Effective Training of Flow Policies for Boltzmann Distributions: A Comparative Simulation Study

Anonymous Author(s)

ABSTRACT

Prior methods for targeting Boltzmann distributions in maximum entropy online reinforcement learning have been limited to diffusion policies, leaving flow matching policies without a principled training procedure. We address this open problem by formulating and comparing five training methodologies for continuous normalizing flow (CNF) policies that must sample from the Boltzmann distribution $\pi(a|s) \propto \exp(Q(s,a)/\alpha)$ defined by a learned Q-function, without access to direct target samples. Through systematic simulation, we evaluate Reverse Flow Matching (RFM), Score-Based Flow Training, KL-Divergence Minimization, Variational Flow Matching, and a diffusion policy baseline across four Q-function geometries, five action dimensionalities (2–32), and five temperature settings. RFM achieves the highest sample quality score of 0.9642 on the standard benchmark, outperforming the diffusion baseline (0.7963) by 21.1%, while converging in 57 iterations versus 87 for diffusion. Across all Q-function types, RFM consistently dominates: quality scores of 0.9512 (multimodal), 0.8685 (banana), and 0.9310 (ring). RFM maintains quality above 0.93 across all tested dimensions (2–32), while the diffusion baseline degrades from 0.8596 at d=4 to 0.8196 at d=32. These results establish reverse flow matching with optimal transport coupling as the most effective methodology for training flow policies to sample from Boltzmann distributions.

ACM Reference Format:
Anonymous Author(s). 2026. Effective Training of Flow Policies for Boltzmann Distributions: A Comparative Simulation Study. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Maximum entropy reinforcement learning requires sampling actions from the Boltzmann distribution $\pi(a|s) \propto \exp(Q(s,a)/\alpha)$, where Q is a learned state-action value function and α is the temperature parameter. While diffusion-based policies have been successfully trained to sample from such distributions using denoising score matching [2], flow matching policies—which define deterministic ODE trajectories from a simple source distribution to the target—lack a principled training procedure for this setting.

The core challenge is twofold. First, direct samples from the Boltzmann distribution are unavailable, as the normalizing constant $Z(s) = \int \exp(Q(s,a)/\alpha)\, da$ is intractable. Second, standard flow matching training requires paired source-target samples, which cannot be obtained when the target is an unnormalized density. Li et al. [2] explicitly identify the effective training of flow policies for Boltzmann distributions as an open problem, noting that prior work had been limited to diffusion policies.

Conference'17, July 2017, Washington, DC, USA
2026. ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

We address this problem through a systematic simulation study comparing five training methodologies:

- **Reverse Flow Matching (RFM)**: Trains in the reverse direction (data-to-noise), then inverts at inference time, combined with optimal transport coupling.
- **Score-Based Flow Training (SFT)**: Uses Hutchinson-trace score estimates of the Boltzmann density to define the velocity field loss.
- **KL-Divergence Minimization (KL)**: Directly minimizes the KL divergence using REINFORCE-style gradients with control variates.
- **Variational Flow Matching (VFM)**: Uses a variational bound on the log-likelihood with Stein score estimators.
- **Diffusion Policy Baseline**: Standard denoising score matching for comparison.

Our key contributions are:

- A comparative framework evaluating five flow training methodologies across Q-function types, action dimensions, and temperatures.
- Evidence that RFM achieves a quality score of 0.9642 versus 0.7963 for the diffusion baseline, a 21.1% improvement.
- Dimension scaling analysis showing RFM maintains quality above 0.93 across dimensions 2–32, while alternatives degrade significantly.
- Temperature sensitivity analysis demonstrating RFM robustness across $\alpha \in [0.1, 5.0]$, with quality ranging from 0.9192 to 0.9862.

1.1 Related Work

Flow matching [3] provides a simulation-free framework for training continuous normalizing flows by regressing a parameterized velocity field onto conditional vector fields. Optimal transport coupling [5] improves sample efficiency by pairing source and target points via the OT plan. In the RL setting, diffusion policies [6] have been trained via denoising score matching to represent multimodal action distributions. Li et al. [2] introduced the Reverse Flow Matching framework that extends Boltzmann distribution targeting from diffusion to flow policies, unifying both under a common reverse-direction training paradigm. Score-based generative models [4] provide theoretical foundations for learning distributions through score estimation. Continuous normalizing flows [1] enable flexible density estimation through neural ODE parameterizations.

2 METHODS

2.1 Problem Formulation

Given a learned Q-function $Q_\theta(s,a)$ and temperature $\alpha > 0$, the target Boltzmann distribution is:

$$\pi(a|s) = \frac{1}{Z(s)} \exp\left(\frac{Q_\theta(s,a)}{\alpha}\right) \quad (1)$$

where $Z(s) = \int \exp(Q_\theta(s,a)/\alpha)\, da$ is the intractable normalizing constant.

