Diffusion Advantage Matching: Unified Diffusion Policy Optimization

Anonymous authors
Paper under double-blind review

Abstract

Diffusion policies can represent multimodal behaviors, but improving them online is challenging because their action likelihoods are intractable. Existing diffusion-RL methods address this difficulty through action gradients, likelihood approximations, or backpropagation through the denoising chain (BPTT), using different formulations with distinct limitations. Motivated by KL-regularized RL, we unify these seemingly different methods under Generalized Optimal Policy Matching (GOPM), a policy-agnostic framework that views soft policy improvement as matching the current policy to a soft-optimal target under a general divergence. For diffusion policies, we instantiate GOPM as Diffusion Advantage Matching (DAM), a theoretically grounded method with a detached estimator that avoids BPTT while preserving the soft-optimal distribution. Under the Log-Variance divergence, DAM remains valid for samplers covering both policy and reference, enabling DAM to support online, offline, and hybrid learning under a single objective; to the best of our knowledge, it is the first diffusion-RL method to do so. Across MuJoCo and DeepMind Control, DAM achieves strong online and offline performance, better preserves multimodal distributions, and delivers strong results in both single-stage and two-stage offline-to-online learning.

1 Introduction

Diffusion models have achieved strong results in offline reinforcement learning (RL), where their capacity to represent complex, multimodal distributions allows them to faithfully capture behavior policies from fixed datasets (Wang et al., 2023; Kang et al., 2023; Hansen-Estruch et al., 2023; Lu et al., 2023; Ding & Jin, 2024). However, extending diffusion policies to the online setting, where the data distribution continually evolves through environment interaction without collapsing this multimodal structure, remains an open problem.

The central challenge is that standard policy-gradient objectives require evaluating or differentiating the action likelihood $\log \pi_\theta(a|s)$, which is not available in closed form for diffusion policies. Existing methods address this challenge in several distinct ways. Firstly, action-gradient methods (Psenka et al., 2024; Yang et al., 2023) bypass the likelihood by improving sampled actions using $\nabla_a Q(s, a)$ and then train the diffusion policy to reproduce these actions. Alternatively, proximity-based methods (Ding et al., 2025) instead approximate or reconstruct policy likelihoods through invertible reparameterizations or flow-matching surrogates, enabling PPO-style updates. These lines of work, however, suffer from mode collapse: optimizing toward individual high-value actions rather than the full soft-optimal distribution can yield suboptimal policies that collapse a multimodal policy onto a subset of rewarding modes, despite auxiliary exploration mechanisms such as entropy regularization or noise injection. KL-regularized RL instead performs distribution-level policy improvement relative to a reference policy π_{ref}, yielding the soft-optimal target $\pi^* \propto \pi_{\text{ref}} \exp(Q/\beta)$. When π_{ref} is uniform, this recovers maximum-entropy RL. This target motivates Q-weighting methods (Ding et al., 2024; Ma et al., 2025), which approximate soft optimal target by reweighting the diffusion loss with $\exp\left(Q\left(s, a\right)/\beta\right)$, encouraging towards high-value actions while retaining the multimodal coverage. Other methods backpropagate through the denoising chain, either to maximize expected Q (Wang et al., 2024) or to optimize a trajectory-level

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

maximum-entropy objective (Celik et al., 2025) , incurring high memory cost and gradient instability.

These approaches therefore differ in both their policy-improvement objectives and how those objectives are made tractable for diffusion policies. We unify these perspectives through **Generalized Optimal Policy Matching (GOPM)**, a policy-agnostic KL-regularized RL framework that formulates soft policy improvement as minimizing a general divergence D between the current policy π and the soft-optimal policy π^*. Under different choices of D, GOPM recovers existing KL-regularized online methods, such as QSM and DIME (Psenka et al., 2024; Celik et al., 2025), as its special cases. For diffusion policies, we instantiate GOPM as Diffusion Advantage Matching (DAM), which directly matches the policy to the soft-optimal target using a detached estimator to avoid BPTT while preserving the full soft-optimal distribution coverage. Furthermore, under the Log-Variance divergence, the objective becomes independent of the sampling distribution subject to standard support conditions, enabling a unified objective for online, offline, and hybrid learning.

Empirically, DAM consistently outperforms existing diffusion-RL methods across MuJoCo and DeepMind Control benchmarks, recovers multimodal action distributions that pathwise methods collapse, and learns from offline, online, and hybrid learning under a single objective.

2 Preliminaries

2.1 Diffusion Models

Diffusion models (Ho et al., 2020; Song et al., 2021) generate data by reversing a noise-injection process. The forward process gradually corrupts an initial data point $x \sim q(x)$ into Gaussian noise $x_K \sim \mathcal{N}(0, \mathrm{I})$ according to a predefined noise schedule $\{\alpha_k, \sigma_k\}_{k=0}^K$:

$$x_k = \alpha_k x + \sigma_k \epsilon \tag{1}$$

where $\epsilon \sim \mathcal{N}(0, \mathrm{I})$. This characterizes the forward distribution $q(x_k|x)$ as a Gaussian $\mathcal{N}\left(\alpha_k x, \sigma_k^2 \mathrm{I}\right)$. Conversely, the backward process starts from $x_K \sim \mathcal{N}(0, \mathrm{I})$ and iteratively denoises x_k back to x_{k-1} by sampling from the parameterized backward transition distribution $p_\theta(x_{k-1}|x_k)$. This distribution is learned to match $q(x_{k-1}|x_k, x)$ by minimizing the KL divergence $D_{\mathrm{KL}}\left(q(x_{k-1}|x_k, x) \| p_\theta(x_{k-1}|x_k)\right)$ for every noise level k. By modeling $p_\theta(x_{k-1}|x_k)$ as a Gaussian distribution and using a reparameterization trick, this objective simplifies to the noise matching loss (Ho et al., 2020):

$$\mathcal{L}_{\mathrm{NM}}(\theta) := \mathbb{E}_{x,k,\epsilon,x_k}\left[w_\epsilon(k) \|\epsilon_\theta(x_k, k) - \epsilon\|_2^2\right] \tag{2}$$

where $\epsilon_\theta(x_k, k)$ denotes the noise prediction of x_k at the noise level k; $x \sim q(x)$, $k \sim \mathcal{U}(1, K)$, $\epsilon \sim \mathcal{N}(0, \mathrm{I})$, $w_\epsilon(k)$ is a timestep weighting function, and x_k is computed using Eq. 1.

