Signed Rectified Flow: Negativity-Controlled Generation

Runlong Liao* **Baiyu Su*** **Lizhang Chen** **Qiang Liu**

The University of Texas at Austin

https://rectifiedflow.github.io/Signed-RF/

Abstract

We introduce *Signed Rectified Flow (Signed RF)*, a generalization of Rectified Flow that targets a signed measure $\pi^{\texttt{sign}} = (1+\alpha)\,\pi^{+} - \alpha\,\pi^{-}$, where $\alpha > 0$, π^{+} represents the distribution to promote, and π^{-} represents the distribution to suppress. Although sampling from a signed measure is not well-defined, Signed RF induces a valid generative process that concentrates on the positive region of $\pi^{\texttt{sign}}$ while provably excluding regions dominated by the negative component. This yields a principled framework for incorporating negative information and exclusion constraints into generative modeling. Theoretically, we analyze the signed continuity equation underlying Signed RF and explain how negative mass creates exclusion barriers through a charged-particle interpretation. Empirically, Signed RF leads to practical adaptive guidance algorithms. Across applications, Signed RF improves the fidelity–diversity trade-off on ImageNet, reduces nearest-neighbor similarity in anti-memorization tests, and reduces adversarial-prompt nudity in SD 3.5 while preserving CLIP and aesthetic scores.



(a) π^{-}: Red mode (b) π^{-}: All modes, π^{+}: Blue mode (c) π^{-}: Data points

(d) π^{-}: *Mickey Mouse, Disney character* (e) π^{-}: *folded paper, origami*

(f) π^{-}: *low-quality, blurry* (g) π^{-}: *sad, depressed* (h) π^{-}: Training Data

Figure 1: Signed Rectified Flow steers samples toward the positive target π^{+} while avoiding regions where the negative target π^{-} dominates. Top row: 2D toy examples. (a) suppresses an undesired mode; (b) strengthens the desired target mode; and (c) avoids generating protected data. Bottom row: image-generation examples. (d) preserves cartoon attributes in π^{+} (round black ears, red shorts, yellow shoes) while avoiding the copyrighted character identity in π^{-}; (e) preserves a paper-crane scene in π^{+} while avoiding a folded-paper appearance in π^{-}; (f)–(g) suppress undesired visual attributes in π^{-}; and (h) mitigates data leakage by repelling samples from the training set.

*Co-first authors.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

1 Introduction

Many generative tasks can be viewed as *promoting* a desirable distribution π^+ while *suppressing* an undesirable one π^-. In conditional generation, this corresponds to producing samples that better match the target condition and avoiding inconsistent outputs [1–8]. In safety and alignment settings, the objective is to generate value-aligned outputs while excluding harmful ones [9–11]. In preference learning, given preferred and dispreferred samples, the goal is to steer generation toward the preferred distribution [12–16]. Across these settings, the central challenge is to design mechanisms that reliably amplify desirable modes while discouraging undesirable ones.

For flow and diffusion models [17–23], the standard approach is to apply guidance [3–8] at sampling time, where trajectories are steered by extrapolating between positive and negative fields [3]. However, such procedures remain largely heuristic: their induced sampling distributions are poorly understood, making it unclear which regions of the data space are ultimately promoted or suppressed.

We propose *signed measures* as a principled mechanism for incorporating negative information into generative modeling. Given a distribution π^+ to favor and π^- to avoid, we consider the signed target

$$\pi^{\texttt{sign}}(\boldsymbol{x}) = (1+\alpha)\,\pi^+(\boldsymbol{x}) - \alpha\,\pi^-(\boldsymbol{x}), \qquad \alpha > 0.$$

The goal is to use negativity to encode more than the mere absence of positive data: the signed target explicitly specifies what the model should avoid. Although $\pi^{\texttt{sign}}$ is not generally a probability distribution and cannot be sampled from in the usual sense, we do not seek to sample from it directly. Instead, we use it as an intermediate object for constructing a valid generative procedure that follows its positive part while avoiding regions dominated by negative mass.

To achieve this, we formally extend Rectified Flow [20, 21] to signed targets by extrapolating the standard RF velocity formula to $\pi^{\texttt{sign}}$. As illustrated in Fig. 2, simulating the resulting dynamics from the source distribution generates samples from a dynamically reachable subset of the positive region of the signed target. Regions in which $\pi^{\texttt{sign}}$ is negative are never sampled, while an additional unsampled *buffer region* separates the reachable and negative regions. Signed RF therefore rectifies the signed target into a valid sampling law by preserving the signed density on its reachable region and assigning zero density elsewhere.

The paper is organized as follows. In Sec. 2.1, we derive Signed RF by extending Rectified Flow from convex mixtures to signed targets. In Sec. 2.2, we characterize the induced sampling law through the notions of reachable and buffer regions, provide a charged-particle interpretation of the dynamics, and establish the main theoretical guarantees, including density preservation on the reachable region and nonpenetration into negative regions. In Sec. 2.3, we derive a practical guidance-form implementation whose strength is determined by a density ratio, estimated either with a learned classifier or through online ODE tracking. Finally, in Sec. 3, we demonstrate Signed RF across a range of applications, including conditional generation, safe generation, trajectory planning, concept suppression, and anti-memorization.

2 Method

2.1 From Rectified Flow to Signed Rectified Flow

Rectified Flow. Let π_1 be the target data distribution on $\mathbb{R}^d$, and let π_0 be a noise source, such as $\mathcal{N}(0, \boldsymbol{I})$. Rectified Flow (RF) [20–23] constructs a linear interpolation $\boldsymbol{X}_t = t\boldsymbol{X}_1 + (1-t)\boldsymbol{X}_0$, where $\boldsymbol{X}_0 \sim \pi_0$ and $\boldsymbol{X}_1 \sim \pi_1$ are independent. This interpolation induces the ODE

$$\frac{\mathrm{d}\boldsymbol{Z}_t}{\mathrm{d}t} = v_t^{\texttt{RF}}(\boldsymbol{Z}_t), \qquad \text{starting from } \boldsymbol{Z}_0 \sim \pi_0,$$

whose velocity field is given by

$$v_t^{\texttt{RF}}(\boldsymbol{x}) = \mathbb{E}\big[\boldsymbol{X}_1 - \boldsymbol{X}_0 \mid \boldsymbol{X}_t = \boldsymbol{x}\big]. \tag{1}$$

Let π_t denote the marginal distribution of $\boldsymbol{X}_t$. A key property of RF is marginal preservation: the ODE solution satisfies $\boldsymbol{Z}_t \sim \pi_t$ for all $t \in [0,1]$, and in particular $\boldsymbol{Z}_1 \sim \pi_1$. In practice, the velocity field $v_t^{\texttt{RF}}$ is approximated by a neural network v^θ trained with

$$\mathcal{L}(\theta) = \int_0^1 \mathbb{E}\Big[\|v^\theta(\boldsymbol{X}_t, t) - (\boldsymbol{X}_1 - \boldsymbol{X}_0)\|^2\Big]\,\mathrm{d}t.$$

Rectified Flow for Convex Mixtures. We next examine how Rectified Flow behaves under mixtures of target distributions. Let (π_t^+, v_t^+) and (π_t^-, v_t^-) denote the RF marginals and velocity fields induced, respectively, by $\boldsymbol{X}_t^+ = (1-t)\boldsymbol{X}_0 + t\boldsymbol{X}_1^+$ and $\boldsymbol{X}_t^- = (1-t)\boldsymbol{X}_0 + t\boldsymbol{X}_1^-$, where $\boldsymbol{X}_0 \sim \pi_0$, $\boldsymbol{X}_1^+ \sim \pi_1^+$, and $\boldsymbol{X}_1^- \sim \pi_1^-$. For the convex mixture $\pi_1^{\texttt{mix}} = (1-w)\pi_1^+ + w\pi_1^-$, with $w \in [0,1]$, linearity of the RF marginal and flux gives $\pi_t^{\texttt{mix}} = (1-w)\pi_t^+ + w\pi_t^-$. Applying Eq. 1 to the mixture gives the RF velocity field

$$v_t^{\texttt{mix}}(\boldsymbol{x}) = \frac{(1-w)\,\pi_t^+(\boldsymbol{x})\,v_t^+(\boldsymbol{x}) + w\,\pi_t^-(\boldsymbol{x})\,v_t^-(\boldsymbol{x})}{(1-w)\,\pi_t^+(\boldsymbol{x}) + w\,\pi_t^-(\boldsymbol{x})}. \tag{2}$$

By the marginal-preservation property of RF, integrating $\dot{\boldsymbol{Z}}_t = v_t^{\texttt{mix}}(\boldsymbol{Z}_t)$ from $\boldsymbol{Z}_0 \sim \pi_0$ yields $\boldsymbol{Z}_t \sim \pi_t^{\texttt{mix}}$ for all $t \in [0,1]$, and in particular $\boldsymbol{Z}_1 \sim \pi_1^{\texttt{mix}}$. For $w \in [0,1]$, $\pi_t^{\texttt{mix}}$ remains a nonnegative, normalized density for all t, and the associated dynamics therefore define a standard probability flow. A natural question, however, is what happens when w is allowed to be negative.

Rectified Flow for Signed Mixtures. We now extend the mixture coefficient beyond the convex regime by setting $w = -\alpha$, where $\alpha > 0$. This gives the signed target $\pi_1^{\texttt{sign}}(\boldsymbol{x}) := (1+\alpha)\pi_1^+(\boldsymbol{x}) - \alpha\pi_1^-(\boldsymbol{x})$, which has unit total mass but need not be nonnegative. By linearity of the RF marginals, the corresponding signed marginal is $\pi_t^{\texttt{sign}}(\boldsymbol{x}) = (1+\alpha)\pi_t^+(\boldsymbol{x}) - \alpha\pi_t^-(\boldsymbol{x})$ for $t \in [0,1]$. Substituting $w = -\alpha$ into Eq. 2 yields the Signed RF velocity field

$$v_t^{\texttt{signRF}}(\boldsymbol{x}) = \frac{(1+\alpha)\,\pi_t^+(\boldsymbol{x})\,v_t^+(\boldsymbol{x}) - \alpha\,\pi_t^-(\boldsymbol{x})\,v_t^-(\boldsymbol{x})}{(1+\alpha)\,\pi_t^+(\boldsymbol{x}) - \alpha\,\pi_t^-(\boldsymbol{x})}. \tag{3}$$

The denominator is precisely $\pi_t^{\texttt{sign}}(\boldsymbol{x})$, and hence $v_t^{\texttt{signRF}}$ is well defined away from the zero set $\Omega_t^0 := \{\boldsymbol{x} : \pi_t^{\texttt{sign}}(\boldsymbol{x}) = 0\}$. Although $\pi_t^{\texttt{sign}}$ may fail to be a probability density, its initial marginal remains the valid source distribution: $\pi_0^{\texttt{sign}} = \pi_0$. We therefore define Signed RF through the source-initialized ODE $\dot{\boldsymbol{Z}}_t = v_t^{\texttt{signRF}}(\boldsymbol{Z}_t)$, with $\boldsymbol{Z}_0 \sim \pi_0$, and denote the law of $\boldsymbol{Z}_t$ by $\pi_t^{\texttt{signRF}}$. As we establish below, almost every source-initialized trajectory remains in the positive region $\Omega_t^+ := \{\boldsymbol{x} : \pi_t^{\texttt{sign}}(\boldsymbol{x}) > 0\}$ and does not cross the singular zero set Ω_t^0. The singular boundary therefore acts as a repulsive barrier for the realized dynamics, allowing the ODE to remain well defined along sampled trajectories even when the signed marginal develops negative regions.

2.2 Sampling Behavior of Signed RF

By construction, $\pi_t^{\texttt{signRF}}$, as the law of the source-initialized ODE, is a valid probability distribution. In contrast, $\pi_t^{\texttt{sign}}$ has unit total mass but may take negative values. Therefore, once negative regions emerge, $\pi_t^{\texttt{signRF}}$ can no longer coincide globally with $\pi_t^{\texttt{sign}}$. This raises the central question: which part of the signed marginal is realized by the source-initialized flow? The one-dimensional example in Fig. 2 illustrates the resulting structure. The black curve shows the signed marginal $\pi_t^{\texttt{sign}}$, while the blue histogram shows the empirical density obtained by simulating the ODE driven by $v_t^{\texttt{signRF}}$ from $\boldsymbol{Z}_0 \sim \pi_0$. Their evolution reveals three key phenomena.

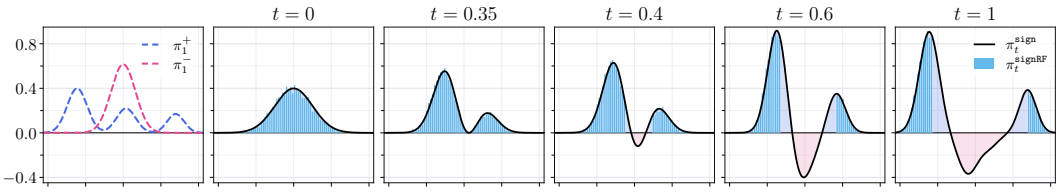


Figure 2: 1D evolution of the signed density $\pi_t^{\texttt{sign}}$ and the induced sampling law $\pi_t^{\texttt{signRF}}$. Signed RF rejects negative regions and matches the signed density on a reachable subset of the positive region.

- **Trajectories remain in the positive region.** Almost every source-initialized trajectory remains in the positive region $\Omega_t^+ := \{\boldsymbol{x} : \pi_t^{\texttt{sign}}(\boldsymbol{x}) > 0\}$ and never enters the negative region $\Omega_t^- := \{\boldsymbol{x} : \pi_t^{\texttt{sign}}(\boldsymbol{x}) < 0\}$. However, these trajectories generally occupy only a subset of Ω_t^+. We define $\Omega_t^r := \mathrm{supp}(\pi_t^{\texttt{signRF}})$ as the *reachable region*, shown in blue in Fig. 2.

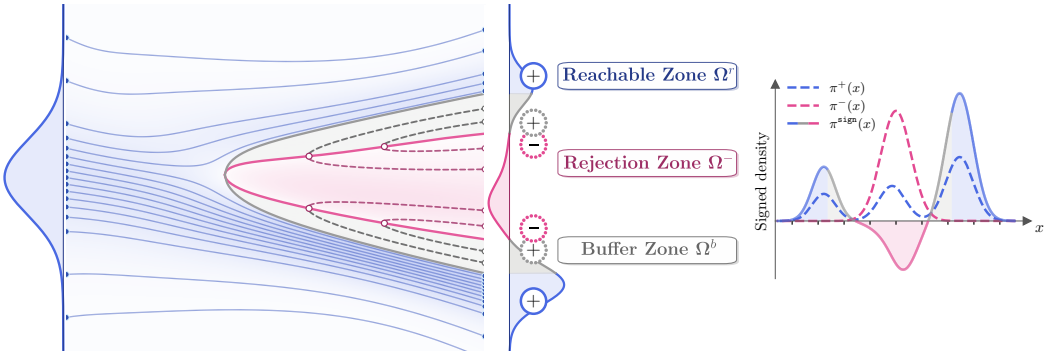


Figure 3: **Signed RF sample-region decomposition.** Left: Signed RF dynamics partition the space into reachable, negative (rejection), and buffer regions. Right: the terminal signed density $\pi_1^{\texttt{sign}}$. The induced sampling law is supported on the reachable region, where it coincides with the signed density, while the negative and buffer regions remain unobserved by the source-initialized sampler.

- **Signed RF rectifies the signed marginal.** On the reachable region Ω_t^r, the sampling density $\pi_t^{\texttt{signRF}}(\boldsymbol{x})$ coincides exactly with the signed marginal $\pi_t^{\texttt{sign}}(\boldsymbol{x})$. Outside this region, no mass is realized by the source-initialized flow. That is,

$$\pi_t^{\texttt{signRF}}(\boldsymbol{x}) = \pi_t^{\texttt{sign}}(\boldsymbol{x})\,\mathbf{1}\{\boldsymbol{x} \in \Omega_t^r\}. \tag{4}$$

- **The buffer region.** The remaining part of the positive region, $\Omega_t^b := \Omega_t^+ \setminus \Omega_t^r$, is not reached by trajectories initialized from π_0. We call it the *buffer region*. Since both $\pi_t^{\texttt{signRF}}$ and $\pi_t^{\texttt{sign}}$ have unit total mass, Eq. 4 implies

$$\int_{\Omega_t^r} \pi_t^{\texttt{sign}}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = 1 \qquad \Longrightarrow \qquad \int_{\Omega_t^b} \pi_t^{\texttt{sign}}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} + \int_{\Omega_t^-} \pi_t^{\texttt{sign}}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = 0.$$

  Thus, the positive mass contained in the buffer region exactly balances the magnitude of the negative mass excluded from the sampling law.

Remark. The identities above characterize the mass decomposition, but the location of the reachable region Ω_t^r is determined implicitly by the flow dynamics. In general, it need not admit a closed-form representation; all that is known is that it lies inside the positive region and is separated from the negative region Ω_t^- by the buffer region. Thus, Signed RF can be viewed as "rectifying" the signed marginal $\pi_t^{\texttt{sign}}$ into a valid probability law $\pi_t^{\texttt{signRF}}$: it preserves the signed density on a reachable subset of the positive region and sets it to zero elsewhere. Equivalently, $\pi_t^{\texttt{signRF}}$ is one of the total-variation-optimal nonnegative approximations to $\pi_t^{\texttt{sign}}$, solving $\min_{\rho \in \mathcal{P}} \mathrm{TV}(\rho, \pi_t^{\texttt{sign}})$ over the set of valid probability distributions $\mathcal{P}$. See Proposition B.1. The corresponding region decomposition is summarized in Fig. 3.

The physical picture. How does the buffer region arise? A useful intuition is to interpret the Signed RF dynamics as the motion of charged particles. Starting from the terminal signed target $\pi_1^{\texttt{sign}}$, trace the dynamics backward toward $t = 0$. Imagine placing a particle at each $\boldsymbol{x}$ at $t = 1$, with charge determined by the sign of $\pi_1^{\texttt{sign}}(\boldsymbol{x})$, and evolving it backward according to $\dot{\boldsymbol{Z}}_t = v_t^{\texttt{signRF}}(\boldsymbol{Z}_t)$. Its fate depends on the region from which it originates.

