DIFFUSION-GUIDED ADVERSARIAL IMITATION LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Generative Adversarial Imitation Learning (GAIL) discriminates expert from learner states with a binary classifier and uses its output as a reward signal, yet its minimax objective is unstable to optimize. Score-based diffusion models, which match distributions by regressing score functions, offer a more stable route to imitation learning. In this paper, we analyze the direct regression of the learner's score against the expert's, and show that it yields biased policy gradients. We then prove that this score-based discrepancy admits an equivalent minimax form and can be optimized with unbiased policy gradients. We instantiate this form as **Diffusion-Guided Adversarial Imitation Learning (DGAIL)**: a vector-field discriminator guided by a pre-trained expert score network is maximized against the policy's trajectories; the policy is updated to minimize the objective. Theoretically, we prove that the game value attained by any policy during training certifies a bound on its true divergence up to explicit approximation errors. This bound further yields an instance-dependent imitation gap for any bounded evaluation cost. To the best of our knowledge, DGAIL is the first method to bring the explicit score signal of diffusion models into adversarial imitation learning. Empirically, DGAIL matches or outperforms strong imitation learning baselines on MuJoCo continuous control tasks.

1 INTRODUCTION

Reinforcement learning (RL) agents typically rely on carefully engineered reward functions, which are difficult to design for many real-world tasks. Imitation learning (IL) bypasses reward engineering by learning directly from expert demonstrations, serving as a fundamental paradigm across autonomous driving, robotics, and large language models (Pomerleau, 1991; Osa et al., 2018; Ouyang et al., 2022; Black et al., 2024). A classical route, behavioral cloning (BC), casts the problem as supervised learning but suffers from compounding errors under state-distribution shift (Ross & Bagnell, 2010). Adversarial imitation learning (AIL) instead matches the learner's state-action distribution to the expert's in an adversarial manner, offering better generalization at the cost of a more complex minimax training procedure (Abbeel & Ng, 2004; Ziebart et al., 2008; Fu et al., 2018; Kostrikov et al., 2019). GAIL (Ho & Ermon, 2016) frames imitation as a sequential generalization of GANs, where the output of a binary classifier over expert and learner data serves as the reward signal for policy optimization. Despite its success, GAIL-style discriminators suffer from a well-documented issue: the Jensen–Shannon (JS) or KL objectives underlying the binary-classifier formulation are unstable to optimize in practice (Arjovsky et al., 2017; Kostrikov et al., 2019), with training loss oscillating and convergence slowing down (Luo et al., 2024).

Modern score-based diffusion models (Ho et al., 2020; Song et al., 2021) provide a complementary resource that AIL methods have not yet leveraged. Such models learn a target distribution by fitting the gradient of its log-density, a vector field known as the *score function*. The generative process depends only on this learned score (Song et al., 2021), and fitting it is a stationary regression objective that affords distribution coverage unlike adversarial training (Dhariwal & Nichol, 2021), yielding markedly more stable optimization in practice. Recently, attempts have been made to bring diffusion models into imitation learning. Lai et al. (2024) and Wang et al. (2024) inject diffusion into the adversarial loop, in essence still running AIL on diffusion steps with the explicit score signal left untapped; Wu et al. (2025) match the student's and expert's scores by supervised regression, departing from the adversarial framework altogether. In this paper, we further explore whether the

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.

score-matching objective can be extended to distribution matching, yielding an imitation learning method built on score functions. We first analyze the direct approach, regressing the learner's score against the expert's and taking the resulting score-based discrepancy as the objective; this objective involves the learner's score $\nabla \log \rho_\pi$ and is therefore π-dependent through the learner occupancy, so a policy-gradient update that treats it as a fixed cost is biased. Our key observation is that the score-based discrepancy admits a variational form, and a minimax game arises naturally: writing the objective as optimization over vector fields removes the learner's score from the cost via integration by parts, leaving a cost that depends only on the expert score and the divergence of the learned vector field. The emerging game has precisely the form of adversarial imitation learning, with the vector field acting as the discriminator and its optimum defining the reward; because this reward is independent of the policy being evaluated, the learner remains trainable by standard policy-gradient methods.

We consolidate these ideas into **Diffusion-Guided Adversarial Imitation Learning (DGAIL)**: a pre-trained diffusion model provides the expert score prior s_ψ; a vector-field network f_ϕ serves as the discriminator, matching the learner and expert distributions over diffused states at multiple noise scales; and the discriminator output is converted into a dual reward on which the policy is updated via RL, the two updates alternating until convergence. Theoretically, building on the adversarial variational form of the score-based divergence, we prove a training-time certificate: the game value attained by any policy bounds its true score-based divergence, with the quality of the score prior, the discriminator capacity, and the stochastic divergence estimation entering as explicit error terms; the bound further transfers to an instance-dependent imitation gap under any bounded evaluation cost, which tightens when the cost varies little over the expert occupancy rather than being a worst-case statement. Empirically, DGAIL matches or outperforms strong AIL baselines and BC on MuJoCo continuous control benchmarks. To the best of our knowledge, it is the first method to bring the explicit score signal of pre-trained diffusion models into adversarial imitation, with the prior kept fixed and the learner's score never estimated.

2 Related Work

Adversarial imitation learning. GAIL (Ho & Ermon, 2016) formulates IL as a two-player game between a policy and a binary discriminator, matching the expert occupancy under the Jensen–Shannon divergence. This formulation has been extended along several axes. Off-policy and sample-efficient variants (DAC (Kostrikov et al., 2019)) reduce the interaction cost of the adversarial loop by reusing a replay buffer. Inverse-RL and Q-based formulations (AIRL (Fu et al., 2018), IQ-Learn (Garg et al., 2021)) interpret the discriminator as an implicit reward or Q-function and optimize the policy against the recovered signal. Distribution-divergence variants replace the JS objective by other divergences: ValueDICE (Kostrikov et al., 2021) optimizes an f-divergence through a change-of-measure identity, and Wasserstein-based objectives (Xiao et al., 2019; Dadashi et al., 2021) exploit the geometry of the state space through the dual formulation of the Wasserstein distance, which often yields more stable gradients. Despite these advances, the discriminators in this family remain scalar classifiers that emit implicit signals, and their training stability and generalization remain a central concern: a naive binary classifier can overfit the finite demonstration set, produce a non-stationary and noisy reward signal, and lead to unstable or degenerate policies during training. A number of works have therefore focused on stabilizing the adversarial loop, including control-theoretic analyses (C-GAIL (Luo et al., 2024)) and boosting-based perspectives (AILBoost (Chang et al., 2024)). In contrast, our discriminator is a vector field whose variational objective is the Fisher divergence between the diffused occupancies: it emits an explicit score-difference signal and requires neither a learned critic nor an estimate of the learner's score.

Diffusion models for imitation. Owing to their ability to represent multi-modal action distributions, diffusion models have found broad application in imitation learning. Chi et al. (2023) train a conditional diffusion model to generate action sequences from observations under behavior-cloning supervision; Janner et al. (2022) plan by directly synthesizing trajectories with a diffusion model; Ajay et al. (2023) cast offline decision making as conditional generative modeling over trajectories; and Wang et al. (2023) use diffusion models as an expressive policy class for offline reinforcement learning. More recently, the application of diffusion models has expanded from policy parameterization to distribution matching between the learner and the expert: Lai et al. (2024) derive an adversarial reward from a diffusion model trained on the expert data, and Wang et al. (2024) build

a discriminator from the noise-prediction error on diffused state–action pairs. In both, the diffusion component still emits a scalar signal under a Jensen–Shannon-type objective, whereas our variational objective is the diffusion Fisher divergence.

3 PRELIMINARIES

3.1 MARKOV DECISION PROCESS AND STATE OCCUPANCY

We consider an episodic Markov decision process (MDP) $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, H, c)$ with state space $\mathcal{S} \subseteq \mathbb{R}^d$, action space $\mathcal{A}$, transition kernel P, horizon H, and a bounded cost $c : \mathcal{S} \to [0, C_{\max}]$. A policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ induces a state occupancy measure $\rho_\pi(s) = \frac{1}{H}\sum_{h=1}^{H} \rho_{\pi,h}(s)$, the time-averaged distribution of states visited by π. We write $V_c^\pi = \mathbb{E}_{s\sim\rho_\pi}[c(s)]$ for the expected cost of π under c. The expert is a policy π_E whose demonstrations $\mathcal{D}_E = \{s^{(i)}\}_{i=1}^N$ are state-only samples from $\rho_E := \rho_{\pi_E}$. Our goal is imitation: find a policy π with small *imitation gap* $V_c^{\pi_E} - V_c^\pi$ for the (unknown) ground-truth cost c (Xu et al., 2020; 2024).

3.2 DIFFUSION MODELS

Diffusion models (Ho et al., 2020; Song et al., 2021) are deep generative models composed of two stochastic processes: a forward noising process that progressively corrupts clean data with Gaussian noise, and a reverse denoising process that inverts this corruption. We adopt this framework to learn a distributional prior over the expert states.

We consider a variance-preserving forward diffusion $q_t(s_t \mid s_0)$ over a discrete set of diffusion times $t = 1, \dots, T$ applied to initial states $s_0 \sim \rho$:

$$s_t = \alpha_t\, s_0 + \sigma_t\, \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, I), \qquad q_t(s_t \mid s_0) = \mathcal{N}\big(s_t;\, \alpha_t\, s_0,\, \sigma_t^2 I\big), \tag{1}$$

for a noise schedule (α_t, σ_t) satisfying $\alpha_0 = 1$, $\sigma_0 = 0$, $\sigma_t > 0$ for at least one diffusion time $t \in (0, T]$, and the boundary $\alpha_T \to 0$, $\sigma_T \to 1$ so that $q_T(\cdot \mid s_0) \approx \mathcal{N}(0, I)$. The reverse (denoising) process is a Markov chain with a parameterized Gaussian transition,

$$p_\theta(s_{t-1} \mid s_t) = \mathcal{N}\big(s_{t-1};\, \mu_\theta(s_t, t),\, \Sigma_t\big), \tag{2}$$

where the mean μ_θ is parameterized by a neural network. Following Ho et al. (2020), we train the reverse process by predicting the forward noise rather than the mean, using the simplified loss

$$\mathcal{L}(\theta) \;=\; \mathbb{E}_{s_0\sim\rho,\, t\sim\mathcal{U}\{1,\dots,T\},\, \varepsilon\sim\mathcal{N}(0,I)} \big\|\varepsilon - \varepsilon_\theta(s_t, t)\big\|^2. \tag{3}$$

What this denoising objective teaches the network is not the density itself: as we formalize below, it is the *score* of the noise-perturbed data distribution.

3.3 THE SCORE FUNCTION AND FISHER DIVERGENCE

Let p be a differentiable density, strictly positive on $\mathbb{R}^d$. Its *score function* is the gradient of the log-density,

$$x \;\mapsto\; \nabla \log p(x) \in \mathbb{R}^d, \tag{4}$$

a vector field pointing in the direction of steepest ascent of the log-density. Since $\log p$ is recovered from its score by line integration up to an additive constant, the score characterizes the distribution as fully as the density itself. Score-based generative models build on this characterization: the drift of the reverse process equation 2 at each noise level depends on the diffused marginal only through its score (Song et al., 2021), so generating from the model amounts to evaluating scores.

The score of a diffused distribution is available through the noise, a relation known as Tweedie's formula. For any distribution ρ over states, let ρ_t denote the marginal of s_t when $s_0 \sim \rho$ is diffused by equation 1; then

$$\nabla \log \rho_t(x) \;=\; -\,\mathbb{E}\big[\varepsilon \mid s_t = x\big] \,/\, \sigma_t, \tag{5}$$

so the conditional mean of the added noise is, up to scale, the negative score. Substituting equation 5 into equation 3 and writing $s_\theta(x, t) := -\,\varepsilon_\theta(x, t) \,/\, \sigma_t$ for the induced score estimate, the denoising loss is revealed to be denoising score matching: for each fixed diffusion time t,

$$\mathbb{E}_{s_0\sim\rho,\, \varepsilon\sim\mathcal{N}(0,I)} \big\|\varepsilon - \varepsilon_\theta(s_t, t)\big\|^2 \;=\; \sigma_t^2\, \mathbb{E}_{s_t\sim\rho_t} \big\|s_\theta(s_t, t) - \nabla \log \rho_t(s_t)\big\|^2 \;+\; \mathrm{const}(t), \tag{6}$$

where $\mathrm{const}(t)$ does not depend on θ. A network trained by equation 3 therefore provides an estimate of the score at every noise level. Fitted on the expert demonstrations by equation 3, the resulting pre-trained network is the *score network*

$$s_\psi(x,t) \;:=\; -\,\varepsilon_\psi(x,t)\,/\,\sigma_t \;\approx\; \nabla \log \rho_{E,t}(x), \tag{7}$$

an estimate of the score of the t-diffused expert marginal $\rho_{E,t}$. Throughout the paper we assume that s_ψ satisfies the bounded-error condition equation 27. The score network is used *only* as a distributional prior on the expert and is never trained during imitation.

Since a distribution is characterized by its score, a natural way to compare two distributions is to compare their score functions pointwise. This is precisely the classical *Fisher divergence*. For two distributions p and q over $\mathbb{R}^d$, it is defined as the expected squared norm of the score difference,

$$D_F(q \,\|\, p) := \mathbb{E}_{x\sim q}\big\|\nabla \log p(x) - \nabla \log q(x)\big\|^2. \tag{8}$$

The Fisher divergence is a strong divergence: if $D_F(q\|p)$ is small, then the KL divergence, Hellinger distance, and total variation distance between q and p are all small (under mild regularity conditions; see Chen et al. (2023)). Our method applies this divergence to the diffused learner and expert occupancies across all noise levels (Section 4). Throughout, we assume the diffused densities are strictly positive and smooth with sufficiently fast boundary decay, so that integration by parts is valid and the relevant scores are square-integrable under the corresponding distributions.