2.2 KL-Regularized Reinforcement Learning

KL-regularized RL (Haarnoja et al., 2017; 2018; Nachum et al., 2017) aims to learn a policy π that maximizes the regularized expected return:

$$\mathcal{J}(\pi) := \mathbb{E}_{\tau\sim\pi}\left[\sum_{t=0}^{\infty} \gamma^t \left(r(s_t, a_t) - \beta D_{\mathrm{KL}}\left(\pi(\cdot|s_t) \| \pi_{\mathrm{ref}}(\cdot|s_t)\right)\right)\right] \tag{3}$$

where $\tau = (s_0, a_0, s_1, a_1, ...)$ denotes a trajectory generated by π, $\gamma \in (0, 1]$ is the discount factor, π_{ref} is a reference policy, and $\beta \geq 0$ controls the strength of the KL regularization.

The KL penalty encourages the learned policy to remain close to the reference policy while still maximizing reward, thereby improving training stability and mitigating premature convergence to sub-optimal solutions. Moreover, when π_{ref} is chosen to be uniform, KL regularization reduces to entropy regularization, promoting exploration and recovering the standard maximum-entropy (MaxEnt) RL objective (Haarnoja et al., 2018).

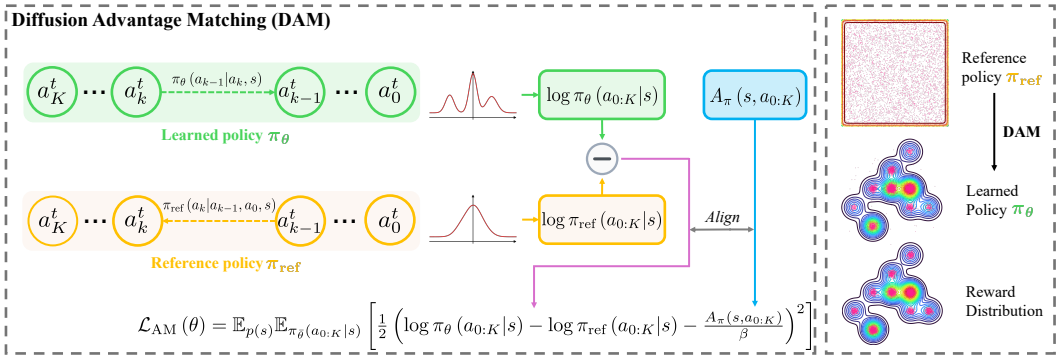

Figure 1: **Overview of Diffusion Advantage Matching (DAM). (Left)** $\mathcal{L}_{\text{AM}}(\theta)$ objective aligns advantage with the policy log-ratio, adjusting the reference policy into the current policy toward optimality. **(Right)** Visualizations demonstrate the resulting policy closely matching the optimal target.

The soft state value function $V_{\pi,\beta}$, soft state-action value function $Q_{\pi,\beta}$, and soft advantage $A_{\pi,\beta}$ are defined as:

$$V_{\pi,\beta}(s) := \mathbb{E}_{\tau\sim\pi}\left[\sum_{t=0}^{\infty}\gamma^t\left(r(s_t,a_t) - \beta\log\frac{\pi(a_t|s_t)}{\pi_{\text{ref}}(a_t|s_t)}\right)\Big|s_0 = s\right] \tag{4}$$

$$Q_{\pi,\beta}(s,a) := \mathbb{E}_{\tau\sim\pi}\left[r(s_0,a_0) + \sum_{t=1}^{\infty}\gamma^t\left(r(s_t,a_t) - \beta\log\frac{\pi(a_t|s_t)}{\pi_{\text{ref}}(a_t|s_t)}\right)\Big|s_0 = s, a_0 = a\right] \tag{5}$$

$$A_{\pi,\beta}(s,a) := Q_{\pi,\beta}(s,a) - V_{\pi,\beta}(s) \tag{6}$$

Details about the soft Bellman expectation equations are provided in Appendix A.2. Due to the convexity of the KL-regularized RL objective when $\beta > 0$, there exists a *unique* optimal policy π^* that admits the closed form expression:

$$\pi^*(a|s) = \pi_{\text{ref}}(a|s)\exp\left(\frac{A_{*,\beta}(s,a)}{\beta}\right) \tag{7}$$

where $V_{*,\beta}(s) := \sup_\pi V_{\pi,\beta}(s)$, $Q_{*,\beta}(s,a) := \sup_\pi Q_{\pi,\beta}(s,a)$, and $A_{*,\beta}(s,a) := Q_{*,\beta}(s,a) - V_{*,\beta}(s)$ denote the optimal soft state value and soft state-action value functions, respectively. $V_{*,\beta}$ and $Q_{*,\beta}$ are related through the soft Bellman optimality equation:

$$V_{*,\beta}(s) = \beta\log\int_{\mathcal{A}}\pi_{\text{ref}}(a|s)\exp\left(\frac{1}{\beta}Q_{*,\beta}(s,a)\right)da \tag{8}$$

Please refer to Appendix A.2 for a detailed mathematical derivations of Eqs. 7 and 8.

3 Generalized Optimal Policy Matching

Inspired by the probabilistic and KL-regularized view of reinforcement learning and control (Levine, 2018; Haarnoja et al., 2017; 2018; Abdolmaleki et al., 2018) we provide a unified probabilistic framework for RL by reinterpreting existing policy optimization algorithms as minimizing a statistical divergence D between current policy π and optimal policy π^*:

$$\mathcal{L}(\pi) := \mathbb{E}_{s\sim p(s)}\left[D\left(\pi(\cdot|s), \pi^*(\cdot|s)\right)\right] \tag{9}$$

We refer to this framework as *Generalized Optimal Policy Matching* (GOPM). Different choices of D recover a variety of existing learning paradigms. For example, choosing D as the forward KL divergence yields the standard behavior cloning (BC) objective, while choosing the Jensen–Shannon (JS) divergence recovers the objective underlying GAIL (Ho & Ermon, 2016), with π^* acting as the expert policy. As shown in Appendix C, several existing KL-regularized or MaxEnt RL algorithms Celik et al. (2025); Psenka et al. (2024); Ma et al.

