Asymptotically Exact Negative Guidance of Diffusion Models via Positive–Unlabeled Learning

Hirohane Takagi[1] **Masashi Sugiyama**[2,1]
[1] The University of Tokyo, Tokyo, Japan
[2] RIKEN AIP, Tokyo, Japan
htakagi@is.s.u-tokyo.ac.jp sugi@k.u-tokyo.ac.jp

Abstract

Negative guidance of diffusion models generates samples that avoid an undesired condition, specified either by reference samples or by a conditional score alongside an unconditional model. Existing methods inject repulsive terms derived from local information at each sample, so the resulting sampler does not target an explicit density, and its bias cannot be systematically reduced by increasing computation. We instead present a negative-guidance method that samples from an explicit target density, which uniformly handles both reference-sample and score-based specifications and suppresses probability mass near the undesired condition with an avoidance strength that parametrizes the target density itself. Realizing this target requires estimating, at each diffusion timestep, how likely each sample is to belong to the undesired condition. We cast this estimation as a Positive–Unlabeled learning problem that admits lightweight online training along the diffusion trajectory, using condition samples as positives and diffusion particles as unlabeled data. Because the guided particles drift from the unconditional marginal, we further correct the residual mismatch via Sequential Monte Carlo, yielding a sampler that is asymptotically exact in the particle limit and that turns any existing negative-guidance method into its proposal kernel. Empirically, our method approaches the ground-truth target on Gaussian mixtures and Pareto-dominates existing methods on trade-offs between avoidance and generation quality, distributional bias, and diversity, including in an image-generation setting with a pretrained score network.

1 Introduction

Diffusion models have become a dominant framework for sampling from a data density $p(x)$ over objects x. Their conditional generation mechanisms, including classifier guidance (CG; 15) and classifier-free guidance (CFG; 25), are now standard across modalities, including images [42], molecules [27], audio [39], decision-making and robotic control [28, 1]. CFG amplifies a *desired* condition by extrapolating between conditional and unconditional scores, and recent work has characterized the induced density and the resulting fidelity–diversity trade-off [54, 10, 7, 38, 56].

In contrast, *negative* guidance aims to steer generation *away from* an undesired condition. This capability is important for AI safety, where harmful outputs should be suppressed [43, 20, 21], and where memorized or copyrighted training data should not be reproduced [8, 46]. It is also important for controllable generation, where constraints are often more naturally expressed by negation than by positive prompts [2, 3]. Throughout this paper, c denotes the undesired condition and $\neg c$ its complement. The conditional probability $p(\neg c \mid x)$ satisfies $p(\neg c \mid x) = 1 - p(c \mid x)$. Thus c plays the opposite role to the conditioning variable in CFG: the goal is to reduce $p(c \mid x)$ for generated samples while preserving the structure of $p(x)$.

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

Existing negative guidance methods add a CFG-like repulsive term to the sampling dynamics of a fixed pretrained diffusion model, either without retraining the score network [20, 51] or by editing internal representations [21, 5]. The repulsion is constructed from either a pretrained conditional score for $p(x \mid c)$ [34], or a finite set of reference samples from $p(x \mid c)$ [29]. These methods share two structural limitations: the repulsive update uses only information local to each sample, so the resulting sampler does not target an explicit density, and the induced bias cannot be reduced by allocating more computation. We formalize these limitations in Section 2.

We instead formulate negative guidance as sampling from the explicit avoidance target

$$p_{\text{target}}(x) \;\propto\; p(x)\,(1 - p(c \mid x))^{\lambda}, \qquad \lambda \geq 0, \tag{1}$$

where λ is a user-controlled avoidance strength that parametrizes the target density itself: $\lambda = 0$ recovers the unconditional density, and $\lambda \to \infty$ concentrates mass on the support of $\neg c$. When the condition c is a semantic concept rather than a deterministic rule, $p(c \mid x)$ varies smoothly with x, and intermediate λ continuously interpolates between fidelity to $p(x)$ and avoidance of c. The remaining challenge is to sample from (1) using only the information available in negative guidance: an unconditional score for $p(x)$ together with either a conditional score for $p(x \mid c)$ or reference samples from $p(x \mid c)$.

Our approach rests on the observation that estimating $p(c \mid x)$ in this setting can be cast as a Positive–Unlabeled (PU) learning problem [14, 18], with samples associated with class c as positives and unconditional diffusion particles as unlabeled data. We pair an online PU estimator of $p(c \mid x)$ with Sequential Monte Carlo (SMC; 13, 11) targeting (1), which corrects the residual mismatch between guided particles and the target marginal. The resulting sampler is asymptotically exact in the particle limit, in the same spirit as test-time diffusion alignment via SMC [52, 45, 31], while keeping per-step PU updates lightweight.

Our contributions.

- We formulate negative guidance as sampling from the explicit avoidance density (1), realized through the finite-step reverse process of a trained diffusion model, for both reference-sample and conditional-score specifications.
- We design an SMC sampler driven by an online PU estimator of $p(c \mid x)$, prove its asymptotic exactness, and show that any existing negative-guidance method can serve as its proposal kernel.
- Empirically, our sampler matches an oracle diffusion sampler with access to the true score for $p(x \mid \neg c)$ on Gaussian mixtures and Pareto-dominates representative baselines on avoidance–quality trade-offs under both specifications, including image generation with a pretrained score network.

2 Existing Negative Guidance Methods

This section reviews representative negative guidance methods, namely sign-flipped CFG (NegCFG; also analyzed as standard negative prompting by 3), Dynamic Negative Guidance (DNG; 34), and the Safe Denoiser [29]. We then identify structural limitations that motivate our formulation.

Diffusion models. We use a variance-preserving diffusion process [48, 26] with forward kernel

$$q_t(x_t \mid x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t} x_0, (1 - \bar{\alpha}_t) I), \qquad t = 1, \ldots, T, \tag{2}$$

where $\mathcal{N}(x; \mu, \Sigma)$ denotes the Gaussian density in x with mean μ and covariance Σ and I is the identity matrix. This kernel induces noisy marginals $p_t(x_t)$ with $p_T(x_T)$ close to $\mathcal{N}(0, I)$. A pretrained noise predictor $\hat{\epsilon}_\theta(x_t, t)$ parameterizes the unconditional score via

$$\hat{\epsilon}_\theta(x_t, t) \approx -\sqrt{1 - \bar{\alpha}_t}\, \nabla_{x_t} \log p_t(x_t),$$

and, when available, a conditional predictor $\hat{\epsilon}_\theta(x_t, t \mid c)$ analogously parameterizes $\nabla_{x_t} \log p_t(x_t \mid c)$. Tweedie's formula [17] gives the standard denoised estimate

$$\hat{x}_0(x_t) = \frac{x_t - \sqrt{1 - \bar{\alpha}_t}\hat{\epsilon}_\theta(x_t, t)}{\sqrt{\bar{\alpha}_t}} \approx \mathbb{E}[x_0 \mid x_t].$$

DDPM (Denoising Diffusion Probabilistic Models) sampling [26] starts from $x_T \sim \mathcal{N}(0, I)$ and applies the reverse kernels $p_\theta(x_{t-1} \mid x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \sigma_t^2 I)$ for $t = T, T-1, \ldots, 1$. All guidance methods reviewed here modify only the mean μ_θ, leaving the variance $\sigma_t^2 I$ unchanged; this is exploited in Section 3. Full closed-form expressions are deferred to Appendix A.1.

Sign-flipped CFG (NegCFG). NegCFG combines unconditional and conditional noise predictors:

$$\hat{\epsilon}_{\mathrm{CFG}}(x_t, t) \leftarrow \hat{\epsilon}_\theta(x_t, t) - \eta \left(\hat{\epsilon}_\theta(x_t, t \mid c) - \hat{\epsilon}_\theta(x_t, t)\right),$$

and throughout this paper $\eta > 0$ consistently denotes repulsion from condition c, while $\eta < 0$ corresponds to standard CFG that amplifies c. Several variants build on this basic rule with stabilization mechanisms that suppress artifacts arising from score extrapolation, though these mechanisms are tailored to specific applications or architectures [43, 9].

Dynamic Negative Guidance (DNG). DNG [34] starts from the exact score decomposition

$$\nabla_{x_t} \log p_t(x_t \mid \neg c) \leftarrow \nabla_{x_t} \log p_t(x_t) + \eta \, \nabla_{x_t} \log\big(1 - p_t(c \mid x_t)\big), \tag{3}$$

which requires the global posterior $p_t(c \mid x_t)$. Applying Bayes' rule yields a CFG-style repulsion scaled by the posterior odds:

$$\nabla_{x_t} \log(1 - p_t(c \mid x_t)) = -\tfrac{p_t(c|x_t)}{1-p_t(c|x_t)} \big(\nabla_{x_t} \log p_t(x_t \mid c) - \nabla_{x_t} \log p_t(x_t)\big).$$

Since $p_t(c \mid x_t)$ is unavailable during sampling, DNG estimates it recursively via a local Gaussian approximation to the reverse transition [37]:

$$\log p_{t-1}(c \mid x_{t-1}) \leftarrow \log p_t(c \mid x_t) - \tfrac{1}{2\sigma_t^2} \left(\|x_{t-1} - \mu_\theta(x_t, t \mid c)\|^2 - \|x_{t-1} - \mu_\theta(x_t, t)\|^2\right), \tag{4}$$

introducing an approximation bias whose magnitude has no known bound. Appendix A.4 supplements this point with the conditions under which DNG becomes exact and how they differ from our setting.

Reference-particle negative guidance. NegCFG and DNG both require a pretrained conditional score $\hat{\epsilon}_\theta(x_t, t \mid c)$, which does not exist for user-specified conditions such as copyrighted or memorized content [29]; in such cases, the condition is instead provided as a finite reference set $\{x_c^i\}_{i=1}^N$. The Safe Denoiser [29] addresses this regime by modifying the denoised estimate itself:

$$\hat{x}_{\mathrm{safe}}(x_t) = \hat{x}_0(x_t) + \frac{\eta}{N} \sum_{i=1}^{N} q_t(x_t \mid x_c^i) \left(\hat{x}_0(x_t) - \frac{\sum_{k=1}^N x_c^k q_t(x_t \mid x_c^k)}{\sum_{j=1}^N q_t(x_t \mid x_c^j)}\right),$$

where $q_t(x_t \mid x_c^i)$ is the forward kernel (2) evaluated as a likelihood (not a sampling density) and $\eta > 0$ controls the repulsion strength. Geometrically, the bracketed term pushes $\hat{x}_0(x_t)$ away from the kernel-weighted mean of the references, and the prefactor amplifies the repulsion when x_t lies near the reference region. In practice, η and the time window over which guidance is applied must be tuned per condition [29, 30], navigating an implicit avoidance–quality trade-off. Shielded Diffusion [32] takes a similar reference-particle approach but sparsifies the repulsion, activating it only when $\hat{x}_0(x_t)$ falls within a fixed radius of the reference region.

Structural limitations. These methods are lightweight and effective in practice, but share a structural limitation: each replaces the global posterior $p_t(c \mid x_t)$ in (3) with a single-particle surrogate, namely a conditional score, a Gaussian-transition ratio, or nearby reference particles. The repulsion-strength hyperparameter exposed by each method thus carries no density-level meaning, and the induced bias cannot be bounded or reduced by additional computation. Section 3 instead estimates $p_t(c \mid x_t)$ from multiple particles via online PU learning and corrects residual bias with SMC, targeting (1) with λ as the sole density-level knob.

3 Proposed Method: PU Steering

The limitations above have a common root: without a global estimate of $p_t(c \mid x_t)$, no per-particle update can target (1) as an explicit density. *PU Steering* closes this gap by combining (i) a tempered SMC path toward (1), (ii) an *online* PU classifier $f_t \approx p_t(c \mid x_t)$ fit from the running particles, and (iii) a particle-density correction that makes those particles unbiased unlabeled data for the PU step. Any existing negative-guidance method plugs in as the proposal kernel, and an analysis-friendly form of the sampler is asymptotically exact for (1) (Section 4). Section 3.1 presents the meta-algorithm; Section 3.2 and Section 3.3 specialize it to the reference-sample and score-based regimes.

3.1 Common Framework

PU Steering combines three components, described in turn below: an SMC backbone, a PU posterior estimator that yields f_t online, and an importance-weight correction that bridges them.

SMC backbone. SMC targets a sequence $\gamma_T, \gamma_{T-1}, \ldots, \gamma_0$ in turn via a weighted particle system, alternating at each step a *proposal* that moves each particle, and a *weight update* that retargets it to the next γ_t. Following test-time alignment of diffusion models via SMC [52, 31] (recalled in Appendix A.3), we deploy SMC along the reverse diffusion with the tempered path

$$\gamma_t(x_t) \;\propto\; p_{\theta,t}(x_t)\,(1 - f_t(x_t))^{\lambda_t}, \qquad 0 = \lambda_T \le \ldots \le \lambda_0 = \lambda, \tag{5}$$

realizing our target (1) at $t = 0$ whenever $f_0 = p_{\theta,0}(c \mid \cdot)$. Here $p_{\theta,t}$ denotes the marginal of the pretrained reverse process in Section 2, so $p_{\theta,t}(x_t) \approx p_t(x_t)$ and $\gamma_T(x_T) = \mathcal{N}(x_T; 0, I)$. All operations of PU Steering below are defined on this path of the pretrained sampler: posteriors and class-conditional densities are those of $p_{\theta,t}$, and p_{target} denotes (1) with p replaced by $p_{\theta,0}$. Unlike [52, 31], where the potential defining γ_t is pre-specified, we estimate f_t online from the running particles. Section 4 shows that the marginal of γ_0 depends only on the terminal f_0, so neither accurate $\nabla_{x_t} f_t$ for all t nor accurate f_t at $t > 0$ is required.

Concretely, M particles are initialized at $t = T$ as $\{x_T^j\}_{j=1}^M \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I)$ with $w_{q,T}^j = 1$ for all j. At each subsequent step, a Gaussian proposal kernel $m_{t-1}(\cdot \mid x_t) = \mathcal{N}(\mu_t^{\text{prop}}, \sigma_t^2 I)$ whose mean μ_t^{prop} is specified per setting in Section 3.2 and Section 3.3, draws x_{t-1}^j; the weight update from γ_t to γ_{t-1} then admits the closed-form recursion

$$\begin{aligned} w_{q,t-1} \;&=\; w_{q,t}\; r_t \frac{(1 - f_{t-1}(x_{t-1}))^{\lambda_{t-1}}}{(1 - f_t(x_t))^{\lambda_t}}, \;\text{ where} \\ r_t \;&:=\; \frac{\exp\!\big(-\|x_{t-1} - \mu_\theta(x_t, t)\|^2 / 2\sigma_t^2\big)}{\exp\!\big(-\|x_{t-1} - \mu_t^{\text{prop}}\|^2 / 2\sigma_t^2\big)} \;=\; \frac{p_\theta(x_{t-1} \mid x_t)}{m_{t-1}(x_{t-1} \mid x_t)}, \end{aligned} \tag{6}$$

i.e., r_t is the density ratio between the pretrained reverse kernel and the proposal; the trailing classifier ratio in (6) accounts for the change of tilting from γ_t to γ_{t-1} (derivation in Appendix A.3).

PU posterior estimator. Estimating $p_{\theta,t}(c \mid x_t)$ is a binary classification problem with positive class c and negative class $\neg c$, but no samples from $p_{\theta,t}(\cdot \mid \neg c)$ are directly available in either of our settings. What both settings provide is (i) positive samples from $p_{\theta,t}(\cdot \mid c)$ and (ii) the pretrained unconditional marginal $p_{\theta,t}$, which mixes the two class-conditional distributions at the prior rate $\pi = p(c)$:

$$p_{\theta,t}(x_t) = \pi\, p_{\theta,t}(x_t \mid c) + (1 - \pi)\, p_{\theta,t}(x_t \mid \neg c).$$

This identity recasts the binary classification risk using only samples from $p(\cdot \mid c)$ and unconditional samples alone, giving the unbiased PU (uPU) risk [16, 41]

$$\mathcal{L}_{\text{uPU}}(f) = \pi\, \mathbb{E}_{p(x|c)}[\ell(f(x), +1)] + \mathbb{E}_{p(x)}[\ell(f(x), -1)] - \pi\, \mathbb{E}_{p(x|c)}[\ell(f(x), -1)], \tag{7}$$

where $\mathbb{E}$ denotes the expectation. This risk equals the standard binary classification risk (derivation in Appendix A.2). Minimizing it with a proper loss yields $f_t \approx p_{\theta,t}(c \mid x_t)$, provided π is known or estimable from domain knowledge. The SMC particles, however, do not follow $p_{\theta,t}$ and thus cannot directly serve as unlabeled samples in (7).

Particle density correction. We correct this mismatch by importance reweighting [49]. The density $\hat{\gamma}_t$ induced by the M particles under the proposal kernel and resampling rule has no closed form. Our key observation concerns the importance weight retargeting it to $p_{\theta,t}$, conceptually defined as

$$w_{\emptyset,t}(x_t) \;:=\; \frac{p_{\theta,t}(x_t)}{\hat{\gamma}_t(x_t)} \qquad (\emptyset \text{ marks the unconditional target}). \tag{8}$$

Setting $\lambda_t \equiv 0$ in (6) collapses γ_t to $p_{\theta,t}$ and the classifier ratio to one, giving its closed form

$$w_{\emptyset,t-1} \;=\; w_{\emptyset,t}\, r_t, \qquad w_{\emptyset,T} = 1; \tag{9}$$

i.e., $w_{\emptyset,t}$ is itself the SMC weight that would target $p_{\theta,t}$ along the *same* particle trajectory as $w_{q,t}$. The two recursions share r_t and differ only in the tilting factor $(1 - f_t)^{\lambda_t}$, so $w_{\emptyset,t}$ is tracked alongside $w_{q,t}$ at no additional cost. Empirically approximating $\mathbb{E}_{p(x)}$ in (7) with $\{(x_t^j, w_{\emptyset,t}^j)\}_{j=1}^M$ restores unbiasedness of the uPU risk used to fit f_t. This bridges the SMC sampler and the PU estimator: (6) drives particles toward γ_t for sampling, while (9) pulls them back to $p_{\theta,t}$ inside the uPU loss for f_t.

Algorithm 1 PU Steering: one reverse-diffusion step ($t \to t-1$).

Require: particles $\{(x_t^j, w_{q,t}^j, w_{\emptyset,t}^j)\}_{j=1}^M$, classifier f_t, schedule entry λ_{t-1}, class prior π.

 # Propose particles via any negative-guidance kernel

1: $x_{t-1}^j \sim m_{t-1}(\cdot \mid x_t^j),\ j = 1, \dots, M$ ▷ e.g., (10)

 # Advance $w_\emptyset$ via the closed-form recursion

2: update $w_{\emptyset,t-1}^j$ toward $p_{\theta,t-1}$ ▷ (9)

 # Refit the PU classifier on new particles

3: Sample positives $\mathcal{P}_{t-1}$ from $p_{\theta,t-1}(\cdot \mid c)$ ▷ *setting-specific*: Section 3.2/Section 3.3

4: $f_{t-1} \leftarrow \arg\min_f \mathcal{L}_{\mathrm{uPU}}(f)$ on $\mathcal{P}_{t-1}$ and $\{(x_{t-1}^j, w_{\emptyset,t-1}^j)\}_{j=1}^M$ ▷ (7)

 # Reweight w_q via the shared closed-form recursion

5: update $w_{q,t-1}^j$ toward γ_{t-1} ▷ (6)

 # Resample only when particles degenerate

6: **if** $\mathrm{ESS}(\{w_{q,t-1}^j\}) < \mathrm{ESS}_{\min}$ **then** resample $\propto w_{q,t-1}^j$ and reweight $w_{\emptyset,t-1}^j$ accordingly

Meta-algorithm. Algorithm 1 summarizes one timestep of PU Steering, which propagates particles through a negative-guidance proposal, refits the PU classifier on the reweighted particles, and updates the target weights before resampling. The full procedure initializes $\{x_T^j\}_{j=1}^M \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I)$, $w_{q,T}^j = w_{\emptyset,T}^j = 1$, and any f_T, and then applies Algorithm 1 for $t = T, T-1, \dots, 1$. The two settings in Section 3.2 and Section 3.3 differ only in the proposal kernel m_{t-1} and in how positives $\mathcal{P}_{t-1}$ are obtained; all other steps are shared.