- **Reachable particles.** Positive particles originating from the reachable region Ω_1^r admit well-defined backward trajectories all the way to $t = 0$. These trajectories remain in the positive region and, collectively, recover the source distribution π_0 at $t = 0$. Reversing them in time yields precisely the trajectories obtained by simulating Signed RF forward from $\boldsymbol{Z}_0 \sim \pi_0$.
- **Buffer and negative particles.** Positive particles originating from the buffer region Ω_1^b, along with negative particles originating from Ω_1^-, do not admit backward trajectories that reach $t = 0$.

Instead, they encounter the moving zero set $\Omega_t^0 \coloneqq \{\boldsymbol{x} : \pi_t^{\texttt{sign}}(\boldsymbol{x}) = 0\}$, where the velocity field becomes singular. In the particle picture, positive and negative particles meet and annihilate at this boundary, reflecting the equal and opposite signed masses of the buffer and negative regions. Equivalently, viewed forward in time, the singular boundary can be interpreted as creating positive–negative particle pairs. The positive branch forms the buffer region, while the negative branch forms Ω_1^-. These "dark" particles are invisible to the source-initialized sampler because they emerge from the singular boundary at intermediate times rather than being transported from π_0.

Continuity equation. Why do $\pi_t^{\texttt{sign}}$ and $\pi_t^{\texttt{signRF}}$ coincide on the reachable region, as in Eq. 4? The key is that the signed marginal and the source-initialized sampling law obey the same continuity equation on this region. Following standard RF theory, each branch satisfies $\partial_t \pi_t^{\pm} + \nabla \cdot (\pi_t^{\pm} v_t^{\pm}) = 0$, which ensures marginal preservation for each branch. Taking the corresponding signed combination yields $\partial_t \pi_t^{\texttt{sign}} + \nabla \cdot \big((1+\alpha)\pi_t^+ v_t^+ - \alpha \pi_t^- v_t^-\big) = 0$. By the definition of $v_t^{\texttt{signRF}}$ in Eq. 3, the signed marginal therefore satisfies

$$\partial_t \pi_t^{\texttt{sign}} + \nabla \cdot \Big(\pi_t^{\texttt{sign}} v_t^{\texttt{signRF}}\Big) = 0.$$

Thus, $\pi_t^{\texttt{sign}}$ is a signed solution of the continuity equation driven by $v_t^{\texttt{signRF}}$. However, this alone does not make it the law of $\boldsymbol{Z}_t$, since $\pi_t^{\texttt{sign}}$ may take negative values.

Now consider the restriction of the signed marginal to the reachable region, $\bar{\pi}_t(\boldsymbol{x}) \coloneqq \pi_t^{\texttt{sign}}(\boldsymbol{x}) \mathbf{1}\{\boldsymbol{x} \in \Omega_t^r\}$. The region Ω_t^r is transported by the same flow, so its moving boundary carries no additional flux. Consequently, $\bar{\pi}_t$ satisfies the same continuity equation and remains a valid probability density. Under Assumption B.2 in Appendix B, it must therefore coincide with the law of the source-initialized ODE: $\pi_t^{\texttt{signRF}}(\boldsymbol{x}) = \pi_t^{\texttt{sign}}(\boldsymbol{x}) \mathbf{1}\{\boldsymbol{x} \in \Omega_t^r\}$. The appendix also establishes that almost every source-initialized trajectory remains in the positive region and never crosses into Ω_t^-.

2.3 Signed Rectified Flow in Practice

For practical implementation, it is useful to rewrite Eq. 3 in a guidance-like form. Define the density ratio $r_t(\boldsymbol{x}) \coloneqq \pi_t^-(\boldsymbol{x})/\pi_t^+(\boldsymbol{x})$ and $\Delta v_t(\boldsymbol{x}) \coloneqq v_t^+(\boldsymbol{x}) - v_t^-(\boldsymbol{x})$. Then, wherever $\pi_t^{\texttt{sign}}(\boldsymbol{x}) \neq 0$,

$$v_t^{\texttt{signRF}}(\boldsymbol{x}) = v_t^+(\boldsymbol{x}) + \lambda_t^\alpha(\boldsymbol{x})\, \Delta v_t(\boldsymbol{x}), \qquad \lambda_t^\alpha(\boldsymbol{x}) = \frac{\alpha\, r_t(\boldsymbol{x})}{(1+\alpha) - \alpha\, r_t(\boldsymbol{x})}. \tag{5}$$

Thus, Signed RF takes v_t^+ as its base field and applies time- and state-dependent guidance in the direction $v_t^+ - v_t^-$. Implementing this velocity requires two ingredients: (i) the branch velocities v_t^+ and v_t^-, obtained either from a single model under different conditionings or from separately trained models; and (ii) the density ratio $r_t(\boldsymbol{x})$, for which we consider two estimators below.

Classifier-based. The ratio $r_t(\boldsymbol{x})$ can be estimated with a binary classifier trained on noisy states $\boldsymbol{x}_t$. Given a balanced dataset with $\boldsymbol{x}_t^+ \sim \pi_t^+$ and $\boldsymbol{x}_t^- \sim \pi_t^-$, both obtained from the standard RF interpolation, one can train $p_t^\phi(y \mid \boldsymbol{x}_t)$ with the binary cross-entropy loss

$$\mathcal{L}(\phi) = -\mathbb{E}_{\boldsymbol{x}_t^+ \sim \pi_t^+}\Big[\log p_t^\phi(y = + \mid \boldsymbol{x}_t^+)\Big] - \mathbb{E}_{\boldsymbol{x}_t^- \sim \pi_t^-}\Big[\log p_t^\phi(y = - \mid \boldsymbol{x}_t^-)\Big].$$

Under balanced sampling, the Bayes-optimal classifier satisfies $p_t^*(y = - \mid \boldsymbol{x})/p_t^*(y = + \mid \boldsymbol{x}) = \pi_t^-(\boldsymbol{x})/\pi_t^+(\boldsymbol{x}) = r_t(\boldsymbol{x})$. Thus, the density ratio in Eq. 5 can be obtained from the classifier odds.

Online ratio tracking. The classifier-based approach requires training an auxiliary model to estimate the density ratio. Alternatively, the ratio can be tracked directly along each Signed RF trajectory. Let $u_t \coloneqq \log r_t(\boldsymbol{z}_t)$, where $\dot{\boldsymbol{z}}_t = v_t^{\texttt{signRF}}(\boldsymbol{z}_t)$. The log-density ratio evolves according to

$$\dot{u}_t = \nabla \cdot \Delta v_t(\boldsymbol{z}_t) + \Delta v_t(\boldsymbol{z}_t)^\top s_t^-(\boldsymbol{z}_t) + \lambda_t^\alpha(\boldsymbol{z}_t)\, \Delta v_t(\boldsymbol{z}_t)^\top \big(s_t^-(\boldsymbol{z}_t) - s_t^+(\boldsymbol{z}_t)\big), \tag{6}$$

where $s_t^\pm = \nabla \log \pi_t^\pm$ are the branch score functions and $u_0 = 0$. For RF with a Gaussian source π_0, Tweedie's formula gives $s_t^\pm(\boldsymbol{x}) = (t\, v_t^\pm(\boldsymbol{x}) - \boldsymbol{x})/(1-t)$ [24]. The divergence term can be evaluated exactly when tractable or estimated using Hutchinson's trace estimator [25]. Jointly integrating Eq. 6 with the Signed RF ODE therefore tracks the density ratio along the trajectory as $r_t(\boldsymbol{z}_t) = \exp(u_t)$.

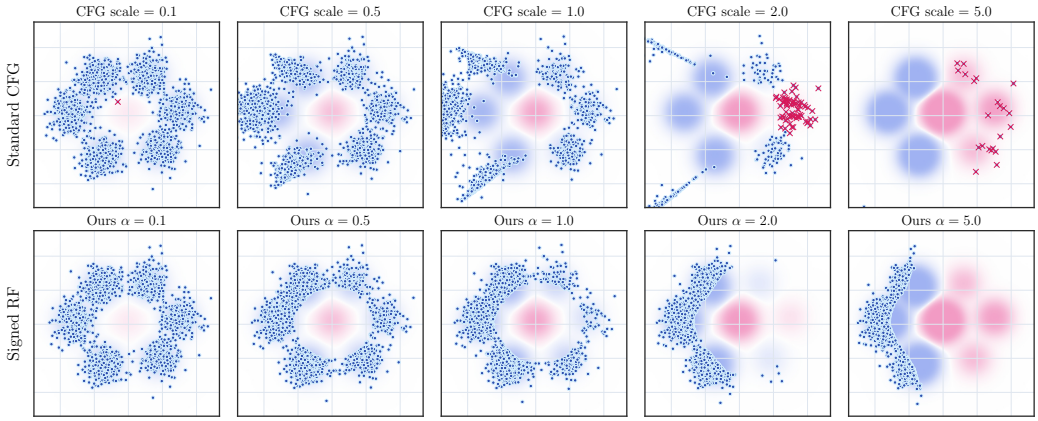


Figure 4: **2D Gaussian-mixture comparison.** Top: constant guidance with varying scales. Bottom: Signed RF with varying α. The background visualizes the signed target $\pi^{\texttt{sign}} = (1+\alpha)\pi^+ - \alpha\pi^-$, with blue and pink indicating positive and negative regions, respectively.

Stabilization. In numerical sampling, finite step sizes may cause the denominator of λ_t^α to become small, leading to excessively large guidance scales. A simple practical modification is to replace the denominator with $\max\big((1+\alpha) - \alpha\, r_t(\boldsymbol{z}_t),\, \varepsilon\big)$ and, optionally, cap λ_t^α at $\lambda_{\max}$.

3 Experiments

3.1 Toy Datasets

Gaussian mixture. We construct 2D Gaussian-mixture toys with positive target π^+ and negative target π^-. We train separate RF models for the two targets, yielding the velocity fields v_t^+ and v_t^-, and compute the density ratio $\pi_t^-(\boldsymbol{x})/\pi_t^+(\boldsymbol{x})$ in closed form. We compare standard constant guidance, $v_t(\boldsymbol{x}) = v_t^+(\boldsymbol{x}) + \omega\big(v_t^+(\boldsymbol{x}) - v_t^-(\boldsymbol{x})\big)$, with Signed RF, $v_t(\boldsymbol{x}) = v_t^+(\boldsymbol{x}) + \lambda_t^\alpha(\boldsymbol{x})\big(v_t^+(\boldsymbol{x}) - v_t^-(\boldsymbol{x})\big)$.

Fig. 4 visualizes the resulting sampling behavior. Constant guidance cannot reliably separate the positive and negative modes: weak guidance leaves samples near negative modes, whereas strong guidance over-repels the trajectories and pushes samples into low-density regions under π^+. In contrast, Signed RF avoids the negative modes while preserving the positive modes and their coverage. Additional Gaussian-mixture configurations in Sec. C.1 exhibit the same behavior.

PointMaze navigation. We use the PointMaze dataset [26, 27] to illustrate the *missing negative data* problem. The offline dataset contains only collision-free demonstrations, from which we train a start–goal-conditioned positive flow v^+. These demonstrations specify where valid trajectories should lie, but provide no explicit supervision for where trajectories must not go. In particular, wall interiors are absent from the dataset, yet their absence alone does not constitute a negative constraint. As a result, model approximation and generalization can still assign probability mass to these unobserved regions, producing wall-crossing trajectories despite training exclusively on valid demonstrations.

Signed RF provides a natural way to supply the missing negative information. We define wall-interior points as a negative distribution, train a negative flow v^- together with a density-ratio classifier, and use Signed RF to repel the planner from these explicitly modeled invalid regions. As shown in Fig. 5, this removes wall crossings while preserving broad coverage of feasible paths. In contrast, constant guidance exposes the usual scale trade-off: weak guidance fails to enforce the constraint, whereas strong guidance over-repels trajectories and reduces path diversity.

3.2 Signed RF as State-Aware CFG

We evaluate ImageNet 256×256 class-conditional generation using a pretrained Rectified Flow model [28, 29]. In this setting, $v_t^+(\boldsymbol{x}_t, c)$ and $v_t^-(\boldsymbol{x}_t)$ are the conditional and unconditional velocity

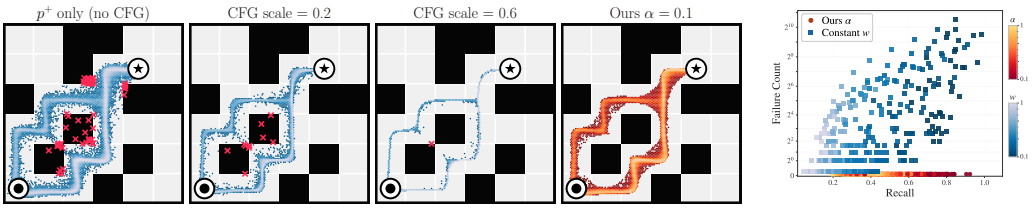


Figure 5: **The missing negative data.** A positive-only planner produces wall-crossing trajectories (a). Constant guidance trades safety for path diversity: weak guidance leaves violations (b), while strong guidance collapses the trajectories (c). Signed RF avoids obstacles while preserving diverse feasible paths (d). The Pareto plot compares the number of failures and path diversity (e).

NFE		FID ↓	IS ↑	Prec.↑	Rec.↑
4	CFG	10.3	224.2	**0.765**	0.311
	Ours	**6.68**	**229.2**	0.750	**0.474**
8	CFG	3.96	284.2	**0.839**	0.481
	Ours	**2.65**	**305.9**	0.805	**0.570**
16	CFG	2.38	274.9	**0.828**	0.566
	Ours	**1.82**	**296.8**	0.801	**0.615**
32	CFG	1.87	263.2	**0.815**	0.598
	Ours	**1.51**	**288.6**	0.800	**0.629**
64	CFG	1.73	272.4	**0.821**	0.601
	Ours	**1.41**	**278.2**	0.797	**0.627**

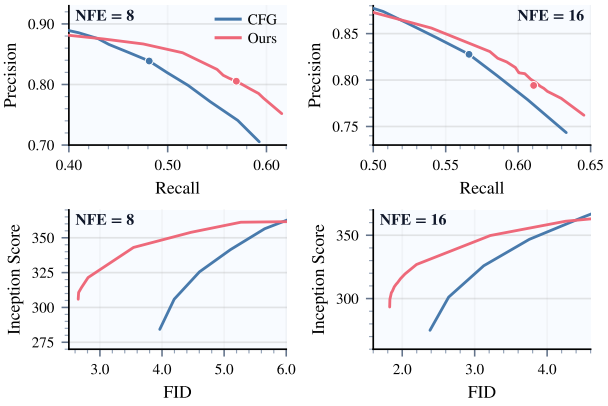


Figure 6: **ImageNet-256 comparison.** Left: best-FID operating points from the parameter sweeps. Right: Pareto frontiers; markers denote the corresponding best-FID points.

fields, respectively. We estimate the density ratio using a timestep- and class-conditioned ViT classifier trained on positive–negative latent pairs. We report FID [30], IS [31], and P/R [32] on 50K generated samples.

As shown in Fig. 6, Signed RF consistently improves FID across all NFE budgets. With 16 NFEs, for example, it reduces FID from 2.38 with CFG to 1.82. The precision–recall curves further show a better fidelity–diversity trade-off: while both methods improve precision at the expense of recall, Signed RF maintains higher recall at comparable precision.

We further verify that the density-ratio-based guidance is stable in practice. Performance varies little across classifier checkpoints and remains robust to moderate choices of the guidance cap $\lambda_{\max}$. The resulting $\lambda_t^{\alpha}(\boldsymbol{x}_t, c)$ is not a fixed guidance scale or time schedule: it varies across seeds within the same class and across classes for the same seed. The auxiliary classifier adds only modest overhead, increasing runtime and peak GPU memory by approximately 10%, while requiring roughly 10% of the original model's pretraining compute in our setup. Additional implementation details, ablations, and comparisons with other guidance methods are provided in Appendix C.2.

3.3 Anti-memorization

We next apply Signed RF to anti-memorization by explicitly modeling the training set as a negative distribution, with the goal of reducing training-data replication while preserving generation quality.

Analytic flow. To explicitly model memorization, we take the negative target to be the empirical training distribution $\pi_1^- = \frac{1}{N}\sum_{i=1}^N \delta_{\boldsymbol{x}^{(i)}}$, where $\mathcal{D} = \{\boldsymbol{x}^{(i)}\}_{i=1}^N$ denotes the training set. The RF transporting the Gaussian source $\pi_0 = \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ to π_1^- admits a closed-form marginal and velocity field:

$$\pi_t^-(\boldsymbol{x}) = \frac{1}{N}\sum_{i=1}^N \mathcal{N}\Big(\boldsymbol{x}; t\boldsymbol{x}^{(i)}, (1-t)^2\boldsymbol{I}\Big), \qquad v_t^-(\boldsymbol{x}) = \sum_{i=1}^N w_i(\boldsymbol{x}, t)\,\frac{\boldsymbol{x}^{(i)} - \boldsymbol{x}}{1-t}, \tag{7}$$

where $w_i(\boldsymbol{x}, t)$ are the normalized Gaussian kernel weights; see Sec. C.3 for details. This defines a training-free negative flow whose terminal distribution is exactly the empirical training distribution. As $t \to 1$, the Gaussian components in π_t^- collapse onto individual training examples, and trajectories under this flow terminate at training samples [20, 24]. Moreover, v_t^- contains the singular factor $1/(1-t)$, inducing strong attraction toward nearby training examples close to terminal time. Constant negative guidance is therefore poorly conditioned, as it directly inherits this terminal singularity.