(2025) emerge as special cases of GOPM. These connections highlight GOPM as a unified framework encompassing both imitation learning and reinforcement learning.

In this work, we focus on the online KL-regularized RL setting. Substituting the closed-form expression of the optimal policy from Eq. 9 gives:

$$\mathcal{L}(\pi) = \mathbb{E}_{s\sim p(s)}\left[D\left(\pi(\cdot|s), \pi_{\text{ref}}(\cdot|s)\exp\left(\frac{A_{*,\beta}(s,\cdot)}{\beta}\right)\right)\right] \tag{10}$$

The main challenge is that the soft optimal advantage function $A_{*,\beta}$ is generally inaccessible in practice. Fortunately, the following theorem, which builds up on the fundamental results in (Haarnoja et al., 2017; 2018; Nachum et al., 2017), shows that the optimal soft advantage $A_{*,\beta}$ can be replaced by the policy-dependent soft advantage $A_{\pi,\beta}$ without changing the optimization target.

Theorem 1. *Consider the KL-regularized RL objective in Eq. 3 with $\beta > 0$. If a policy π satisfies $\pi(a|s) \propto \pi_{ref}(a|s)\exp\left(\frac{A_{\pi,\beta}(s,a)}{\beta}\right)$ for all (s,a) then π is π^*.*

Proof. The proof follows from the Gibbs variational principle; see Appendix A.3. □

Theorem 1 establishes that π^* is the unique fixed point of the self-consistency relation between π and $A_{\pi,\beta}$, motivating the *Generalized Advantage Matching* (GAM) objective:

$$\mathcal{L}(\pi) = \mathbb{E}_{p(s)}\left[D\left(\pi(\cdot|s), \pi_{\text{ref}}(\cdot|s)\exp\left(\frac{A_{\pi,\beta}(s,\cdot)}{\beta}\right)\right)\right] \tag{11}$$

When D is the reverse KL divergence, the GAM objective becomes:

$$\mathcal{L}(\pi) = \mathbb{E}_{p(s)}\left[D_{\text{KL}}\left(\pi(\cdot|s)\middle\|\pi_{\text{ref}}(\cdot|s)\exp\left(\frac{A_{\pi,\beta}(s,\cdot)}{\beta}\right)\right)\right] \tag{12}$$

Minimizing this objective encourages $\pi(\cdot|s) \propto \pi_{\text{ref}}(\cdot|s)\exp\left(\frac{A_{\pi,\beta}(s,\cdot)}{\beta}\right)$, which leads to the optimality of π according to Theorem 1. When π is parameterized as π_θ, the following Advantage Matching (AM) objective can be optimized instead:

$$\mathcal{L}_{\text{AM}}(\theta) = \mathbb{E}_{p(s)}\mathbb{E}_{\pi_{\bar{\theta}}(a|s)}\left[\frac{1}{2}\left(\log\pi_\theta(a|s) - \log\pi_{\text{ref}}(a|s) - \frac{A_{\pi,\beta}(s,a)}{\beta}\right)^2\right] \tag{13}$$

where $\pi_{\bar{\theta}}$ denotes the learned policy with gradients stopped. As shown in Appendix A.5, $\mathcal{L}_{\text{AM}}(\theta)$ is equivalent in terms of optimization to $\mathbb{E}_{p(s)}\left[D_{\text{KL}}\left(\pi_\theta(\cdot|s)\middle\|\pi_{\text{ref}}(\cdot|s)\exp\left(\frac{A_{\pi,\beta}(s,\cdot)}{\beta}\right)\right)\right]$.

In practice, the soft advantage $A_{\pi,\beta}(s,a)$ is often approximated by a parameterized critic (Section 4.1). Training therefore alternates between (i) updating the policy by minimizing the AM objective with the approximated soft advantage, and (ii) updating the critic network to estimate the soft advantage of the latest policy. This alternating optimization closely mirrors policy iteration in Soft Actor-Critic (SAC), where the AM objective serves as the policy improvement step, as formalized in Theorem 2 (which is closely related to Lemma 2 of Haarnoja et al. (2018)).

Theorem 2. *Let π_{old} be any policy and define:*

$$\pi_{new}(\cdot|s) = \underset{\pi}{argmin} D_{KL}\left(\pi(\cdot|s)\middle\|\pi_{ref}(\cdot|s)\exp\left(\frac{A_{\pi_{old},\beta}(s,\cdot)}{\beta}\right)\right) \tag{14}$$

Then $Q_{\pi_{new}}(s,a) \geq Q_{\pi_{old}}(s,a)$ for all (s,a).

Proof. See Appendix 2. □

4 Advantage Matching for Diffusion Policies

In continuous control, policy parameterization plays a crucial role. Explicit policies, such as Gaussian distributions with diagonal covariance matrices (Haarnoja et al., 2018) are computationally convenient but often lack sufficient expressiveness to capture complex dependencies among action dimensions. More expressive alternatives represent $\pi_\theta(a|s)$ implicitly using state-conditional diffusion or flow models (Janner et al., 2022; Hansen-Estruch et al., 2023; Wang et al., 2023), enabling rich multimodal action distributions. However, their implicit nature makes evaluating $\log \pi_\theta(a|s)$ difficult.

To overcome this limitation for diffusion policies, we reinterpret an action as a denoising trajectory rather than a single clean action and refer to our method as *Diffusion Advantage Matching (DAM)*. Specifically, instead of representing an action by a_0, we consider the sequence $a_{0:K} := (a_0, \ldots, a_k, \ldots, a_K)$ where a_k denotes the latent (noisy) action at diffusion step k. Applying the Advantage Matching (AM) objective to trajectory-level actions yields:

$$\mathcal{L}_{\text{AM}}^{\text{online}}(\theta) = \mathbb{E}_{p(s)}\mathbb{E}_{\pi_{\bar{\theta}}(a_{0:K}|s)}\left[\frac{1}{2}\left(\log \pi_\theta(a_{0:K}|s) - \log \pi_{\text{ref}}(a_{0:K}|s) - \frac{A_{\pi,\beta}(s, a_{0:K})}{\beta}\right)^2\right] \tag{15}$$

Since rewards are only received after executing the final clean action a_0, $A_{\pi,\beta}(s, a_{0:K})$ reduces to $A_{\pi,\beta}(s, a_0)$.

