Score-Regularized Joint Sampling with Importance Weights for Flow Matching

Xinshuang Liu Runfa Blark Li Shaoxiu Wei Truong Nguyen

University of California, San Diego
San Diego, CA, USA
xil235@ucsd.edu rul002@ucsd.edu shwei@ucsd.edu tqn001@ucsd.edu
https://XinshuangL.github.io/SRIW-Flow

Abstract

Flow matching models effectively represent complex distributions, yet estimating expectations of functions of their outputs remains challenging under limited sampling budgets. Independent sampling often yields high-variance estimates, especially when rare but high-impact outcomes dominate the expectation. We propose a **non-IID sampling framework** that jointly draws multiple samples to cover diverse, salient regions of a flow matching model's generative distribution. To balance diversity and quality, we introduce a **score-based regularization for the diversity mechanism** (SR), which uses the score function, *i.e.*, the gradient of the log probability, to ensure samples are pushed apart within high-density regions of the data manifold, mitigating off-manifold drift. To enable unbiased estimation when desired, we further develop an approach for **importance weighting of non-IID flow samples** by learning a residual velocity field that reproduces the marginal distribution of the non-IID samples and by evolving importance weights *along trajectories*. Empirically, our method produces diverse, high-quality samples and accurate importance-weight estimates and debiased expectation estimates, advancing the reliable characterization of flow matching model outputs.

1 INTRODUCTION

Flow matching models [Lipman et al., 2023] are powerful tools for representing complex distributions. Beyond drawing individual samples, many applications require expectations of *functions* of model outputs [Favero et al., 2025]. Formally, for a fixed, pre-trained flow matching model with distribution $p(x)$ and a task-defined functional $f : \mathcal{X} \to \mathbb{R}^d$, we seek

$$\mu = \mathbb{E}_{X\sim p}\big[f(X)\big]. \tag{1}$$

As an illustrative example, consider estimating the expected rating of a text-to-image model under a fixed prompt, where $f(X)$ rates a generated image X (*e.g.*, a user rating) and μ is the model's expected score. A common practice is IID Monte Carlo with $X^{(1)}, \ldots, X^{(n)} \sim p$,

$$\hat{\mu}_{\text{IID}} = \frac{1}{n}\sum_{i=1}^{n} f\big(X^{(i)}\big). \tag{2}$$

However, when sampling is costly, n is small and $\hat{\mu}_{\text{IID}}$ can have high variance, especially for rare but high-impact outcomes (*e.g.*, occasional surprising or disappointing images). To cover such modes under a fixed budget, we draw the n samples *jointly* and encourage them to be diverse. Enforcing diversity, however, changes each sample's marginal distribution, so a plain average is biased; *importance weighting* corrects this, recovering an unbiased estimate of μ. This motivates pairing *joint, diversity-enhancing sampling* with *importance weighting*.

From joint sampling to unbiased estimation. Let $X^{(1:n)} \sim p_{\text{joint}}$ denote n samples drawn *once* from a joint distribution that encourages diversity. Denote by $p'(x)$ the marginal density of any single element under p_{joint}. An unbiased estimator is then (see Appendix A for details)

$$\hat{\mu}_{\text{N-IID}} = \frac{1}{n}\sum_{i=1}^{n} w\big(X^{(i)}\big)\, f\big(X^{(i)}\big), \quad w(x) = \frac{p(x)}{p'(x)}. \tag{3}$$

Thus, a *good* non-IID strategy for estimating μ must satisfy two goals: **(G1) Diversity with quality**—the jointly drawn set covers p's support while keeping each sample on-manifold; and **(G2) Unbiasedness**—we can accurately estimate the per-sample importance weights $w(x)$.

Limitations of existing diverse samplers. Recent work explores non-IID sampling to improve diversity. *Particle Guidance* (PG) [Corso et al., 2024] couples diffusion

Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence (UAI 2026), PMLR 337:3957–3978.

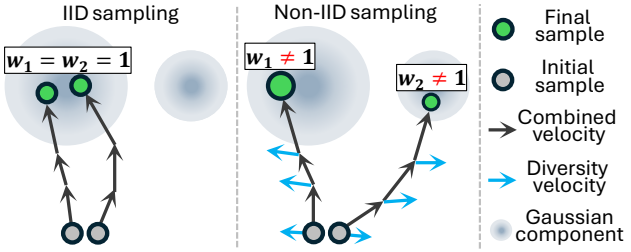

Figure 1: **Illustration of importance-weighted non-IID sampling.** Under IID sampling, both samples are likely drawn from the same dominant mode. In contrast, diversity velocity encourages samples to diverge along their trajectories, leading to coverage of multiple modes. To correct the resulting sampling bias, importance weights are required; intuitively, $w_1 > 1 > w_2$, since non-IID sampling draws the second sample from a minor mode.

trajectories via a repulsive potential, while *DiverseFlow* (DF) [Morshed and Boddeti, 2025] injects diversity into flow matching via a determinantal point process (DPP) objective [Kulesza et al., 2012]. Both can be viewed as augmenting the sampling dynamics with a *diversity velocity* that pushes concurrent trajectories apart. When jointly sampling n trajectories $X_t^{(1:n)}$ from a flow matching model, these methods solve n coupled ODEs

$$\dot{X}_t^{(i)} = v\big(X_t^{(i)}, t\big) + u\big(X_t^{(i)}, X_t^{(-i)}, t\big), \quad X_0^{(i)} \overset{\text{IID}}{\sim} p_0, \tag{4}$$

where i indexes sample trajectories, $X_t^{(-i)}$ denotes the set excluding $X_t^{(i)}$, $t \in [0,1]$, v is the pre-trained flow-matching velocity field (parameters omitted since they remain fixed), and u computes a diversity velocity that pushes samples apart. Although effective at increasing diversity, these methods face a trade-off: strong diversity velocities improve spread but risk pushing samples into low-density, off-manifold regions; weak velocities preserve quality but limit diversity gains. Moreover, these approaches do *not* provide the importance weights needed to form unbiased estimators of expectations like μ, so equal-weighted averages generally introduce bias.

Overview of our approach. We propose a *score-regularized non-IID sampling framework with importance weights* that achieves both (G1) and (G2) (Figure 1). For a pre-trained flow matching model v and n coupled trajectories $\{X_t^{(i)}\}_{i=1}^n$, we *constrain the direction* of g using the model's score $\nabla_x \log p(x \mid t)$ so diversity pushes samples *within* high-density regions (along the data manifold) rather than ejecting them off-manifold. Since g will be normalized in practice, removing components that push samples off-manifold will amplify on-manifold components. This yields sets that are both diverse and high-quality, alleviating the core trade-off observed in prior work. To obtain unbiased estimates, we provide (to our knowledge) the *first* method to compute importance weights for non-IID, jointly sampled outputs from flow-matching models. Directly estimating the marginal $p'(x)$ of the joint sampler is challenging because $X^{(1:n)}$ is drawn only once. We address this by learning a lightweight *residual velocity* $r_\phi(x,t)$ such that the perturbed flow

$$\dot{X}_t = v(X_t, t) + r_\phi(X_t, t), \quad X_0 \sim p_0, \tag{5}$$

matches the *marginal* distribution induced by the diversity-coupled sampler at $t = 1$. Since rectified-flow models [Liu et al., 2023] are widely used in recent flow matching work [Esser et al., 2024, Labs et al., 2025, Labs, 2024], we focus on rectified flows here and include discussions for general flow-matching models in Section 3.2. For rectified flows, learning such a residual velocity field is sufficient for importance-weight estimation, and a smaller network for r_ϕ than for v can be used to reduce training and inference cost. Once trained for a given joint-sampling configuration, the residual can be reused across many test evaluations, so the training cost is amortized across the expectations and thus is negligible.

Empirical validation. We evaluate our method comprehensively. For accurate diagnosis, we first test on a Gaussian mixture model, where the true density (under IID sampling) is available in closed form. Our method improves sample diversity and quality and yields more accurate importance-weight estimates and debiases expectation estimation. We then evaluate advanced conditional flow-matching models: Stable Diffusion 3.5 Medium [Esser et al., 2024] for text-to-image generation and FLUX.1-Fill-dev [Labs et al., 2025, Labs, 2024] for image inpainting, where our method improves coverage of the output distribution as a plug-in via score-based regularization. Across these tasks, our approach produces diverse, high-quality samples and accurate importance-weight estimates that debias expectation estimation, and demonstrates potential gains for large models.

Contribution and Significance. We propose a non-IID sampling framework for flow matching. We constrain diversity directions using model scores to maintain sample quality while enhancing coverage. For unbiased expectation estimation, we further introduce a novel method to compute importance weights for non-IID flow samples. The correctness of the proposed method is theoretically proven, followed by comprehensive empirical validation of its effectiveness. By open-sourcing our code, this work facilitates the management of the diversity–quality trade-off of flow matching generative models and effective importance weight estimation for unbiased expectation estimation, all of which are fundamental and important in flow matching generative models.

2 PRELIMINARIES & RELATED WORK

2.1 FLOW MATCHING MODELS

Flow matching models [Lipman et al., 2023] generate a sample X_1 by integrating an ODE from a base distribution to the target distribution

$$\dot{X}_t = v_\theta(X_t, t), \quad X_0 \sim p_0\,, \tag{6}$$

from $t = 0$ to $t = 1$, where $v_\theta(x, t)$ is a learned time-dependent velocity field and p_0 is a tractable base density (*e.g.*, a standard Gaussian). The ODE can be solved by a first-order integrator such as Euler's method:

$$X_{t_{i+1}} = X_{t_i} + (t_{i+1} - t_i)\, v_\theta(X_{t_i}, t_i), \tag{7}$$

where $0 = t_0 < \cdots < t_N = 1$ and N is the number of integration steps. The evolution of the sample density $p_t(x)$ at a fixed x is given by the continuity equation:

$$\partial_t p_t(x) = -\nabla_x \cdot \big(v_\theta(x, t)p_t(x)\big), \quad p_{t=0}(x) = p_0(x). \tag{8}$$

One widely used variant of flow matching is the *rectified flow* model [Liu et al., 2023], adopted by Stable Diffusion 3 [Esser et al., 2024] due to its effectiveness. In rectified flow, given a sample $X_0 \sim p_0$ and a target sample X_1 (data), the intermediate state is defined by linear interpolation $X_t = (1 - t)\, X_0 + t\, X_1$. The velocity field is then trained by

$$\min_\theta \mathbb{E}_{X_0, X_1, t}\Big[\big\|(X_1 - X_0) - v_\theta(X_t, t)\big\|_2^2\Big], \tag{9}$$

which leads to $v_\theta(x, t) \approx \mathbb{E}[\, X_1 - X_0 \mid X_t = x\,]$, meaning the learned velocity at any point x approximates the expected remaining change needed to reach the target. An important consequence is that the model's score function $s(x, t)$ can be computed in closed form from the velocity field and the current sample (Appendix B)

$$s(x, t) := \nabla_x \log p_\theta(x \mid t) \approx \frac{t\, v_\theta(x, t) - x}{1 - t}. \tag{10}$$

When we use a trained velocity field, this identity is an approximation. We will later use this relationship to guide diversity in our sampling method.