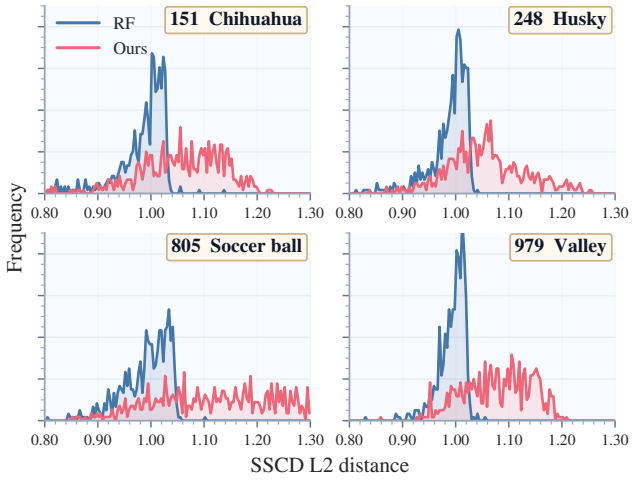
Data Repulsive Flow. The analytic flow in Eq. 7 provides a natural negative branch for repelling generated trajectories from the training set. For each class c, we define *Data Repulsive Flow* using the pretrained class-conditional flow as $v_t^+(\boldsymbol{x}, c)$ and the class-specific empirical analytic flow as $v_t^-(\boldsymbol{x}, c)$. We track the density ratio online using Eq. 6. Specifically, the log-ratio is initialized as $u_0 = 0$ and updated at each Euler step using v_t^+, v_t^-, the RF score function [24, 33], and a Hutchinson estimate of $\nabla \cdot (v_t^+ - v_t^-)$ [24, 25].

We evaluate memorization using a stress test. Since memorization events are rare under random sampling, aggregate statistics may be dominated by low-risk seeds and obscure copying behavior. For class c, we first generate 50K samples from the base model with $\omega = 1.0$, rank their initial seeds by the nearest-neighbor SSCD L_2 distance of the resulting samples to the training set [34], and retain the 600 highest-risk seeds. All methods are then evaluated using this same seed set for a controlled comparison. We additionally report standard generation metrics on 50K randomly sampled seeds.

Fig. 7 shows representative results on the high-risk seeds. The base model generates images highly similar to their nearest training examples, in some cases reproducing dataset-specific artifacts such as borders and watermarks. Data Repulsive Flow suppresses these artifacts, preserves class semantics, and produces samples that are visibly farther from their nearest training images. Quantitatively, the SSCD L_2 distributions in Fig. 8 shift toward larger distances. Compared with SPELL [35], Data Repulsive Flow achieves a better protection–quality trade-off. At $\alpha = 1.0$, it nearly matches the P_{05} protection level of SPELL-50 while retaining substantially better generation quality, with FID 2.03 versus 7.41. Increasing α to 7.5 further strengthens protection with only mild FID degradation and nearly unchanged IS. Additional per-class sweeps in Sec. C.3 show that protection improves smoothly with α; the same trend holds under the latent L_2 protection metric, which is not used to select the high-risk seeds.



Figure 7: **Memorization examples on ImageNet.** Each column uses the same initial seed, with the second and fourth rows showing the nearest training images. The base model closely reproduces training examples, including dataset-specific artifacts such as borders, watermarks, and backgrounds. In contrast, Data Repulsive Flow suppresses these artifacts while preserving class semantics, producing samples that are visibly less similar to their nearest training images.

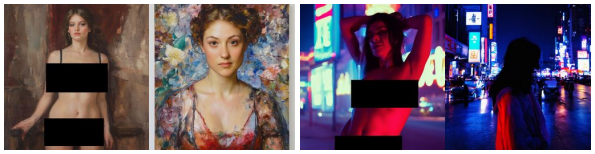


Method	FID ↓	IS ↑	Prec.↑	Rec.↑
Base	2.07	242.4	0.790	**0.627**
$\alpha=1.0$	**2.03**	246.0	0.796	0.624
$\alpha=3.0$	2.11	**246.2**	**0.799**	0.614
$\alpha=7.5$	2.47	245.9	0.797	0.600
SPELL-45	2.59	233.2	0.771	0.614
SPELL-50	7.41	189.7	0.680	0.563

Method	P_{05} ↑	P_{10} ↑	P_{25} ↑	Mean
Base	0.921	0.944	0.982	0.993
$\alpha=1.0$	0.939	0.962	0.992	1.024
$\alpha=3.0$	0.956	0.975	1.005	1.045
$\alpha=7.5$	**0.964**	**0.987**	**1.020**	**1.066**
SPELL-45	0.919	0.943	0.981	0.994
SPELL-50	0.940	0.963	1.003	1.049

Figure 8: **Quantitative comparison for anti-memorization on ImageNet.** Left: SSCD L_2 nearest-neighbor histograms for four classes on the same high-risk seeds. Upper right: generation quality under the standard 50K-sample evaluation. Lower right: memorization statistics for class 248, summarized by the 5%, 10%, and 25% quantiles and mean SSCD L_2 distance (higher is better).

Method	Ring-A-Bell		COCO-30K	
	ASR (%) ↓	TR ↓	CLIP ↑	AES ↑
SD 3.5-M	15.19	0.180	31.85	5.373
SAFREE	11.39	0.146	31.44	5.363
Safe Denoiser	12.66	0.157	31.44	5.373
Ours	**6.33**	**0.125**	31.57	5.371



(a) Baseline (b) Ours (c) Baseline (d) Ours

Figure 9: **Quantitative comparison for nudity prevention on SD 3.5-Medium.** Left: quantitative results on Ring-A-Bell and COCO-30K. Signed RF reduces unsafe generation while maintaining comparable image quality. Right: qualitative comparisons.

3.4 Concept Suppression

Nudity prevention. We next apply Signed RF to nudity prevention. Starting from a pretrained text-to-image model[2], we train two LoRA branches conditioned on the same unsafe prompt: a positive branch v^+ that models safe outputs and a negative branch v^- that models the corresponding unsafe outputs. The training data are constructed from paired prompts in ViSU-Text [37]. Although fine-tuning v^+ toward safe outputs reduces unsafe generation, residual unsafe modes inherited from the base model may remain. Signed RF explicitly models these modes through v^- and suppresses them during sampling.

We evaluate on Ring-A-Bell [38], a benchmark of adversarial prompts designed to elicit nude content. We compare with SAFREE [39] and Safe Denoiser [40], using the same backbone fine-tuned on safe outputs for all methods. We report attack success rate (ASR) and toxicity rate (TR), together with CLIP and aesthetic scores (AES) on COCO prompts to measure generation quality. As shown in Fig. 9, Signed RF reduces ASR from 15.19% to 6.33% and TR from 0.180 to 0.125, while maintaining comparable CLIP and AES scores.

Concept suppression. We further study Signed RF in qualitative concept-suppression settings. Given a pretrained text-to-image model [41], we obtain v^+ and v^- by conditioning the same model on positive and negative prompts, respectively, and track the density ratio online using Eq. 6.

Results are shown in Fig. 1 and Fig. 10. For IP protection, the positive prompt describes the desired visual attributes without naming the protected character, whereas the negative prompt explicitly specifies the identity to suppress. Signed RF removes identity-specific features while preserving the

[2] `stabilityai/stable-diffusion-3.5-medium` [36].

requested appearance, pose, and composition. For example, it suppresses the recognizable character identity in Fig. 1(d) while retaining the round black ears, red shorts, yellow shoes, and white gloves. Similarly, in Fig. 10(e), it removes Spider-Man-specific elements, such as the spider emblem, while preserving the pose, colors, and detailed suit structure. Signed RF also suppresses broader semantic attributes: it replaces a statue-like interpretation with a living angel in Fig. 10(b) and removes red tones while preserving the landscape structure in Fig. 10(d). Full prompts are provided in Sec. C.4.



(a) π^-: *trees, grass* (b) π^-: *statue, pedestal* (c) π^-: *white-male suit*

(d) π^-: *red, crimson* (e) π^-: *Spider-Man* (f) π^-: *old white male*

Figure 10: Qualitative comparison of concept suppression on Z-Image.

4 Conclusion

We introduced Signed Rectified Flow, an extension of Rectified Flow that incorporates explicit negative information through the signed target $(1 + \alpha)\pi^+ - \alpha\pi^-$. We showed that the source-initialized flow remains in a dynamically reachable subset of the positive region, where its sampling density matches the signed marginal, while the zero set acts as a barrier separating the realized trajectories from negative regions. This yields a state-dependent guidance rule whose strength is determined by the density ratio between the positive and negative distributions.

Signed RF can be implemented using either a learned density-ratio classifier or online ratio tracking. Across diverse settings, it resolves the missing negative data problem in trajectory planning, improves the fidelity–diversity trade-off in ImageNet class-conditional generation, reduces training-data replication, and suppresses unsafe or undesired concepts while preserving generation quality. Overall, Signed RF provides a principled and practical framework for incorporating explicit negative distributions into generative flows.

Acknowledgments and Disclosure of Funding

This work was supported in part by the Institute for Foundations of Machine Learning (IFML). The authors acknowledge the Texas Advanced Computing Center (TACC) at The University of Texas at Austin for providing computational resources that have contributed to the research results reported within this paper. The authors declare no competing interests.

References

[1] Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021.

[2] Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207*, 2021.

[3] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

[4] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

[5] 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*, pages 10684–10695, 2022.

[6] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

[7] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

[8] Runlong Liao, Jian Yu, Baiyu Su, Chi Zhang, Lizhang Chen, and Qiang Liu. Momentum guidance: Plug-and-play guidance for flow models. *arXiv preprint arXiv:2602.20360*, 2026.

[9] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

[10] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

[11] Patrick Schramowski, Manuel Brack, Björn Deiseroth, and Kristian Kersting. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22522–22531, 2023.

[12] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

[13] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

[14] Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. In *The Twelfth International Conference on Learning Representations*, 2024.

[15] Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8228–8238, 2024.

[16] Kimin Lee, Hao Liu, Moonkyung Ryu, Olivia Watkins, Yuqing Du, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, and Shixiang Shane Gu. Aligning text-to-image models using human feedback. *arXiv preprint arXiv:2302.12192*, 2023.

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

[18] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

[19] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[20] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

[21] Qiang Liu. Rectified flow: A marginal preserving approach to optimal transport. *arXiv preprint arXiv:2209.14577*, 2022.

[22] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

[23] Michael S Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. *arXiv preprint arXiv:2209.15571*, 2022.

[24] Qiang Liu. Let Us Flow Together. Lecture notes, `https://www.cs.utexas.edu/~lqiang/PDF/flow_book.pdf`, 2024. Version dated December 24, 2024.

[25] Michael F Hutchinson. A stochastic estimator of the trace of the influence matrix for laplacian smoothing splines. *Communications in Statistics-Simulation and Computation*, 18(3):1059–1076, 1989.

[26] Omar G. Younis, Rodrigo Perez-Vicente, John U. Balis, Will Dudley, Alex Davey, and Jordan K Terry. Minari. Zenodo, version 0.5.0, `https://doi.org/10.5281/zenodo.13767625`, September 2024.

[27] Justin Fu, Aviral Kumar, Ofir Nachum, George Tucker, and Sergey Levine. D4rl: Datasets for deep data-driven reinforcement learning. *arXiv preprint arXiv:2004.07219*, 2020.

[28] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[29] Qiang Liu, Runlong Liao, Bo Liu, and Xixi Hu. PyTorch RectifiedFlow. `https://github.com/lqiang67/rectified-flow`, 2024.

[30] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

[31] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016.

[32] Mehdi SM Sajjadi, Olivier Bachem, Mario Lucic, Olivier Bousquet, and Sylvain Gelly. Assessing generative models via precision and recall. *Advances in neural information processing systems*, 31, 2018.

[33] Xixi Hu, Runlong Liao, Keyang Xu, Bo Liu, Yeqing Li, Eugene Ie, Hongliang Fei, and Qiang Liu. Improving rectified flow with boundary conditions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18177–18186, 2025.

[34] Ed Pizzi, Sreya Dutta Roy, Sugosh Nagavara Ravindra, Priya Goyal, and Matthijs Douze. A self-supervised descriptor for image copy detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14532–14542, 2022.

[35] Michael Kirchhof, James Thornton, Louis Béthune, Pierre Ablin, Eugene Ndiaye, and Marco Cuturi. Shielded diffusion: Generating novel and diverse images using sparse repellency. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 30911–30942, 2025.

[36] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

[37] Samuele Poppi, Tobia Poppi, Federico Cocchi, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Safe-CLIP: Removing NSFW Concepts from Vision-and-Language Models. In *Proceedings of the European Conference on Computer Vision*, 2024.

[38] Yu-Lin Tsai, Chia-Yi Hsu, Chulin Xie, Chih-Hsun Lin, Jia-You Chen, Bo Li, Pin-Yu Chen, Chia-Mu Yu, and Chun-Ying Huang. Ring-a-bell! how reliable are concept removal methods for diffusion models? *arXiv preprint arXiv:2310.10012*, 2023.

[39] Jaehong Yoon, Shoubin Yu, Vaidehi Patil, Huaxiu Yao, and Mohit Bansal. SAFREE: Training-free and adaptive guard for safe text-to-image and video generation. In *The Thirteenth International Conference on Learning Representations*, 2025.

[40] Mingyu Kim, Dongjun Kim, Amman Yusuf, Stefano Ermon, and Mijung Park. Training-free safe denoisers for safe use of diffusion models. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

[41] Z-Image Team, Huanqia Cai, Sihan Cao, Ruoyi Du, Peng Gao, Aiming Hao, Steven C. H. Hoi, Zhaohui Hou, Shijie Huang, Dengyang Jiang, Yuming Jiang, Xin Jin, Liangchen Li, et al. Z-Image: An efficient image generation foundation model with single-stream diffusion transformer. *arXiv preprint arXiv:2511.22699*, 2025.

[42] Arwen Bradley and Preetum Nakkiran. Classifier-free guidance is a predictor-corrector. *Transactions on Machine Learning Research*, 2025.

[43] Badr Moufad, Yazid Janati, Alain Durmus, Ahmed Ghorbel, Eric Moulines, and Jimmy Olsson. Conditional diffusion models with classifier-free gibbs-like guidance. *arXiv preprint arXiv:2505.21101*, 2025.

[44] Tero Karras, Miika Aittala, Tuomas Kynkäänniemi, Jaakko Lehtinen, Timo Aila, and Samuli Laine. Guiding a diffusion model with a bad version of itself. *Advances in Neural Information Processing Systems*, 37:52996–53021, 2024.

[45] Tuomas Kynkäänniemi, Miika Aittala, Tero Karras, Samuli Laine, Timo Aila, and Jaakko Lehtinen. Applying guidance in a limited interval improves sample and distribution quality in diffusion models. In *Advances in Neural Information Processing Systems*, 2024.

[46] Hyungjin Chung, Jeongsol Kim, Geon Yeong Park, Hyelin Nam, and Jong Chul Ye. Cfg++: Manifold-constrained classifier free guidance for diffusion models. In *The Thirteenth International Conference on Learning Representations*, 2025.

[47] Seyedmorteza Sadat, Otmar Hilliges, and Romann M. Weber. Eliminating oversaturation and artifacts of high guidance scales in diffusion models. In *The Thirteenth International Conference on Learning Representations*, 2025.

[48] Weichen Fan, Amber Yijia Zheng, Raymond A. Yeh, and Ziwei Liu. Cfg-zero*: Improved classifier-free guidance for flow matching models. *arXiv preprint arXiv:2503.18886*, 2025.

[49] Cheng Jin, Zhenyu Xiao, Chutao Liu, and Yuantao Gu. Angle domain guidance: Latent diffusion requires rotation rather than extrapolation. In *International Conference on Machine Learning*, pages 28187–28212. PMLR, 2025.

[50] Felix Koulischer, Johannes Deleu, Gabriel Raya, Thomas Demeester, and Luca Ambrogioni. Dynamic negative guidance of diffusion models. In *The Thirteenth International Conference on Learning Representations*, 2025.

[51] Yilun Du, Shuang Li, and Igor Mordatch. Compositional visual generation with energy-based models. In *Advances in Neural Information Processing Systems*, 2020.

[52] Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B. Tenenbaum. Compositional visual generation with composable diffusion models. In *European Conference on Computer Vision*, 2022.

[53] Manuel Brack, Felix Friedrich, Dominik Hintersdorf, Lukas Struppek, Patrick Schramowski, and Kristian Kersting. Sega: Instructing text-to-image models using semantic guidance. In *Advances in Neural Information Processing Systems*, 2023.

[54] Rohit Gandikota, Joanna Materzynska, Jaden Fiotto-Kaufman, and David Bau. Erasing concepts from diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

[55] Rohit Gandikota, Hadas Orgad, Yonatan Belinkov, Joanna Materzynska, and David Bau. Unified concept editing in diffusion models. In *Winter Conference on Applications of Computer Vision (WACV)*, 2024.

[56] Shilin Lu, Zilan Wang, Leyang Li, Yanzhu Liu, and Adams Wai-Kin Kong. Mace: Mass concept erasure in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[57] Yimeng Zhang, Xin Chen, Jinghan Jia, Yihua Zhang, Chongyu Fan, Jiancheng Liu, Mingyi Hong, Ke Ding, and Sijia Liu. Defensive unlearning with adversarial training for robust concept erasure in diffusion models. In *Advances in Neural Information Processing Systems*, 2024.

[58] Bartosz Cywiński and Kamil Deja. SAeuron: Interpretable concept unlearning in diffusion models with sparse autoencoders. In *Forty-Second International Conference on Machine Learning*, 2025.

[59] Nicholas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramèr, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. In *32nd USENIX Security Symposium (USENIX Security 23)*, pages 5253–5270, 2023.

[60] Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Diffusion art or digital forgery? investigating data replication in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[61] Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Understanding and mitigating copying in diffusion models. In *Advances in Neural Information Processing Systems*, 2023.

[62] Yuxin Wen, Yuchen Liu, Chen Chen, and Lingjuan Lyu. Detecting, explaining, and mitigating memorization in diffusion models. In *The Twelfth International Conference on Learning Representations*, 2024.

[63] Tim Dockhorn, Tianshi Cao, Arash Vahdat, and Karsten Kreis. Differentially private diffusion models. *Transactions on Machine Learning Research*, 2023.

[64] Giannis Daras, Kulin Shah, Yuval Dagan, Aravind Gollakota, Alexandros G. Dimakis, and Adam Klivans. Ambient diffusion: Learning clean distributions from corrupted data. In *Advances in Neural Information Processing Systems*, 2023.

[65] Anubhav Jain, Yuya Kobayashi, Takashi Shibuya, Yuhta Takida, Nasir Memon, Julian Togelius, and Yuki Mitsufuji. Classifier-free guidance inside the attraction basin may cause memorization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

[66] Seyedmorteza Sadat, Jakob Buhmann, Derek Bradley, Otmar Hilliges, and Romann M. Weber. Cads: Unleashing the diversity of diffusion models through condition-annealed sampling. In *The Twelfth International Conference on Learning Representations*, 2024.