Negative guidance ultimately requires a particle set in which the undesired condition c is suppressed, not a weighted empirical distribution dominated by a few high-weight particles. *Resampling* enforces this: ancestors $a(j) \propto w_{q,t-1}^j$ are drawn, each particle is replaced by $x_{t-1}^{a(j)}$, $w_{q,t-1}^j$ is reset to 1, and $w_{\emptyset,t-1}^j$ is rescaled to $w_{\emptyset,t-1}^{a(j)} / w_{q,t-1}^{a(j)}$ to preserve the ratio interpretation (8). Following [11], we trigger resampling only when the effective sample size $\mathrm{ESS}(\{w_{q,t-1}^j\}) := \big(\sum_j w_{q,t-1}^j\big)^2 / \sum_j (w_{q,t-1}^j)^2$ falls below $\mathrm{ESS}_{\min}$, since unnecessary resampling adds Monte Carlo noise and computational cost without benefit.

3.2 Reference-Based Setting

We instantiate Algorithm 1 when the undesired condition c is specified by a finite reference set $\{x_c^i\}_{i=1}^N$ of clean samples. The setting-specific slots are the positive set $\mathcal{P}_{t-1}$ in Step 3 and the proposal mean μ_t^{prop} in Step 1. Detailed pseudocode is given as Algorithm 2 in Appendix B.

PU classifier. Only the terminal f_0 controls the marginal of γ_0 (Section 3.1), so f_t for $t > 0$ need not be highly accurate. Classification of c is also intrinsically easier on clean samples than on noised ones, since the forward kernel of (2) destroys class-discriminative structure as t grows. Both observations motivate the parameterization

$$f_t(x_t) \ := \ g_0\big(\hat{x}_0(x_t)\big), \qquad g_0(x_0) \approx p_{\theta,0}(c \mid x_0).$$

At $t = 0$, $\hat{x}_0(x_0) = x_0$ gives $f_0 \equiv g_0$ and $\gamma_0(x_0) \propto p_{\theta,0}(x_0)\big(1 - g_0(x_0)\big)^\lambda$, recovering p_{target}.

Because g_0 acts on inputs in the denoised space, the uPU loss must be estimated with both positives and unlabeled samples projected through $\hat{x}_0$. We accordingly set

$$\mathcal{P}_{t-1} = \big\{\hat{x}_0(\tilde{x}_{t-1}^i)\big\}_{i=1}^N, \qquad \mathcal{U}_{t-1} = \big\{\big(\hat{x}_0(x_{t-1}^j),\ w_{\emptyset,t-1}^j\big)\big\}_{j=1}^M,$$

where $\tilde{x}_{t-1}^i \sim q_{t-1}(\cdot \mid x_c^i)$ is a forward-noised reference, so that positives and unlabeled samples are aligned on the same $\hat{x}_0$-pushforward at each level t and reduce to the original references at $t = 0$. Since g_0 is the only trainable object, the per-step refit reduces to a few warm-started gradient steps on its parameters, with $f_{t-1} = g_0 \circ \hat{x}_0$ updated implicitly.

SMC proposal. *(i) CG-style proposal driven by f_t.*

$$\mu_t^{\mathrm{prop}} \ := \ \mu_\theta(x_t, t) \ + \ \sigma_t^2 \, \lambda_{t-1} \, \nabla_{x_t} \log\big(1 - f_t(x_t)\big). \tag{10}$$

Under $f_t = g_0 \circ \hat{x}_0$ the gradient routes through Tweedie's formula, making (10) a CG variant that lifts a clean-domain predictor through $\hat{x}_0$[12, 4, 55]. The difference is that no accurate predictor is available at the start of sampling in our setting, so g_0 is fit online via PU learning.

(ii) Existing reference-based guidance as proposal. Methods such as the Safe Denoiser [29] provide efficient guidance but target no explicit density. Using their guidance as μ_t^{prop} leaves the per-step update intact while the SMC layer ensures asymptotic exactness for p_{target}.

3.3 Score-Based Setting

We instantiate Algorithm 1 when c is specified by a conditional noise predictor $\hat{\epsilon}_\theta(\cdot, \cdot \mid c)$ but no reference set is given. The parameterization $f_t = g_0 \circ \hat{x}_0$ and the two proposal options of Section 3.2 carry over unchanged. The non-trivial step is constructing $\mathcal{P}_{t-1}$ in Step 3 of Algorithm 1 without samples from $p_0(\cdot \mid c)$. Our key observation is that the particle-density correction of Section 3.1 extends directly to the conditional marginal once its score is available, supplying positives from the same particle trajectory. Detailed pseudocode is given as Algorithm 3 in Appendix B.

Positives via importance reweighting. Applying the construction of $w_{\emptyset,t}$ in Section 3.1 to the conditional target gives an analogous weight that retargets the particles to $p_{\theta,t}(\cdot \mid c)$,

$$w_{c,t}(x_t) \;:=\; \frac{p_{\theta,t}(x_t \mid c)}{\hat{\gamma}_t(x_t)}.$$

The same particles serve as positives and unlabeled data under different weights:

$$\mathcal{P}_{t-1} = \big\{\big(\hat{x}_0(x_{t-1}^j),\, w_{c,t-1}^j\big)\big\}_{j=1}^M, \quad \mathcal{U}_{t-1} = \big\{\big(\hat{x}_0(x_{t-1}^j),\, w_{\emptyset,t-1}^j\big)\big\}_{j=1}^M.$$

Sharing the particle set across the two terms of $\mathcal{L}_{\mathrm{uPU}}$ preserves unbiasedness in the particle limit (Section 4). Step 2 of Algorithm 1 then carries a recursion for $w_{c,t}$ alongside the existing $w_{\emptyset,t}$ update; both share the same closed-form recursion as in Section 3.1, with $\mu_\theta(x_t, t)$ replaced by $\mu_\theta(x_t, t \mid c)$.

SMC proposal. Both options of Section 3.2 extend directly: *(i)* the CG-style mean (10) driven by the f_t, or *(ii)* an existing score-based negative-guidance method, such as DNG [34], plugged in as μ_t^{prop}, with our SMC layer correcting its density-level bias against p_{target}. When the proposal already evaluates the conditional score, the $w_{c,t}$ recursion reuses that evaluation at no extra cost.

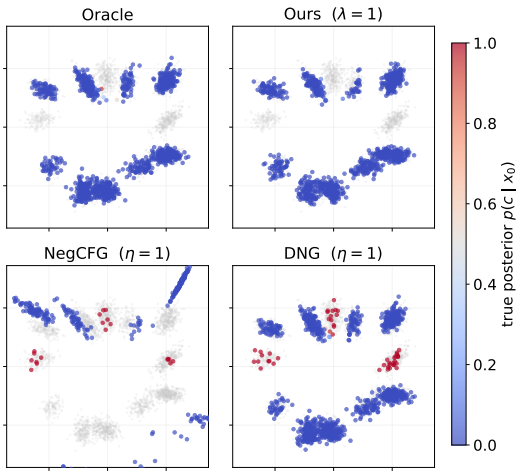
4 Theoretical Analysis

We summarize the convergence result for the reference-based instantiation of Algorithm 1 (Section 3.2) and extend the result to the score-based setting (Section 3.3). Formal statements and proofs are in Appendix C.

For implementation efficiency, Algorithm 1 uses a single particle system to both train f_t and form the final SMC estimator. For analytical clarity, we instead train f_t on an independent auxiliary system $\{x_t^k\}_{k=1}^{M'}$ driven by the same kernels, which removes the dependence between the classifier and the main SMC particles $\{(x_t^j, w_{q,t}^j)\}_{j=1}^M$, whose weighted empirical measure we denote by $\widehat{\nu}_t^{M|M',N}$. We establish asymptotic exactness in the resulting infinite-particle limit $\lim_{M',N\to\infty} \lim_{M\to\infty}$, in which the inner $M \to \infty$ controls the main-system SMC error at frozen classifiers and the outer $M', N \to \infty$ drives the classifier estimation error to zero.

Assumptions. Let $\mathcal{G}_t$ denote the function class over which f_t is fit. We require four conditions on $\mathcal{G}_t$ and on the posterior $p_{\theta,t}(c \mid x)$. *(i) Uniform convergence on empirical positive samples*: $\sup_{g\in\mathcal{G}_t} \big|\frac{1}{N}\sum_{i=1}^N g(X_i^+) - \mathbb{E}_{p_{\theta,t}(\cdot|c)}[g]\big| \to 0$ almost surely as $N \to \infty$, with $X_i^+ \overset{\text{i.i.d.}}{\sim} p_{\theta,t}(\cdot \mid c)$. *(ii) Realizability of the Bayes logit*: $\log\{p_{\theta,t}(c \mid x)/(1 - p_{\theta,t}(c \mid x))\} \in \mathcal{G}_t$. *(iii) Two-sided non-degeneracy*: $\delta \le p_{\theta,t}(c \mid x) \le 1 - \delta$ for some $\delta \in (0, 1)$. *(iv) Total boundedness* of $\mathcal{G}_t$ in the supremum norm.

Reference-based setting. The N i.i.d. references $\{x_c^i\}_{i=1}^N$ supply the positive term of the uPU risk (7), and the M' auxiliary particles supply the unlabeled term. We argue in three steps. **Step 1**,



Method	Posterior leak ↓	W_2 ↓
Oracle	0.012±0.010	1.5±0.3
Uncond (η=0)	0.253±0.016	3.6±0.6
NegCFG (η=0.5)	0.068±0.008	9.7±2.4
NegCFG (η=1)	0.018±0.005	20.5±6.3
DNG (η=1)	0.052±0.010	1.6±0.3
DNG (η=2)	0.015±0.005	3.5±0.7
Ours	**0.012±0.013**	**1.8±0.6**

Table 1: Comparison with baselines across repulsion strengths η, mean±std over 50 trials. Uncond corresponds to NegCFG/DNG at η=0, which reduces to unconditional DDPM from $p(x)$. Posterior leak $\mathbb{E}[p(c|x_0)]$ and W_2 measure 2-Wasserstein distance to ground-truth. Baselines show a trade-off between leak and W_2, while ours matches the Oracle in leak without inflating W_2.

Figure 1: Generated samples on one trial. Each point is a sample in its first two coordinates, colored by the true posterior $p(c \mid x) \in [0, 1]$ (red: $p(c \mid x)$=1, blue: 0). Oracle and Ours avoid the condition c and cover all remaining modes. NegCFG suppresses leak by collapsing onto a few remaining modes, and DNG keeps coverage but retains residual samples on the avoided modes.

fix timestep t and assume that the unlabeled empirical average converges uniformly to $\mathbb{E}_{p_{\theta,t}}$ over $\mathcal{G}_t$. Then uPU risk minimization [41], combined with (i) and (ii), yields L^1-convergence $f_t \to p_{\theta,t}(c \mid \cdot)$ as $M', N \to \infty$. We will invoke this at $t = 0$. **Step 2**, this convergence does hold: under (iii) and (iv), and with a mild proposal regularization (Appendix C.2.1) that keeps the importance weights bounded, the weighted empirical measure of the auxiliary particles $\{(x_t^k, w_{\emptyset,t}^k)\}_{k=1}^{M'}$ converges uniformly over $\mathcal{G}_t$ to $p_{\theta,t}$ by the SMC law of large numbers (LLN) in [11]. **Step 3**, the tempering factors $(1 - f_t)^{\lambda_t}$ enter the recursion (6) as ratios between consecutive times, so the cumulative product telescopes and only the terminal f_0 enters γ_0, similar to [31]. Combining these three steps yields Theorem 1.

Theorem 1 (Asymptotic exactness, informal version of Theorem 11 in Appendix C.3)**.** *Consider Algorithm 1 with the auxiliary M' system introduced above. For every bounded measurable $\varphi : \mathcal{X}_0 \to \mathbb{R}$,*

$$\lim_{M',N\to\infty} \lim_{M\to\infty} \widehat{\nu}_0^{M|M',N}(\varphi) \;=\; p_{\text{target}}(\varphi) \quad \textit{almost surely,}$$

under (i)–(iv) for the reference-based setting.

Extension to score-based setting. Without external references, the positive term of the uPU risk (7) is supplied by the same M' auxiliary particles, re-weighted by $w_{c,t}$ and $w_{\emptyset,t}$. The argument mirrors the reference-based one, with only **Step 2** changed. By an analogous SMC LLN [11], the weighted empirical measure of $\{(x_t^k, w_{c,t}^k)\}_{k=1}^{M'}$ converges uniformly over $\mathcal{G}_t$ to $p_{\theta,t}(\cdot \mid c)$ as $M' \to \infty$. Combined with the unlabeled-term convergence from **Step 2** of the reference-based setting, this gives uniform convergence of both terms of the uPU risk and hence $f_t \to p_{\theta,t}(c \mid \cdot)$ as $M' \to \infty$ at each timestep t. **Step 1** and **Step 3** then carry over unchanged, yielding $\widehat{\nu}_0^{M|M'}(\varphi) \to p_{\text{target}}(\varphi)$ under (ii)–(iv), for every bounded measurable φ (Theorem 16 in Appendix C.4).

5 Experiments

This section summarizes our experimental results in three settings of increasing scale. Section 5.1 uses the score-based setting (Section 3.3) on a 3-dimensional Gaussian mixture model (GMM), where closed-form densities enable rigorous metric evaluation and direct visualization of the full distribution. Section 5.2 uses the reference-based setting (Section 3.2) on CIFAR-10 image generation with a pretrained score network, under different specifications of the avoidance condition c (a class or set of classes to avoid). Section 5.3 applies the score-based setting to text-to-image generation with Stable Diffusion. All three compare against the baselines from Section 2, with full experimental details

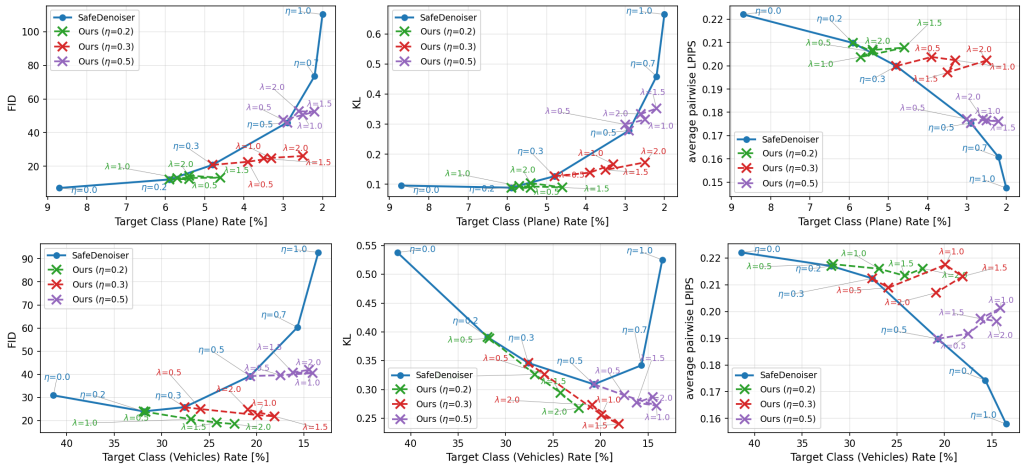


Figure 2: Avoidance rate vs. generation quality on CIFAR-10. Top: plane. Bottom: four-class vehicles. Columns plot target-class rate against FID, KL, and average pairwise LPIPS (left to right); lower is better except for LPIPS. The Safe Denoiser sweeps η. Ours uses the Safe Denoiser as its proposal kernel, fixes $\eta \in \{0.2, 0.3, 0.5\}$, and sweeps λ, tracing the Pareto frontier in all panels.

deferred to Appendix D. To align the evaluation protocol with these baselines, we treat the final SMC output after intermediate resampling as particles with unit weight rather than as weighted particles.[1]

5.1 GMM with Score-Based Setting

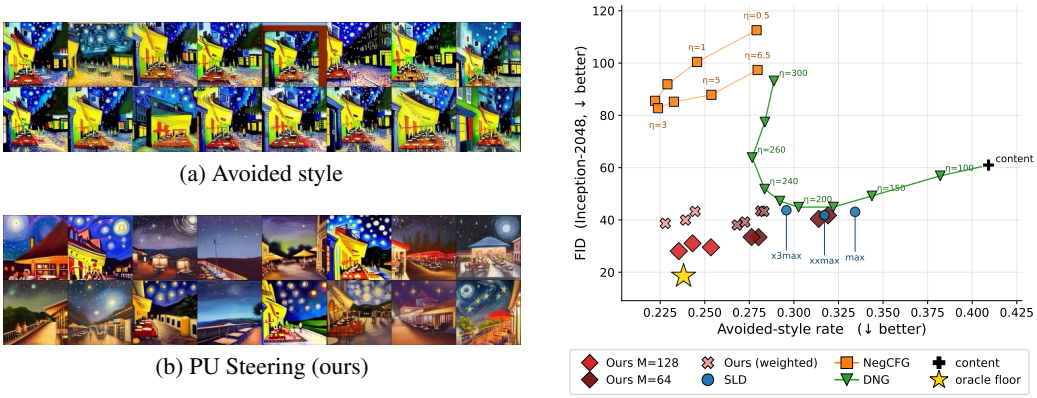
Setup. The unconditional density is a 3-dimensional GMM of 12 modes, with the avoided condition c assigned to 3 modes and $\neg c$ to the other 9. The layout of the GMM is randomly drawn. We define the ground-truth target by setting $\lambda{=}1$ in (1), giving $p_{\text{target}}{=}p(\cdot \mid \neg c)$. All methods use a common $T{=}100$ DDPM discretization with $M{=}1024$ samples. The Oracle alone has access to the true score for the target $\nabla \log p_t(x_t \mid \neg c)$, serving as the achievable upper bound. We compare against NegCFG and DNG from Section 2, and our method uses a kernel-based PU classifier [41]. We report posterior leak $\mathbb{E}[p(c \mid x_0)]$ (average posterior of c at generated samples, lower is stronger avoidance) in closed form, and the 2-Wasserstein distance (W_2, calculated by [19]) to 3000 i.i.d. draws from $p(\cdot \mid \neg c)$. Results are averaged over 50 trials with the GMM layout and DDPM sampling seed both randomized.

Results. Figure 1 shows the generated samples from each method on one trial. NegCFG collapses onto a few $\neg c$-modes, and DNG retains samples on the avoided c-modes, both consistent with the limitations of local repulsion away from c discussed in Section 2. In contrast, our method preserves all nine $\neg c$-modes and matches the Oracle, leveraging global information across particles, consistent with Theorem 1. Table 1 reports the same trend quantitatively over 50 trials. Ours matches the Oracle in posterior leak (both 0.012) and stays within one Oracle standard deviation in W_2 ($1.8{\pm}0.6$ vs. $1.5{\pm}0.3$), while NegCFG and DNG exhibit a trade-off between leak and W_2. Full hyperparameter sweeps are in Appendix D.1.

5.2 CIFAR-10 with Reference-Based Setting

Setup. We use a pretrained UNet for the unconditional CIFAR-10 diffusion model from [40], and run DDPM sampling with a batch size of $M{=}128$ over $T{=}250$ timesteps. The avoided condition c is one or more CIFAR-10 classes. We evaluate plane (single class) and the four vehicle classes (plane, car, ship, truck). For each class in c, we draw $N{=}200$ references from the training set. The PU discriminator f_t is a linear probe on pooled UNet bottleneck features. Because N and M are both small here, we replace the uPU loss of Section 3 with binary cross-entropy (BCE), treating the references as positives and the SMC particles as soft negatives. After resampling, with so few SMC particles in pixel space, those satisfying c are quickly depleted, so the particle set increasingly behaves as a true negative set as t decreases, making BCE an effective proxy in this regime.

[1] Code and experimental results are available at `https://github.com/htkg/pu-steering`.



(a) Avoided style

(b) PU Steering (ours)

Figure 3: Results of Van Gogh style avoidance on Stable Diffusion v1.4. Left: generated examples of (a) the avoided style and (b) ours at $M{=}128$, $\lambda{=}16$, each showing 16 randomly selected images. These are the top two rows of the random 8×8 grids in Figure 6 of Appendix D.3, which also shows the other methods. Right: avoided-style rate against FID; lower left is better. The baselines sweep their strengths, and ours uses SLD as its proposal kernel, reaching the oracle-floor rate with the lowest FID while keeping the prompted scene. Full results are in Appendix D.3.

Evaluation metrics. We aggregate 80 seeds $\times$ 128 samples per configuration. Fréchet Inception Distance (*FID*, 24) is computed against the CIFAR-10 training set from non-avoided classes, using the implementation of [44]. Each generated image is then classified by an independently trained ResNet-18 [23] with $\sim$95% test accuracy into the class histogram $\hat{p}$, from which we report the *target-class rate* on avoided classes (lower is stronger avoidance) and the *KL* divergence of $\hat{p}$ from a uniform over non-avoided classes (lower is more uniform). As an auxiliary check, we also report average pairwise Learned Perceptual Image Patch Similarity (LPIPS; 58) via AlexNet [35] on 256 random samples (higher is more diverse).