The above objective corresponds to minimizing the (reverse) KL divergence over full denoising action trajectories. By the data processing inequality, this trajectory-level KL divergence *upper bounds* the KL divergence defined over clean actions alone (Eq. 13). A complete proof is provided in Appendix A.6. At convergence, it enforces the condition $\pi_\theta(a_0|s) \propto \pi_{\text{ref}}(a_0|s) \exp\left(\frac{A_{\pi,\beta}(s,a_0)}{\beta}\right)$ which, by Theorem 1, uniquely characterizes the optimal KL-regularized policy $\pi^*(a_0|s)$.

The policy $\pi_\theta(a_{0:K}|s)$ is typically defined as a reverse-time diffusion process:

$$\pi_\theta(a_{0:K}|s) = q(a_K) \prod_{k=1}^{K} \pi_\theta(a_{k-1}|a_k, s) \tag{16}$$

where $q(a_K) = \mathcal{N}(0, \mathrm{I})$. Meanwhile, the reference policy $\pi_{\text{ref}}(a_{0:K}|s)$ can be expressed in either of two ways:

$$\text{Formula 1:} \quad \pi_{\text{ref}}(a_{0:K}|s) = q(a_K) \prod_{k=1}^{K} \pi_{\text{ref}}(a_{k-1}|a_k, s) \tag{17}$$

$$\text{Formula 2:} \quad \pi_{\text{ref}}(a_{0:K}|s) = \pi_{\text{ref}}(a_0|s) \prod_{k=1}^{K} \pi_{\text{ref}}(a_k|a_{k-1}, a_0, s) \tag{18}$$

Formula 1 mirrors the decomposition of $\pi_\theta(a_{0:K}|s)$ and is the natural choice when π_{ref} is itself a pretrained diffusion policy from which π_θ is finetuned. Formula 2, on the other hand, is typically used when π_θ is trained from scratch and π_{ref} is instead a predefined, non-parametric diffusion process conditioned on the clean action a_0. A common instantiation of the latter sets $\pi_{\text{ref}}(a_0|s)$ to a uniform distribution and $\pi_{\text{ref}}(a_k|a_{k-1}, a_0, s)$ to a Gaussian distribution, which recovers the MaxEnt diffusion policy RL setting (Haarnoja et al., 2018).

4.1 Soft Advantage Estimation

To estimate $A_{\pi,\beta}(s, a_{0:K})$, we leverage the soft Bellman expectation equation:

$$\begin{aligned} A_{\pi,\beta}(s, a_{0:K}) &= Q_{\pi,\beta}(s, a_{0:K}) - \mathbb{E}_{\pi(a'_{0:K}|s)}\left[Q_{\pi,\beta}(s, a'_{0:K}) - \beta \log \frac{\pi(a'_{0:K}|s)}{\pi_{\text{ref}}(a'_{0:K}|s)}\right] \\ &= Q_{\pi,\beta}(s, a_0) - \mathbb{E}_{\pi(a'_{0:K}|s)}\left[Q_{\pi,\beta}(s, a'_0) - \beta \log \frac{\pi(a'_{0:K}|s)}{\pi_{\text{ref}}(a'_{0:K}|s)}\right] \end{aligned}$$

This expression naturally suggests a leave-one-out (LOO) Monte Carlo estimator (Salimans & Knowles, 2014; Kool et al., 2019), which provides an unbiased and low-variance estimate of the soft advantage. Specifically, for a given state s, we sample N independent denoising action trajectories $\left\{a_{0:K}^{(n)}\right\}_{n=1}^{N}$ from the current policy π. The soft advantage of the n-th trajectory is then approximated as:

$$A_{\pi,\beta}\left(s, a_{0:K}^{(n)}\right) \approx Q_{\pi,\beta}\left(s, a_0^{(n)}\right) - \frac{1}{N-1}\sum_{i\neq n}\left(Q_{\pi,\beta}\left(s, a_0^{(i)}\right) - \beta \log \frac{\pi\left(a_{0:K}^{(i)}|s\right)}{\pi_{\text{ref}}\left(a_{0:K}^{(i)}|s\right)}\right) \quad (19)$$

As discussed in the previous section, the trajectory-level log-policy ratio admits a tractable factorization and can therefore be computed efficiently. To estimate $Q_{\pi,\beta}$ we train a parameterized critic network Q_ϕ by minimizing the soft Bellman residual:

$$\mathcal{L}_{\text{critic}}(\phi) = \mathbb{E}_{(s,a)\sim\mathcal{B}}\left[\left(Q_\phi(s,a) - \mathcal{T}_\pi Q_{\bar{\phi}}(s,a)\right)^2\right] \quad (20)$$

where $\mathcal{B}$ denotes the replay buffer, $Q_{\bar{\phi}}$ is Q_ϕ with gradients stopped, and $\mathcal{T}_\pi$ is the soft Bellman operator defined as:

$$\mathcal{T}_\pi Q_{\bar{\phi}}(s,a_0) = r(s,a_0) + \gamma \mathbb{E}_{p(s'|s,a_0)}\mathbb{E}_{\pi(a'_{0:K}|s')}\left[Q_{\bar{\phi}}(s', a'_0) - \beta \log \frac{\pi(a'_{0:K}|s')}{\pi_{\text{ref}}(a'_{0:K}|s')}\right]$$

When the environment dynamics are deterministic, the next state s' is uniquely determined by (s, a_0). The Bellman target can therefore be approximated using a single denoising action trajectory $a'_{0:K}$ sampled from the current policy π:

$$\mathcal{T}_\pi Q_{\bar{\phi}}(s,a_0) \approx r(s,a_0) + \gamma\left(Q_{\bar{\phi}}(s', a'_0) - \beta \log \frac{\pi(a'_{0:K}|s')}{\pi_{\text{ref}}(a'_{0:K}|s')}\right)$$

Training proceeds by alternating between policy updates using Eq. 15 and critic updates using Eq. 20. The full training algorithm is provided in Algorithm 1.