4 Method: Diffusion-Guided Adversarial Imitation

4.1 Score-Guided Distribution Matching over Diffused States

Matching two distributions can be reduced to matching their score functions over diffused states. Imitation can therefore be cast as the search for a policy whose diffused state distribution matches the expert's across all noise scales,

$$\inf_{\pi\in\Pi} D_F(\rho_\pi \,\|\, \rho_E) \;:=\; \inf_{\pi\in\Pi} \mathbb{E}_{s\sim\rho_\pi}\, \mathbb{E}_{t\sim\mathcal{U}([0,T])}\, \mathbb{E}_{s_t\sim q_t(\cdot|s)}\big\|\nabla \log \rho_{E,t}(s_t) - \nabla \log \rho_{\pi,t}(s_t)\big\|_2^2, \tag{9}$$

where D_F is the *diffusion Fisher divergence* between the learner's and expert's *diffused* state distributions and $q_t(\cdot \mid s)$ is the forward diffusion kernel of equation 1 (variance-preserving Gaussian noise of scale t applied to the clean state s); in practice the expert score $\nabla \log \rho_{E,t}$ is provided by a pre-trained score network $s_\psi(s_t,t) \approx \nabla \log \rho_{E,t}$. Minimizing D_F matches the learner's distribution to the expert's across noise scales, inheriting the robustness of multi-scale score matching.

Why a direct score-matching cost fails. The first strategy one might consider is to use the divergence directly as a cost and optimize it by policy gradient. Consider the per-step reward given by the negative squared score gap at each diffused state, in which the expert score $s_{E,t}$ is replaced in practice by the pre-trained s_ψ,

$$r_\theta(x_t,t) \;=\; -\big\|s_{\pi_\theta,t}(x_t) - s_{E,t}(x_t)\big\|_2^2, \qquad s_{\pi_\theta,t}(x) \;:=\; \nabla_x \log \rho_{\pi_\theta,t}(x). \tag{10}$$

The trajectory return $R_\theta(\tau) := \frac{1}{T}\sum_{t=1}^{T} r_\theta(x_t,t)$, which averages the per-step reward over the diffusion steps, induces the objective $J(\theta) = \mathbb{E}_{\tau\sim p_\theta}[R_\theta(\tau)] = -D_F(\rho_\pi \,\|\, \rho_E)$. The gradient then splits into two terms,

$$\nabla_\theta J \;=\; \underbrace{\mathbb{E}_{\tau\sim p_\theta}\big[R_\theta(\tau)\,\nabla_\theta \log p_\theta(\tau)\big]}_{g_{\mathrm{PG}}} \;+\; \underbrace{\mathbb{E}_{\tau\sim p_\theta}\big[\nabla_\theta R_\theta(\tau)\big]}_{g_{\mathrm{dep}}}. \tag{11}$$

The first term is the standard policy-gradient term. Substituting equation 10 into the second term gives

$$g_{\mathrm{dep}} \;=\; \mathbb{E}_{\tau\sim p_\theta}\big[\nabla_\theta R_\theta(\tau)\big] \;=\; -\frac{2}{T}\sum_{t=1}^{T} \mathbb{E}_{x_t\sim\rho_{\pi_\theta,t}}\Big[\big(s_{\pi_\theta,t}(x_t) - s_{E,t}(x_t)\big)^\top \nabla_\theta s_{\pi_\theta,t}(x_t)\Big], \tag{12}$$

where $\rho_{\pi_\theta,t}$ is the marginal of x_t under $p_\theta(\tau)$. The remaining derivative is a mixed parameter–state Hessian of the log diffused occupancy,

$$\nabla_\theta s_{\pi_\theta,t}(x) \;=\; \nabla_\theta \nabla_x \log \rho_{\pi_\theta,t}(x), \tag{13}$$

which no trajectory-level estimator provides. A standard policy-gradient update that keeps only g_{PG} is therefore biased, with bias exactly g_{dep}. The direct score-matching cost of equation 9 therefore cannot be optimized by off-the-shelf RL: it is π-dependent through the learner score, and one must explicitly account for how that score changes with the policy.

We resolve this difficulty by deriving an *adversarial variational form* of this divergence, in which integration by parts removes the learner's score from the cost: the resulting objective depends on the policy only through the sampled states, and is therefore amenable to standard policy-gradient optimization. Our derivation relies on two elementary identities, which we state first. Throughout, we assume densities are smooth and vanish sufficiently fast at the boundary so that the relevant integrals are well-defined.

Lemma 1. *For a probability density $q(x)$ and a vector field $f(x)$, assuming the product $q(x)f(x)$ vanishes at the boundary of the domain,*

$$\mathbb{E}_q\big[\nabla \log q(x)^\top f(x)\big] = -\mathbb{E}_q\big[\nabla \cdot f(x)\big]. \tag{14}$$

Proof. See Appendix A.

Lemma 2. *For any vector $a \in \mathbb{R}^d$,*

$$\|a\|^2 = \sup_{f \in \mathbb{R}^d} \big(2f^\top a - \|f\|^2\big), \tag{15}$$

with the supremum attained at $f = a$.

Proof. See Appendix A.

Applying Lemma 2 pointwise inside the expectation, and assuming the pointwise maximizer $f^\star(x_t, t) := \nabla \log \rho_{E,t}(x_t) - \nabla \log \rho_{\pi,t}(x_t)$ lies in $\mathcal{F}$ (the precise condition is given in Appendix A.1), we obtain

$$\begin{aligned} D_F(\rho_\pi \,\|\, \rho_E) &= \sup_{f \in \mathcal{F}} \mathbb{E}_{s \sim \rho_\pi, t, s_t}\Big[2f(x_t,t)^\top \big(\nabla \log \rho_{E,t}(x_t) - \nabla \log \rho_{\pi,t}(x_t)\big) - \|f(x_t,t)\|^2\Big] \\ &= \sup_{f \in \mathcal{F}} \mathbb{E}_{s \sim \rho_\pi, t, s_t}\Big[2f(x_t,t)^\top \nabla \log \rho_{E,t}(x_t) + 2\,\nabla \cdot f(x_t,t) - \|f(x_t,t)\|^2\Big], \end{aligned} \tag{16}$$

where the second line applies Lemma 1: integration by parts trades the learner-score term for the divergence of the discriminator, $-2\,\mathbb{E}_{x_t \sim \rho_{\pi,t}}[f^\top \nabla \log \rho_{\pi,t}] = 2\,\mathbb{E}_{x_t \sim \rho_{\pi,t}}[\nabla \cdot f]$. We state this formally as follows.

Proposition 1 (Adversarial Variational Form of the Diffusion Fisher Divergence)**.** *Assume the function class $\mathcal{F}$ contains the score difference $x_t \mapsto \nabla \log \rho_{E,t}(x_t) - \nabla \log \rho_{\pi,t}(x_t)$ (Appendix A.1). Then for every policy π,*

$$D_F(\rho_\pi \,\|\, \rho_E) = \sup_{f \in \mathcal{F}} \mathbb{E}_{s \sim \rho_\pi, t, s_t}\Big[2f(x_t,t)^\top \nabla \log \rho_{E,t}(x_t) + 2\,\nabla \cdot f(x_t,t) - \|f(x_t,t)\|^2\Big]. \tag{17}$$

Proof. See Appendix A. ■

Taking the infimum over policies on both sides of equation 17 turns the imitation objective equation 9 into the DGAIL game,

$$\inf_{\pi \in \Pi} D_F(\rho_\pi \,\|\, \rho_E) \;=\; \inf_{\pi \in \Pi} \sup_{f \in \mathcal{F}} \mathbb{E}_{s \sim \rho_\pi,\, t,\, s_t}\Big[2f(x_t,t)^\top \nabla \log \rho_{E,t}(x_t) + 2\,\nabla \cdot f(x_t,t) - \|f(x_t,t)\|^2\Big]. \tag{18}$$

Crucially, the integrand of equation 18 no longer depends on the policy through the learner score: it involves only (i) the expert score $\nabla \log \rho_{E,t}$, provided by the pre-trained diffusion model s_ψ, and (ii) the divergence of the learned vector field f. The policy enters only through the sampled distribution, so the anomalous term in equation 11 vanishes and standard policy-gradient estimators apply.

4.2 Training Objective and Algorithm

In practice, we work with a discrete set of diffusion times $t = 1, \dots, T$, which yields the discrete-time diffusion Fisher divergence, the uniform average of the per-level score errors over the T noise levels (the discrete counterpart of equation 9):

$$D_F(\rho_\pi \,\|\, \rho_E) \;:=\; \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{s\sim\rho_\pi}\, \mathbb{E}_{s_t\sim q_t(\cdot|s)} \big\|\nabla \log \rho_{E,t}(s_t) - \nabla \log \rho_{\pi,t}(s_t)\big\|^2, \tag{19}$$

and the population adversarial objective of Prop. 1

$$\mathcal{J}(\pi, f) \;:=\; \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{x_t\sim\rho_{\pi,t}} \Big[2f(x_t,t)^\top s_{E,t}(x_t) + 2\,\nabla_{x_t}\cdot f(x_t,t) - \|f(x_t,t)\|_2^2\Big], \tag{20}$$

where $s_{E,t}$ denotes the true expert score and $x_t \sim \rho_{\pi,t}$ the diffused learner state at time t. In practice, we replace the expert score with the pre-trained network s_ψ, yielding the practical objective

$$\mathcal{J}_\psi(\pi, f) \;:=\; \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{x_t\sim\rho_{\pi,t}} \Big[2f(x_t,t)^\top s_{\psi}(x_t,t) + 2\,\nabla_{x_t}\cdot f(x_t,t) - \|f(x_t,t)\|_2^2\Big]. \tag{21}$$

The discriminator is updated by maximizing equation 21, while the policy minimizes the resulting adversarial value through the dual reward introduced below; this two-player structure is inherited from the variational form of Prop. 1.

We solve equation 21 by alternating between discriminator and policy updates; the full procedure is summarized in Algorithm 1 in Appendix F. The vector field $f_\phi(x,t)$ is parameterized by a neural network. For the discriminator update, the objective involves the divergence $\nabla \cdot f_\phi(x_t,t)$, which is expensive to compute exactly in high dimensions. Note that the divergence of a vector field equals the trace of its Jacobian matrix,

$$\nabla \cdot f_\phi(x,t) \;=\; \mathrm{tr}\big(J_{f_\phi}(x,t)\big), \qquad J_{f_\phi}(x,t) := \nabla_x f_\phi(x,t) \in \mathbb{R}^{d\times d}, \tag{22}$$

so computing it exactly would require evaluating all d diagonal entries of the Jacobian. Hutchinson's estimator (Hutchinson, 1990) provides an unbiased estimate of a trace from a single vector–Jacobian product,

$$\mathrm{tr}\big(J\big) \;=\; \mathbb{E}_{\xi\sim\mathcal{N}(0,I)}\big[\xi^\top J\xi\big] \;=\; \mathbb{E}_{\xi\sim\mathcal{N}(0,I)}\big[\xi^\top \nabla_x\big(f_\phi(x,t)^\top\xi\big)\big], \tag{23}$$

where the second equality follows because $\nabla_x(f_\phi^\top\xi) = J^\top\xi$ and $\mathrm{tr}(J) = \mathbb{E}_\xi[\xi^\top J\xi]$ since $\mathbb{E}_\xi[\xi\xi^\top] = I$. This makes the divergence term as cheap as a standard backward pass of the vector-field network. The computational cost and the effect of the number of Hutchinson vectors on estimator stability are discussed in Appendix C.3.

With this estimator, we define the *score-guided operator*

$$T_\phi(x_t,t) \;=\; 2\Big(\xi^\top\nabla_{x_t}\big(f_\phi(x_t,t)^\top\xi\big) + f_\phi(x_t,t)^\top s_\psi(x_t,t)\Big) - \|f_\phi(x_t,t)\|^2, \tag{24}$$

which is the Monte Carlo estimate of the integrand in equation 21 with the divergence replaced by its Hutchinson estimator (23): the first term estimates $2\nabla\cdot f_\phi$, the second couples the field with the expert score s_ψ (the score-guided signal), and the third is a quadratic regularizer inherited from the variational form. The discriminator is updated by maximizing $\mathcal{L}(\phi) = \mathbb{E}[T_\phi(x_t,t)]$.

For the policy update, the *dual reward* is the negative of the (time-averaged) discriminator objective,

$$r(x) \;=\; -\mathbb{E}_{t,\xi}\big[T_\phi(x_t,t)\big], \tag{25}$$

and the policy is updated with any off-the-shelf RL algorithm (e.g., PPO or SAC) to maximize cumulative dual rewards. Intuitively, the discriminator learns the direction of maximal distribution mismatch, and the policy is pushed away from states where the expert score disagrees with the learner's.

4.3 Theoretical Analysis

In this section we demonstrate that (i) minimizing the diffusion Fisher divergence is equivalent to matching the expert occupancy (Thm. 1); (ii) the game value attained by *any* policy certifies a bound on its true diffusion Fisher divergence, up to the three approximation components specific to DGAIL (Thm. 2), which yields a bound for the learned policy in which the residual optimization error appears as one further explicit term (Cor. 1); and (iii) this divergence guarantee transfers to an instance-dependent bound on the imitation gap under any bounded evaluation cost (Thm. 3).

Theorem 1. *For any policy $\pi \in \Pi$,*

$$D_F(\rho_\pi \,\|\, \rho_E) = 0 \iff \rho_\pi = \rho_E. \tag{26}$$

The detailed proof is presented in Appendix B.

By Proposition 1, the global minimum of the population DGAIL objective is attained exactly by policies matching the expert occupancy, making the variational form meaningful as an imitation objective.