[67] Masashi Sugiyama, Taiji Suzuki, and Takafumi Kanamori. *Density Ratio Estimation in Machine Learning*. Cambridge University Press, 2012.

[68] Michael U. Gutmann and Aapo Hyvärinen. Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. *Journal of Machine Learning Research*, 13:307–361, 2012.

[69] Kristy Choi, Chenlin Meng, Yang Song, and Stefano Ermon. Density ratio estimation via infinitesimal classification. In *International Conference on Artificial Intelligence and Statistics*, 2022.

[70] Dongjun Kim, Yeongmin Kim, Se Jung Kwon, Wanmo Kang, and Il-Chul Moon. Refining generative process with discriminator guidance in score-based diffusion models. In *International Conference on Machine Learning*, 2023.

[71] Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems*, 2018.

[72] Will Grathwohl, Ricky T. Q. Chen, Jesse Bettencourt, Ilya Sutskever, and David Duvenaud. Ffjord: Free-form continuous dynamics for scalable reversible generative models. In *International Conference on Learning Representations*, 2019.

[73] Fu-Yun Wang, Yunhao Shui, Jingtan Piao, Keqiang Sun, and Hongsheng Li. Diffusion-npo: Negative preference optimization for better preference-aligned generation of diffusion models. In *The Thirteenth International Conference on Learning Representations*, 2025.

[74] Michael S Albergo, Nicholas M Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *arXiv preprint arXiv:2303.08797*, 2023.

[75] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[76] Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. In *The Thirteenth International Conference on Learning Representations*, 2025.

[77] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[78] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023.

[79] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

A Related Work

Inference-time guidance and its target distribution. Classifier guidance augments the diffusion score with the gradient of an external classifier, whereas classifier-free guidance (CFG) forms an affine combination of conditional and unconditional model predictions [3, 4]. Both use a user-specified guidance scale. Recent theory shows that CFG should not generally be interpreted as exact sampling from a simple tilt of the clean-data distribution: its effect can depend on the sampler, admit a predictor–corrector interpretation, or require additional correction terms to realize a prescribed tilted target [42, 43]. Other work changes the reference branch itself, for example by guiding with a weaker version of the same model [44].

Signed RF asks a complementary question: given distributions π^+ and π^-, what RF dynamics are induced by the signed marginal $\pi_t^{\texttt{sign}} = (1+\alpha)\pi_t^+ - \alpha\pi_t^-$? Applying the RF construction to this signed marginal yields a guidance-like velocity whose strength is determined locally by the density ratio between the negative and positive branches. Once α is fixed, the effective local scale is therefore induced by the signed target rather than prescribed as a fixed schedule. Once the two branch RFs and α are fixed, they determine the signed flux and velocity away from the zero set. Under the assumptions of Appendix B, the zero set forms a dynamical boundary underlying the nonpenetration, reachable-region, and buffer-region behavior analyzed in Sec. 2.2.

Adaptive and geometry-aware guidance. Several methods refine constant-scale CFG by changing when or how guidance is applied. Guidance Interval restricts it to a selected range of timesteps [45]. CFG++ modifies the update to better respect the data manifold and reduce mode collapse [46]. Adaptive Projected Guidance decomposes the update into components parallel and orthogonal to the conditional prediction and down-weights the parallel component associated with oversaturation [47]. CFG-Zero suppresses early-step guidance and adds a scalar correction for inaccurate velocity estimates [48]. Momentum Guidance constructs a trajectory-dependent signal using an exponential

moving average of past velocities [8], while Angle Domain Guidance emphasizes angular rather than radial changes to avoid latent-norm amplification [49].

These methods adapt the schedule, geometry, normalization, or temporal aggregation of a guidance update. Signed RF is adaptive for a different reason: its scale is derived from $r_t(\boldsymbol{x}) = \pi_t^-(\boldsymbol{x})/\pi_t^+(\boldsymbol{x})$, and therefore grows where the current state is locally more compatible with the negative branch. Moreover, $(1+\alpha) - \alpha r_t(\boldsymbol{x})$ is the signed-density factor, so its zero set is a structural boundary rather than only a numerical singularity. Learned ratio estimates and clipping approximate this idealized field in practice; under the stated assumptions, the strict nonpenetration result applies to the exact Signed RF dynamics.

Dynamic negative guidance and negative prompting. Dynamic Negative Guidance (DNG) and Training-Free Safe Denoisers are closely related adaptive negative-guidance methods [40, 50]. DNG derives a state-dependent suppression rule from posterior probabilities for class removal. Safe Denoisers derive an adaptive denoiser from a binary safe–unsafe partition. Signed RF instead permits a general, potentially overlapping pair (π^+, π^-), constructs the signed target $(1+\alpha)\pi^+ - \alpha\pi^-$, and analyzes the resulting zero boundary, reachable region, and buffer region.

Negative prompting and semantic-guidance variants also use subtractive directions during sampling, while broader compositional-generation methods combine multiple energy or score terms [51–53]. Signed RF provides a target-level construction for positive–negative distribution pairs and derives the associated dynamics.

Concept erasure, safety, and controllable suppression. Concept-erasure methods such as ESD, UCE, MACE, and adversarially robust unlearning alter model weights or attention parameters [54–57]. SAeUron instead trains a sparse autoencoder and ablates concept-associated activations at inference time [58]. Other inference-time safety methods retain the base generator and intervene through safety guidance, prompt or representation filtering, or adaptive safe denoisers [11, 39, 40].

Signed RF belongs to the inference-time guidance family. The parameter α controls the overall degree of suppression, while $r_t(\boldsymbol{x})$ determines where strong repulsion is needed. Guidance is consequently weak where the negative-to-positive density ratio is small and grows as that ratio increases.

Memorization and anti-memorization. Diffusion models can reproduce or closely imitate training examples, and such replication can be detected with nearest-neighbor and copy-detection metrics such as SSCD [34, 59–62]. Training-time mitigations include DP-SGD and learning from corrupted observations [63, 64]. At inference time, guidance can be deferred near attraction basins or samples can be repelled from protected reference sets [35, 65]. CADS instead anneals the condition to improve sample diversity; it is relevant to guidance-induced collapse but is not a direct anti-memorization method [66].

In our anti-memorization instantiation, we take the negative branch to be the empirical training distribution. For a finite dataset, the Gaussian-source RF marginal and velocity toward the empirical measure admit closed forms, yielding an analytic negative flow whose trajectories are attracted toward training examples. Signed RF subtracts this branch at the measure level rather than applying pairwise repulsion to individual references. Near terminal time, the empirical-flow velocity contains the singular factor $1/(1-t)$; Signed RF modulates its strength using the local density ratio instead of a fixed scale.

Density ratios, discriminator guidance, and preference learning. Estimating $r_t(\boldsymbol{x}) = \pi_t^-(\boldsymbol{x})/\pi_t^+(\boldsymbol{x})$ connects Signed RF to density-ratio estimation and noise-contrastive estimation [67, 68]. With balanced positive and negative samples, a Bayes-optimal binary classifier's odds recover the ratio. DRE-∞ estimates an endpoint density ratio by integrating learned time scores across a continuum of bridge distributions [69]; this is related to, but distinct from, tracking a ratio along each sampling trajectory. Discriminator Guidance similarly uses a classifier-derived signal to refine generation [70], while continuous-normalizing-flow likelihood methods supply the divergence identities used for trajectory-wise log-ratio tracking [25, 71, 72].

DPO uses relative policy log-likelihoods, while Diffusion-DPO adapts the objective using an ELBO-derived surrogate for diffusion likelihoods [13, 15]. Diffusion-NPO fine-tunes a diffusion model using negative preferences [73], whereas DDPO optimizes expected reward with policy gradients [14].

These are training-time alignment methods. Signed RF instead uses a positive–negative density ratio at sampling time to set the local guidance strength and determine the valid sampling law induced by the signed target.

B Proofs and Additional Derivations

This appendix provides the proofs and derivations deferred from Sec. 2.2 and Sec. 2.3. Let π_1^+ and π_1^- be two target distributions, and let (π_t^+, v_t^+) and (π_t^-, v_t^-) denote the corresponding RF marginals and velocity fields induced from the source π_0. For $\alpha > 0$, define the signed marginal

$$\pi_t^{\texttt{sign}}(\boldsymbol{x}) = (1+\alpha)\,\pi_t^+(\boldsymbol{x}) - \alpha\,\pi_t^-(\boldsymbol{x}).$$

We partition the state space according to the sign of $\pi_t^{\texttt{sign}}$:

$$\Omega_t^+ \coloneqq \{\boldsymbol{x} : \pi_t^{\texttt{sign}}(\boldsymbol{x}) > 0\}, \qquad \Omega_t^0 \coloneqq \{\boldsymbol{x} : \pi_t^{\texttt{sign}}(\boldsymbol{x}) = 0\}, \qquad \Omega_t^- \coloneqq \{\boldsymbol{x} : \pi_t^{\texttt{sign}}(\boldsymbol{x}) < 0\}.$$

The Signed RF velocity field is well defined on $\Omega_t^+ \cup \Omega_t^-$ and becomes singular on the moving zero set Ω_t^0. Under the regularity assumptions stated below, it is locally smooth away from Ω_t^0, so standard ODE theory applies up to the first time a trajectory approaches this zero set. Since both RF branches share the same source distribution, $\pi_0^{\texttt{sign}} = (1+\alpha)\pi_0 - \alpha\pi_0 = \pi_0 > 0$, and hence all source-initialized trajectories begin in Ω_0^+.

Derivation of the Signed RF velocity. For each branch, $\pi_t^{\pm}(\boldsymbol{x})$ denotes the RF marginal density and $\pi_t^{\pm}(\boldsymbol{x})v_t^{\pm}(\boldsymbol{x})$ denotes the corresponding probability flux. Both quantities depend linearly on the terminal distribution. Therefore, for the signed target $\pi_1^{\texttt{sign}} = (1+\alpha)\pi_1^+ - \alpha\pi_1^-$, the induced signed marginal and flux are

$$\pi_t^{\texttt{sign}}(\boldsymbol{x}) = (1+\alpha)\,\pi_t^+(\boldsymbol{x}) - \alpha\,\pi_t^-(\boldsymbol{x}), \tag{8}$$

$$\pi_t^{\texttt{sign}}(\boldsymbol{x})v_t^{\texttt{signRF}}(\boldsymbol{x}) = (1+\alpha)\,\pi_t^+(\boldsymbol{x})v_t^+(\boldsymbol{x}) - \alpha\,\pi_t^-(\boldsymbol{x})v_t^-(\boldsymbol{x}). \tag{9}$$

As in standard RF, the velocity field is obtained by dividing the flux by the marginal density. Thus, wherever $\pi_t^{\texttt{sign}}(\boldsymbol{x}) \neq 0$,

$$v_t^{\texttt{signRF}}(\boldsymbol{x}) = \frac{(1+\alpha)\,\pi_t^+(\boldsymbol{x})v_t^+(\boldsymbol{x}) - \alpha\,\pi_t^-(\boldsymbol{x})v_t^-(\boldsymbol{x})}{(1+\alpha)\,\pi_t^+(\boldsymbol{x}) - \alpha\,\pi_t^-(\boldsymbol{x})}. \tag{10}$$

This recovers Eq. 3 in the main text.

To obtain the guidance form, define $r_t(\boldsymbol{x}) \coloneqq \pi_t^-(\boldsymbol{x})/\pi_t^+(\boldsymbol{x})$ and $\Delta v_t(\boldsymbol{x}) \coloneqq v_t^+(\boldsymbol{x}) - v_t^-(\boldsymbol{x})$. Dividing the numerator and denominator of Eq. 10 by $\pi_t^+(\boldsymbol{x})$ gives

$$\begin{aligned} v_t^{\texttt{signRF}}(\boldsymbol{x}) &= \frac{(1+\alpha)v_t^+(\boldsymbol{x}) - \alpha\,r_t(\boldsymbol{x})v_t^-(\boldsymbol{x})}{(1+\alpha) - \alpha\,r_t(\boldsymbol{x})} \\ &= v_t^+(\boldsymbol{x}) + \frac{\alpha\,r_t(\boldsymbol{x})}{(1+\alpha) - \alpha\,r_t(\boldsymbol{x})}\Delta v_t(\boldsymbol{x}). \end{aligned}$$

Hence,

$$v_t^{\texttt{signRF}}(\boldsymbol{x}) = v_t^+(\boldsymbol{x}) + \lambda_t^\alpha(\boldsymbol{x})\Delta v_t(\boldsymbol{x}), \qquad \lambda_t^\alpha(\boldsymbol{x}) = \frac{\alpha\,r_t(\boldsymbol{x})}{(1+\alpha) - \alpha\,r_t(\boldsymbol{x})}. \tag{11}$$

The denominator satisfies

$$(1+\alpha) - \alpha\,r_t(\boldsymbol{x}) = \frac{\pi_t^{\texttt{sign}}(\boldsymbol{x})}{\pi_t^+(\boldsymbol{x})}.$$

Therefore, assuming $\pi_t^+(\boldsymbol{x}) > 0$, it vanishes precisely on the signed zero set Ω_t^0, where the Signed RF velocity becomes singular.

Proposition B.1 (TV-optimal probability approximation)**.** *Fix $t < 1$, and define the total negative mass of the signed marginal by*

$$M_t^- \coloneqq -\int_{\Omega_t^-} \pi_t^{\texttt{sign}}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}.$$

Then, among all probability densities q,

$$\inf_{\substack{q(\boldsymbol{x})\geq 0 \\ \int q(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}=1}} \frac{1}{2}\int_{\mathbb{R}^d} \left|q(\boldsymbol{x}) - \pi_t^{\texttt{sign}}(\boldsymbol{x})\right| \mathrm{d}\boldsymbol{x} = M_t^-.$$

Moreover, q is optimal if and only if

$$q(\boldsymbol{x}) = 0 \quad \textit{on } \Omega_t^0 \cup \Omega_t^-, \qquad 0 \leq q(\boldsymbol{x}) \leq \pi_t^{\texttt{sign}}(\boldsymbol{x}) \quad \textit{on } \Omega_t^+,$$

up to sets of measure zero.

In particular, if $S_t \subseteq \Omega_t^+$ satisfies

$$\int_{S_t} \pi_t^{\texttt{sign}}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = 1,$$

then

$$q_{S_t}(\boldsymbol{x}) = \pi_t^{\texttt{sign}}(\boldsymbol{x})\mathbf{1}\{\boldsymbol{x} \in S_t\}$$

is TV-optimal. By Theorem B.4, the Signed RF sampling law corresponds to $S_t = \Omega_t^+$, and is therefore one such optimal approximation.

Proof. Since $\pi_t^{\texttt{sign}}$ has total mass one, its positive region contains total mass

$$\int_{\Omega_t^+} \pi_t^{\texttt{sign}}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = 1 + M_t^-.$$

Now let q be any probability density. Define

$$A \coloneqq \int_{\Omega_t^0 \cup \Omega_t^-} q(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$$

as the probability mass that q assigns outside the positive region, and define

$$E \coloneqq \int_{\Omega_t^+} \big(q(\boldsymbol{x}) - \pi_t^{\texttt{sign}}(\boldsymbol{x})\big)_+\,\mathrm{d}\boldsymbol{x}$$

as the amount by which q exceeds the signed density inside the positive region.

Because q has mass $1 - A$ on Ω_t^+,

$$\int_{\Omega_t^+} \big(\pi_t^{\texttt{sign}}(\boldsymbol{x}) - q(\boldsymbol{x})\big)\,\mathrm{d}\boldsymbol{x} = (1 + M_t^-) - (1 - A) = M_t^- + A.$$

Therefore,

$$\int_{\Omega_t^+} \left|q(\boldsymbol{x}) - \pi_t^{\texttt{sign}}(\boldsymbol{x})\right| \mathrm{d}\boldsymbol{x} = M_t^- + A + 2E.$$

On $\Omega_t^0 \cup \Omega_t^-$, we have $\pi_t^{\texttt{sign}}(\boldsymbol{x}) \leq 0$ and $q(\boldsymbol{x}) \geq 0$, so

$$\int_{\Omega_t^0 \cup \Omega_t^-} \left|q(\boldsymbol{x}) - \pi_t^{\texttt{sign}}(\boldsymbol{x})\right| \mathrm{d}\boldsymbol{x} = A + M_t^-.$$

Combining the two regions gives

$$\frac{1}{2}\int_{\mathbb{R}^d} \left|q(\boldsymbol{x}) - \pi_t^{\texttt{sign}}(\boldsymbol{x})\right| \mathrm{d}\boldsymbol{x} = M_t^- + A + E \geq M_t^-.$$

Equality holds precisely when $A = 0$ and $E = 0$, meaning that q assigns no mass outside Ω_t^+ and never exceeds $\pi_t^{\texttt{sign}}$ inside Ω_t^+. This proves the stated characterization.

Finally, any $q_{S_t}(\boldsymbol{x}) = \pi_t^{\texttt{sign}}(\boldsymbol{x})\mathbf{1}\{\boldsymbol{x} \in S_t\}$ with unit total mass satisfies these two conditions and is therefore TV-optimal. □

Assumption B.2 (Regularity of the signed flow)**.** Let π_1^+ and π_1^- be two target distributions with finite first moments, and suppose that both RF branches use the Gaussian source $\pi_0 = \mathcal{N}(\mathbf{0}, \boldsymbol{I})$. Let (π_t^+, v_t^+) and (π_t^-, v_t^-) denote the corresponding RF marginals and velocity fields.

We assume that the signed marginal

$$(t, \boldsymbol{x}) \longmapsto \pi_t^{\texttt{sign}}(\boldsymbol{x}) = (1+\alpha)\pi_t^+(\boldsymbol{x}) - \alpha\pi_t^-(\boldsymbol{x})$$

and the signed probability flux

$$(t, \boldsymbol{x}) \longmapsto (1+\alpha)\pi_t^+(\boldsymbol{x})v_t^+(\boldsymbol{x}) - \alpha\pi_t^-(\boldsymbol{x})v_t^-(\boldsymbol{x})$$

are locally C^1 on $(0,1) \times \mathbb{R}^d$ and, together with their spatial gradients, extend continuously to $[0,1) \times \mathbb{R}^d$. Let

$$\mathcal{D} := \left\{(t, \boldsymbol{x}) : 0 < t < 1,\ \pi_t^{\texttt{sign}}(\boldsymbol{x}) = 0,\ \nabla\pi_t^{\texttt{sign}}(\boldsymbol{x}) = 0\right\}$$

denote the degenerate points of the zero set. We further assume:

(i) *Avoidance of degenerate zeros.* For π_0-almost every $\boldsymbol{z}_0$, the Signed RF trajectory started at $\boldsymbol{z}_0$ has no limit point in $\mathcal{D}$ at the end of its maximal interval of existence.