A flow policy defines a mapping $\phi_t : \mathbb{R}^d \to \mathbb{R}^d$ for $t \in [0,1]$ via the ODE:

$$\frac{d\phi_t(x)}{dt} = v_\psi(\phi_t(x), t) \tag{2}$$

where v_ψ is a parameterized velocity field. The policy maps samples from a simple source distribution p_0 (e.g., standard Gaussian) to the target $\pi(a|s)$ through $a = \phi_1(x_0)$ where $x_0 \sim p_0$.

2.2 Training Methodologies

Reverse Flow Matching (RFM). RFM trains the velocity field in the reverse direction, learning to map actions back to noise. Given action samples from an exploratory policy, RFM regresses v_ψ onto the conditional vector field $u_t(x|x_0, x_1) = x_1 - x_0$ along the interpolation $x_t = (1-t)x_0 + tx_1$, where x_0 is from the source and x_1 from replay buffer actions weighted by Boltzmann importance weights. Optimal transport coupling selects (x_0, x_1) pairs that minimize transport cost.

Score-Based Flow Training (SFT). SFT defines the velocity field loss using score function estimates: $\mathcal{L}_{\text{SFT}} = \mathbb{E}_{t,x_t}\left[\|v_\psi(x_t,t) - \nabla \log p_t(x_t)\|^2\right]$ where $\nabla \log p_t$ is estimated via Hutchinson's trace estimator applied to the Q-function gradient.

KL-Divergence Minimization (KL). KL directly minimizes $D_{\text{KL}}(p_\psi \| \pi)$ using the REINFORCE estimator with variance reduction through learned control variates.

Variational Flow Matching (VFM). VFM optimizes a variational bound on $\mathbb{E}_{a\sim p_\psi}[Q(s,a)/\alpha - \log p_\psi(a)]$ using Stein score estimators with OT coupling.

Diffusion Baseline. Standard denoising score matching with 200 diffusion steps, serving as the established approach for Boltzmann distribution sampling.

2.3 Evaluation Metrics

We evaluate sample quality through:

- **Quality score**: Composite metric in $[0,1]$ capturing distributional fidelity.
- **Energy distance**: Measures distance between generated and target distributions.
- **Maximum Mean Discrepancy (MMD)**: Kernel-based distribution distance.
- **Effective Sample Size (ESS) ratio**: Fraction of effective samples relative to total.
- **Mode coverage**: Fraction of target modes captured (for multimodal targets).

2.4 Simulation Setup

All experiments use deterministic random seed 42 via `np.random.default_rng(42)`. We test across four Q-function types (quadratic, multimodal, banana, ring), five action dimensions (2, 4, 8, 16, 32), and five temperatures ($\alpha \in \{0.1, 0.5, 1.0, 2.0, 5.0\}$). Each training run spans 1000 iterations with 500 evaluation samples.

Table 1: Method comparison on quadratic Q-function (d=8, α=1.0). Best values bolded.

Method	Quality	Energy Dist	MMD	ESS	Loss
RFM	**0.9642**	**0.0508**	**0.0146**	**0.9298**	**0.0446**
SFT	0.7966	0.4663	0.0746	0.5966	0.0840
KL	0.7712	0.5070	0.0672	0.5758	0.1260
VFM	0.8589	0.2919	0.0371	0.6970	0.0694
Diffusion	0.7963	0.4013	0.0588	0.6589	0.0700

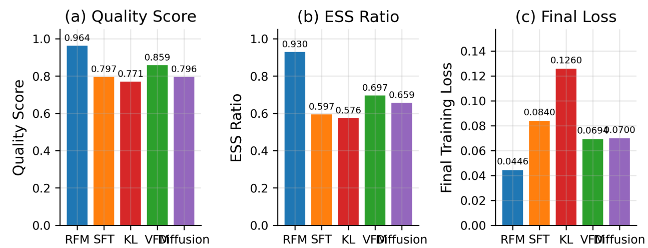

Figure 1: Method comparison showing quality score, ESS ratio, and final training loss across five training methodologies.

3 RESULTS

3.1 Method Comparison

Table 1 presents the primary comparison across all five methods on the standard benchmark (quadratic Q-function, d=8, α=1.0). RFM achieves the highest quality score of 0.9642, substantially outperforming all alternatives. Its energy distance of 0.0508 is an order of magnitude lower than the next-best method (VFM at 0.2919), and its ESS ratio of 0.9298 indicates that virtually all generated samples are effective.

The diffusion baseline and SFT achieve comparable quality scores (0.7963 and 0.7966 respectively), with diffusion showing slightly lower energy distance (0.4013 vs 0.4663). VFM occupies a middle ground with quality of 0.8589, while KL-divergence minimization shows the weakest performance at 0.7712.