Results. We first run the Safe Denoiser [29] sweeping its repulsion strength η. We then run our method (Section 3.2), which uses the Safe Denoiser as its proposal μ_t^{prop}, at three fixed values $\eta \in \{0.2, 0.3, 0.5\}$ while sweeping the terminal tempering strength λ, so that both methods are compared on a shared η axis. Figure 2 plots target-class rate against each quality metric for plane and four-class vehicles avoidance. At each fixed η, our λ-sweep Pareto-dominates the Safe Denoiser at the same η on FID and KL across the full operating range, and is no worse on LPIPS over the bulk of it. A weak negative guidance leaves residual mass on the avoided classes that the classifier and resampling then correct, raising the avoidance rate without the quality degradation incurred by stronger negative guidance. The added machinery beyond the Safe Denoiser is one linear probe on UNet features and ESS-based adaptive resampling on the classifier-driven weights. Per-class results for cat, frog, and ship and regularization ablations are in Appendix D.2.

5.3 Stable Diffusion with Score-Based Setting

Setup. We use Stable Diffusion v1.4 [42], a pretrained text-to-image latent diffusion model, with classifier-free guidance at scale 7.5 as the sampler of the base density p_θ, running a batch size of $M{=}128$ over $T{=}50$ timesteps. The prompt `"a painting of a cafe terrace at night under a starry sky"` often yields Van Gogh's style, which we specify as the avoided condition c to suppress. Appending `"in the style of Van Gogh"` to the prompt supplies the conditional score, which generates images such as Figure 3(a), and we use it to run the score-based setting of Section 3.3. The PU classifier f_t is a kernel logistic model on $\hat{x}_0(x_t)$. Its positives and unlabeled data are $\hat{x}_0$ of the same particles under the style-appended and original prompts, respectively, so they share the prompted content and differ mainly in the style. This contrast keeps the fit stable, so we use the uPU loss (7) as written, unlike the BCE in Section 5.2.

Evaluation metrics. We aggregate 2048 images per configuration, generated with distinct seeds. An independently trained detector, a logistic head on a style embedding [47], assigns each image a probability of the avoided style, and we report its mean as the *avoided-style rate* (lower is stronger

avoidance). As no samples of the target are available, we approximate them by filtering: from 8192 images of the prompt without avoidance, the 2048 least like the style under the detector form the reference set, against which FID and KID [6] are computed (lower is better). Applying the same filter to an independent pool generated with disjoint seeds gives the *oracle floor*, a reference point for an ideal avoidance rather than a lower bound.

Results. We first run SLD [43], a negative-guidance method specialized to Stable Diffusion, and DNG, each sweeping its strength. We then run our method (Section 3.3), which uses SLD as its proposal μ_t^{prop}, while sweeping λ. Figure 3 (right) plots the avoided-style rate against FID. DNG traces an avoidance–quality trade-off: at its best strengths it is comparable to SLD, and stronger guidance only degrades FID. NegCFG lowers the rate only at FIDs above 80. Our method at $M{=}128$, $\lambda{=}16$ reaches a rate of 0.235, at the oracle floor of 0.238, with the lowest FID of 28.10, improving both avoidance and quality beyond SLD and DNG. Halving the particle count to $M{=}64$ gives a rate of 0.276 and an FID of 33.46, so additional particles improve both. Qualitatively, Figure 3 (left) shows that our samples keep the prompted scene while suppressing the avoided style and spreading over a variety of painting styles. The Stable Diffusion results are thus consistent with the CIFAR-10 results in Section 5.2. The full setup and results are in Appendix D.3.

6 Limitations and Future Work

Scope of the experiments. Our experiments cover several settings up to image-generation scale (Section 5). Extending the framework to the broader range of applications that motivate negative guidance (Section 1) is left open. Our setting also differs from SMC-based diffusion alignment [52, 45, 31], which sharpens particles toward a desired condition. PU Steering instead targets everything other than the avoided condition c, which is harder to evaluate because no single reference distribution is available beyond a labeled validation set. A parallel line of recent work develops SMC and other samplers that route interacting particles to a target along the diffusion path [53, 50], training-free and with convergence guarantees on multi-modal targets. These samplers are demonstrated on low- to moderate-dimensional settings ($d < 100$), and scaling them to image-scale dimensions is left open. PU Steering reaches CIFAR-10 and Stable Diffusion by combining a pretrained UNet with an existing negative-guidance kernel [29, 43] as the SMC proposal, and the same recipe applies to any pretrained diffusion model with such a kernel. This connects principled probabilistic samplers to alignment of modern generative models. We discuss future directions for improving and specializing our framework below, which would also let us examine the effect of scaling the particle count.

Variance of the importance weights. A known challenge in SMC is weight degeneracy, in which the variance of the importance weights grows and the ESS decreases [11]. This can lead to frequent resampling, and may eventually result in a few common ancestors dominating the particle set. For a fixed target path, [31] derives a proposal kernel that minimizes the one-step weight variance. In our case, the target path itself depends on the online classifier f_t, and inaccuracies in f_t further affect the weight variance. Extending such an analysis is left for future work. Task-specific regularization and other practical stabilization techniques are natural directions in this line.

Stabilizing the online PU classifier. Our method trains the classifier f_t in an online manner. Recent PU learning literature offers stabilization techniques for limited positive samples, small-batch variability, and high-capacity models such as neural networks that are prone to overfitting [33, 22]. Online and importance-weighted PU learning further address streaming data and weighted unlabeled samples [57, 36]. Incorporating these techniques may further stabilize our method.

7 Conclusion

In this paper, we addressed negative guidance of diffusion models, the task of steering generation away from an undesired condition. We recast the problem as sampling from an explicit avoidance density and realized it via tempered SMC driven by an online PU classifier, where particles exploit global information across samples rather than local repulsion. Our asymptotic exactness analysis implies that additional computation reduces distributional bias and improves generation quality. On the practical side, we showed that augmenting existing guidance methods with a lightweight classifier and resampling improves the avoidance–quality trade-off.

Acknowledgments and Disclosure of Funding

MS was supported by JST ASPIRE Grant Number JPMJAP25B1.

References

[1] Anurag Ajay, Yilun Du, Abhi Gupta, Joshua B. Tenenbaum, Tommi S. Jaakkola, and Pulkit Agrawal. Is Conditional Generative Modeling all you need for Decision Making? In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=sP1fo2K9DFG.

[2] Mohammadreza Armandpour, Ali Sadeghian, Huangjie Zheng, Amir Sadeghian, and Mingyuan Zhou. Re-imagine the Negative Prompt Algorithm: Transform 2D Diffusion into 3D, alleviate Janus problem and Beyond, 2023. URL https://arxiv.org/abs/2304.04968.

[3] Yuanhao Ban, Ruochen Wang, Tianyi Zhou, Minhao Cheng, Boqing Gong, and Cho-Jui Hsieh. Understanding the Impact of Negative Prompts: When and How Do They Take Effect? In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol, editors, *Computer Vision – ECCV 2024*, pages 190–206, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-73024-5.

[4] Arpit Bansal, Hong-Min Chu, Avi Schwarzschild, Roni Sengupta, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Universal Guidance for Diffusion Models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=pzpWBbnwiJ.

[5] Shristi Das Biswas, Arani Roy, and Kaushik Roy. CURE: Concept Unlearning via Orthogonal Representation Editing in Diffusion Models. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen, editors, *Advances in Neural Information Processing Systems*, volume 38, pages 82539–82563. Curran Associates, Inc., 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/769736dfbf6a1f64b4d2ab5c82c3d5e2-Paper-Conference.pdf.

[6] Mikołaj Bińkowski, Dougal J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=r1lUOzWCW.

[7] Arwen Bradley and Preetum Nakkiran. Classifier-Free Guidance is a Predictor-Corrector. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=zrWNtzSZsf.

[8] Nicolas 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, Anaheim, CA, August 2023. USENIX Association. ISBN 978-1-939133-37-3. URL https://www.usenix.org/conference/usenixsecurity23/presentation/carlini.

[9] Dar-Yen Chen, Hmrishav Bandyopadhyay, Kai Zou, and Yi-Zhe Song. Normalized Attention Guidance: Universal Negative Guidance for Diffusion Models. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen, editors, *Advances in Neural Information Processing Systems*, volume 38, pages 137412–137446. Curran Associates, Inc., 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/c8aefe45deba3639af7120b141b87b50-Paper-Conference.pdf.

[10] Muthu Chidambaram, Khashayar Gatmiry, Sitan Chen, Holden Lee, and Jianfeng Lu. What does guidance do? A fine-grained analysis in a simple setting. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 84968–85005. Curran Associates, Inc., 2024. doi: 10.52202/079017-2698. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/9a3942c235daa9c5f62a8598ae81a946-Paper-Conference.pdf.

[11] Nicolas Chopin and Omiros Papaspiliopoulos. *An Introduction to Sequential Monte Carlo*. Springer International Publishing, 2020. ISBN 9783030478452. doi: 10.1007/978-3-030-47845-2. URL http://dx.doi.org/10.1007/978-3-030-47845-2.

[12] Hyungjin Chung, Jeongsol Kim, Michael Thompson Mccann, Marc Louis Klasky, and Jong Chul Ye. Diffusion Posterior Sampling for General Noisy Inverse Problems. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=OnD9zGAGT0k.

[13] Pierre Del Moral, Arnaud Doucet, and Ajay Jasra. Sequential Monte Carlo Samplers. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 68(3):411–436, 06 2006. ISSN 1369-7412. doi: 10.1111/j.1467-9868.2006.00553.x. URL https://doi.org/10.1111/j.1467-9868.2006.00553.x.

[14] François Denis, Rémi Gilleron, and Fabien Letouzey. Learning from positive and unlabeled examples. *Theoretical Computer Science*, 348(1):70–83, 2005. ISSN 0304-3975. doi: https://doi.org/10.1016/j.tcs.2005.09.007. URL https://www.sciencedirect.com/science/article/pii/S0304397505005256. Algorithmic Learning Theory (ALT 2000).

[15] Prafulla Dhariwal and Alexander Nichol. Diffusion Models Beat GANs on Image Synthesis. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 8780–8794. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/49ad23d1ec9fa4bd8d77d02681df5cfa-Paper.pdf.

[16] Marthinus C. du Plessis, Gang Niu, and Masashi Sugiyama. Analysis of Learning from Positive and Unlabeled Data. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. URL https://proceedings.neurips.cc/paper_files/paper/2014/file/f032bc3f1eb547f716df87edb523b8f0-Paper.pdf.

[17] Bradley Efron. Tweedie's Formula and Selection Bias. *Journal of the American Statistical Association*, 106(496):1602–1614, 2011. ISSN 01621459. URL http://www.jstor.org/stable/23239562.

[18] Charles Elkan and Keith Noto. Learning classifiers from only positive and unlabeled data. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, page 213–220, New York, NY, USA, 2008. Association for Computing Machinery. ISBN 9781605581934. doi: 10.1145/1401890.1401920. URL https://doi.org/10.1145/1401890.1401920.

[19] Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z. Alaya, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, Léo Gautheron, Nathalie T.H. Gayraud, Hicham Janati, Alain Rakotomamonjy, Ievgen Redko, Antoine Rolet, Antony Schutz, Vivien Seguy, Danica J. Sutherland, Romain Tavenard, Alexander Tong, and Titouan Vayer. POT: Python Optimal Transport. *Journal of Machine Learning Research*, 22(78):1–8, 2021. URL http://jmlr.org/papers/v22/20-451.html.

[20] 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 (ICCV)*, pages 2426–2436, October 2023.

[21] Rohit Gandikota, Hadas Orgad, Yonatan Belinkov, Joanna Materzyńska, and David Bau. Unified Concept Editing in Diffusion Models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 5111–5120, January 2024.

[22] Saurabh Garg, Yifan Wu, Alexander Smola, Sivaraman Balakrishnan, and Zachary Lipton. Mixture Proportion Estimation and PU Learning: A Modern Approach. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 8532–8544. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/47b4f1bfdf6d298682e610ad74b37dca-Paper.pdf.

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[24] 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. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/8a1d694707eb0fefe65871369074926d-Paper.pdf.

[25] Jonathan Ho and Tim Salimans. Classifier-Free Diffusion Guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. URL https://openreview.net/forum?id=qw8AKxfYbI.

[26] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf.

[27] Emiel Hoogeboom, Víctor Garcia Satorras, Clément Vignac, and Max Welling. Equivariant Diffusion for Molecule Generation in 3D. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 8867–8887. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/hoogeboom22a.html.

[28] Michael Janner, Yilun Du, Joshua Tenenbaum, and Sergey Levine. Planning with Diffusion for Flexible Behavior Synthesis. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 9902–9915. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/janner22a.html.

[29] Mingyu Kim, Dongjun Kim, Amman Yusuf, Stefano Ermon, and Mi Jung Park. Training-Free Safe Denoisers for Safe Use of Diffusion Models. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen, editors, *Advances in Neural Information Processing Systems*, volume 38, pages 136004–136035. Curran Associates, Inc., 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/c66e1fcc9691aae706250638f36f681b-Paper-Conference.pdf.

[30] Mingyu Kim, Young-Heon Kim, and Mijung Park. Safety-Guided Flow (SGF): A Unified Framework for Negative Guidance in Safe Generation. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=EA80Zib9UI.

[31] Sunwoo Kim, Minkyu Kim, and Dongmin Park. Test-time Alignment of Diffusion Models without Reward Over-optimization. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=vi3DjUhFVm.

[32] Michael Kirchhof, James Thornton, Louis Béthune, Pierre Ablin, Eugene Ndiaye, and Marco Cuturi. Shielded Diffusion: Generating Novel and Diverse Images using Sparse Repellency. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu, editors, *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 30911–30942. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/kirchhof25a.html.

[33] Ryuichi Kiryo, Gang Niu, Marthinus C du Plessis, and Masashi Sugiyama. Positive-Unlabeled Learning with Non-Negative Risk Estimator. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.,

2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/7cce53cf90577442771720a370c3c723-Paper.pdf.

[34] 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. URL https://openreview.net/forum?id=6p74UyAdLa.

[35] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In F. Pereira, C.J. Burges, L. Bottou, and K. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. URL https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

[36] Atsutoshi Kumagai, Tomoharu Iwata, Hiroshi Takahashi, Taishi Nishiyama, and Yasuhiro Fujiwara. Importance-weighted Positive-unlabeled Learning for Distribution Shift Adaptation. In Yingzhen Li, Stephan Mandt, Shipra Agrawal, and Emtiyaz Khan, editors, *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pages 1576–1584. PMLR, 03–05 May 2025. URL https://proceedings.mlr.press/v258/kumagai25a.html.

[37] Alexander C. Li, Mihir Prabhudesai, Shivam Duggal, Ellis Brown, and Deepak Pathak. Your Diffusion Model is Secretly a Zero-Shot Classifier. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2206–2217, October 2023.

[38] Gen Li and Yuchen Jiao. Provable Efficiency of Guidance in Diffusion Models for General Data Distribution. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu, editors, *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 35034–35046. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/li25as.html.

[39] Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D Plumbley. AudioLDM: Text-to-Audio Generation with Latent Diffusion Models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 21450–21474. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/liu23f.html.

[40] Alexander Quinn Nichol and Prafulla Dhariwal. Improved Denoising Diffusion Probabilistic Models. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8162–8171. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/nichol21a.html.

[41] Marthinus Du Plessis, Gang Niu, and Masashi Sugiyama. Convex Formulation for Learning from Positive and Unlabeled Data. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1386–1394, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/plessis15.html.

[42] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-Resolution Image Synthesis With Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, June 2022.

[43] 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 (CVPR)*, pages 22522–22531, June 2023.

[44] Maximilian Seitzer. pytorch-fid: FID Score for PyTorch. https://github.com/mseitzer/pytorch-fid, August 2020. Version 0.3.0.

[45] Raghav Singhal, Zachary Horvitz, Ryan Teehan, Mengye Ren, Zhou Yu, Kathleen Mckeown, and Rajesh Ranganath. A General Framework for Inference-time Scaling and Steering of Diffusion Models. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu, editors, *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 55810–55827. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/singhal25b.html.

[46] 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 (CVPR)*, pages 6048–6058, June 2023.

[47] Gowthami Somepalli, Anubhav Gupta, Kamal Gupta, Shramay Palta, Micah Goldblum, Jonas Geiping, Abhinav Shrivastava, and Tom Goldstein. Measuring style similarity in diffusion models. *arXiv preprint arXiv:2404.01292*, 2024.

[48] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-Based Generative Modeling through Stochastic Differential Equations. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=PxTIG12RRHS.

[49] Masashi Sugiyama and Motoaki Kawanabe. *Machine Learning in Non-Stationary Environments: Introduction to Covariate Shift Adaptation*. The MIT Press, 03 2012. ISBN 9780262017091. doi: 10.7551/mitpress/9780262017091.001.0001. URL https://doi.org/10.7551/mitpress/9780262017091.001.0001.

[50] Hirohane Takagi and Atsushi Nitanda. Alternating Diffusion for Proximal Sampling with Zeroth Order Queries. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=NjjRuJuMTd.

[51] Hiroshi Takahashi, Tomoharu Iwata, Atsutoshi Kumagai, Yuuki Yamanaka, and Tomoya Yamashita. Positive-Unlabeled Diffusion Models for Preventing Sensitive Data Generation. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=jKcZ4hF4s5.

[52] Luhuan Wu, Brian Trippe, Christian Naesseth, David Blei, and John Cunningham. Practical and Asymptotically Exact Conditional Sampling in Diffusion Models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 31372–31403. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/63e8bc7bbf1cfea36d1d1b6538aecce5-Paper-Conference.pdf.

[53] Luhuan Wu, Yi Han, Christian Andersson Naesseth, and John Cunningham. Reverse Diffusion Sequential Monte Carlo Samplers. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen, editors, *Advances in Neural Information Processing Systems*, volume 38, pages 69445–69459. Curran Associates, Inc., 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/6490ab6ef8eb7116c6c1889795d21748-Paper-Conference.pdf.

[54] Yuchen Wu, Minshuo Chen, Zihao Li, Mengdi Wang, and Yuting Wei. Theoretical insights for diffusion guidance: A case study for Gaussian mixture models. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 53291–53327. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/wu24b.html.

[55] Jiwen Yu, Yinhuai Wang, Chen Zhao, Bernard Ghanem, and Jian Zhang. FreeDoM: Training-Free Energy-Guided Conditional Diffusion Model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 23174–23184, October 2023.

[56] Zhiyuan Zhan, Liuzhuozheng Li, and Masashi Sugiyama. Understanding Guidance Scale in Diffusion Models from a Geometric Perspective. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. URL https://openreview.net/forum?id=nfHimL6g8G.

[57] Chuang Zhang, Chen Gong, Tengfei Liu, Xun Lu, Weiqiang Wang, and Jian Yang. Online Positive and Unlabeled Learning. In Christian Bessiere, editor, *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 2248–2254. International Joint Conferences on Artificial Intelligence Organization, 7 2020. doi: 10.24963/ijcai.2020/311. URL https://doi.org/10.24963/ijcai.2020/311. Main track.

[58] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

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: Our claims match the theoretical and experimental results.

   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 in Section 6

   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: We provide full assumptions and complete proofs in the Appendix C.

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: All settings of the experiments are documented in Section 5 and Appendix D.

   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: Code and per-experiment documentation are provided in the supplementary material. A cleaned-up version is available at `https://github.com/htkg/pu-steering`.

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: The details of the implementation are presented in Section 5 and Appendix D.

   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: [Yes]

   Justification: For experiments on CPUs in Section 5.1, results are reported as mean $\pm$ standard deviation (1-sigma) over a sufficient number of trials. For experiments on GPUs in Section 5.2, only the mean over a large sample size is reported, and error bars are not included mainly due to the lack of computational resources.

   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: We explain the compute resources we used for experiments in Appendix D.

   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 conforms with the NeurIPS Code of Ethics.

   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: Broader impacts are discussed in Appendix E.

    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: [N/A]

    Justification: We release no new pretrained model or dataset.

    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: We cite the original references of the assets used in the paper, and document the assets actually used in Appendix D and the files of the supplementary material.

    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: [N/A]

    Justification: We do not release new pretrained models or datasets.

    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.

    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 crowdsourcing or research with human subjects.

    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 a component of the core methodology. We use LLMs only for writing polish and coding assistance.

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.

A Extended Preliminaries

This appendix provides standard background and derivations on three prerequisite topics—diffusion models [48, 26], Positive–Unlabeled (PU) learning [16, 41], and Diffusion Alignment as Sampling (DAS; 31)—that the main text (Section 2 and Section 3) treats briefly.

A.1 Background on Diffusion Models

Diffusion models [26, 48] are a class of generative models that learn to generate data by reversing a gradual noising process applied to training samples.

