the actor only (critic kept at h_Q=256) to avoid the collapse caused by widening the Q-network; SAC scaled uses h=512 for both actor and critic. Approximate parameter counts are for the Ant-v4 obs/act dimensions. Mean $\pm$ std over 5 seeds at 1M training steps. Base (h=256) and ESAC rows are reproduced from Table 1 for direct comparison. Bold marks the highest mean per environment.

Method	Ant-v4	Walker2d-v4	Hopper-v4	HalfCheetah-v4
SAC, h=256 (0.43M)	2480 ± 820	3110 ± 950	$\mathbf{3110 \pm 345}$	8745 ± 415
SAC scaled, h=512 (1.64M)	4658 ± 940	3772 ± 1290	2574 ± 1103	7340 ± 3097
SDAC, h=256 (0.43M)	798 ± 190	3176 ± 782	2472 ± 993	$\mathbf{11949 \pm 659}$
SDAC scaled, h_π=512 (0.84M)	719 ± 168	3275 ± 1180	2584 ± 716	11478 ± 392
ESAC (Diff.), K=5 (1.45M)	4043 ± 617	4172 ± 483	2284 ± 1147	11310 ± 728
ESAC (FM), K=5 (1.45M)	$\mathbf{5646 \pm 1001}$	$\mathbf{4305 \pm 529}$	2759 ± 437	10328 ± 1119

reasons for the observed plateau. The takeaway for practitioners is that the ensemble pays for itself when the task is known or suspected to admit multiple competing solutions, and a single-policy backbone is the safer default on tightly-constrained, densely-rewarded skills.

I.1. Setup and Results

We evaluate on five MetaWorld v3 tasks: `reach-v3`, `drawer-close-v3`, `door-open-v3`, `button-press-topdown-v3`, and `peg-insert-side-v3`. The first three are trained for 0.2M environment steps and the latter two for 0.5M steps, in line with the per-task training budgets typically reported on this benchmark. We compare four methods. SAC (Haarnoja et al., 2018) is the Gaussian-actor baseline. SDAC (Ma et al., 2025) is the single-policy (K=1) diffusion actor-critic that serves as the ESAC backbone. ESAC (Diff.) and ESAC (FM) are our K=5 ensembles with diffusion and flow-matching backbones respectively, both with TIDE exploration and β_0=2.0. All methods use the same MetaWorld auto-detected single-environment collection mode (no vectorisation), 32 action particles, and 5 seeds per (method, task). The reported success rate is the binary per-episode success indicator emitted by the MetaWorld API, averaged over the final 50 training episodes, with similar conclusions holding for 100- and 200-episode windows.

Table 5 reports the final success rates, Figure 8 visualizes the same numbers as a bar chart, and Figure 9 reports the training-time success-rate curves. The qualitative pattern across the five tasks divides cleanly. On `drawer-close-v3` and `button-press-topdown-v3` every method including SAC reaches 100 percent within the training budget, leaving no headroom for any algorithm to differentiate. On `door-open-v3` both SDAC (98.8 ± 2.4) and the two ESAC variants (94.4 ± 11.2 for diffusion, 98.0 ± 4.5 for flow matching) reach near-perfect success while SAC (25.2 ± 38.7) collapses on four of five seeds, confirming that the diffusion prior is the dominant contributor on this task and that the additional ensemble structure adds nothing measurable. On `reach-v3` all three diffusion-based methods cluster around 80 percent within seed noise.

The exception is `peg-insert-side-v3`, the hardest task in the suite and the only one where ESAC underperforms its backbone. SDAC reaches 99.6 ± 0.8 percent, while ESAC (Diff.) reaches 87.6 ± 24.8 percent and ESAC (FM) reaches 88.0 ± 17.9 percent. The gap is 12 percentage points with roughly 25 times larger seed variance, and one ESAC seed bottoms out at 38 percent, dragging down the mean. SAC, which lacks the diffusion prior entirely, reaches only