4.2 Extension to Offline and Hybrid RL

The online AM objective with the LOO estimator (Eq. 19) can be regarded as an instance of the Log-Variance loss Richter et al. (2020) where the sampling policy r taken to be the current policy $\pi_{\bar{\theta}}$:

$$\mathcal{L}_{\text{AM}}^{\text{online}}(\theta) \equiv \mathbb{E}_{p(s)}\left[\frac{1}{2}\text{Var}_{r=\pi_{\bar{\theta}}}\left(\log \frac{\pi_\theta(a|s)}{\pi_{\text{ref}}(a|s)\exp\left(\frac{Q_{\pi,\beta}(s,a)}{\beta}\right)}\right)\right] \quad (21)$$

Building on this perspective, we extend the framework to the offline RL setting via the loss:

$$\mathcal{L}_{\text{AM}}^{\text{offline}}(\theta) = \mathbb{E}_{(s,a)\sim\mathcal{D}}\left[\frac{1}{2}\text{Var}_{r_\alpha(a_{0:K}|s,a)}\left(\log \frac{\pi_\theta(a_{0:K}|s)}{\pi_{\text{ref}}(a_{0:K}|s)} - \frac{Q_{\phi,\beta}(s,a_0)}{\beta}\right)\right] \quad (22)$$

where $\mathcal{D}$ is the offline dataset and the sampling policy $r_\alpha(a_{0:K}|s,a)$ is defined as:

$$r_\alpha(a_{0:K}|s,a) = \alpha\pi_{\bar{\theta}}(a_{0:K}|s) + (1-\alpha)\,\delta_{a_0=a}\prod_{k=1}^{K}\pi_{\text{ref}}(a_k|a_{k-1},a_0,s)$$

Under r_α, an action trajectory $a_{0:K}$ is, with probability α, drawn from the learned diffusion policy π_θ by denoising from noise, and with probability $(1-\alpha)$, constructed as a forward-noised trajectory starting from the offline action a. Sampling from r_α stablizes training and encourages exploration beyond the offline data.

Since $Q_{\pi,\beta}$ is difficult to estimate reliably in the offline setting, we treat $Q_{\phi,\beta}$ as a surrogate for $Q_{*,\beta}$ and learn it via IQL on $\mathcal{D}$. Under this formulation, $\mathcal{L}_{\text{AM}}^{\text{offline}}(\theta)$ can be seen as a realization of the loss in Eq. 10. Implementation details are given in Appendix B.1.

	Humanoid	Walker2D	HalfCheetah	Ant	DogRun	DogTrot
DIPO	4966 ± 663	2968 ± 1258	6773 ± 325	1417 ± 642	135.1 ± 16.4	158.4 ± 23.3
QSM	4092 ± 2438	826 ± 483	8212 ± 596	2119 ± 163	88.8 ± 8.1	114.9 ± 9.1
DPMD	3947 ± 1194	917 ± 842	8090 ± 648	4643 ± 1034	451.9 ± 52.6	641.8 ± 51.2
QVPO	2111 ± 1293	177 ± 153	5582 ± 892	985 ± 109	14.7 ± 2.4	16.8 ± 9.9
DACER	2660 ± 1455	919 ± 665	6015 ± 419	3277 ± 2646	395.1 ± 29.9	728.7 ± 94.5
SDAC	5017 ± 1739	1394 ± 556	6734 ± 778	528 ± 8	140.4 ± 3.8	246.9 ± 23.3
DIME	12700 ± 242	2535 ± 2049	11261 ± 784	5196 ± 899	602.6 ± 90.1	752.7 ± 65.3
DAM	$\mathbf{13078 \pm 221}$	$\mathbf{4103 \pm 1364}$	$\mathbf{12232 \pm 474}$	$\mathbf{5848 \pm 528}$	$\mathbf{740.1 \pm 57.6}$	$\mathbf{895.0 \pm 14.2}$

Table 1: Quantitative comparison of DAM with online diffusion policy baselines. Values denote mean score ± std over 5 seeds. Best per task in **bold**.

Hybrid RL In the hybrid setting, $Q_{\pi,\beta}$ can instead be estimated from trajectories sampled by π_θ. We therefore combine the online and offline state distributions and optimize:

$$\mathcal{L}_{\text{AM}}^{\text{hybrid}}(\theta) = \lambda \mathcal{L}_{\text{AM}}^{\text{online}}(\theta) + (1-\lambda)\,\tilde{\mathcal{L}}_{\text{AM}}^{\text{offline}}(\theta) \tag{23}$$

where $\lambda \in [0,1]$ is the online data mixing coefficient, and $\tilde{\mathcal{L}}_{\text{AM}}^{\text{offline}}$ take the same form as Eq. 22 but uses the current soft critic $Q_{\pi,\beta}$ inplace of the frozen IQL critic.

5 Experiment

Experimental setup We evaluate DAM across online, offline, and hybrid reinforcement learning settings. For *online learning*, we consider 8 MuJoCo Gymnasium-v4 locomotion tasks Towers et al. (2025) and 7 challenging and high-dimensional DMC Dog/Humanoid tasks Tassa et al. (2018). For *offline and hybrid learning*, we consider 3 tasks: HalfCheetah, Hopper, and Walker2d, each associated with the D4RL-v2 Medium, Medium-Replay, and Medium-Expert datasets, following Fu et al. (2020). Unless otherwise specified, we train for 300k environment interactions on MuJoCo and 1M interactions on Humanoid-v4/DMC. All baselines are reproduced using their official implementations under matched evaluation protocols, with additional implementation details provided in Appendix G.

5.1 Main Results

5.1.1 Online Learning

DAM achieves the highest final return across MuJoCo benchmarks (Table 1), with full results provided in Appendix E. More importantly, DAM's advantage over our primary baseline, DIME, becomes more pronounced on challenging and high-dimensional DMC Dog tasks (Figure 2). Because DAM and DIME use the same diffusion architecture and reverse-KL objective, DAM's faster convergence and higher returns isolate the policy-gradient estimator as the key difference. In Appendix $E.3$, we analyze this difference and show how DAM's detached estimator avoids DIME's BPTT approach, which creates an optimization bottleneck.