The practical algorithm, however, does not have direct access to the ideal population objective, and we analyze how each of the resulting approximation errors propagates into the bound on the divergence achieved by the learned policy.

The first source is the approximation of the expert score by the pre-trained network s_ψ. For a policy π, write $f^\star_\pi(x,t) := s_{E,t}(x) - s_{\pi,t}(x)$ for the ideal variational maximizer; the score-approximation error at π is

$$\epsilon_{\text{score}}(\pi)^2 \;:=\; \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{x_t\sim\rho_{\pi,t}} \big\| s_\psi(x_t,t) - s_{E,t}(x_t) \big\|^2. \tag{27}$$

The second source is the restriction of the discriminator to the parametric class $\mathcal{F}$, whose approximation error and magnitude at π are

$$\epsilon_{\mathcal{F}}(\pi)^2 \;:=\; \inf_{f\in\mathcal{F}} \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{x_t\sim\rho_{\pi,t}} \big\| f(x_t,t) - f^\star_\pi(x_t,t) \big\|^2, \tag{28}$$

$$B_f(\pi)^2 \;:=\; \sup_{f\in\mathcal{F}} \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{x_t\sim\rho_{\pi,t}} \|f(x_t,t)\|^2. \tag{29}$$

We assume these three quantities are finite at the occupancies at which they are invoked. The third source is the stochastic estimation of the divergence with Hutchinson's estimator, whose uniform estimation error $\epsilon_{\text{H}}(M,\delta)$ is quantified in Theorem 2. The Jacobian $J_f(x,t) := \nabla_x f(x,t)$ is likewise assumed to satisfy, at those occupancies,

$$\sup_{f\in\mathcal{F}} \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{x_t\sim\rho_{\pi,t}} \big\| J_f(x_t,t) \big\|_F^2 \;\le\; B_J^2, \qquad \sup_{f\in\mathcal{F}} \sup_{x,t} \big\| J_f(x,t) \big\|_{\text{op}} \;\le\; B_{\text{op}}. \tag{30}$$

Theorem 2. *Let $\epsilon_{\text{H}}(M,\delta)$ denote a uniform (over $f \in \mathcal{F}$) estimation error for the stochastic divergence estimator constructed with M independent Hutchinson vectors, with failure probability at most δ, and let $\widetilde{G}_{\psi,M}(\pi) := \sup_{f\in\mathcal{F}} \widetilde{\mathcal{J}}_{\psi,M}(\pi,f)$ denote the attained game value, i.e. the practical objective equation 21 with the divergence replaced by its Hutchinson estimate. Then for* every *policy π, with probability at least $1-\delta$,*

$$D_F(\rho_\pi \,\|\, \rho_E) \;\le\; \widetilde{G}_{\psi,M}(\pi) \;+\; 2B_f(\pi)\,\epsilon_{\text{score}}(\pi) + \epsilon_{\mathcal{F}}(\pi)^2 + \epsilon_{\text{H}}(M,\delta). \tag{31}$$

The three additive terms are exactly the three approximation sources intrinsic to DGAIL: expert-score approximation, vector-field approximation, and stochastic divergence estimation.

The detailed proof is presented in Appendix C.

Theorem 2 says that the attained game value is a *certificate* for the quantity we actually care about: whatever the policy optimizer did, the true diffusion Fisher divergence of the resulting policy is bounded by the value it attains plus three terms that do not depend on the optimizer. Remark 1 discusses what is additionally needed to read this certificate off a training run. Specializing Theorem 2 to the learned policy and bounding the attained game value by that of the expert comparator turns the residual suboptimality of the policy optimizer into one further explicit term.

Corollary 1. *Let $\hat{\pi}$ be the policy returned by the training procedure and let*

$$\epsilon_{\text{opt}} \;:=\; \widetilde{G}_{\psi,M}(\hat{\pi}) \;-\; \inf_{\pi\in\Pi} \widetilde{G}_{\psi,M}(\pi) \;\geq\; 0 \tag{32}$$

denote its optimization error on the practical objective. Assume the expert occupancy is realizable, $\pi_E \in \Pi$ with $\rho_{\pi_E} = \rho_E$. Then, with probability at least $1-\delta$,

$$D_F(\rho_{\hat{\pi}} \,\|\, \rho_E) \;\leq\; \epsilon_{\text{DGAIL}} \;:=\; 2B_f(\hat{\pi})\,\epsilon_{\text{score}}(\hat{\pi}) + \epsilon_{\text{score}}(\pi_E)^2 + \epsilon_{\mathcal{F}}(\hat{\pi})^2 + 2\epsilon_{\text{H}}(M,\delta) + \epsilon_{\text{opt}}. \tag{33}$$

The score-approximation error enters twice and at two different occupancies: to first order at the learner, where the pre-trained network is queried during training, and to second order at the expert, where it measures the residual value of the game at the comparator. If the expert occupancy is not realizable by Π, the expert-side term is replaced by a projection term and the bound becomes

$$D_F(\rho_{\hat{\pi}}\|\rho_E) \leq \inf_{\pi\in\Pi} \Big\{ 2D_F(\rho_\pi\|\rho_E) + 2\epsilon_{\text{score}}(\pi)^2 \Big\} + 2B_f(\hat{\pi})\,\epsilon_{\text{score}}(\hat{\pi}) + \epsilon_{\mathcal{F}}(\hat{\pi})^2 + 2\epsilon_{\text{H}}(M,\delta) + \epsilon_{\text{opt}}, \tag{34}$$

the infimum being over the sum, since a comparator is good only if it is simultaneously close to the expert occupancy and well covered by the pre-trained score network.

Corollary 2. *If $\epsilon_{\text{score}} \to 0$, $\epsilon_{\mathcal{F}} \to 0$, $\epsilon_{\text{H}}(M,\delta) \to 0$ and $\epsilon_{\text{opt}} \to 0$ along a sequence of training problems, then $D_F(\rho_{\hat{\pi}}\|\rho_E) \to 0$, and by Theorem 1, $\rho_{\hat{\pi}} \to \rho_E$. Hence DGAIL is consistent for expert occupancy matching as the approximation sources intrinsic to its formulation vanish. In particular, for any fixed confidence level δ, $\epsilon_{\text{H}}(M,\delta) \to 0$ as $M \to \infty$ at the rate derived in Appendix C.3, so the stochastic divergence error can be reduced by increasing the number of Hutchinson vectors.*

Corollary 1 bounds a divergence between *diffused* occupancies, which is not by itself a statement about imitation. We close the analysis by converting it into a guarantee on the quantity of interest, the gap in expected cost under an unknown evaluation cost, using the stability of the reverse diffusion dynamics. Concretely, the diffusion Fisher divergence controls the Hellinger distance between the learner and expert occupancies, and the Hellinger distance in turn controls the cost expectation gap.

Theorem 3. *Write*

$$\epsilon_{\text{occ}} \;:=\; C_{\text{diff}}\,\epsilon_{\text{DGAIL}} + \epsilon_{\text{mix}}, \tag{35}$$

where ϵ_{DGAIL} is the divergence error of Corollary 1, equation 33, which includes the optimization error ϵ_{opt} of the policy optimizer, $\epsilon_{\text{mix}} = O\big((m+d)e^{-T}\big)$ is the diffusion mixing error, and $C_{\text{diff}} = O(T)$ depends on the diffusion time horizon and the score regularity. Assume in addition that the diffused scores are Lipschitz, that the occupancy second moment is bounded by m, and that the score discrepancy is integrable in the diffusion time. Then for any *bounded evaluation cost $c : \mathcal{S} \to [0, C_{\max}]$, with probability at least $1-\delta$,*

$$\big|\mathbb{E}_{\rho_{\hat{\pi}}}[c] - \mathbb{E}_{\rho_E}[c]\big| \;\leq\; \min\Big\{ C_{\max},\;\; c_1\sqrt{\text{Var}_{\rho_E}(c)\,\epsilon_{\text{occ}}} + c_2\,C_{\max}\,\epsilon_{\text{occ}} \Big\}, \tag{36}$$

where c_1, c_2 are constants.

The detailed proof is presented in Appendix D.

5 Experimental Validation

This section validates our theoretical findings on continuous-control imitation; implementation details are provided in Appendix F, and the code is available at `https://anonymous.4open.science/r/fisheradv-7736/`.

5.1 Experimental Set-up

We evaluate our method on five MuJoCo tasks: HalfCheetah-v2, Hopper-v2, Walker2d-v2, Ant-v2, and Humanoid-v2. Each algorithm is evaluated over five random seeds. We consider three categories of baselines. The first is behavioral cloning (BC) (Pomerleau, 1991). For AIL, we include GAIL (Ho & Ermon, 2016), DAC (Kostrikov et al., 2019), DiffAIL (Wang et al., 2024), and DRAIL (Lai et al., 2024). For Q-based IL, we include ValueDICE (Kostrikov et al., 2021), IQ-Learn (Garg et al., 2021), Reg IQ-Learn (the official implementation with the additional regularizer), and LS-IQ (Al-Hafez et al., 2023). The expert policy π_E is included as an upper-bound reference.

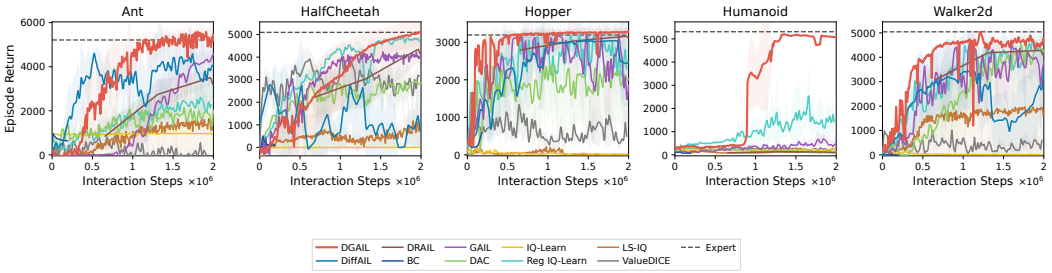


Figure 1: Training curves on five MuJoCo tasks (mean $\pm$ std over seeds). The dashed line denotes the expert return.

5.2 Experimental Results

Results on MuJoCo Tasks. Figure 1 shows the learning curves of DGAIL and the AIL and Q-based baselines on the five MuJoCo tasks; we draw three conclusions. First, DGAIL matches or approaches the expert return on all five tasks, confirming that the variational form of Prop. 1 supports policy optimization as effectively as classical adversarial objectives. Second, DGAIL consistently outperforms the Q-based baselines (IQ-Learn, Reg IQ-Learn, LS-IQ, ValueDICE) on the four harder tasks (Hopper, Walker2d, Ant, Humanoid), the practical benefit of the explicit Fisher signal over learned critics. On Humanoid, the highest-dimensional task, DGAIL stays at the random-policy level until roughly 0.85M steps and then transitions abruptly to a plateau close to the expert, while Reg IQ-Learn, the strongest baseline on this task, stays below half of the expert return throughout, plausibly because the policy-independent dual reward equation 25 remains informative where critics built on learner–expert overlap saturate. On Ant, no Q-based method reaches half of the expert return, whereas DGAIL crosses it around 0.6M steps and stays at or above it. Third, compared with the AIL baselines GAIL and DiffAIL, DGAIL trains more stably: GAIL reaches the expert return only intermittently and never on Humanoid, and DiffAIL, although often the fastest early learner, collapses midway on Walker2d or plateaus below the expert with sustained oscillations on Ant and HalfCheetah, whereas DGAIL converges to the expert return on all five tasks.

5.3 Additional Results

Data efficiency. On Ant, HalfCheetah, and Walker2d, DGAIL is already close to the expert with a single trajectory, whereas BC and IQ-Learn remain substantially below the expert across the budget range of $\{1, 3, 5\}$ trajectories, consistent with the distribution-matching guarantee of Theorem 2 (Figure 4 in Appendix E.3).

Divergence estimator. A natural concern is that the stochastic divergence estimate in equation 24, which enters both the discriminator objective and the dual reward, requires extensive sampling. We give empirical guarantees (Appendices E.1 and E.2), validating the theoretical analysis of Appendix C.3: the certificate of Theorem 2 remains valid for any number of probes, and training is essentially unchanged with a small probe budget. The computational overhead is therefore acceptable.

6 Conclusion

We introduced the adversarial variational form of the Fisher divergence and showed that it makes the explicit score-difference signal tractable for imitation learning: via integration by parts, the learner's score is removed from the discriminator's objective. The resulting reward is π-independent, so the policy remains trainable by unbiased policy gradient, and the minimax objective requires only a pre-trained expert score network and a divergence-estimated vector field. We proved that the game value DGAIL attains during training certifies a bound on the true divergence, which transfers to an instance-dependent imitation gap under any bounded evaluation cost. Empirically, DGAIL matches or outperforms strong AIL baselines on MuJoCo continuous control.

Reproducibility Statement

All theoretical results stated in the main text are proved in full in Appendices A–F, including the adversarial variational identity (Prop. 1), the identifiability theorem (Thm. 1), and the objective-level approximation guarantee (Thm. 2, Cor. 1) with all auxiliary lemmas and assumptions. A summary of the training procedure is given in Algorithm 1, the experimental setup is described in Section 5.1, and complete hyperparameters are reported in Appendix F. The source code will be released upon publication to facilitate exact reproduction of all experiments.

AI Use Statement

In this work, we did not use generative AI tools for any research subtask for which disclosure is required by the ICLR 2027 AI Policy for Authors: the theoretical analysis, the proofs, and the design of the methodology and experiments were carried out entirely by the authors. We used generative AI tools only for tasks for which disclosure is recommended, namely minor code debugging, polishing the presentation of the text, and assisting with the LaTeX formatting of the manuscript. We have reviewed all AI-assisted work and take full responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

Ethics Statement

This work studies imitation learning in simulated continuous-control environments (MuJoCo). It does not involve human subjects or personal data, and we are not aware of direct negative societal consequences. Because learned imitation policies inherit the behavior of their demonstrations, deployable applications should audit the expert data for unsafe or biased behaviors; the reliance on a pre-trained expert score network also means that its failure modes are inherited by the learned policy.