(i) Forward process. Let $\{\alpha_t\}_{t=1}^T$ be a noise schedule with $\alpha_t = 1 - \beta_t \in (0,1)$ and cumulative product $\bar{\alpha}_t = \prod_{s=1}^t \alpha_s$. The *forward Markov chain* is

$$q(x_t \mid x_{t-1}) \;=\; \mathcal{N}\big(x_t;\, \sqrt{\alpha_t}\, x_{t-1},\; \beta_t I\big), \qquad t = 1, \dots, T,$$

which by Gaussian convolution admits the closed-form conditional marginal

$$q_t(x_t \mid x_0) \;=\; \mathcal{N}\big(x_t;\, \sqrt{\bar{\alpha}_t}\, x_0,\; (1-\bar{\alpha}_t) I\big).$$

The schedule is chosen so that $\bar{\alpha}_T \approx 0$, ensuring $p_T(x_T) \approx \mathcal{N}(x_T;\, 0,\, I)$.

(ii) Denoising score matching and the noise-predictor identity. The *score* of the time-t marginal is $\nabla_{x_t} \log p_t(x_t)$. Diffusion models are trained with the *denoising score matching* objective [48, 26]

$$\min_\theta \;\; \mathbb{E}_{x_0 \sim p,\, \epsilon \sim \mathcal{N}(0,I),\, t \sim \mathrm{Unif}\{1,\dots,T\}} \Big[\big\| \hat{\epsilon}_\theta(x_t, t) - \epsilon \big\|^2 \Big], \qquad x_t \;=\; \sqrt{\bar{\alpha}_t}\, x_0 + \sqrt{1-\bar{\alpha}_t}\, \epsilon.$$

The optimum $\hat{\epsilon}_\theta^\star$ predicts the noise ϵ that was added to x_0 to produce x_t, and differentiating the log of the Gaussian density $q_t(x_t \mid x_0)$ gives $\nabla_{x_t} \log q_t(x_t \mid x_0) = -\epsilon / \sqrt{1-\bar{\alpha}_t}$. Taking the conditional expectation over $x_0 \mid x_t$ and using $\nabla_{x_t} \log p_t(x_t) = \mathbb{E}_{x_0 \mid x_t}[\nabla_{x_t} \log q_t(x_t \mid x_0)]$ yields the identity

$$\hat{\epsilon}_\theta^\star(x_t, t) \;=\; -\sqrt{1-\bar{\alpha}_t}\, \nabla_{x_t} \log p_t(x_t).$$

(iii) Tweedie's formula. For $Y = \sqrt{\bar{\alpha}_t}\, X + \sqrt{1-\bar{\alpha}_t}\, Z$ with $Z \sim \mathcal{N}(0, I)$, Tweedie's formula [17] gives the posterior mean

$$\mathbb{E}[X \mid Y] \;=\; \frac{Y + (1-\bar{\alpha}_t)\, \nabla_Y \log p_Y(Y)}{\sqrt{\bar{\alpha}_t}}.$$

Substituting $Y = x_t$, $X = x_0$ and the score identity from (ii) yields

$$\hat{x}_0(x_t) \;=\; \frac{x_t - \sqrt{1-\bar{\alpha}_t}\, \hat{\epsilon}_\theta(x_t, t)}{\sqrt{\bar{\alpha}_t}} \;\approx\; \mathbb{E}[x_0 \mid x_t].$$

Many downstream tasks require evaluating a function g_0 that is defined on clean samples x_0, but one only has access to noisy samples x_t during the reverse process. Using $\hat{x}_0(x_t)$ as a proxy for x_0, one can *lift* g_0 to noisy-sample space by defining $f_t(x_t) := g_0(\hat{x}_0(x_t))$. This lifting technique is used in, e.g., Diffusion Posterior Sampling [12] and the Twisted Diffusion Sampler [52].

(iv) Reverse process and DDPM parameterization. By Bayes' rule the ground-truth reverse transition conditioned on the clean sample is Gaussian,

$$q(x_{t-1} \mid x_t, x_0) \;=\; \mathcal{N}\big(x_{t-1};\, \tilde{\mu}_t(x_t, x_0),\; \tilde{\sigma}_t^2 I\big), \quad \tilde{\sigma}_t^2 \;=\; \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\, \beta_t.$$

The DDPM parameterization [26] approximates this by plugging in the Tweedie estimate for x_0, yielding the learned reverse kernel

$$p_\theta(x_{t-1} \mid x_t) \;=\; \mathcal{N}\big(x_{t-1};\, \mu_\theta(x_t, t),\; \sigma_t^2 I\big),$$

where

$$\mu_\theta(x_t, t) \;=\; \frac{1}{\sqrt{\alpha_t}} \Big(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\, \hat{\epsilon}_\theta(x_t, t) \Big), \qquad \sigma_t^2 \;=\; \tilde{\sigma}_t^2.$$

Unconditional generation ("ancestral sampling" [48]) then initializes $x_T \sim \mathcal{N}(0, I)$ and iteratively draws $x_{t-1} \sim p_\theta(\cdot \mid x_t)$ for $t = T, T-1, \dots, 1$.

A.2 PU Learning

PU binary classification considers a setting where *positive* samples are drawn from $p(x \mid c)$ and *unlabeled* samples from the marginal $p(x)$, which mixes c and $\neg c$ in proportion $\pi = p(c)$, but no labeled $\neg c$ samples are available. The goal is to train a binary classifier with label $y = +1$ for class c and $y = -1$ for $\neg c$.

We now derive the *unbiased PU (uPU) risk*. Given the data decomposition $p(x) = \pi\, p(x \mid c) + (1 - \pi)\, p(x \mid \neg c)$ with class prior $\pi \in (0, 1)$ and a per-sample loss $\ell(\cdot, y)$, the standard fully supervised (Positive–Negative, PN) risk is

$$\mathcal{L}_{\mathrm{PN}}(f) = \pi\, \mathbb{E}_{p(x|c)}[\ell(f(x), +1)] + (1 - \pi)\, \mathbb{E}_{p(x|\neg c)}[\ell(f(x), -1)].$$

Rearranging the data decomposition $p(x) = \pi\, p(x \mid c) + (1-\pi)\, p(x \mid \neg c)$ expresses the unobservable negative-class term as $(1 - \pi)\, p(x \mid \neg c) = p(x) - \pi\, p(x \mid c)$, replacing it with the marginal $p(x)$ and the positive class $p(x \mid c)$. [16, 41] apply this substitution at the level of expectations to obtain the exact identity

$$(1 - \pi)\, \mathbb{E}_{p(x|\neg c)}[\ell(f(x), -1)] = \mathbb{E}_{p(x)}[\ell(f(x), -1)] - \pi\, \mathbb{E}_{p(x|c)}[\ell(f(x), -1)].$$

Substituting yields the *unbiased PU (uPU) risk* (7),

$$\mathcal{L}_{\mathrm{uPU}}(f) = \pi\, \mathbb{E}_{p(x|c)}[\ell(f(x), +1)] + \mathbb{E}_{p(x)}[\ell(f(x), -1)] - \pi\, \mathbb{E}_{p(x|c)}[\ell(f(x), -1)],$$

which can be estimated from P- and U-samples alone. When ℓ is a proper loss (e.g. logistic) and the hypothesis class is sufficiently expressive, $\mathcal{L}_{\mathrm{uPU}}$ shares its population minimizer with $\mathcal{L}_{\mathrm{PN}}$; the minimizer $f^\star$ is calibrated to $p(c \mid x)$ and therefore serves as a drop-in replacement for a supervised classifier.

A.3 Diffusion Alignment as Sampling (DAS)

A *reward* $R : \mathcal{X}_0 \to \mathbb{R}$ assigns a scalar score to each clean sample x_0 according to how well it satisfies a desired property. DAS [31] draws samples from a reward-tilted target [31, Eq. (3)] $p_{\mathrm{DAS}}(x_0) \propto p_{\theta,0}(x_0) \exp(\lambda R(x_0))$ for guidance strength $\lambda > 0$. Because exact reward-conditioned marginals along the reverse process are intractable, DAS defines tractable intermediate targets [31, Eq. (11)]

$$\gamma_t(x_t) \;\propto\; p_{\theta,t}(x_t) \exp\big(\lambda_t R(\hat{x}_0(x_t))\big), \qquad 0 = \lambda_T \le \cdots \le \lambda_0 = \lambda,$$

which lift the data-space reward to noisy-sample space via the Tweedie estimate $\hat{x}_0(x_t)$. Sampling from this path uses SMC with a weighted particle system $\{(x_t^i, w_t^i)\}_{i=1}^M$, initialized with $w_T^i = 1$ (since $\lambda_T = 0$ makes $\gamma_T \propto p_{\theta,T}$, so particles drawn from $\mathcal{N}(0, I) = p_{\theta,T}$ already follow γ_T exactly). *Remark on notation.* [31] parameterize the guidance strength as $1/\alpha$ with a schedule $\lambda_t \in [0, 1]$; we use the equivalent form $\lambda_t \in [0, \lambda]$ with $\lambda := 1/\alpha$, matching the main text. (Here α is the DAS hyperparameter, distinct from the diffusion noise schedule α_t of Appendix A.1.)

Proposal kernel. The *locally optimal proposal* m_{t-1}^* is the kernel that minimizes the conditional variance $\mathrm{Var}(w_{t-1} \mid x_t)$ of the importance weight [31, Proposition 1]. Because sampling from m_{t-1}^* exactly is non-trivial, DAS approximates it [31, Eq. (13)] with a Gaussian,

$$m_{t-1}(x_{t-1} \mid x_t) \;:=\; \mathcal{N}\Big(x_{t-1};\, \underbrace{\mu_\theta(x_t, t) + \sigma_t^2\, \lambda_{t-1}\, \nabla_{x_t} R(\hat{x}_0(x_t))}_{\mu_t^{\mathrm{prop}}},\; \sigma_t^2 I\Big),$$

which shifts the DDPM mean by the gradient of the lifted reward.

Weight update derivation. The general SMC importance weight is

$$w_{t-1}(x_{t-1}) \;\propto\; w_t(x_t)\, \frac{\gamma_{t-1}(x_{t-1})\, L_t(x_t \mid x_{t-1})}{\gamma_t(x_t)\, m_{t-1}(x_{t-1} \mid x_t)},$$

where $L_t(x_t \mid x_{t-1})$ is a *backward kernel* that reverses the forward step. DAS chooses [31, Eq. (9)]

$$L_t(x_t \mid x_{t-1}) \;=\; \frac{p_\theta(x_{t-1} \mid x_t)\, p_{\theta,t}(x_t)}{p_{\theta,t-1}(x_{t-1})}.$$

Substituting $\gamma_t(x_t) = p_{\theta,t}(x_t)\exp(\lambda_t R(\hat{x}_0(x_t)))$ (dropping the common normalizing constants, which cancel in ratios) and the backward kernel:

$$\frac{\gamma_{t-1}(x_{t-1})\,L_t(x_t \mid x_{t-1})}{\gamma_t(x_t)} = \frac{p_{\theta,t-1}(x_{t-1})\exp(\lambda_{t-1}R(\hat{x}_0(x_{t-1})))}{p_{\theta,t}(x_t)\exp(\lambda_t R(\hat{x}_0(x_t)))} \cdot \frac{p_\theta(x_{t-1} \mid x_t)\,p_{\theta,t}(x_t)}{p_{\theta,t-1}(x_{t-1})}$$
$$= \frac{\exp(\lambda_{t-1}R(\hat{x}_0(x_{t-1})))\,p_\theta(x_{t-1} \mid x_t)}{\exp(\lambda_t R(\hat{x}_0(x_t)))}.$$

Since $p_\theta(\cdot \mid x_t)$ and $m_{t-1}(\cdot \mid x_t)$ share covariance $\sigma_t^2 I$, their ratio reduces to

$$r_t \;:=\; \frac{p_\theta(x_{t-1} \mid x_t)}{m_{t-1}(x_{t-1} \mid x_t)} \;=\; \exp\Big(\frac{1}{2\sigma_t^2}\big(\|x_{t-1}-\mu_t^{\text{prop}}\|^2 - \|x_{t-1}-\mu_\theta\|^2\big)\Big),$$

giving the weight recursion [31, Eq. (14)]

$$w_{t-1}(x_{t-1}) \;=\; w_t(x_t)\,\frac{\exp(\lambda_{t-1}R(\hat{x}_0(x_{t-1})))}{\exp(\lambda_t R(\hat{x}_0(x_t)))}\,r_t.$$

Asymptotic exactness. Under standard regularity assumptions [31, Appendix C.2.2], the weighted particle system is asymptotically exact [31, Proposition 8, Appendix C.2]: for any $\varphi \in \mathcal{C}_b(\mathcal{X}_0)$,

$$\sum_{i=1}^{M} \frac{w_0^i}{\sum_{k=1}^{M} w_0^k}\,\varphi(x_0^i) \;\xrightarrow{\text{a.s.}}\; p_{\text{DAS}}(\varphi) \qquad \text{as } M \to \infty.$$

Extension to our density path. Our method instead uses the density path (5) in Section 3.1, $\gamma_t(x_t) \propto p_{\theta,t}(x_t)(1-f_t(x_t))^{\lambda_t}$, with the optional parameterization $f_t(x_t) := g_0(\hat{x}_0(x_t))$ of Section 3.2, where g_0 is a PU classifier estimating $p_{\theta,0}(c \mid x_0)$. This corresponds to instantiating the DAS framework with $R(x_0) \leftarrow \log(1-g_0(x_0))$, since

$$\exp\big(\lambda_t R(\hat{x}_0(x_t))\big) \;=\; \exp\big(\lambda_t \log(1-g_0(\hat{x}_0(x_t)))\big) \;=\; (1-f_t(x_t))^{\lambda_t}.$$

Substituting into the weight recursion above yields

$$w_{t-1}(x_{t-1}) \;=\; w_t(x_t)\;r_t\;\frac{(1-f_{t-1}(x_{t-1}))^{\lambda_{t-1}}}{(1-f_t(x_t))^{\lambda_t}},$$

which is the recursion (6).

A.4 On the Exactness of Dynamic Negative Guidance

DNG [34] is exact under three conditions: exact scores, the limit $T \to \infty$, and repulsion strength $\eta = 1$. This subsection examines the last two.

The posterior recursion is exact only as $T \to \infty$. The guided score (3) needs the marginal posterior $p_t(c \mid x_t)$, whereas the recursion (4) is built from the trajectory $x_T, \ldots, x_t$ generated so far. For two reverse chains started at the same $x_T \sim \mathcal{N}(0, I)$, one unconditional and one conditioned on c, Bayes' rule gives

$$p(c \mid x_T, \ldots, x_t) \;=\; p(c)\prod_{s=t+1}^{T}\frac{p(x_{s-1} \mid x_s, c)}{p(x_{s-1} \mid x_s)} \;=\; p_t(c \mid x_t)\;\underbrace{\frac{p(x_T, \ldots, x_{t+1} \mid x_t, c)}{p(x_T, \ldots, x_{t+1} \mid x_t)}}_{(\star)}. \qquad (11)$$

The logarithm of the middle expression is exactly (4), so the recursion tracks the trajectory posterior, which equals the marginal one only when $(\star) = 1$. Koulischer et al. [34] prove $(\star) = 1$ in their Appendix C as a consequence of the Markov assumption, the forward chain carrying no condition. The argument presumes exact time reversal, which holds only as $T \to \infty$, the reverse kernel being Gaussian only for small steps. Table 4 measures the residual bias on the Gaussian mixture, where the scores are exact.

The tempered density is not guaranteed as a data-level target. At $\eta = 1$ the guided score is that of $p_t(\cdot \mid \neg c)$, the forward marginal of the data density $p_0(\cdot \mid \neg c)$, so the sampler reverses a genuine forward diffusion. At $\eta > 1$ the guidance instead follows $\nabla_{x_t} \log\big[p_t(x_t)\big(1 - p_t(c \mid x_t)\big)^{\eta}\big]$, aiming at the tempered densities in score form, and Footnote 2 of Koulischer et al. [34] states that this implies sampling from them. The implication does not hold in general: raising the second factor to a power does not commute with the integral defining the forward marginal, so the tempered densities need not be the forward marginals of any single data density. Theorem 3.2 of Bradley and Nakkiran [7] turns the same gap into a counterexample for CFG, where the sampler provably does not produce the powered density at a given guidance strength.

The analysis of Section 4 is stated under the complementary conditions: it is relative to the pretrained model's own reverse process, so it assumes no exact score, it is carried out on the discrete chain at finite T, and it holds for every avoidance strength $\lambda \geq 0$, with the remaining error controlled by the particle counts.

B Algorithms

Algorithm 2 gives detailed pseudocode for the reference-based setting (Section 3.2); Algorithm 3 for the score-based setting (Section 3.3). The following notation supplements Section 2.

- **Conditional Tweedie denoiser.** $\hat{x}_0(x_t \mid c) := \big(x_t - \sqrt{1-\bar{\alpha}_t}\,\hat{\epsilon}_\theta(x_t, t \mid c)\big)/\sqrt{\bar{\alpha}_t}$ is the conditional analog of $\hat{x}_0(x_t)$. Algorithm 3 uses it for the positive set because conditioning on c yields a sharper denoised view at intermediate t, easing discrimination. Both $\hat{x}_0(x_t)$ and $\hat{x}_0(x_t \mid c)$ reduce to the identity at $t{=}0$.
- **uPU oracle.** $\mathcal{O}_{\mathrm{uPU}}(\pi, \mathcal{P}, \mathcal{U}, f)$ denotes a few gradient steps minimizing the uPU risk (7) with positives $\mathcal{P}$, weighted unlabeled set $\mathcal{U}$, class prior π, warm-started from classifier f. At $t{=}T$ the oracle is called without f (cold start).
- **Time-varying class prior.** Both algorithms accept a schedule $\{\pi_t\}_{t=0}^{T}$. Setting $\pi_t \equiv \pi$ recovers the main-text formulation. A varying π_t is straightforward in practice since the theory requires only f_0 to be accurate at $t{=}0$ (Section 4).

Algorithm 2 Negative Alignment via SMC with Tweedie–uPU: Reference-Based Setting

Require: Pretrained DDPM reverse mean μ_θ and Tweedie denoiser $\hat{x}_0$ (see Section 2), noise schedule $\{\beta_t\}_{t=1}^T$, particle count M, resampling threshold $\mathrm{ESS}_{\min}$, class priors $\{\pi_t\}_{t=0}^T$, uPU oracle $\mathcal{O}_{\mathrm{uPU}}$, reference sampler producing $\{x_{c,t}^{(i)}\}_{i=1}^N$ at all t, proposal mean $\mu^{\mathrm{prop}}(x_t, t, \lambda, f)$ (default: (10)), tempering: precomputed $\{\lambda_t\}_{t=0}^T$ ($0 = \lambda_T \le \cdots \le \lambda_0$) **or** ESS target $\mathrm{ESS}_{\mathrm{tar}}$ for adaptive tempering.

