SCORE-CALIBRATED FLOW FOR SAMPLING FROM UNNORMALIZED DENSITIES WITH APPLICATIONS TO GENERATIVE ONLINE REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Diffusion and flow models provide expressive policy classes for online reinforcement learning (RL), enabling multimodal behaviors and improved performance. However, training these policies remains challenging: the critic specifies the desired policy as an unnormalized Boltzmann density but does not provide direct samples from it. Many existing methods rely on importance sampling to construct training signals, which can suffer from high variance, increasing computational cost and destabilizing training. In this paper, we propose *Score-Calibrated Flow* (SCF), a simple and efficient algorithm for training generative models to sample from unnormalized densities without importance sampling or backpropagation through the sampling trajectory. Our key idea is to learn the desired flow by enforcing self-consistency, bypassing the challenging estimation of the target posterior mean. By jointly exploiting the prescribed target score and the structure of flow matching, we establish these self-consistency requirements as *score-calibrated optimality conditions*, first for the terminal density and then for the trainable velocity field. We prove that their unique solutions are, respectively, the target density and the ideal flow model that conditional flow matching (CFM) would recover if target samples were available. We formulate the velocity condition as a fixed-point equation and exploit its conditional-expectation structure to construct a stop-gradient objective for enforcing it. The resulting training procedure retains the scalable sample–interpolate–regress structure of conditional flow matching despite the absence of target samples, using endpoints generated by the current flow. For online RL, the critic gradient supplies the target score at the generated actions, yielding a direct approach to actor training. Experiments on RL benchmarks demonstrate that SCF matches or improves upon state-of-the-art generative-policy baselines, while substantially reducing training time.

1 INTRODUCTION

Diffusion and flow models learn expressive distributions by transforming initial noise through stochastic or deterministic dynamics (Sohl-Dickstein et al., 2015; Ho et al., 2020; Lipman et al., 2023). They are appealing as policy classes since several distinct actions may be valuable in the same state, whereas a restrictive distribution may fail to represent these alternatives. Greater policy expressiveness can also translate into improved performance. These advantages have motivated the use of generative policies in imitation learning (Chi et al., 2025) and offline reinforcement learning (RL) (Wang et al., 2023). In these settings, expert demonstrations or offline datasets provide the samples needed in the standard training recipe for generative models. In online RL, however, direct samples from the target density that policy improvement asks the actor to learn are unavailable.

The difficulty is particularly clear in maximum-entropy RL (Haarnoja et al., 2018). At a given state s, soft policy improvement targets the Boltzmann policy $\pi^Q(a|s) \propto \exp(\tau Q(s,a))$, where Q is the critic and $\tau > 0$ is the inverse temperature. Soft actor-critic (SAC) approximates this target within a Gaussian policy family using a mode-seeking objective. A diffusion or flow policy offers a richer representation, but standard training methods such as conditional flow matching (CFM) require samples from the Boltzmann policy, whereas the critic provides only evaluations of $Q(s,a)$ and its

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.

action gradient $\nabla_a Q(s, a)$. More broadly, the challenge is *learning to sample from unnormalized densities*, rather than learning from target samples as in standard generative training.

Existing methods for generative online RL can be grouped into four categories (see Appendix A for a detailed discussion of related work). *Direct backpropagation* (Wang et al., 2024; Lyu et al., 2025; Celik et al., 2025) optimizes critic-based objectives by differentiating through the entire sampling trajectory, which incurs substantial computational cost and can destabilize training. *Value-guided sampling* (Psenka et al., 2024; Wei et al., 2026) uses critic information to guide sampling and exploration through stochastic or deterministic dynamics. *Weighted regression* (Ding et al., 2024; Fan et al., 2025; Ma et al., 2025) trains the policy on candidate actions using value-dependent weights, adding sampling overhead and making the training target depend on the proposal distribution and weighting rule. These approaches rely on surrogate objectives or sampling approximations that can introduce bias relative to exact soft policy improvement. In contrast, *Boltzmann targeting* (Dong et al., 2025; Ma et al., 2025; Jain et al., 2025; Li et al., 2026) derives training signals directly from the Boltzmann density, recovering the desired policy at the ideal population optimum. In practice, however, the required posterior means are typically estimated through multi-particle importance sampling. This incurs computational overhead at each update and can produce high-variance estimates that destabilize training.

In this paper, we propose *Score-Calibrated Flow* (SCF), a simple and efficient algorithm for learning to sample from unnormalized densities. Our key idea is to learn the desired flow by enforcing self-consistency, bypassing the challenging estimation of the target posterior mean. By jointly exploiting the prescribed target score and the structure of flow matching, we establish these self-consistency requirements as *score-calibrated optimality conditions*. We first derive conditions that characterize the target density, then develop corresponding conditions for the trainable velocity field.

The velocity condition takes the form of a fixed-point equation whose conditional-expectation structure enables training with self-generated samples. Specifically, SCF generates an endpoint with the current flow, interpolates it with independent source noise, and constructs a regression target whose conditional mean equals the right-hand side of the fixed-point equation. This retains the scalable sample–interpolate–regress structure of CFM despite the absence of target samples, without importance sampling or backpropagation through the sampling trajectory.

Our contributions are summarized as follows.

1. **Score-calibrated optimality conditions.** We establish necessary and sufficient self-consistency conditions for both the terminal density and the trainable velocity field. Their unique solutions are, respectively, the target density and the ideal flow model that CFM would recover if target samples were available.
2. **Scalable training from self-generated samples.** We construct a sample-wise regression target whose conditional mean equals the score-calibration operator applied to the current velocity. We use this target in a stop-gradient objective to enforce the velocity condition. Each target requires only one self-generated endpoint and one target-score evaluation, avoiding importance sampling and backpropagation through the sampling trajectory.
3. **Principled design and strong RL performance.** We derive coefficient schedules that cancel source-end singular factors and handle bounded actions through an exact change of variables that preserves the intended target density. Experiments on online RL benchmarks demonstrate that SCF matches or improves upon state-of-the-art generative-policy baselines, with substantial savings in training time.

2 Background

2.1 Sampling from Unnormalized Densities

We consider sampling from a target density on $\mathbb{R}^d$ defined by a continuously differentiable potential $Q : \mathbb{R}^d \to \mathbb{R}$ and an inverse temperature $\tau > 0$:

$$\pi(x) = \frac{\exp(\tau Q(x))}{\displaystyle\int_{\mathbb{R}^d} \exp(\tau Q(z))\,\mathrm{d}z}. \tag{1}$$

The normalizing integral is assumed finite, and the target score is $\nabla \log \pi(x) = \tau \nabla Q(x)$. Our goal is to learn a sampler for π given access to evaluations of $Q(x)$ and its gradient $\nabla Q(x)$, but not to the normalizing constant or direct target samples.

We first develop our method in this general setting, where $x \in \mathbb{R}^d$ denotes a point in data space. We later specialize it to RL, reusing Q to denote the state–action value function $Q(s, a)$. At each fixed state s, this function serves as the sampling potential over actions a.

2.2 Flow Matching

Flow matching learns a time-dependent velocity field that transports a source density p_0 to a target density p_1 (Lipman et al., 2023). We adopt the standard setting of a Gaussian source $p_0 = \mathcal{N}(0, I)$, an independent coupling of endpoints $X_0 \sim p_0$ and $X_1 \sim p_1$, and linear interpolation:

$$X_t = \alpha_t X_1 + \beta_t X_0, \qquad t \in [0, 1]. \tag{2}$$

The interpolation schedules $\alpha, \beta \in C^1([0, 1])$ satisfy $\alpha_0 = \beta_1 = 0$, $\alpha_1 = \beta_0 = 1$, and $\alpha_t, \beta_t > 0$ for $t \in (0, 1)$.

For $t \in (0, 1)$, one-sided conditioning on $X_1 = x_1$ gives the conditional probability path

$$p_{t|1}(x_t|x_1) = \mathcal{N}(x_t; \alpha_t x_1, \beta_t^2 I), \tag{3}$$

with associated conditional velocity

$$v_{t|1}(x_t|x_1) = \dot{\alpha}_t x_1 + \dot{\beta}_t \frac{x_t - \alpha_t x_1}{\beta_t}. \tag{4}$$

Conditional flow matching (CFM) uses this velocity as a regression target to train a parameterized field v^θ:

$$\mathcal{L}_{\mathrm{CFM}}(\theta) := \mathbb{E}_{t\sim \mathrm{U}(0,1),\, X_1\sim p_1,\, X_t\sim p_{t|1}(\cdot|X_1)} \left[\left\| v_t^\theta(X_t) - v_{t|1}(X_t|X_1) \right\|_2^2 \right].$$

The population minimizer is the conditional mean of the regression target given (t, X_t).

To generate samples with the learned field v^θ, we solve the ordinary differential equation (ODE)

$$\frac{\mathrm{d}Y_t}{\mathrm{d}t} = v_t^\theta(Y_t), \qquad Y_0 \sim p_0, \qquad t \in [0, 1].$$

Here Y_t denotes the sampling trajectory, while X_0, X_t, X_1 are reserved for the interpolant used to construct training pairs. With perfect CFM training, the terminal state Y_1 has density p_1. This sample–interpolate–regress construction requires direct target samples $X_1 \sim p_1$, which are unavailable in the unnormalized-density setting considered in this paper.

2.3 Maximum-Entropy Reinforcement Learning

We consider a Markov decision process (MDP) $(\mathcal{S}, \mathcal{A}, P, r, d_0, \gamma)$, with state and action spaces $\mathcal{S}, \mathcal{A}$, transition kernel P, reward function r, initial state distribution d_0, and discount factor $\gamma \in (0, 1)$. In the RL setting, $\pi(a|s)$ denotes a policy. Maximum-entropy RL (Haarnoja et al., 2018) augments the expected discounted return with policy entropy $\mathcal{H}(\pi(\cdot|s)) = -\mathbb{E}_{a\sim\pi(\cdot|s)}[\log \pi(a|s)]$, weighted by $\frac{1}{\tau}$.

For a policy π, the regularized self-consistency operator is given by

$$(\mathcal{B}^\pi Q)(s, a) = r(s, a) + \gamma \mathbb{E}_{s'\sim P(\cdot|s,a)} \left[\mathbb{E}_{a'\sim\pi(\cdot|s')}[Q(s', a')] + \frac{1}{\tau}\mathcal{H}(\pi(\cdot|s')) \right].$$

The Bellman operator is given by $(\mathcal{B}Q)(s, a) = \max_\pi\{(\mathcal{B}^\pi Q)(s, a)\}$. The soft Q-function Q^π and its optimal counterpart Q^* satisfy $Q^\pi = \mathcal{B}^\pi Q^\pi$ and $Q^* = \mathcal{B}Q^*$.

Holding $Q = Q^\pi$ fixed, soft policy improvement maximizes $\mathbb{E}_{a\sim\pi(\cdot|s)}[Q(s, a)] + \frac{1}{\tau}\mathcal{H}(\pi(\cdot|s))$ at each state, yielding the critic-prescribed Boltzmann policy

$$\pi^Q(a|s) = \frac{\exp(\tau Q(s, a))}{\displaystyle\int_{\mathcal{A}} \exp(\tau Q(s, a'))\, \mathrm{d}a'}. \tag{5}$$

Thus, soft policy improvement requires sampling from an unnormalized density specified by the critic.

3 SCORE-CALIBRATED FLOW

In this section, we develop Score-Calibrated Flow (SCF) for learning generative models to sample from unnormalized densities. The key paradigm shift enabling SCF is to reinterpret the flow matching framework as a one-to-one correspondence between terminal densities and their associated ideal flow models, rather than solely as a procedure for learning from target samples. Each terminal density determines the population minimizer of the CFM loss, which is the velocity field that CFM would learn if samples from that density were available. Conversely, pushing the source density through the resulting flow induces the terminal density.

We exploit the two directions of this correspondence to develop our optimality conditions. First, we view the flow model as a function of its terminal density and derive a score-calibrated condition that characterizes the target density through the flow it induces. We then let the trainable velocity field determine its own terminal density through the pushforward map, yielding a self-consistency condition directly on the velocity. This condition takes the form of a fixed-point equation whose conditional-expectation structure enables scalable regression using self-generated samples. Finally, we develop principled design choices to address practical challenges, including numerical singularities and bounded domains, and demonstrate how to apply SCF to generative online RL.

3.1 SCORE-CALIBRATED OPTIMALITY CONDITION FOR TERMINAL DENSITY

We retain the standard choices from Section 2.2: a Gaussian source $p_0 = \mathcal{N}(0, I)$, independent coupling, and the linear interpolant (2). Under these choices, the conditional probability path $p_{t|1}$ and conditional velocity $v_{t|1}$ in (3)–(4) are also fixed.

In contrast, the marginal probability path and endpoint posterior depend on the terminal density. For a candidate density ρ, we take $X_1 \sim \rho$ and use a superscript to make this dependence explicit:

$$p_t^\rho(x_t) := \int_{\mathbb{R}^d} \rho(x_1) p_{t|1}(x_t|x_1)\,\mathrm{d}x_1, \qquad p_{1|t}^\rho(x_1|x_t) := \frac{\rho(x_1)p_{t|1}(x_t|x_1)}{p_t^\rho(x_t)}.$$

The latter is the posterior density of X_1 given the interpolant $X_t = x_t$. It determines the ideal flow model associated with ρ:

$$v_t^\rho(x_t) := \mathbb{E}_{X_1 \sim p_{1|t}^\rho(\cdot|x_t)}[v_{t|1}(x_t|X_1)]. \tag{6}$$

This is the velocity field that CFM would learn from samples of ρ. Its flow generates the marginal path p_t^ρ and transports the source density to ρ. Thus, the posterior, marginal probability path, and marginal velocity vary together with the terminal density, while the conditional probability path and conditional velocity remain unchanged.

To relate the flow model to the terminal score, define

$$\kappa_t := \frac{\beta_t}{\alpha_t}(\dot\alpha_t\beta_t - \alpha_t\dot\beta_t), \qquad t \in (0,1). \tag{7}$$

We assume $\kappa_t > 0$, equivalently $\frac{\mathrm{d}}{\mathrm{d}t}\log\left(\frac{\alpha_t}{\beta_t}\right) > 0$. This requires the ratio $\frac{\alpha_t}{\beta_t}$ to increase along the noise-to-data path, a condition commonly satisfied by standard diffusion and flow schedules. The following identity expresses v_t^ρ in terms of $\nabla \log \rho$.