2.2 ENHANCING DIVERSITY OF FLOW MATCHING MODELS

When drawing multiple samples concurrently from a generative model, a key goal is to ensure they cover different high-probability regions of p_θ—in other words, to encourage *diversity*. Recent methods for diffusion [Rombach et al., 2022] and flow [Lipman et al., 2023] models have explored non-IID sampling strategies to increase diversity among n joint samples [Corso et al., 2024, Morshed and Boddeti, 2025, Liu et al., 2026]. The core idea is to introduce a *diversity objective* $h(X_t^{(1:n)})$, where higher h indicates a more diverse set.

A common construction is to define h using normalized pairwise distances K between samples:

$$K_{ij} = \frac{D_{ij}}{\mathrm{med}(D)}, \quad D_{ij} = \big\|x_t^{(i)} - x_t^{(j)}\big\|_2^2, \tag{11}$$

where D_{ij} is the squared distance between samples i and j at time t, and $\mathrm{med}(D)$ is the median of all D_{ij} with $i \neq j$ (treated as a fixed constant to avoid affecting gradients). In practice, the diversity term can be computed using the predicted final samples $\hat{x}_1^{(i)} = x_t^{(i)} + v(x_t^{(i)}, t)(1-t)$ [Morshed and Boddeti, 2025]. We adopt this formulation in all experiments; we omit it from analytical discussion, as it does not affect the conclusions.

To increase h during sampling dynamics, one can compute the gradient of h with respect to each sample as

$$g\big(X_t^{(i)}, X_t^{(-i)}, t\big) = \nabla_{X_t^{(i)}}\, h\big(X_t^{(1:n)}\big), \tag{12}$$

and use a scaled version as the *diversity velocity* u (as in Eq. (4)):

$$u\big(X_t^{(i)}, X_t^{(-i)}, t\big) = \gamma\big(X_t^{(1:n)}, t\big) g\big(X_t^{(i)}, X_t^{(-i)}, t\big). \tag{13}$$

We set

$$\gamma\big(X_t^{(1:n)}, t\big) = \lambda\sqrt{1-t}\, \frac{\Big\|v(X_t^{(1:n)}, t)\Big\|_2}{\Big\|\nabla_{X_t^{(1:n)}} h\big(X_t^{(1:n)}\big)\Big\|_2}, \tag{14}$$

where $v(X_t^{(1:n)}, t)$ is the concatenation of velocities for all $X_t^{(i)}$ and λ controls the intensity. As $t \to 1$, the factor $\sqrt{1-t}$ shrinks the diversity velocity to preserve quality, and the ratio of norms keeps u comparable in magnitude to v. *This augmentation spreads the samples but faces a trade-off*: if u is too strong, samples may be pushed off-manifold (quality drops); if too weak, diversity gains are limited. Moreover, *existing approaches do not provide importance weights for jointly drawn samples; equal-weighted averages generally yield biased estimates.*

3 METHODOLOGY

Our approach addresses both (G1) *diversity with quality* and (G2) *unbiasedness* by modifying the sampling of a pre-trained flow. We generate a batch of n samples jointly, $X^{(1:n)} = (X^{(1)}, \ldots, X^{(n)})$, from velocity $v(x, t)$ with base distribution $p_0(x)$. We encourage diversity while preserving on-manifold quality and assign an importance weight to each sample so that the expectation estimator remains unbiased (as in Eq. (3)). We achieve this with two

components: (1) *score-based diversity-velocity regularization* that pushes samples apart primarily along high-density directions, and (2) a *residual velocity for importance weighting* added to v to model each sample's marginal distribution under the joint sampler.

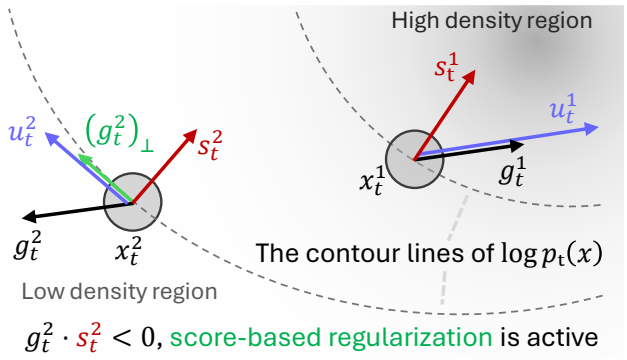

Figure 2: Illustration of our score-based regularization for the diversity mechanism.

3.1 SCORE-BASED DIVERSITY VELOCITY REGULARIZATION

Gradient of a diversity term. As in Section 2.2, define a diversity objective $h\big(X_t^{(1:n)}\big)$ and compute

$$g\big(X_t^{(i)}, X_t^{(-i)}, t\big) = \nabla_{X_t^{(i)}} h\big(X_t^{(1:n)}\big) \tag{15}$$

for each i. Adding a normalized version of g as a diversity velocity increases h during integration. The core trade-off remains: strong diversity improves spread but risks low-density drift; weak diversity preserves quality but limits gains.

Keeping samples within the data manifold. A diversity-driven move $x_t \mapsto x_t + \delta x$ departs from the manifold if it reduces the log-density at time t, *i.e.*, $\log p_t(x_t + \delta x) < \log p_t(x_t)$. To avoid this, constrain the move to satisfy

$$\delta x \cdot s(x,t) \geq 0, \tag{16}$$

where $s(x,t) := \nabla_x \log p_t(x)$ is the score. Decompose

$$g \;=\; g_\parallel + g_\perp, \tag{17}$$

with respect to the unit score direction $\hat{s} = s/\|s\|_2$, where $g_\parallel = (g \cdot \hat{s})\,\hat{s}$ and $g_\perp = g - g_\parallel$. Regularize as

$$g_{\text{reg}} = \alpha(t)\, g_\parallel + g_\perp, \tag{18}$$

with $\alpha(t) = 1$ when $g \cdot \hat{s} \geq 0$ (toward higher density). When $g \cdot \hat{s} < 0$ (toward off-manifold), we consider:

- *Soft*: $\alpha(t) = 1 - t$ (weak early, strong near $t = 1$);
- *Hard*: $\alpha(t) = 0$ (remove the harmful component).

For rectified flows, the score is computed directly from v via Eq. (10). Since $v(x,t)$ is already computed at each step, applying score-based regularization incurs negligible additional overhead. After computing g_{reg} for all samples and jointly normalizing them, we obtain diversity velocities u with *amplified on-manifold components* and reduced (or removed) off-manifold components.

Illustration. In Figure 2, we show two joint samples x_t^1 and x_t^2. Since the scores satisfy $s_t^1 \cdot g_t^1 > 0$ and $s_t^2 \cdot g_t^2 < 0$, regularization is active only for the second sample. Thus, $(g_t^1)_{\text{reg}} = g_t^1$ and $(g_t^2)_{\text{reg}} = (g_t^2)_\perp$ (the hard version). After normalization, the diversity velocities u_1, u_2 push samples apart while avoiding departures from the manifold.

3.2 IMPORTANCE WEIGHT ESTIMATION

Representing the non-IID marginal with a residual velocity. To obtain unbiased estimates, we compute importance weights for non-IID, jointly sampled outputs. Because the joint draw $X^{(1:n)}$ is realized once, the induced single-sample marginal $p'(x)$ is difficult to evaluate directly. We learn a lightweight *residual velocity* $r_\phi(x,t)$ such that the perturbed flow

$$\dot{X}_t = v(X_t,t) + r_\phi(X_t,t), \quad X_0 \sim p_0, \tag{19}$$

matches the *marginal* distribution induced by the diversity-coupled sampler at $t = 1$. Since rectified-flow models [Liu et al., 2023] are widely used [Esser et al., 2024, Labs et al., 2025, Labs, 2024], we focus on rectified flows, where learning such a residual velocity field is sufficient for importance-weight estimation, and a small network (compared to that for v) can be used to reduce training and inference cost.

Our method is also suitable for general flow matching models, but score functions then need to be learned, *e.g.*, via score matching [Song et al., 2021]. We summarize the three ODEs used throughout: the standard Original ODE (Definition 1), the Joint ODE we use for non-IID sampling (Definition 2), and our learned Marginal ODE for importance weighting (Definition 3).

Definition 1 (Original ODE)**.**

$$\dot{X}_t = v(X_t, t), \quad X_0 \sim p_0.$$

We denote its probability density as $p_t(x)$.

Definition 2 (Joint ODEs)**.**

$$\dot{X}_t^{(i)} = v\big(X_t^{(i)}, t\big) + u\big(X_t^{(i)}, X_t^{(-i)}, t\big), \quad X_0^{(i)} \overset{\text{IID}}{\sim} p_0.$$

We denote its marginal probability density as $p'_t(x)$. We use this setup for sampling.

Definition 3 (Marginal ODE)**.**

$$\dot{X}_t = v(X_t, t) + r_\phi(X_t, t), \quad X_0 \sim p_0.$$

We denote its probability density as $p''_{\phi,t}(x)$ and its score as $s''_\phi(x,t) := \nabla_x \log p''_{\phi,t}(x)$. We train r_ϕ such that $p''_{\phi,1}(x) \approx p'_1(x)$ for x generated by the joint ODEs.

Density evolution along the sampling path. We aim to compute the importance weight

$$w(x) = \frac{p_1(x)}{p'_1(x)} \approx \frac{p_1(x)}{p''_{\phi,1}(x)}. \tag{20}$$

To compute this, we define a time-dependent ratio

$$w_{\phi,t}(x) \;:=\; \frac{p_t(x)}{p''_{\phi,t}(x)}, \tag{21}$$

so that $w_{\phi,1}(x) \approx w(x)$ and $w_{\phi,0}(x) = 1$. From Eq. (8), we derive the evolution of $\log w_{\phi,t}(x)$ at a fixed position x, as given in Theorem 1. Since sampling proceeds along the coupled dynamics in Definition 2,

$$\dot{X}_t^{(i)} = v(X_t^{(i)}, t) + u(X_t^{(i)}, X_t^{(-i)}, t), \; i = 1, \ldots, n, \tag{22}$$

combining this with Theorem 1 yields Theorem 2, which characterizes how $w_{\phi,t}(X_t^{(i)})$ evolves along each trajectory. As discussed in Section 3.3, estimating the density ratio along trajectories addresses key limitations of fixed-position estimation, and we therefore adopt this as our default method. For rectified flows, the computation simplifies further—requiring no additional score function estimation—as shown in Corollary 1. All proofs are provided in Appendix C.

Theorem 1 (Evolution of the Importance Weight at Fixed Position)**.** *At a fixed position x, the evolution of $\log w_{\phi,t}(x) = \log p_t(x) - \log p''_{\phi,t}(x)$ is*

$$\begin{aligned}\partial_t \log w_{\phi,t}(x) = \nabla_x \cdot r_\phi(x,t) + s''_\phi(x,t) \cdot r_\phi(x,t) \\ + v(x,t) \cdot \big(s''_\phi(x,t) - s(x,t)\big).\end{aligned}$$

Intuitively, this tracks the density ratio at a fixed point x as time advances; from $w_{\phi,0}(x) = 1$, integrating to $t = 1$ recovers the weight $w(x)$.

Theorem 2 (Evolution of the Importance Weight Along the Trajectory)**.** *Along the sample trajectory, the evolution of $\log w_{\phi,t}(X_t^{(i)})$ is*

$$\begin{aligned}&\frac{\mathrm{d}}{\mathrm{d}t} \log w_{\phi,t}(X_t^{(i)}) \\ &= \nabla_x \cdot r_\phi(X_t^{(i)}, t) + s''_\phi(X_t^{(i)}, t) \cdot r_\phi(X_t^{(i)}, t) \\ &- \big(s''_\phi(X_t^{(i)}, t) - s(X_t^{(i)}, t)\big) \cdot u(X_t^{(i)}, X_t^{(-i)}, t).\end{aligned}$$

Here the ratio is tracked along each sample's own trajectory. We adopt this as our default estimator, since integrating along the sampled path avoids evaluating r_ϕ off its training distribution (Section 3.3).