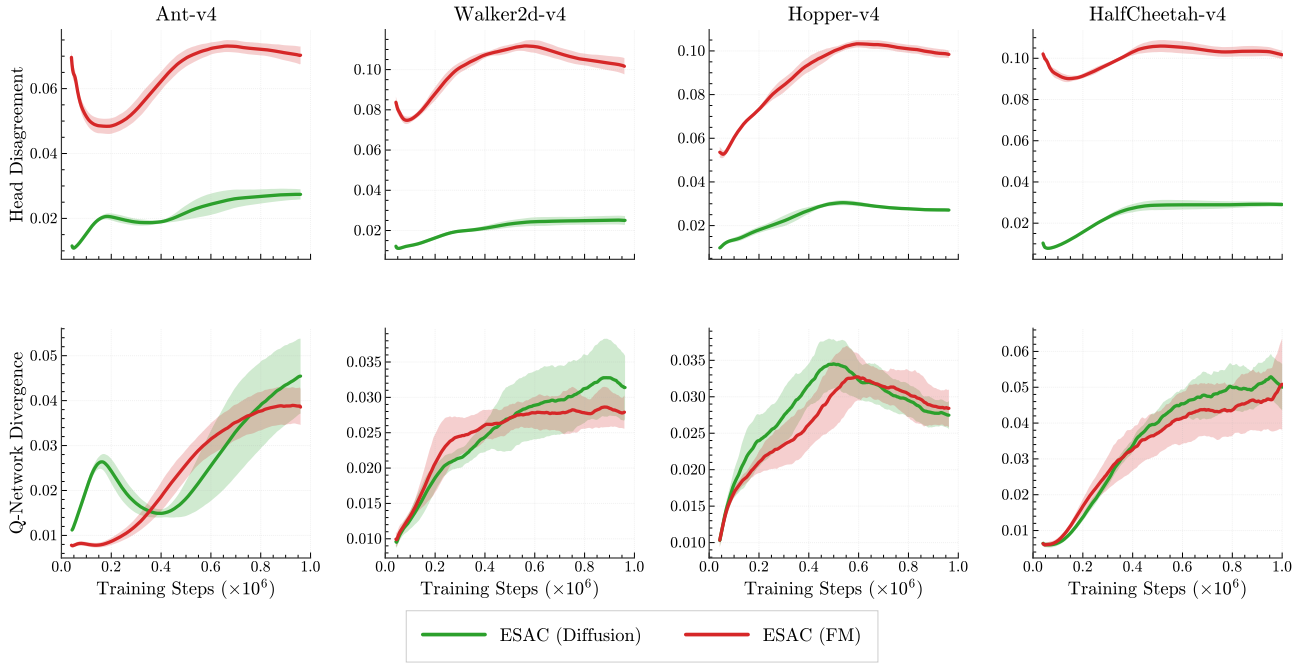

Figure 7. Ensemble metrics over training. Top row, per-head action disagreement at the final denoising step. Bottom row, Q-network divergence across the K independent Q-networks.

Table 5. Final success rate (%) on MetaWorld v3 manipulation tasks, mean $\pm$ standard deviation across 5 seeds, computed over the last 50 training episodes. **Bold** marks the highest mean per task.

Task	SAC	SDAC	ESAC (Diff.)	ESAC (FM)
`reach-v3`	78.8 ± 39.5	79.6 ± 39.8	$\mathbf{80.8 \pm 36.4}$	71.0 ± 40.8
`drawer-close-v3`	$\mathbf{100.0 \pm 0.0}$	99.2 ± 1.0	$\mathbf{100.0 \pm 0.0}$	$\mathbf{100.0 \pm 0.0}$
`door-open-v3`	25.2 ± 38.7	$\mathbf{98.8 \pm 2.4}$	94.4 ± 11.2	98.0 ± 4.5
`button-press-topdown-v3`	$\mathbf{100.0 \pm 0.0}$	$\mathbf{100.0 \pm 0.0}$	$\mathbf{100.0 \pm 0.0}$	$\mathbf{100.0 \pm 0.0}$
`peg-insert-side-v3`	60.0 ± 49.0	$\mathbf{99.6 \pm 0.8}$	87.6 ± 24.8	88.0 ± 17.9

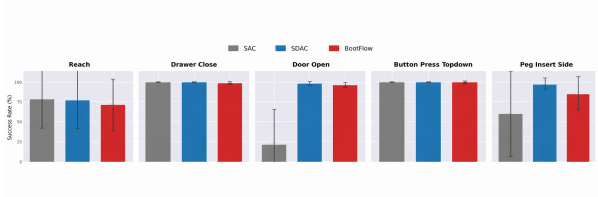

Figure 8. Final success rate on five MetaWorld v3 manipulation tasks, mean $\pm$ std across 5 seeds.