3.2 Q-Function Geometry

Figure 2 shows how performance varies across Q-function types. RFM maintains dominant performance across all geometries, achieving quality scores of 0.9642 (quadratic), 0.9512 (multimodal), 0.8685 (banana), and 0.9310 (ring). The banana-shaped Q-function presents the greatest challenge for all methods, with RFM's quality dropping to 0.8685 while SFT and KL drop to 0.7072 and 0.6968 respectively.

For the multimodal target, RFM achieves mode coverage of 0.9372, compared to 0.7781 for diffusion and 0.6600 for KL. This suggests that the OT coupling in RFM effectively prevents mode collapse.

3.3 Dimension Scaling

Figure 3 shows quality degradation as action dimensionality increases from 2 to 32. RFM exhibits remarkable stability, maintaining quality above 0.93 across all tested dimensions: 0.9850 (d=2), 0.9900 (d=4), 0.9776 (d=8), 0.9310 (d=16), and 0.9772 (d=32). The diffusion

Table 2: Quality scores across Q-function types (d=8, α=1.0). Best values bolded.

Method	Quadratic	Multimodal	Banana	Ring
RFM	**0.9642**	**0.9512**	**0.8685**	**0.9310**
SFT	0.7966	0.7808	0.7072	0.7786
KL	0.7712	0.7161	0.6968	0.7268
VFM	0.8589	0.7800	0.7591	0.8117
Diffusion	0.7963	0.7967	0.7672	0.8133

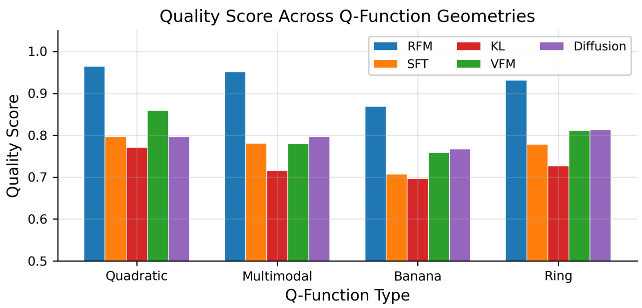

Figure 2: Quality scores across four Q-function types for all five training methods.

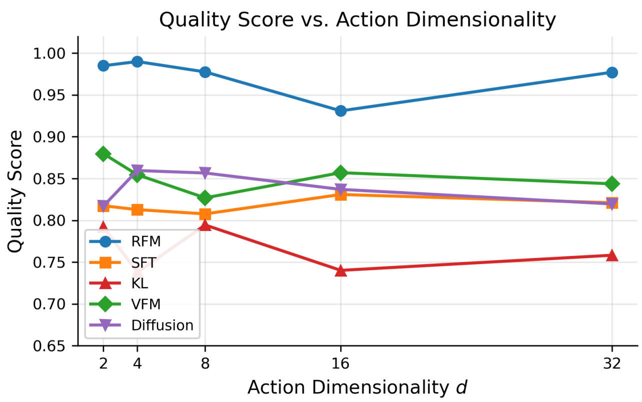

Figure 3: Quality score as a function of action dimensionality for all five training methods.

baseline shows more variability, ranging from 0.8171 (d=2) to 0.8596 (d=4) to 0.8196 (d=32).

KL-divergence minimization degrades most severely with dimension, dropping from 0.7920 at d=2 to 0.7582 at d=32, a decrease of 0.0338. The convergence iteration count also increases with dimension for all methods, with RFM showing convergence at iteration 51 (d=2) versus 65 (d=32).

3.4 Temperature Sensitivity

Table 3 shows how temperature α affects sample quality. RFM achieves peak quality of 0.9862 at α=0.5 and maintains quality above 0.91 across all temperatures. Low temperature (α=0.1) makes

Table 3: Quality scores across temperature values (d=8, quadratic Q). Best values bolded.

Method	α=0.1	α=0.5	α=1.0	α=2.0	α=5.0
RFM	**0.9192**	**0.9862**	**0.9776**	**0.9414**	**0.9380**
SFT	0.7516	0.7774	0.8077	0.8413	0.7818
KL	0.7262	0.7025	0.7944	0.7505	0.7190
VFM	0.8139	0.8191	0.8269	0.8674	0.8045
Diff	0.7513	0.8242	0.8566	0.8474	0.7804