Corollary 1 (Evolution under Rectified Flows)**.** *Suppose v and $v + r_\phi$ are rectified flows, then Theorem 2 simplifies without learning additional score functions:*

$$\begin{aligned}&\frac{\mathrm{d}}{\mathrm{d}t} \log w_{\phi,t}(X_t^{(i)}) \\ &= \nabla_x \cdot r_\phi(X_t^{(i)}, t) + \tfrac{t v(X_t^{(i)}, t) - X_t^{(i)}}{1-t} \cdot r_\phi(X_t^{(i)}, t) \\ &+ \tfrac{t}{1-t}\big(r_\phi(X_t^{(i)}, t) - u\big(X_t^{(i)}, X_t^{(-i)}, t\big)\big) \cdot r_\phi(X_t^{(i)}, t).\end{aligned}$$

Learning the residual velocity. We train a *rectified flow* to represent the non-IID marginal induced by the diversity-enhanced ODEs. Rather than re-learning v, we learn only the residual r_ϕ, which together with v forms the marginal velocity $v + r_\phi$. We adopt the rectified-flow objective

$$\min_\phi \mathbb{E}_{X_0, X_1, t}\Big[\big\|(X_1 - X_0) - v(X_t, t) - r_\phi(X_t, t)\big\|_2^2\Big], \tag{23}$$

where $X_0 \sim p_0$, X_1 is sampled from the diversity-enhanced non-IID sampler, and $X_t = (1-t)\, X_0 + t\, X_1$. The training target does not require external labels since it is created by our sampler. In practice, a relatively lightweight neural network for r_ϕ can be used to accelerate both training and inference, because it is based on v (See Appendix F for more details).

3.3 DISCUSSIONS

We discuss two variations to justify our design choices.

Variation 1. Directly sampling from the marginal velocity. Since r_ϕ is trained, one can sample independently from $v + r_\phi$ (the ODE in Definition 3), which matches the marginal distribution of diversity-enhanced samples. However, IID batches drawn from $v+r_\phi$ are typically less diverse than those generated by the coupled ODEs in Definition 2, which explicitly enhance diversity across concurrent trajectories (see Table 1).

Variation 2. Estimating importance weights via fixed-position evolution (for samples at $t = 1$). In Theorem 1, we estimate the importance weight at the fixed position for each sample as

$$\log w_{\phi,1}(x) = \int_0^1 \Big(\nabla_x \cdot r_\phi(x,t) + s''_\phi(x,t) \cdot r_\phi(x,t) + v(x,t) \cdot \big(s''_\phi(x,t) - s(x,t)\big) \Big) \mathrm{dt}, \tag{24}$$

which, given v and $v + r_\phi$ are rectified flows, further simplifies (Proof in Appendix C.4):

$$\log w_{\phi,1}(x) = \int_0^1 \Big(\nabla_x \cdot r_\phi(x,t) + \frac{1}{1-t} r_\phi(x,t) \cdot \big(2tv(x,t) + tr_\phi(x,t) - x\big) \Big) \mathrm{dt}. \tag{25}$$

However, in practice, the pair $(X_1^{(i)}, t)$ might be unlikely to occur along the real sampling path when $t < 1$, because $X_1^{(i)}$ is the sample for time $t = 1$. In conditional generation, *i.e.*, generating an image for the text "a cat", $\mathbb{E}X_1^{(i)} \neq \mathbf{0}$ but $\mathbb{E}X_0^{(i)} = \mathbf{0}$, which means $X_1^{(i)}$ can be rare among real samples at $t = 0$, *i.e.*, $p_0(X_1^{(i)}) \approx 0$. This implies that the training data of r_ϕ cannot cover this situation, potentially leading to an undefined output $r_\phi(X_1^{(i)}, t)$ for small t and worsening performance. Our trajectory-based estimator instead integrates *along the actual path* and avoids out-of-distribution inputs to r_ϕ (see Table 2).

4 EXPERIMENTS

We first evaluate on a **Gaussian mixture** with known density, enabling precise diagnosis of diversity, quality, and estimation of importance weights and expectations. We then assess *plug-in* benefits on large flow models—**Stable Diffusion 3.5 Medium** [Esser et al., 2024] for text-to-image generation and **FLUX.1-Fill-dev** [Labs et al., 2025, Labs, 2024] for image inpainting—to study distribution coverage in complex image settings. Diversity velocities and implementation details for each experiment are in Appendix D and Appendix E, respectively.

4.1 EXPERIMENTS ON GAUSSIAN MIXTURE

The experiment uses a mixture of ten Gaussians with centers uniformly placed on a unit circle in the first two dimensions. Each trial jointly samples ten points.

4.1.1 Sample Diversity and Quality

Setup. We use a sparse 8-D near-planar distribution with variances of 0.01 on the first two coordinates and 0.0001 on the remaining six (mimicking a sparse data distribution). We run 10,000 trials, evaluating several diversity objectives and testing whether adding our *score-based regularization* (SR) improves performance. For *diversity*, we assign each sample to its nearest Gaussian and report *mode coverage*: the number of distinct modes captured by the 10 joint samples. We also report a *Marginal* variant by pooling joint samples across repetitions, randomly regrouping them into sets of 10, and computing the resulting coverage—this mimics IID sampling from the *marginal* of the diversity-enhanced sampler. For *quality*, we report $\log p$ (density at each sample computed under the original Gaussian mixture) and RMSE to the nearest mode.

Results. In Table 1, across all diversity objectives, adding SR (soft or hard) *significantly* improves quality ($\log p$ increases; RMSE decreases) while preserving joint mode coverage. For example, with DPP, $\log p$ and RMSE improve from 7.16 and 0.527 to 19.27 and 0.253 under SR-hard, while joint coverage remains almost unchanged ($8.55 \rightarrow 8.57$). By contrast, DiverseFlow (DPP+DF) improves quality but *reduces* diversity ($8.55 \rightarrow 6.89$), evidencing a diversity–quality trade-off that our score-based regularization alleviates. Finally, IID sampling from the diversity-enhanced *marginal* fails to recover joint-diversity benefits (*e.g.*, DPP joint coverage 8.55 vs. marginal 6.51, close to original IID 6.51), justifying our design choice of *joint* sampling via coupled ODEs (Definition 2) rather than IID from the diversity-enhanced marginal (Definition 3).

4.1.2 Importance Weight & Expectation Estimation

Setup. For accurate ground truth, we use a 2D Gaussian mixture, shifted by 1 along the first axis to induce a moderate distribution shift, where each component k has a weight proportional to 2^{-k}. Non-IID joint sets are sampled using a Harmonic DPP with soft SR. The ground-truth importance weights are computed as the ratio between the original density and the non-IID marginal estimated via Local-Likelihood Density Estimation (LLDE) [Loader, 1996, Hjort and Jones, 1996] (see Appendix H). We train the residual velocity r_ϕ on $1{,}000$ joint sets (each containing 10 trajectories) and evaluate on $10{,}000$ sets. We compare three types of estimators: (i) *trajectory-based* evolution (ours, Theorem 2); (ii) *fixed-position* evolution at a terminal

Method	SR	Mode coverage↑ Joint	Mode coverage↑ Marginal	$\log p \uparrow$	RMSE ↓
IID	-	-	6.51(2)	20.31(1)	0.139(1)
PG	✗	7.50(2)	6.51(2)	7.28(7)	0.526(1)
	soft	7.56(2)	6.51(2)	14.94(4)	0.386(1)
	hard	**7.59(2)**	6.52(2)	**19.93(1)**	**0.227(1)**
Chebyshev	✗	8.39(1)	6.51(2)	7.74(6)	0.517(1)
	soft	**8.60(1)**	6.52(2)	15.15(3)	0.379(1)
	hard	8.58(1)	6.52(2)	**19.90(1)**	**0.227(1)**
Reciprocal	✗	9.45(1)	6.51(2)	7.91(6)	0.514(1)
	soft	**9.55(1)**	6.51(2)	14.71(4)	0.390(1)
	hard	**9.55(1)**	6.51(2)	**19.34(2)**	**0.247(1)**
Log-barrier	✗	9.44(1)	6.52(2)	7.50(7)	0.521(1)
	soft	9.62(1)	6.53(2)	15.08(3)	0.380(1)
	hard	**9.63(1)**	6.52(2)	**19.84(1)**	**0.227(1)**
DPP+DF	-	6.89(2)	6.50(2)	**20.96(1)**	**0.121(1)**
DPP	✗	8.55(2)	6.51(2)	7.16(6)	0.527(1)
	soft	8.56(2)	6.50(2)	14.25(3)	0.402(1)
	hard	**8.57(2)**	6.50(2)	19.27(1)	0.253(1)

Table 1: **Sampling diversity and quality results for the Gaussian mixture experiment.** Mode coverage, $\log p$, and RMSE are reported as mean(uncertainty) with 95% confidence intervals. SR (soft and hard) is our score-based regularization, applied on top of each diversity method. For each metric, the best result within each method group is shown in **bold**; the IID reference is excluded.

state (Theorem 1); and (iii) density baselines—k-Nearest Neighbors Density Estimation (kNN) [Loftsgaarden and Quesenberry, 1965], Gaussian Kernel Density Estimation (KDE) [Silverman, 1986], and Multivariate Gaussian Fit (MGF) [Bishop and Nasrabadi, 2007] (see Appendix I).

Importance-weight accuracy. We report squared error (SE) and ranking metrics—Kendall's τ_b [Kendall, 1945], Spearman's ρ [Spearman, 1987], and graded AP (gAP) [Moffat et al., 2022] (see Appendix J)—computed on the top 50% most-confident ground truth. In Table 2, density baselines exhibit large SE and weak ranking, while the *trajectory-based* estimator consistently outperforms the *fixed-position* variant, since path integration avoids out-of-distribution inputs to r_ϕ.

Function expectation. We estimate $\mu{=}\mathbb{E}f(X)$, where $f(x)$ maps each sample to a one-hot vector indicating the index of its nearest Gaussian. Ground-truth μ is computed from $1{,}000{,}000$ IID samples drawn from the original flow-matching model, yielding the probability that a generated sample belongs to each Gaussian. We evaluate on joint samples via Jensen–Shannon (JS) divergence [Lin, 1991], which remains well-defined when some predicted probabilities are

	SE ↓	$\tau_b \uparrow$	$\rho \uparrow$	gAP ↑
KDE	8.0(1)	0.348(6)	0.449(7)	0.736(3)
kNN	9.9(1)	0.327(6)	0.424(8)	0.734(3)
MGF	8.8(1)	0.327(6)	0.427(8)	0.736(3)
Ours†	3.8(1)	0.498(5)	0.621(6)	0.794(2)
Ours	**2.6(1)**	**0.548(5)**	**0.668(6)**	**0.805(2)**

Table 2: **Importance-weight estimation results for the Gaussian mixture experiment.** SE and ranking metrics are reported as mean(uncertainty) with 95% confidence intervals. † indicates the variation of our method that estimates the importance-weight evolution at a fixed, final sample position. The best-performing method for each metric is shown in **bold**.

zero. Our *trajectory-based* weighting debiases the equal-weighted non-IID average (JS $0.088{\to}0.071$) and improves over spending the same budget on additional IID draws (0.077); the *fixed-position* variant (0.078) lies between. The density baselines attain the lowest JS here, but this is potentially because they are additionally given the analytic, closed-form target density p_1, the true density our estimator never accesses; the full comparison is in Appendix I.1.