(ii) *Non-explosion.* For π_0-almost every $\boldsymbol{z}_0$, the Signed RF trajectory started at $\boldsymbol{z}_0$ is bounded on $I \cap [0, T]$ for each $T < 1$, where I is its maximal interval of existence.

For the common Gaussian source $\pi_0 = \mathcal{N}(\mathbf{0}, \boldsymbol{I})$, Tweedie's identity gives, for each RF branch,

$$\pi_t^{\pm}(\boldsymbol{x})v_t^{\pm}(\boldsymbol{x}) = \frac{\boldsymbol{x}}{t}\,\pi_t^{\pm}(\boldsymbol{x}) + \frac{1-t}{t}\,\nabla\pi_t^{\pm}(\boldsymbol{x}), \qquad 0 < t < 1.$$

Taking the same signed combination as in the definition of $\pi_t^{\texttt{sign}}$ yields

$$(1+\alpha)\pi_t^+(\boldsymbol{x})v_t^+(\boldsymbol{x}) - \alpha\pi_t^-(\boldsymbol{x})v_t^-(\boldsymbol{x}) = \frac{\boldsymbol{x}}{t}\,\pi_t^{\texttt{sign}}(\boldsymbol{x}) + \frac{1-t}{t}\,\nabla\pi_t^{\texttt{sign}}(\boldsymbol{x}).$$

Therefore, at any nondegenerate point of the zero set, where $\pi_t^{\texttt{sign}}(\boldsymbol{x}) = 0$ and $\nabla\pi_t^{\texttt{sign}}(\boldsymbol{x}) \neq 0$,

$$\left(\nabla\pi_t^{\texttt{sign}}(\boldsymbol{x})\right)^\top \left[(1+\alpha)\pi_t^+(\boldsymbol{x})v_t^+(\boldsymbol{x}) - \alpha\pi_t^-(\boldsymbol{x})v_t^-(\boldsymbol{x})\right] = \frac{1-t}{t}\left\|\nabla\pi_t^{\texttt{sign}}(\boldsymbol{x})\right\|^2 > 0.$$

Thus, the signed flux points locally toward increasing $\pi_t^{\texttt{sign}}$, i.e., toward the positive side of the zero set. This is the barrier property used in the nonpenetration proof below.

Proposition B.3 (Signed RF remains in the positive region)**.** *Suppose that Assumption B.2 holds. Then, for π_0-almost every $\boldsymbol{z}_0$, the Signed RF trajectory started at $\boldsymbol{z}_0$ exists on $[0,1)$ and satisfies $\boldsymbol{z}_t \in \Omega_t^+$ for all $t \in [0,1)$. Consequently, the law $\pi_t^{\texttt{signRF}}$ of the source-initialized solution satisfies*

$$\pi_t^{\texttt{signRF}}(\Omega_t^+) = 1, \qquad \textit{for every } t \in [0,1).$$

Proof. Let $\boldsymbol{z}_t$ be a source-initialized Signed RF trajectory. Since both RF branches share the same Gaussian source,

$$\pi_0^{\texttt{sign}}(\boldsymbol{z}_0) = (1+\alpha)\pi_0(\boldsymbol{z}_0) - \alpha\pi_0(\boldsymbol{z}_0) = \pi_0(\boldsymbol{z}_0) > 0.$$

Thus, every trajectory starts in the positive region Ω_0^+.

Define the signed probability current

$$j_t^{\texttt{sign}}(\boldsymbol{x}) := (1+\alpha)\pi_t^+(\boldsymbol{x})v_t^+(\boldsymbol{x}) - \alpha\pi_t^-(\boldsymbol{x})v_t^-(\boldsymbol{x}).$$

By linearity of the two branch continuity equations,

$$\partial_t\pi_t^{\texttt{sign}} + \nabla\cdot j_t^{\texttt{sign}} = 0.$$

Moreover, wherever $\pi_t^{\texttt{sign}}(\boldsymbol{x}) \neq 0$,

$$v_t^{\texttt{signRF}}(\boldsymbol{x}) = \frac{j_t^{\texttt{sign}}(\boldsymbol{x})}{\pi_t^{\texttt{sign}}(\boldsymbol{x})}.$$

Therefore, as long as the trajectory remains in Ω_t^+, the chain rule gives

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\pi_t^{\texttt{sign}}(\boldsymbol{z}_t) &= \partial_t\pi_t^{\texttt{sign}}(\boldsymbol{z}_t) + \big(\nabla\pi_t^{\texttt{sign}}(\boldsymbol{z}_t)\big)^\top \dot{\boldsymbol{z}}_t \\ &= -\nabla\cdot j_t^{\texttt{sign}}(\boldsymbol{z}_t) + \frac{\big(\nabla\pi_t^{\texttt{sign}}(\boldsymbol{z}_t)\big)^\top j_t^{\texttt{sign}}(\boldsymbol{z}_t)}{\pi_t^{\texttt{sign}}(\boldsymbol{z}_t)}. \end{aligned} \tag{12}$$

Since $v_t^{\texttt{signRF}} = j_t^{\texttt{sign}}/\pi_t^{\texttt{sign}}$ and its spatial Jacobian are continuous on the relatively open set $U \coloneqq \{(t,\boldsymbol{x}) : 0 \le t < 1,\ \pi_t^{\texttt{sign}}(\boldsymbol{x}) > 0\}$, the Signed RF ODE has a unique maximal solution $\boldsymbol{z}_t$ in U, defined on some interval $[0, t_{\max})$ with $0 < t_{\max} \le 1$; if $t_{\max} < 1$, then $(t, \boldsymbol{z}_t)$ eventually leaves every compact subset of U as $t \uparrow t_{\max}$. Fix $\boldsymbol{z}_0$ in the full-measure set on which Assumption B.2(i) and (ii) hold, and suppose, for contradiction, that $t_{\max} < 1$.

By Assumption B.2(ii), $\boldsymbol{z}_t$ stays in a closed ball $\bar{B}_R$ for $t < t_{\max}$. Hence the set K of limit points of $(t, \boldsymbol{z}_t)$ as $t \uparrow t_{\max}$ is a nonempty compact subset of $\{t_{\max}\} \times \bar{B}_R$. By the escape property, $K \cap U = \emptyset$; since $\pi_t^{\texttt{sign}}(\boldsymbol{z}_t) > 0$, continuity gives $\pi_{t_{\max}}^{\texttt{sign}}(\boldsymbol{x}) = 0$ for every $(t_{\max}, \boldsymbol{x}) \in K$, and $\pi_t^{\texttt{sign}}(\boldsymbol{z}_t) \to 0$ as $t \uparrow t_{\max}$.

By Assumption B.2(i), $K \cap \mathcal{D} = \emptyset$, so $\nabla\pi_{t_{\max}}^{\texttt{sign}}(\boldsymbol{x}) \neq 0$ on K. By the Gaussian-source identity derived above,

$$\big(\nabla\pi_{t_{\max}}^{\texttt{sign}}(\boldsymbol{x})\big)^\top j_{t_{\max}}^{\texttt{sign}}(\boldsymbol{x}) = \frac{1 - t_{\max}}{t_{\max}} \left\|\nabla\pi_{t_{\max}}^{\texttt{sign}}(\boldsymbol{x})\right\|^2 > 0 \qquad \text{for } (t_{\max}, \boldsymbol{x}) \in K.$$

By continuity and compactness of K, there exist an open neighborhood N of K and constants $c > 0$ and $M < \infty$ such that

$$\big(\nabla\pi_t^{\texttt{sign}}(\boldsymbol{x})\big)^\top j_t^{\texttt{sign}}(\boldsymbol{x}) \ge c, \qquad \left|\nabla\cdot j_t^{\texttt{sign}}(\boldsymbol{x})\right| \le M, \qquad (t,\boldsymbol{x}) \in N,$$

and some $\delta > 0$ with $(t, \boldsymbol{z}_t) \in N$ for $t \in (t_{\max} - \delta, t_{\max})$. Equation 12 therefore implies

$$\frac{\mathrm{d}}{\mathrm{d}t}\pi_t^{\texttt{sign}}(\boldsymbol{z}_t) \ge -M + \frac{c}{\pi_t^{\texttt{sign}}(\boldsymbol{z}_t)}, \qquad t \in (t_{\max} - \delta, t_{\max}).$$

As $\pi_t^{\texttt{sign}}(\boldsymbol{z}_t) \to 0$, the right-hand side becomes positive, so the positive function $t \mapsto \pi_t^{\texttt{sign}}(\boldsymbol{z}_t)$ is increasing near $t_{\max}$ and cannot tend to 0, a contradiction.

Hence $t_{\max} = 1$ for π_0-almost every $\boldsymbol{z}_0$, and $\pi_t^{\texttt{signRF}}(\Omega_t^+) = 1$ for every $t \in [0, 1)$. □

Theorem B.4 (Sampling law of Signed RF)**.** *Suppose that Assumption B.2 holds. For $t < 1$, let G_t be the set of initial points $\boldsymbol{z}_0$ whose Signed RF trajectory exists on $[0, t]$, let $\Phi_t(\boldsymbol{z}_0)$ denote its position at time t, and define the reachable region by*

$$\Omega_t^r \coloneqq \big\{\Phi_t(\boldsymbol{z}_0) : \boldsymbol{z}_0 \in G_t\big\}.$$

Then, for every $t < 1$, the density of the source-initialized Signed RF samples is

$$\pi_t^{\texttt{signRF}}(\boldsymbol{x}) = \pi_t^{\texttt{sign}}(\boldsymbol{x})\, \mathbf{1}\{\boldsymbol{x} \in \Omega_t^r\}.$$

In particular,

$$\int_{\Omega_t^r} \pi_t^{\texttt{sign}}(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} = 1.$$

Proof. Recall that the signed probability current is

$$j_t^{\texttt{sign}}(\boldsymbol{x}) = (1+\alpha)\pi_t^+(\boldsymbol{x})v_t^+(\boldsymbol{x}) - \alpha\pi_t^-(\boldsymbol{x})v_t^-(\boldsymbol{x}).$$

By linearity of the two branch continuity equations,

$$\partial_t\pi_t^{\texttt{sign}} + \nabla\cdot j_t^{\texttt{sign}} = 0.$$

Moreover, wherever $\pi_t^{\texttt{sign}}(\boldsymbol{x}) \neq 0$,

$$j_t^{\texttt{sign}}(\boldsymbol{x}) = \pi_t^{\texttt{sign}}(\boldsymbol{x}) v_t^{\texttt{signRF}}(\boldsymbol{x}).$$

Therefore, on the positive region,

$$\partial_t \pi_t^{\texttt{sign}} + \nabla \cdot \left(\pi_t^{\texttt{sign}} v_t^{\texttt{signRF}}\right) = 0.$$

By Proposition B.3, $\pi_0(G_t) = 1$ for every $t < 1$. Since the Signed RF velocity and its spatial Jacobian are continuous on the relatively open set where $\pi_t^{\texttt{sign}} > 0$, standard ODE theory implies that G_t is open and that Φ_t is a C^1 diffeomorphism from G_t onto Ω_t^r.

Let

$$J_t(\boldsymbol{z}_0) \coloneqq |\det D\Phi_t(\boldsymbol{z}_0)|$$

denote the Jacobian determinant of the flow map. Along the trajectory $\Phi_t(\boldsymbol{z}_0)$, Liouville's formula gives

$$\frac{\mathrm{d}}{\mathrm{d}t} J_t(\boldsymbol{z}_0) = J_t(\boldsymbol{z}_0)\, \nabla \cdot v_t^{\texttt{signRF}}\big(\Phi_t(\boldsymbol{z}_0)\big).$$

Combining this identity with the continuity equation yields

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\Big[\pi_t^{\texttt{sign}}\big(\Phi_t(\boldsymbol{z}_0)\big) J_t(\boldsymbol{z}_0)\Big] &= \Big[\partial_t \pi_t^{\texttt{sign}} + \big(\nabla \pi_t^{\texttt{sign}}\big)^\top v_t^{\texttt{signRF}} + \pi_t^{\texttt{sign}} \nabla \cdot v_t^{\texttt{signRF}}\Big]_{\boldsymbol{x}=\Phi_t(\boldsymbol{z}_0)} J_t(\boldsymbol{z}_0) \\
&= \Big[\partial_t \pi_t^{\texttt{sign}} + \nabla \cdot \left(\pi_t^{\texttt{sign}} v_t^{\texttt{signRF}}\right)\Big]_{\boldsymbol{x}=\Phi_t(\boldsymbol{z}_0)} J_t(\boldsymbol{z}_0) \\
&= 0. \qquad (13)
\end{aligned}$$

Thus, signed mass is conserved along the flow map. Since $\Phi_0(\boldsymbol{z}_0) = \boldsymbol{z}_0$, $J_0(\boldsymbol{z}_0) = 1$, and

$$\pi_0^{\texttt{sign}} = (1+\alpha)\pi_0 - \alpha\pi_0 = \pi_0,$$

Eq. 13 gives

$$\pi_t^{\texttt{sign}}\big(\Phi_t(\boldsymbol{z}_0)\big) J_t(\boldsymbol{z}_0) = \pi_0(\boldsymbol{z}_0).$$

Now let f be any compactly supported test function. By the definition of the source-initialized sampling law,

$$\int_{\mathbb{R}^d} f(\boldsymbol{x}) \pi_t^{\texttt{signRF}}(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} = \int_{G_t} f\big(\Phi_t(\boldsymbol{z}_0)\big) \pi_0(\boldsymbol{z}_0)\, \mathrm{d}\boldsymbol{z}_0.$$

Substituting the conservation identity above and applying the change-of-variables formula gives

$$\begin{aligned}
\int_{\mathbb{R}^d} f(\boldsymbol{x}) \pi_t^{\texttt{signRF}}(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} &= \int_{G_t} f\big(\Phi_t(\boldsymbol{z}_0)\big) \pi_t^{\texttt{sign}}\big(\Phi_t(\boldsymbol{z}_0)\big) J_t(\boldsymbol{z}_0)\, \mathrm{d}\boldsymbol{z}_0 \\
&= \int_{\Omega_t^r} f(\boldsymbol{x}) \pi_t^{\texttt{sign}}(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x}.
\end{aligned}$$

Since this holds for every test function f,

$$\pi_t^{\texttt{signRF}}(\boldsymbol{x}) = \pi_t^{\texttt{sign}}(\boldsymbol{x})\, \mathbf{1}\{\boldsymbol{x} \in \Omega_t^r\}$$

almost everywhere. Both sides are integrable; integrating this identity over $\mathbb{R}^d$ (equivalently, applying it to compactly supported cutoffs converging to 1 and using dominated convergence) gives

$$\int_{\Omega_t^r} \pi_t^{\texttt{sign}}(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} = 1.$$

□

Remark B.5 (Discussion of the assumptions). The regularity in Assumption B.2 ensures that the continuity equation, the Signed RF ODE, and the Jacobian calculation above are well defined. Such assumptions are standard in continuous-time flow analyses [22, 23, 71, 74].

Away from $\mathcal{D}$, the zero set is locally a smooth boundary, and the Gaussian-source identity makes the signed current point toward its positive side. Degenerate zeros cannot be excluded altogether: where a negative component is born or two components merge, $\pi_t^{\texttt{sign}}$ has a critical point with value 0. For example, for $d = 1$, $\pi_1^+ = \mathcal{N}(0,1)$, $\pi_1^- = \mathcal{N}(0, 0.2^2)$, and $\alpha = 1$, a negative component is born at $\boldsymbol{x}_* = 0$ at $t_* \approx 0.654$, where $\nabla \pi_{t_*}^{\texttt{sign}}(\boldsymbol{x}_*) = 0$. By symmetry, $\boldsymbol{z}_t \equiv 0$ is a trajectory and lies in Ω_t^-

after t_*. For t close to t_*, the velocity near 0 points away from 0, and one-dimensional trajectories cannot cross $\boldsymbol{z}_t \equiv 0$, so every other trajectory avoids $(t_*, \boldsymbol{x}_*)$ and condition (i) holds. Condition (ii) rules out finite-time blow-up, the only other way a solution can cease to exist before $t = 1$.

We state the theorem for $t < 1$ because the Gaussian interpolation smooths the intermediate marginals whenever $t < 1$, whereas the terminal distributions at $t = 1$ may be nonsmooth or singular.

Online density-ratio tracking. We derive the evolution of the density ratio along a source-initialized Signed RF trajectory $\boldsymbol{z}_t$. Recall

$$r_t(\boldsymbol{x}) = \frac{\pi_t^-(\boldsymbol{x})}{\pi_t^+(\boldsymbol{x})}, \qquad u_t \coloneqq \log r_t(\boldsymbol{z}_t).$$

Since the two branches share the same source distribution, $r_0(\boldsymbol{z}_0) = 1$, and therefore $u_0 = 0$.

Let $s_t^\pm(\boldsymbol{x}) \coloneqq \nabla \log \pi_t^\pm(\boldsymbol{x})$ denote the branch score functions. Each branch satisfies the continuity equation

$$\partial_t \pi_t^\pm + \nabla\cdot\big(\pi_t^\pm v_t^\pm\big) = 0.$$

Dividing by $\pi_t^\pm$ gives

$$\partial_t \log \pi_t^\pm = -\nabla\cdot v_t^\pm - (v_t^\pm)^\top s_t^\pm.$$

Along the Signed RF trajectory,

$$\dot{\boldsymbol{z}}_t = v_t^+(\boldsymbol{z}_t) + \lambda_t^\alpha(\boldsymbol{z}_t)\Delta v_t(\boldsymbol{z}_t), \qquad \Delta v_t \coloneqq v_t^+ - v_t^-,$$

where

$$\lambda_t^\alpha(\boldsymbol{z}_t) = \frac{\alpha\exp(u_t)}{(1+\alpha) - \alpha\exp(u_t)}.$$

Using $u_t = \log \pi_t^-(\boldsymbol{z}_t) - \log \pi_t^+(\boldsymbol{z}_t)$ and applying the chain rule yields

$$\begin{aligned}\dot{u}_t = {}& \nabla\cdot \Delta v_t(\boldsymbol{z}_t) + \Delta v_t(\boldsymbol{z}_t)^\top s_t^-(\boldsymbol{z}_t) \\ & + \lambda_t^\alpha(\boldsymbol{z}_t)\,\Delta v_t(\boldsymbol{z}_t)^\top\big(s_t^-(\boldsymbol{z}_t) - s_t^+(\boldsymbol{z}_t)\big).\end{aligned} \tag{14}$$

This recovers Eq. 6 in the main text.