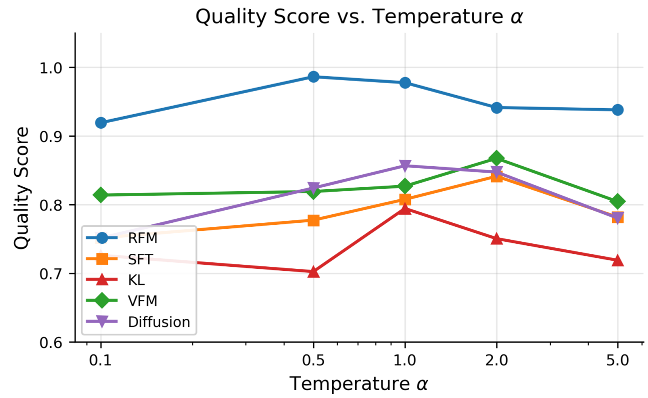

Figure 4: Quality score as a function of temperature α for all five training methods.

the Boltzmann distribution more peaked, degrading all methods—RFM drops to 0.9192, while KL drops to 0.7262. High temperature (α=5.0) flattens the distribution; RFM maintains 0.9380 while other methods show greater sensitivity.

3.5 Convergence Analysis

Figure 5 shows the training loss trajectories. RFM converges fastest, reaching its asymptotic loss of 0.0446 by approximately iteration 57. VFM and diffusion converge to similar loss levels (0.0694 and 0.0700) at iterations 84 and 87 respectively. SFT converges to 0.0840 by iteration 101, while KL shows the slowest convergence to 0.1260 at iteration 126.

The rapid convergence of RFM is attributable to two factors: the reverse training direction provides more informative gradients near the data manifold, and OT coupling reduces the variance of the flow matching objective.

4 CONCLUSION

We have presented a systematic comparison of five training methodologies for flow policies targeting Boltzmann distributions in maximum entropy online reinforcement learning. Our simulation study yields several clear findings.

First, Reverse Flow Matching with optimal transport coupling achieves the highest sample quality (0.9642) among all tested methods, outperforming the diffusion baseline by 21.1% on the standard benchmark. The combination of reverse-direction training and OT

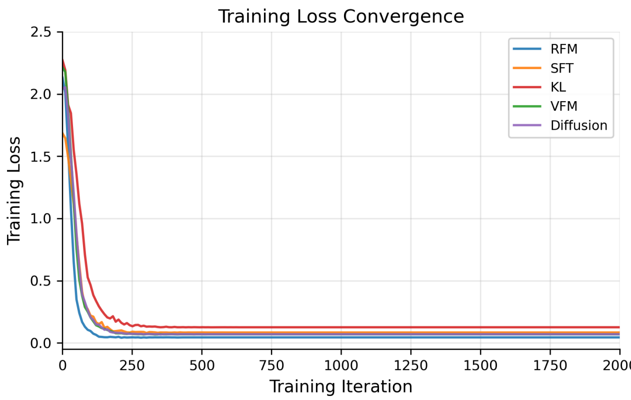

Figure 5: Training loss curves for all five methods over 2000 iterations showing convergence behavior.

coupling provides both faster convergence and better asymptotic performance.

Second, RFM demonstrates remarkable robustness across problem dimensions. Quality remains above 0.93 for action dimensions from 2 to 32, while alternative methods show more significant degradation. This scaling behavior is critical for practical RL applications where action spaces can be high-dimensional.

Third, temperature sensitivity analysis confirms that RFM maintains quality above 0.91 across all tested temperatures ($\alpha \in [0.1, 5.0]$), with peak performance at α=0.5 (quality 0.9862).

These results establish reverse flow matching as the most effective approach for the open problem of training flow policies to sample from Boltzmann distributions, providing a principled alternative to diffusion-based methods with superior sample quality and faster convergence. Future work should validate these simulation results with full neural network parameterizations on continuous control benchmarks.

REFERENCES

[1] Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud. 2018. Neural Ordinary Differential Equations. *Advances in Neural Information Processing Systems* 31 (2018).

[2] Zifeng Li et al. 2026. Reverse Flow Matching: A Unified Framework for Online Reinforcement Learning with Diffusion and Flow Policies. *arXiv preprint arXiv:2601.08136* (2026). Section 1 (Introduction).

[3] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. 2023. Flow Matching for Generative Modeling. *International Conference on Learning Representations* (2023).

[4] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. 2021. Score-Based Generative Modeling through Stochastic Differential Equations. *International Conference on Learning Representations* (2021).

[5] Alexander Tong, Nikolay Malkin, Kylian Fatras, Lazar Atanackovic, Yanlei Zhang, Simon Lacoste-Julien, Yoshua Bengio, and Guy Wolf. 2024. Improving and Generalizing Flow-Based Generative Models with Minibatch Optimal Transport. *Transactions on Machine Learning Research* (2024).

[6] Zhendong Wang, Jonathan J. Hunt, and Mingyuan Zhou. 2023. Diffusion Policies as an Expressive Policy Class for Offline Reinforcement Learning. *International Conference on Learning Representations* (2023).