Summary. Our along-trajectory estimator outperforms its fixed-position variant and the density baselines in importance-weight estimation; used downstream, these weights debias the equal-weighted non-IID average and improve over additional IID sampling.

4.2 EXPERIMENTS ON TEXT-TO-IMAGE GENERATION

We next test Stable Diffusion 3.5 Medium [Esser et al., 2024] for text-to-image generation. We study whether using diverse samples can better cover the IID samples S_{IID} with a fixed budget of samples S_{Rep}, measured by the **coverage radius**

$$\mathcal{R}(S_{\text{IID}}|S_{\text{Rep}}) = \max_{I_1 \in S_{\text{IID}}} \min_{I_2 \in S_{\text{Rep}}} d^2(I_1, I_2), \qquad (26)$$

where $d(I_1, I_2)$ is the distance between two images; to avoid evaluation bias, we do not introduce any embedding model and instead use the distance between their spatial means in latent space. This metric quantifies worst-case geometric coverage in the latent-space representation, which is more meaningful than a perceptual score. The text prompts are:

- T_1: "a fish"
- T_2: "a realistic fish"
- T_3: "a cat"
- T_4: "a releastic cat"
- T_5: "something outside"
- T_6: "someone ahead"

They cover different scenarios: $T_1{\to}T_2$ (coarse→fine conditions), $T_3{\to}T_4$ (low-quality conditions, *e.g.*, typos), and T_5, T_6 (ambiguous conditions). For each prompt, we generate

	SR	T_1	T_2	T_3	T_4	T_5	T_6	Mean
IID	-	22.6(4)	22.5(4)	18.6(2)	19.1(2)	20.5(2)	22.0(2)	20.9(2)
DPP	✗	19.2(3)	18.8(4)	15.3(2)	15.7(2)	17.5(2)	18.8(2)	17.5(2)
	soft	19.0(3)	18.6(4)	15.2(2)	15.6(2)	17.4(2)	18.7(2)	17.4(2)
	hard	**18.7(3)**	**18.4(4)**	**15.1(2)**	**15.5(2)**	**17.2(2)**	**18.4(2)**	**17.2(2)**
PG	✗	18.1(3)	17.0(3)	14.5(2)	14.9(2)	17.6(2)	19.1(2)	16.9(2)
	soft	18.0(3)	16.9(3)	14.5(2)	14.8(2)	17.6(2)	19.0(2)	16.8(2)
	hard	**17.7(3)**	**16.7(3)**	**14.4(2)**	**14.7(2)**	**17.4(2)**	**18.8(2)**	**16.6(2)**
DF	✗	18.9(3)	18.6(4)	15.2(2)	15.6(2)	17.4(2)	18.6(2)	17.4(2)
	soft	18.9(3)	18.5(4)	15.1(2)	15.5(2)	17.3(2)	18.6(2)	17.3(2)
	hard	**18.7(3)**	**18.4(4)**	**15.1(2)**	**15.5(2)**	**17.2(2)**	**18.4(2)**	**17.2(2)**

Table 3: **Coverage radius results on text-to-image generation.** Values are mean(uncertainty) with 95% confidence intervals. SR (soft and hard) is our score-based regularization, applied on top of each diversity method.

Method	SR	T_1	T_2	T_3	T_4	T_5	T_6	T_7	T_8	T_9	T_{10}	Mean
IID	-	2.47(4)	0.22(1)	0.09(1)	0.15(1)	0.46(2)	1.43(5)	0.07(1)	0.09(1)	0.25(1)	0.11(1)	0.53(1)
DPP	✗	2.20(4)	0.41(1)	0.09(1)	0.17(1)	0.35(1)	0.95(5)	0.21(1)	0.13(1)	0.28(1)	0.17(1)	0.50(1)
	soft	**2.19(4)**	0.38(1)	0.09(1)	0.16(1)	0.33(1)	**0.89(4)**	0.20(1)	0.12(1)	0.25(1)	0.16(1)	0.48(1)
	hard	2.26(3)	**0.17(1)**	**0.07(1)**	**0.11(1)**	**0.30(2)**	1.11(5)	**0.10(1)**	**0.08(1)**	**0.19(1)**	**0.09(1)**	**0.45(1)**
PG	✗	2.14(3)	0.53(2)	0.11(1)	0.21(1)	0.49(2)	0.63(4)	0.25(1)	0.16(1)	0.34(1)	0.20(1)	0.51(1)
	soft	**2.13(3)**	0.52(2)	0.10(1)	0.21(1)	0.47(2)	**0.63(4)**	0.24(1)	0.15(1)	0.32(1)	0.19(1)	0.50(1)
	hard	2.21(3)	**0.21(1)**	**0.08(1)**	**0.15(1)**	**0.32(1)**	0.93(5)	**0.15(1)**	**0.10(1)**	**0.18(1)**	**0.10(1)**	**0.44(1)**
DF	✗	2.20(4)	0.40(1)	0.09(1)	0.17(1)	0.34(1)	0.92(4)	0.21(1)	0.13(1)	0.27(1)	0.17(1)	0.49(1)
	soft	**2.19(4)**	0.36(1)	0.09(1)	0.16(1)	0.32(1)	**0.89(5)**	0.19(1)	0.12(1)	0.24(1)	0.16(1)	0.47(1)
	hard	2.24(3)	**0.17(1)**	**0.07(1)**	**0.11(1)**	**0.30(2)**	1.10(5)	**0.10(1)**	**0.08(1)**	**0.20(1)**	**0.09(1)**	**0.45(1)**

Table 4: **Coverage radius results on image inpainting.** Values are mean(uncertainty) with 95% confidence intervals. SR (soft and hard) is our score-based regularization, applied on top of each diversity method.

10,000 images for S_{IID} and 10 images for each S_{Rep} (1,000 repetitions) under evaluation.

Table 3 shows that introducing diversity terms reduces the coverage radius for all prompts T_1 to T_6 relative to IID (taking S_{Rep} as 10 IID samples), indicating improved coverage under a fixed budget. Adding our score-based regularization (SR) further improves performance across all prompts, suggesting that maintaining on-manifold quality increases sample efficiency (fewer "wasted" samples). Qualitative results are shown in Figure 3.

4.3 EXPERIMENTS ON IMAGE INPAINTING

We additionally evaluate performance on image inpainting, which is more constrained than text-to-image, to test whether improvements persist. We use FLUX.1-Fill-dev [Labs et al., 2025, Labs, 2024]. We select 10 square, high-quality images from the test split of the Open Images Dataset V7 [Kuznetsova et al., 2020] and apply a mask with four squares to inpaint. For each image, we generate $|S_{\text{IID}}| = 1{,}000$ inpaintings and evaluate $|S_{\text{Rep}}| = 10$ jointly sampled inpaintings (100 repetitions). The coverage radius (Eq. (26)) is computed over average latent features *within the masked regions*.

Table 4 shows that diversity terms improve the mean coverage radius across images relative to IID. Similar to text-to-image generation, adding SR improves the mean coverage radius for every method, though SR-hard can regress on a few inputs (*e.g.*, T_1 and T_6), indicating that maintaining on-manifold quality increases sample efficiency. Qualitative results are shown in Figure 4.

5 CONCLUSION

We introduced an *importance-weighted non-IID sampling* framework for flow matching that (i) applies *score-based regularization for diversity velocity* to push concurrent trajectories apart while keeping them on-manifold, and (ii) computes *importance weights* along trajectories by learning a residual velocity. On a Gaussian mixture, SR improves

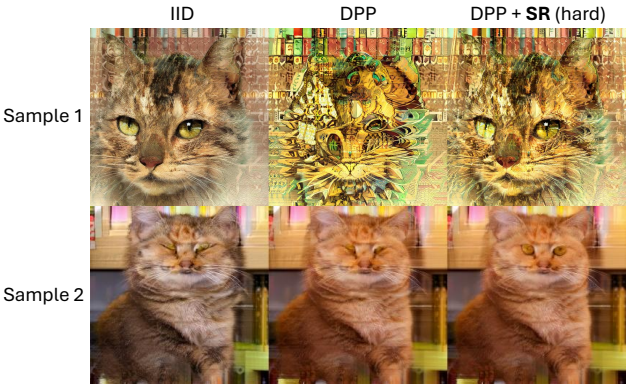

Figure 3: **Qualitative results for text-to-image generation.** Although diverse, **Sample 1** from DPP appears unreasonable. Adding SR (hard) preserves diversity while making it reasonable. For **Sample 2**, SR (hard) refines the cat's eyes.

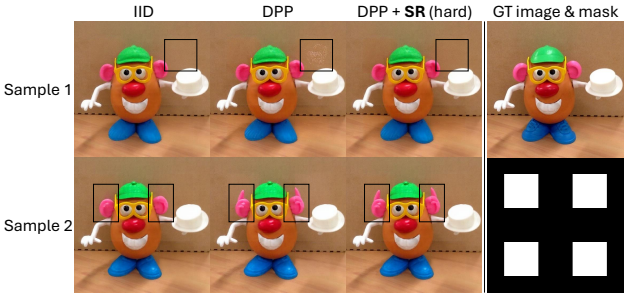

Figure 4: **Qualitative results for image inpainting.** Two samples are shown from the joint samples, with all methods sharing identical initialization. In **Sample 1**, DPP introduces artifacts (highlighted by the black rectangle) that are removed by SR (hard). While enhancing quality, DPP+SR retains the diversity of DPP, as illustrated in **Sample 2**.

sample quality while preserving diversity, and our importance weighting yields the most accurate importance-weight estimates, improving expectation estimation over both equal weighting and additional IID sampling. On Stable Diffusion 3.5 Medium and FLUX.1-Fill-dev, diversity with SR consistently reduces the mean sample coverage radius.

References

Christopher M. Bishop and Nasser M. Nasrabadi. *Pattern Recognition and Machine Learning*. *J. Electronic Imaging*, 16(4):049901, 2007. doi: 10.1117/1.2819119. URL https://doi.org/10.1117/1.2819119.

Gabriele Corso, Yilun Xu, Valentin De Bortoli, Regina Barzilay, and Tommi S. Jaakkola. Particle guidance: non-i.i.d. diverse sampling with diffusion models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=KqbCvIFBY7.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=FPnUhsQJ5B.

Gian Mario Favero, Parham Saremi, Emily Kaczmarek, Brennan Nichyporuk, and Tal Arbel. Conditional diffusion models are medical image classifiers that provide explainability and uncertainty for free. *CoRR*, abs/2502.03687, 2025. doi: 10.48550/ARXIV.2502.03687. URL https://doi.org/10.48550/arXiv.2502.03687.

Nils Lid Hjort and M. C. Jones. Locally parametric nonparametric density estimation. *Annals of Statistics*, 24(4):1619–1647, 1996. doi: 10.1214/aos/1032298288.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *CoRR*, abs/2207.12598, 2022. doi: 10.48550/ARXIV.2207.12598. URL https://doi.org/10.48550/arXiv.2207.12598.