Initialize $\{x_T^{(j)}\}_{j=1}^M \sim \mathcal{N}(0, I)$, $\quad w_{q,T}^{(j)} \leftarrow 1$, $\quad w_{\emptyset,T}^{(j)} \leftarrow 1$, $\quad \lambda_T \leftarrow 0$

$\mathcal{P}_T = \{\hat{x}_0(x_{c,T}^{(i)})\}_{i=1}^N$, $\quad \mathcal{U}_T = \{(\hat{x}_0(x_T^{(j)}), w_{\emptyset,T}^{(j)})\}_{j=1}^M$ ▷ Initialize classifier

$f_T \leftarrow \mathcal{O}_{\mathrm{uPU}}(\pi_T, \mathcal{P}_T, \mathcal{U}_T)$

for $t = T, T-1, \ldots, 1$ **do**

  if adaptive tempering **then**

    Find $\delta_t \ge 0$ s.t. $\mathrm{ESS}(\{w_{q,t}^{(j)}(1 - f_t(\hat{x}_0(x_t^{(j)})))^{\delta_t}\}) \approx \mathrm{ESS}_{\mathrm{tar}}$; $\quad \lambda_{t-1} \leftarrow \min(\lambda_t + \delta_t,\ \lambda_0)$

  else

    $\lambda_{t-1} \leftarrow$ precomputed value

  end if

  for $j = 1, \ldots, M$ **do** ▷ Proposal

    $\mu_t^{(j)} = \mu^{\mathrm{prop}}(x_t^{(j)}, t, \lambda_{t-1}, f_t)$

      // default: $\mu_\theta(x_t^{(j)}, t) + \sigma_t^2 \lambda_{t-1} \nabla_{x_t} \log(1 - f_t(\hat{x}_0(x_t^{(j)})))$

    $x_{t-1}^{(j)} \sim \mathcal{N}(\mu_t^{(j)}, \sigma_t^2 I)$

    $r_t^{(j)} \leftarrow \exp\Big(\frac{1}{2\sigma_t^2}\big(\|x_{t-1}^{(j)} - \mu_t^{(j)}\|^2 - \|x_{t-1}^{(j)} - \mu_\theta(x_t^{(j)}, t)\|^2\big)\Big)$

    $w_{\emptyset,t-1}^{(j)} \leftarrow w_{\emptyset,t}^{(j)}\, r_t^{(j)}$ ▷ Update $w_\emptyset$

  end for

  $\mathcal{P}_{t-1} = \{\hat{x}_0(x_{c,t-1}^{(i)})\}_{i=1}^N$, $\quad \mathcal{U}_{t-1} = \{(\hat{x}_0(x_{t-1}^{(j)}), w_{\emptyset,t-1}^{(j)})\}_{j=1}^M$ ▷ Refit classifier

  $f_{t-1} \leftarrow \mathcal{O}_{\mathrm{uPU}}(\pi_{t-1}, \mathcal{P}_{t-1}, \mathcal{U}_{t-1}, f_t)$

  for $j = 1, \ldots, M$ **do** ▷ Update w_q (6)

    $$w_{q,t-1}^{(j)} \leftarrow w_{q,t}^{(j)}\, r_t^{(j)} \frac{\big(1 - f_{t-1}(\hat{x}_0(x_{t-1}^{(j)}))\big)^{\lambda_{t-1}}}{\big(1 - f_t(\hat{x}_0(x_t^{(j)}))\big)^{\lambda_t}}$$

  end for

  if $\mathrm{ESS}(\{w_{q,t-1}^{(j)}\}) < \mathrm{ESS}_{\min}$ **then** ▷ Resampling

    Sample ancestors $a(1{:}M)$ using $\{w_{q,t-1}^{(j)}\}$

    for $j = 1, \ldots, M$ **do**

      $$x_{t-1}^{(j)} \leftarrow x_{t-1}^{(a(j))}, \quad w_{\emptyset,t-1}^{(j)} \leftarrow \frac{w_{\emptyset,t-1}^{(a(j))}}{w_{q,t-1}^{(a(j))}}, \quad w_{q,t-1}^{(j)} \leftarrow 1$$

    end for

  end if

end for

return $\{x_0^{(j)}, w_{q,0}^{(j)}\}_{j=1}^M$

Algorithm 3 Negative Alignment via SMC with Tweedie–uPU: Score-Based Setting

Require: Pretrained DDPM reverse means $\mu_\theta(x_t, t)$ and $\mu_\theta(x_t, t \mid c)$, Tweedie denoisers $\hat{x}_0(x_t)$ and $\hat{x}_0(x_t \mid c)$ (see Section 2 and above), noise schedule $\{\beta_t\}_{t=1}^T$, particle count M, resampling threshold $\mathrm{ESS}_{\min}$, class priors $\{\pi_t\}_{t=0}^T$, uPU oracle $\mathcal{O}_{\mathrm{uPU}}$, proposal mean $\mu^{\mathrm{prop}}(x_t, t, \lambda, f)$ (default: (10)), tempering: precomputed $\{\lambda_t\}_{t=0}^T$ ($0 = \lambda_T \le \cdots \le \lambda_0$) **or** ESS target $\mathrm{ESS}_{\mathrm{tar}}$ for adaptive tempering; initialize $\lambda_T \leftarrow 0$.

Initialize $\{x_T^{(j)}\}_{j=1}^M \sim \mathcal{N}(0, I)$, $w_{q,T}^{(j)} \leftarrow 1$, $w_{\emptyset,T}^{(j)} \leftarrow 1$, $w_{c,T}^{(j)} \leftarrow 1$ $\forall j$

$\mathcal{P}_T = \{(\hat{x}_0(x_T^{(j)} \mid c), w_{c,T}^{(j)})\}_{j=1}^M$, $\mathcal{U}_T = \{(\hat{x}_0(x_T^{(j)}), w_{\emptyset,T}^{(j)})\}_{j=1}^M$

$f_T \leftarrow \mathcal{O}_{\mathrm{uPU}}(\pi_T, \mathcal{P}_T, \mathcal{U}_T)$ ▷ Initialize classifier

for $t = T, T-1, \ldots, 1$ **do**

  if adaptive tempering **then**

    Find $\delta_t \ge 0$ s.t. $\mathrm{ESS}(\{w_{q,t}^{(j)}(1 - f_t(\hat{x}_0(x_t^{(j)})))^{\delta_t}\}) \approx \mathrm{ESS}_{\mathrm{tar}}$; $\lambda_{t-1} \leftarrow \min(\lambda_t + \delta_t,\ \lambda_0)$

  else

    $\lambda_{t-1} \leftarrow$ precomputed value

  end if

  for $j = 1, \ldots, M$ **do** ▷ Proposal

    $\mu_t^{(j)} = \mu^{\mathrm{prop}}(x_t^{(j)}, t, \lambda_{t-1}, f_t)$

      // default: $\mu_\theta(x_t^{(j)}, t) + \sigma_t^2 \lambda_{t-1} \nabla_{x_t} \log(1 - f_t(\hat{x}_0(x_t^{(j)})))$

    $x_{t-1}^{(j)} \sim \mathcal{N}(\mu_t^{(j)}, \sigma_t^2 I)$

    $r_t^{(j)} \leftarrow \exp\Big(\frac{1}{2\sigma_t^2}\big(\|x_{t-1}^{(j)} - \mu_t^{(j)}\|^2 - \|x_{t-1}^{(j)} - \mu_\theta(x_t^{(j)}, t)\|^2\big)\Big)$

    $w_{\emptyset,t-1}^{(j)} \leftarrow w_{\emptyset,t}^{(j)}\, r_t^{(j)}$ ▷ Update $w_\emptyset$

    $w_{c,t-1}^{(j)} \leftarrow w_{c,t}^{(j)} \exp\Big(\frac{1}{2\sigma_t^2}\big(\|x_{t-1}^{(j)} - \mu_t^{(j)}\|^2 - \|x_{t-1}^{(j)} - \mu_\theta(x_t^{(j)}, t \mid c)\|^2\big)\Big)$ ▷ Update w_c

  end for

  $\mathcal{P}_{t-1} = \{(\hat{x}_0(x_{t-1}^{(j)} \mid c), w_{c,t-1}^{(j)})\}_{j=1}^M$, $\mathcal{U}_{t-1} = \{(\hat{x}_0(x_{t-1}^{(j)}), w_{\emptyset,t-1}^{(j)})\}_{j=1}^M$

  $f_{t-1} \leftarrow \mathcal{O}_{\mathrm{uPU}}(\pi_{t-1}, \mathcal{P}_{t-1}, \mathcal{U}_{t-1}, f_t)$ ▷ Refit classifier

  for $j = 1, \ldots, M$ **do** ▷ Update w_q (6)

    $w_{q,t-1}^{(j)} \leftarrow w_{q,t}^{(j)}\, r_t^{(j)} \dfrac{(1 - f_{t-1}(\hat{x}_0(x_{t-1}^{(j)})))^{\lambda_{t-1}}}{(1 - f_t(\hat{x}_0(x_t^{(j)})))^{\lambda_t}}$

  end for

  if $\mathrm{ESS}(\{w_{q,t-1}^{(j)}\}) < \mathrm{ESS}_{\min}$ **then** ▷ Resampling

    Sample ancestors $a(1{:}M)$ using $\{w_{q,t-1}^{(j)}\}$

    for $j = 1, \ldots, M$ **do**

      $x_{t-1}^{(j)} \leftarrow x_{t-1}^{(a(j))}$, $w_{\emptyset,t-1}^{(j)} \leftarrow \dfrac{w_{\emptyset,t-1}^{(a(j))}}{w_{q,t-1}^{(a(j))}}$, $w_{c,t-1}^{(j)} \leftarrow \dfrac{w_{c,t-1}^{(a(j))}}{w_{q,t-1}^{(a(j))}}$, $w_{q,t-1}^{(j)} \leftarrow 1$

    end for

  end if

end for

return $\{x_0^{(j)}, w_{q,0}^{(j)}\}_{j=1}^M$

C Proof and Technical Detail

This appendix provides formal statements and proofs supporting Section 4. The object of analysis is the meta-algorithm Algorithm 1, in its reference-based instantiation (Section 3.2) and score-based instantiation (Section 3.3). We identify the avoidance target (1) p_{target} with the terminal Feynman–Kac measure

$$\nu_0(x) \;\propto\; (1-\eta_0(x))^{\lambda_0}\, p_{\theta,0}(x), \qquad \eta_0(x) := p_{\theta,0}(c \mid x),$$

and show that the weighted empirical measure $\widehat{\nu}_0^{M|M',N}$ produced by Algorithm 1 converges to ν_0 almost surely in the iterated limit $\lim_{M',N\to\infty}\lim_{M\to\infty}$.

For analytical clarity we adopt a proof-oriented variant in which classifier estimation is decoupled from the final SMC step. The classifier f_t is fit on an auxiliary particle system of size M' that shares the kernels of, but is independent of, the main system of size M. Positive samples for the PU risk are the N external references in the reference-based setting, and are supplied by the auxiliary system itself in the score-based setting. The proposal kernel is left arbitrary in the algorithm and regularized within the proof so that the importance weights remain uniformly bounded.

The argument has three steps. (i) Appendix C.1 establishes consistency of the weighted PU classifier at each fixed diffusion time, taking the positive and unlabeled samples as exogenous inputs satisfying suitable uniform laws of large numbers. (ii) The SMC construction underlying Algorithm 1 is shown to satisfy these uniform laws for both $p_{\theta,t}$ and $p_{\theta,t}(\cdot \mid c)$. (iii) A plug-in argument, combined with the telescoping of the tilting factors $(1-f_t)^{\lambda_t}$ in the recursion (6), yields convergence of the main particle system to ν_0. The reference-based instantiation serves as the master argument. The score-based instantiation follows by the same scheme, with the external positive samples replaced by the auxiliary particles reweighted by $w_{c,t}$, which retargets them to $p_{\theta,t}(\cdot \mid c)$ as in Section 3.3.

C.1 Consistency of Weighted PU Logistic Regression at a Fixed Diffusion Time

Fix a diffusion time t and write

$$\eta_t(x) := p_{\theta,t}(c \mid x), \qquad x \in \mathcal{X}_t,$$

for the conditional class posterior at time t. Let P_t denote the marginal law of x_t, let

$$\pi_t := P_t(c) \in (0,1),$$

and let P_t^+ denote the conditional law of x_t given c.

We observe:

- positive samples $X_1^+,\dots,X_N^+ \overset{\text{i.i.d.}}{\sim} P_t^+$;
- unlabeled samples $U_1,\dots,U_M$ together with nonnegative importance weights $w_t(U_j)$.

Define the normalized weights

$$\bar{w}_{t,j} := \frac{w_t(U_j)}{\sum_{k=1}^M w_t(U_k)}.$$

Let $\mathcal{G}_t$ be a class of measurable functions $g : \mathcal{X}_t \to \mathbb{R}$, and define

$$\ell_1(z) := \log(1+e^{-z}), \qquad \ell_0(z) := \log(1+e^{z}), \qquad \sigma(z) := \frac{1}{1+e^{-z}}.$$

For $g \in \mathcal{G}_t$, define the weighted empirical PU risk

$$\begin{aligned}\widehat{L}_{t,M,N}(g) &:= \pi_t \frac{1}{N}\sum_{i=1}^N \ell_1\big(g(X_i^+)\big) + \sum_{j=1}^M \bar{w}_{t,j}\,\ell_0\big(g(U_j)\big) - \pi_t \frac{1}{N}\sum_{i=1}^N \ell_0\big(g(X_i^+)\big) \\ &= \; -\pi_t \frac{1}{N}\sum_{i=1}^N g(X_i^+) + \sum_{j=1}^M \bar{w}_{t,j}\,\ell_0\big(g(U_j)\big). \end{aligned} \tag{12}$$

This is the logistic objective of the standard PU learning [41], with the unlabeled empirical average replaced by an importance-weighted one. Correspondingly, define the population PU risk

$$\begin{aligned} L_t^{\mathrm{PU}}(g) &:= \pi_t \, \mathbb{E}_{P_t^+}[\ell_1(g(X))] + \mathbb{E}_{P_t}[\ell_0(g(X))] - \pi_t \, \mathbb{E}_{P_t^+}[\ell_0(g(X))] \\ &= \; -\pi_t \, \mathbb{E}_{P_t^+}[g(X)] + \mathbb{E}_{P_t}[\ell_0(g(X))]. \end{aligned} \tag{13}$$

We impose the following assumptions.

(A1) **Positive-sample uniform law of large numbers.** There exists a function class $\mathcal{H}_t$ containing $\{g : g \in \mathcal{G}_t\} \cup \{\ell_1 \circ g : g \in \mathcal{G}_t\} \cup \{\ell_0 \circ g : g \in \mathcal{G}_t\}$ such that, as $N \to \infty$,

$$\sup_{h \in \mathcal{H}_t} \left| \frac{1}{N} \sum_{i=1}^{N} h(X_i^+) - \mathbb{E}_{P_t^+}[h(X)] \right| \to 0 \qquad \text{a.s.}$$

(A2) **Weighted unlabeled uniform convergence.** For the same class $\mathcal{H}_t$, as $M \to \infty$,

$$\sup_{h \in \mathcal{H}_t} \left| \sum_{j=1}^{M} \bar{w}_{t,j} h(U_j) - \mathbb{E}_{P_t}[h(X)] \right| \to 0 \qquad \text{a.s.}$$

(A3) **Realizability of the Bayes logit.** The posterior satisfies $\eta_t(x) \in (0,1)$ for P_t-a.e. x, and the Bayes logit $g_t^\star(x) := \log \frac{\eta_t(x)}{1-\eta_t(x)}$ belongs to $\mathcal{G}_t$.

Finally, assume that for every (M, N) under consideration, a measurable empirical minimizer exists:

$$\widehat{g}_{t,M,N} \in \arg\min_{g \in \mathcal{G}_t} \widehat{L}_{t,M,N}(g).$$

Define the induced probabilistic discriminator by $\widehat{f}_{t,M,N}(x) := \sigma(\widehat{g}_{t,M,N}(x))$.

Theorem 2 (Consistency of weighted PU logistic discrimination at a fixed diffusion time)**.** *Suppose* (A1)–(A3) *hold. Then, along any sequence* (M, N) *such that* $M \to \infty$ *and* $N \to \infty$*,*

$$\|\widehat{f}_{t,M,N} - \eta_t\|_{L^1(P_t)} \to 0 \qquad \textit{a.s.}$$

Proof. Let

$$\begin{aligned} \Delta_{M,N} &:= \sup_{g \in \mathcal{G}_t} \left| \widehat{L}_{t,M,N}(g) - L_t^{\mathrm{PU}}(g) \right| \\ &\leq \pi_t \sup_{g \in \mathcal{G}_t} \left| \frac{1}{N} \sum_{i=1}^{N} g(X_i^+) - \mathbb{E}_{P_t^+}[g(X)] \right| \; + \sup_{g \in \mathcal{G}_t} \left| \sum_{j=1}^{M} \bar{w}_{t,j} \ell_0(g(U_j)) - \mathbb{E}_{P_t}[\ell_0(g(X))] \right|. \end{aligned}$$

By Assumptions (A1) and (A2), $\Delta_{M,N} \to 0$ a.s.

We first rewrite the population PU risk. Let P_t^- denote the conditional law of x_t given $\neg c$. Since $P_t = \pi_t P_t^+ + (1-\pi_t) P_t^-$ and $\pi_t \, dP_t^+ = \eta_t \, dP_t$, we have

$$L_t^{\mathrm{PU}}(g) = \mathbb{E}_{P_t}[\eta_t \, \ell_1(g) + (1-\eta_t) \, \ell_0(g)] \, .$$

Writing $q_g = \sigma(g)$ and using $\ell_1(g) = -\log q_g$, $\ell_0(g) = -\log(1-q_g)$, this becomes

$$L_t^{\mathrm{PU}}(g) = \mathbb{E}_{P_t}[-\eta_t \log q_g - (1-\eta_t) \log(1-q_g)] \, .$$

Hence

$$L_t^{\mathrm{PU}}(g) - L_t^{\mathrm{PU}}(g_t^\star) = \mathbb{E}_{P_t}[\mathrm{KL}(\mathrm{Bern}(\eta_t) \, \| \, \mathrm{Bern}(q_g))] \, ,$$

and by Pinsker's inequality and Jensen's inequality,

$$L_t^{\mathrm{PU}}(g) - L_t^{\mathrm{PU}}(g_t^\star) \geq 2 \, \mathbb{E}_{P_t}\left[|\eta_t - q_g|^2\right] \geq 2 \, \|q_g - \eta_t\|_{L^1(P_t)}^2.$$

Next, since $\widehat{g}_{t,M,N}$ minimizes $\widehat{L}_{t,M,N}$ over $\mathcal{G}_t$,

$$\widehat{L}_{t,M,N}(\widehat{g}_{t,M,N}) \leq \widehat{L}_{t,M,N}(g_t^\star).$$

Hence

$$\begin{aligned} L_t^{\mathrm{PU}}(\widehat{g}_{t,M,N}) - L_t^{\mathrm{PU}}(g_t^\star) &= \Big(L_t^{\mathrm{PU}}(\widehat{g}_{t,M,N}) - \widehat{L}_{t,M,N}(\widehat{g}_{t,M,N})\Big) \\ &\quad + \Big(\widehat{L}_{t,M,N}(\widehat{g}_{t,M,N}) - \widehat{L}_{t,M,N}(g_t^\star)\Big) \\ &\quad + \Big(\widehat{L}_{t,M,N}(g_t^\star) - L_t^{\mathrm{PU}}(g_t^\star)\Big) \\ &\le \Big|L_t^{\mathrm{PU}}(\widehat{g}_{t,M,N}) - \widehat{L}_{t,M,N}(\widehat{g}_{t,M,N})\Big| + \Big|\widehat{L}_{t,M,N}(g_t^\star) - L_t^{\mathrm{PU}}(g_t^\star)\Big| \\ &\le 2 \sup_{g \in \mathcal{G}_t} \big|\widehat{L}_{t,M,N}(g) - L_t^{\mathrm{PU}}(g)\big| = 2\Delta_{M,N}. \end{aligned}$$

Since $\Delta_{M,N} \to 0$ a.s., it follows that

$$L_t^{\mathrm{PU}}(\widehat{g}_{t,M,N}) - L_t^{\mathrm{PU}}(g_t^\star) \to 0 \qquad \text{a.s.}$$

On the other hand, from the lower bound established above, applied with $g = \widehat{g}_{t,M,N}$, we have

$$L_t^{\mathrm{PU}}(\widehat{g}_{t,M,N}) - L_t^{\mathrm{PU}}(g_t^\star) \ge 2\,\|q_{\widehat{g}_{t,M,N}} - \eta_t\|_{L^1(P_t)}^2 = 2\,\|\widehat{f}_{t,M,N} - \eta_t\|_{L^1(P_t)}^2.$$

Therefore,

$$0 \le 2\,\|\widehat{f}_{t,M,N} - \eta_t\|_{L^1(P_t)}^2 \le L_t^{\mathrm{PU}}(\widehat{g}_{t,M,N}) - L_t^{\mathrm{PU}}(g_t^\star) \to 0 \qquad \text{a.s.}$$

and hence

$$\|\widehat{f}_{t,M,N} - \eta_t\|_{L^1(P_t)} \to 0 \qquad \text{a.s.}$$

□

C.2 Verification of Dual-Weight SMC Construction

This subsection establishes the theoretical validity of the dual-weight SMC construction. We show that, under a regularized common proposal, the resulting particle system simultaneously provides: (i) a consistent approximation of the unconditional marginal $p_{\theta,t}$, and (ii) a particle approximation of the tilted targets ν_t associated with the negative-alignment tempering.

The key ingredients are: a common proposal yielding uniformly bounded incremental weights, a telescoping structure of the associated Feynman–Kac models, and a ratio transform linking the tilted and unconditional marginals.

These results together imply that the particle system supplies the weighted empirical measure required in Assumption (A2).

We introduce the following assumption, used throughout the remainder of the appendix.