Proposition 3.1. *Let ρ be a positive, normalized C^1 density satisfying the regularity condition (21) in Appendix B.1. For every $t \in (0,1)$ and $x_t \in \mathbb{R}^d$,*

$$v_t^\rho(x_t) = \frac{\dot\alpha_t}{\alpha_t}x_t + \frac{\kappa_t}{\alpha_t}\mathbb{E}_{X_1 \sim p_{1|t}^\rho(\cdot|x_t)}[\nabla \log \rho(X_1)]. \tag{8}$$

Applying (8) to the target density π, whose score is $\tau\nabla Q$, gives

$$v_t^\pi(x_t) = \frac{\dot\alpha_t}{\alpha_t}x_t + \frac{\kappa_t}{\alpha_t}\mathbb{E}_{X_1 \sim p_{1|t}^\pi(\cdot|x_t)}[\tau\nabla Q(X_1)]. \tag{9}$$

We now take a different perspective on this equation: rather than using the target density to describe its associated flow, we seek to characterize the target density through the flow it induces under the CFM construction. Specifically, we treat the terminal density ρ as the unknown, allowing it to determine both the marginal velocity v^ρ and the posterior $p_{1|t}^\rho$, while keeping the prescribed target score $\tau\nabla Q$ fixed. The following theorem establishes the resulting self-consistency equation as a score-calibrated optimality condition for the terminal density.

Theorem 3.2. *Let π be the target density defined in (1). Let ρ be a positive, normalized C^1 density satisfying regularity conditions (21) and (22) in Appendix B.1. Define the density optimality condition*

$$v_t^\rho(x_t) = \frac{\dot\alpha_t}{\alpha_t} x_t + \frac{\kappa_t}{\alpha_t} \mathbb{E}_{X_1 \sim p^\rho_{1|t}(\cdot|x_t)}[\tau \nabla Q(X_1)] \quad \textit{for every } x_t \in \mathbb{R}^d. \tag{10}$$

Then

$$\rho = \pi \quad \Leftrightarrow \quad (10) \textit{ holds for every } t \in (0,1) \quad \Leftrightarrow \quad (10) \textit{ holds for some } t \in (0,1).$$

The sufficiency of $\rho = \pi$ follows directly from (9), which holds at every interior time. The necessity is nontrivial: satisfying (10) for every x_t at even a single $t \in (0,1)$ forces $\rho = \pi$. Thus, requiring the marginal velocity v^ρ to agree with the velocity reconstructed from the prescribed target score $\tau \nabla Q$ under the posterior $p^\rho_{1|t}$ uniquely identifies $\rho = \pi$. Enforcing this relation for every x_t at a single interior time is equivalent to enforcing it throughout the interpolation path.

Generalized optimality condition for terminal density. The preceding condition identifies the desired density through its associated flow model. We next generalize its form without changing its solution, introducing coefficients that will later allow us to address numerical singularities and adjust the sample-wise training signal.

First, for any $\lambda_t \neq 0$, we can equivalently write (10) as

$$v_t^\rho(x_t) = (1-\lambda_t) v_t^\rho(x_t) + \lambda_t \left(\frac{\dot\alpha_t}{\alpha_t} x_t + \frac{\kappa_t}{\alpha_t} \mathbb{E}_{X_1 \sim p^\rho_{1|t}(\cdot|x_t)}[\tau \nabla Q(X_1)] \right). \tag{11}$$

The coefficient λ_t controls the weight of the score-based term while preserving $\rho = \pi$ as the unique solution. Next, recall that the marginal velocity is the conditional expectation of $v_{t|1}$. We can therefore express part of the remaining velocity term through this relationship: for any $\zeta_t \in \mathbb{R}$,

$$(1-\lambda_t) v_t^\rho(x_t) = (1-\lambda_t-\zeta_t) v_t^\rho(x_t) + \zeta_t \mathbb{E}_{X_1 \sim p^\rho_{1|t}(\cdot|x_t)}[v_{t|1}(x_t|X_1)]. \tag{12}$$

Substituting (12) into (11) gives the following generalized optimality condition.

Proposition 3.3. *Adopt the notation and assumptions of Theorem 3.2. Let $\lambda_t \neq 0$ and $\zeta_t \in \mathbb{R}$. Define the generalized density optimality condition*

$$\begin{aligned} v_t^\rho(x_t) =& (1-\lambda_t-\zeta_t) v_t^\rho(x_t) + \zeta_t \mathbb{E}_{X_1 \sim p^\rho_{1|t}(\cdot|x_t)}[v_{t|1}(x_t|X_1)] \\ &+ \lambda_t \left(\frac{\dot\alpha_t}{\alpha_t} x_t + \frac{\kappa_t}{\alpha_t} \mathbb{E}_{X_1 \sim p^\rho_{1|t}(\cdot|x_t)}[\tau \nabla Q(X_1)] \right) \quad \textit{for every } x_t \in \mathbb{R}^d. \end{aligned} \tag{13}$$

Then

$$\rho = \pi \quad \Leftrightarrow \quad (13) \textit{ holds for every } t \in (0,1) \quad \Leftrightarrow \quad (13) \textit{ holds for some } t \in (0,1).$$

The generalized condition retains the same unique solution while providing flexibility in the eventual training construction. The coefficient λ_t controls the strength of score calibration, while ζ_t determines how much of the marginal velocity is represented through the conditional velocity.

The optimality conditions for ρ provide an important first step by characterizing the desired terminal density. However, the object we train is the velocity field, not the density itself. To obtain an optimality condition directly on the velocity, we exploit the reverse direction of the correspondence: a velocity field determines its terminal density through the pushforward of the source. Using this generated density to determine the posterior leads to a self-consistency condition on the trainable velocity.

3.2 Score-Calibrated Optimality Condition for Velocity

For a velocity field v, let $\Phi^v_{0\to t}$ be the flow map of $\dot Y_t = v_t(Y_t)$ and define its terminal density by

$$\rho^v := (\Phi^v_{0\to 1})_\# p_0,$$

where $\#$ denotes pushforward. The regularity requirements for well-posedness of this definition are specified in Appendix B.1.

We now seek a velocity field that satisfies the score-based relation in (10), with the posterior determined by its own terminal density. To express this self-consistency requirement, define the *score-calibration operator* by

$$(\mathcal{I}v)_t(x_t) := \frac{\dot{\alpha}_t}{\alpha_t}x_t + \frac{\kappa_t}{\alpha_t}\mathbb{E}_{X_1\sim p^{\rho^v}_{1|t}(\cdot|x_t)}[\tau\nabla Q(X_1)] \tag{14}$$

and consider the fixed-point equation $v = \mathcal{I}v$. The same field supplies the velocity on the left and generates the terminal density used to form the posterior on the right. The desired field v^π satisfies this equation by (9). The following theorem establishes that this condition is both necessary and sufficient to identify the optimal velocity.

Theorem 3.4. *Let π be the target density defined in (1), and assume that the regularity condition (21) holds for $\rho = \pi$. Assume that v^π and v are admissible, as specified in Appendix B.1. Assume that v satisfies (23) and (24). Then, $v = \mathcal{I}v$ holds on $(0,1)\times\mathbb{R}^d$ if and only if $\rho^v = \pi$ and $v = v^\pi$.*

The nontrivial implication is that self-consistency forces both $\rho^v = \pi$ and $v = v^\pi$. Thus, the condition does not merely identify the desired terminal density; it uniquely determines the ideal flow model that CFM would learn if target samples were available. This completes the transition from an optimality condition on the density to one directly on the trainable velocity.

Generalized optimality condition for velocity. We next extend the two-coefficient construction in (13) to the velocity formulation. The explicit marginal-velocity term is replaced by the trainable field v, while its generated density ρ^v determines both posterior expectations. This gives the *generalized score-calibration operator*

$$\begin{aligned}(\mathcal{I}_{\lambda,\zeta}v)_t(x_t) :=& (1-\lambda_t-\zeta_t)v_t(x_t) + \zeta_t\mathbb{E}_{X_1\sim p^{\rho^v}_{1|t}(\cdot|x_t)}[v_{t|1}(x_t|X_1)] \\ &+ \lambda_t\left(\frac{\dot{\alpha}_t}{\alpha_t}x_t + \frac{\kappa_t}{\alpha_t}\mathbb{E}_{X_1\sim p^{\rho^v}_{1|t}(\cdot|x_t)}[\tau\nabla Q(X_1)]\right).\end{aligned} \tag{15}$$

The corresponding optimality condition is $v = \mathcal{I}_{\lambda,\zeta}v$. It combines the current velocity, the conditional-velocity expectation, and the target-score expression, with coefficients λ_t and ζ_t controlling their contributions. Setting $\lambda_t = 1$ and $\zeta_t = 0$ recovers the original condition $v = \mathcal{I}v$. The following theorem gives a coefficient condition under which the generalized velocity condition also uniquely identifies the desired flow model.

Theorem 3.5. *Assume the hypotheses of Theorem 3.4 on v and π, and let t_0 be as in (23). Let $\lambda_t > 0$ and $\zeta_t \geq 0$ be measurable, finite schedules on $(0,1)$ satisfying*

$$\int_0^{t_0} \frac{\lambda_t}{\lambda_t+\zeta_t}\frac{\kappa_t}{\beta_t^2}\,\mathrm{d}t = \infty. \tag{16}$$

Then $v = \mathcal{I}_{\lambda,\zeta}v$ holds on $(0,1)\times\mathbb{R}^d$ if and only if $\rho^v = \pi$ and $v = v^\pi$.

Under the conditions of Theorem 3.5, the original and generalized velocity equations have the same unique solution v^π. The coefficients therefore provide freedom to weight the conditional-velocity and target-score terms while preserving the desired flow model.

To clarify the relationship with the density formulation, consider a field $v = v^\rho$ obtained from the CFM construction. Its flow generates the terminal density ρ, so both operators use the posterior $p^\rho_{1|t}$. Moreover, by (6), the conditional expectation of $v_{t|1}$ under $p^\rho_{1|t}(\cdot|x_t)$ equals the input velocity $v^\rho_t(x_t)$. Consequently, the original velocity condition becomes (10), and the generalized velocity condition becomes (13). Consequently, under the assumptions of Theorem 3.2 and for $\lambda_t \neq 0$ on $(0,1)$,

$$\mathcal{I}v^\rho = v^\rho \quad\Longleftrightarrow\quad \mathcal{I}_{\lambda,\zeta}v^\rho = v^\rho \quad\Longleftrightarrow\quad \rho = \pi.$$

For a general trainable field v, the reduction above no longer applies, since v need not coincide with the ideal flow model associated with its generated density. Theorem 3.5 establishes a stronger characterization: under the stated assumptions and coefficient condition (16), the generalized velocity condition forces both $\rho^v = \pi$ and $v = v^\pi$, without presupposing that the candidate field is obtained through the CFM construction.

The preceding velocity theorems require a growth bound near the source and a weighted exponential moment of the generated density. The latter can be difficult to verify since ρ^v is defined implicitly by the sampling dynamics. We provide another result (Proposition B.5 in Appendix B.3) that gives sufficient conditions stated directly in terms of v and the target score $\tau\nabla Q$: an integrable linear-growth bound over the full sampling interval, together with polynomial growth of the target score, implies both requirements without evaluating ρ^v.

We now turn the generalized velocity condition into a training objective. Its fixed-point form asks the current prediction to equal the conditional expectation defining $\mathcal{I}_{\lambda,\zeta}v$. Rather than evaluate this expectation explicitly, we construct a sample-wise target from self-generated endpoints and independent source noise. Its conditional mean equals the operator output, so its difference from the current prediction provides a sampled correction for the fixed-point residual. The next subsection uses this construction in a stop-gradient objective.

3.3 Learning from Self-Generated Samples

We seek to enforce the generalized velocity condition through its fixed-point residual,

$$(\mathcal{I}_{\lambda,\zeta}v)_t(x_t) - v_t(x_t) = 0. \tag{17}$$

The conditional-expectation structure of the generalized score-calibration operator provides a sample-wise training signal. For an admissible field v, combine the two expectations in (15) and place the current-prediction term inside the same expectation. The resulting integrand defines the target

$$\mathcal{Y}_{v,\lambda,\zeta}(t,x_t,x_1) := \zeta_t v_{t|1}(x_t|x_1) + (1-\lambda_t-\zeta_t)v_t(x_t) + \lambda_t\left(\frac{\dot\alpha_t}{\alpha_t}x_t + \frac{\kappa_t}{\alpha_t}\tau\nabla Q(x_1)\right). \tag{18}$$

By construction, whenever the expectation exists,

$$\mathbb{E}_{X_1\sim p_{1|t}^{\rho^v}(\cdot|x_t)}[\mathcal{Y}_{v,\lambda,\zeta}(t,x_t,X_1)] = (\mathcal{I}_{\lambda,\zeta}v)_t(x_t).$$

For the current network v^θ, let $\rho^\theta := \rho^{v^\theta}$. Generate Y_1 with the current flow and set $X_1 := Y_1 \sim \rho^\theta$. Independently draw fresh source noise $X_0 \sim p_0$ and $t \sim \mathrm{U}(0,1)$, then form $X_t = \alpha_t X_1 + \beta_t X_0$. By Bayes' rule, the conditional density of X_1 given $(t, X_t = x_t)$ is $p_{1|t}^{\rho^\theta}(\cdot|x_t)$.

To seek a zero of the fixed-point residual in (17), we use the sampled discrepancy as a correction signal, treating the target as fixed during differentiation. This gives the stop-gradient objective

$$\mathcal{L}_{\lambda,\zeta}^{\mathrm{SCF}}(\theta) := \mathbb{E}_{t\sim\mathrm{U}(0,1),\, X_1\sim\rho^{\mathrm{sg}(\theta)},\, X_t\sim p_{t|1}(\cdot|X_1)}\left[\left\|v_t^\theta(X_t) - \mathrm{sg}(\mathcal{Y}_{v^\theta,\lambda,\zeta}(t,X_t,X_1))\right\|_2^2\right].$$

Here sg denotes the stop-gradient operator, which preserves the target value but blocks derivatives through all its terms.

It remains to specify the coefficients λ_t and ζ_t in the SCF objective. A key consideration is the singular behavior of the score-based term in (18) as $t \to 0$. The coefficients must therefore cancel this singularity for numerical stability while satisfying (16) to preserve fixed-point uniqueness. We provide concrete coefficient choices in Appendix B.4 and empirically demonstrate their effectiveness through ablation studies in Appendix D.

3.4 Bounded Domains

The preceding construction learns a sampler on $\mathbb{R}^d$. To accommodate bounded domains, such as the action spaces in many RL problems, we map an unconstrained latent variable into the desired domain and account for the change of variables in the target density. This allows us to retain the optimality conditions and training objective in latent space.

Consider a target $\pi(a) \propto e^{\tau Q(a)}$ on $(-1,1)^d$ and parameterize $a = \tanh x$ coordinatewise, with $x \in \mathbb{R}^d$. We use a tilde to denote the corresponding density in latent spaces. The change-of-variables formula gives

$$\widetilde\pi(x) = \pi(\tanh x)\prod_{i=1}^d(1-\tanh^2 x_i) \propto e^{\tau Q(\tanh x)}\prod_{i=1}^d(1-\tanh^2 x_i). \tag{19}$$

Applying $\tanh$ ensures bounded outputs, but matching the intended density additionally requires the Jacobian factor above. The following proposition gives the resulting score and establishes the equivalence between the latent and bounded-domain sampling problems.