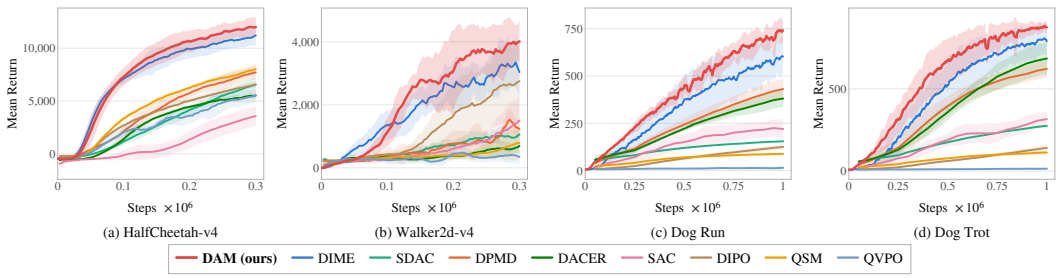

Figure 2: Online learning curves on hard MuJoCo (HalfCheetah-v4, Walker2d-v4) and DMC (Dog Run, Dog Trot) tasks. DAM consistently outperforms remaining baselines.

Task	CQL	IQL	IDQL	QGPO	DAM
hopper-medium-replay	95.0	94.7	92.1	96.9	**100.24±0.38**
hopper-medium	58.5	64.3	65.4	**98.0**	89.24 ± 3.29
hopper-medium-expert	105.4	91.5	**108.6**	108.1	102.43 ± 0.62
halfcheetah-medium-replay	44.2	44.2	45.8	47.6	**48.35±2.00**
halfcheetah-medium	47.2	47.4	51.0	54.1	**55.45±1.15**
halfcheetah-medium-expert	91.6	86.7	**95.9**	93.5	87.92±1.28
walker2d-medium-replay	77.2	73.9	85.1	84.4	**97.87±0.91**
walker2d-medium	72.5	78.3	82.5	86.0	**88.74±0.74**
walker2d-medium-expert	108.8	109.6	**112.7**	110.7	109.98 ± 0.41

Method	Fwd. KL ↓	Rev. KL ↓	JSD ↓	MMD ↓
DACER	15.26	1.81	0.457	0.756
QVPO	4.92	1.35	0.364	0.298
DPMD	9.28	1.35	0.366	0.371
QSM	13.56	1.40	0.384	0.352
DIME	0.519	0.337	0.094	0.039
DAM	**0.174**	**0.146**	**0.039**	**0.024**

Table 2: **(a)** Final evaluation return (mean ± std over 5 seeds) for offline D4RL-v2 settings and **(b)** Coverage metrics between generated and true 9-GMM samples. Best results are in **bold** and second-best are underlined.

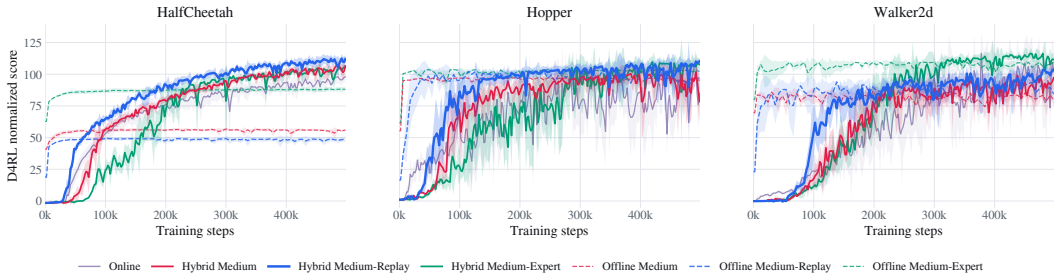

Figure 3: **DAM supports online, offline, and hybrid training.** Learning curves on HalfCheetah, Hopper, and Walker2d over 500k training steps on 3 D4RL datasets.

5.1.2 Offline Learning

As expected in offline RL, Medium-Expert datasets often yield high absolute returns due to their high-quality trajectories. More revealingly, DAM attains positive margins over strongest baselines on Medium-Replay datasets, while trailing IDQL and QGPO on Medium-Expert datasets. This pattern is consistent with the Log-Variance support condition (Eq. 21): replay data provides broader state-action coverage, offering better support for the policy and reference distributions and enabling more faithful distributional matching, whereas narrower expert support violates the condition more severely.

5.1.3 Hybrid Learning

Figure 3 shows that hybrid training consistently improves beyond the corresponding offline solutions through online interaction. The benefit depends on the offline data used in training: Medium-Replay provides stronger early guidance through its broader state-action coverage, whereas Medium-Expert's narrower support leads to slower initial improvement, consistent with the support condition discussed in Section 4.2. This is all achieved in a single-stage hybrid learning without separate offline pretraining and online fine-tuning. Despite this single-stage training, DAM achieves competitive performance with two-stage offline-to-online baselines while using significantly less training steps (Appendix (4)). A detailed comparison of DAM's one- and two-stage training is examined in Appendix E.2, with the effect of λ provided in Section E.3.

5.1.4 Distributional Coverage

Theorem 1 enforces $\pi \propto \pi_{\text{ref}} \exp(A_{\pi,\beta}/\beta)$ pointwise, providing a direct distributional target. We test this on a 2D nine-component Gaussian mixture (9-GMM) with *unequal* weights against diffusion online RL methods under identical experimental settings. Quantitatively, DAM achieves the lowest error on all four metrics (Table 2). Qualitatively, Figure 4 shows that DAM recovers all nine modes while preserving their unequal relative densities, whereas prior standard RL and MaxEnt-RL methods either collapse or produce overly diffuse samples. The effect of β on the coverage–concentration trade-off is analyzed in Section 5.2.

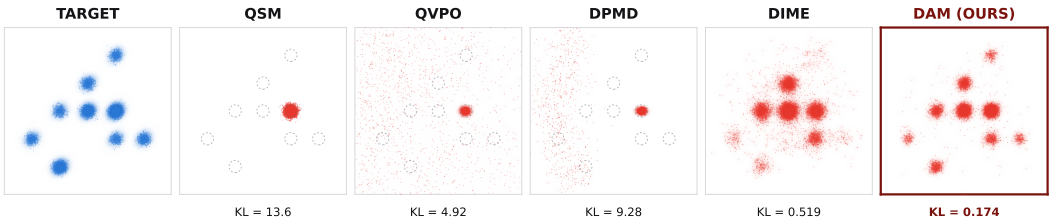

Figure 4: Distributional coverage on 9-GMM: DAM vs. MaxEnt RL baselines.