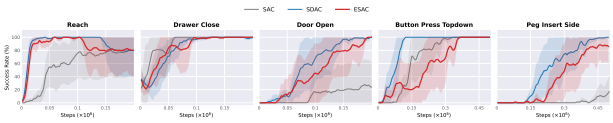

Figure 9. Training-time success rate on the five MetaWorld v3 tasks, rolling mean over 20 episodes with shaded ± 1 std bands across 5 seeds.

60.0 ± 49.0 percent.

The qualitative pattern across the five tasks rules out a tuning artefact and points to a structural mismatch between ESAC's exploration mechanism and the MetaWorld benchmark. Locomotion benchmarks contain multiple competing gait modes such as galloping versus trotting or three-leg versus four-leg support, and ensemble heads can converge to genuinely different solutions in those landscapes. MetaWorld v3 tasks reward a single contact-rich trajectory class consisting of reaching, grasping, aligning, and finally either pushing or inserting, and with dense shaped rewards pointing strongly toward this single trajectory at every step, the induced Q-landscape is closer to convex than multimodal. An ensemble of $K{=}5$ heads in this regime converges to slight perturbations of the same policy rather than discovering distinct strategies, which is the manipulation-side analogue of the HalfCheetah result in our main locomotion experiments where the four SDAC-family methods all converge to within 14 percent of each other because no method can extract gains from a unimodal landscape.

This unimodality interacts with benchmark saturation to amplify the apparent plateau. Three of the five tasks reach 99 percent or above for SDAC alone, namely `drawer-close-v3`, `button-press-topdown-v3`, and `door-open-v3`, leaving no room for any method to improve on a saturated ceiling. The remaining two tasks, `reach-v3` and `peg-insert-side-v3`, are the only real sources of differentiation, and one of them is dominated by seed noise rather than method differences. A benchmark in which most tasks are saturated by the strongest baseline cannot resolve method-level differences regardless of the underlying algorithmic merit, so even if ESAC's ensemble structure were providing a marginal benefit on these tasks, the measurement instrument would be unable to detect it.

I.2. Why ensemble exploration plateaus on manipulation

The peg-insertion regression is the only task where the gap is large enough to be interpreted as a real algorithmic effect rather than measurement noise, and the explanation comes from the same axis that drives the locomotion gains in the opposite direction. Peg-insertion requires a tight, precisely-timed grasp, align, and insert sequence in which sharpness of policy execution dominates breadth of policy coverage. A single-policy actor like SDAC can concentrate all of its representational capacity on that sequence and reach 99.6 percent. A $K{=}5$ ensemble distributes capacity across modes that do not exist on this task, and the 12 percentage-point drop is the direct cost of ensemble fragmentation on a sharpness-dominated objective rather than a failure of the underlying SDAC backbone. This is the dual of the Ant-v4 result, where a multimodal Q-landscape rewards exactly the same coverage that hurts here.

The MetaWorld results are therefore not a failure mode of ESAC but a regime characterization. Exploration-by-ensemble methods are most useful on tasks whose Q-landscape contains multiple high-value basins of attraction, exactly the regime in which a single-policy actor risks committing prematurely to a local mode. Locomotion benchmarks such as Ant-v4 and Humanoid-v4 provide that regime, while goal-directed single-solution manipulation benchmarks largely do not. A natural extension is multi-solution manipulation, for example bin-picking with multiple valid grasps or insertion with multiple valid approach angles, where the same ensemble structure should convert latent multimodality into faster mode discovery rather than fragmented heads. We leave this evaluation to future work.