References

Pieter Abbeel and Andrew Y. Ng. Apprenticeship learning via inverse reinforcement learning. In *International Conference on Machine Learning (ICML)*, 2004.

Anurag Ajay, Yilun Du, Abhi Gupta, Joshua B. Tenenbaum, Tommi S. Jaakkola, and Sergey Levine. Is conditional generative modeling all you need for decision-making? In *International Conference on Learning Representations (ICLR)*, 2023.

Firas Al-Hafez, Davide Tateo, Oleg Arenz, Guoping Zhao, and Jan Peters. LS-IQ: Implicit reward regularization for inverse reinforcement learning. In *International Conference on Learning Representations (ICLR)*, 2023.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning (ICML)*, 2017.

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, et al. π_0: A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024.

Jonathan D. Chang, Dhruv Sreenivas, Yingbing Huang, Kianté Brantley, and Wen Sun. Adversarial imitation learning via boosting. In *International Conference on Learning Representations (ICLR)*, 2024.

Sitan Chen, Sinho Chewi, Jerry Li, Yuanzhi Li, Adil Salim, and Anru R Zhang. Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions. In *International Conference on Learning Representations (ICLR)*, 2023.

Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin Burchfiel, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. In *Robotics: Science and Systems (RSS)*, 2023.

Robert Dadashi, Léonard Hussenot, Matthieu Geist, and Olivier Pietquin. Primal Wasserstein imitation learning. In *International Conference on Learning Representations (ICLR)*, 2021.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Justin Fu, Katie Luo, and Sergey Levine. Learning robust rewards with adversarial inverse reinforcement learning. In *International Conference on Learning Representations (ICLR)*, 2018.

Divyansh Garg, Shuvam Chakraborty, Chris Cundy, Jiaming Song, and Stefano Ermon. IQ-Learn: Inverse soft-Q learning for imitation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

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 (ICML)*, 2018.

Jonathan Ho and Stefano Ermon. Generative adversarial imitation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

M. F. Hutchinson. A stochastic estimator of the trace of the influence matrix for Laplacian smoothing splines. *Communications in Statistics – Simulation and Computation*, 19(2):433–450, 1990.

Michael Janner, Yilun Du, Joshua B. Tenenbaum, and Sergey Levine. Planning with diffusion for flexible behavior synthesis. In *International Conference on Machine Learning (ICML)*, 2022.

Ilya Kostrikov, Kumar Krishna Agrawal, Debidatta Dwibedi, Sergey Levine, and Jonathan Tompson. Discriminator-actor-critic: Addressing sample inefficiency and reward bias in adversarial imitation learning. In *International Conference on Learning Representations (ICLR)*, 2019.

Ilya Kostrikov, Ofir Nachum, and Jonathan Tompson. Imitation learning via off-policy distribution matching. In *International Conference on Learning Representations (ICLR)*, 2021.

Chun-Mao Lai, Hsiang-Chun Wang, Ping-Chun Hsieh, Yu-Chiang Frank Wang, Min-Hung Chen, and Shao-Hua Sun. Diffusion-reward adversarial imitation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Tianjiao Luo, Tim Pearce, Huayu Chen, Jianfei Chen, and Jun Zhu. C-GAIL: Stabilizing generative adversarial imitation learning with control theory. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Takayuki Osa, Joni Pajarinen, Gerhard Neumann, J. Andrew Bagnell, Pieter Abbeel, and Jan Peters. An algorithmic perspective on imitation learning. *Foundations and Trends in Robotics*, 7(1–2): 1–179, 2018.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Dean A. Pomerleau. Efficient training of artificial neural networks for autonomous navigation. *Neural Computation*, 3(1):88–97, 1991.

Stéphane Ross and Drew Bagnell. Efficient reductions for imitation learning. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2010.

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

Bingzheng Wang, Yan Zhang, Teng Pang, Guoqiang Wu, and Yilong Yin. DiffAIL: Diffusion adversarial imitation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2024.

Zhendong Wang, Jonathan J. Hunt, and Mingyuan Zhou. Diffusion policies as an expressive policy class for offline reinforcement learning. In *International Conference on Learning Representations (ICLR)*, 2023.

Runzhe Wu, Yiding Chen, Gokul Swamy, Kiante Brantley, and Wen Sun. Diffusing states and matching scores: A new framework for imitation learning. In *International Conference on Learning Representations (ICLR)*, 2025.

Huang Xiao, Michael Herman, Joerg Wagner, Sebastian Ziesche, Jalal Etesami, and Thai Hong Linh. Wasserstein adversarial imitation learning. *arXiv preprint arXiv:1906.08113*, 2019.

Tian Xu, Ziniu Li, and Yang Yu. Error bounds of imitating policies and environments. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Tian Xu, Zhilong Zhang, Ruishuo Chen, Yihao Sun, and Yang Yu. Provably and practically efficient adversarial imitation learning with general function approximation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Brian D. Ziebart, Andrew Maas, J. Andrew Bagnell, and Anind K. Dey. Maximum entropy inverse reinforcement learning. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2008.

A Proof of Proposition 1: Adversarial Variational Form

A.1 Exchanging Expectation and Supremum

The identity in Proposition 1 exchanges the expectation over (s, t, s_t) with the supremum over $\mathcal{F}$. For each (x_t, t), define the pointwise quantity

$$\ell_f(x_t, t) \;:=\; 2f(x_t, t)^\top \big(\nabla \log \rho_{E,t}(x_t) - \nabla \log \rho_{\pi,t}(x_t)\big) - \|f(x_t, t)\|^2.$$

By Lemma 2, for every (x_t, t) the pointwise supremum is attained at $f^\star(x_t, t) := \nabla \log \rho_{E,t}(x_t) - \nabla \log \rho_{\pi,t}(x_t)$, giving $\sup_{f\in\mathcal{F}} \ell_f(x_t, t) = \|\nabla \log \rho_{E,t}(x_t) - \nabla \log \rho_{\pi,t}(x_t)\|^2$ when $f^\star \in \mathcal{F}$. The exchange then follows from a one-line argument: the pointwise maximizer $f^\star$ lies in $\mathcal{F}$, hence

$$\sup_{f\in\mathcal{F}} \mathbb{E}[\ell_f] \;\geq\; \mathbb{E}[\ell_{f^\star}] \;=\; \mathbb{E}\big[\sup_{f\in\mathcal{F}} \ell_f\big] \;=\; D_F(\rho_\pi \,\|\, \rho_E),$$

while the standard "sup outside $\mathbb{E}$" bound gives the reverse inequality $\sup_{f\in\mathcal{F}} \mathbb{E}[\ell_f] \;\leq\; \mathbb{E}[\sup_{f\in\mathcal{F}} \ell_f]$. Combining the two directions yields the equality in equation 17. No additional structure on $\mathcal{F}$ (e.g. linearity or convexity of the objective in f) is needed.

A.2 Proof of Lemma 1 (Integration by Parts)

Proof. Expanding the expectation in terms of the density and using the identity $\nabla \log q(x) = \nabla q(x)/q(x)$, we have

$$\mathbb{E}_q\big[\nabla \log q(x)^\top f(x)\big] = \int q(x) \Big(\frac{\nabla q(x)}{q(x)}\Big)^\top f(x)\, dx = \int \nabla q(x)^\top f(x)\, dx.$$

It remains to rewrite $\int \nabla q(x)^\top f(x)\, dx$, which cannot be evaluated directly since q is arbitrary. The key tool is the divergence theorem (Gauss's theorem): for any vector field $F(x)$ on a domain Ω with smooth boundary $\partial\Omega$ and outward unit normal $n(x)$,

$$\int_\Omega \nabla \cdot F(x)\, dx \;=\; \oint_{\partial\Omega} F(x) \cdot n(x)\, dS.$$

We apply it to the vector field $F(x) = q(x)f(x)$ over $\Omega = \mathbb{R}^d$. The boundary integral $\oint_{\partial\mathbb{R}^d} q(x)f(x) \cdot n(x)\, dS$ vanishes because $q(x)f(x)$ is assumed to vanish sufficiently fast at infinity,

so the total flux of qf through the enclosing surface is zero; hence $\int_{\mathbb{R}^d} \nabla \cdot \big(q(x)f(x)\big)\, dx = 0$. Expanding the integrand by the product rule $\nabla \cdot (qf) = \nabla q^\top f + q\, \nabla \cdot f$,

$$\int \nabla q(x)^\top f(x)\, dx = -\int q(x)\, \nabla \cdot f(x)\, dx.$$

Substituting this back and recognizing the right-hand side as an expectation under q,

$$\mathbb{E}_q\big[\nabla \log q(x)^\top f(x)\big] = -\mathbb{E}_q\big[\nabla \cdot f(x)\big],$$

which is equation 14. ■

A.3 Proof of Lemma 2 (Variational Identity)

Proof. The function $x \mapsto y^\top x - \frac{1}{2}\|x\|^2$ is strictly concave and attains its maximum at the unique stationary point $x = y$, where its value is $\frac{1}{2}\|y\|^2$. Rescaling by 2 and identifying $a \equiv y$, $f \equiv x$ yields equation 15. ■

A.4 Proof of Proposition 1

Proof. For every $f \in \mathcal{F}$ and every (x_t, t), Lemma 2 gives $2f^\top a - \|f\|^2 \le \|a\|^2$ with $a := \nabla \log \rho_{E,t} - \nabla \log \rho_{\pi,t}$. Taking the expectation yields $\mathbb{E}[\ell_f] \le D_F(\rho_\pi \| \rho_E)$, hence $\sup_{f \in \mathcal{F}} \mathbb{E}[\ell_f] \le D_F(\rho_\pi \| \rho_E)$.

For the reverse inequality, by the condition in Appendix A.1, the pointwise maximizer $f^\star = \nabla \log \rho_{E,t} - \nabla \log \rho_{\pi,t}$ belongs to $\mathcal{F}$, and $\mathbb{E}[\ell_{f^\star}] = D_F(\rho_\pi \| \rho_E)$ by Lemma 2. Combining the two directions yields equation 17. The form in equation 16 follows from Lemma 1, which rewrites the learner-score term as the divergence of f. ■

B Proof of Theorem 1: Identifiability of the Diffusion Fisher Divergence

Proof. We first show that $D_F = 0$ implies $\rho_\pi = \rho_E$. By definition,

$$D_F(\rho_\pi \| \rho_E) \;=\; \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{\rho_{\pi,t}} \big\| s_{\pi,t} - s_{E,t} \big\|^2.$$

Since the integrand is nonnegative, $D_F = 0$ implies $s_{\pi,t}(x) = s_{E,t}(x)$ for almost every (t, x) under the corresponding diffused learner distribution. By the regularity assumptions of Section 3 the densities are strictly positive, so the equality extends to the full support. Fix $t = t_0$ for which the equality holds and for which the non-degeneracy condition of Section 3.2 gives $\sigma_{t_0} > 0$. Then $\nabla \log \rho_{\pi,t_0}(x) = \nabla \log \rho_{E,t_0}(x)$, so that

$$\nabla\big(\log \rho_{\pi,t_0} - \log \rho_{E,t_0}\big) = 0.$$

Under the positivity and connectedness of the support,

$$\log \rho_{\pi,t_0}(x) - \log \rho_{E,t_0}(x) = C_{t_0}, \qquad \rho_{\pi,t_0} = e^{C_{t_0}} \rho_{E,t_0},$$

for a constant C_{t_0}. Both sides are probability densities, so integrating gives $e^{C_{t_0}} = 1$ and hence $\rho_{\pi,t_0} = \rho_{E,t_0}$.

To extend the equality from t_0 to the original distributions, we use the injectivity of the Gaussian convolution. Under the forward-diffusion form of Section 3.2, the diffused densities are Gaussian convolutions of the original ones up to the scaling α_{t_0}:

$$\widehat{\rho_{\pi,t_0}}(\omega) \;=\; \widehat{\rho_\pi}(\alpha_{t_0}\omega)\, \exp\Big(-\tfrac{1}{2}\sigma_{t_0}^2 \|\omega\|^2\Big),$$

and similarly for the expert. The left-hand sides coincide by the previous step; the Gaussian factor is strictly positive for all ω, so

$$\widehat{\rho_\pi}(\alpha_{t_0}\omega) \;=\; \widehat{\rho_E}(\alpha_{t_0}\omega).$$

Provided $\alpha_{t_0} \neq 0$, the change of variables $\xi = \alpha_{t_0}\omega$ gives $\widehat{\rho_\pi}(\xi) = \widehat{\rho_E}(\xi)$ for all ξ, and the Fourier uniqueness theorem yields $\rho_\pi = \rho_E$.

Conversely, if $\rho_\pi = \rho_E$, the diffusion operation preserves equality of the marginals and their scores, so $D_F(\rho_\pi \| \rho_E) = 0$. ■

C Proofs of Theorem 2 and Corollary 1

This appendix provides the auxiliary lemmas and detailed proofs for the objective-level guarantee. The regularity and non-degeneracy conditions on the diffusion setting are stated in Section 3, the instance-dependent approximation quantities attached to the components of DGAIL are defined in Section 4.3, and the realizability assumption is stated in Corollary 1. Throughout, $\epsilon_{\mathrm{score}}(\pi)$, $\epsilon_{\mathcal{F}}(\pi)$ and $B_f(\pi)$ are always evaluated at the occupancy of the policy in their argument; no condition is imposed uniformly over Π.