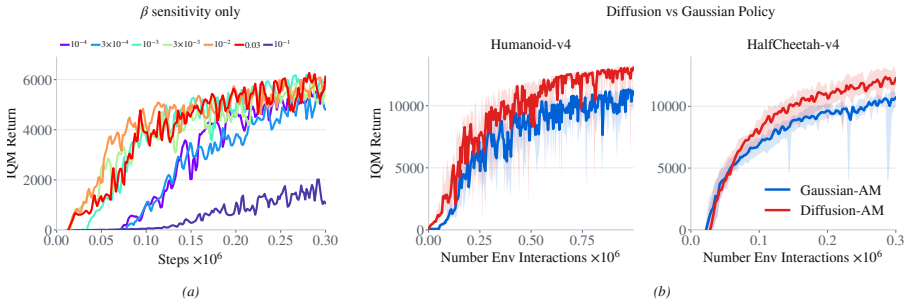

Figure 5: **DAM ablation analysis**. **(a)** Effect of β on Ant-v4. **(b)** Diffusion versus Gaussian policies on Humanoid-v4 and HalfCheetah-v4.

5.2 Ablation Studies

We examine the effects of KL-regularized coefficient β and diffusion policies here. Further ablations are provided in Appendix E.

Impact of KL-regularized coefficient β: We train multiple variants of DAM for $\beta \in [10^{-4}, 10^{-1}]$, and visualize their learning curves in Fig. 5a. We observe that large β (e.g., 0.1) imposes strong KL regularization, constraining the policy toward the reference distribution and preventing effective policy improvement throughout training. In contrast, very small β values ($\beta \in \left[3 \times 10^{-4}, 10^{-3}\right]$ yield a highly concentrated soft-improvement target that focuses on current high-advantage actions, hindering exploration for early learning and slowing the discovery of high-return policies.

Analysis of Diffusion vs Gaussian Policies: To isolate the effect of the policy parameterization, we instantiate Generalized Advantage Matching with a Gaussian policy (Gaussian-AM) using Eq. 13 matching implementation details. Fig. 5 shows that Gaussian-AM consistently converges slower and achieves lower returns than Diffusion-AM, which highlights the benefit of a more expressive policy class.

6 Conclusion

We present Diffusion Advantage Matching (DAM) as a diffusion-policy instantiation of Generalized Optimal Policy Matching (GOPM) framework, unifying existing approaches under a common divergence-based formulation. DAM matches policy log-ratios to soft advantages with a detached estimator, avoiding backpropagation through the denoising chain while preserving the soft-optimal distribution. This formulation further decouples the optimization objective from the sampling distribution, allowing the same objective to support online, offline, and hybrid learning.

7 AI Disclosure

This paper uses AI for: formulating questions for surveys or interviews, create or modify scientific figures or images, suggest experimental parameters, create or edit software code, creation of artifacts, draft parts of a research paper, transcribe recordings of research material, summarize or analyse existing literature, discover research topics or identify gaps, edit a research paper to improve readability, identify relevant literature, format references, propose a title or keywords for a research paper.

References

Abbas Abdolmaleki, Jost Tobias Springenberg, Yuval Tassa, Rémi Munos, Nicolas Heess, and Martin A. Riedmiller. Maximum a posteriori policy optimisation. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL https://openreview.net/forum?id=S1ANxQW0b.

Ha Manh Bui, Metod Jazbec, Eric T. Nalisnick, and Anqi Liu. Efficient and uncertainty-aware diffusion framework for offline-to-online reinforcement learning. *CoRR*, abs/2605.30776, 2026. doi: 10.48550/ARXIV.2605.30776. URL https://doi.org/10.48550/arXiv.2605.30776.

Onur Celik, Zechu Li, Denis Blessing, Ge Li, Daniel Palenicek, Jan Peters, Georgia Chalvatzaki, and Gerhard Neumann. DIME: diffusion-based maximum entropy reinforcement learning. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, volume 267 of *Proceedings of Machine Learning Research*. PMLR / OpenReview.net, 2025. URL https://proceedings.mlr.press/v267/celik25a.html.

Tianyu Chen, Zhendong Wang, and Mingyuan Zhou. Diffusion policies creating a trust region for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 37: 50098–50125, 2024.

Shutong Ding, Ke Hu, Zhenhao Zhang, Kan Ren, Weinan Zhang, Jingyi Yu, Jingya Wang, and Ye Shi. Diffusion-based reinforcement learning via q-weighted variational policy optimization. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/6111371a868af8dcfba0f96ad9e25ae3-Abstract-Conference.html.

Shutong Ding, Ke Hu, Shan Zhong, Haoyang Luo, Weinan Zhang, Jingya Wang, Jun Wang, and Ye Shi. Genpo: Generative diffusion models meet on-policy reinforcement learning. In Danielle Belgrave, Cheng Zhang, Laura N. Montoya, Hsuan-Tien Lin, Razvan Pascanu, Piotr Koniusz, Marzyeh Ghassemi, Nancy Chen, Iván Vladimir Meza Ruíz, and Arturo Loaiza-Bonilla (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025. URL http://papers.nips.cc/paper_files/paper/2025/hash/bd31bfd4caa85bffe07a35568182cdfa-Abstract-Conference.html.

Zihan Ding and Chi Jin. Consistency models as a rich and efficient policy class for reinforcement learning. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=v8jdwkUNXb.

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

Matthieu Geist, Bruno Scherrer, and Olivier Pietquin. A theory of regularized markov decision processes. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2160–2169. PMLR, 2019. URL http://proceedings.mlr.press/v97/geist19a.html.

Tuomas Haarnoja, Haoran Tang, Pieter Abbeel, and Sergey Levine. Reinforcement learning with deep energy-based policies. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW,*

Australia, 6-11 August 2017, volume 70 of *Proceedings of Machine Learning Research*, pp. 1352–1361. PMLR, 2017. URL http://proceedings.mlr.press/v70/haarnoja17a.html.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International Conference on Machine Learning*, pp. 1861–1870. PMLR, 2018.