(A4) **Non-degeneracy of the conditional class posterior.** There exists $\delta \in (0,1)$ such that for all $t \in \{0,\dots,T\}$,

$$\delta \le \eta_t(x) \le 1 - \delta \qquad \text{for } p_{\theta,t}\text{-a.e. } x.$$

C.2.1 Regularized Common Proposal

Reverse kernels and tilted targets. Fix a diffusion horizon $T \in \mathbb{N}$. For each $t \in \{0,1,\dots,T\}$, let $p_{\theta,t}$ denote the unconditional marginal law at time t, and define the unconditional reverse diffusion kernel by

$$K_{t-1}(x_t, dx_{t-1}) := p_\theta(dx_{t-1} \mid x_t), \qquad t = 1,\dots,T.$$

Let $\{\lambda_t\}_{t=0}^T$ be a deterministic tempering schedule satisfying $0 = \lambda_T \le \lambda_{T-1} \le \cdots \le \lambda_0 < \infty$. Fix a measurable version $\eta_t^\sharp$ of η_t, and define the clipped posterior

$$\bar{\eta}_t(x) := \min\{1-\delta, \max\{\eta_t^\sharp(x), \delta\}\} \quad \in [\delta, 1-\delta] \quad \text{for all } x.$$

Define

$$\omega_t(x) := \big(1 - \bar{\eta}_t(x)\big)^{\lambda_t}, \qquad \nu_t(dx) := \frac{\omega_t(x)\, p_{\theta,t}(dx)}{\int \omega_t(x)\, p_{\theta,t}(dx)}.$$

By Assumption (A4), we have $\bar{\eta}_t(x) = \eta_t(x)$ for $p_{\theta,t}$-a.e. x, so the definition of ν_t is unchanged. Since $\lambda_T = 0$, we have $\omega_T \equiv 1$ and $\nu_T = p_{\theta,T}$. Under Assumption (A4),

$$\delta^{\lambda_0} \le \omega_t(x) \le 1 \qquad \text{for all } x. \tag{14}$$

Regularized proposal. Let $\widetilde{M}_{t-1}(x_t, dx_{t-1})$ be any measurable Markov kernel (the nominal guided proposal). Fix $\rho \in (0,1)$ and define the *regularized common proposal*

$$M^{\mathrm{reg}}_{t-1}(x_t, dx_{t-1}) := (1-\rho)\,\widetilde{M}_{t-1}(x_t, dx_{t-1}) + \rho\, K_{t-1}(x_t, dx_{t-1}). \tag{15}$$

Lemma 3 (Regularized common proposal)**.** *For any nominal proposal $\widetilde{M}_{t-1}$ and any $\rho \in (0,1)$, the mixture kernel M^{reg}_{t-1} satisfies*

$$K_{t-1}(x_t, \cdot) \ll M^{\mathrm{reg}}_{t-1}(x_t, \cdot), \qquad \frac{dK_{t-1}(x_t, \cdot)}{dM^{\mathrm{reg}}_{t-1}(x_t, \cdot)}(x_{t-1}) \le \rho^{-1}$$

for $M^{\mathrm{reg}}_{t-1}(x_t, \cdot)$-a.e. x_{t-1} and every x_t.

Proof. By construction, $M^{\mathrm{reg}}_{t-1}(x_t, A) \ge \rho\, K_{t-1}(x_t, A)$ for every measurable set A. Absolute continuity is immediate, and the Radon–Nikodym bound follows. □

C.2.2 Feynman–Kac Models and Path Measures

Define the unconditional and tilted incremental weights on the common proposal. For $t = 1, \dots, T$, let

$$G^{\emptyset}_{t-1}(x_t, x_{t-1}) := \frac{dK_{t-1}(x_t, \cdot)}{dM^{\mathrm{reg}}_{t-1}(x_t, \cdot)}(x_{t-1}), \tag{16}$$

and

$$G^{q}_{t-1}(x_t, x_{t-1}) := G^{\emptyset}_{t-1}(x_t, x_{t-1}) \frac{\omega_{t-1}(x_{t-1})}{\omega_t(x_t)}. \tag{17}$$

Lemma 4 (Uniform boundedness of both potentials)**.** *Under Assumption* (A4)*,*

$$0 \le G^{\emptyset}_{t-1}(x_t, x_{t-1}) \le \rho^{-1}, \qquad 0 \le G^{q}_{t-1}(x_t, x_{t-1}) \le \rho^{-1}\delta^{-\lambda_0},$$

for all $t = 1, \dots, T$, for all x_t, and for $M^{\mathrm{reg}}_{t-1}(x_t, \cdot)$-a.e. x_{t-1}.

Proof. The bound on $G^{\emptyset}$ is Lemma 3. By Eq. (14), $\omega_{t-1}(x_{t-1}) \le 1$ and $\omega_t(x_t)^{-1} \le \delta^{-\lambda_0}$, giving $G^{q}_{t-1} \le \rho^{-1}\delta^{-\lambda_0}$. □

Path measures. Define the common proposal path measure

$$\mathbf{M}(dx_{0:T}) := p_{\theta,T}(dx_T) \prod_{s=1}^{T} M^{\mathrm{reg}}_{s-1}(x_s, dx_{s-1}),$$

and the unnormalized path weights

$$W^{a}_{t}(x_{t:T}) := \prod_{s=t+1}^{T} G^{a}_{s-1}(x_s, x_{s-1}), \qquad a \in \{\emptyset, q\},$$

for $t = 0, \dots, T$ (empty products equal 1). Define the associated unnormalized path measures $\mathbf{\Gamma}^{a}_{t}(dx_{t:T}) := W^{a}_{t}(x_{t:T})\, p_{\theta,T}(dx_T) \prod_{s=t+1}^{T} M^{\mathrm{reg}}_{s-1}(x_s, dx_{s-1})$.

Proposition 5 (Identification of the path models)**.** *For every $t \in \{0, \dots, T\}$:*

(i) *the x_t-marginal of $\mathbf{\Gamma}^{\emptyset}_{t}$ is $p_{\theta,t}$;*

(ii) *the x_t-marginal of $\mathbf{\Gamma}^{q}_{t}$ is $\omega_t(x)\, p_{\theta,t}(dx)$, so the normalized marginal is ν_t.*

Proof. Part (i). Cancellation of proposal densities gives

$$\mathbf{\Gamma}^{\emptyset}_{t}(dx_{t:T}) = p_{\theta,T}(dx_T) \prod_{s=t+1}^{T} K_{s-1}(x_s, dx_{s-1}),$$

the unconditional reverse diffusion path law. Its x_t-marginal is $p_{\theta,t}$.

Part (ii). Using the definition (17), the product of tilted-to-unconditional ratios telescopes:

$$\prod_{s=t+1}^{T} \frac{\omega_{s-1}(x_{s-1})}{\omega_s(x_s)} = \frac{\omega_t(x_t)}{\omega_T(x_T)} = \omega_t(x_t),$$

since $\omega_T \equiv 1$. Therefore $W_t^q(x_{t:T}) = W_t^\emptyset(x_{t:T})\,\omega_t(x_t)$, and taking the x_t-marginal yields $\omega_t(x_t)\,p_{\theta,t}(dx_t)$. □

Ratio recovery. Define $R_t(x) := \omega_t(x)^{-1}$. By Eq. (14), $1 \le R_t(x) \le \delta^{-\lambda_0}$.

Lemma 6 (Ratio identity)**.** *For every bounded measurable φ,*

$$p_{\theta,t}(\varphi) = \frac{\nu_t(R_t\varphi)}{\nu_t(R_t)}. \tag{18}$$

Proof. By definition of ν_t, $\nu_t(R_t\varphi) = \frac{\int R_t(x)\varphi(x)\,\omega_t(x)\,p_{\theta,t}(dx)}{\int \omega_t(x)\,p_{\theta,t}(dx)} = \frac{\int \varphi(x)\,p_{\theta,t}(dx)}{\int \omega_t(x)\,p_{\theta,t}(dx)}$ since $R_t\,\omega_t \equiv 1$. Similarly, $\nu_t(R_t) = 1/\int \omega_t\,dp_{\theta,t}$. Taking the ratio gives the identity (18). □

C.2.3 Particle Approximation and Uniform Convergence

We introduce one further regularity assumption:

(A5) **Total boundedness of the class in (A1)–(A2).** For the same class $\mathcal{H}_t$ as in Assumptions (A1)–(A2), $\mathcal{H}_t$ is uniformly bounded and totally bounded under the supremum norm $\|\cdot\|_\infty$.

Remark. Assumption (A4) implies that the true posterior satisfies $\eta_t(x) \in [\delta, 1-\delta]$, so the Bayes logit $g_t^\star(x) = \log \frac{\eta_t(x)}{1-\eta_t(x)}$ is uniformly bounded. Hence, without excluding the target minimizer, one may work with a uniformly bounded hypothesis class $\mathcal{G}_t$ containing $g_t^\star$. For such a class, the logistic loss transforms $\ell_1 \circ g$ and $\ell_0 \circ g$ are also uniformly bounded, and it is therefore natural to impose total boundedness on the enlarged class $\mathcal{H}_t$ used in Assumptions (A1)–(A2).

SMC convergence for the tilted model. Consider an SMC sampler with multinomial resampling targeting the Feynman–Kac model with proposal $\{M_{t-1}^{\mathrm{reg}}\}$ and potentials $\{G_{t-1}^q\}$. Let $\nu_t^M := \sum_{j=1}^M \bar{w}_{q,t}^j\,\delta_{X_t^j}$ denote the normalized particle approximation of ν_t. Since the potentials are uniformly bounded (Lemma 4), the standard strong law of large numbers (LLN) for particle filters with multinomial resampling [11, Prop. 11.4] gives, for each fixed t and each bounded measurable φ,

$$\nu_t^M(\varphi) \xrightarrow{\text{a.s.}} \nu_t(\varphi).$$

To verify Assumption (A2), we require convergence that is uniform over a class of test functions. We therefore strengthen the above pointwise convergence to a uniform one over $\mathcal{H}_t$.

Lemma 7 (Uniform SMC convergence over a totally bounded class)**.** *Under Assumption* (A5)*,*

$$\sup_{h\in\mathcal{H}_t} |\nu_t^M(h) - \nu_t(h)| \xrightarrow{\text{a.s.}} 0.$$

Proof. Let $\varepsilon > 0$ and take a finite ε-net $\{h_1, \dots, h_K\} \subset \mathcal{H}_t$ in $\|\cdot\|_\infty$. For each $h \in \mathcal{H}_t$, choose h_i with $\|h - h_i\|_\infty \le \varepsilon$. Then

$$\sup_{h\in\mathcal{H}_t} |\nu_t^M(h) - \nu_t(h)| \le \max_{1\le i\le K} |\nu_t^M(h_i) - \nu_t(h_i)| + 2\varepsilon.$$

Each term converges to zero almost surely by the SMC LLN. Since the maximum is over finitely many i, it also converges to zero almost surely. Letting $M \to \infty$ and then $\varepsilon \downarrow 0$ yields the claim. □

Recovering $p_{\theta,t}$ via ratio transform. Define the particle-based estimator

$$\widehat{P}_t^M(\varphi) := \frac{\nu_t^M(R_t\varphi)}{\nu_t^M(R_t)}.$$

Theorem 8 (SMC verification of weighted uniform convergence)**.** *Assume* (A4) *and* (A5)*. Then for each fixed t,*

$$\sup_{h\in\mathcal{H}_t} \left|\widehat{P}_t^M(h) - p_{\theta,t}(h)\right| \xrightarrow{\text{a.s.}} 0.$$

In particular, defining

$$\widetilde{w}_{t,j}^M := \frac{\bar{w}_{q,t}^j\, R_t(X_t^j)}{\sum_{k=1}^M \bar{w}_{q,t}^k\, R_t(X_t^k)},$$

we have

$$\widehat{P}_t^M(h) = \sum_{j=1}^M \widetilde{w}_{t,j}^M\, h(X_t^j),$$

and hence Assumption (A2) *holds for the weighted sample $\{(X_t^j, \widetilde{w}_{t,j}^M)\}_{j=1}^M$ and the class $\mathcal{H}_t$.*

Proof. Fix t. By the SMC LLN, for every bounded measurable φ,

$$\nu_t^M(\varphi) \xrightarrow{\text{a.s.}} \nu_t(\varphi).$$

In particular,

$$\nu_t^M(R_t) \xrightarrow{\text{a.s.}} \nu_t(R_t).$$

Since $R_t \geq 1$ by Assumption (A4) and Eq. (14), we have $\nu_t(R_t) \geq 1$. Hence, on an event of probability one, there exists M_0 such that for all $M \geq M_0$,

$$\nu_t^M(R_t) \geq \frac{1}{2}.$$

Next, since R_t is bounded and $\mathcal{H}_t$ is totally bounded, the class $R_t\mathcal{H}_t := \{R_t h : \ h \in \mathcal{H}_t\}$ is uniformly bounded and totally bounded. An ε-net argument identical to that in Lemma 7 yields

$$\sup_{h\in\mathcal{H}_t} |\nu_t^M(R_t h) - \nu_t(R_t h)| \xrightarrow{\text{a.s.}} 0.$$

Using the ratio identity

$$p_{\theta,t}(h) = \frac{\nu_t(R_t h)}{\nu_t(R_t)},$$

we write

$$\widehat{P}_t^M(h) - p_{\theta,t}(h) = \frac{\nu_t^M(R_t h) - \nu_t(R_t h)}{\nu_t^M(R_t)} + \nu_t(R_t h)\left(\frac{1}{\nu_t^M(R_t)} - \frac{1}{\nu_t(R_t)}\right).$$

Therefore, for all sufficiently large M,

$$\sup_{h\in\mathcal{H}_t} |\widehat{P}_t^M(h) - p_{\theta,t}(h)| \leq 2 \sup_{h\in\mathcal{H}_t} |\nu_t^M(R_t h) - \nu_t(R_t h)| + \sup_{h\in\mathcal{H}_t} |\nu_t(R_t h)| \left|\frac{1}{\nu_t^M(R_t)} - \frac{1}{\nu_t(R_t)}\right|.$$

Since

$$\sup_{h\in\mathcal{H}_t} |\nu_t(R_t h)| \leq \|R_t\|_\infty \sup_{h\in\mathcal{H}_t} \|h\|_\infty < \infty,$$

the first term converges to zero almost surely, and the second term also converges to zero almost surely because $\nu_t^M(R_t) \to \nu_t(R_t) > 0$. This proves the result. □

Combining Theorems 2 and 8, we obtain the following.

Corollary 9 (Fixed-time classifier consistency)**.** *Under Assumptions* (A1) *and* (A3)–(A5)*, and assuming that a measurable empirical minimizer exists for each (M, N), for each fixed t and along any sequence with $M, N \to \infty$,*

$$\left\|\widehat{f}_{t,M,N} - \eta_t\right\|_{L^1(P_t)} \longrightarrow 0 \qquad \textit{a.s.}$$

Proof. By Theorem 8, the weighted empirical measure constructed from $\{(X_t^j, \widetilde{w}_{t,j}^M)\}_{j=1}^M$ satisfies Assumption (A2). The result then follows from Theorem 2 under Assumptions (A1) and (A3). □

Lemma 10 (Asymptotic exactness under oracle classifiers)**.** *Assume* (A4)*, and suppose the classifiers are oracle: $f_t = \eta_t$ for all t. Then the SMC algorithm with potentials $\{G_{t-1}^q\}$ and multinomial resampling satisfies, for every bounded measurable φ,*

$$\nu_0^M(\varphi) \xrightarrow{\text{a.s.}} \nu_0(\varphi), \qquad M \to \infty.$$

Proof. Under oracle classifiers, the potentials $\{G_{t-1}^q\}$ are deterministic and, by Lemma 4, uniformly bounded. Therefore the standard strong law for SMC with multinomial resampling [11, Prop. 11.4] applies and yields the claim. □

C.3 Asymptotic Exactness of the Reference-Based Instantiation with Learned Classifiers

We now combine the consistency result of Appendix C.1 with the SMC approximation of Appendix C.2 to establish asymptotic exactness of the reference-based instantiation of Algorithm 1 when the classifiers are learned during sampling. The proof-oriented variant introduced at the start of this appendix renders the learned classifiers exogenous, so the two results apply without modification.

For each $t = 0, \dots, T$, let $\widehat{f}_t^{M',N}$ denote the classifier trained at time t from the auxiliary system, using M' unlabeled particles and N samples from $p_{\theta,t}(\cdot \mid c)$. Define its clipped version by

$$\bar{f}_t^{M',N}(x) := \min\{1-\delta, \max(\widehat{f}_t^{M',N}(x), \delta)\},$$

where $\delta \in (0,1)$ is the constant in Assumption (A4), and set

$$\widehat{\omega}_t^{M',N}(x) := \big(1 - \bar{f}_t^{M',N}(x)\big)^{\lambda_t}.$$

The clipping yields $\delta \le \bar{f}_t^{M',N}(x) \le 1-\delta$ for all x, and hence $\delta^{\lambda_0} \le \widehat{\omega}_t^{M',N}(x) \le 1$.

The auxiliary system enters the analysis only through Assumption (A2), namely as a weighted approximation of $p_{\theta,t}$. Two constructions suffice. The first is an auxiliary particle system run with the regularized Feynman–Kac construction of Appendix C.2, independent of the main one. The second is direct simulation of the unconditional diffusion with unit weights. Both yield a consistent approximation of $p_{\theta,t}$ and satisfy Assumption (A2).

We emphasize that the specific form of the proposal kernel is not essential for the asymptotic argument. What matters is that the associated importance weights define the correct Feynman–Kac change of measure and remain uniformly bounded. Accordingly, we allow an arbitrary measurable proposal kernel $\widehat{M}_{t-1}^{M',N}(x_t, \cdot)$ and regularize it by

$$\widehat{M}_{t-1}^{M',N,\mathrm{reg}}(x_t, dx_{t-1}) := (1-\rho)\, \widehat{M}_{t-1}^{M',N}(x_t, dx_{t-1}) + \rho\, K_{t-1}(x_t, dx_{t-1}),$$

where K_{t-1} is the reference reverse kernel from Appendix C.2 and $\rho \in (0,1)$ is fixed. As in that section, this mixture guarantees

$$K_{t-1}(x_t, \cdot) \ll \widehat{M}_{t-1}^{M',N,\mathrm{reg}}(x_t, \cdot), \qquad \frac{dK_{t-1}(x_t,\cdot)}{d\widehat{M}_{t-1}^{M',N,\mathrm{reg}}(x_t,\cdot)}(x_{t-1}) \le \rho^{-1},$$

so the Radon–Nikodym component of the incremental weight is uniformly bounded. Thus, the role of the proposal in this subsection is purely instrumental: any regularized proposal for which the corresponding potentials are bounded is admissible.

Given the learned classifiers, define the plug-in incremental weights by

$$\begin{aligned} \widehat{G}_{t-1}^{q,M',N}(x_t, x_{t-1}) &:= \frac{dK_{t-1}(x_t,\cdot)}{d\widehat{M}_{t-1}^{M',N,\mathrm{reg}}(x_t,\cdot)}(x_{t-1}) \frac{\widehat{\omega}_{t-1}^{M',N}(x_{t-1})}{\widehat{\omega}_t^{M',N}(x_t)}, \\ \widehat{G}_{t-1}^{\emptyset,M',N}(x_t, x_{t-1}) &:= \frac{dK_{t-1}(x_t,\cdot)}{d\widehat{M}_{t-1}^{M',N,\mathrm{reg}}(x_t,\cdot)}(x_{t-1}). \end{aligned} \tag{19}$$

The second family tracks the unconditional model, while the first tracks the tilted model obtained by replacing the oracle factor $(1-\eta_t)^{\lambda_t}$ with its learned plug-in counterpart. The analysis therefore

depends only on the behavior of the weights in Eq. (19), not on any particular choice of proposal mechanism.

Finally, define the target density at terminal time by

$$\nu_0(dx) := \frac{(1-\eta_0(x))^{\lambda_0}\, p_{\theta,0}(dx)}{\int (1-\eta_0(x))^{\lambda_0}\, p_{\theta,0}(dx)}.$$

Since the fixed-time consistency result applies at each t, in particular at $t = 0$, the asymptotic behavior of the algorithm is determined by whether the SMC procedure correctly tracks the corresponding Feynman–Kac model. Under the above construction, this is ensured by the boundedness of the incremental weights together with the standard SMC convergence results.

Theorem 11 (Asymptotic exactness of the reference-based instantiation in the analysis variant)**.** *Assume* (A1) *and* (A3)–(A5)*. Consider the proof-oriented variant described above, in which*

- *for each $t = 0, \dots, T$, the classifier $\widehat{f}_t^{M',N}$ is trained from an auxiliary sample independent of the particle system used in the final SMC approximation; and*
- *the final particle system uses the regularized proposal $\widehat{M}_{t-1}^{M',N,\mathrm{reg}}$ and plug-in potentials $\{\widehat{G}_{t-1}^{q,M',N}\}_{t=1}^{T}$ defined in Eq.* (19)*.*

Then, for every bounded measurable $\varphi : \mathcal{X}_0 \to \mathbb{R}$,

$$\lim_{M',N\to\infty} \lim_{M\to\infty} \widehat{\nu}_0^{M\,|\,M',N}(\varphi) = \nu_0(\varphi) \qquad \textit{a.s.}$$

This is the formal version of Theorem 1 for the reference-based instantiation of Section 3.2.