Proposition 3.6. *Let $Q \in C^1((-1,1)^d)$ and let π be a probability density on $(-1,1)^d$ satisfying $\pi(a) \propto e^{\tau Q(a)}$. The latent target has score*

$$\nabla_x \log \widetilde{\pi}(x) = \tau \nabla_x Q(\tanh x) - 2 \tanh x. \tag{20}$$

Moreover, $Y_1 \sim \widetilde{\pi}$ if and only if $\tanh Y_1 \sim \pi$.

We therefore apply SCF in latent space with target $\widetilde{\pi}$, replacing $\tau \nabla Q$ in the optimality conditions and regression target with the latent score in (20). Interpolation and training otherwise remain unchanged; during sampling, we apply $\tanh$ to the generated endpoint to obtain a sample in the original bounded domain.

4 Applications to Generative Online Reinforcement Learning

We apply SCF within an off-policy actor–critic framework. The overall procedure is summarized in Algorithm 1 in Appendix C. Without loss of generality, we take $\mathcal{A} = (-1,1)^d$ after affine rescaling of action bounds. The flow policy is parameterized by $v^\theta(\cdot, s)$, and $\pi^\theta(\cdot|s)$ denotes its induced action density. Sampling $a \sim \pi^\theta(\cdot|s)$ consists of solving

$$\frac{\mathrm{d}Y_t}{\mathrm{d}t} = v_t^\theta(Y_t, s), \qquad Y_0 \sim p_0,$$

and returning $a = \tanh Y_1$.

Critic training. We maintain two critics Q^{ω_1} and Q^{ω_2}, together with target networks parameterized by $\bar{\omega}_1$ and $\bar{\omega}_2$. Let $\mathcal{D}$ denote the replay buffer. Each critic, indexed by $i \in \{1,2\}$, minimizes

$$\mathcal{L}_Q(\omega_i) := \mathbb{E}_{(s,a,r,s') \sim \mathcal{D},\, a' \sim \pi^\theta(\cdot|s')} \left[\left(Q^{\omega_i}(s,a) - \mathrm{sg}\left(r + \gamma \min_{j \in \{1,2\}} Q^{\bar{\omega}_j}(s',a') \right) \right)^2 \right].$$

Actor training. Holding the state and critics fixed, SCF trains the flow policy toward the Boltzmann density $\pi^Q(\cdot|s)$ defined in (5). We use $Q(s,a) := \min_{i \in \{1,2\}} Q^{\omega_i}(s,a)$ as the sampling potential. Let $\widetilde{\pi}^\theta(\cdot|s)$ denote the density of the latent endpoint Y_1, so that $\pi^\theta(\cdot|s) = (\tanh)_\# \widetilde{\pi}^\theta(\cdot|s)$. The corresponding latent target $\widetilde{\pi}^Q(\cdot|s)$ is obtained from $\pi^Q(\cdot|s)$ through the change of variables in Section 3.4, with score $\nabla_x \log \widetilde{\pi}^Q(x|s) = \tau \nabla_x Q(s, \tanh x) - 2 \tanh x$. At each fixed state s, $\widetilde{\pi}^Q(\cdot|s)$ and $\widetilde{\pi}^\theta(\cdot|s)$ correspond, respectively, to the target density π and generated density ρ^θ in the general formulation developed in the preceding sections. The resulting actor loss is given by

$$\mathcal{L}^{\mathrm{actor}}_{\lambda,\zeta}(\theta) := \mathbb{E}_{s \sim \mathcal{D},\, t \sim \mathrm{U}(0,1),\, X_1 \sim \widetilde{\pi}^{\mathrm{sg}(\theta)}(\cdot|s),\, X_t \sim p_{t|1}(\cdot|X_1)} \left[\left\| v_t^\theta(X_t, s) - \mathrm{sg}(\mathcal{Y}_{v^\theta,\lambda,\zeta}(t, X_t, X_1; s)) \right\|_2^2 \right].$$

5 Experiments

5.1 Toy Example

We first validate SCF on a toy example, $\pi(x) \propto \exp(\tau Q(x))$, $x \in \mathbb{R}^2$, an equally weighted mixture of eight Gaussian components arranged on a circle. We compare with four recent generative samplers for unnormalized densities: Adjoint Sampling (AS) (Havens et al., 2025), Adjoint Schrödinger Bridge Sampler (ASBS) (Liu et al., 2025), Bridge Matching Sampler (BMS) (Blessing et al., 2026), and Flow Sampling (FS) (Havens et al., 2026). The results are summarized in Figure 1. Experimental details are provided in Appendix C.1. SCF achieves the lowest discrepancy from ground-truth samples while being the most efficient in terms of training and inference.

5.2 RL Tasks

We evaluate SCF on two sets of tasks: widely used continuous-control tasks from the DeepMind Control Suite (Tassa et al., 2018) and challenging humanoid-control tasks from HumanoidBench (Sferrazza et al., 2024). We compare SCF with soft actor-critic (SAC) (Haarnoja et al., 2018),

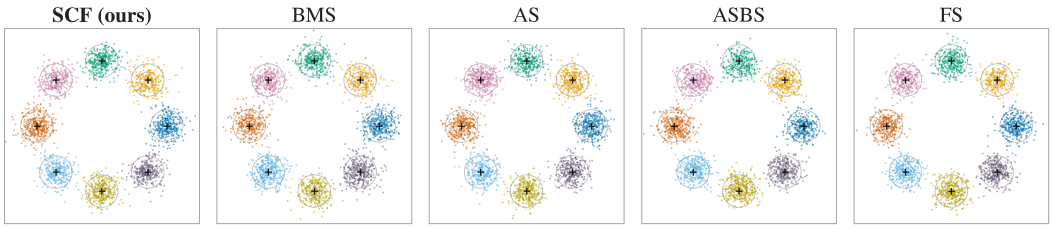

Figure 1: Visualization of samples from the five trained models on the eight-Gaussian target. SCF achieves the lowest discrepancy from ground-truth samples while being the cheapest to train and to sample from. The detailed descriptions are provided in Appendix C.1.

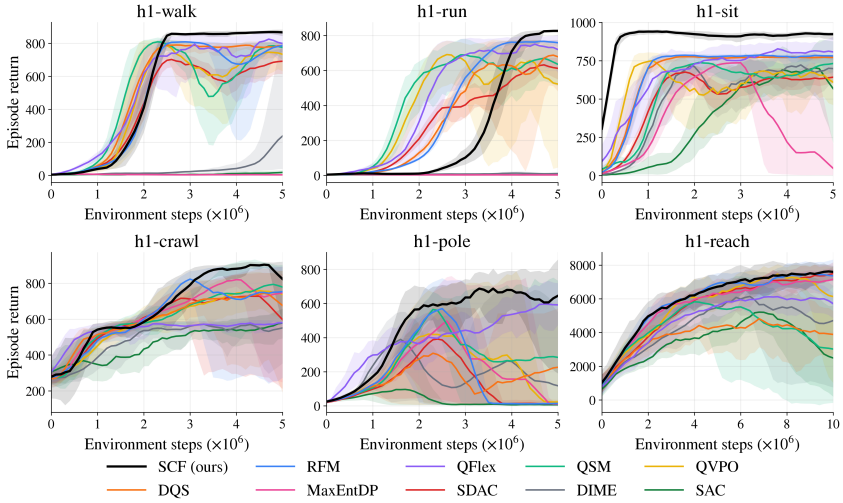

Figure 2: Training curves on six tasks from HumanoidBench. The solid lines correspond to the mean across five seeds, and the shaded regions represent minimum-to-maximum range.

a standard maximum-entropy RL algorithm with Gaussian policies, and eight strong generative-policy baselines spanning the four categories introduced in Section 1: DIME (Celik et al., 2025), Q-score matching (QSM) (Psenka et al., 2024), QFlex (Wei et al., 2026), Q-weighted variational policy optimization (QVPO) (Ding et al., 2024), MaxEntDP (Dong et al., 2025), soft diffusion actor-critic (SDAC) (Ma et al., 2025), diffusion Q-sampling (DQS) (Jain et al., 2025), and reverse flow matching (RFM) (Li et al., 2026). Experimental details are provided in Appendix C.2.

The training curves for HumanoidBench tasks are shown in Figure 2. The training curves for DeepMind Control Suite tasks are deferred to Figure 3 in Appendix C.2. The results demonstrate that SCF consistently ranks among the top-performing methods across all tasks. Table 2 in Appendix C.2 reports the wall-clock time per training iteration for each algorithm, with SCF requiring less time than all generative-policy baselines.

6 Conclusion

In this paper, we proposed Score-Calibrated Flow (SCF), a simple and efficient method for learning to sample from unnormalized densities with applications to generative online RL. By shifting from target posterior-mean estimation to self-consistency enforcement, we derived score-calibrated optimality conditions for both terminal densities and velocity fields. Their conditional-expectation structure yields a stop-gradient regression objective using self-generated samples. Experiments demonstrated strong performance and computational efficiency of SCF.

AI USE STATEMENT

We used AI tools to improve the readability of the manuscript and assist with code implementation. The authors reviewed all AI-assisted work and take full responsibility for the final manuscript and code.

ETHICS STATEMENT

This work does not involve human subjects or personal data. All experiments use synthetic data or publicly available simulators.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, the pseudocode of SCF and implementation details are provided in Appendix C, and the source code used in our experiments is available to reviewers upon request.

REFERENCES

Tara Akhound-Sadegh, Jarrid Rector-Brooks, Joey Bose, Sarthak Mittal, Pablo Lemos, Cheng-Hao Liu, Marcin Sendera, Siamak Ravanbakhsh, Gauthier Gidel, Yoshua Bengio, Nikolay Malkin, and Alexander Tong. Iterated denoising energy matching for sampling from Boltzmann densities. In *International Conference on Machine Learning*, volume 235, pp. 760–786, 2024.

Denis Blessing, Lorenz Richter, Julius Berner, Egor Malitskiy, and Gerhard Neumann. Bridge matching sampler: Scalable sampling via generalized fixed-point diffusion matching. In *International Conference on Machine Learning*, 2026.

Onur Celik, Zechu Li, Denis Blessing, Ge Li, Daniel Palenicek, Jan Peters, Georgia Chalvatzaki, and Gerhard Neumann. DIME: Diffusion-based maximum entropy reinforcement learning. In *International Conference on Machine Learning*, volume 267, pp. 6958–6977, 2025.

Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, 44(10–11):1684–1704, 2025.

Shutong Ding, Ke Hu, Zhenhao Zhang, Kan Ren, Weinan Zhang, Jingyi Yu, Jingya Wang, and Ye Shi. Diffusion-based reinforcement learning via Q-weighted variational policy optimization. In *Advances in Neural Information Processing Systems*, volume 37, pp. 53945–53968, 2024.

Xiaoyi Dong, Jian Cheng, and Xi Sheryl Zhang. Maximum entropy reinforcement learning with diffusion policy. In *International Conference on Machine Learning*, volume 267, pp. 13963–13983, 2025.

Jiajun Fan, Shuaike Shen, Chaoran Cheng, Yuxin Chen, Chumeng Liang, and Ge Liu. Online reward-weighted fine-tuning of flow matching with Wasserstein regularization. In *International Conference on Learning Representations*, 2025.

Gerald B. Folland. *Real Analysis: Modern Techniques and Their Applications*. John Wiley & Sons, second edition, 1999.

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*, volume 80, pp. 1861–1870, 2018.

Aaron J. Havens, Benjamin Kurt Miller, Bing Yan, Carles Domingo-Enrich, Anuroop Sriram, Daniel S. Levine, Brandon M. Wood, Bin Hu, Brandon Amos, Brian Karrer, Xiang Fu, Guan-Horng Liu, and Ricky T. Q. Chen. Adjoint sampling: Highly scalable diffusion samplers via adjoint matching. In *International Conference on Machine Learning*, volume 267, pp. 22204–22237, 2025.

Aaron J. Havens, Brian Karrer, and Neta Shaul. Flow sampling: Learning to sample from unnormalized densities via denoising conditional processes. In *International Conference on Machine Learning*, 2026.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020.

Vineet Jain, Tara Akhound-Sadegh, and Siamak Ravanbakhsh. Sampling from energy-based policies using diffusion. *Reinforcement Learning Journal*, 6:2291–2307, 2025.

Zeyang Li, Sunbochen Tang, and Navid Azizan. Reverse flow matching: A unified framework for online reinforcement learning with diffusion and flow policies. In *International Conference on Machine Learning*, 2026.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023.

Guan-Horng Liu, Jaemoo Choi, Yongxin Chen, Benjamin Kurt Miller, and Ricky T. Q. Chen. Adjoint Schrödinger bridge sampler. In *Advances in Neural Information Processing Systems*, volume 38, pp. 15673–15708, 2025.

Lei Lyu, Yunfei Li, Yu Luo, Fuchun Sun, Tao Kong, Jiafeng Xu, and Xiao Ma. Flow-based policy for online reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 38, pp. 93967–93990, 2025.

Haitong Ma, Tianyi Chen, Kai Wang, Na Li, and Bo Dai. Efficient online reinforcement learning for diffusion policy. In *International Conference on Machine Learning*, volume 267, pp. 41837–41853, 2025.

Frank Noé, Simon Olsson, Jonas Köhler, and Hao Wu. Boltzmann generators: Sampling equilibrium states of many-body systems with deep learning. *Science*, 365(6457):eaaw1147, 2019.

George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, 22(57):1–64, 2021.

Michael Psenka, Alejandro Escontrela, Pieter Abbeel, and Yi Ma. Learning a diffusion model policy from rewards via Q-score matching. In *International Conference on Machine Learning*, volume 235, pp. 41163–41182, 2024.

Walter Rudin. *Real and Complex Analysis*. McGraw-Hill, third edition, 1987.

Carmelo Sferrazza, Dun-Ming Huang, Xingyu Lin, Youngwoon Lee, and Pieter Abbeel. HumanoidBench: Simulated humanoid benchmark for whole-body locomotion and manipulation. In *Robotics: Science and Systems*, 2024.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, volume 37, pp. 2256–2265, 2015.

Yuval Tassa, Yotam Doron, Alistair Muldal, Tom Erez, Yazhe Li, Diego de Las Casas, David Budden, Abbas Abdolmaleki, Josh Merel, Andrew Lefrancq, Timothy Lillicrap, and Martin Riedmiller. DeepMind control suite. *arXiv preprint arXiv:1801.00690*, 2018.

Francisco Vargas, Will Grathwohl, and Arnaud Doucet. Denoising diffusion samplers. In *International Conference on Learning Representations*, 2023.