For RF with the Gaussian source $\pi_0 = \mathcal{N}(\mathbf{0}, \boldsymbol{I})$, the branch scores satisfy

$$s_t^\pm(\boldsymbol{x}) = \frac{t\,v_t^\pm(\boldsymbol{x}) - \boldsymbol{x}}{1-t}, \qquad s_t^-(\boldsymbol{x}) - s_t^+(\boldsymbol{x}) = -\frac{t}{1-t}\Delta v_t(\boldsymbol{x}).$$

Substituting these identities into Eq. 14 gives the velocity-only update

$$\dot{u}_t = \nabla\cdot \Delta v_t(\boldsymbol{z}_t) + \frac{\Delta v_t(\boldsymbol{z}_t)^\top\big(t\,v_t^-(\boldsymbol{z}_t) - \boldsymbol{z}_t\big)}{1-t} - \lambda_t^\alpha(\boldsymbol{z}_t)\frac{t}{1-t}\big\|\Delta v_t(\boldsymbol{z}_t)\big\|^2. \tag{15}$$

Thus, the density ratio can be tracked jointly with the Signed RF trajectory by solving the augmented ODE

$$\begin{cases}\dot{\boldsymbol{z}}_t = v_t^+(\boldsymbol{z}_t) + \lambda_t^\alpha(\boldsymbol{z}_t)\Delta v_t(\boldsymbol{z}_t), \\ \dot{u}_t = \nabla\cdot \Delta v_t(\boldsymbol{z}_t) + \dfrac{\Delta v_t(\boldsymbol{z}_t)^\top\big(t\,v_t^-(\boldsymbol{z}_t) - \boldsymbol{z}_t\big)}{1-t} - \lambda_t^\alpha(\boldsymbol{z}_t)\dfrac{t}{1-t}\big\|\Delta v_t(\boldsymbol{z}_t)\big\|^2,\end{cases}$$

with

$$\boldsymbol{z}_0 \sim \pi_0, \qquad u_0 = 0, \qquad \lambda_t^\alpha(\boldsymbol{z}_t) = \frac{\alpha\exp(u_t)}{(1+\alpha) - \alpha\exp(u_t)}.$$

The density ratio along the trajectory is then recovered as $r_t(\boldsymbol{z}_t) = \exp(u_t)$. The divergence $\nabla\cdot \Delta v_t(\boldsymbol{z}_t)$ can be evaluated exactly when tractable or estimated using Hutchinson's trace estimator.

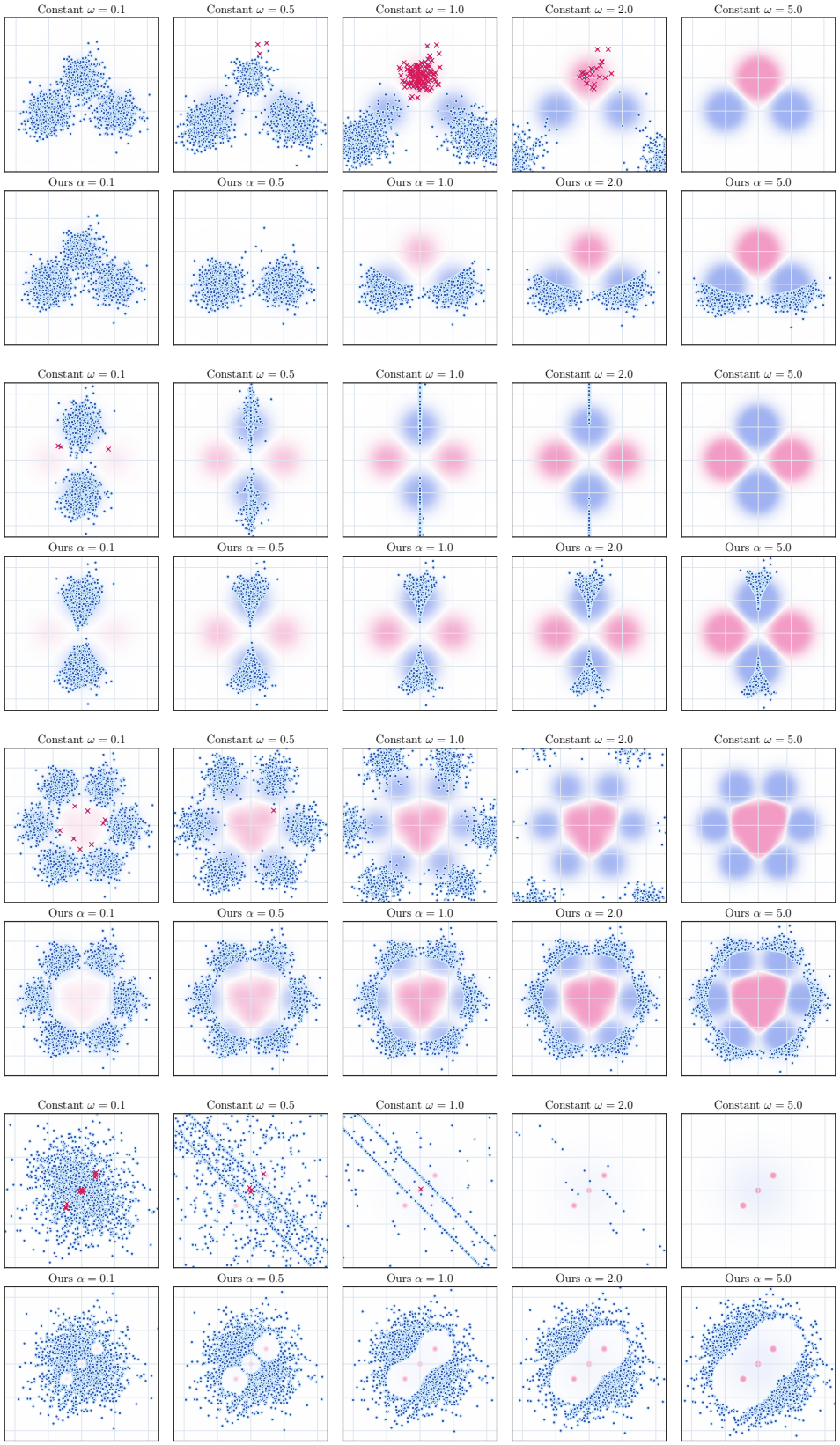


Figure 11: Additional 2D examples across multiple setups.

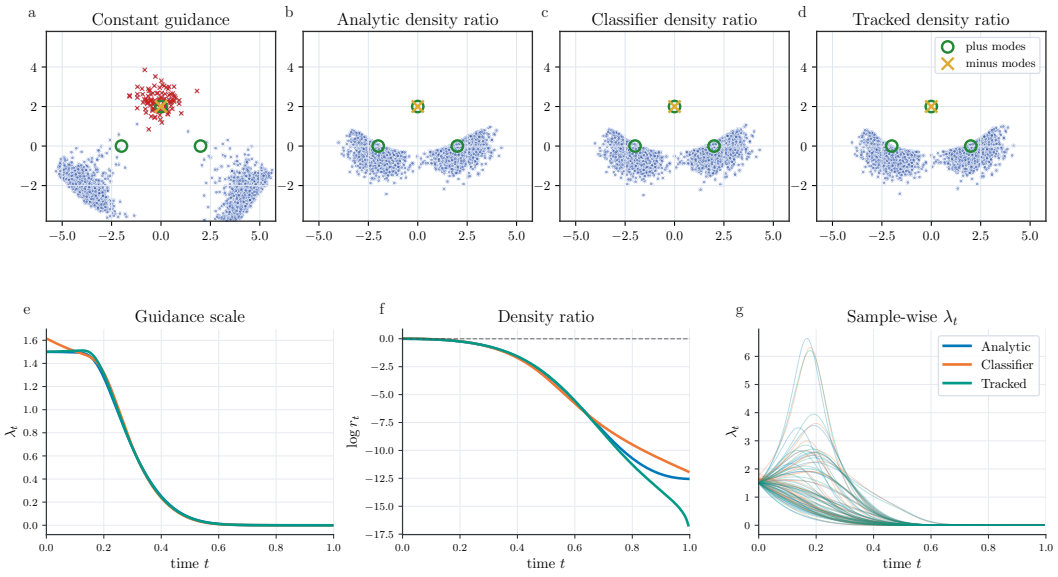


Figure 12: **Density-ratio estimation ablation.** Columns compare constant guidance, the analytic ratio, the classifier-based estimate, and the online-tracked estimate, using the same initial Gaussian noise and velocity fields v_t^+ and v_t^-. Panels (e)–(g) show the mean guidance scale λ_t^α, the mean log-density ratio $\log r_t$, and the sample-wise guidance scales along individual trajectories. Despite small estimation errors, both practical estimators produce samples close to those obtained with the analytic ratio, indicating robustness to moderate density-ratio estimation error.

C Additional Experimental Details

C.1 Additional 2D Gaussian Mixture Experiments

Additional results on 2D toys. Each toy dataset consists of a positive target π_1^+ and a negative target π_1^-, both defined as mixtures of isotropic Gaussians in $\mathbb{R}^2$ with specified component means, weights, and variances. We train separate MLP-based RF models on samples from the two targets, yielding the velocity fields v_t^+ and v_t^-. Each velocity model is a four-layer MLP with hidden width 256, SiLU activations, and a linear two-dimensional output layer. We train the models for 10,000 steps using Adam [75], with a learning rate of 10^{-3} and batch size 1024.

For Gaussian-mixture targets, the intermediate RF marginals π_t^+ and π_t^- remain Gaussian mixtures, allowing the density ratio to be evaluated in closed form. In particular, consider a Gaussian component $\boldsymbol{X}_1 \sim \mathcal{N}(\boldsymbol{\mu}_k, \sigma_k^2 \boldsymbol{I})$ and the source $\boldsymbol{X}_0 \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$. Under the RF interpolation $\boldsymbol{X}_t = t\boldsymbol{X}_1 + (1-t)\boldsymbol{X}_0$, its intermediate marginal is $\boldsymbol{X}_t \sim \mathcal{N}\big(t\boldsymbol{\mu}_k,\ \big((1-t)^2 + t^2\sigma_k^2\big)\boldsymbol{I}\big)$. Therefore,

$$r_t(\boldsymbol{x}) = \frac{\pi_t^-(\boldsymbol{x})}{\pi_t^+(\boldsymbol{x})} = \frac{\displaystyle\sum_{k=1}^{K_-} w_k^- \,\mathcal{N}\big(\boldsymbol{x};\, t\boldsymbol{\mu}_k^-,\, \big((1-t)^2 + t^2(\sigma_k^-)^2\big)\boldsymbol{I}\big)}{\displaystyle\sum_{k=1}^{K_+} w_k^+ \,\mathcal{N}\big(\boldsymbol{x};\, t\boldsymbol{\mu}_k^+,\, \big((1-t)^2 + t^2(\sigma_k^+)^2\big)\boldsymbol{I}\big)}. \tag{16}$$

We use this analytic ratio as ground truth, allowing us to study the behavior of Signed RF independently of density-ratio estimation error. Fig. 11 shows additional visualizations across a range of geometries. In all cases, sampling starts from the same Gaussian source and uses Euler integration with 200 steps. The background heatmap visualizes the target signed density at $t = 1$, $\pi_1^{\texttt{sign}} = (1+\alpha)\,\pi_1^+ - \alpha\,\pi_1^-$; blue denotes positive regions and pink denotes negative regions.

Density-ratio estimation analysis. We evaluate the accuracy and stability of the density-ratio estimators. In the same 2D setting, we train a density-ratio classifier with the same depth and

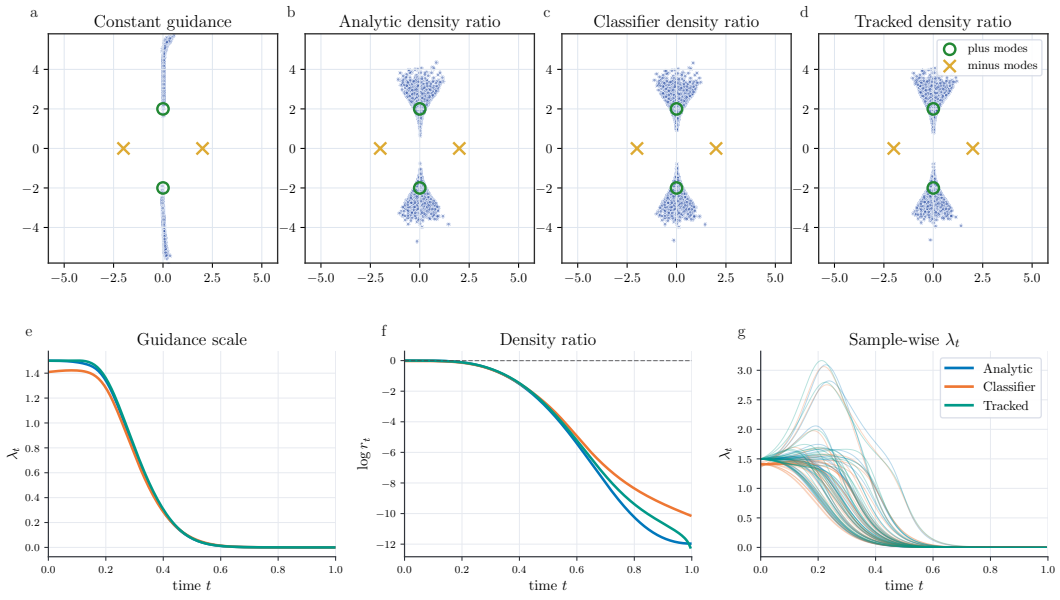


Figure 13: Additional density-ratio estimation ablation in a different setup.

hidden width as the velocity networks. The classifier outputs a single logit and is trained with binary cross-entropy on balanced samples from the positive and negative branches. We use Adam for $10{,}000$ steps, with a learning rate of 10^{-3} and batch size 2048.

Using the closed-form ratio in Eq. 16 as ground truth, we compare three variants of Signed RF: the analytic ratio, the classifier-based estimate, and the online-tracked estimate. All variants use the same initial Gaussian noise and the same velocity fields v_t^+ and v_t^-, differing only in how $r_t(\boldsymbol{x})$ is obtained during sampling. Results are shown in Figs. 12, 13, 14, and 15. Panel (e) shows the mean effective guidance scale λ_t^α; the three variants agree closely, except near the beginning of the trajectory. Panel (f) shows the mean $\log r_t$. Both practical estimators remain close to the analytic ratio over most of the trajectory, although the tracked estimate deviates slightly near the end, likely because of accumulated numerical integration error. Panel (g) shows the sample-wise guidance scales, which vary substantially across trajectories and toy configurations. This confirms that Signed RF guidance is state- and sample-dependent rather than a fixed time schedule.

Despite the observed differences in r_t and λ_t^α, the analytic, classifier-based, and tracked variants produce visually similar samples. These results suggest that Signed RF is robust to moderate density-ratio estimation error in these examples, with both practical estimators closely approaching the sampling behavior obtained using the exact ratio.

C.2 Additional ImageNet-256 Details

Comparison with related methods. Table 1 extends the comparison to several related guidance methods, including ADG [49], CFG++ [46], and MG [8]. We report these results in the appendix because these methods are not derived from the same signed-target formulation and therefore are not direct baselines for our theoretical framework. Nevertheless, among methods using the same RF backbone without additional representation alignment, Signed RF achieves comparable or lower FID. We report REPA [76] separately because it introduces vision foundation-model alignment and uses a substantially larger training budget. Our RF backbone is trained for 320 epochs without such alignment, whereas REPA is trained for 800 epochs with VFM alignment. Despite this difference, Signed RF achieves a comparable FID with substantially fewer sampling steps: 1.41 at 64 NFE, compared with 1.42 at 250 NFE for REPA.

Implementation details. We use the pretrained ImageNet 256×256 model released with the Rectified Flow implementation [29] and perform sampling in the latent space of a KL-16 autoencoder.

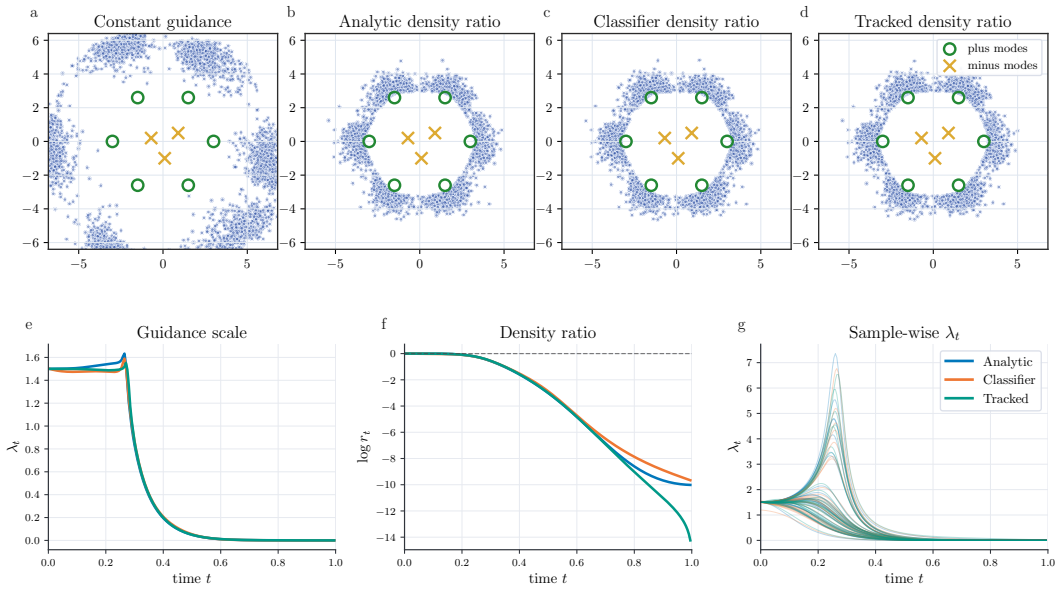


Figure 14: Additional density-ratio estimation ablation in a different setup.