Proof. We separate the total error into: (I) the SMC approximation error for a fixed target, and (II) the error induced by replacing the true target with a plug-in estimate. For fixed (M', N), let

$$\widetilde{\nu}_0^{M',N}(dx) := \frac{\widehat{\omega}_0^{M',N}(x)\, p_{\theta,0}(dx)}{\int \widehat{\omega}_0^{M',N}(x)\, p_{\theta,0}(dx)}$$

denote the terminal target induced by the learned classifiers, and let $\widehat{\nu}_0^{M\,|\,M',N}$ denote the weighted empirical output of the main particle system with M particles constructed using $\{\widehat{f}_t^{M',N}\}_{t=0}^{T}$. We decompose the error as

$$\widehat{\nu}_0^{M\,|\,M',N}(\varphi) - \nu_0(\varphi) = \underbrace{\left(\widehat{\nu}_0^{M\,|\,M',N}(\varphi) - \widetilde{\nu}_0^{M',N}(\varphi)\right)}_{\text{(I): SMC error}} + \underbrace{\left(\widetilde{\nu}_0^{M',N}(\varphi) - \nu_0(\varphi)\right)}_{\text{(II): plug-in target error}}. \tag{20}$$

Step 1: SMC error for the main particle system. Fix (M', N). Since the classifiers are trained from an auxiliary sample independent of the main particle system, the learned quantities

$$\left\{\widehat{\omega}_t^{M',N},\ \widehat{M}_{t-1}^{M',N,\mathrm{reg}},\ \widehat{G}_{t-1}^{q,M',N},\ \widehat{G}_{t-1}^{\emptyset,M',N}\right\}$$

are external to the main SMC approximation. Therefore, conditional on the auxiliary sample, the main particle system is an ordinary Feynman–Kac model with deterministic proposal kernels and deterministic potentials.

By Lemma 3, $0 \le \widehat{G}_{t-1}^{\emptyset,M',N} \le \rho^{-1}$. Moreover, by construction of the clipped classifier used to define $\widehat{\omega}_t^{M',N}$, we have $\delta^{\lambda_0} \le \widehat{\omega}_t^{M',N}(x) \le 1$ for all t, x. Therefore

$$0 \le \widehat{G}_{t-1}^{q,M',N}(x_t, x_{t-1}) \le \rho^{-1}\delta^{-\lambda_0}.$$

Thus the standard strong law for SMC with multinomial resampling [11, Prop. 11.4] applies to the main particle system for this fixed environment, and yields

$$\widehat{\nu}_0^{M\,|\,M',N}(\varphi) - \widetilde{\nu}_0^{M',N}(\varphi) \to 0 \qquad \text{a.s. as } M \to \infty,$$

conditionally on the auxiliary sample. Since this holds for almost every auxiliary realization, it also holds jointly almost surely.

Step 2: Identification of the induced target. Applying the same telescoping argument as in Proposition 5(ii) to the learned tempering factors,

$$\prod_{s=1}^{T} \frac{\widehat{\omega}_{s-1}^{M',N}(x_{s-1})}{\widehat{\omega}_{s}^{M',N}(x_s)} = \frac{\widehat{\omega}_{0}^{M',N}(x_0)}{\widehat{\omega}_{T}^{M',N}(x_T)} = \widehat{\omega}_{0}^{M',N}(x_0),$$

since $\lambda_T = 0$ implies $\widehat{\omega}_T^{M',N} \equiv 1$. Therefore the terminal marginal induced by the learned Feynman–Kac model is exactly $\widetilde{\nu}_0^{M',N}$. In particular, the induced terminal target depends only on $\widehat{\omega}_0^{M',N}$, although the intermediate classifiers may still affect the finite-particle behavior through the proposal and the incremental-weight decomposition.

Step 3: Plug-in convergence of the induced target. By construction of the auxiliary sample, Assumption (A2) holds at $t = 0$. Therefore Corollary 9 applies at $t = 0$, and yields

$$\|\widehat{f}_0^{M',N} - \eta_0\|_{L^1(p_{\theta,0})} \to 0 \qquad \text{a.s. as } M', N \to \infty.$$

Let

$$\mathrm{clip}(u) := \min\{1-\delta, \max(u, \delta)\}, \qquad u \in \mathbb{R}.$$

Then $\mathrm{clip}(\cdot)$ is 1-Lipschitz, and by Assumption (A4) we have $\mathrm{clip}(\eta_0) = \eta_0$ a.s.

Since $\bar{f}_0^{M',N} = \mathrm{clip}(\widehat{f}_0^{M',N})$, it follows that

$$\|\bar{f}_0^{M',N} - \eta_0\|_{L^1(p_{\theta,0})} \le \|\widehat{f}_0^{M',N} - \eta_0\|_{L^1(p_{\theta,0})} \to 0 \qquad \text{a.s.}$$

Now the map $u \mapsto (1-u)^{\lambda_0}$ is Lipschitz on the compact interval $[\delta, 1-\delta]$, so

$$\|\widehat{\omega}_0^{M',N} - \omega_0\|_{L^1(p_{\theta,0})} \to 0 \qquad \text{a.s. as } M', N \to \infty.$$

Since $\widehat{\omega}_0^{M',N}(x) \ge \delta^{\lambda_0}$ and $\omega_0(x) \ge \delta^{\lambda_0}$ for all x, the corresponding normalizing constants are bounded away from zero. Therefore, for every bounded measurable φ,

$$\widetilde{\nu}_0^{M',N}(\varphi) \to \nu_0(\varphi) \qquad \text{a.s. as } M', N \to \infty.$$

This shows that term (II) in Eq. (20) vanishes.

Step 4: Conclusion. For each fixed (M', N), Step 1 gives $\widehat{\nu}_0^{M \mid M',N}(\varphi) - \widetilde{\nu}_0^{M',N}(\varphi) \to 0$ a.s. as $M \to \infty$. Step 3 yields $\widetilde{\nu}_0^{M',N}(\varphi) \to \nu_0(\varphi)$ a.s. as $M', N \to \infty$. Hence,

$$\lim_{M',N\to\infty} \lim_{M\to\infty} \widehat{\nu}_0^{M \mid M',N}(\varphi) = \nu_0(\varphi) \qquad \text{a.s.},$$

which proves the claim. □

Relation to DAS and conceptual interpretation. Theorem 11 is closely related to the asymptotic exactness result for DAS [31], but applies in a setting where the reward is replaced by a learned, data-dependent discriminator.

- **(i) Learned, time-dependent potentials.** In DAS, the reward R is fixed and known; here, the potentials depend on learned classifiers $\widehat{f}_t^{M',N}$ that are both data-dependent and time-dependent. Despite this additional randomness, asymptotic exactness is preserved.
- **(ii) Dual-weight tracking and ratio recovery.** The algorithm tracks two sets of weights (w_q and $w_\emptyset$) on the same proposal. The $w_\emptyset$ weights recover the unconditional marginal $p_{\theta,t}$ via a ratio transform that cancels the tilting factor algebraically, regardless of classifier accuracy.
- **(iii) Cancellation of intermediate errors.** The telescoping structure ensures that the terminal marginal depends only on $\widehat{\omega}_0$, not on intermediate classifiers. Combined with the ratio recovery in (ii), in the particle limit, classifier errors at intermediate times $t > 0$ do not affect the unlabeled sample quality nor the terminal output density.

C.4 Extension to the Score-Based Instantiation: Dual Weighted Particle Systems

We now treat the score-based instantiation, in which the positive density $P_t^+ = p_{\theta,t}(\cdot \mid c)$ is approximated by a second weighted particle system on the same proposal trajectories (cf. outline at the start of this appendix). Throughout this subsection, we retain all notation from Appendix C.2 and Appendix C.3.

Conditional reverse kernel. For $t = 1, \dots, T$, define $K_{t-1}^c(x_t, dx_{t-1}) := p_\theta(dx_{t-1} \mid x_t, c)$. Since both conditional and unconditional diffusion marginals share the same Gaussian limit, we take $p_{\theta,T}(\cdot \mid c) = p_{\theta,T}$.

Three-way common proposal. Let $\widetilde{M}_{t-1}$ be any nominal guided proposal. For $\rho_\emptyset, \rho_c \in (0,1)$ with $\rho_\emptyset + \rho_c < 1$, define

$$M_{t-1}^{\mathrm{alg2}}(x_t, dx_{t-1}) := (1 - \rho_\emptyset - \rho_c)\, \widetilde{M}_{t-1}(x_t, dx_{t-1}) + \rho_\emptyset\, K_{t-1}(x_t, dx_{t-1}) + \rho_c\, K_{t-1}^c(x_t, dx_{t-1}). \tag{21}$$

Lemma 12 (Three-way common proposal)**.** *The kernels* $\{M_{t-1}^{\mathrm{alg2}}\}$ *satisfy*

$$\frac{dK_{t-1}(x_t, \cdot)}{dM_{t-1}^{\mathrm{alg2}}(x_t, \cdot)} \le \rho_\emptyset^{-1}, \qquad \frac{dK_{t-1}^c(x_t, \cdot)}{dM_{t-1}^{\mathrm{alg2}}(x_t, \cdot)} \le \rho_c^{-1}.$$

Proof. By construction, $M_{t-1}^{\mathrm{alg2}} \ge \rho_\emptyset K_{t-1}$ and $M_{t-1}^{\mathrm{alg2}} \ge \rho_c K_{t-1}^c$. The Radon–Nikodym bounds follow. □

Three Feynman–Kac models. Define

$$G_{t-1}^\emptyset(x_t, x_{t-1}) := \frac{dK_{t-1}(x_t, \cdot)}{dM_{t-1}^{\mathrm{alg2}}(x_t, \cdot)}(x_{t-1}), \tag{22}$$

$$G_{t-1}^c(x_t, x_{t-1}) := \frac{dK_{t-1}^c(x_t, \cdot)}{dM_{t-1}^{\mathrm{alg2}}(x_t, \cdot)}(x_{t-1}), \tag{23}$$

$$G_{t-1}^q(x_t, x_{t-1}) := G_{t-1}^\emptyset(x_t, x_{t-1}) \frac{\omega_{t-1}(x_{t-1})}{\omega_t(x_t)}. \tag{24}$$

Lemma 13 (Uniform boundedness of three potentials)**.** *Under* (A4) *(with the full two-sided bound),*

$$0 \le G_{t-1}^\emptyset \le \rho_\emptyset^{-1}, \qquad 0 \le G_{t-1}^c \le \rho_c^{-1}, \qquad 0 \le G_{t-1}^q \le \rho_\emptyset^{-1} \delta^{-\lambda_0}.$$

Proof. The first two bounds follow from Lemma 12; the third follows as in Lemma 4. □

Positive-side particle approximation. All three Feynman–Kac models are built on the same proposal trajectories $X_t^1, \dots, X_t^M$, but the corresponding target marginals differ. To justify a uniform approximation of the positive conditional law $P_t^+ = p_{\theta,t}(\cdot \mid c)$, we proceed exactly as in Theorem 8: rather than invoking a direct SMC law of large numbers for the w_c-weighted empirical measure, we recover P_t^+ from the tilted target ν_t by a ratio transform.

Recall that

$$\nu_t(dx) = \frac{\omega_t(x)\, p_{\theta,t}(dx)}{\int \omega_t(x)\, p_{\theta,t}(dx)}, \qquad P_t^+(dx) = p_{\theta,t}(dx \mid c),$$

and define

$$S_t(x) := \frac{\eta_t(x)}{\omega_t(x)}.$$

Since $\eta_t(x) \in [\delta, 1-\delta]$ under Assumption (A4) and $\omega_t(x) = (1 - \eta_t(x))^{\lambda_t}$ $p_{\theta,t}$-a.e., we have

$$0 \le S_t(x) \le \delta^{-\lambda_0} \qquad p_{\theta,t}\text{-a.e.}$$

Hence S_t is bounded, and therefore the class $S_t\mathcal{H}_t := \{S_t h : h \in \mathcal{H}_t\}$ is uniformly bounded and totally bounded whenever $\mathcal{H}_t$ is. Define the particle-based estimator

$$\widehat{P}_t^{+,M}(h) := \frac{\nu_t^M(S_t h)}{\nu_t^M(S_t)}, \qquad h \in \mathcal{H}_t, \qquad \text{where} \qquad \nu_t^M := \sum_{j=1}^M \bar{w}_{q,t}^j \, \delta_{X_t^j}$$

is the tilted particle approximation from Appendix C.2. Equivalently, if we define

$$\bar{w}_{c,t}^j := \frac{\bar{w}_{q,t}^j \, S_t(X_t^j)}{\sum_{k=1}^M \bar{w}_{q,t}^k \, S_t(X_t^k)}, \qquad j = 1, \ldots, M, \qquad \text{then} \qquad \widehat{P}_t^{+,M}(h) = \sum_{j=1}^M \bar{w}_{c,t}^j \, h(X_t^j).$$

Theorem 14 (SMC verification of positive-side weighted uniform convergence)**.** *Assume* (A4) *and* (A5)*. Then for each fixed t,*

$$\sup_{h \in \mathcal{H}_t} \left| \widehat{P}_t^{+,M}(h) - P_t^+(h) \right| \xrightarrow[M\to\infty]{\text{a.s.}} 0.$$

Equivalently,

$$\sup_{h \in \mathcal{H}_t} \left| \sum_{j=1}^M \bar{w}_{c,t}^j \, h(X_t^j) - P_t^+(h) \right| \xrightarrow[M\to\infty]{\text{a.s.}} 0.$$

Proof. Fix t. Since the tilted potentials $\{G_{t-1}^q\}$ are uniformly bounded by Lemma 13, the same SMC law of large numbers used in Appendix C.2 applies to ν_t^M. Therefore, for every bounded measurable φ,

$$\nu_t^M(\varphi) \xrightarrow{\text{a.s.}} \nu_t(\varphi).$$

We first record the positive-side ratio identity. For every bounded measurable h,

$$\nu_t(S_t h) = \frac{\int S_t(x) h(x) \omega_t(x) \, p_{\theta,t}(dx)}{\int \omega_t(x) \, p_{\theta,t}(dx)} = \frac{\int \eta_t(x) h(x) \, p_{\theta,t}(dx)}{\int \omega_t(x) \, p_{\theta,t}(dx)}.$$

Since $\eta_t(x) = p_{\theta,t}(c \mid x)$ and $\int \eta_t(x) \, p_{\theta,t}(dx) = P_t(c)$, we obtain

$$\nu_t(S_t h) = \frac{P_t(c) \, P_t^+(h)}{\int \omega_t(x) \, p_{\theta,t}(dx)}, \qquad \nu_t(S_t) = \frac{P_t(c)}{\int \omega_t(x) \, p_{\theta,t}(dx)}.$$

Hence

$$P_t^+(h) = \frac{\nu_t(S_t h)}{\nu_t(S_t)}.$$

Next, since S_t is bounded and $\mathcal{H}_t$ is totally bounded, the class

$$S_t\mathcal{H}_t := \{S_t h : h \in \mathcal{H}_t\}$$

is also uniformly bounded and totally bounded in the supremum norm. Therefore, exactly as in Lemma 7, $\sup_{h \in \mathcal{H}_t} \left| \nu_t^M(S_t h) - \nu_t(S_t h) \right| \xrightarrow{\text{a.s.}} 0$. In particular, $\nu_t^M(S_t) \xrightarrow{\text{a.s.}} \nu_t(S_t)$.

Moreover, $\nu_t(S_t) > 0$ because

$$\nu_t(S_t) = \frac{P_t(c)}{\int \omega_t(x) \, p_{\theta,t}(dx)},$$

and $P_t(c) \in (0,1)$ by the standing assumptions of Appendix C.1. Hence, on an event of probability one, there exists M_0 such that $\nu_t^M(S_t) \geq \nu_t(S_t)/2 > 0$ for all $M \geq M_0$.

Using the ratio identity above, for all sufficiently large $M \geq M_0$ we have

$$\widehat{P}_t^{+,M}(h) - P_t^+(h) = \frac{\nu_t^M(S_t h) - \nu_t(S_t h)}{\nu_t^M(S_t)} + \nu_t(S_t h) \left(\frac{1}{\nu_t^M(S_t)} - \frac{1}{\nu_t(S_t)} \right).$$

Therefore,

$$\sup_{h \in \mathcal{H}_t} \left| \widehat{P}_t^{+,M}(h) - P_t^+(h) \right| \leq \frac{2}{\nu_t(S_t)} \sup_{h \in \mathcal{H}_t} \left| \nu_t^M(S_t h) - \nu_t(S_t h) \right| + \sup_{h \in \mathcal{H}_t} \left| \nu_t(S_t h) \right| \left| \frac{1}{\nu_t^M(S_t)} - \frac{1}{\nu_t(S_t)} \right|.$$

Since

$$\sup_{h\in\mathcal{H}_t}|\nu_t(S_t h)| \le \|S_t\|_\infty \sup_{h\in\mathcal{H}_t}\|h\|_\infty < \infty,$$

the first term converges to zero almost surely by the uniform convergence over $S_t\mathcal{H}_t$, and the second term also converges to zero almost surely because $\nu_t^M(S_t) \to \nu_t(S_t) > 0$. This proves

$$\sup_{h\in\mathcal{H}_t}\left|\widehat{P}_t^{+,M}(h) - P_t^+(h)\right| \to 0 \qquad \text{a.s.}$$

Finally, by the definition of $\bar{w}_{c,t}^j$, $\widehat{P}_t^{+,M}(h) = \sum_{j=1}^M \bar{w}_{c,t}^j\, h(X_t^j)$, so the equivalent weighted-empirical formulation follows. □

Dual weighted PU consistency. Using the positive-side estimator from Theorem 14 and the unlabeled estimator from Theorem 8, define the dual weighted empirical PU risk by

$$\widehat{L}_{t,M}^{\mathrm{dual}}(g) := -\pi_t \sum_{j=1}^M \bar{w}_{c,t}^j\, g(X_t^j) + \sum_{j=1}^M \widetilde{w}_t^j\, \ell_0\big(g(X_t^j)\big), \qquad g \in \mathcal{G}_t, \tag{25}$$

where $\{\bar{w}_{c,t}^j\}_{j=1}^M$ are the normalized weights defining the positive conditional estimator $\widehat{P}_t^{+,M}(h) = \sum_{j=1}^M \bar{w}_{c,t}^j\, h(X_t^j)$, and $\{\widetilde{w}_t^j\}_{j=1}^M$ are the normalized weights defining the ratio-recovered unlabeled estimator $\widehat{P}_t^M(h) = \sum_{j=1}^M \widetilde{w}_t^j\, h(X_t^j)$. Its population counterpart is

$$L_t^{\mathrm{PU}}(g) := -\pi_t\, P_t^+\big(g\big) + P_t\big(\ell_0 \circ g\big),$$

which is exactly the same population PU risk as in Appendix C.1.

Assume, as in Appendix C.1, that for every M under consideration there exists a measurable empirical minimizer $\widehat{g}_{t,M}^{\mathrm{dual}} \in \arg\min_{g\in\mathcal{G}_t} \widehat{L}_{t,M}^{\mathrm{dual}}(g)$, and define $\widehat{f}_{t,M}^{\mathrm{dual}}(x) := \sigma\big(\widehat{g}_{t,M}^{\mathrm{dual}}(x)\big)$.

Theorem 15 (Fixed-time consistency under dual weighted particles)**.** *Under* (A3)–(A5)*, for each fixed t,*

$$\big\|\widehat{f}_{t,M}^{\mathrm{dual}} - \eta_t\big\|_{L^1(P_t)} \xrightarrow[M\to\infty]{\text{a.s.}} 0.$$

Proof. Fix t. By Theorem 14, for the class $\mathcal{H}_t$, $\sup_{h\in\mathcal{H}_t}\left|\widehat{P}_t^{+,M}(h) - P_t^+(h)\right| \xrightarrow{\text{a.s.}} 0$. By Theorem 8, $\sup_{h\in\mathcal{H}_t}\left|\widehat{P}_t^M(h) - P_t(h)\right| \xrightarrow{\text{a.s.}} 0$. Since $\mathcal{H}_t$ contains $\mathcal{G}_t$ and $\{\ell_0 \circ g : g \in \mathcal{G}_t\}$ by Assumption (A1) (inherited from Appendix C.1), it follows that

$$\begin{aligned}
\Delta_M^{\mathrm{dual}} &:= \sup_{g\in\mathcal{G}_t}\left|\widehat{L}_{t,M}^{\mathrm{dual}}(g) - L_t^{\mathrm{PU}}(g)\right| \\
&\le \pi_t \sup_{g\in\mathcal{G}_t}\left|\widehat{P}_t^{+,M}(g) - P_t^+(g)\right| + \sup_{g\in\mathcal{G}_t}\left|\widehat{P}_t^M(\ell_0\circ g) - P_t(\ell_0\circ g)\right| \\
&\xrightarrow{\text{a.s.}} 0.
\end{aligned}$$

Thus the dual weighted empirical risk satisfies the same uniform convergence property as the empirical risk in Theorem 2. Moreover, its population target is the same risk L_t^{PU}. Therefore the proof of Theorem 2 applies verbatim with $\Delta_{M,N}$ replaced by Δ_M^{dual}. Using Assumption (A3) and the existence of the measurable empirical minimizer $\widehat{g}_{t,M}^{\mathrm{dual}}$, we obtain

$$\big\|\widehat{f}_{t,M}^{\mathrm{dual}} - \eta_t\big\|_{L^1(P_t)} \xrightarrow{\text{a.s.}} 0.$$

□

Proof-oriented variant and its asymptotic exactness. The proof-oriented variant of Appendix C.3 carries over: at each t, the classifier is trained from an auxiliary system, here a dual-weight particle system of size M' independent of the main one.