Yinuo Wang, Likun Wang, Yuxuan Jiang, Wenjun Zou, Tong Liu, Xujie Song, Wenxuan Wang, Liming Xiao, Jiang Wu, Jingliang Duan, and Shengbo Eben Li. Diffusion actor-critic with entropy regulator. In *Advances in Neural Information Processing Systems*, volume 37, pp. 54183–54204, 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*, 2023.

Yunyue Wei, Chenhui Zuo, and Yanan Sui. Scalable exploration for high-dimensional continuous control via value-guided flow. In *International Conference on Learning Representations*, 2026.

Qinsheng Zhang and Yongxin Chen. Path integral sampler: A stochastic control approach for sampling. In *International Conference on Learning Representations*, 2022.

A Related Work

A.1 Generative Online Reinforcement Learning

We organize existing methods by how they construct supervision signals for actor training.

Direct backpropagation. DACER (Wang et al., 2024), FlowRL (Lyu et al., 2025), and DIME (Celik et al., 2025) differentiate critic-based objectives through the sampling trajectory, incorporating additional approximations and regularization to make policy optimization more stable. Specifically, DACER estimates policy entropy by fitting Gaussian mixture models to sampled actions, FlowRL uses a flow-matching surrogate for Wasserstein regularization toward high-value replay-buffer behaviors, and DIME optimizes a tractable lower bound on the maximum-entropy objective. Despite these measures, backpropagation through the full sampling trajectory incurs substantial computational and memory costs and poses a nonnegligible risk of training instability.

Value-guided sampling. QSM (Psenka et al., 2024) trains a diffusion score network using critic gradients at noisy actions, while QFlex (Wei et al., 2026) performs critic-gradient ascent from a learned source and fits a flow model to the resulting endpoints. Unlike our method, QSM and QFlex do not establish an exact characterization that uniquely identifies both the prescribed Boltzmann density and its velocity from CFM construction. Their guidance relies on repeated action-refinement steps, and evaluating the critic at noisy actions can expose it to inputs where its estimates are unreliable. We instead derive regression targets from score-calibrated optimality conditions and evaluate the critic gradient at clean actions. Each target requires only one endpoint-score evaluation, and the continuous-time formulation is independent of the number of sampling steps, without requiring an additional multi-step guidance procedure.

Weighted regression. QVPO (Ding et al., 2024) trains a diffusion policy with Q-dependent weights and entropy regularization. Related approaches include reward-weighted flow matching (Fan et al., 2025) and diffusion policy mirror descent (Ma et al., 2025). These methods use importance sampling to construct weighted regression objectives and therefore remain susceptible to high-variance training signals when the importance weights become concentrated. Generating and evaluating multiple candidate actions further increases computational cost. Moreover, their induced training densities depend on the proposal distribution and weighting rule and generally differ from the prescribed Boltzmann target π^Q, so even exact optimization may not recover the desired policy.

Boltzmann targeting. MaxEntDP (Dong et al., 2025) and SDAC (Ma et al., 2025) use Q-weighted noise signals to train noise-prediction and score networks, respectively. DQS (Jain et al., 2025) adapts iterated denoising energy matching (Akhound-Sadegh et al., 2024) to train a score network using Q-weighted averages of critic gradients. RFM (Li et al., 2026) identifies these noise-expectation and gradient-expectation methods as instances of a shared posterior-mean estimation problem, and provides a principled way to combine both families through Stein control variates for variance reduction. These formulations target the Boltzmann density at the ideal population optimum, but their practical estimators rely on importance sampling, incurring computational overhead and estimator variance. SCF instead enforces score-calibrated self-consistency through regression on self-generated endpoints, without explicitly estimating the target posterior mean.

A.2 Learning to Sample from Unnormalized Densities

Learning generative samplers from unnormalized densities has been extensively studied, with representative approaches based on normalizing flows (Noé et al., 2019; Papamakarios et al., 2021) and diffusion models (Zhang & Chen, 2022; Vargas et al., 2023; Akhound-Sadegh et al., 2024). We focus on recent matching-based methods that avoid both importance-weighted training targets and backpropagation through sampling trajectories.

Adjoint Sampling (AS) (Havens et al., 2025) derives an endpoint-gradient regression objective from stochastic optimal control. Adjoint Schrödinger Bridge Sampler (ASBS) (Liu et al., 2025) extends this construction beyond memoryless reference processes by alternating between learning a diffusion drift and a terminal corrector. Bridge Matching Sampler (BMS) (Blessing et al., 2026) derives

diffusion-matching targets from a fixed-point iteration perspective. Flow Sampling (FS) (Havens et al., 2026) constructs regression targets by combining the conditional velocity with an endpoint-score correction.

SCF grounds its training rule in optimality conditions that uniquely characterize the target density and ideal flow velocity under mild regularity requirements. BMS and FS adopt a fixed-point iteration perspective on training with self-generated samples, but do not establish that the desired sampler is the unique fixed point (Blessing et al., 2026; Havens et al., 2026). AS establishes fixed-point uniqueness only under the additional assumption of reciprocal-projection invariance (Havens et al., 2025). This is closer in spirit to our density optimality condition, where the flow is prescribed by the candidate density. Our velocity optimality condition goes further: it rules out spurious fixed points without presupposing a density–flow consistency. ASBS establishes convergence of its alternating updates under the assumption that each stage reaches its critical point, but does not establish uniqueness of the joint critical points of its coupled drift–corrector matching equations (Liu et al., 2025). SCF instead derives necessary and sufficient conditions directly for the terminal density and trainable velocity. Their unique solutions are the target density π and the ideal velocity v^π that CFM would learn from target samples. Moreover, the four methods above learn the drifts of stochastic differential equations (SDEs). Their sampling procedures rely on stochastic simulation, which may require a large number of function evaluations (NFEs), increasing computational costs during both training and inference. FS illustrates a structural limitation of this formulation: its target-score correction is proportional to the squared diffusion coefficient, so setting this coefficient to zero eliminates explicit target-score supervision (Havens et al., 2026). The remaining objective reduces to CFM on self-generated endpoints, which reproduces the current density without identifying the prescribed target. Thus, simply removing the diffusion noise does not yield a training objective that ensures the learned velocity field generates the prescribed target density through ODE sampling. In contrast, SCF derives its velocity regression objective from score-calibrated optimality conditions, retaining target-score supervision without relying on stochastic sampling dynamics.

B Proofs of Theoretical Results

We state the mild regularity conditions in Appendix B.1 and provide the proofs in the subsequent subsections. For convenience, we restate each proposition or theorem from the main text before its proof.

B.1 Regularity Conditions

The schedules satisfy the assumptions in Section 2.2 and $\kappa_t > 0$ from (7). Identities with divisions by α_t or β_t are stated for $t \in (0,1)$. Velocity equalities are understood on $(0,1) \times \mathbb{R}^d$.

Score identities. For every density ρ to which Proposition 3.1 is applied, assume positivity, normalization, and C^1 regularity. For every $t \in (0,1)$ and $x_t \in \mathbb{R}^d$, assume

$$\mathbb{E}_{X_1 \sim p^\rho_{1|t}(\cdot|x_t)}[\|X_1\|_2] < \infty, \qquad \mathbb{E}_{X_1 \sim p^\rho_{1|t}(\cdot|x_t)}[\|\nabla \log \rho(X_1)\|_2] < \infty. \tag{21}$$

For every density ρ to which Theorem 3.2 is applied, additionally assume, for every $t \in (0,1)$ and $x_t \in \mathbb{R}^d$, that

$$\mathbb{E}_{X_1 \sim p^\rho_{1|t}(\cdot|x_t)}[\|\nabla Q(X_1)\|_2] < \infty. \tag{22}$$

Velocity fields. We call a velocity field v *admissible* if its source-initialized ODE

$$\frac{\mathrm{d}}{\mathrm{d}t}\Phi^v_{0\to t}(x_0) = v_t(\Phi^v_{0\to t}(x_0)), \qquad \Phi^v_{0\to 0}(x_0) = x_0,$$

has a unique solution on $[0,1]$ with measurable flow maps, its terminal pushforward has a positive C^1 density ρ^v, and the conditional moments defining the score-calibration operators are finite.

For a population CFM field v^ρ, assume that its ODE transports p_0 along p^ρ_t to ρ. These requirements also apply to v^π when asserting existence of the target flow.

For every admissible field v to which Theorem 3.4 is applied, assume that, for some $t_0 \in (0,1)$, $\delta > 0$, and nonnegative measurable functions a, b on $(0, t_0)$,

$$\|v_t(x_t)\|_2 \le a_t + b_t \|x_t\|_2, \qquad \frac{\alpha_t}{\beta_t^2} a_t \in L^1(0,t_0), \qquad b \in L^1(0,t_0), \tag{23}$$

for almost every $t \in (0,t_0)$ and x_t with $\|x_t\|_2 \le \frac{\beta_t^2}{\alpha_t}\delta$. Additionally assume that, for some $\varepsilon > 0$,

$$\int_{\mathbb{R}^d} e^{\varepsilon\|x\|_2} \rho^v(x)\big(1 + \|\tau \nabla Q(x)\|_2\big)\,\mathrm{d}x < \infty. \tag{24}$$

B.2 Density Optimality Conditions

Proposition B.1. *Let ρ be a positive, normalized C^1 density satisfying the condition (21). For every $t \in (0,1)$ and $x_t \in \mathbb{R}^d$,*

$$\int_{\mathbb{R}^d} \nabla_{x_1} p^{\rho}_{1|t}(x_1|x_t)\,\mathrm{d}x_1 = 0.$$

Proof. Fix $t \in (0,1)$ and $x_t \in \mathbb{R}^d$. For every $R > 0$, denote $B_R := \{x \in \mathbb{R}^d : \|x\|_2 \le R\}$ by the closed ball of radius R. Choose a smooth cutoff function $\varphi_1 \in C^\infty(\mathbb{R}^d)$ such that

$$0 \le \varphi_1 \le 1, \qquad \varphi_1 \equiv 1 \text{ on } B_1, \qquad \operatorname{supp}\varphi_1 \subset B_2, \qquad C := \sup_{x\in\mathbb{R}^d} \|\nabla\varphi_1(x)\|_2 < \infty,$$

where supp denotes the support set of a function. For every $R > 0$, define $\varphi_R(x) := \varphi_1\left(\frac{x}{R}\right)$. Then

$$0 \le \varphi_R \le 1, \qquad \varphi_R \equiv 1 \text{ on } B_R, \qquad \operatorname{supp}\varphi_R \subset B_{2R}, \qquad \sup_{x\in\mathbb{R}^d} \|\nabla\varphi_R(x)\|_2 \le \frac{C}{R}.$$

Since φ_R has compact support, integrating by parts yields

$$\int_{\mathbb{R}^d} \varphi_R(x_1)\nabla_{x_1} p^{\rho}_{1|t}(x_1|x_t)\,\mathrm{d}x_1 = -\int_{\mathbb{R}^d} p^{\rho}_{1|t}(x_1|x_t)\nabla\varphi_R(x_1)\,\mathrm{d}x_1.$$

For every $R > 0$, the integrand on the left-hand side satisfies

$$\left\|\varphi_R(x_1)\nabla_{x_1} p^{\rho}_{1|t}(x_1|x_t)\right\|_2 \le \left\|\nabla_{x_1} p^{\rho}_{1|t}(x_1|x_t)\right\|_2 \quad \text{for every } x_1.$$

By Bayes' rule and differentiation of the Gaussian likelihood,

$$\nabla_{x_1} p^{\rho}_{1|t}(x_1|x_t) = p^{\rho}_{1|t}(x_1|x_t)\left[\nabla\log\rho(x_1) + \frac{\alpha_t}{\beta_t^2}(x_t - \alpha_t x_1)\right].$$

Consequently, (21) gives

$$\begin{aligned}
\int_{\mathbb{R}^d} \left\|\nabla_{x_1} p^{\rho}_{1|t}(x_1|x_t)\right\|_2 \mathrm{d}x_1 =& \mathbb{E}_{X_1\sim p^{\rho}_{1|t}(\cdot|x_t)}\left[\left\|\nabla\log\rho(X_1) + \frac{\alpha_t}{\beta_t^2}(x_t - \alpha_t X_1)\right\|_2\right] \\
\le& \mathbb{E}_{X_1\sim p^{\rho}_{1|t}(\cdot|x_t)}\left[\|\nabla\log\rho(X_1)\|_2 + \frac{\alpha_t^2}{\beta_t^2}\|X_1\|_2\right] + \frac{\alpha_t}{\beta_t^2}\|x_t\|_2 \\
<& \infty.
\end{aligned}$$

Since

$$\left\|\int_{\mathbb{R}^d} p^{\rho}_{1|t}(x_1|x_t)\nabla\varphi_R(x_1)\,\mathrm{d}x_1\right\|_2 \le \frac{C}{R}\left|\int_{\mathbb{R}^d} p^{\rho}_{1|t}(x_1|x_t)\,\mathrm{d}x_1\right| = \frac{C}{R} \to 0 \quad \text{as } R\to\infty,$$

Since $\varphi_R(x_1) \to 1$ for every x_1 and the preceding integrable bound is independent of R, Lebesgue's dominated convergence theorem (Rudin, 1987, Theorem 1.34), applied componentwise, gives

$$\begin{aligned}
\int_{\mathbb{R}^d} \nabla_{x_1} p^{\rho}_{1|t}(x_1|x_t)\,\mathrm{d}x_1 =& \lim_{R\to\infty}\int_{\mathbb{R}^d} \varphi_R(x_1)\nabla_{x_1} p^{\rho}_{1|t}(x_1|x_t)\,\mathrm{d}x_1 \\
=& -\lim_{R\to\infty}\int_{\mathbb{R}^d} p^{\rho}_{1|t}(x_1|x_t)\nabla\varphi_R(x_1)\,\mathrm{d}x_1 \\
=& 0.
\end{aligned}$$

This proves the proposition. □

Proposition 3.1. *Let ρ be a positive, normalized C^1 density satisfying the regularity condition (21) in Appendix B.1. For every $t \in (0,1)$ and $x_t \in \mathbb{R}^d$,*

$$v_t^\rho(x_t) = \frac{\dot{\alpha}_t}{\alpha_t} x_t + \frac{\kappa_t}{\alpha_t} \mathbb{E}_{X_1 \sim p^\rho_{1|t}(\cdot|x_t)}[\nabla \log \rho(X_1)]. \tag{8}$$

Proof. Fix $t \in (0,1)$ and $x_t \in \mathbb{R}^d$. By Bayes' rule,

$$\log p^\rho_{1|t}(x_1|x_t) = \log \rho(x_1) + \log p_{t|1}(x_t|x_1) - \log p^\rho_t(x_t).$$