Maurice G Kendall. The treatment of ties in ranking problems. *Biometrika*, 33(3):239–251, 1945.

Leslie Kish. *Survey Sampling*. John Wiley & Sons, New York, 1965.

Alex Kulesza, Ben Taskar, et al. Determinantal point processes for machine learning. *Foundations and Trends® in Machine Learning*, 5(2–3):123–286, 2012.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper R. R. Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. The open images dataset V4. *Int. J. Comput. Vis.*, 128(7):1956–1981, 2020. doi: 10.1007/S11263-020-01316-Z. URL https://doi.org/10.1007/s11263-020-01316-z.

Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space, 2025. URL https://arxiv.org/abs/2506.15742.

Jianhua Lin. Divergence measures based on the shannon entropy. *IEEE Trans. Inf. Theory*, 37(1):145–151, 1991. doi: 10.1109/18.61115. URL https://doi.org/10.1109/18.61115.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=XVjTT1nw5z.

Xinshuang Liu, Runfa Blark Li, and Truong Nguyen. Consistency-preserving diverse video generation. *arXiv preprint arXiv:2602.15287*, 2026.

Clive R. Loader. Local likelihood density estimation. *Annals of Statistics*, 24(4):1602–1618, 1996. doi: 10.1214/aos/1032298287.

Don O Loftsgaarden and Charles P Quesenberry. A nonparametric estimate of a multivariate density function. *The Annals of Mathematical Statistics*, 36(3):1049–1051, 1965.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

John C Mason and David C Handscomb. *Chebyshev polynomials*. Chapman and Hall/CRC, 2002.

Alistair Moffat, Joel Mackenzie, Paul Thomas, and Leif Azzopardi. A flexible framework for offline effectiveness metrics. In Enrique Amigó, Pablo Castells, Julio Gonzalo, Ben Carterette, J. Shane Culpepper, and Gabriella Kazai, editors, *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, pages 578–587. ACM, 2022. doi: 10.1145/3477495.3531924. URL https://doi.org/10.1145/3477495.3531924.

Mashrur M. Morshed and Vishnu Boddeti. Diverseflow: Sample-efficient diverse mode coverage in flows. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 23303–23312. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.02170. URL https://openaccess.thecvf.com/content/CVPR2025/html/Morshed_DiverseFlow_Sample-Efficient_Diverse_Mode_Coverage_in_Flows_CVPR_2025_paper.html.

Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, June 2022.

Bernard W. Silverman. *Density Estimation for Statistics and Data Analysis*. Springer, 1986. ISBN 978-1-4899-3324-9. doi: 10.1007/978-1-4899-3324-9. URL https://doi.org/10.1007/978-1-4899-3324-9.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=PxTIG12RRHS.

Charles Spearman. The proof and measurement of association between two things. *The American journal of psychology*, 100(3/4):441–471, 1987.

Abraham Wald. Note on the consistency of the maximum likelihood estimate. *The Annals of Mathematical Statistics*, 20(4):595–601, 1949.

M. P. Wand and M. C. Jones. *Kernel Smoothing*, volume 60 of *Monographs on Statistics and Applied Probability*. Chapman and Hall/CRC, London, 1995. doi: 10.1201/b14876.

Jiwen Yu, Yinhuai Wang, Chen Zhao, Bernard Ghanem, and Jian Zhang. Freedom: Training-free energy-guided conditional diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23174–23184, 2023.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

Score-Regularized Joint Sampling with Importance Weights for Flow Matching

Xinshuang Liu Runfa Blark Li Shaoxiu Wei Truong Nguyen

University of California, San Diego
San Diego, CA, USA
xil235@ucsd.edu rul002@ucsd.edu shwei@ucsd.edu tqn001@ucsd.edu
https://XinshuangL.github.io/SRIW-Flow

A IMPORTANCE SAMPLING

For completeness, we briefly introduce importance sampling here. While drawing non-IID samples can improve coverage of the support, the marginal density of each sample is generally biased. To estimate the expectation of a test function f as

$$\mathbb{E}_{X\sim p}[f(X)] \approx \frac{1}{n}\sum_{i=1}^{n} f\Big(X^{(i)}\Big), \quad X^{(i)} \overset{\text{IID}}{\sim} p, \tag{27}$$

one requires IID samples from p. Under non-IID sampling, the samples are drawn from a joint distribution

$$(X^{(1)}, \ldots, X^{(n)}) \sim p_{\text{joint}}. \tag{28}$$

This can increase diversity but introduces bias that must be corrected with importance weights:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} w\Big(X^{(i)}\Big)\; f\Big(X^{(i)}\Big). \tag{29}$$

We seek w such that

$$\mathbb{E}_{(X^{(1)},\ldots,X^{(n)})\sim p_{\text{joint}}}\hat{\mu} = \mathbb{E}_{X\sim p}[f(X)]. \tag{30}$$

It is straightforward to verify that

$$w(x) = \frac{p(x)}{p'(x)}, \tag{31}$$

suffices, where p' is the marginal density of p_{joint},

$$p'_i(x) = \int p_{\text{joint}}\big(x^{(1)}, \ldots, x^{(n)}\big)\Big|_{x^{(i)}=x} \mathrm{d}x^{(-i)}, \tag{32}$$

and, assuming exchangeability,

$$p'(x) := p'_1(x) = \cdots = p'_n(x). \tag{33}$$

Proof. Substituting the proposed weights, we obtain

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} w\Big(X^{(i)}\Big)\; f\Big(X^{(i)}\Big), \quad w(x) = \frac{p(x)}{p'(x)}. \tag{34}$$

Thus,

$$\begin{aligned}
&\mathbb{E}_{(X^{(1)},\dots,X^{(n)})\sim p_{\text{joint}}}[\hat{\mu}] \\
&= \int \frac{1}{n}\sum_{i=1}^{n} f(x^{(i)})\, w(x^{(i)})\, p_{\text{joint}}\big(x^{(1:n)}\big)\, \mathrm{d}x^{(1:n)} \\
&= \frac{1}{n}\sum_{i=1}^{n} \int f(x^{(i)})\, w(x^{(i)})\, p_i'(x^{(i)})\, \mathrm{d}x^{(i)} \\
&= \frac{1}{n}\sum_{i=1}^{n} \int f(x^{(i)})\, p(x^{(i)})\, \mathrm{d}x^{(i)} \\
&= \mathbb{E}_{X\sim p}\big[f(X)\big].
\end{aligned} \tag{35}$$

B RELATIONSHIP BETWEEN THE SCORE AND THE VELOCITY IN RECTIFIED FLOW

For completeness, we include the known relationship between the score and the velocity in rectified flow here.

Setup. Let $X_0 \sim p_0 = \mathcal{N}(0, I)$ and $X_1 \sim p_{\text{data}}$ independent. Given X_1 and X_0, $X_t = (1-t)X_0 + tX_1$ with $t \in (0,1)$. Denote by p_t the density of X_t and by $s(x,t) := \nabla_x \log p_t(x)$ its score.

Marginal density of X_t. Solving for X_0 gives

$$X_0 = \tfrac{X_t - tX_1}{1-t}. \tag{36}$$

For fixed X_1, the map $x_0 \mapsto (1-t)x_0 + tX_1$ has Jacobian $(1-t)I$ and determinant $(1-t)^d$. By change of variables,

$$p_{X_t|X_1}(x \mid X_1) = (1-t)^{-d}\, p_0\Big(\tfrac{x-tX_1}{1-t}\Big), \tag{37}$$

hence

$$p_t(x) = \mathbb{E}_{X_1}\left[(1-t)^{-d} p_0\Big(\tfrac{x-tX_1}{1-t}\Big)\right]. \tag{38}$$

Score as a conditional mean. Differentiating Eq. (38) and using $\nabla \log p_0(z) = -z$ for $\mathcal{N}(0, I)$ gives

$$\begin{aligned}
&\nabla_x \log p_t(x) \\
&= \frac{\mathbb{E}_{X_1}\left[(1-t)^{-d} p_0(\cdot)\, \nabla_x \log p_0\big(\tfrac{x-tX_1}{1-t}\big)\right]}{\mathbb{E}_{X_1}[(1-t)^{-d} p_0(\cdot)]} \\
&= -\frac{1}{1-t}\, \frac{\mathbb{E}_{X_1}\left[(1-t)^{-d} p_0(\cdot)\, \frac{x-tX_1}{1-t}\right]}{\mathbb{E}_{X_1}[(1-t)^{-d} p_0(\cdot)]} \\
&= -\frac{1}{1-t}\, \mathbb{E}\big[X_0 \,\big|\, X_t = x\big].
\end{aligned} \tag{39}$$

Thus

$$\mathbb{E}[X_0 \mid X_t = x] = -(1-t)\, s(x,t). \tag{40}$$

Linear constraint on conditional means. Taking conditional expectations of $X_t = (1-t)X_0 + tX_1$ given $X_t = x$,

$$x = (1-t)\, \mathbb{E}[X_0 \mid X_t = x] + t\, \mathbb{E}[X_1 \mid X_t = x]. \tag{41}$$

Let

$$m_0(x,t) := \mathbb{E}[X_0 \mid X_t = x] \tag{42}$$

and

$$m_1(x,t) := \mathbb{E}[X_1 \mid X_t = x]. \tag{43}$$

Using Eq. (40) in Eq. (41) gives

$$t\, m_1(x,t) = x + (1-t)^2 s(x,t). \tag{44}$$

Velocity identity and conclusion. The rectified-flow velocity field is the conditional mean displacement,

$$\begin{aligned} v_\theta(x,t) &= \mathbb{E}[X_1 - X_0 \mid X_t = x] \\ &= m_1(x,t) - m_0(x,t) \\ &= m_1(x,t) + (1-t)s(x,t), \end{aligned} \tag{45}$$

where the last equality uses Eq. (40). Multiplying Eq. (45) by t and substituting Eq. (44) gives

$$t\, v_\theta(x,t) = x + (1-t)s(x,t), \tag{46}$$

which implies

$$s(x,t) = \frac{t\, v_\theta(x,t) - x}{1-t}. \tag{47}$$

C EVOLUTION OF IMPORTANCE WEIGHT

C.1 EVOLUTION AT A FIXED POSITION

Represent the distribution as the outcome X_1 at $t = 1$ of an ODE

$$\dot{X}_t = \tilde{v}(X_t, t), \quad X_0 \sim p_0, \tag{48}$$

where $\tilde{v}$ is either the original flow-matching velocity v or the marginal velocity under non-IID sampling, $v + r_\phi$. From the continuity equation

$$\partial_t \tilde{p}_t(x) = -\nabla_x \cdot \big(\tilde{v}(x,t)\tilde{p}_t(x)\big), \tag{49}$$

we obtain

$$\begin{aligned} &\partial_t \log \tilde{p}_t(x) \\ &= -\frac{1}{\tilde{p}_t(x)} \nabla_x \cdot \big(\tilde{v}(x,t)\tilde{p}_t(x)\big) \\ &= -\nabla_x \cdot \tilde{v}(x,t) - \tilde{v}(x,t) \cdot \frac{\nabla_x \tilde{p}_t(x)}{\tilde{p}_t(x)} \\ &= -\nabla_x \cdot \tilde{v}(x,t) - \tilde{v}(x,t) \cdot \nabla_x \log \tilde{p}_t(x) \\ &= -\nabla_x \cdot \tilde{v}(x,t) - \tilde{v}(x,t) \cdot \tilde{s}(x,t). \end{aligned} \tag{50}$$

Thus,

$$\partial_t \log p_t(x) = -\nabla_x \cdot v(x,t) - v(x,t) \cdot s(x,t), \tag{51}$$

and

$$\begin{aligned} \partial_t \log p''_{\phi,t}(x) = &-\nabla_x \cdot \big(v(x,t) + r_\phi(x,t)\big) \\ &- \big(v(x,t) + r_\phi(x,t)\big) \cdot s''_\phi(x,t), \end{aligned} \tag{52}$$

where p, s denote the original flow and p''_ϕ, s''_ϕ denote the marginal under diversity. Subtracting Eq. (52) from Eq. (51) gives

$$\begin{aligned} \partial_t \log w_t(x) &= \partial_t \log p_t(x) - \partial_t \log p''_{\phi,t}(x) \\ &= \nabla_x \cdot r_\phi(x,t) + s''_\phi(x,t) \cdot r_\phi(x,t) \\ &\quad + v(x,t) \cdot \big(s''_\phi(x,t) - s(x,t)\big). \end{aligned} \tag{53}$$

which matches Theorem 1.