C.1 Notation and Setup

Throughout this appendix, an expectation over diffusion times is the joint expectation under $t \sim \mathcal{U}\{1, \dots, T\}$ and $x_t \sim \rho_{\pi,t}$, that is, $\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{x_t \sim \rho_{\pi,t}}[\cdot]$, matching the discrete definition equation 19 of the divergence.

For compactness define, for a policy π,

$$D(\pi) := D_F(\rho_\pi \| \rho_E), \qquad G(\pi) := \sup_{f \in \mathcal{F}} \mathcal{J}(\pi, f), \qquad G_\psi(\pi) := \sup_{f \in \mathcal{F}} \mathcal{J}_\psi(\pi, f), \tag{37}$$

and let $\widetilde{G}_{\psi,M}(\pi) := \sup_{f \in \mathcal{F}} \widetilde{\mathcal{J}}_{\psi,M}(\pi, f)$ denote the population game value built with the Hutchinson estimator of Appendix C.3. The Hutchinson perturbation is controlled by

$$\left|\widetilde{G}_{\psi,M}(\pi) - G_\psi(\pi)\right| \;\le\; \epsilon_{\mathrm{H}}(M, \delta), \tag{38}$$

which holds with probability at least $1 - \delta$ at the occupancies at which it is invoked; this is the uniform (over $\mathcal{F}$) bound of Lemma 5(iii). All quantities $\epsilon_{\mathrm{score}}(\pi)$, $\epsilon_{\mathcal{F}}(\pi)$ and $B_f(\pi)$ are the instance-dependent quantities defined in Section 4.3, evaluated at the occupancy of the policy indicated by their argument. No condition is imposed uniformly over Π.

C.2 Auxiliary Lemmas for Theorem 2

The following three lemmas quantify the score approximation error, the game value at the expert comparator, and the stochastic divergence estimation error.

Lemma 3 (Score perturbation)**.** *Let $\mathcal{J}_\psi$ denote the population adversarial objective obtained by replacing the true expert score $s_{E,t}$ with s_ψ in equation 20. For any policy π at which $\epsilon_{\mathrm{score}}(\pi)$ and $B_f(\pi)$ are finite, and every $f \in \mathcal{F}$,*

$$\left|\mathcal{J}_\psi(\pi, f) - \mathcal{J}(\pi, f)\right| \;\le\; 2B_f(\pi)\,\epsilon_{\mathrm{score}}(\pi). \tag{39}$$

Consequently,

$$\left|\sup_{f \in \mathcal{F}} \mathcal{J}_\psi(\pi, f) \;-\; \sup_{f \in \mathcal{F}} \mathcal{J}(\pi, f)\right| \;\le\; 2B_f(\pi)\,\epsilon_{\mathrm{score}}(\pi). \tag{40}$$

Proof. The divergence and quadratic terms are identical in the two objectives, so $\mathcal{J}_\psi(\pi, f) - \mathcal{J}(\pi, f) = 2\mathbb{E}[f^\top(s_\psi - s_{E,t})]$, the expectation being taken over $t \sim \mathcal{U}\{1, \dots, T\}$ and $x_t \sim \rho_{\pi,t}$. By Cauchy–Schwarz and the definitions equation 27 and equation 29, $\left|\mathcal{J}_\psi - \mathcal{J}\right| \le 2(\mathbb{E}\|f\|^2)^{1/2}(\mathbb{E}\|s_\psi - s_{E,t}\|^2)^{1/2} \le 2B_f(\pi)\,\epsilon_{\mathrm{score}}(\pi)$. Taking the supremum over $f \in \mathcal{F}$ yields equation 40. ■

Lemma 4 (Game value at the expert comparator)**.** *Under the regularity assumptions of Section 3 and the realizability assumption $\rho_{\pi_E} = \rho_E$ of Corollary 1, the learned-score objective at the expert policy satisfies*

$$\sup_{f \in \mathcal{F}} \mathcal{J}_\psi(\pi_E, f) \;\le\; \epsilon_{\mathrm{score}}(\pi_E)^2. \tag{41}$$

More generally, for any policy π,

$$\sup_{f \in \mathcal{F}} \mathcal{J}_\psi(\pi, f) \;\le\; 2\epsilon_{\mathrm{score}}(\pi)^2 + 2D_F(\rho_\pi \| \rho_E). \tag{42}$$

Proof. By integration by parts under $\rho_{\pi,t}$ (Lem. 1), $\mathbb{E}[2\nabla \cdot f] = -2\,\mathbb{E}[f^\top s_{\pi,t}]$, hence $\mathcal{J}_\psi(\pi, f) = \mathbb{E}\big[2f^\top(s_\psi - s_{\pi,t}) - \|f\|^2\big]$. Applying the pointwise variational inequality $2f^\top a - \|f\|^2 \le \|a\|^2$ with $a = s_\psi - s_{\pi,t}$ (Lem. 2) gives $\sup_{f\in\mathcal{F}} \mathcal{J}_\psi(\pi, f) \le \mathbb{E}\|s_\psi - s_{\pi,t}\|^2$. For $\pi = \pi_E$, the realizability assumption implies $s_{\pi_E,t} = s_{E,t}$, and the definition equation 27 of $\epsilon_{\text{score}}(\pi_E)$ then gives equation 41. For general π, the inequality $\|s_\psi - s_{\pi,t}\|^2 \le 2\|s_\psi - s_{E,t}\|^2 + 2\|s_{E,t} - s_{\pi,t}\|^2$ together with equation 27 and the definition of D_F in equation 19 give equation 42. ■

Lemma 5 (Hutchinson divergence estimation)**.** *For $f \in \mathcal{F}$, let $J_f(x,t) := \nabla_x f(x,t)$ denote the Jacobian of the vector field, and let $\xi_1, \dots, \xi_M \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I)$. Define the Hutchinson estimator*

$$\widehat{\operatorname{div}}_M f(x,t) \;:=\; \frac{1}{M}\sum_{m=1}^{M} \xi_m^\top J_f(x,t)\xi_m. \tag{43}$$

Then:

1. *The estimator is unbiased: $\mathbb{E}_\xi\big[\widehat{\operatorname{div}}_M f(x,t)\big] = \nabla \cdot f(x,t)$.*
2. *For Gaussian Hutchinson vectors, writing $J_f^{\text{sym}} := \frac{1}{2}(J_f + J_f^\top)$,*
$$\operatorname{Var}_\xi\big[\widehat{\operatorname{div}}_M f(x,t)\big] \;=\; \frac{2}{M}\big\|J_f^{\text{sym}}(x,t)\big\|_F^2 \;\le\; \frac{2}{M}\|J_f(x,t)\|_F^2. \tag{44}$$
3. *Under the Jacobian bounds of Section 4.3 and a covering-number condition on $\mathcal{F}$, there exist universal constants $c_1, c_2 > 0$ such that, with probability at least $1-\delta$,*
$$\epsilon_{\text{H}}(M,\delta) \;\le\; c_1 B_J \sqrt{\frac{\log(2/\delta)}{M}} + c_2 B_{\text{op}} \frac{\log(2/\delta)}{M}, \tag{45}$$
where $\epsilon_{\text{H}}(M,\delta)$ is the uniform (over $f \in \mathcal{F}$) stochastic divergence perturbation used in Theorem 2. The first term is the dominant $M^{-1/2}$ variance contribution, while the second is the higher-order concentration term.

Proof. For any matrix J_f, the cyclic property of the trace gives $\mathbb{E}_\xi[\xi^\top J_f \xi] = \mathbb{E}_\xi[\operatorname{tr}(J_f \xi\xi^\top)] = \operatorname{tr}(J_f) = \nabla\cdot f$, which proves unbiasedness. Moreover, $\xi^\top J_f \xi = \xi^\top J_f^{\text{sym}} \xi$, and a centered Gaussian quadratic form with symmetric matrix A satisfies $\operatorname{Var}(\xi^\top A\xi) = 2\|A\|_F^2$; hence $\operatorname{Var}\big[\widehat{\operatorname{div}}_M f\big] = \frac{2}{M}\|J_f^{\text{sym}}\|_F^2 \le \frac{2}{M}\|J_f\|_F^2$. The concentration bound follows from the sub-exponential concentration of Gaussian quadratic forms (Hanson–Wright) together with the Jacobian bounds of Section 4.3; the uniformity over $f \in \mathcal{F}$ follows from a standard covering argument over the discriminator class, which contributes only a logarithmic factor. ■

The rates of Lemma 5 are verified empirically on a trained diffusion model in Appendix E.1.

C.3 Stochastic Divergence Estimation

This section analyzes the Hutchinson estimator used to evaluate the divergence of the vector field, which is the source of stochasticity specific to the DGAIL objective beyond the generic finite-sample estimation of expectations from learner rollouts.

C.3.1 Population Hutchinson Approximation

For a fixed policy π, define the population objective using the finite Hutchinson estimator by

$$\widetilde{\mathcal{J}}_{\psi,M}(\pi, f) \;:=\; \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{x_t\sim\rho_{\pi,t}}\Big[2f(x_t,t)^\top s_\psi(x_t,t) + 2\,\widehat{\operatorname{div}}_M f(x_t,t) - \|f(x_t,t)\|_2^2\Big], \tag{46}$$

where $\widehat{\operatorname{div}}_M f$ is the estimator of Lemma 5. The only randomness in $\widetilde{\mathcal{J}}_{\psi,M}$ is therefore induced by the M Hutchinson vectors; no finite learner-sample approximation is introduced in this definition. By Lemma 5(i),

$$\mathbb{E}_\xi\big[\widetilde{\mathcal{J}}_{\psi,M}(\pi, f)\big] \;=\; \mathcal{J}_\psi(\pi, f). \tag{47}$$

This formulation isolates the stochasticity specific to the divergence computation from the generic empirical-process error associated with estimating expectations from finite learner rollouts.

C.3.2 Dependence on the Number of Hutchinson Vectors

By Lemma 5(ii) and the Jacobian bound of Section 4.3, the variance of the divergence estimator satisfies

$$\mathrm{Var}\big[\widehat{\mathrm{div}}_M f\big] \;\le\; \frac{2B_J^2}{M}. \tag{48}$$

Consequently, increasing M decreases the standard deviation of the divergence estimator at the standard Monte Carlo rate $O(M^{-1/2})$. The corresponding divergence-estimation error obeys

$$\epsilon_{\mathrm{H}}(M,\delta) \;=\; O\Big(B_J\sqrt{\frac{\log(1/\delta)}{M}} + B_{\mathrm{op}}\frac{\log(1/\delta)}{M}\Big), \tag{49}$$

so the number of Hutchinson vectors provides a direct computational–stability trade-off: increasing M reduces stochastic fluctuations but increases the computational cost of each discriminator evaluation linearly in M. Because $\widetilde{\mathcal{J}}_{\psi,M}$ averages the T per-level estimation errors, the bound equation 49 applies to the averaged game value as stated; when the per-level errors are independent across noise levels, the averaging contributes an additional $\sqrt{T}$ tightening. In our implementation, a small number of Hutchinson vectors is therefore sufficient in principle, while larger values can be used when lower estimator variance is preferred. These rates are validated empirically on a trained diffusion model in Appendix E.1.

C.4 The Exact Discriminator Identity

For a fixed policy π, set $a_\pi(x,t) := s_{E,t}(x) - s_{\pi,t}(x)$. Using $2f^\top a - \|f\|^2 = \|a\|^2 - \|f-a\|^2$ pointwise, we obtain

$$\mathcal{J}(\pi,f) \;=\; D_F(\rho_\pi \,\|\, \rho_E) - \mathbb{E}\big[\|f - f_\pi^\star\|_2^2\big], \tag{50}$$

where $f_\pi^\star = a_\pi$. Consequently,

$$D_F(\rho_\pi \,\|\, \rho_E) - \sup_{f\in\mathcal{F}} \mathcal{J}(\pi,f) \;=\; \inf_{f\in\mathcal{F}} \mathbb{E}\big[\|f - f_\pi^\star\|_2^2\big], \tag{51}$$

and by the discriminator-approximation condition of Section 4.3, $0 \le D_F(\rho_\pi\|\rho_E) - \sup_{f\in\mathcal{F}}\mathcal{J}(\pi,f) \le \epsilon_{\mathcal{F}}^2$. The discriminator approximation error is therefore exactly the minimum mean-square error with which the neural vector field approximates the true score difference.