The last term is independent of x_1. Differentiating the Gaussian conditional density therefore gives

$$\nabla_{x_1} \log p^\rho_{1|t}(x_1|x_t) = \nabla \log \rho(x_1) + \frac{\alpha_t}{\beta_t^2}(x_t - \alpha_t x_1). \tag{25}$$

Proposition B.1 implies

$$\mathbb{E}_{X_1 \sim p^\rho_{1|t}(\cdot|x_t)} \left[\nabla_{x_1} \log p^\rho_{1|t}(X_1|x_t) \right] = \int_{\mathbb{R}^d} \nabla_{x_1} p^\rho_{1|t}(x_1|x_t)\, \mathrm{d}x_1 = 0.$$

Consequently, taking the expectation of (25) under the same posterior $p^\rho_{1|t}(\cdot|x_t)$ gives

$$\mathbb{E}_{X_1 \sim p^\rho_{1|t}(\cdot|x_t)}[\nabla \log \rho(X_1)] = -\frac{\alpha_t}{\beta_t^2} x_t + \frac{\alpha_t^2}{\beta_t^2} \mathbb{E}_{X_1 \sim p^\rho_{1|t}(\cdot|x_t)}[X_1].$$

Then, the conditional-velocity formula and the definition of κ_t imply

$$\begin{aligned} v_t^\rho(x_t) &= \frac{\dot{\beta}_t}{\beta_t} x_t + \frac{\alpha_t \kappa_t}{\beta_t^2} \mathbb{E}_{X_1 \sim p^\rho_{1|t}(\cdot|x_t)}[X_1] \\ &= \left(\frac{\dot{\beta}_t}{\beta_t} + \frac{\kappa_t}{\beta_t^2} \right) x_t + \frac{\kappa_t}{\alpha_t} \mathbb{E}_{X_1 \sim p^\rho_{1|t}(\cdot|x_t)}[\nabla \log \rho(X_1)]. \end{aligned}$$

Substituting $\frac{\dot{\beta}_t}{\beta_t} + \frac{\kappa_t}{\beta_t^2} = \frac{\dot{\alpha}_t}{\alpha_t}$ yields (8). □

Proposition B.2. *Let $h \in L^1(\mathbb{R}^d; \mathbb{R}^d)$ satisfy*

$$\int_{\mathbb{R}^d} e^{\omega^\top x} h(x)\, \mathrm{d}x = 0, \qquad \int_{\mathbb{R}^d} e^{\omega^\top x} \|h(x)\|_2\, \mathrm{d}x < \infty$$

for every $\omega \in \mathbb{R}^d$. Then $h = 0$ almost everywhere.

Proof. By the injectivity of the Fourier transform on $L^1(\mathbb{R}^d)$ (Folland, 1999, Corollary 8.27), applied componentwise, it suffices to show that

$$\mathscr{F}[h](\omega) := \int_{\mathbb{R}^d} e^{-\mathrm{i}\omega^\top x} h(x)\, \mathrm{d}x$$

is zero for every $\omega \in \mathbb{R}^d$.

Fix $\omega \in \mathbb{R}^d$ and define $F_\omega : \mathbb{C} \to \mathbb{C}^d$ by

$$F_\omega(z) := \int_{\mathbb{R}^d} e^{z\omega^\top x} h(x)\, \mathrm{d}x.$$

This integral is well-defined for every $z \in \mathbb{C}$, since

$$\int_{\mathbb{R}^d} \left\| e^{z\omega^\top x} h(x) \right\|_2 \mathrm{d}x = \int_{\mathbb{R}^d} e^{(\mathrm{Re}\, z)\omega^\top x} \|h(x)\|_2\, \mathrm{d}x < \infty.$$

By assumption, $F_\omega(z) = 0$ for every $z \in \mathbb{R}$, while

$$\mathscr{F}[h](\omega) = F_\omega(-\mathrm{i}).$$

It therefore remains to verify that F_ω is entire componentwise: the identity theorem will then extend its vanishing on the real axis to all of $\mathbb{C}$, giving $F_\omega(-\mathrm{i}) = 0$.

To establish this, observe that for every $R > 0$,

$$\int_{\mathbb{R}^d} \sup_{|z|\le R} \left\| e^{z\omega^\top x} h(x) \right\|_2 \, \mathrm{d}x \le \int_{\mathbb{R}^d} e^{R|\omega^\top x|} \, \|h(x)\|_2 \, \mathrm{d}x$$
$$\le \int_{\mathbb{R}^d} \left(e^{R\omega^\top x} + e^{-R\omega^\top x} \right) \|h(x)\|_2 \, \mathrm{d}x < \infty.$$

The integrand is measurable in x and entire in z for each x. Hence the locally dominated form of the holomorphic parameter-integral theorem, a consequence of Rudin (1987, Chapter 10, Exercise 16), implies that F_ω is holomorphic componentwise on $|z| < R$. Since $R > 0$ is arbitrary, F_ω is entire componentwise.

Then, applying the identity theorem componentwise gives $F_\omega \equiv 0$ on $\mathbb{C}$. Thus, $\mathscr{F}[h](\omega) = F_\omega(-\mathrm{i}) = 0$. Since $\omega \in \mathbb{R}^d$ was arbitrary, Fourier injectivity yields $h = 0$ almost everywhere. □

Theorem 3.2. *Let π be the target density defined in (1). Let ρ be a positive, normalized C^1 density satisfying regularity conditions (21) and (22) in Appendix B.1. Define the density optimality condition*

$$v_t^\rho(x_t) = \frac{\dot\alpha_t}{\alpha_t} x_t + \frac{\kappa_t}{\alpha_t} \mathbb{E}_{X_1 \sim p^\rho_{1|t}(\cdot|x_t)}[\tau \nabla Q(X_1)] \quad \textit{for every } x_t \in \mathbb{R}^d. \tag{10}$$

Then

$$\rho = \pi \quad \Leftrightarrow \quad \text{(10) holds for every } t \in (0,1) \quad \Leftrightarrow \quad \text{(10) holds for some } t \in (0,1).$$

Proof. Suppose $\rho = \pi$, then $\nabla \log \rho = \tau \nabla Q$; hence, (22) and Proposition 3.1 give (10) for every $t \in (0,1)$. The implication from every interior time to some interior time is immediate.

Suppose (10) holds for some $t \in (0,1)$. Since $\frac{\kappa_t}{\alpha_t} > 0$, Proposition 3.1 gives

$$\mathbb{E}_{X_1 \sim p^\rho_{1|t}(\cdot|x_t)}[\nabla \log \rho(X_1) - \tau \nabla Q(X_1)] = 0.$$

Since $p_t^\rho(x_t) > 0$, expanding the posterior and multiplying by this marginal density yields

$$\int_{\mathbb{R}^d} p_{t|1}(x_t|x_1)\rho(x_1)(\nabla \log \rho(x_1) - \tau \nabla Q(x_1)) \, \mathrm{d}x_1 = 0 \quad \text{for every } x_t \in \mathbb{R}^d. \tag{26}$$

We now show that (26) forces the candidate and target scores to coincide. Define

$$h_t(x_1) := \rho(x_1)(\nabla \log \rho(x_1) - \tau \nabla Q(x_1)) \exp\left(-\frac{\alpha_t^2}{2\beta_t^2} \|x_1\|_2^2 \right).$$

Since

$$p_{t|1}(x_t|x_1) = \mathcal{N}(x_t; \alpha_t x_1, \beta_t^2 I) \propto \exp\left(-\frac{1}{2\beta_t^2} \|x_t - \alpha_t x_1\|_2^2 \right),$$

(26) becomes

$$\int_{\mathbb{R}^d} h_t(x_1) \exp\left(\frac{\alpha_t}{\beta_t^2} x_t^\top x_1 \right) \mathrm{d}x_1 = 0 \quad \text{for every } x_t \in \mathbb{R}^d.$$

Then, (21) and (22) imply that

$$\int_{\mathbb{R}^d} \|h_t(x_1)\|_2 \exp\left(\frac{\alpha_t}{\beta_t^2} x_t^\top x_1 \right) \mathrm{d}x_1 < \infty \quad \text{for every } x_t \in \mathbb{R}^d.$$

Taking $x_t = 0$ in the preceding integrability bound gives $h_t \in L^1(\mathbb{R}^d; \mathbb{R}^d)$. Moreover, $\frac{\alpha_t}{\beta_t^2} x_t$ ranges over $\mathbb{R}^d$ as x_t varies. Therefore, Proposition B.2 gives

$$h_t(x_1) = 0 \quad \text{for a.e. } x_1.$$

By the positivity of ρ and the continuity of $\nabla \log \rho$, ∇Q, this implies

$$\nabla \log \rho(x_1) = \tau \nabla Q(x_1) = \nabla \log \pi(x_1) \quad \text{for every } x_1 \in \mathbb{R}^d.$$

The normalization of ρ and π then gives $\rho = \pi$. □

Proposition 3.3. *Adopt the notation and assumptions of Theorem 3.2. Let $\lambda_t \neq 0$ and $\zeta_t \in \mathbb{R}$. Define the generalized density optimality condition*

$$v_t^\rho(x_t) = (1-\lambda_t-\zeta_t)v_t^\rho(x_t) + \zeta_t \mathbb{E}_{X_1\sim p^\rho_{1|t}(\cdot|x_t)}[v_{t|1}(x_t|X_1)] + \lambda_t\left(\frac{\dot\alpha_t}{\alpha_t}x_t + \frac{\kappa_t}{\alpha_t}\mathbb{E}_{X_1\sim p^\rho_{1|t}(\cdot|x_t)}[\tau\nabla Q(X_1)]\right) \quad \text{for every } x_t\in\mathbb{R}^d. \tag{13}$$

Then

$$\rho=\pi \quad\Leftrightarrow\quad (13)\text{ holds for every } t\in(0,1) \quad\Leftrightarrow\quad (13)\text{ holds for some } t\in(0,1).$$

Proof. By (6) and $\lambda_t\neq 0$, (13) is equivalent to (10). Therefore, this proposition is a straightforward conclusion of Theorem 3.2. □

B.3 Velocity Optimality Conditions

Proposition B.3. *Let π be the target density in (1), and let ρ be a positive, normalized C^1 density satisfying*

$$\int_{\mathbb{R}^d} e^{\omega^\top x}\rho(x)\left(\tau\nabla Q(x)+\omega\right)\,\mathrm{d}x = 0 \tag{27}$$

for every ω in a neighborhood of $0\in\mathbb{R}^d$, and

$$\int_{\mathbb{R}^d} e^{\varepsilon\|x\|_2}\rho(x)\left(\|\tau\nabla Q(x)\|_2+1\right)\,\mathrm{d}x<\infty \tag{28}$$

for some $\varepsilon>0$. Then $\rho=\pi$.

Proof. For every $\omega\in\mathbb{R}^d$, consider the Fourier transforms

$$\mathscr{F}[\rho\tau\nabla Q](\omega) := \int_{\mathbb{R}^d} e^{-\mathrm{i}\omega^\top x}\rho(x)\tau\nabla Q(x)\,\mathrm{d}x, \qquad \mathscr{F}[\rho](\omega) := \int_{\mathbb{R}^d} e^{-\mathrm{i}\omega^\top x}\rho(x)\,\mathrm{d}x.$$

Both integrals converge absolutely by (28). To prove $\rho=\pi$, it suffices to establish the score identity

$$\rho\tau\nabla Q = \nabla\rho. \tag{29}$$

The formal Fourier differentiation identity $\mathscr{F}[\nabla\rho](\omega) = \mathrm{i}\omega\mathscr{F}[\rho](\omega)$ suggests a Fourier-based argument. However, we have not assumed that $\nabla\rho$ is integrable, so its Fourier transform is not yet justified as an absolutely convergent integral. We therefore proceed in two steps: first, we establish

$$\mathscr{F}[\rho\tau\nabla Q](\omega) = \mathrm{i}\omega\mathscr{F}[\rho](\omega) \quad \text{for every } \omega\in\mathbb{R}^d; \tag{30}$$

then, we recover (29) from this Fourier identity using compactly supported smooth test functions.

Fix a nonzero $\omega\in\mathbb{R}^d$ and define F_ω on the strip $|\mathrm{Re}\,z|<\frac{\varepsilon}{\|\omega\|_2}$ by

$$F_\omega(z) := \int_{\mathbb{R}^d} e^{z\omega^\top x}\rho(x)\left(\tau\nabla Q(x)+z\omega\right)\,\mathrm{d}x.$$

This integral is well-defined for every z in the strip by (28). Applying (27) at $s\omega$ gives $F_\omega(s)=0$ for every real s sufficiently close to 0, while

$$F_\omega(-\mathrm{i}) = \mathscr{F}[\rho\tau\nabla Q](\omega) - \mathrm{i}\omega\mathscr{F}[\rho](\omega).$$

To obtain the desired Fourier identity (30), it therefore remains to verify that F_ω is holomorphic componentwise on the strip.

To establish this, fix a compact subset K of the strip and choose $R>0$ such that $|z|\le R$ for every $z\in K$. Then (28) implies

$$\int_{\mathbb{R}^d}\sup_{z\in K}\left\|e^{z\omega^\top x}\rho(x)\left(\tau\nabla Q(x)+z\omega\right)\right\|_2\mathrm{d}x \le \int_{\mathbb{R}^d} e^{\varepsilon\|x\|_2}\rho(x)\left(\|\tau\nabla Q(x)\|_2+R\,\|\omega\|_2\right)\mathrm{d}x<\infty.$$

The integrand is measurable in x and entire in z for each x. Therefore, the locally dominated form of the holomorphic parameter-integral theorem, a consequence of Rudin (1987, Chapter 10, Exercise 16), implies that F_ω is holomorphic componentwise on the strip.

The identity theorem (Rudin, 1987, Corollary of Theorem 10.18), applied componentwise, now gives $F_\omega \equiv 0$ throughout the connected strip. In particular, $F_\omega(-\mathrm{i}) = 0$ gives

$$\mathscr{F}[\rho\tau\nabla Q](\omega) = \mathrm{i}\omega\mathscr{F}[\rho](\omega).$$

This holds for every $\omega \neq 0$. For $\omega = 0$, the same identity follows from (27) at the origin. Thus (30) holds for every $\omega \in \mathbb{R}^d$.