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

Jonathan Ho and Stefano Ermon. Generative adversarial imitation learning. In Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pp. 4565–4573, 2016. URL https://proceedings.neurips.cc/paper/2016/hash/cc7e2b878868cbae992d1fb743995d8f-Abstract.html.

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

Michael Janner, Yilun Du, Joshua B. Tenenbaum, and Sergey Levine. Planning with diffusion for flexible behavior synthesis. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato (eds.), *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pp. 9902–9915. PMLR, 2022. URL https://proceedings.mlr.press/v162/janner22a.html.

Bingyi Kang, Xiao Ma, Chao Du, Tianyu Pang, and Shuicheng Yan. Efficient diffusion policies for offline reinforcement learning. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/d45e0bfb5a39477d56b55c0824200008-Abstract-Conference.html.

Wouter Kool, Herke van Hoof, and Max Welling. Buy 4 reinforce samples, get a baseline for free! *ICLR 2019 workshop: Deep RL Meets Structured Prediction*, 2019.

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

Sergey Levine. Reinforcement learning and control as probabilistic inference: Tutorial and review. *CoRR*, abs/1805.00909, 2018. URL http://arxiv.org/abs/1805.00909.

Cheng Lu, Huayu Chen, Jianfei Chen, Hang Su, Chongxuan Li, and Jun Zhu. Contrastive energy prediction for exact energy-guided diffusion sampling in offline reinforcement learning. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 22825–22855. PMLR, 2023. URL https://proceedings.mlr.press/v202/lu23d.html.

Haitong Ma, Tianyi Chen, Kai Wang, Na Li, and Bo Dai. Efficient online reinforcement learning for diffusion policy. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, volume 267 of *Proceedings of Machine Learning Research*. PMLR / OpenReview.net, 2025. URL https://proceedings.mlr.press/v267/ma25d.html.

Ofir Nachum, Mohammad Norouzi, Kelvin Xu, and Dale Schuurmans. Bridging the gap between value and policy based reinforcement learning. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 2775–2785, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/facf9f743b083008a894eee7baa16469-Abstract.html.

Michael Psenka, Alejandro Escontrela, Pieter Abbeel, and Yi Ma. Learning a diffusion model policy from rewards via q-score matching. In Ruslan Salakhutdinov, Zico Kolter, Katherine A. Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, volume 235 of *Proceedings of Machine Learning Research*, pp. 41163–41182. PMLR / OpenReview.net, 2024. URL https://proceedings.mlr.press/v235/psenka24a.html.

Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley Series in Probability and Statistics. Wiley, 1994. ISBN 978-0-47161977-2. doi: 10.1002/9780470316887. URL https://doi.org/10.1002/9780470316887.

Lorenz Richter, Ayman Boustati, Nikolas Nüsken, Francisco Ruiz, and Omer Deniz Akyildiz. Vargrad: a low-variance gradient estimator for variational inference. *Advances in Neural Information Processing Systems*, 33:13481–13492, 2020.

Tim Salimans and David A Knowles. On using control variates with stochastic approximation for variational bayes and its connection to stochastic linear regression. *arXiv preprint arXiv:1401.1022*, 2014.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *CoRR*, abs/2402.03300, 2024. doi: 10.48550/ARXIV.2402.03300. URL https://doi.org/10.48550/arXiv.2402.03300.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

Yuval Tassa, Yotam Doron, Alistair Muldal, Tom Erez, Yazhe Li, Diego de Las Casas, David Budden, Abbas Abdolmaleki, Josh Merel, Andrew Lefrancq, Timothy P. Lillicrap, and Martin A. Riedmiller. Deepmind control suite. *CoRR*, abs/1801.00690, 2018. URL http://arxiv.org/abs/1801.00690.

Mark Towers, Ariel Kwiatkowski, John U. Balis, Gianluca De Cola, Tristan Deleu, Manuel Goulão, Andreas Kallinteris, Markus Krimmel, Arjun KG, Rodrigo Perez-Vicente, J. K. Terry, Andrea Pierré, Sander Schulhoff, Jun Jet Tai, Hannah Tan, and Omar G. Younis. Gymnasium: A standard interface for reinforcement learning environments. In Danielle Belgrave, Cheng Zhang, Laura N. Montoya, Hsuan-Tien Lin, Razvan Pascanu, Piotr Koniusz, Marzyeh Ghassemi, Nancy Chen, Iván Vladimir Meza Ruíz, and Arturo Loaiza-Bonilla (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025. URL http://papers.nips.cc/paper_files/paper/2025/hash/d7ff1795e8527f6443371c3933bdb52b-Abstract-Datasets_and_Benchmarks_Track.html.

Siddarth Venkatraman, Moksh Jain, Luca Scimeca, Minsu Kim, Marcin Sendera, Mohsin Hasan, Luke Rowe, Sarthak Mittal, Pablo Lemos, Emmanuel Bengio, Alexandre Adam, Jarrid Rector-Brooks, Yoshua Bengio, Glen Berseth, and Nikolay Malkin. Amortizing intractable inference in diffusion models for vision, language, and control. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver,*

BC, Canada, December 10 - 15, 2024, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/8b21a7ea42cbcd1c29a7a88c444cce45-Abstract-Conference.html.

Yinuo Wang, Likun Wang, Yuxuan Jiang, Wenjun Zou, Tong Liu, Xujie Song, Wenxuan Wang, Liming Xiao, Jiang Wu, Jingliang Duan, and Shengbo Li. Diffusion actor-critic with entropy regulator. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/6174c67b136621f3f2e4a6b1d3286f6b-Abstract-Conference.html.

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

Long Yang, Zhixiong Huang, Fenghao Lei, Yucun Zhong, Yiming Yang, Cong Fang, Shiting Wen, Binbin Zhou, and Zhouchen Lin. Policy representation via diffusion probability model for reinforcement learning. *CoRR*, abs/2305.13122, 2023. doi: 10.48550/ARXIV.2305.13122. URL https://doi.org/10.48550/arXiv.2305.13122.