C.5 Proof of Theorem 2

Proof. Fix any policy π. We first bound the error caused by restricting the discriminator class. By the exact variational identity of Appendix C.4,

$$\mathcal{J}(\pi,f) \;=\; D(\pi) - \mathbb{E}\big[\|f - f_\pi^\star\|_2^2\big], \qquad f_\pi^\star := s_{E,t} - s_{\pi,t}, \tag{52}$$

so that taking the supremum over $f \in \mathcal{F}$ and using the definition equation 28 of $\epsilon_{\mathcal{F}}(\pi)$,

$$0 \;\le\; D(\pi) - G(\pi) \;=\; \inf_{f\in\mathcal{F}} \mathbb{E}\big[\|f - f_\pi^\star\|_2^2\big] \;=\; \epsilon_{\mathcal{F}}(\pi)^2, \tag{53}$$

and hence

$$D(\pi) \;\le\; G(\pi) + \epsilon_{\mathcal{F}}(\pi)^2. \tag{54}$$

Next, we account for the approximation of the expert score. Lemma 3, applied at the occupancy of π, gives

$$G(\pi) \;\le\; G_\psi(\pi) + 2B_f(\pi)\,\epsilon_{\mathrm{score}}(\pi), \tag{55}$$

and therefore

$$D(\pi) \;\le\; G_\psi(\pi) + 2B_f(\pi)\,\epsilon_{\mathrm{score}}(\pi) + \epsilon_{\mathcal{F}}(\pi)^2. \tag{56}$$

Finally we account for the Hutchinson estimator. By equation 38, $G_\psi(\pi) \le \widetilde{G}_{\psi,M}(\pi) + \epsilon_{\mathrm{H}}$, hence

$$D(\pi) \;\le\; \widetilde{G}_{\psi,M}(\pi) + 2B_f(\pi)\,\epsilon_{\mathrm{score}}(\pi) + \epsilon_{\mathcal{F}}(\pi)^2 + \epsilon_{\mathrm{H}}, \tag{57}$$

which is equation 31. Every step above is an inequality between the true divergence and the attained game value of the *same* policy π; no property of the procedure that produced π has been used. ■

C.6 Proof of Corollary 1

Proof. Apply Theorem 2 at $\pi = \hat{\pi}$ and bound the attained game value $\widetilde{G}_{\psi,M}(\hat{\pi})$ by that of a comparator. By the definition equation 32 of the optimization error and the realizability assumption $\pi_E \in \Pi$,

$$\widetilde{G}_{\psi,M}(\hat{\pi}) \;=\; \inf_{\pi\in\Pi} \widetilde{G}_{\psi,M}(\pi) + \epsilon_{\mathrm{opt}} \;\le\; \widetilde{G}_{\psi,M}(\pi_E) + \epsilon_{\mathrm{opt}} \;\le\; G_\psi(\pi_E) + \epsilon_{\mathrm{H}} + \epsilon_{\mathrm{opt}}, \tag{58}$$

where the last step applies equation 38 at the expert comparator. It remains to bound the game value at the comparator: by Lemma 4,

$$G_\psi(\pi_E) \;\le\; \epsilon_{\mathrm{score}}(\pi_E)^2. \tag{59}$$

Substituting equation 59 into equation 58 and the result into equation 57 at $\pi = \hat{\pi}$ gives

$$D_F(\rho_{\hat{\pi}} \| \rho_E) \;\le\; 2B_f(\hat{\pi})\,\epsilon_{\mathrm{score}}(\hat{\pi}) + \epsilon_{\mathrm{score}}(\pi_E)^2 + \epsilon_{\mathcal{F}}(\hat{\pi})^2 + 2\epsilon_{\mathrm{H}}(M,\delta) + \epsilon_{\mathrm{opt}}, \tag{60}$$

which is equation 33. Without the realizability assumption, applying the general branch of Lemma 4 at an arbitrary comparator $\pi \in \Pi$ in place of equation 59 gives $\widetilde{G}_{\psi,M}(\hat{\pi}) \le 2\epsilon_{\mathrm{score}}(\pi)^2 + 2D_F(\rho_\pi\|\rho_E) + \epsilon_{\mathrm{H}} + \epsilon_{\mathrm{opt}}$. Since this holds for every $\pi \in \Pi$, we may take the infimum of the right-hand side; the two comparator-dependent terms must be minimized jointly, which gives the projection form equation 34 stated in the corollary. ■

Remark 1 (Reading the certificate off a training run). Theorem 2 bounds the true divergence by the attained game value $\widetilde{G}_{\psi,M}(\pi) = \sup_{f\in\mathcal{F}} \widetilde{\mathcal{J}}_{\psi,M}(\pi, f)$. Turning this into a quantity that can be read off a training log requires the inner maximization to be solved, since the logged value $\widetilde{\mathcal{J}}_{\psi,M}(\pi, f_\phi)$ at the current discriminator *lower*-bounds $\widetilde{G}_{\psi,M}(\pi)$ and therefore does not by itself certify anything; the gap is the inner-optimization error. This assumption is of a different nature from the global optimality of the policy: by the exact identity of Appendix C.4, maximizing over f is precisely the least-squares regression $\min_{f\in\mathcal{F}} \mathbb{E}\|f - f^\star_\pi\|^2$ of the discriminator onto the score difference, so the inner problem is a standard supervised-learning problem rather than a minimax one.

Remark 2 (Scope of the guarantee). Theorem 2 is unconditional in the policy: it converts an *attained* game value into a bound on the true diffusion Fisher divergence, and is therefore informative even for a policy that is far from optimal. Corollary 1 then charges the residual suboptimality of the policy optimizer to the single explicit term ϵ_{opt}, so that the four contributions to the final bound—expert-score approximation, discriminator-class approximation, stochastic divergence estimation, and policy optimization—are separated rather than assumed away. Two effects remain outside the decomposition. The first is the finite-sample error of estimating the expectations in $\widetilde{\mathcal{J}}_{\psi,M}$ from learner rollouts and of fitting s_ψ from a finite demonstration set; the latter is implicit in $\epsilon_{\mathrm{score}}$, which is a population quantity. The second is the inner-maximization gap discussed in Remark 1, which separates the logged discriminator value from $\widetilde{G}_{\psi,M}$. Neither is specific to diffusion-guided adversarial imitation.

D Proof of Theorem 3: Diffusion Stability and the Imitation Gap

D.1 Diffusion-Stability Analysis: From Score Discrepancy to the Imitation Gap

This section proves the two lemmas that bridge the distribution-matching guarantee of Corollary 1 to the imitation gap. Throughout, $D_H^2(P,Q) := \int(\sqrt{p} - \sqrt{q})^2\,dx$ denotes the squared Hellinger distance between densities p and q, and $\rho_{\pi,t}$ the marginal of the forward diffusion of ρ_π at time t. The key observation is that the diffusion Fisher divergence measures the drift discrepancy of the reverse diffusion processes of the two distributions, so that standard diffusion stability results (Girsanov's theorem and the exponential convergence of the forward process to the standard Gaussian) transfer it into a distributional discrepancy. Throughout this section the Girsanov argument is run in continuous diffusion time $t \in (0, T]$; the discrete-time divergence equation 19 is the uniform average of the per-level score discrepancies over the grid $\{1, \dots, T\}$, so the continuous-time integral below equals T times it up to a Riemann discretization error that is absorbed into the constants. The discrete-time objective of Section 4.2 thus enters the analysis only through the error term $\epsilon_{\mathrm{DGAIL}}$ of Theorem 3.

Diffusion regularity for the performance analysis. Let $\Pi_E := \Pi \cup \{\pi_E\}$. For every $\pi \in \Pi_E$, the diffused score $s_{\pi,t}(x) := \nabla \log \rho_{\pi,t}(x)$ is Lipschitz continuous in x with a finite constant for every $t \in (0, T]$, the second moment of the state occupancy is bounded, $\mathbb{E}_{s\sim\rho_\pi}\left[\|s\|_2^2\right] \le m$ for some $m > 0$, and the score discrepancy $t \mapsto \mathbb{E}_{x_t\sim\rho_{\pi,t}}\|s_{\pi,t}(x_t) - s_{E,t}(x_t)\|_2^2$ is integrable on $(0, T]$. This condition is used only for the performance analysis of Theorem 3, not for Theorems 1 and 2. The expert is included in Π_E because the expert-side mixing term of Lemma 6 requires a second-moment bound on ρ_E as well; under the realizability assumption of Corollary 1 this is already implied by $\pi_E \in \Pi$. The integrability requirement is what licenses the Girsanov change of measure on all of $(0, T]$ and rules out a non-integrable score blow-up as $t \to 0$; see Remark 3 for the variant that replaces it by early stopping.

Diffusion parameterization used in this section. The general variance-preserving schedule (α_t, σ_t) of Section 3.2 is specialized here to the standard Ornstein–Uhlenbeck instance $\alpha_t = e^{-t}$, $\sigma_t^2 = 1 - e^{-2t}$, i.e. the forward process $dx_t = -x_t\,dt + \sqrt{2}\,dB_t$ with stationary law $\gamma = \mathcal{N}(0, I)$. The explicit constants below ($C_{\mathrm{diff}} = 2T$ and the e^{-T} mixing rate) refer to this parameterization; a general schedule enters only through a deterministic time change of the diffusion clock and rescales these constants accordingly.

Lemma 6. *Under the diffusion regularity condition stated above,*

$$D_H^2(\rho_{\hat\pi}, \rho_E) \;\le\; C_{\mathrm{diff}}\,\epsilon_{\mathrm{DGAIL}} + C_{\mathrm{mix}}\,e^{-T}, \tag{61}$$

where $C_{\mathrm{diff}} = O(T)$ and $C_{\mathrm{mix}} = O(m + d)$.

Proof. Consider the two reverse-time SDEs with the common initial distribution $\rho_{\hat\pi,T}$,

$$dx_t = \big(x_t + 2s_{\hat\pi,T-t}(x_t)\big)\,dt + \sqrt{2}\,dB_t, \qquad x_0 \sim \rho_{\hat\pi,T}, \tag{62}$$

$$dy_t = \big(y_t + 2s_{E,T-t}(y_t)\big)\,dt + \sqrt{2}\,dB_t, \qquad y_0 \sim \rho_{\hat\pi,T}, \tag{63}$$

where $s_{\pi,t}$ and $s_{E,t}$ denote the true scores of the diffused learner and expert distributions and B_t is a standard Brownian motion, and let ν_E be the terminal distribution of the second process. The first process is the exact reverse of the learner's forward diffusion and therefore has terminal distribution $\rho_{\hat\pi}$.

By Girsanov's theorem, the KL divergence between the two path measures of a diffusion with coefficient σ^2 equals $\frac{1}{2\sigma^2}\,\mathbb{E}\int_0^T \|\Delta b_t\|^2\,dt$, where Δb_t is the drift difference; here $\sigma^2 = 2$, so the prefactor is $\frac{1}{4}$ rather than $\frac{1}{2}$. The change of measure is licensed by the integrability part of the diffusion regularity condition. Since the marginal of the first process at reverse time t is $\rho_{\hat\pi,T-t}$ and the drifts differ by $\Delta b_t = 2(s_{\hat\pi,T-t} - s_{E,T-t})$, so that $\frac{1}{4}\|\Delta b_t\|^2 = \|s_{\hat\pi,T-t} - s_{E,T-t}\|^2$, we obtain

$$D_{\mathrm{KL}}\big(\mathrm{law}(x)\,\|\,\mathrm{law}(y)\big) = T\,\mathbb{E}_{t\sim\mathcal{U}(0,T)}\,\mathbb{E}_{x_t\sim\rho_{\hat\pi,t}}\big\|s_{\hat\pi,t}(x_t) - s_{E,t}(x_t)\big\|^2 = T\,D_F(\rho_{\hat\pi}\|\rho_E). \tag{64}$$

Mapping a path to its terminal coordinate is deterministic, so the data-processing inequality gives $D_{\mathrm{KL}}(\rho_{\hat\pi}\|\nu_E) \le D_{\mathrm{KL}}(\mathrm{law}(x)\|\mathrm{law}(y))$, and $D_H^2 \le D_{\mathrm{KL}}$ then yields $D_H^2(\rho_{\hat\pi}, \nu_E) \le T\,D_F(\rho_{\hat\pi}\|\rho_E)$.

To relate ν_E to ρ_E, note that both are produced by the same reverse-diffusion kernel (drift s_E) from the initial distributions $\rho_{\hat\pi,T}$ and $\rho_{E,T}$, respectively. By the data-processing inequality for the Hellinger distance and its triangle inequality against the standard Gaussian γ,

$$D_H^2(\nu_E, \rho_E) \;\le\; D_H^2(\rho_{\hat\pi,T}, \rho_{E,T}) \;\le\; 2D_H^2(\rho_{\hat\pi,T}, \gamma) + 2D_H^2(\rho_{E,T}, \gamma) \;\le\; C'_{\mathrm{mix}}\,e^{-T}, \tag{65}$$

where the last step uses the exponential convergence of the forward diffusion to γ under the bounded second moment, $D_{\mathrm{KL}}(\rho_T\|\gamma) = O((m + d)e^{-T})$, and $D_H^2 \le D_{\mathrm{KL}}$. This is where the diffusion regularity condition is needed for the expert as well: the term $D_H^2(\rho_{E,T}, \gamma)$ requires a second-moment bound on ρ_E, which is why the assumption is stated over $\Pi_E = \Pi \cup \{\pi_E\}$.

Combining the two bounds with the triangle inequality for the Hellinger distance gives

$$D_H^2(\rho_{\hat\pi}, \rho_E) \;\le\; 2D_H^2(\rho_{\hat\pi}, \nu_E) + 2D_H^2(\nu_E, \rho_E) \;\le\; 2T\,D_F(\rho_{\hat\pi}\|\rho_E) + 2C'_{\mathrm{mix}}\,e^{-T}. \tag{66}$$

Setting $C_{\mathrm{mix}} := 2C'_{\mathrm{mix}} = O(m + d)$ and $C_{\mathrm{diff}} := 2T$, the claim follows from Corollary 1, which bounds $D_F(\rho_{\hat\pi}\|\rho_E)$ by $\epsilon_{\mathrm{DGAIL}}$ on an event of probability at least $1 - \delta$; the bound equation 61 holds on that same event. ∎

Remark 3. The integrability condition of the diffusion regularity assumption is what allows the Girsanov argument to run on all of $(0, T]$. If instead the score discrepancy is only controlled away from the origin, $t \geq t_{\min} > 0$, then running the same argument on $[t_{\min}, T]$ gives

$$D_H^2(\rho_{\hat{\pi}}, \rho_E) \;\leq\; 2(T - t_{\min})\, D_F(\rho_{\hat{\pi}} \| \rho_E) + C_{\mathrm{mix}}\, e^{-T} + \epsilon_{\mathrm{early}}(t_{\min}),$$

where $\epsilon_{\mathrm{early}}(t_{\min}) \to 0$ as $t_{\min} \to 0$ accounts for stopping the reverse process early. This is the standard early-stopping term of diffusion sampling analyses and changes neither the form of equation 61 nor that of Theorem 3.