It remains to recover the score identity (29). Fix any smooth function $\varphi : \mathbb{R}^d \to \mathbb{R}$ with compact support, and denote

$$\mathscr{F}[\varphi](\omega) := \int_{\mathbb{R}^d} e^{-\mathrm{i}\omega^\top x}\varphi(x)\,\mathrm{d}x.$$

Since φ is smooth with compact support, we have (Folland, 1999, Corollary 8.23)

$$\int_{\mathbb{R}^d} (1 + \|\omega\|_2)\,|\mathscr{F}[\varphi](\omega)|\,\mathrm{d}\omega < \infty.$$

Together with the integrability of ρ and $\rho\tau\nabla Q$, this justifies the applications of Fubini's theorem below. Fourier inversion (Folland, 1999, Theorem 8.26) and Fubini's theorem, followed by (30) at $-\omega$, give

$$\begin{aligned}\int_{\mathbb{R}^d} \varphi(x)\rho(x)\tau\nabla Q(x)\,\mathrm{d}x &= \frac{1}{(2\pi)^d}\int_{\mathbb{R}^d} \mathscr{F}[\varphi](\omega)\mathscr{F}[\rho\tau\nabla Q](-\omega)\,\mathrm{d}\omega \\ &= -\frac{1}{(2\pi)^d}\int_{\mathbb{R}^d} \mathrm{i}\omega\mathscr{F}[\varphi](\omega)\mathscr{F}[\rho](-\omega)\,\mathrm{d}\omega.\end{aligned}$$

The Fourier differentiation identity (Folland, 1999, Theorem 8.22) applies to the test function φ and gives $\mathscr{F}[\nabla\varphi](\omega) = \mathrm{i}\omega\mathscr{F}[\varphi](\omega)$. Hence,

$$\begin{aligned}-\frac{1}{(2\pi)^d}\int_{\mathbb{R}^d} \mathrm{i}\omega\mathscr{F}[\varphi](\omega)\mathscr{F}[\rho](-\omega)\,\mathrm{d}\omega &= -\frac{1}{(2\pi)^d}\int_{\mathbb{R}^d} \mathscr{F}[\nabla\varphi](\omega)\mathscr{F}[\rho](-\omega)\,\mathrm{d}\omega \\ &= -\int_{\mathbb{R}^d} \rho(x)\nabla\varphi(x)\,\mathrm{d}x \\ &= \int_{\mathbb{R}^d} \varphi(x)\nabla\rho(x)\,\mathrm{d}x.\end{aligned}$$

Here the second equality follows from Fourier inversion applied componentwise to $\nabla\varphi$ and Fubini's theorem. The last equality is integration by parts, justified by $\rho \in C^1$ and the compact support of φ; no global integrability of $\nabla\rho$ is required.

Therefore, for every smooth function $\varphi : \mathbb{R}^d \to \mathbb{R}$ with compact support,

$$\int_{\mathbb{R}^d} \varphi(x)\rho(x)\tau\nabla Q(x)\,\mathrm{d}x = \int_{\mathbb{R}^d} \varphi(x)\nabla\rho(x)\,\mathrm{d}x.$$

This implies (29) almost everywhere. The continuity of $\nabla\rho$ and $\rho\tau\nabla Q$ then gives (29) pointwise on $\mathbb{R}^d$. By positivity of ρ,

$$\nabla \log \rho = \tau\nabla Q = \nabla \log \pi.$$

The normalization of ρ and π therefore yields $\rho = \pi$. □

Proposition B.4. *Let ρ be a positive, normalized C^1 density on $\mathbb{R}^d$. Assume that (24) holds with ρ^v replaced by ρ for some $\varepsilon > 0$. Let $\varphi \in C(\mathbb{R}^d; \mathbb{R}^d)$ satisfy*

$$\int_{\mathbb{R}^d} e^{\xi^\top x}\rho(x)\,\|\varphi(x)\|_2\,\mathrm{d}x < \infty \qquad \textit{for every } \xi \in \mathbb{R}^d \textit{ with } \|\xi\|_2 < \varepsilon.$$

Then, for every fixed $\xi \in \mathbb{R}^d$ with $\|\xi\|_2 < \varepsilon$,

$$\lim_{t\downarrow 0} \mathbb{E}_{X_1 \sim p^\rho_{1|t}(\cdot|c_t\xi)}[\varphi(X_1)] = \frac{\int_{\mathbb{R}^d} e^{\xi^\top x}\rho(x)\varphi(x)\,\mathrm{d}x}{\int_{\mathbb{R}^d} e^{\xi^\top x}\rho(x)\,\mathrm{d}x}, \tag{31}$$

where $c_t := \frac{\beta_t^2}{\alpha_t}$. *The expectations and integrals in* (31) *are absolutely convergent, and the limiting denominator is strictly positive.*

In particular, (24) *guarantees the required integrability for both* $\varphi(x) = x$ *and* $\varphi(x) = \tau\nabla Q(x)$. *Consequently, for every fixed* $\xi \in \mathbb{R}^d$ *with* $\|\xi\|_2 < \varepsilon$,

$$\lim_{t\downarrow 0} \mathbb{E}_{X_1\sim p^{\rho}_{1|t}(\cdot|c_t\xi)}[X_1] = \frac{\int_{\mathbb{R}^d} e^{\xi^\top x}\rho(x)x\,\mathrm{d}x}{\int_{\mathbb{R}^d} e^{\xi^\top x}\rho(x)\,\mathrm{d}x}, \tag{32a}$$

$$\lim_{t\downarrow 0} \mathbb{E}_{X_1\sim p^{\rho}_{1|t}(\cdot|c_t\xi)}[\tau\nabla Q(X_1)] = \frac{\int_{\mathbb{R}^d} e^{\xi^\top x}\rho(x)\tau\nabla Q(x)\,\mathrm{d}x}{\int_{\mathbb{R}^d} e^{\xi^\top x}\rho(x)\,\mathrm{d}x}. \tag{32b}$$

Both limits are finite. In particular, for each such ξ, *both posterior means are bounded for all sufficiently small* t.

Proof. Fix $\xi \in \mathbb{R}^d$ with $\|\xi\|_2 < \varepsilon$. Expanding the Gaussian likelihood at $x_t = c_t\xi$ gives

$$-\frac{1}{2\beta_t^2}\|c_t\xi - \alpha_t x_1\|_2^2 = -\frac{\beta_t^2}{2\alpha_t^2}\|\xi\|_2^2 + \xi^\top x_1 - \frac{\alpha_t^2}{2\beta_t^2}\|x_1\|_2^2.$$

The first term is independent of x_1 and cancels in posterior normalization, yielding

$$p^{\rho}_{1|t}(x_1|c_t\xi) = \frac{\exp\left(\xi^\top x_1 - \frac{\alpha_t^2}{2\beta_t^2}\|x_1\|_2^2\right)\rho(x_1)}{\int_{\mathbb{R}^d}\exp\left(\xi^\top x - \frac{\alpha_t^2}{2\beta_t^2}\|x\|_2^2\right)\rho(x)\,\mathrm{d}x}. \tag{33}$$

We first establish

$$\lim_{t\downarrow 0}\int_{\mathbb{R}^d}\exp\left(\xi^\top x - \frac{\alpha_t^2}{2\beta_t^2}\|x\|_2^2\right)\rho(x)\,\mathrm{d}x = \int_{\mathbb{R}^d} e^{\xi^\top x}\rho(x)\,\mathrm{d}x, \tag{34a}$$

$$\lim_{t\downarrow 0}\int_{\mathbb{R}^d}\exp\left(\xi^\top x - \frac{\alpha_t^2}{2\beta_t^2}\|x\|_2^2\right)\rho(x)\varphi(x)\,\mathrm{d}x = \int_{\mathbb{R}^d} e^{\xi^\top x}\rho(x)\varphi(x)\,\mathrm{d}x. \tag{34b}$$

Both integrands are measurable since ρ and φ are continuous. Since α, β are continuous with $\alpha_0 = 0$ and $\beta_0 = 1$, we have $\frac{\alpha_t}{\beta_t} \to 0$ as $t \downarrow 0$. Thus, for every $x \in \mathbb{R}^d$,

$$\lim_{t\downarrow 0}\exp\left(\xi^\top x - \frac{\alpha_t^2}{2\beta_t^2}\|x\|_2^2\right)\rho(x) = e^{\xi^\top x}\rho(x),$$
$$\lim_{t\downarrow 0}\exp\left(\xi^\top x - \frac{\alpha_t^2}{2\beta_t^2}\|x\|_2^2\right)\rho(x)\varphi(x) = e^{\xi^\top x}\rho(x)\varphi(x).$$

For every $t \in (0,1)$ and $x \in \mathbb{R}^d$,

$$0 < \exp\left(\xi^\top x - \frac{\alpha_t^2}{2\beta_t^2}\|x\|_2^2\right)\rho(x) \le e^{\xi^\top x}\rho(x) \le e^{\varepsilon\|x\|_2}\rho(x),$$
$$\left\|\rho(x)\exp\left(\xi^\top x - \frac{\alpha_t^2}{2\beta_t^2}\|x\|_2^2\right)\varphi(x)\right\|_2 \le e^{\xi^\top x}\rho(x)\|\varphi(x)\|_2.$$

The right-hand sides are independent of t. The scalar bound is integrable by (24), and the vector bound is integrable by the assumption on φ. Therefore, Lebesgue's dominated convergence theorem (Rudin, 1987, Theorem 1.34), applied componentwise to the vector-valued integral, implies (34).

The same bounds ensure absolute convergence of both integrals for every $t \in (0,1)$ and of their limits. Since $\rho > 0$, the denominator in (33) and its limit are strictly positive. Taking the ratio therefore gives (31).

It remains to verify the integrability assumption for the two stated choices of φ. For $\varphi(x) = x$, the strict inequality $\|\xi\|_2 < \varepsilon$ gives

$$\int_{\mathbb{R}^d} e^{\xi^\top x}\rho(x)\|x\|_2\,\mathrm{d}x \le \int_{\mathbb{R}^d} e^{\|\xi\|_2\|x\|_2}\rho(x)\|x\|_2\,\mathrm{d}x \le \frac{1}{\varepsilon - \|\xi\|_2}\int_{\mathbb{R}^d} e^{\varepsilon\|x\|_2}\rho(x)\,\mathrm{d}x < \infty,$$

where we used $re^{-(\varepsilon-\|\xi\|_2)r} \le (\varepsilon - \|\xi\|_2)^{-1}$ for $r \ge 0$. For $\varphi(x) = \tau\nabla Q(x)$,

$$\int_{\mathbb{R}^d} e^{\xi^\top x}\rho(x)\,\|\tau\nabla Q(x)\|_2\,\mathrm{d}x \le \int_{\mathbb{R}^d} e^{\varepsilon\|x\|_2}\rho(x)\,\|\tau\nabla Q(x)\|_2\,\mathrm{d}x < \infty.$$

Both bounds follow from (24). Applying (31) to these two continuous functions proves (32). Since their limits are finite, both posterior means are bounded for all sufficiently small t. □

Theorem 3.4. *Let π be the target density defined in (1), and assume that the regularity condition (21) holds for $\rho = \pi$. Assume that v^π and v are admissible, as specified in Appendix B.1. Assume that v satisfies (23) and (24). Then, $v = \mathcal{I}v$ holds on $(0,1)\times\mathbb{R}^d$ if and only if $\rho^v = \pi$ and $v = v^\pi$.*

Proof. Suppose $\rho^v = \pi$ and $v = v^\pi$. By (9) and (14),

$$v_t^\pi(x_t) = \frac{\dot\alpha_t}{\alpha_t}x_t + \frac{\kappa_t}{\alpha_t}\mathbb{E}_{X_1\sim p^\pi_{1|t}(\cdot|x_t)}[\tau\nabla Q(X_1)] = (\mathcal{I}v^\pi)_t(x_t)$$

for every $t\in(0,1)$ and $x_t\in\mathbb{R}^d$. Therefore, $v^\pi = \mathcal{I}v^\pi$. It remains to prove the converse direction.

Assume $v = \mathcal{I}v$. For notational simplicity, denote $\rho := \rho^v$ and $c_t := \frac{\beta_t^2}{\alpha_t}$. Fix $\xi\in\mathbb{R}^d$ with $\|\xi\|_2 < \min\{\delta, \frac{\varepsilon}{2}\}$, where δ and ε are from (23)–(24) in Appendix B.1. By Proposition B.4, the limit

$$\lim_{t\downarrow 0}\left(\xi + \mathbb{E}_{X_1\sim p^\rho_{1|t}(\cdot|c_t\xi)}[\tau\nabla Q(X_1)]\right) = \frac{\int_{\mathbb{R}^d} e^{\xi^\top x}\rho(x)\,(\xi+\tau\nabla Q(x))\,\mathrm{d}x}{\int_{\mathbb{R}^d} e^{\xi^\top x}\rho(x)\,\mathrm{d}x} \in \mathbb{R}^d \tag{35}$$

exists. We first prove that this limit is zero, then apply Proposition B.3 to identify ρ, and finally use the fixed-point equation to identify v.

The assumption $v = \mathcal{I}v$ gives

$$v_t(c_t\xi) = \frac{\dot\alpha_t\beta_t^2}{\alpha_t^2}\xi + \frac{\kappa_t}{\alpha_t}\mathbb{E}_{X_1\sim p^\rho_{1|t}(\cdot|c_t\xi)}[\tau\nabla Q(X_1)].$$

Using (7), we obtain

$$\begin{aligned}\frac{\alpha_t}{\beta_t^2}v_t(c_t\xi) - \frac{\dot\beta_t}{\beta_t}\xi &= \left(\frac{\dot\alpha_t}{\alpha_t} - \frac{\dot\beta_t}{\beta_t}\right)\xi + \frac{\kappa_t}{\beta_t^2}\mathbb{E}_{X_1\sim p^\rho_{1|t}(\cdot|c_t\xi)}[\tau\nabla Q(X_1)]\\ &= \frac{\kappa_t}{\beta_t^2}\left(\xi + \mathbb{E}_{X_1\sim p^\rho_{1|t}(\cdot|c_t\xi)}[\tau\nabla Q(X_1)]\right).\end{aligned}$$

Since $t_0 < 1$, (23) implies

$$\begin{aligned}&\int_0^{t_0}\frac{\kappa_t}{\beta_t^2}\left\|\xi + \mathbb{E}_{X_1\sim p^\rho_{1|t}(\cdot|c_t\xi)}[\tau\nabla Q(X_1)]\right\|_2\mathrm{d}t\\ &= \int_0^{t_0}\left\|\frac{\alpha_t}{\beta_t^2}v_t(c_t\xi) - \frac{\dot\beta_t}{\beta_t}\xi\right\|_2\mathrm{d}t\\ &\le \int_0^{t_0}\left(\frac{\alpha_t}{\beta_t^2}\|v_t(c_t\xi)\|_2 + \left|\frac{\dot\beta_t}{\beta_t}\right|\|\xi\|_2\right)\mathrm{d}t\\ &\le \int_0^{t_0}\left(\frac{\alpha_t}{\beta_t^2}(a_t + b_tc_t\|\xi\|_2) + \left|\frac{\dot\beta_t}{\beta_t}\right|\|\xi\|_2\right)\mathrm{d}t\\ &= \int_0^{t_0}\frac{\alpha_t}{\beta_t^2}a_t\,\mathrm{d}t + \|\xi\|_2\int_0^{t_0}\left(b_t + \left|\frac{\dot\beta_t}{\beta_t}\right|\right)\mathrm{d}t < \infty.\end{aligned} \tag{36}$$

On the other hand, (7) implies

$$\int_0^{t_0}\frac{\kappa_t}{\beta_t^2}\,\mathrm{d}t = \log\frac{\alpha_t}{\beta_t}\Big|_{t=0}^{t_0} = \infty.$$

Suppose, for contradiction, that the limit in (35) is nonzero. Then there exist $m > 0$ and $t_1 \in (0, t_0)$ such that

$$\left\| \xi + \mathbb{E}_{X_1 \sim p^{\rho}_{1|t}(\cdot|c_t\xi)}[\tau \nabla Q(X_1)] \right\|_2 \geq m \quad \text{for every } t \in (0, t_1).$$

Consequently, the integral in (36) is at least $m \int_0^{t_1} \frac{\kappa_t}{\beta_t^2} \, \mathrm{d}t = \infty$, which leads to a contradiction. Therefore, the limit (35) is zero. Multiplying by the strictly positive limiting denominator in (35), and using the arbitrariness of ξ, gives

$$\int_{\mathbb{R}^d} e^{\xi^\top x} \rho(x) \left(\xi + \tau \nabla Q(x) \right) \mathrm{d}x = 0 \quad \text{for every } \|\xi\|_2 < \min\left\{ \delta, \frac{\varepsilon}{2} \right\}. \tag{37}$$

Applying Proposition B.3 with (24) and (37) gives $\rho = \pi$. Substituting this equality into $v = \mathcal{I} v$ and using (9) yields $v = v^\pi$. □

Theorem 3.5. *Assume the hypotheses of Theorem 3.4 on v and π, and let t_0 be as in (23). Let $\lambda_t > 0$ and $\zeta_t \geq 0$ be measurable, finite schedules on $(0, 1)$ satisfying*

$$\int_0^{t_0} \frac{\lambda_t}{\lambda_t + \zeta_t} \frac{\kappa_t}{\beta_t^2} \, \mathrm{d}t = \infty. \tag{16}$$

Then $v = \mathcal{I}_{\lambda,\zeta} v$ holds on $(0, 1) \times \mathbb{R}^d$ if and only if $\rho^v = \pi$ and $v = v^\pi$.