C.2 EVOLUTION ALONG A TRAJECTORY

As in Definition 2, samples evolve according to

$$\dot{X}_t^{(i)} = v\big(X_t^{(i)}, t\big) + u\big(X_t^{(i)}, X_t^{(-i)}, t\big),\ X_0^{(i)} \overset{\text{IID}}{\sim} p_0, \tag{54}$$

where i indexes trajectories, $X_t^{(-i)}$ excludes $X_t^{(i)}$, v is the pre-trained flow, and u is a diversity velocity. For the importance weight along trajectory i,

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\log w_{\phi,t}(X_t^{(i)}) &= \nabla_x \log w_{\phi,t}(X_t^{(i)})\cdot \dot{X}_t^{(i)} \\ &\quad + \partial_t \log w_{\phi,t}(X_t^{(i)}),\end{aligned} \tag{55}$$

where

$$\begin{aligned}\nabla_x \log w_{\phi,t}(x) &= \nabla_x \log p_t(x) - \nabla_x \log p''_{\phi,t}(x) \\ &= s(x,t) - s''_\phi(x,t).\end{aligned} \tag{56}$$

Combining Eq. (55), Eq. (53), Eq. (56) yields

$$\begin{aligned}&\frac{\mathrm{d}}{\mathrm{d}t}\log w_{\phi,t}(X_t^{(i)}) \\ &= \big(s(X_t^{(i)},t) - s''_\phi(X_t^{(i)},t)\big)\cdot\big(v\big(X_t^{(i)},t\big) \\ &\quad + u\big(X_t^{(i)},X_t^{(-i)},t\big)\big) + \nabla_x\cdot r_\phi(X_t^{(i)},t) \\ &\quad + s''_\phi(X_t^{(i)},t)\cdot r_\phi(X_t^{(i)},t) \\ &\quad + v(X_t^{(i)},t)\cdot\big(s''_\phi(X_t^{(i)},t) - s(X_t^{(i)},t)\big) \\ &= \nabla_x\cdot r_\phi(X_t^{(i)},t) + s''_\phi(X_t^{(i)},t)\cdot r_\phi(X_t^{(i)},t) \\ &\quad - \big(s''_\phi(X_t^{(i)},t) - s(X_t^{(i)},t)\big)\cdot u\big(X_t^{(i)},X_t^{(-i)},t\big),\end{aligned} \tag{57}$$

which matches Theorem 2.

C.3 EVOLUTION ALONG A TRAJECTORY FOR RECTIFIED FLOWS

For rectified flows, Appendix B yields

$$s(x,t) = \frac{tv(x,t)-x}{1-t}, \tag{58}$$

and

$$s''_\phi(x,t) = \frac{t\big(v(x,t)+r_\phi(x,t)\big)-x}{1-t}. \tag{59}$$

Thus,

$$s''_\phi(x,t) - s(x,t) = \frac{t}{1-t}r_\phi(x,t). \tag{60}$$

Substituting Eq. (58), Eq. (59), and Eq. (60) into 57 gives

$$\begin{aligned}&\frac{\mathrm{d}}{\mathrm{d}t}\log w_{\phi,t}(X_t^{(i)}) \\ &= \nabla_x\cdot r_\phi(X_t^{(i)},t) \\ &\quad + \frac{t\big(v(X_t^{(i)},t)+r_\phi(X_t^{(i)},t)\big)-X_t^{(i)}}{1-t}\cdot r_\phi(X_t^{(i)},t) \\ &\quad - \frac{t}{1-t}r_\phi(X_t^{(i)},t)\cdot u\big(X_t^{(i)},X_t^{(-i)},t\big) \\ &= \nabla_x\cdot r_\phi(X_t^{(i)},t) + \frac{tv(X_t^{(i)},t)-X_t^{(i)}}{1-t}\cdot r_\phi(X_t^{(i)},t) \\ &+ \tfrac{t}{1-t}\big(r_\phi(X_t^{(i)},t) - u\big(X_t^{(i)},X_t^{(-i)},t\big)\big)\cdot r_\phi(X_t^{(i)},t).\end{aligned} \tag{61}$$

which matches Corollary 1.

C.4 EVOLUTION AT A FIXED POSITION FOR RECTIFIED FLOWS

Combining Eq. (51) and Eq. (52) with the rectified-flow identity in Appendix B, we obtain

$$\begin{aligned}
&\partial_t \log w_t(x) \\
&= \nabla_x \cdot r_\phi(x,t) + v(x,t) \cdot \frac{t}{1-t} r_\phi(x,t) \\
&\quad + \frac{t\big(v(x,t)+r_\phi(x,t)\big)-x}{1-t} \cdot r_\phi(x,t) \\
&= \nabla_x \cdot r_\phi(x,t) \\
&\quad + \frac{1}{1-t} r_\phi(x,t) \cdot \big(2tv(x,t) + tr_\phi(x,t) - x\big).
\end{aligned} \tag{62}$$

which matches Eq. (25).

D DIVERSITY OBJECTIVES USED IN EXPERIMENTS

To ensure generality, we include several diversity objectives in our experiments.

- **Determinantal Point Process (DPP)** [Kulesza et al., 2012]:
  $h\big(x^{(1:n)}\big) = \log\det\big(\frac{\exp(-K)}{\exp(-K)+I}\big)$,
- **Harmonic DPP**:
  $h\big(x^{(1:n)}\big) = \log\det\big(\frac{K_{\mathrm{H}}}{K_{\mathrm{H}}+I}\big)$,
- **Particle Guidance (PG)** [Corso et al., 2024]:
  $h\big(x^{(1:n)}\big) = -\sum_{i,j}\exp(-K_{ij})$,
- **Chebyshev**:
  $h\big(x^{(1:n)}\big) = -\sum_{r=1}^{R}\Big[1+\frac{2}{n}\sum_{i<j}T_r\big(\exp(-K_{ij})\big)\Big]^2$,
- **Log-barrier**:
  $h\big(x^{(1:n)}\big) = \sum_{i<j}\log\big(1-\exp(-K_{ij})\big)$,
- **Reciprocal**:
  $h\big(x^{(1:n)}\big) = -\sum_{i<j}\frac{1}{K_{ij}}$,

where K is defined in Eq. (11), T_r is the r-th Chebyshev polynomial [Mason and Handscomb, 2002], and K_{H} is defined as

$$K_{\mathrm{H}} = \frac{1}{\sum_{r=0}^{n-1} a_r}\sum_{r=0}^{n-1} a_r\, T_r(L_{\cos}), \tag{63}$$

with $a_0{=}1$, $a_r{=}2\big(1-\frac{r}{n}\big)$ when $r{\geq}1$, $L_{\cos,ij}{=}\langle\hat{x}^{(i)},\hat{x}^{(j)}\rangle$, and $\hat{x}^{(i)}{=}x^{(i)}/\|x^{(i)}\|$. Gradients are normalized in implementation, so constant scalings are omitted. For the strength parameter λ of the diversity velocity (see Eq. (14)), we set $\lambda = 1$ for Section 4.1, $\lambda = 0.1$ for Section 4.2, and $\lambda = 0.3$ for Section 4.3.

E EXPERIMENT IMPLEMENTATION DETAILS

We first summarize the shared protocol for the two image-generation tasks in Table 5. Within each task, all methods use a *matched* protocol: the same conditions, number of sampling steps, joint sample size n, number of repetitions, and IID reference size, with a common diversity strength λ for the diversity-based methods, so that any difference in coverage radius is attributable to the methods themselves rather than to differences in the evaluation setup. Classifier-Free Guidance (CFG) is disabled for all image methods to isolate the effect of SR, and for inpainting the coverage radius is computed only over the masked regions.

For the Gaussian mixture experiment (Section 4.1), we use 100 flow-matching steps with uniformly spaced time intervals and an Euler solver. For this task, we train the residual velocity for 100 epochs, sampling a random $t \in [0,1]$ for each

Task	Conditions	Steps	Sample size	Diversity strength	Repetitions	IID reference size
Text-to-image	6 text prompts	28	10	0.1	1,000/prompt	10,000/prompt
Inpainting	10 masked images	28	10	0.3	100/input	1,000/input

Table 5: Consolidated protocol for the image-generation experiments.

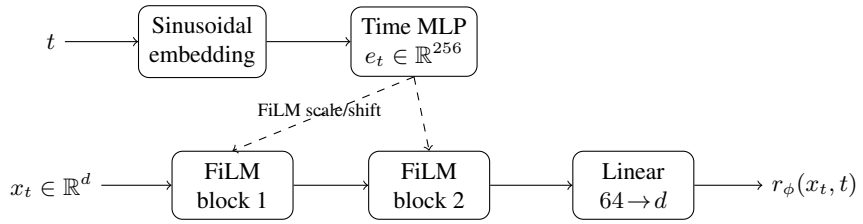

Figure 5: **Architecture of the residual velocity r_ϕ.** A sinusoidal time embedding feeds a time MLP that produces a shared feature $e_t \in \mathbb{R}^{256}$, which modulates both FiLM-conditioned blocks (dashed) through per-layer scale and shift. Each block contains two width-64 linear layers, each followed by FiLM and a SiLU activation; a final linear layer maps back to $\mathbb{R}^d$. Only the first input projection and the output layer depend on the data dimension d.

example. We use the AdamW [Loshchilov and Hutter, 2019] optimizer with a learning rate of 10^{-4}, weight decay of 10^{-3}, and a batch size of $1{,}000$.

For both text-to-image generation (Section 4.2) and image inpainting (Section 4.3), we use 28 uniformly spaced sampling steps with an Euler solver and disable CFG [Ho and Salimans, 2022] for all experiments, so that the measured differences isolate the effect of SR. The SR projection rule still applies with CFG enabled; one can use a proxy score direction read off from the CFG-guided velocity via Eq. (10), which we leave for future work. We compute diversity objectives using the spatial mean of latent features. For inpainting, we further adopt a fixed text prompt, "a realistic photo," across all input images.

F PARAMETERIZATION AND ARCHITECTURE OF RESIDUAL VELOCITY

The marginal ODE (Definition 3) requires the marginal velocity v' of the non-IID, jointly drawn samples. Rather than learn v' from scratch, we parameterize it as $v' = v + r_\phi$, reusing the pre-trained velocity v and learning the correction r_ϕ. When the diversity-velocity magnitude is small, the joint trajectories (Definition 2) interact weakly, so v' stays close to v and the correction $r_\phi = v' - v$ is small; learning this small residual rather than a fresh velocity improves training efficiency and potentially enables a smaller network than v.