Lemma 7. *For any bounded cost $c : \mathcal{S} \to [0, C_{\max}]$ and any two distributions P, Q on $\mathcal{S}$,*

$$\big|\mathbb{E}_P[c] - \mathbb{E}_Q[c]\big| \;\leq\; c_1 \sqrt{\mathrm{Var}_Q(c)}\, D_H(P, Q) + c_2\, C_{\max}\, D_H^2(P, Q), \tag{67}$$

with the explicit absolute constants $c_1 = c_2 = 2$, written in lower case to distinguish them from the concentration constants of the Hutchinson bound equation 45. No smallness condition on $D_H(P, Q)$ is required.

Proof. Write $\Delta := \mathbb{E}_P[c] - \mathbb{E}_Q[c]$ and centre the cost at the Q-mean, $\varphi := c - \mathbb{E}_Q[c]$. Since both c and $\mathbb{E}_Q[c]$ lie in $[0, C_{\max}]$ we have $\|\varphi\|_\infty \leq C_{\max}$, and by construction $\mathbb{E}_Q[\varphi^2] = \mathrm{Var}_Q(c)$ and $\mathbb{E}_P[\varphi^2] = \mathrm{Var}_P(c) + \Delta^2$. Set

$$S \;:=\; \mathbb{E}_P[\varphi^2] + \mathbb{E}_Q[\varphi^2] \;=\; \mathrm{Var}_P(c) + \mathrm{Var}_Q(c) + \Delta^2. \tag{68}$$

The proof applies one and the same Cauchy–Schwarz estimate twice: once to φ, bounding Δ in terms of S, and once to φ^2, bounding S in terms of $\mathrm{Var}_Q(c)$ alone.

We first bound Δ in terms of S. Since $\int (p - q)\, dx = 0$, subtracting the constant $\mathbb{E}_Q[c]$ from the cost leaves Δ unchanged, and factoring $p - q = (\sqrt{p} - \sqrt{q})(\sqrt{p} + \sqrt{q})$ gives $\Delta = \int (\sqrt{p} - \sqrt{q})(\sqrt{p} + \sqrt{q})\, \varphi\, dx$. By Cauchy–Schwarz and $(\sqrt{p} + \sqrt{q})^2 \leq 2(p + q)$,

$$\Delta^2 \;\leq\; D_H^2(P, Q) \int \varphi^2 (\sqrt{p} + \sqrt{q})^2\, dx \;\leq\; 2\, D_H^2(P, Q)\, S. \tag{69}$$

We next bound S in terms of $\mathrm{Var}_Q(c)$. Applying the same factorization and Cauchy–Schwarz to φ^2 in place of φ, and then using $\varphi^4 \leq C_{\max}^2 \varphi^2$,

$$\begin{aligned} \big|\mathbb{E}_P[\varphi^2] - \mathbb{E}_Q[\varphi^2]\big| \;&=\; \Big| \int \varphi^2 (\sqrt{p} - \sqrt{q})(\sqrt{p} + \sqrt{q})\, dx \Big| \\ &\leq\; D_H(P, Q) \sqrt{2 \int \varphi^4 (p + q)\, dx} \;\leq\; \sqrt{2}\, C_{\max}\, D_H(P, Q) \sqrt{S}. \end{aligned} \tag{70}$$

This is the step at which the boundedness of c enters. It is essential that the difference of the second moments be controlled by D_H directly, in the same way as Δ itself: the pointwise inequality $c^2 \leq C_{\max} c$ bounds $\mathbb{E}_P[c^2]$ and $\mathbb{E}_Q[c^2]$ separately but carries no information about the size of their difference relative to Δ. Indeed two distributions can have equal c-means, so $\Delta = 0$, while their c-variances differ by as much as $C_{\max}^2/4$, so no bound of the form $|\mathbb{E}_P[c^2] - \mathbb{E}_Q[c^2]| \lesssim C_{\max}|\Delta|$ can hold.

Substituting $\mathbb{E}_P[\varphi^2] \leq \mathbb{E}_Q[\varphi^2] + \sqrt{2}\, C_{\max} D_H \sqrt{S}$ into equation 68 gives $S \leq 2\mathrm{Var}_Q(c) + \sqrt{2}\, C_{\max} D_H \sqrt{S}$, which is a quadratic inequality $u^2 - \sqrt{2}\, C_{\max} D_H\, u - 2\mathrm{Var}_Q(c) \leq 0$ in $u = \sqrt{S} \geq 0$. Its solution set is

$$u \;\leq\; \tfrac{1}{2}\Big(\sqrt{2}\, C_{\max} D_H + \sqrt{2C_{\max}^2 D_H^2 + 8\mathrm{Var}_Q(c)}\Big) \;\leq\; \sqrt{2}\, C_{\max} D_H + \sqrt{2\mathrm{Var}_Q(c)},$$

where the last step uses $\sqrt{a + b} \leq \sqrt{a} + \sqrt{b}$.

Combining with equation 69, which reads $|\Delta| \leq \sqrt{2}\, D_H\, u$,

$$|\Delta| \;\leq\; \sqrt{2}\, D_H \Big(\sqrt{2}\, C_{\max} D_H + \sqrt{2\mathrm{Var}_Q(c)}\Big) \;=\; 2\sqrt{\mathrm{Var}_Q(c)}\, D_H(P, Q) + 2\, C_{\max}\, D_H^2(P, Q),$$

which is equation 67 with $c_1 = c_2 = 2$. ∎

Remark 4. Because the second estimate controls the difference of the second moments by D_H rather than by $|\Delta|$, the argument closes without any restriction on $D_H(P, Q)$, and the constants do not degrade as D_H^2 approaches its maximal value. The bound is nevertheless informative only in the small-distance regime: D_H^2 is used here in its unnormalized form $\int(\sqrt{p} - \sqrt{q})^2 dx \in [0, 2]$, so once D_H^2 is of order one the right-hand side of equation 67 exceeds $C_{\max}$ and the trivial bound $|\mathbb{E}_P[c] - \mathbb{E}_Q[c]| \le C_{\max}$ is the stronger of the two. This is why Theorem 3 retains the minimum of its two branches.

D.2 Proof of Theorem 3

Proof. Because c takes values in $[0, C_{\max}]$, the gap is at most $C_{\max}$, which is the first branch of equation 36.

For the second branch, work on the event of probability at least $1 - \delta$ on which Corollary 1 holds. Lemma 6 converts the score discrepancy into an occupancy discrepancy: under the diffusion regularity condition,

$$D_H^2(\rho_{\hat{\pi}}, \rho_E) \;\le\; C_{\text{diff}}\, \epsilon_{\text{DGAIL}} + \epsilon_{\text{mix}} \;=\; \epsilon_{\text{occ}}.$$

Lemma 7, which applies to any pair of distributions, then converts the occupancy discrepancy into a cost gap,

$$\left|\mathbb{E}_{\rho_{\hat{\pi}}}[c] - \mathbb{E}_{\rho_E}[c]\right| \;\le\; c_1 \sqrt{\mathrm{Var}_{\rho_E}(c)}\, D_H(\rho_{\hat{\pi}}, \rho_E) + c_2\, C_{\max}\, D_H^2(\rho_{\hat{\pi}}, \rho_E),$$

applied with $P = \rho_{\hat{\pi}}$ and $Q = \rho_E$ so that the variance factor is the expert-side variance $\mathrm{Var}_{\rho_E}(c)$. Both terms on the right are nondecreasing in $D_H(\rho_{\hat{\pi}}, \rho_E)$, so substituting $D_H(\rho_{\hat{\pi}}, \rho_E) \le \sqrt{\epsilon_{\text{occ}}}$ gives the second branch of equation 36. ■

Choice of the diffusion horizon. The constant $C_{\text{diff}} = O(T)$ grows in the diffusion horizon while the mixing error $\epsilon_{\text{mix}} = O((m + d)e^{-T})$ decays in it, so the two are balanced by $T = \Theta(\log(1/\epsilon_{\text{DGAIL}}))$, at which point $\epsilon_{\text{occ}} = O(\epsilon_{\text{DGAIL}} \log(1/\epsilon_{\text{DGAIL}}))$. Since $\mathrm{Var}_{\rho_E}(c) \le C_{\max}^2/4$ for $c \in [0, C_{\max}]$, the first term of equation 36 dominates in the small-error regime, giving $\left|\mathbb{E}_{\rho_{\hat{\pi}}}[c] - \mathbb{E}_{\rho_E}[c]\right| = \tilde{O}\big(\sqrt{\epsilon_{\text{DGAIL}}}\big)$, the tilde hiding the logarithmic factor.

E Additional Experimental Results

This appendix collects additional experimental results that complement the theoretical analysis of the main text.

E.1 Empirical Validation of the Hutchinson Divergence Estimator

Lemma 5 controls the stochastic divergence estimation error through the exact Jacobian statistics of the vector field. We verify that the estimator used in our implementation—the VJP-based Hutchinson estimator of Section 4, which never forms J_f explicitly—attains these rates in practice.

Setup. We fix a frozen expert diffusion model on the DeepMind Control Suite cheetah-run task (observation dimension 17) and a discriminator vector field f_ϕ trained with the objective of Section 4 for 500 gradient steps and then frozen. We freeze 16 states (x, t), with x drawn from states visited by a random policy and t from the diffusion timestep grid, and compute the exact divergence $\nabla \cdot f_\phi(x, t) = \mathrm{Tr}(J_f(x, t))$ from the explicit 17×17 Jacobian obtained by automatic differentiation. The single-probe VJP estimator used during training matches the explicit quadratic form $\xi^\top J_f \xi$ to a mean absolute error of 1.6×10^{-7}. For each $M \in \{1, 2, 4, 8, 16, 32\}$—a range covering the operating points $M = 8$ (discriminator training) and $M = 32$ (reward computation) reported in Appendix F—we draw 2048 independent redraws of the M Gaussian probes and record $\widehat{\mathrm{div}}_M f$ at every frozen state.

Results. Figure 2(a) shows that the raw errors are centered at zero at every M—the pooled mean error stays below 10^{-2} in magnitude (for instance 9×10^{-3} at $M = 1$, against a standard deviation of 1.13)—confirming the unbiasedness of Lemma 5(i), while the error spread contracts visibly with

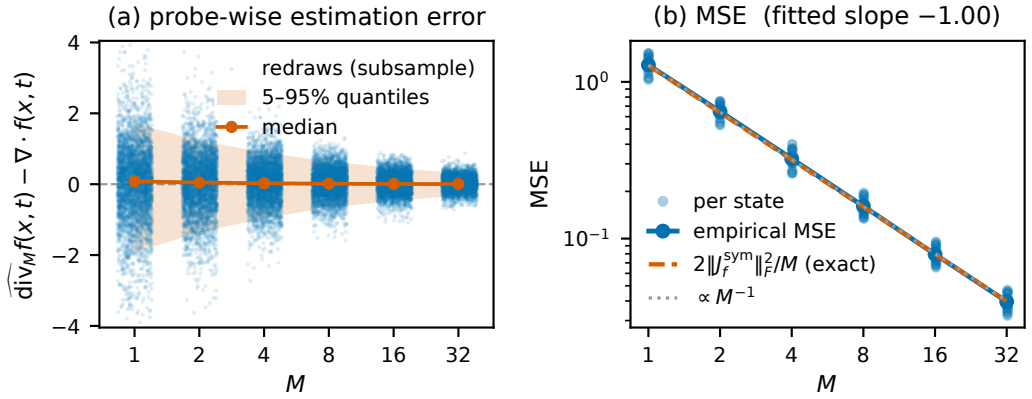


Figure 2: Empirical validation of Lemma 5 on the DeepMind Control Suite cheetah-run task (observation dimension 17), using the same VJP-based estimator as in training. (a) Raw estimation error $\widehat{\mathrm{div}}_M f(x,t) - \nabla \cdot f(x,t)$ at each M over 2048 independent probe redraws (a uniform subsample is shown) at 16 frozen states; the median and 5%–95% quantile band shrink at the Monte Carlo rate while remaining centered at zero. (b) Mean squared error against the exact divergence: light dots are individual frozen states, the solid line is their mean, and the dashed line is the exact Gaussian prediction $2\,\|J_f^{\mathrm{sym}}\|_F^2/M$ evaluated per state from the explicit Jacobian, with no fitted constant. The fitted log-log slope is -1.00.

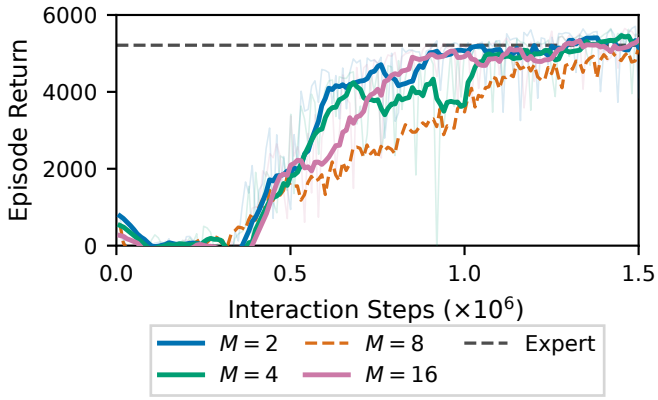


Figure 3: Training curves for different numbers of probes M on the Ant-v2 task.

M. In Figure 2(b), the empirical mean squared error matches the exact Gaussian prediction of Lemma 5(ii) within 2% at every M, and the fitted log-log slope of the MSE is -1.00. Taking square roots, the standard deviation of the estimator decays at the Monte Carlo rate $O(M^{-1/2})$ of Appendix C.3, that is, the leading term of the uniform error $\epsilon_{\mathrm{H}}(M,\delta)$ of Lemma 5(iii) that enters Theorem 2. The exact variance constant is therefore informative well beyond the asymptotic regime: the theory predicts both the rate and the level of the estimation error at the operating points used in training.