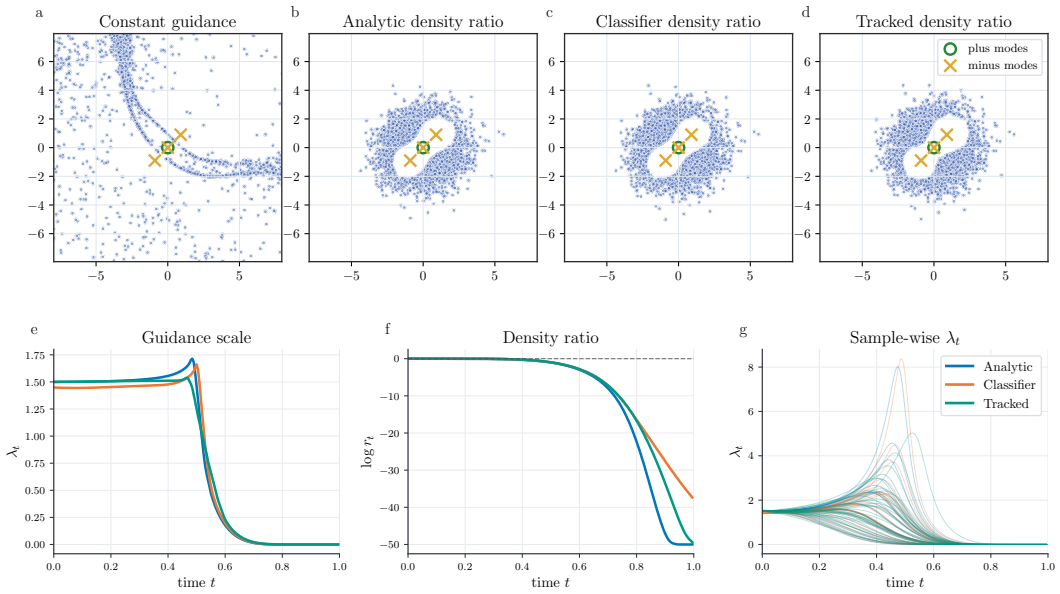


Figure 15: Additional density-ratio estimation ablation in a different setup.

Table 1: System-level ImageNet-256 comparison with related guidance baselines.

NFE	Method	FID ↓	Prec. ↑	Rec. ↑
Same RF backbone (320 epochs; no VFM alignment)				
16	CFG	2.38	0.828	0.566
	ADG	2.32	0.827	0.570
	CFG++	2.62	0.842	0.572
	MG	1.85	0.820	0.600
	Signed RF	**1.82**	0.801	0.615
32	CFG	1.87	0.815	0.598
	ADG	2.00	0.835	0.582
	CFG++	1.71	0.809	0.624
	MG	1.71	0.810	0.620
	Signed RF	**1.51**	0.800	0.629
64	CFG	1.73	0.821	0.601
	ADG	1.85	0.793	0.634
	CFG++	3.39	0.838	0.573
	MG	1.60	0.810	0.620
	Signed RF	**1.41**	0.797	0.627
VFM-aligned representation (800 epochs)				
250	REPA	1.42	0.800	0.650

At each Euler step, evaluating the model with class label c produces the conditional velocity $v_t^+(\boldsymbol{x}_t, c)$, while evaluating it with the null class token $c_\varnothing$ produces the unconditional velocity $v_t^-(\boldsymbol{x}_t)$.

The CFG baseline uses the constant-guidance field $\hat{v}_t^{\mathrm{CFG}}(\boldsymbol{x}_t, c) = v_t^+(\boldsymbol{x}_t, c) + \omega\big(v_t^+(\boldsymbol{x}_t, c) - v_t^-(\boldsymbol{x}_t)\big)$. Signed RF uses the same guidance direction but replaces the constant scale ω with the state-dependent scale $\lambda_t^\alpha(\boldsymbol{x}_t, c)$.

The direction $v_t^+ - v_t^-$ serves two related purposes. First, it strengthens class conditioning by moving away from directions shared across the broader unconditional distribution. Second, because the unconditional branch averages over many classes and modes, it acts as a smoother reference field. Extrapolating away from this reference can compensate for the smoothing effect of neural velocity predictions [8]. Thus, the unconditional branch is not treated as an undesirable distribution to be removed; instead, it provides a reference direction for steering samples toward better class alignment and perceptual quality.

We estimate the density ratio using a separate timestep- and class-conditioned ViT classifier [77] trained in the same latent space. Following the classifier-based estimator in Sec. 2.3, the classifier $p_t^\phi(y \mid \boldsymbol{x}_t, c)$ predicts the branch label $y \in \{-, +\}$. With balanced positive and negative sampling,

$$\hat{r}_t(\boldsymbol{x}_t, c) = \frac{p_t^\phi(- \mid \boldsymbol{x}_t, c)}{p_t^\phi(+ \mid \boldsymbol{x}_t, c)}, \qquad \lambda_t^\alpha(\boldsymbol{x}_t, c) = \frac{\alpha\, \hat{r}_t(\boldsymbol{x}_t, c)}{(1+\alpha) - \alpha\, \hat{r}_t(\boldsymbol{x}_t, c)}.$$

The classifier training examples are designed to capture both class misalignment and poor sample quality. Positive endpoints $\boldsymbol{x}_1^+$ are ImageNet latents paired with their ground-truth class labels. Negative endpoints $\boldsymbol{x}_1^-$ include class-aligned but low-quality samples generated by an earlier model checkpoint without CFG, as well as samples paired with an incorrect conditioning label. The former teaches the classifier to identify over-smoothed or low-quality states, while the latter teaches it to detect class-inconsistent states.

For each positive–negative pair, we sample $t \sim \mathrm{Unif}[0, 1]$ and $\boldsymbol{x}_0 \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$, and construct

$$\boldsymbol{x}_t^+ = (1-t)\boldsymbol{x}_0 + t\boldsymbol{x}_1^+, \qquad \boldsymbol{x}_t^- = (1-t)\boldsymbol{x}_0 + t\boldsymbol{x}_1^-.$$

The two branches share the same $(\boldsymbol{x}_0, t)$ within each pair, reducing variation unrelated to the endpoint distributions.

A single classifier is shared across all ImageNet classes and receives the timestep and class label as conditioning. Positive samples are paired with their ground-truth labels. For negative samples,

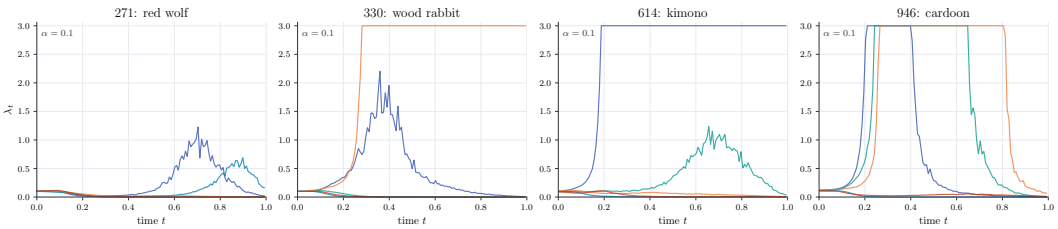


Figure 16: Effective guidance scale λ_t^α along trajectories for four ImageNet classes. Each color corresponds to the same initial seed shared across panels. Despite identical seeds, the guidance profiles differ substantially across classes, illustrating the nontrivial and class-dependent behavior of the guidance.

the condition is set to the associated sample class with probability 0.9 and to a randomly selected different class with probability 0.1. The classifier has 12 Transformer layers, a hidden dimension of 768, 12 attention heads, and AdaLN-style timestep and class conditioning [78], giving it roughly the capacity of a `DiT-B` model. We train it from scratch using AdamW [79], with a learning rate of 2×10^{-4}, weight decay 0.01, global batch size 1024, and EMA rate 0.999.

During sampling, we apply the stabilization procedures described in Sec. 2.3. Specifically, we clip $\log \hat{r}_t$ to $[-20, 20]$, lower-bound the denominator by $\max\big((1+\alpha) - \alpha\hat{r}_t,\, 10^{-3}\big)$, and cap the effective guidance scale according to $\lambda_t^\alpha \leftarrow \min(\lambda_t^\alpha, \lambda_{\max})$. All ImageNet comparisons use $50{,}000$ class-balanced samples. We use the same initial Gaussian seeds across methods and NFEs, and report the best-FID operating point from the corresponding sweep over α or ω.

Runtime and computational cost. We train the density-ratio classifier for 250K steps on 8 NVIDIA GH200 GPUs using bf16 mixed precision and `torch.compile`. Training takes approximately 10 hours. For comparison, the pretrained XL flow backbone requires approximately 60 hours on 16 GPUs. Thus, the ratio classifier adds substantially less training cost than the generative backbone.

At inference time, the main overhead of Signed RF comes from evaluating the ratio classifier at each Euler step; the scalar computations required to obtain $\hat{r}_t$ and λ_t^α are negligible in comparison. We measure this overhead on a single NVIDIA GH200 GPU using bf16 mixed precision, `torch.compile`, batch size 100, and 64 Euler steps. We exclude checkpoint loading, compilation warmup, VAE decoding, and image writing from the timing. Under this setup, CFG generates 13.85 images/s, corresponding to 72.22 ms/image, with 4.57 GB of peak GPU memory. Signed RF generates 12.44 images/s, corresponding to 80.38 ms/image, with 5.05 GB of peak GPU memory. This amounts to an 11.3% increase in sampling time, or 8.16 additional ms/image, together with 0.48 GB of additional peak memory. Overall, the additional training and inference costs of Signed RF are small relative to those of the underlying generative model.

Stability analysis. We first evaluate sensitivity to the training stage of the density-ratio classifier. We use checkpoints collected after 100K, 160K, 200K, and 250K updates. For each checkpoint and NFE, we sweep α and report the operating point with the best FID in Table 2. Performance varies only slightly across checkpoints. The FID ranges are 1.824–1.845 at 16 NFE, 1.536–1.550 at 32 NFE, and 1.436–1.473 at 64 NFE. IS, precision, and recall exhibit similarly small variations. These results indicate that Signed RF does not require the ratio classifier to be stopped at a narrowly selected training checkpoint.

We next study sensitivity to the guidance cap $\lambda_{\max}$, which limits the effective guidance when $(1+\alpha) - \alpha\hat{r}_t$ becomes small because of ratio-estimation or numerical error. Using the 250K classifier and fixing $\alpha = 0.08$, Table 3 shows that moderate caps produce similar results. At 32 NFE, $\lambda_{\max} \in \{3, 5, 10\}$ gives FID 1.518, 1.510, and 1.545, respectively; at 64 NFE, the corresponding FIDs are 1.413, 1.431, and 1.477. Fig. 16 visualizes the effective guidance scale along trajectories for several classes using shared initial seeds. The guidance profiles vary across both time and class, confirming that Signed RF applies nontrivial state-dependent guidance rather than a fixed schedule. Overall, Signed RF is robust to both the classifier training checkpoint and the choice of a moderate guidance cap.

Table 2: ImageNet classifier-checkpoint stability. All runs use $\lambda_{\max}$ fixed to 10.

NFE	Step	α	FID ↓	IS ↑	Prec. ↑	Rec. ↑
16	100K	0.10	1.845	299.7	0.796	0.615
	160K	0.10	1.827	295.2	0.802	0.616
	200K	0.10	1.824	296.8	0.801	0.615
	250K	0.08	1.827	293.3	0.794	0.611
32	100K	0.10	1.536	286.6	0.795	0.632
	160K	0.10	1.550	290.7	0.800	0.622
	200K	0.10	1.549	294.3	0.800	0.626
	250K	0.07	1.536	290.5	0.794	0.635
64	100K	0.10	1.436	278.9	0.792	0.638
	160K	0.10	1.473	287.3	0.799	0.627
	200K	0.10	1.471	291.4	0.799	0.627
	250K	0.07	1.469	289.1	0.792	0.633

Table 3: ImageNet guidance-cap stability. We use the 250K classifier, fix α to 0.08, and vary $\lambda_{\max}$.

NFE	$\lambda_{\max}$	FID ↓	IS ↑	Prec. ↑	Rec. ↑
16	3	1.918	275.8	0.793	0.615
	5	1.845	289.0	0.797	0.612
	10	1.827	293.3	0.794	0.611
32	3	1.518	278.4	0.798	0.630
	5	1.510	288.6	0.800	0.629
	10	1.545	293.6	0.796	0.625
64	3	1.413	278.2	0.797	0.627
	5	1.431	286.6	0.797	0.627
	10	1.477	292.0	0.795	0.632

C.3 Additional Analytic Negative-Flow Details

Analytic flow under a Gaussian source. Let π_1 be the empirical distribution supported on $\{\boldsymbol{x}_1^{(i)}\}_{i=1}^N \subset \mathbb{R}^D$:

$$\pi_1(\boldsymbol{x}) = \frac{1}{N}\sum_{i=1}^{N} \delta_{\boldsymbol{x}_1^{(i)}}(\boldsymbol{x}).$$

Consider the RF interpolation $\boldsymbol{X}_t = t\boldsymbol{X}_1 + (1-t)\boldsymbol{X}_0, \qquad \boldsymbol{X}_0 \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$. The corresponding velocity field is $v_t(\boldsymbol{x}) = \mathbb{E}[\boldsymbol{X}_1 - \boldsymbol{X}_0 \,|\, \boldsymbol{X}_t = \boldsymbol{x}]$. For $0 \le t < 1$, conditioned on $\boldsymbol{X}_1 = \boldsymbol{x}_1^{(i)}$, the intermediate state follows

$$\boldsymbol{X}_t \sim \mathcal{N}\Big(t\boldsymbol{x}_1^{(i)}, (1-t)^2\boldsymbol{I}\Big).$$

Averaging over the empirical data points therefore gives the time-t marginal

$$\pi_t(\boldsymbol{x}) = \frac{1}{N}\sum_{i=1}^{N} \mathcal{N}\Big(\boldsymbol{x};\, t\boldsymbol{x}_1^{(i)}, (1-t)^2\boldsymbol{I}\Big). \tag{17}$$

Given $\boldsymbol{X}_t = \boldsymbol{x}$, the posterior probability that the endpoint is $\boldsymbol{x}_1^{(i)}$ is

$$w_i(\boldsymbol{x}, t) \coloneqq \mathbb{P}\Big(\boldsymbol{X}_1 = \boldsymbol{x}_1^{(i)} \,\Big|\, \boldsymbol{X}_t = \boldsymbol{x}\Big),$$

where

$$w_i(\boldsymbol{x}, t) = \frac{\exp\Big(-\|\boldsymbol{x} - t\boldsymbol{x}_1^{(i)}\|^2 / \big(2(1-t)^2\big)\Big)}{\displaystyle\sum_{j=1}^{N} \exp\Big(-\|\boldsymbol{x} - t\boldsymbol{x}_1^{(j)}\|^2 / \big(2(1-t)^2\big)\Big)}. \tag{18}$$

Using $\boldsymbol{X}_0 = (\boldsymbol{X}_t - t\boldsymbol{X}_1)/(1-t)$, we obtain

$$\begin{aligned} v_t(\boldsymbol{x}) &= \mathbb{E}[\boldsymbol{X}_1 - \boldsymbol{X}_0 \mid \boldsymbol{X}_t = \boldsymbol{x}] \\ &= \mathbb{E}\left[\frac{\boldsymbol{X}_1 - \boldsymbol{x}}{1-t} \,\middle|\, \boldsymbol{X}_t = \boldsymbol{x}\right] \\ &= \sum_{i=1}^{N} w_i(\boldsymbol{x}, t)\frac{\boldsymbol{x}_1^{(i)} - \boldsymbol{x}}{1-t}. \end{aligned} \tag{19}$$

Thus, the exact velocity is obtained by moving toward the posterior-weighted average of the empirical endpoints.

For numerical evaluation, the corresponding log marginal can be written as

$$\log \pi_t(\boldsymbol{x}) = \log \sum_{i=1}^{N} \exp\left(-\frac{\|\boldsymbol{x} - t\boldsymbol{x}_1^{(i)}\|^2}{2(1-t)^2}\right) - \log N - \frac{D}{2}\log(2\pi) - D\log(1-t). \tag{20}$$

These density, weight, velocity, and log-density formulas apply for $t < 1$; at $t = 1$, the terminal law is the empirical measure π_1.

PyTorch implementation. The implementation below evaluates the analytic velocity for a reference set of data latents with shape $[N, C, H, W]$ and a batch of intermediate states $\boldsymbol{x}_t$ with shape $[B, C, H, W]$. After flattening the latents, it computes the $B \times N$ matrix of squared distances $\|\boldsymbol{x}_t - t\boldsymbol{x}_1^{(i)}\|^2$. A softmax over the reference points gives the posterior weights in Eq. 18, from which the posterior-weighted endpoint mean and the velocity are obtained. When requested, the same unnormalized log weights are reused to evaluate $\log \pi_t(\boldsymbol{x}_t)$ through a log-sum-exp operation.