The residual velocity r_ϕ is a lightweight time-conditioned FiLM MLP with about 0.24M parameters, used in the Gaussian-mixture importance-weight experiments (Section 4.1; training details in Appendix E). We detail its architecture and parameter count here for reproducibility.

Architecture. A sinusoidal embedding of t is passed through a time MLP to a shared feature $e_t \in \mathbb{R}^{256}$, which conditions two FiLM blocks [Perez et al., 2018], each with two width-64 linear layers, followed by a linear layer back to $\mathbb{R}^d$. For each layer, e_t produces a FiLM scale s and shift b that modulate the activation as $h \leftarrow h \odot (1 + s_{\max} \tanh(s)) + b$ with $s_{\max} = 1$, after which a SiLU nonlinearity is applied; the scale and shift projections are zero-initialized, so the conditioning starts from the identity. Figure 5 summarizes the architecture.

Parameter count. Table 6 lists the per-module parameter count. Only the input projection and the output layer depend on the data dimension d, so the total, $242{,}944 + 129\,d$, grows linearly in d and stays close to 0.24M across the dimensions we use.

Table 6: Trainable parameter count of the residual velocity r_ϕ given the data dimension d.

Module	Parameter count
Time MLP	98,816
Block 1 with FiLM	$64d + 70{,}016$
Block 2 with FiLM	74,112
Output layer	$65d$
Total	$\mathbf{242{,}944 + 129\,d}$

Table 7: **Perceptual cross-check of the text-to-image coverage radius (lower is better).** The "Latent" column is the model-free latent coverage radius (Mean column of Table 3); the "LPIPS-Alex" column recomputes it (Eq. (26)) on the same samples under an AlexNet LPIPS distance. Values are mean(uncertainty) with 95% confidence intervals. SR (soft and hard) is our method.

Method	SR	Latent↓	LPIPS-Alex↓
IID	-	20.9(2)	3.877(36)
DPP	✗	17.5(2)	3.829(36)
	soft	17.4(2)	3.829(36)
	hard	17.2(2)	3.829(36)
PG	✗	16.9(2)	3.873(37)
	soft	16.8(2)	3.872(37)
	hard	16.6(2)	3.870(37)
DF	✗	17.4(2)	3.832(36)
	soft	17.3(2)	3.831(36)
	hard	17.2(2)	3.829(36)

G ADDITIONAL EXPERIMENTS AND ANALYSES

Perceptual cross-check with LPIPS. To check that our latent coverage metric does not understate perceptual diversity, we recompute it under a perceptual LPIPS distance [Zhang et al., 2018] (LPIPS-Alex; Table 7); the two broadly agree at the IID-versus-diverse level.

Score-based regularization vs. scalar annealing. SR changes the *direction* of the diversity velocity, not its magnitude. Could rescaling the magnitude schedule $\beta(t)$ (the $\sqrt{1-t}$ time factor of γ in Eq. (14); default $\beta(t) = \sqrt{1-t}$) match its performance instead? On the 8-D Gaussian setup (Table 1), no swept $\beta(t)$ shape reaches the DPP+SR-hard reference (Table 8). This is a rough comparison, as the sweep (500 repetitions) and the reference (10,000 trials) are separate runs; still, since rescaling alone cannot match SR, its benefit appears to come from the directional, score-aware correction, not magnitude tuning.

Sensitivity to the diversity strength λ. The strength λ (Eq. (14)) governs the diversity–quality trade-off: too small gives little diversity, too large pushes samples off-manifold and lowers quality. We sweep it for DPP+SR-hard on Stable Diffusion 3.5 Medium [Esser et al., 2024] text-to-image over T_1–T_6, reporting the LPIPS-Alex radius against an IID reference (Table 9; a separate 30-repetition run, so its values are not comparable to Table 3). The radius traces a broad U-shape: it stays at or below IID for λ up to about 0.15 (minimum near 0.075) and degrades for large λ (clearly by 0.3), so moderate λ is robust and the default $\lambda = 0.1$ is a safe choice.

Choice of baselines and related work. We compare against DPP, Particle Guidance (PG) [Corso et al., 2024], and DiverseFlow (DF) [Morshed and Boddeti, 2025], each evaluated with and without our SR, under the same matched, fixed-budget protocol (Tables 1 and 3). A related line is training-free guidance such as FreeDoM [Yu et al., 2023], which steers a frozen pre-trained diffusion model toward a target condition; it guides a single sample, whereas we couple n samples for mutual diversity and recover importance weights. We do not treat methods that construct or train the generative flow itself as baselines, since they are not focused on multi-sample diversity.

Diversity velocity / schedule	Mode cov. ↑	$\log p$ ↑	RMSE ↓
DPP, no SR, $\beta = \sqrt{1-t}$ (default)	8.53	7.077	0.529
DPP, no SR, $\beta =$ cosine	8.48	11.349	0.443
DPP, no SR, $\beta = \max(1-t, 0)$ (linear)	8.44	15.925	0.325
DPP, no SR, $\beta =$ sigmoidal	8.33	17.149	0.286
DPP+SR-hard (ours)	8.57	19.270	0.253

Table 8: **Score-based regularization vs. scalar annealing schedules.** The first four rows disable SR and vary only the scalar magnitude schedule $\beta(t)$; the last row is the DPP+SR-hard reference from Table 1.

DPP+SR-hard (ours)	LPIPS-Alex ↓	Δ vs. IID	Beats IID?
IID (reference)	3.902	-	-
$\lambda = 0.01$	3.880	-0.022	✓
$\lambda = 0.025$	3.849	-0.053	✓
$\lambda = 0.05$	3.826	-0.076	✓
$\lambda = 0.075$	**3.816**	$\mathbf{-0.086}$	✓
$\lambda = 0.1$ (default)	3.854	-0.048	✓
$\lambda = 0.15$	3.851	-0.051	✓
$\lambda = 0.2$	3.923	$+0.021$	✗
$\lambda = 0.3$	4.218	$+0.316$	✗

Table 9: **Sensitivity to the diversity strength** λ for DPP+SR-hard on Stable Diffusion 3.5 Medium text-to-image, averaged over T_1–T_6 (LPIPS-Alex radius, lower is better; 30 repetitions, B=10). Δ is the gap to the IID reference and "Beats IID?" its sign at point estimates. The empirical minimum ($\lambda = 0.075$) is in **bold**; the default $\lambda = 0.1$ is marked.

H CONSTRUCTION OF GROUND-TRUTH DENSITY

We obtain ground-truth densities for the non-IID marginal by repeatedly resampling from the non-IID dataset and pooling the draws into a single large set. Given a sample set $\mathcal{Y} = \{y_i\}_{i=1}^{B}$, $y_i \in \mathbb{R}^d$, we evaluate densities at each sample $x \in \mathcal{Y}$ using a leave-one-out strategy—excluding x from its own neighborhood—so that n=$B-1$ samples contribute to the estimate. Letting I_d denote the $d\times d$ identity, we follow prior work [Loader, 1996, Hjort and Jones, 1996] and estimate $\log p'(x)$ by fitting a local quadratic model to $\log p'$ using kernel-weighted moments computed across multiple kNN-defined bandwidths (scales h). We then combine the resulting per-scale log-density estimates via inverse-variance weighting.

kNN neighborhoods and bandwidth selection. We use the neighbor-count grid

$$\mathcal{K} = \{10, 15, 25, 35, 50\} \cap \{1, \dots, n\}. \tag{64}$$

For each $k \in \mathcal{K}$, we sort all samples by their distance to x, after removing x itself. Let $r_k(x)$ be the distance to the k-th nearest neighbor, and define the balloon bandwidth

$$h = \max\{r_k(x),\, h_{\min}\}, \quad h_{\min} = 10^{-12}. \tag{65}$$

To stabilize the moment estimates, we include more than k neighbors:

$$K_{\text{use}}(k) = \min\big\{\, n,\ \max(3k,\ k{+}8)\,\big\}. \tag{66}$$

Let $\mathcal{N}_x^{(k)}$ be the indices of the $K_{\text{use}}(k)$ nearest neighbors (with self-exclusion applied).

Kernel-weighted moments at scale h. Let the Gaussian *base* kernel be $K(u) = (2\pi)^{-d/2}\exp(-\|u\|_2^2/2)$ and define the scaled kernel $K_h(u) = h^{-d}K(u/h)$, so that $K_h(u) = (2\pi)^{-d/2}h^{-d}\exp(-\|u\|_2^2/(2h^2))$. For offsets $u_i = y_i - x$ with $i \in \mathcal{N}_x^{(k)}$, set $w_i = K_h(u_i)$, and set w_i=0 for $i \notin \mathcal{N}_x^{(k)}$. Define the weighted moments (sums taken over i=$1, \dots, n$)

$$m_0 = \sum_i w_i,\ \ m_1 = \sum_i w_i\, u_i,\ \ M_2 = \sum_i w_i\, u_i u_i^\top, \tag{67}$$

and the empirical mean and covariance

$$\mu = \frac{m_1}{m_0}, \quad S = \frac{M_2}{m_0} - \mu\mu^\top. \tag{68}$$

Local log-density at scale h. Under the quadratic approximation $\log p'(x+u) \approx a + b^\top u + \frac{1}{2} u^\top C u$, profiling the local likelihood yields the closed-form estimate

$$\begin{aligned} \widehat{a}_h(x) = \log\Big(\frac{m_0}{n}\Big) + d \log h \\ - \tfrac{1}{2} \operatorname{logdet} S - \tfrac{1}{2}\, \mu^\top S^{-1} \mu, \end{aligned} \tag{69}$$

and

$$\widehat{p}'_h(x) = \exp\{\widehat{a}_h(x)\}, \tag{70}$$

where $\widehat{a}_h(x)$ is the estimated local log-density at location x and scale h.

Uncertainty at a single scale. For the *unit-bandwidth* Gaussian kernel [Wand and Jones, 1995], the kernel roughness is

$$R(K) = \int K(u)^2 \, \mathrm{d}u = (4\pi)^{-d/2}. \tag{71}$$

Using Kish's effective sample size [Kish, 1965],

$$n_{\text{eff}}(x, h) = \frac{(\sum_i w_i)^2}{\sum_i w_i^2}, \tag{72}$$

a Wald approximation [Wald, 1949] leads to

$$\operatorname{Var}(\widehat{a}_h(x)) \approx \frac{R(K)}{n_{\text{eff}}(x, h)\, h^d\, \widehat{p}'_h(x)}, \tag{73}$$

and

$$\begin{aligned} I_h(x) &= \operatorname{Var}(\widehat{a}_h(x))^{-1} \\ &= \frac{n_{\text{eff}}(x, h)\, h^d\, \widehat{p}'_h(x)}{R(K)}, \end{aligned} \tag{74}$$

where $I_h(x)$ is the per-scale information (inverse variance) associated with $\widehat{a}_h(x)$.

Multi-scale aggregation. We apply numerical checks at each scale h and exclude any scale whose kernel weights are numerically unreliable. Covariance estimates are stabilized by symmetrization and a small ridge to enforce positive definiteness prior to inversion.