E.2 Performance versus the Number of Hutchinson Probes

The main results use $M{=}8$ Hutchinson probes for the discriminator update. To test how far this budget can be stretched, we retrain DGAIL on Ant-v2 with the discriminator-side count set to $M \in \{2, 4, 16\}$, holding the pre-trained expert score network and the discriminator class fixed; we refer to each configuration by its discriminator-side count.

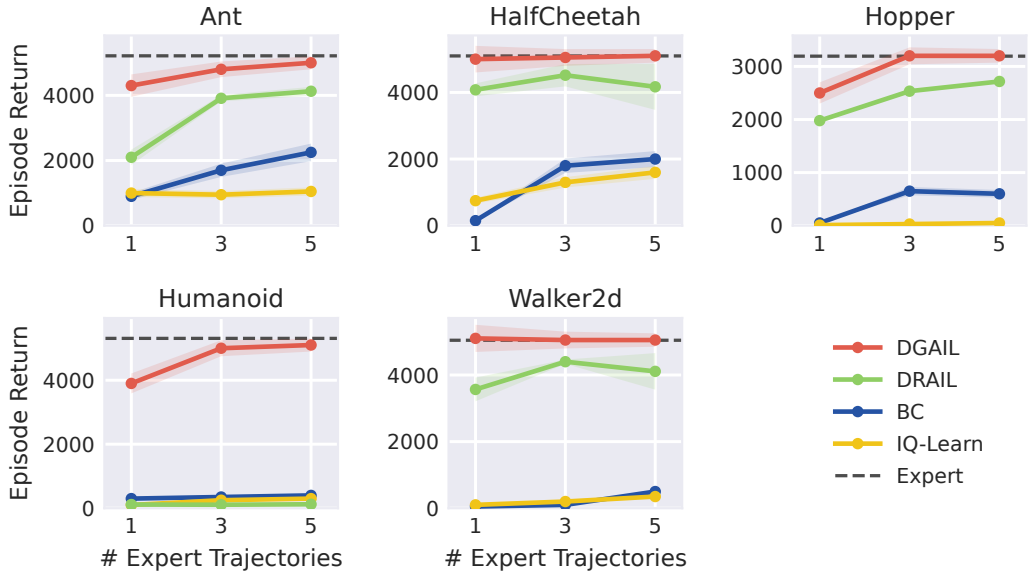


Figure 4: Episode return as the number of expert trajectories varies over $\{1, 3, 5\}$ on the five MuJoCo tasks; shaded bands indicate ± 1 standard deviation over seeds. The dashed line denotes the expert return.

Figure 3 reports the learning curves over the first 1.5M steps. All three configurations exceed the expert return of 5,214 from roughly 0.9M steps onward. The $M{=}2$ configuration rises fastest and reaches 5,705 at the end of the window; $M{=}4$ and $M{=}16$ reach 5,497 and 5,445. Over the full budget, the peaks of the three runs agree within 2% (5,741–5,848). Stretching or shrinking the probe budget by a factor of two to four therefore leaves training performance essentially unchanged, confirming in policy space what Appendix E.1 establishes at the estimator level: a small number of Hutchinson probes already yields an accurate divergence estimate. As in the main experiments, returns beyond this window are noisy and eventually degrade; performance is therefore read over the window shown.

E.3 Performance versus the Number of Expert Trajectories

The main results use a per-task demonstration budget (Table 1). To examine data efficiency directly, we retrain DGAIL and the BC, IQ-Learn, and DRAIL baselines with 1, 3, and 5 expert trajectories on each MuJoCo task of Section 5.1; for DGAIL, the pre-trained score network s_ψ, the discriminator class, and the Hutchinson estimator are kept unchanged. Figure 4 reports the final episode return of each method, averaged over seeds, together with the expert return of the underlying task.

On Ant, HalfCheetah, and Walker2d, DGAIL is already close to the expert with a single trajectory, which is precisely the budget used in the main experiments. Hopper and Humanoid benefit from additional demonstrations and approach the expert level by five trajectories, again consistent with their main-experiment budgets. The baselines display the expected data-efficiency deficit: BC and IQ-Learn remain substantially below the expert across the entire budget range, and DRAIL improves steadily with more demonstrations but collapses on Humanoid at every budget. Overall, DGAIL's performance degrades gracefully rather than catastrophically as the demonstration budget shrinks, in line with the low-demonstration-regime behavior of Section 5.2 and the distribution-matching guarantee of Theorem 2: the discriminator is driven by the diffusion Fisher signal rather than by a learned critic, so policy optimization retains a reliable learning signal even when demonstrations are scarce.

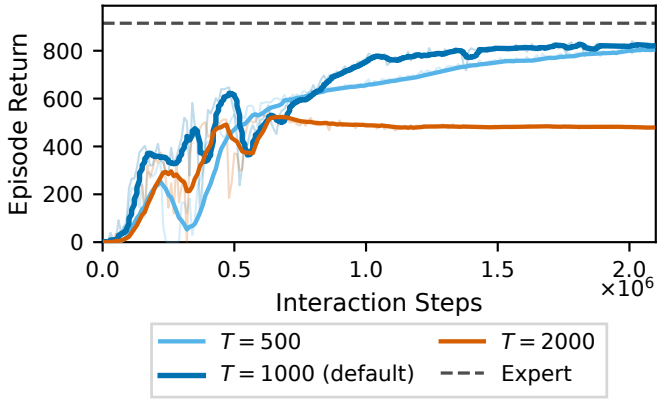


Figure 5: Effect of the number of diffusion steps T on cheetah-run. Faint lines are raw evaluations and solid lines a rolling mean over the 2.1M-step budget. The default T=1000 and the halved T=500 reach the same plateau near the expert, whereas T=2000 plateaus systematically lower. The dashed line denotes the expert return.

E.4 Sensitivity to the Number of Diffusion Steps

The number of diffusion steps T determines how finely the score path of the expert is discretized, and the expert score network s_ψ must be retrained whenever T changes, because the noise schedule is tied to it. We retrain s_ψ and rerun the full DGAIL pipeline on cheetah-run with $T \in \{500, 1000, 2000\}$, keeping the discriminator, the Hutchinson estimator (M=8 during training and M=32 for the reward), and the 2.1M-step budget unchanged. Figure 5 reports the learning curves; each run is summarized by its peak return and the median return over the last 100k steps, since individual evaluations are noisy.

The resulting picture is not monotone in T. Halving the default horizon costs nothing: T=500 reaches a peak of 817 and a last-100k median of 806, on par with the T=1000 default (841 and 820), and both settle on the same plateau close to the expert return of 915. Doubling the horizon hurts: T=2000 peaks at 547 after 0.63M steps and then locks into a plateau near 480, about 340 below the default.

This non-monotonicity is the pattern that the error decomposition of Theorem 2 suggests. A finer discretization of the expert score path is the force that would favor larger T, and it explains why T=500 is already indistinguishable from the default: once the grid is fine enough, refining it further buys little. It is not the only force, however. The bound carries the score-approximation error $\epsilon_{\text{score}}(\pi)$ of equation 27—an average over the T conditioning levels of the diffusion process—as an additive penalty, and s_ψ must be retrained for each T under a fixed architecture, data budget, and training schedule. Enlarging T makes the score regression harder: the network must be accurate on a grid of noise levels twice as dense, so the per-level error grows beyond the mere rescaling of the average, and through Theorem 2 this enters the achievable divergence directly. Past the default, this inflation outweighs the marginal discretization gain and the net effect of increasing T turns negative, which is what the T=2000 run exhibits. The default T=1000 is therefore a sound operating point: past the regime where discretization matters, and before the growth of the score error dominates. All runs use a single seed, so these conclusions are indicative rather than definitive.

F Implementation Details

This appendix documents the implementation of DGAIL: the full training procedure is summarized in Algorithm 1, and the hyperparameters of the main MuJoCo experiments (Section 5) are reported in the remaining subsections, as configured in the released training scripts. The shared RL backbone follows the SAC framework (Haarnoja et al., 2018); entries that deviate from the Dual Q-DM setup are reported at face value and are not adjusted to match that reference.

The vector-field discriminator $f_\phi(x, t)$ is a two-layer MLP (one hidden layer of 256 units, LeakyReLU) that takes the diffused state x_t and a sinusoidal embedding of the diffusion time t as input; the divergence $\nabla \cdot f_\phi$ is estimated with Hutchinson's estimator using 8 noise samples per sample for discriminator training and 32 for the reward; the policy is trained with soft actor–critic (Haarnoja et al., 2018). The expert score network s_ψ is pre-trained once on the state-only demonstrations as a denoising diffusion model in the DDPM framework (Ho et al., 2020) and kept fixed during imitation. We adopt the ε-prediction parameterization, with the score recovered via $s_\psi(x_t, t) = -\varepsilon_\psi(x_t, t)/\sigma_t$, where $\sigma_t = \sqrt{1 - \bar{\alpha}_t}$ under the variance-preserving schedule of Section 3.2, and s_ψ shares its architecture with the discriminator. Complete hyperparameters are given in the tables below.

F.1 Training Algorithm

Algorithm 1 instantiates the alternating discriminator–policy updates described in Section 4.2.

Algorithm 1 DGAIL (Diffusion-Guided Adversarial Imitation Learning)

Require: Pre-trained expert score network $s_\psi(x, t)$ (from a diffusion model)
Require: Initialized policy π_θ, vector-field network $f_\phi(x, t)$
Require: Hutchinson noise $p(\xi) = \mathcal{N}(0, I)$
for iteration $k = 1, 2, \ldots, K$ **do**
  Collect learner trajectories $\mathcal{D}^\pi$ by rolling out π_θ.
  // Discriminator (vector field) update
  for several gradient steps **do**
    Sample $t \sim \mathrm{Uniform}\{1, \ldots, T\}$.
    Sample $x \sim \mathcal{D}^\pi$ and diffuse to x_t via the forward process.
    Sample $\xi \sim p(\xi)$.
    Compute the score-guided operator

$$T_\phi(x_t, t) = 2\Big(\xi^\top \nabla_{x_t}\big(f_\phi(x_t, t)^\top \xi\big) + f_\phi(x_t, t)^\top s_\psi(x_t, t)\Big) - \|f_\phi(x_t, t)\|^2.$$

    Update ϕ by maximizing $\mathcal{L}(\phi) = \mathbb{E}[T_\phi(x_t, t)]$.
  end for
  // Policy (generator) update
  Define the step-wise dual reward $r(x) = -\mathbb{E}_{t,\xi}[T_\phi(x_t, t)]$.
  Update π_θ via RL (soft actor–critic (Haarnoja et al., 2018)) to maximize cumulative rewards.
end for

F.2 Expert Trajectories

Table 1: Number of expert trajectories for each task.

Task	Expert Trajectories
Ant-v2	1
HalfCheetah-v2	1
Hopper-v2	5
Humanoid-v2	5
Walker2d-v2	1

F.3 Score Network and Diffusion

The score network s_ψ and the discriminator f_ϕ share one MLP architecture (time embedding followed by concatenation). MuJoCo main experiments use hidden dimension 256, one hidden layer, time-embedding dimension 16, and LeakyReLU activations.

Table 2: Score network and diffusion configuration.

Parameter	Value
s_ψ architecture	2-layer MLP: Linear(d+16, 256) $\to$ LeakyReLU $\to$ Linear(256, d)
Time embedding	sinusoidal $\mathrm{PosEmb}(16)$ $\to$ Linear(16, 32) $\to$ LeakyReLU $\to$ Linear(32, 16), concatenated
Training epochs	4000
Optimizer / lr	AdamW, lr 5×10^{-3}, weight decay 0.01 (PyTorch default)
Batch size	1024
Objective	ε-prediction; score recovered as $s_\psi(x_t, t) = -\varepsilon_\psi(x_t, t)/\sigma_t$
Diffusion steps T	1000 (HuggingFace `DDPMScheduler`)
Noise schedule	DDPM linear: $\beta_1 = 10^{-4}$, $\beta_T = 0.02$; $a_t = 1-\beta_t$, $\bar{a}_t = \prod_{s\le t} a_s$, giving the equation 1 coefficients $\alpha_t = \sqrt{\bar{a}_t}$, $\sigma_t = \sqrt{1-\bar{a}_t}$

F.4 Discriminator, Hutchinson Estimator, and Reward

Table 3: Discriminator f_ϕ, Hutchinson estimator, and dual reward.

Parameter	Value
f_ϕ architecture	same MLP as s_ψ: 1 hidden layer of 256, LeakyReLU, sinusoidal time embedding
Optimizer / lr	AdamW, lr 1×10^{-5}, weight decay 0.01
Update schedule	every 5 outer epochs, one dataloader epoch ($\approx$ 98 gradient steps, batch 1024)
Hutchinson vectors M	training: 8; reward: 32
Hutchinson noise	$\mathcal{N}(0, I)$
Implementation	VJP: $\langle \varepsilon, J_f^\top \varepsilon \rangle$ via `autograd.grad` of $\langle f, \varepsilon \rangle$
Reward sampling	32 times t and 32 Hutchinson vectors per state

F.5 RL Backbone

Table 4: SAC backbone of the main runs.

Parameter	Value
Algorithm	SAC with learnable temperature
Temperature	$\alpha_0 = 0.1$, lr 10^{-4}, target entropy $-\lvert\mathcal{A}\rvert$
Discount γ	0.99
Replay buffer	10^8
Batch size	1024
Optimizer	Adam ($\beta = (0.9, 0.999)$)
Actor	2 layers $\times$ 1024, ReLU, Tanh-squashed Gaussian, $\log\sigma \in [-5, 2]$; lr 10^{-4}
Critic	2 layers $\times$ 1024, ReLU, Double-Q, target $\tau = 0.005$ updated every 2 steps; lr 10^{-4}

Baselines use the same set of expert demonstrations (subsampled per each method's official procedure) and are otherwise configured as in their original papers.