Proof. For notational simplicity, denote $\rho := \rho^v$ and $c_t := \frac{\beta_t^2}{\alpha_t}$. Since $\lambda_t + \zeta_t > 0$, denote $\bar{\lambda}_t := \frac{\lambda_t}{\lambda_t + \zeta_t} \in (0, 1]$. Then, $v = \mathcal{I}_{\lambda,\zeta} v$ is equivalent to

$$v_t(x_t) = \left(1 - \bar{\lambda}_t\right) \mathbb{E}_{X_1 \sim p^{\rho}_{1|t}(\cdot|x_t)}[v_{t|1}(x_t|X_1)] + \bar{\lambda}_t \left(\frac{\dot{\alpha}_t}{\alpha_t} x_t + \frac{\kappa_t}{\alpha_t} \mathbb{E}_{X_1 \sim p^{\rho}_{1|t}(\cdot|x_t)}[\tau \nabla Q(X_1)] \right). \tag{38}$$

Suppose $v = v^\pi$ and $\rho = \pi$. Then, $v_t^\pi(x_t) = \mathbb{E}_{X_1 \sim p^{\pi}_{1|t}(\cdot|x_t)}[v_{t|1}(x_t|X_1)]$ and $v^\pi = \mathcal{I} v^\pi$ imply (38), which gives $v^\pi = \mathcal{I}_{\lambda,\zeta} v^\pi$.

Suppose $v = \mathcal{I}_{\lambda,\zeta} v$, which gives (38). In the following, we prove $v = v^\pi$ and $\rho = \pi$ following the proof of Theorem 3.4. Fix $\xi \in \mathbb{R}^d$ with $\|\xi\|_2 < \min\left\{\delta, \frac{\varepsilon}{2}\right\}$, where δ and ε are from (23)–(24) in Appendix B.1. By Proposition B.4, the limit

$$\lim_{t \downarrow 0} \left(\xi + \mathbb{E}_{X_1 \sim p^{\rho}_{1|t}(\cdot|c_t\xi)}[\tau \nabla Q(X_1)] \right) = \frac{\int_{\mathbb{R}^d} e^{\xi^\top x} \rho(x) \left(\xi + \tau \nabla Q(x) \right) \mathrm{d}x}{\int_{\mathbb{R}^d} e^{\xi^\top x} \rho(x) \, \mathrm{d}x} \in \mathbb{R}^d \tag{39}$$

exists. We first prove that this limit is zero, then apply Proposition B.3 to identify ρ, and finally use the fixed-point equation to identify v.

The assumption $v = \mathcal{I}_{\lambda,\zeta} v$ gives

$$\frac{\alpha_t}{\beta_t^2} v_t\left(c_t \xi\right) - \frac{\dot{\beta}_t}{\beta_t} \xi = \left(1 - \bar{\lambda}_t\right) \frac{\alpha_t^2 \kappa_t}{\beta_t^4} \mathbb{E}_{X_1 \sim p^{\rho}_{1|t}(\cdot|c_t\xi)}[X_1] + \bar{\lambda}_t \frac{\kappa_t}{\beta_t^2} \left(\xi + \mathbb{E}_{X_1 \sim p^{\rho}_{1|t}(\cdot|c_t\xi)}[\tau \nabla Q(X_1)] \right).$$

Following (36), the condition (23) implies

$$\int_0^{t_0} \left\| \frac{\alpha_t}{\beta_t^2} v_t(c_t \xi) - \frac{\dot{\beta}_t}{\beta_t} \xi \right\|_2 \mathrm{d}t < \infty.$$

Note that

$$\frac{\alpha_t^2 \kappa_t}{\beta_t^4} = \frac{1}{2} \frac{\mathrm{d}}{\mathrm{d}t} \left(\frac{\alpha_t^2}{\beta_t^2} \right), \qquad \int_0^{t_0} \frac{\alpha_t^2 \kappa_t}{\beta_t^4} \, \mathrm{d}t = \frac{\alpha_{t_0}^2}{2 \beta_{t_0}^2} < \infty.$$

By Proposition B.4, the existence of the limit (32a) implies

$$\int_0^{t_0} \left\| \left(1 - \bar{\lambda}_t\right) \frac{\alpha_t^2 \kappa_t}{\beta_t^4} \mathbb{E}_{X_1 \sim p^{\rho}_{1|t}(\cdot|c_t\xi)}[X_1] \right\|_2 \mathrm{d}t < \infty.$$

Therefore,

$$\int_0^{t_0} \left\| \bar{\lambda}_t \frac{\kappa_t}{\beta_t^2} \left(\xi + \mathbb{E}_{X_1 \sim p^{\rho}_{1|t}(\cdot|c_t\xi)} [\tau \nabla Q(X_1)] \right) \right\|_2 \mathrm{d}t < \infty. \tag{40}$$

Suppose, for contradiction, that the limit in (39) is nonzero. Then there exist $m > 0$ and $t_1 \in (0, t_0)$ such that

$$\left\| \xi + \mathbb{E}_{X_1 \sim p^{\rho}_{1|t}(\cdot|c_t\xi)} [\tau \nabla Q(X_1)] \right\|_2 \geq m \quad \text{for every } t \in (0, t_1).$$

Since $0 < \bar{\lambda}_t \leq 1$ and $\frac{\kappa_t}{\beta_t^2}$ is continuous on $[t_1, t_0]$, condition (16) implies

$$\int_0^{t_1} \bar{\lambda}_t \frac{\kappa_t}{\beta_t^2} \, \mathrm{d}t = \infty.$$

Consequently, the integral in (40) is at least $m \int_0^{t_1} \bar{\lambda}_t \frac{\kappa_t}{\beta_t^2} \, \mathrm{d}t = \infty$, which leads to a contradiction. Therefore, the limit (39) is zero. Multiplying by the strictly positive limiting denominator in (39), and using the arbitrariness of ξ, gives

$$\int_{\mathbb{R}^d} e^{\xi^\top x} \rho(x) \left(\xi + \tau \nabla Q(x) \right) \mathrm{d}x = 0 \quad \text{for every } \|\xi\|_2 < \min\left\{ \delta, \frac{\varepsilon}{2} \right\}.$$

Applying Proposition B.3 gives $\rho = \pi$. By (6) and (9), both posterior expressions in (38) then equal $v_t^\pi(x_t)$. Hence,

$$v_t(x_t) = (1 - \bar{\lambda}_t) v_t^\pi(x_t) + \bar{\lambda}_t v_t^\pi(x_t) = v_t^\pi(x_t)$$

for every $t \in (0, 1)$ and $x_t \in \mathbb{R}^d$. □

Proposition B.5. *Let π be the target density defined in (1), and assume that the regularity condition (21) holds for $\rho = \pi$. Assume that v^π and v are admissible, as specified in Appendix B.1. Suppose that there exist nonnegative measurable functions $a, b \in L^1(0, 1)$ such that*

$$\|v_t(x_t)\|_2 \leq a_t + b_t \|x_t\|_2 \quad \textit{for almost every } t \in (0, 1) \textit{ and every } x_t \in \mathbb{R}^d. \tag{41}$$

Assume additionally that there exist constants $C_Q \in [0, \infty)$ and $m \geq 0$ such that

$$\|\tau \nabla Q(x)\|_2 \leq C_Q (1 + \|x\|_2^m) \qquad \textit{for every } x \in \mathbb{R}^d. \tag{42}$$

Then (23) holds for every $t_0 \in (0, 1)$, and (24) holds for every $\varepsilon > 0$. Consequently, $v = \mathcal{I}v$ holds on $(0, 1) \times \mathbb{R}^d$ if and only if $\rho^v = \pi$ and $v = v^\pi$.

Proof. Based on Theorem 3.4, it remains to verify conditions (23)–(24).

For every $t_0 \in (0, 1)$, $\lim_{t \downarrow 0} \frac{\alpha_t}{\beta_t^2} = 0$ and $\frac{\alpha_t}{\beta_t^2}$ is bounded on $[0, t_0]$; hence, $a \in L^1(0, 1)$ implies $\frac{\alpha_t}{\beta_t^2} a_t \in L^1(0, t_0)$. Therefore, (41) implies (23).

Along an ODE trajectory Y_t of v, the following inequality holds for almost every $t \in (0, 1)$:

$$\frac{\mathrm{d}}{\mathrm{d}t} \|Y_t\|_2 \leq \|v_t(Y_t)\|_2 \leq a_t + b_t \|Y_t\|_2 .$$

Denote $\|a\|_{L^1} = \int_0^1 |a_t| \, \mathrm{d}t$ and $\|b\|_{L^1} = \int_0^1 |b_t| \, \mathrm{d}t$. Grönwall's inequality therefore yields

$$\|Y_1\|_2 \leq \exp\left(\int_0^1 b_t \, \mathrm{d}t \right) \left(\|Y_0\|_2 + \int_0^1 a_t \, \mathrm{d}t \right) \leq e^{\|b\|_{L^1}} \left(\|Y_0\|_2 + \|a\|_{L^1} \right).$$

Therefore, for every $\varepsilon > 0$, (42) implies

$$\begin{aligned}
&\mathbb{E}_{Y_1 \sim \rho^v} \left[e^{\varepsilon \|Y_1\|_2} \left(1 + \|\tau \nabla Q(Y_1)\|_2 \right) \right] \\
\leq &\mathbb{E}_{Y_1 \sim \rho^v} \left[e^{\varepsilon \|Y_1\|_2} \left(1 + C_Q (1 + \|Y_1\|_2^m) \right) \right] \\
\leq &\mathbb{E}_{Y_0 \sim p_0} \left[\exp\left(\varepsilon e^{\|b\|_{L^1}} \left(\|Y_0\|_2 + \|a\|_{L^1} \right) \right) \left(1 + C_Q + C_Q e^{m \|b\|_{L^1}} \left(\|Y_0\|_2 + \|a\|_{L^1} \right)^m \right) \right] \\
< &\infty,
\end{aligned}$$

which gives (24).

According to Theorem 3.4, $v = \mathcal{I}v$ holds on $(0, 1) \times \mathbb{R}^d$ if and only if $\rho^v = \pi$ and $v = v^\pi$. □

B.4 Coefficient Selection

The generalized velocity optimality condition and training objective leave the coefficient schedules unspecified. We select them to control numerical singularities in the regression target while preserving the uniqueness guarantee.

Singularity cancellation. The score-based contribution in (18) contains the factors $\lambda_t\dot{\alpha}_t/\alpha_t$ and $\lambda_t\kappa_t/\alpha_t$. Using the definition of κ_t, the latter becomes

$$\lambda_t\frac{\kappa_t}{\alpha_t} = \frac{\lambda_t}{\alpha_t^2}\beta_t(\dot{\alpha}_t\beta_t - \alpha_t\dot{\beta}_t).$$

The inverse powers of α_t can cause numerical difficulties near $t = 0$. Choosing $\lambda_t = \eta\alpha_t^2$, with $\eta > 0$, cancels these factors in both coefficients.

Proposition B.6. *Fix $\eta > 0$. For $\lambda_t = \eta\alpha_t^2$,*

$$\lambda_t\frac{\dot{\alpha}_t}{\alpha_t} = \eta\alpha_t\dot{\alpha}_t, \qquad \lambda_t\frac{\kappa_t}{\alpha_t} = \eta\beta_t(\dot{\alpha}_t\beta_t - \alpha_t\dot{\beta}_t). \tag{43}$$

The two coefficients in (43) are continuous and bounded on $[0, 1]$.

Proof. Substituting $\lambda_t = \eta\alpha_t^2$ and the definition (7) gives (43). Since $\alpha, \beta \in C^1([0, 1])$, the expressions on the right-hand sides extend continuously to $[0, 1]$ and are bounded. □

Preserving fixed-point uniqueness. Theorem 3.5 requires the integral condition (16). The following proposition gives a sufficient condition.