Given the valid per-scale estimates $\widehat{a}_h(x)$ and their information weights $I_h(x)$, we aggregate via inverse-variance weighting:

$$\widehat{a}(x) = \frac{\sum_h I_h(x)\, \widehat{a}_h(x)}{\sum_h I_h(x)}, \quad \widehat{p}'(x) = \exp\{\widehat{a}(x)\}. \tag{75}$$

Only scales that pass the numerical checks contribute to the final estimate.

I BASELINE DENSITY ESTIMATORS

kNN density [Loftsgaarden and Quesenberry, 1965]. This method estimates $p'(x)$ by the k-NN ball volume. Let $r_k(x)$ be the distance to the k-th (k=5) nearest *other* sample in $X = \{x^{(j)}\}_{j=1}^B \subset \mathbb{R}^d$, and let $V_d = \pi^{d/2}/\Gamma(\frac{d}{2}+1)$ be the unit d-ball volume:

$$\widehat{p}_{\text{kNN}}(x) = \frac{k}{B\, V_d \left(r_k(x)\right)^d}. \tag{76}$$

Ours	Ours†	kNN	KDE	MGF	Equal	IID
0.071(1)	0.078(2)	0.052(1)	0.044(1)	0.051(1)	0.088(1)	0.077(1)

Table 10: **Jensen–Shannon divergence for expectation estimation** (lower is better). Values are mean(uncertainty) with 95% confidence intervals. † is our fixed-position variant; "Equal" weights the non-IID samples equally and "IID" uses additional IID samples. The density-ratio estimators (kNN, KDE, MGF) are oracle-fed the analytic target density p_1 (see text).

Gaussian KDE with Silverman's rule [Silverman, 1986]. With isotropic bandwidth $h > 0$,

$$\widehat{p}_{\text{KDE}}(x) = \frac{1}{B}\sum_{j=1}^{B} \frac{\exp\Big(-\frac{\|x-x^{(j)}\|_2^2}{2h^2}\Big)}{(2\pi)^{d/2}h^d}, \tag{77}$$

$$h = \sigma\left(\frac{4}{d+2}\right)^{\frac{1}{d+4}} B^{-\frac{1}{d+4}}, \tag{78}$$

$$\sigma = \tfrac{1}{d}\sum_{\ell=1}^{d} \text{std}(X_{:\ell}). \tag{79}$$

Multivariate Gaussian fit (MGF) [Bishop and Nasrabadi, 2007]. This method fits $N(\mu, \Sigma)$ by maximum likelihood on all B samples:

$$\begin{aligned} \mu_{\text{ML}} &= \tfrac{1}{B}\sum_{j=1}^{B} x^{(j)}, \\ \Sigma_{\text{ML}} &= \tfrac{1}{B}\sum_{j=1}^{B} (x^{(j)}-\mu_{\text{ML}})(x^{(j)}-\mu_{\text{ML}})^\top. \end{aligned} \tag{80}$$

Evaluate the log-density at each $x^{(i)}$ under the same fitted model:

$$\begin{aligned} &\log \widehat{p}_{\text{MGF}}(x^{(i)}) \\ &= -\tfrac{1}{2}\Big(d\log(2\pi) + \log|\Sigma_{\text{ML}}|\Big) \\ &\quad - \tfrac{1}{2}\big(x^{(i)}-\mu_{\text{ML}}\big)^\top \Sigma_{\text{ML}}^{-1}\big(x^{(i)}-\mu_{\text{ML}}\big). \end{aligned} \tag{81}$$

I.1 DOWNSTREAM EXPECTATION ESTIMATION

For the expectation $\mu{=}\mathbb{E}f(X)$ (Section 4.1), evaluated by Jensen–Shannon (JS) divergence (Table 10), our trajectory-based weighting debiases the equal-weighted non-IID average and improves over additional IID samples. The density baselines reach lower JS, but they are given the analytic target density p_1 as their weight numerator (an oracle, available in closed form only on this example and never used by our estimator); the JS metric then rewards recovering p_1, even though their per-weight error and ranking are weaker than ours (Table 2).

J METRICS FOR RANKING

Notation. Consider a query with n items. Let p_i denote the predicted score (the estimated importance weight in our experiment) and g_i the ground-truth grade (the ground-truth importance weight in our experiment).

Kendall's τ_b [Kendall, 1945]. For every unordered pair (i, j), define $s_p = \text{sign}(p_i - p_j)$ and $s_g = \text{sign}(g_i - g_j)$. Here, s_p and s_g summarize how the pair is ordered by the prediction and by the ground truth, respectively ($+1$ if the first item is larger, -1 if smaller, 0 if tied). Kendall's τ_b measures agreement between two orderings while accounting for ties. Count concordant

Figure 6: **Illustrative cases where SR changes the sample significantly.** Two selected text-to-image cases (base sampler: DPP) that illustrate SR's effect. Each row shows, at a matched seed, the IID reference, the base sampler without SR, and with SR-hard. In these cases SR recovers more plausible content while retaining the diversity the base sampler introduced.

pairs C ($s_p = s_g \in \{\pm 1\}$), discordant pairs D ($s_p = -s_g \in \{\pm 1\}$), and one-sided ties $T_p = \#\{(i,j) : s_p = 0,\ s_g \neq 0\}$, $T_g = \#\{(i,j) : s_g = 0,\ s_p \neq 0\}$. Pairs tied in both lists ($s_p = s_g = 0$) are ignored. Then

$$\tau_b = \frac{C - D}{\sqrt{(C + D + T_p)\,(C + D + T_g)}}, \tag{82}$$

with 0 as the result if the denominator is 0. Values range from -1 (complete disagreement) to $+1$ (perfect agreement), with 0 indicating no correlation.

Spearman's ρ [Spearman, 1987]. Spearman's ρ assesses the strength of a monotonic relationship between two rankings. Assign 1-based ranks (1 = best) with average-tied ranks and compute the Pearson correlation between predicted and ground-truth ranks:

$$\rho \;=\; \mathrm{corr}\big(\mathbf{r}_{\text{pred}},\, \mathbf{r}_{\text{gt}}\big).$$

Here, $\rho = 1$ indicates identical rankings up to a monotonic transformation, $\rho = -1$ indicates perfect inverse order, and $\rho \approx 0$ indicates no monotonic association.

Graded Average Precision (gAP) [Moffat et al., 2022]. Graded AP extends Average Precision to graded (multi-level) relevance. First, sort the data by predicted scores. Normalize per query to $[0, 1]$: shift if any grade is negative; then, if the maximum exceeds 1, divide all grades by that maximum. Define

$$\mathrm{Prec@}i = \frac{\sum_{j=1}^{i} g_j}{i}, \quad \mathrm{gAP} = \frac{\sum_{i=1}^{n} g_i \,\mathrm{Prec@}i}{\sum_{i=1}^{n} g_i}. \tag{83}$$

When $g_i \in \{0, 1\}$, gAP reduces to standard AP. Larger values indicate that higher-graded items are concentrated near the top of the ranking.

K ADDITIONAL QUALITATIVE SAMPLES

Figure 6 shows selected text-to-image cases where score-based regularization (SR) changes the sample significantly.

L ILLUSTRATIVE EXAMPLE OF JOINT ODE TRAJECTORIES

Figure 7 visualizes joint ODE trajectories on a 2D Gaussian mixture with three components of mixture weights $1/6$, $1/3$, and $1/2$. Under the IID baseline (no diversity objective), trajectories collapse to the highest-weight component, as

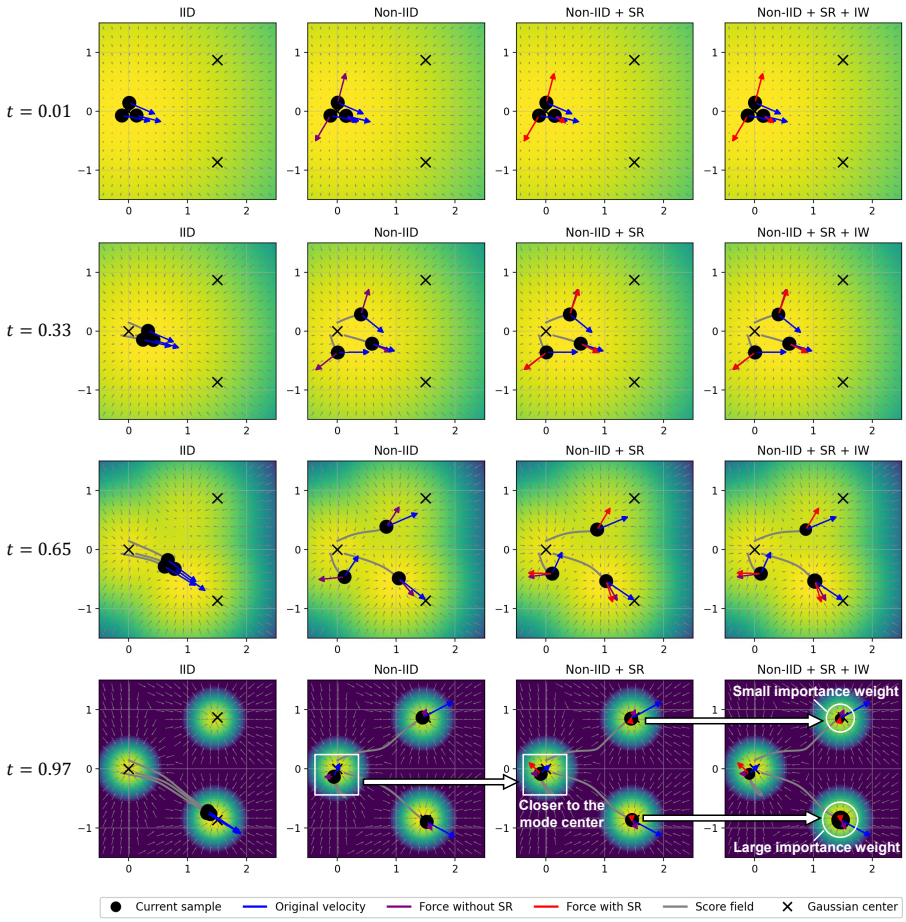

Figure 7: **Joint ODE trajectories on a three-component Gaussian mixture.** All methods start from identical initial states. "IID" is the baseline without a diversity objective and collapses to the highest-weight mode. "Non-IID (DPP)" uses a determinantal point process objective to encourage sample diversity, enabling the three trajectories to discover all three modes. "SR" denotes our score-based diversity-velocity regularization (the soft version is used in this example). Combining DPP+SR pulls samples toward the underlying modes while preserving coverage (white rectangle). "IW" denotes importance-weight estimation. The full method (DPP+SR+IW) further assigns an importance weight to each trajectory; larger markers indicate higher estimated weight (white circles). For visualization, arrow lengths are rescaled nonlinearly while preserving their relative ordering. The background shows the relative value of $\log p(x)$, where p is the target density corresponding to the IID ODE; yellow indicates higher density and purple indicates lower density.

expected. Introducing a DPP diversity objective yields *Non-IID* trajectories that collectively cover all three modes, improving coverage under a fixed trajectory budget. Augmenting DPP with our score-based diversity-velocity regularization (SR) draws trajectories closer to the mode centers while preserving coverage (see the white rectangle). Finally, importance weighting (IW) assigns a weight to each trajectory. In the full method (DPP+SR+IW), the sample closest to the highest-weight component receives the largest weight (white circle) and that closest to the lowest-weight component receives the smallest, correcting the Non-IID sampling bias and yielding an unbiased estimator with respect to the IID target.