PyTorch implementation of the Gaussian analytic flow

```python
class GaussianAnalyticFlow:
    def __init__(self, data):
        # data: [N, C, H, W], e.g., ImageNet autoencoder latents.
        self.data_shape = tuple(data.shape[1:])
        self.data = data.detach().flatten(1).float()
        self.data_sq_norm = self.data.square().sum(dim=1)

    @torch.no_grad()
    def velocity_and_logp(self, x_t, t, return_logp=False):
        # x_t: [B, C, H, W]
        # t: scalar tensor or tensor of shape [B]
        batch_size = x_t.shape[0]
        if t.ndim == 0:
            t = t.expand(batch_size)
        t = t.to(device=x_t.device, dtype=torch.float32)

        x_shape = x_t.shape
        x_flat = x_t.flatten(1).float()
        _, latent_dim = x_flat.shape

        data = self.data.to(
            device=x_t.device,
            dtype=x_flat.dtype,
        )
        data_sq_norm = self.data_sq_norm.to(
            device=x_t.device,
            dtype=x_flat.dtype,
        )
        num_data = data.shape[0]

        # Compute ||x_t - t x_1^{(i)}||^2 for every batch--data pair.
        x_sq_norm = x_flat.square().sum(dim=1)
        x_data_inner_product = x_flat @ data.T
        sq_dist = (
            x_sq_norm[:, None]
            - 2.0 * t[:, None] * x_data_inner_product
            + t.square()[:, None] * data_sq_norm[None, :]
        )
```

```
        one_minus_t = (1.0 - t).clamp_min(1e-12)
        log_weights = (
            -0.5
            * sq_dist
            / one_minus_t.square()[:, None]
        )
        weights = torch.softmax(log_weights, dim=1)

        # Posterior mean E[X_1 | X_t = x_t].
        endpoint_mean = weights @ data
        velocity_flat = (
            endpoint_mean - x_flat
        ) / one_minus_t[:, None]
        velocity = velocity_flat.reshape(x_shape)

        if not return_logp:
            return velocity, None

        log_mixture = (
            torch.logsumexp(log_weights.float(), dim=1)
            - math.log(float(num_data))
        )
        log_normalizer = (
            -0.5 * latent_dim * math.log(2.0 * math.pi)
            - latent_dim * torch.log(one_minus_t.float())
        )
        logp = log_mixture + log_normalizer

        return velocity, logp
```

ImageNet anti-memorization sampler. For the anti-memorization experiments in Sec. 3.3, we construct a separate analytic negative branch for each ImageNet class. Given the training latents of class c,

$$\mathcal{D}_c = \{\boldsymbol{x}_1^{(i)}\}_{i=1}^{N_c} \subset \mathbb{R}^D, \qquad \pi_{1,c}^-(\boldsymbol{x}) = \frac{1}{N_c}\sum_{i=1}^{N_c} \delta_{\boldsymbol{x}_1^{(i)}}(\boldsymbol{x}),$$

Eqs. 17, 19, and 20 provide the corresponding marginal, negative velocity v_t^-, and log-density $\log \pi_{t,c}^-$ in closed form. We augment each class-specific reference set $\mathcal{D}_c$ with horizontally flipped training latents.

The positive velocity v_t^+ is provided by a pretrained class-conditional ImageNet flow model using CFG with $\omega = 1.0$. Because the positive-branch density is not available in closed form, we use the online-tracked ratio described in Sec. 2.3. For each trajectory, we maintain $u_t \approx \log r_t(\boldsymbol{x}_t) = \log \frac{\pi_{t,c}^-(\boldsymbol{x}_t)}{\pi_{t,c}^+(\boldsymbol{x}_t)}$, with $u_0 = 0$, since both branches share the same Gaussian source.

At each Euler step, we evaluate the pretrained positive velocity v_t^+ and the analytic negative velocity v_t^-. We then recover the branch scores using the Gaussian-source RF identity $s_t^\pm(\boldsymbol{x}) = \frac{t\, v_t^\pm(\boldsymbol{x}) - \boldsymbol{x}}{1-t}$ [24], and update u_t according to Eq. 14. The only required divergence term is $\nabla\cdot\left(v_t^+ - v_t^-\right)$, which we estimate using the Hutchinson trace estimator [25]. We apply the same ratio clipping, denominator lower bound, and guidance cap used in the ImageNet-256 experiments above.

Evaluation protocol. For the targeted memorization stress test, we first generate a pool of 50K base-model samples for each selected class and compute the nearest-neighbor distance from each sample to the corresponding class-specific training set. We rank samples using the L_2 distance between SSCD embeddings, which is more sensitive to copied visual details than pixel-space or latent-space distances in our ImageNet experiments.

For each class, we retain the 600 highest-risk initial seeds, defined as those producing the smallest SSCD L_2 nearest-neighbor distances under the base model. Starting from exactly these same initial noise samples, we then rerun the base model, the SPELL baselines, and Data Repulsive Flow. This shared-seed protocol ensures that differences in memorization are attributable to the sampling method rather than to the selected initial states.

Table 4: Additional per-class α sweeps for ImageNet anti-memorization. For each class, we report nearest-neighbor SSCD L_2 and latent L_2 distances on the same 600 high-risk initial seeds used by the base model and Data Repulsive Flow. Larger values indicate lower similarity to the nearest training example. The seeds are selected by base-model SSCD L_2; latent L_2 is reported as an additional metric. For each metric we report the lower-tail P_{05}, P_{10}, and P_{25} quantiles and the mean.

Class	Method	α	SSCD $L_2 \uparrow$				Latent $L_2 \uparrow$			
			P_{05}	P_{10}	P_{25}	Mean	P_{05}	P_{10}	P_{25}	Mean
151 (Chihuahua)	Base	–	0.921	0.974	1.017	1.033	47.27	51.66	58.39	62.98
	Ours	0.1	0.928	0.976	1.020	1.040	48.38	51.96	58.95	63.22
	Ours	1.0	0.944	0.991	1.033	1.068	48.83	53.83	60.35	64.21
	Ours	5.0	0.983	1.013	1.063	1.107	54.28	57.04	61.66	65.23
	Ours	7.5	**0.993**	**1.023**	**1.073**	**1.119**	**54.50**	**57.72**	**61.93**	**65.59**
248 (Eskimo dog, husky)	Base	–	0.921	0.944	0.982	0.993	54.88	56.99	60.59	63.81
	Ours	1.0	0.939	0.962	0.992	1.024	56.31	58.15	61.29	64.28
	Ours	3.0	0.956	0.975	1.005	1.045	57.71	58.78	61.62	64.63
	Ours	5.0	0.958	0.979	1.012	1.056	57.69	59.28	61.87	64.99
	Ours	7.5	**0.964**	**0.987**	**1.020**	**1.066**	**57.99**	**59.91**	**62.36**	**65.36**
555 (fire engine, fire truck)	Base	–	0.967	0.978	0.998	1.010	54.09	55.58	57.54	59.69
	Ours	0.1	0.968	0.980	0.998	1.013	54.27	55.62	57.54	59.72
	Ours	1.0	0.969	0.983	1.002	1.031	54.42	55.97	57.67	59.80
	Ours	5.0	0.978	0.997	1.026	1.070	55.19	56.24	58.27	60.41
	Ours	7.5	**0.983**	**1.004**	**1.039**	**1.081**	**55.49**	**56.55**	**58.54**	**60.72**
805 (soccer ball)	Base	–	0.854	0.873	0.904	0.929	48.96	50.17	52.80	59.07
	Ours	0.1	0.852	0.875	0.911	0.936	49.17	50.50	53.05	59.40
	Ours	1.0	0.865	0.885	0.922	0.977	50.15	51.64	54.47	60.62
	Ours	5.0	0.890	0.919	0.968	1.050	51.45	53.91	57.39	62.82
	Ours	7.5	**0.902**	**0.939**	**0.993**	**1.073**	**52.67**	**54.44**	**58.22**	**63.78**
975 (lakeside, lakeshore)	Base	–	0.949	0.992	1.042	1.061	40.69	43.14	46.66	51.38
	Ours	0.1	0.964	1.004	1.051	1.079	42.06	44.01	47.31	51.85
	Ours	1.0	0.981	1.036	1.079	1.136	43.03	45.05	48.66	52.98
	Ours	5.0	1.025	1.065	1.149	1.186	43.99	45.88	49.70	53.33
	Ours	7.5	**1.039**	**1.089**	**1.164**	**1.195**	**44.23**	**45.94**	**50.05**	**53.64**

We report the 5th, 10th, and 25th percentiles, together with the mean, of the SSCD L_2 nearest-neighbor distance. Larger values indicate that the generated samples are farther from the training set and therefore exhibit less training-data similarity. For the standard generation-quality evaluation, we instead generate 50K class-balanced samples from random initial noise, while sharing the same initial seeds across all compared methods.

Table 4 extends the class-248 results from the main paper to five ImageNet classes and several values of α. Across all evaluated classes, increasing α consistently raises the lower-tail nearest-neighbor distances. This shows that Data Repulsive Flow provides a smooth and interpretable control over the strength of training-data repulsion, rather than exhibiting an abrupt change at a particular setting.

We additionally report nearest-neighbor distances in the autoencoder latent space. These distances are not used to select the high-risk seeds and therefore serve as an independent evaluation metric. Their consistent increase supports the conclusion that the observed effect is not specific to the SSCD L_2 metric used for seed selection and primary reporting.

Runtime and memory cost. The analytic negative branch requires no additional training. For a batch of size B, a reference set of size N, and latent dimension D, its inference cost is dominated by computing the state–reference Gram matrix and the responsibility-weighted endpoint average. Both operations scale as $\mathcal{O}(BND)$. The memory cost consists primarily of the reference matrix, which requires $\mathcal{O}(ND)$ storage, and the temporary $B \times N$ matrices of logits and posterior responsibilities.

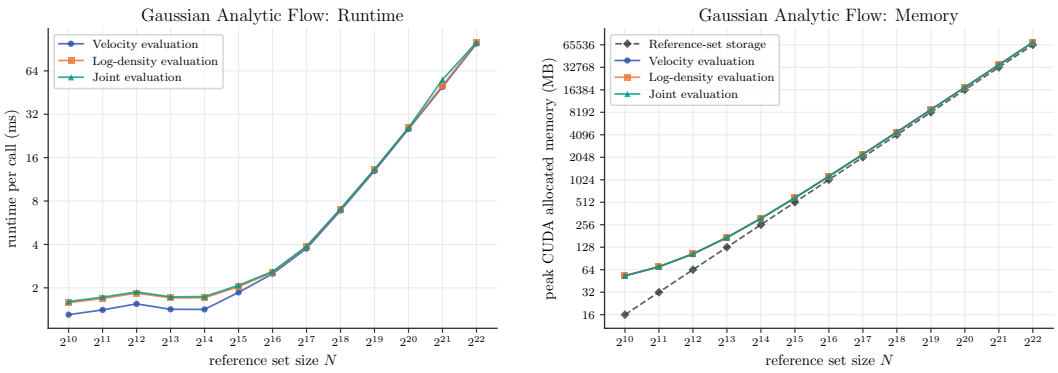


Figure 17: Runtime and peak CUDA memory of the vectorized Gaussian analytic flow on an NVIDIA GH200 GPU. The benchmark sweeps the reference-set size N at fixed ImageNet latent dimension $D = 4096$. Joint velocity and log-density evaluation is nearly as cheap as either quantity alone because the same squared-distance logits are reused.

In our class-conditional ImageNet experiments, each class contains approximately 1.3K reference latents, or 2.6K after horizontal-flip augmentation. At this scale, the analytic branch is small relative to the pretrained flow backbone in both runtime and memory.

Fig. 17 benchmarks the analytic branch on a single NVIDIA GH200 GPU with the ImageNet latent dimension $D = 4096$. Velocity-only, log-density-only, and joint evaluation have nearly identical runtime because all three reuse the same $B \times N$ distance logits. At the class-level reference-set sizes used in our experiments, each evaluation takes only a few milliseconds and requires substantially less than 0.1 GB of additional memory.

Both runtime and reference storage grow approximately linearly with N. At $N = 2^{18}$, evaluation remains below 8 ms. At $N = 2^{22}$, it takes on the order of 10^2 ms, while storing the reference set in fp32 requires 64 GB. These measurements confirm that the class-specific analytic branch is not a computational bottleneck in our ImageNet anti-memorization experiments. For substantially larger protected datasets, the reference set can be processed in chunks while incrementally accumulating the log-sum-exp normalizer and responsibility-weighted endpoint average. This keeps memory bounded at the cost of additional passes over the reference set.

C.4 Additional Concept Suppression Details

For the main-text concept-suppression examples, the positive prompt defines the desired visual content and the negative prompt defines the concept or attribute to suppress. We use the following prompt pairs verbatim.

1. **Positive:** A high-aesthetic anime portrait, upper body, night street background, expressive eyes, polished shading, clean linework, masterpiece, best quality
   Negative: lowres, blurry, bad anatomy, bad hands, missing fingers, extra digits, jpeg artifacts, cropped, low quality, watermark, signature
2. **Positive:** A clean contemporary house, strong geometric lines, large glass panels, minimalist architecture photography, crisp daylight, centered framing, unobstructed surroundings, pure architectural focus
   Negative: trees, bushes, leaves, greenery, vines, overgrown plants, grass, lawn, meadow, turf, wildflowers, landscaped garden
3. **Positive:** A single large paper crane flying low over a quiet lake at dawn, close-up, centered composition, the crane fills much of the frame, poetic, minimalist, soft mist
   Negative: origami, folded paper, craft, crease
4. **Positive:** A solitary young woman sitting by a large train window during rain, soft city lights outside, reflective atmosphere, cinematic portrait
   Negative: sad, crying, depressed, grief, despair, tears, misery

5. **Positive:** A cheerful cartoon mouse with perfectly round black ears, red shorts, yellow shoes, white gloves, simple face, vintage mascot design, standing confidently against a clean pastel background
   Negative: Mickey Mouse, Disney character
6. **Positive:** A pale angel standing silently in a winter garden, marble-like skin, white wings dusted with snow, stillness, ethereal realism
   Negative: stone statue, sculpture pedestal, carved marble, museum artifact
7. **Positive:** A world-class scientist presenting a breakthrough discovery in a sleek research lab, confident, inspiring, magazine-cover composition
   Negative: old white man, male stereotype, lab coat cliche, Einstein-like appearance
8. **Positive:** A visionary CEO standing in a glass office overlooking a futuristic city, calm authority, elegant natural light
   Negative: middle-aged white man in dark suit, corporate cliche, Wall Street stereotype
9. **Positive:** A dynamic cinematic shot of a lean teenage superhero crouching high above a city, wearing an iconic red-and-blue skin-tight suit defined by black webbing lines, a bold spider emblem on the chest, and oversized white pointed eye lenses on a full-face mask, athletic, agile, youthful, with a strong sense of motion and classic urban superhero energy
   Negative: Spider-Man, Marvel, Peter Parker, superhero franchise
10. **Positive:** Sweeping tulip fields under soft spring sunlight, wide-angle landscape view, richly layered flowers, fresh air, elegant color harmony, cinematic countryside photography
    Negative: red, crimson, scarlet, ruby tones

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and Sec. 2.2 state the signed-measure formulation, the theoretical zone-decomposition guarantee, and the empirical claims (ImageNet fidelity–diversity, anti-memorization, and nudity prevention), all of which are substantiated in Sec. 3 and Appendix B.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations are discussed inline where they arise, including the assumptions underlying the exact Signed RF analysis in Sec. 2.2 and Appendix B, the approximate nature of the practical implementation in Sec. 2.3, and the scope of the empirical evaluation given the computational cost of large-scale experiments.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Sec. 2.2 summarizes the sampling result and its scope, while Theorem B.4, the full assumption set, complete proofs, and supporting derivations are provided in Appendix B.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Sec. 3 together with Sec. C.2, Sec. C.3, and Sec. C.4 document the architectures, datasets, training and sampling hyperparameters, classifier setup, evaluation protocol, and verbatim prompts used for the main experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We use publicly available datasets (ImageNet, D4RL PointMaze, ViSU-Text, Ring-A-Bell, COCO) and publicly released backbones (the Rectified Flow ImageNet model and Stable Diffusion 3.5), and Sec. C.2 and Sec. C.3 provide the full algorithmic specification and a reference PyTorch implementation of the analytic negative flow sufficient to reproduce the main results. We will release the code and trained checkpoints.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Optimizers, learning rates, batch sizes, training steps, EMA settings, classifier architecture, sampler configurations, etc., are reported in Sec. C.2 and Sec. C.3; the 2D toy details are reported in Sec. C.1.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

Justification: We report single-run point estimates, as each setting requires costly backbone training and large-sample evaluation (50K samples for FID/IS/P/R; thousands of safety prompts for ASR/TR/CLIP/AES). This follows standard reporting practice in flow- and diffusion-based generative modeling.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Sec. C.2 reports the GPU type (NVIDIA GH200), training duration, sampler throughput, and peak memory for the ImageNet experiments, and Sec. C.3 reports runtime and memory for the analytic negative branch.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research uses publicly available datasets and pretrained models, involves no human subjects, and the safety and IP-protection applications in Sec. 3 are designed to reduce, rather than amplify, harmful outputs of generative models.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.

- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The applications studied in Sec. 3, namely anti-memorization, nudity prevention on adversarial prompts, and concept suppression for IP protection, directly target positive societal use of generative models, while the negative-conditioning mechanism could in principle be repurposed to suppress benign rather than harmful content, a dual-use consideration we acknowledge.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: We do not release any new foundation model or scraped dataset. The ImageNet experiments use the public ImageNet dataset, which is widely regarded as safe for research use. The safety experiments adapt the publicly released Stable Diffusion 3.5 and Z-Image checkpoints via lightweight LoRA branches, so the safeguards already attached to those base models (license terms, built-in safety filters, usage guidelines) carry over to our adapted checkpoints. Moreover, the proposed method is itself a safeguarding mechanism: the negative-conditioning branches in Sec. 3 are designed to suppress nudity on adversarial prompts, thereby strengthening rather than weakening the safety profile of the underlying generators.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.

- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All assets are cited at their point of use in Sec. 3 and used under their terms: ImageNet (ImageNet terms of access, non-commercial research use); D4RL PointMaze (datasets CC BY 4.0, code Apache 2.0), accessed through Minari (MIT); ViSU-Text (CC BY-NC 4.0); Ring-A-Bell (MIT); the PyTorch RectifiedFlow codebase and its ImageNet checkpoint (MIT); Stable Diffusion 3.5 Medium (Stability AI Community License); Z-Image (Apache 2.0); and MS-COCO (annotations CC BY 4.0, images under the Flickr Terms of Use).

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The proposed Signed RF method, the density-ratio classifier, the analytic negative flow, and the LoRA-based safety branches are documented in Sec. 2.3 and Sec. C.2–C.4, including a reference PyTorch implementation of the analytic flow.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or research with human subjects; all experiments use public datasets and pretrained models.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve human subjects, so IRB review is not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs are not part of the core methodology. Agentic coding assistants were used as general-purpose engineering tools to help implement experiments, but they do not constitute an original component of the proposed method; all algorithmic contributions were developed and verified by the authors.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.