Proposition B.7. *Adopt the assumptions of Theorem 3.4 on v and π. Let $\lambda_t > 0$ and $\zeta_t \geq 0$ be measurable, finite schedules on $(0, 1)$. If there exists a constant $C \in (0, \infty)$ such that $\zeta_t \leq C\lambda_t$ for all sufficiently small $t > 0$, then (16) holds. Consequently, $v = \mathcal{I}_{\lambda,\zeta}v$ holds on $(0, 1) \times \mathbb{R}^d$ if and only if $\rho^v = \pi$ and $v = v^\pi$.*

Proof. For notational simplicity, denote $\bar{\lambda}_t := \frac{\lambda_t}{\lambda_t+\zeta_t}$. Suppose $\zeta_t \leq C\lambda_t$ for some finite $C \in (0, \infty)$ and every $t \in (0, t_1)$, where $t_1 \in (0, t_0)$. Then, $\bar{\lambda}_t \geq \frac{1}{1+C} > 0$ for $t \in (0, t_1)$. Therefore,

$$\int_0^{t_0} \bar{\lambda}_t\frac{\kappa_t}{\beta_t^2}\,\mathrm{d}t \geq \int_0^{t_1} \bar{\lambda}_t\frac{\kappa_t}{\beta_t^2}\,\mathrm{d}t \geq \frac{1}{1+C}\int_0^{t_1}\frac{\kappa_t}{\beta_t^2}\,\mathrm{d}t = \infty,$$

which gives (16). Hence, under the assumptions of Theorem 3.5, $v = \mathcal{I}_{\lambda,\zeta}v$ holds on $(0, 1) \times \mathbb{R}^d$ if and only if $\rho^v = \pi$ and $v = v^\pi$. □

We impose $\zeta_t \leq C\lambda_t$, with the following choice:

$$\lambda_t = \eta\alpha_t^2, \qquad \zeta_t = \eta\min\{C\alpha_t^2, \beta_t^2\}, \qquad \eta > 0. \tag{44}$$

Proposition B.7 ensures the uniqueness guarantee of the velocity optimality condition.

B.5 Bounded Domains

Proposition 3.6. *Let $Q \in C^1((-1, 1)^d)$ and let π be a probability density on $(-1, 1)^d$ satisfying $\pi(a) \propto e^{\tau Q(a)}$. The latent target has score*

$$\nabla_x \log\widetilde{\pi}(x) = \tau\nabla_x Q(\tanh x) - 2\tanh x. \tag{20}$$

Moreover, $Y_1 \sim \widetilde{\pi}$ if and only if $\tanh Y_1 \sim \pi$.

Proof. The coordinatewise map $x \mapsto \tanh x$ is a continuously differentiable bijection from $\mathbb{R}^d$ onto $(-1, 1)^d$, with inverse $\operatorname{arctanh}$. Its Jacobian is diagonal with positive entries $1 - \tanh^2 x_i$.

The change-of-variables formula therefore gives (19), including the normalizing constant. Taking logarithms,

$$\log \widetilde{\pi}(x) = \tau Q(\tanh x) + \sum_{i=1}^{d} \log(1 - \tanh^2 x_i) - \log \int_{(-1,1)^d} \exp(\tau Q(a))\, \mathrm{d}a.$$

The final term is independent of x. The gradient of the first term is

$$\nabla_x[\tau Q(\tanh x)] = \tau \operatorname{diag}(1 - \tanh^2 x)\ \nabla_a Q(a)|_{a=\tanh x}\,.$$

For each coordinate, differentiating the log-Jacobian term gives

$$\frac{\mathrm{d}}{\mathrm{d}x_i} \log(1 - \tanh^2 x_i) = \frac{-2 \tanh x_i (1 - \tanh^2 x_i)}{1 - \tanh^2 x_i} = -2 \tanh x_i.$$

Adding these gradients proves (20). In particular, the Jacobian correction is not multiplied by the inverse temperature.

Finally, if Y_1 has density $\widetilde{\pi}$, applying the change-of-variables formula to $\tanh Y_1$ cancels the Jacobian factor in (19), yielding density π. Conversely, if $\tanh Y_1$ has density π, applying $\operatorname{arctanh}$ uniquely determines its preimage density to be $\widetilde{\pi}$. This proves both directions of the distributional equivalence. The same argument applies at each fixed state when $Q(a)$ is replaced by $Q(s, a)$. □

C Experiment Details and Results

C.1 Toy Example

Target density. The target is $\pi(x) \propto \exp(\tau Q(x))$ on $\mathbb{R}^2$ with $\tau = 1$ and

$$Q(x) = \log\left[\frac{1}{8}\sum_{k=0}^{7} \exp\left(-\frac{1}{2}\,\|x - \mu_k\|_2^2\right)\right], \qquad \mu_k = \begin{pmatrix} 8\cos(2\pi k/8) \\ 8\sin(2\pi k/8) \end{pmatrix},$$

an equally weighted mixture of eight unit-variance Gaussians whose means lie on a circle.

Settings. All five methods are integrated into one PyTorch codebase and share the network, optimizer, batch size, training budget, and evaluation protocol to ensure a fair comparison. During training, every algorithm has access only to the target score $\tau\nabla Q(x)$; ground-truth samples are used for evaluation only. All methods are given the same budget of network updates and target-score queries. We use a time-conditioned multilayer perceptron with 4 hidden layers of 128 units, Adam with learning rate 5×10^{-4}, and batch size 512. AS, ASBS, and BMS are implemented following their open-source implementations, and FS following its paper. SCF uses $\alpha_t = t$, $\beta_t = 1 - t$, and the coefficients (44) with $\eta = 1$ and $C = 10^4$. For ASBS, the update budget is split equally between its corrector matching and adjoint matching. For sampling, SCF uses 50 NFEs, while the baselines use 100 NFEs. All runs are conducted on a desktop with an NVIDIA GeForce RTX 5090 GPU and an Intel Core Ultra 9 285K CPU.

Metrics. We evaluate 20 checkpoints per run, evenly spaced in model updates, each with 5×10^4 generated samples. The checkpoint with the smallest W_2 is selected and re-evaluated on fresh generated and reference samples, and we report the mean and standard deviation of the resulting values over five seeds.

- **Mode total variation distance (mode TVD).** Each generated sample is assigned to the component with the nearest mean, and $\widehat{w}_k$ denotes the fraction of samples assigned to component k. The mode TVD $\frac{1}{2}\sum_{k=0}^{7} \left|\widehat{w}_k - \frac{1}{8}\right|$ measures how evenly the eight modes are covered.
- **2-Wasserstein distance (W_2).** Computed exactly under the squared Euclidean cost between 2000 generated samples and 2000 target samples; we repeat the computation 20 times with nonoverlapping subsets of generated samples and fresh target samples, and report the median. As a finite-sample estimate it does not vanish even for perfect samples: between two independent sets of target samples it equals 0.719, which serves as the reference floor.

Algorithm 1: Online Reinforcement Learning with Score-Calibrated Flow

Input: network parameters $\theta, \omega_1, \omega_2$; target network parameters $\bar{\omega}_1 \leftarrow \omega_1, \bar{\omega}_2 \leftarrow \omega_2$; inverse temperature τ; learning rates $\ell_{\text{actor}}, \ell_{\text{critic}}$; target smoothing coefficient ν; replay buffer $\mathcal{D} \leftarrow \varnothing$.

for *each iteration* **do**
  for *each environment step* **do**
    Sample action $a \sim \pi^\theta(\cdot|s)$ and execute it in the environment;
    Observe next state s' and reward r;
    Store transition $\mathcal{D} \leftarrow \mathcal{D} \cup \{(s, a, r, s')\}$.
  end
  for *each gradient step* **do**
    Sample a mini-batch from $\mathcal{D}$;
    Update critics $\omega_i \leftarrow \omega_i - \ell_{\text{critic}} \nabla_{\omega_i} \mathcal{L}_Q(\omega_i)$ for $i \in \{1, 2\}$;
    Update actor $\theta \leftarrow \theta - \ell_{\text{actor}} \nabla_\theta \mathcal{L}^{\text{actor}}_{\lambda,\zeta}(\theta)$;
    Update target networks $\bar{\omega}_i \leftarrow \nu\omega_i + (1 - \nu)\bar{\omega}_i$ for $i \in \{1, 2\}$.
  end
end

Table 1: Shared settings of the RL experiments.

	DeepMind Control Suite	HumanoidBench
Parallel environments	1	50
Environment steps per training step	1	50
Replay buffer size	2.5×10^5	1×10^6
Batch size	256	256
Hidden layers	2×256	3×256
Normalization	none	batch renormalization
Optimizer	Adam	Adam
Actor learning rate	3×10^{-4}	3×10^{-4}
Critic learning rate	1×10^{-3}	3×10^{-4}
Discount factor	0.99	0.99
Target smoothing coefficient	0.005	0.005

Results. The results are summarized in Figure 1. Black crosses mark the component means, the circles mark one and two standard deviations around each mean, and samples are colored by their nearest mean. Each panel reports the mode TVD, the W_2 distance, and the wall-clock training time as mean $\pm$ standard deviation over five seeds (lower is better). SCF achieves the lowest discrepancy on both mode TVD and W_2 while being the most efficient in terms of training and inference.

C.2 RL Tasks

Shared settings. For each set of RL tasks, all algorithms are integrated into one JAX codebase and share the training loop, replay buffer, networks, and optimizer, summarized in Table 1, to ensure a fair comparison. On the DeepMind Control Suite we build on the codebase of RFM (Li et al., 2026) (`https://github.com/azizanlab/ReverseFlowMatching`); on HumanoidBench we build on the codebase of QFlex (Wei et al., 2026) (`https://github.com/LNSGroup/Qflex`). Each algorithm is implemented following its open-source implementation. All training and evaluation runs are conducted on a desktop with an NVIDIA GeForce RTX 5090 GPU and an Intel Core Ultra 9 285K CPU.

Method-specific settings. SCF uses the latent parameterization of Section 3.4 with $\alpha_t = t$, $\beta_t = 1 - t$, and the coefficients (44) with $\eta = 1$ and $C = 10^4$. The overall procedure is summarized in Algorithm 1. For sampling, each baseline uses the NFEs of its original implementation: 10 for RFM, 16 for DIME, and 20 for the others; SCF uses 10. For evaluation, each baseline also follows its own choice of best-of-n action selection, in which n candidate actions are sampled and the one with the highest critic value is executed: RFM, QVPO, MaxEntDP, SDAC, and DIME use $n = 32$,

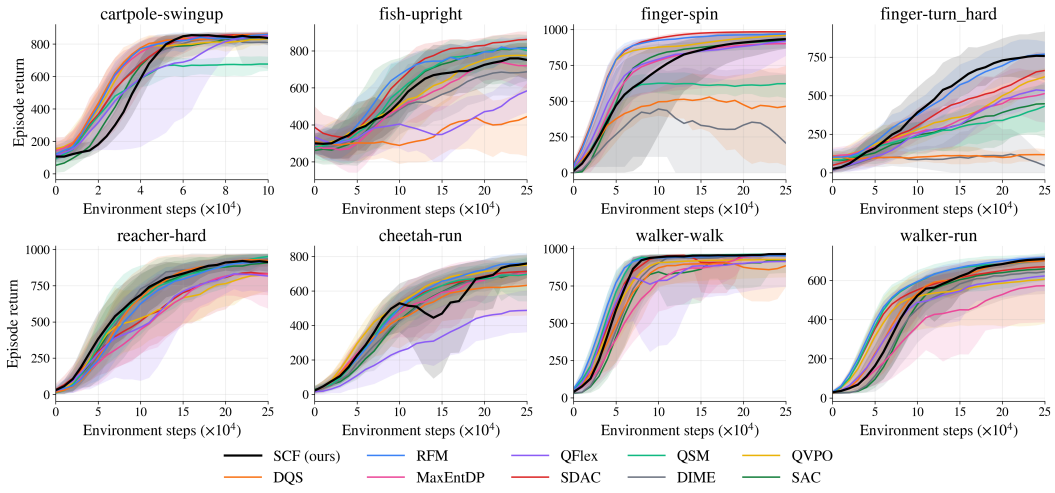

Figure 3: Training curves on eight tasks from the DeepMind Control Suite. The solid lines correspond to the mean across five seeds, and the shaded regions represent minimum-to-maximum range.

Table 2: Milliseconds per training iteration on walker-run (DeepMind Control Suite) and h1-run (HumanoidBench), measured on one NVIDIA GeForce RTX 5090 with batch size 256 and averaged over 1000 steps.

Task	SAC	SCF (ours)	QSM	RFM	QFlex	SDAC	DIME	QVPO	MaxEntDP	DQS
walker-run	0.47	0.87	1.08	2.67	1.94	2.43	3.24	4.40	6.11	7.51
h1-run	1.02	1.12	1.21	2.08	2.99	3.15	3.33	4.58	6.28	11.66

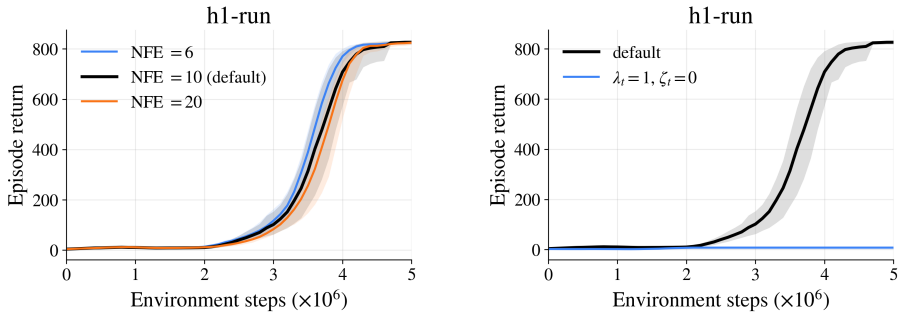

Figure 4: Training curves of SCF on the h1-run task under different design choices. Left: experiments on NFEs for sampling. Right: experiments on the coefficients of the regression target. The solid lines correspond to the mean across five seeds, and the shaded regions represent minimum-to-maximum range.

whereas DQS, QSM, and QFlex use $n = 1$. SCF uses $n = 32$ on the DeepMind Control Suite and $n = 1$ on HumanoidBench.

Results. The training curves are shown in Figures 2 and 3. The wall-clock time per training iteration for each method is summarized in Table 2. The results demonstrate that SCF consistently ranks among the top-performing methods across all tasks, while also being the most efficient in the evaluated generative online RL algorithms.

D Ablation Studies

In this section, we analyze how different design choices affect the performance of SCF. We conduct the experiments on the h1-run task from HumanoidBench. All variants follow the settings of Appendix C.2 and are trained with five seeds; the training curves are shown in Figure 4.

First, we vary the NFEs used to sample actions during both training and evaluation. SCF uses 10 NFEs by default; we compare with 6 and 20 NFEs. As shown in Figure 4 (left), the three settings perform similarly, indicating that the algorithm is robust to the choice of NFEs.

Second, we study the coefficient choices in the regression target (18). Our default choice (44) combines the conditional velocity, current velocity, and score-based term, with $\lambda_t = \eta\alpha_t^2$ canceling the inverse powers of α_t in the score-based contribution (Proposition B.6). As an ablation, we set $\lambda_t = 1$ and $\zeta_t = 0$, leaving only the score-based term. Since this term is unbounded as $t \to 0$, we sample t uniformly from $[0.05, 1]$ to avoid the singularity. Nevertheless, this variant fails to learn: for every seed, the regression loss exceeds 10^{24} within the first 2×10^6 environment steps, and the return never improves (Figure 4, right). These results confirm that the singularity cancellation developed in Appendix B.4 is both theoretically well motivated and practically essential for stable training.