For each t, let $\widehat{f}_t^{M'}$ denote the classifier trained from the auxiliary system, and define its clipped version by

$$\bar{f}_t^{M'}(x) := \min\{1-\delta, \max(\widehat{f}_t^{M'}(x), \delta)\},$$

where $\delta \in (0,1)$ is the constant in Assumption (A4), and set

$$\widehat{\omega}_t^{M'}(x) := \big(1 - \bar{f}_t^{M'}(x)\big)^{\lambda_t}.$$

This yields $\delta^{\lambda_0} \le \widehat{\omega}_t^{M'}(x) \le 1$ for all t and x.

Let the main particle system use these clipped classifiers as exogenous objects, together with any measurable proposal kernels regularized as in Appendix C.4, and denote by $\widehat{\nu}_0^{M \mid M'}$ its weighted empirical output at time 0. Further define the plug-in terminal target

$$\widetilde{\nu}_0^{M'}(dx) := \frac{\widehat{\omega}_0^{M'}(x)\, p_{\theta,0}(dx)}{\int \widehat{\omega}_0^{M'}(x)\, p_{\theta,0}(dx)}.$$

Theorem 16 (Asymptotic exactness of the score-based instantiation in the analysis variant)**.** *Assume* (A3)–(A5)*. Consider the proof-oriented variant described above, in which the classifiers are trained from an auxiliary dual-weight particle system independent of the main particle system used in the final SMC approximation. Assume moreover that, for each auxiliary size M' and each t, a measurable empirical minimizer defining $\widehat{f}_t^{M'}$ exists.*

Then, for every bounded measurable $\varphi : \mathcal{X}_0 \to \mathbb{R}$,

$$\lim_{M' \to \infty} \lim_{M \to \infty} \widehat{\nu}_0^{M \mid M'}(\varphi) = \nu_0(\varphi) \qquad \textit{a.s.},$$

where

$$\nu_0(dx) := \frac{(1 - \eta_0(x))^{\lambda_0}\, p_{\theta,0}(dx)}{\int (1 - \eta_0(x))^{\lambda_0}\, p_{\theta,0}(dx)}.$$

This establishes the extension of Theorem 1 to the score-based instantiation of Section 3.3, as stated informally at the end of Section 4.

Proof. Fix a bounded measurable test function φ. As in the proof of Theorem 11, decompose

$$\widehat{\nu}_0^{M \mid M'}(\varphi) - \nu_0(\varphi) = \underbrace{\Big(\widehat{\nu}_0^{M \mid M'}(\varphi) - \widetilde{\nu}_0^{M'}(\varphi)\Big)}_{\text{(I): SMC error}} + \underbrace{\Big(\widetilde{\nu}_0^{M'}(\varphi) - \nu_0(\varphi)\Big)}_{\text{(II): plug-in target error}}. \tag{26}$$

Step 1: SMC error for fixed learned classifiers. Fix M'. Because the auxiliary particle system is independent of the main one, the collection

$$\Big\{\bar{f}_t^{M'}, \widehat{\omega}_t^{M'}\Big\}_{t=0}^{T}$$

is deterministic from the viewpoint of the main particle system. Hence, conditional on the auxiliary system, the main particle system is an ordinary Feynman–Kac model with deterministic proposal kernels and deterministic potentials.

By the three-way regularized common-proposal construction of Appendix C.4, the Radon–Nikodym part of the incremental weights is uniformly bounded. Since

$$\delta^{\lambda_0} \le \widehat{\omega}_t^{M'}(x) \le 1 \qquad \text{for all } t, x,$$

the corresponding plug-in tilted potentials are also uniformly bounded. Therefore the standard strong law for SMC with multinomial resampling applies, exactly as in Appendix C.3, and yields

$$\widehat{\nu}_0^{M \mid M'}(\varphi) - \widetilde{\nu}_0^{M'}(\varphi) \xrightarrow[M \to \infty]{\text{a.s.}} 0$$

conditionally on the auxiliary system, hence jointly almost surely.

Step 2: Identification of the induced terminal target. For fixed M', the plug-in tilted potentials have the same telescoping structure as in the reference-based instantiation:

$$\prod_{s=1}^{T} \frac{\widehat{\omega}_{s-1}^{M'}(x_{s-1})}{\widehat{\omega}_s^{M'}(x_s)} = \frac{\widehat{\omega}_0^{M'}(x_0)}{\widehat{\omega}_T^{M'}(x_T)}.$$

Since $\lambda_T = 0$, we have $\widehat{\omega}_T^{M'} \equiv 1$. Hence the terminal marginal induced by the learned Feynman–Kac model is exactly

$$\widetilde{\nu}_0^{M'}(dx) = \frac{\widehat{\omega}_0^{M'}(x)\, p_{\theta,0}(dx)}{\int \widehat{\omega}_0^{M'}(x)\, p_{\theta,0}(dx)}.$$

As in the reference-based instantiation, intermediate classifier errors may affect the proposal mechanism and finite-particle performance, but not the form of the induced terminal target.

Step 3: Plug-in convergence of the terminal target. By Theorem 15 applied at $t = 0$,

$$\|\widehat{f}_0^{M'} - \eta_0\|_{L^1(P_0)} \xrightarrow[M'\to\infty]{\text{a.s.}} 0.$$

Since the clipping map

$$\mathrm{clip}(u) := \min\{1-\delta, \max(u, \delta)\}$$

is 1-Lipschitz and, by Assumption (A4), $\eta_0(x) \in [\delta, 1-\delta]$ for P_0-a.e. x, we have $\mathrm{clip}(\eta_0) = \eta_0$ P_0-a.e., and therefore

$$\|\bar{f}_0^{M'} - \eta_0\|_{L^1(P_0)} \le \|\widehat{f}_0^{M'} - \eta_0\|_{L^1(P_0)} \xrightarrow[M'\to\infty]{\text{a.s.}} 0.$$

Now the map

$$u \mapsto (1-u)^{\lambda_0}$$

is Lipschitz on the compact interval $[\delta, 1-\delta]$. Hence

$$\|\widehat{\omega}_0^{M'} - \omega_0\|_{L^1(P_0)} \xrightarrow[M'\to\infty]{\text{a.s.}} 0, \qquad \omega_0(x) := (1-\eta_0(x))^{\lambda_0}.$$

Moreover,

$$\widehat{\omega}_0^{M'}(x) \ge \delta^{\lambda_0}, \qquad \omega_0(x) \ge \delta^{\lambda_0},$$

so the corresponding normalizing constants are uniformly bounded away from zero. Therefore, for every bounded measurable φ,

$$\widetilde{\nu}_0^{M'}(\varphi) \xrightarrow[M'\to\infty]{\text{a.s.}} \nu_0(\varphi).$$

Thus term (II) in Eq. (26) vanishes.

Step 4: Conclusion. For each fixed M', Step 1 gives

$$\lim_{M\to\infty} \widehat{\nu}_0^{M\,|\,M'}(\varphi) = \widetilde{\nu}_0^{M'}(\varphi) \qquad \text{a.s.}$$

Step 3 gives

$$\widetilde{\nu}_0^{M'}(\varphi) \xrightarrow[M'\to\infty]{\text{a.s.}} \nu_0(\varphi).$$

Combining these two limits proves

$$\lim_{M'\to\infty} \lim_{M\to\infty} \widehat{\nu}_0^{M\,|\,M'}(\varphi) = \nu_0(\varphi) \qquad \text{a.s.},$$

as claimed. □

D Experimental Details and Additional Results

This appendix expands Section 5 with full setups and additional results. Section D.1 covers the score-based instantiation (Section 3.3, Algorithm 3), and Section D.2 the reference-based instantiation (Section 3.2, Algorithm 2). Section D.3 scales the score-based instantiation to Stable Diffusion. Code and per-experiment documentation are provided in the supplementary material.

Compute. GMM experiments run on CPU (Intel Xeon Max 9480). CIFAR-10 and Stable Diffusion experiments run on a single NVIDIA GH200 GPU. Python and library versions are pinned in the supplementary material.

D.1 GMM with Score-Based Setting

This appendix gives the full specification of the GMM benchmark in Section 5.1. The closed-form structure of the GMM yields exact expressions for conditional and unconditional scores $\nabla_{x_t} \log p_t(x_t \mid \cdot)$, which lets us evaluate samplers against ground-truth quantities without confounding from a learned score model.

D.1.1 Setup

Data density and class partition. The unconditional density $p(x)$ is a 12-component GMM in $\mathbb{R}^3$, generated as follows. **Mode means.** For $i = 0, \dots, 11$, let $\tilde{\theta}_i = 2\pi i/12 + \delta_i$ with $\delta_i \sim U(-0.2, 0.2)$, $r_i = 10(1 + \varepsilon_i)$ with $\varepsilon_i \sim U(-0.3, 0.3)$, and $z_i \sim U(-1, 1)$, all independent. Set $\tilde{\mu}_i = (r_i \cos\tilde{\theta}_i,\ r_i \sin\tilde{\theta}_i,\ z_i)^\top$ and center the means by $\mu_i = \tilde{\mu}_i - \frac{1}{12}\sum_j \tilde{\mu}_j$. **Mode covariances.** For each mode i, define base unit vectors $\mathbf{e}_r = (\cos\tilde{\theta}_i, \sin\tilde{\theta}_i, 0)^\top$ and $\mathbf{e}_t = (-\sin\tilde{\theta}_i, \cos\tilde{\theta}_i, 0)^\top$. Draw $\rho \sim U(0, 0.3)$, $\alpha \sim U(0, 2\pi)$, and set $\phi = \alpha\rho$. Rotate the base frame in the $(\mathbf{e}_r, \mathbf{e}_t)$ plane: $\mathbf{u}_0 = \cos\phi\, \mathbf{e}_r + \sin\phi\, \mathbf{e}_t$ and $\mathbf{v}_0 = -\sin\phi\, \mathbf{e}_r + \cos\phi\, \mathbf{e}_t$. Perturb the first axis with $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, 0.01\, I_3)$ and normalize: $\mathbf{u} = (\mathbf{u}_0 + \boldsymbol{\xi})/\|\mathbf{u}_0 + \boldsymbol{\xi}\|$. Re-orthogonalize via Gram–Schmidt: $\mathbf{v} = (\mathbf{v}_0 - (\mathbf{v}_0^\top \mathbf{u})\mathbf{u})/\|\mathbf{v}_0 - (\mathbf{v}_0^\top \mathbf{u})\mathbf{u}\|$ and $\mathbf{w} = \mathbf{u} \times \mathbf{v}$. Draw eigenvalues $\lambda_r \sim U(1.0, 2.5)$, $\lambda_t \sim U(0.3, 0.8)$, $\lambda_z \sim U(0.5, 2.0)$ independently, and set $\Sigma_i = R_i\, \mathrm{diag}(\lambda_r, \lambda_t, \lambda_z)\, R_i^\top$ with $R_i = [\mathbf{u}\ \ \mathbf{v}\ \ \mathbf{w}]$. **Weights and class partition.** Draw $(w_0, \dots, w_{11}) \sim \mathrm{Dir}(5 \cdot \mathbf{1}_{12})$. Select an index set $\mathcal{P} \subset \{0, \dots, 11\}$ of size 3 uniformly without replacement, and let $\mathcal{N}$ be its complement. Renormalize within each class: $w_i^+ = w_i / \sum_{j \in \mathcal{P}} w_j$ for $i \in \mathcal{P}$ and $w_i^- = w_i / \sum_{j \in \mathcal{N}} w_j$ for $i \in \mathcal{N}$. The joint density is $p(x) = \pi \sum_{i \in \mathcal{P}} w_i^+\, \mathcal{N}(x; \mu_i, \Sigma_i) + (1 - \pi) \sum_{i \in \mathcal{N}} w_i^-\, \mathcal{N}(x; \mu_i, \Sigma_i)$ with class prior $\pi = 3/12 = 0.25$. All randomness above is controlled by a single integer layout seed.

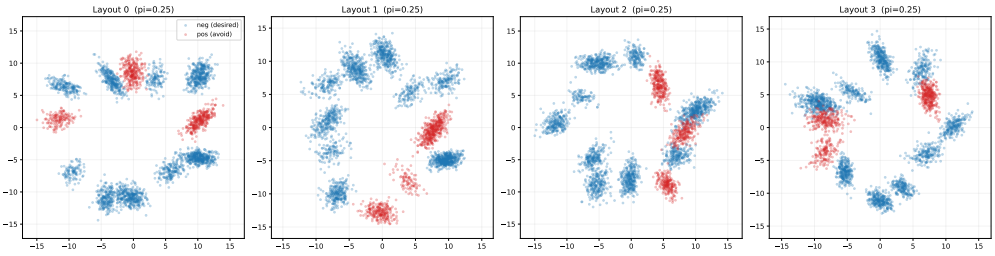


Figure 4: Four GMM layouts drawn at random. Each panel projects 3000 unconditional samples onto the xy-plane. Red points are samples from the avoided component density $p(\cdot \mid c)$ and blue points are samples from $p(\cdot \mid \neg c)$.

Diffusion process and ground-truth quantities. We use the cosine schedule of variance-preserving DDPM [40], offset $s = 0.008$, $\beta_{\max} = 0.999$, and $T = 100$ timesteps. Each marginal $p_t(\cdot \mid c)$ and $p_t(\cdot \mid \neg c)$ is a closed-form GMM with means $\sqrt{\bar{\alpha}_t}\mu_k$ and covariances $\bar{\alpha}_t \Sigma_k + (1 - \bar{\alpha}_t) I$, so the conditional scores and the posterior $p_t(c \mid x_t)$ are analytic. DNG and ours in Algorithm 3 receive the unconditional score $\nabla_{x_t} \log p_t(x_t)$ and the conditional score $\nabla_{x_t} \log p_t(x_t \mid c)$.

Table 2: NegCFG η sweep, 50 trials, mean±std. Mode coverage$= \mathbb{P}[\min_{k\in\neg c} \|x - \mu_k\| < 5]$.

η	Leak (mean) $\downarrow$	$W_2 \downarrow$	Mode coverage $\uparrow$
0.0 (uncond)	0.253 ± 0.016	3.6 ± 0.6	1.00
0.1	0.193 ± 0.014	2.7 ± 0.4	1.00
0.2	0.150 ± 0.013	3.5 ± 0.7	1.00
0.5	0.068 ± 0.008	9.7 ± 2.4	0.99
0.7	0.041 ± 0.007	14.1 ± 3.8	0.97
1.0	0.018 ± 0.005	20.5 ± 6.3	0.94
1.5	0.005 ± 0.002	34.6 ± 17.3	0.85
2.0	0.001 ± 0.001	50.9 ± 34.5	0.74
3.0	0.000 ± 0.000	102.0 ± 107.2	0.50

Table 3: DNG sweep, 50 trials. The main-text comparison uses $(\lambda_0, \tau) = (1, 1)$.

	$\tau = 1.0$		$\tau = 0.2$	
λ_0	Leak $\downarrow$	$W_2 \downarrow$	Leak $\downarrow$	$W_2 \downarrow$
0.5	0.101 ± 0.013	2.2 ± 0.4	0.164 ± 0.013	2.3 ± 0.4
1.0	0.052 ± 0.010	1.6 ± 0.3	0.105 ± 0.011	2.1 ± 0.3
2.0	0.015 ± 0.005	3.5 ± 0.7	0.040 ± 0.006	4.6 ± 0.7
5.0	0.001 ± 0.001	7.8 ± 1.1	0.002 ± 0.001	9.5 ± 1.1
10.0	0.000 ± 0.000	11.1 ± 1.2	0.000 ± 0.000	12.9 ± 1.6
20.0	0.000 ± 0.000	14.6 ± 1.6	0.000 ± 0.000	15.7 ± 1.9

Methods and sweep grids. We compare four samplers below.

- **Oracle.** DDPM sampling with $\nabla_{x_t} \log p_t(x_t \mid \neg c)$.
- **NegCFG.** Repulsion strength $\eta \in \{0, 0.1, 0.2, 0.5, 0.7, 1.0, 1.5, 2.0, 3.0\}$.
- **DNG** [34]. The CFG strength is $-\eta\ \hat{p}_t(c|x_t)/(1 - \hat{p}_t(c|x_t))$. We follow their implementation in [34], which estimates $\hat{p}_t(c|x_t)$ by $\log \hat{p}_{t-1}(c|x_{t-1}) \leftarrow \log \hat{p}_t(c|x_t) - \frac{\tau}{2\sigma_t^2}\left(\|x_{t-1} - \mu_\theta(x_t, t|c)\|^2 - \|x_{t-1} - \mu_\theta(x_t, t)\|^2\right) + \frac{\delta}{2\sigma_t^2}$, introducing the hyperparameters τ and δ in place of the original update (4). We set $\hat{p}_T(c \mid x_T) = \pi$ and $\delta = 0$, and sweep $\lambda_0 \in \{0.5, 1, 2, 5, 10, 20\}$ at $\tau \in \{0.2, 1.0\}$; the main-text comparison uses $(\lambda_0, \tau) = (1, 1)$, written as "DNG ($\eta$=1)" there.
- **Ours.** Algorithm 3 with $\mu_t^{\text{prop}} = \mu_\theta(x_t, t) + \sigma_t^2 \lambda_{t-1} \nabla_{x_t} \log(1 - f_t(\hat{x}_0(x_t)))$ of (10).

Online PU classifier. $f_t = g_0 \circ \hat{x}_0$ is a kernel logistic regression in $\hat{x}_0$ space, refit at every step. The dictionary at step t comprises n_p=128 centers subsampled from the conditional Tweedie estimates and n_u=64 from the unconditional ones, taken over the current proposal particles. The bandwidth uses the median heuristic scaled by 0.3. Weights are obtained by L-BFGS minimization of the uPU logistic risk (7) with an RKHS-norm penalty $\frac{\lambda_{\text{reg}}}{2}\|g\|^2_{\mathcal{H}_k}$, $\lambda_{\text{reg}} = 10^{-4}$; the uPU class prior is annealed as $\pi_t = 0.02 + 0.13 \exp(-5(1 - t/T))$.

SMC parameters. Tempering of λ_t is adaptive: $\delta_t \in [0, 1.5/T]$ is found by binary search so that the ESS of $\{w^j_{q,t}(1 - f_t(\hat{x}_0(x^j_t)))^{\delta_t}\}$ matches the target $0.9M$, and $\lambda_{t-1} \leftarrow \min(\lambda_t + \delta_t, 1)$. ESS-based resampling [11] fires below $0.9M$ and is disabled for the final 5 steps.

D.1.2 Results

Baseline sweeps. Tables 2 and 3 report the full sweeps for NegCFG and DNG. Both methods exhibit a clear trade-off between leak and W_2. NegCFG drives leak monotonically toward zero as η grows, but W_2 inflates and mode coverage drops below 1 for $\eta \geq 0.7$, indicating mode collapse onto a few $\neg c$-modes. DNG shows the same trade-off at both $\tau \in \{0.2, 1.0\}$, and the configuration closest to the Oracle leak ($\lambda_0 = 2, \tau = 1$) still attains $W_2 = 3.5$, more than twice the Oracle. These patterns are consistent with the local-repulsion limitations discussed in Section 2.

Table 4: Timestep-count ablation for DNG on the Gaussian mixture, mean±std over 15 trials, at $(\lambda_0, \tau) = (1, 1)$ and M=1024. The Oracle is recomputed at each T, and Gap is the difference between the two leak columns.

T	Leak ↓	Oracle leak	Gap
100	0.053 ± 0.010	0.012	0.041
200	0.040 ± 0.009	0.012	0.028
500	0.028 ± 0.006	0.010	0.018
1000	0.023 ± 0.008	0.011	0.012
2000	0.020 ± 0.009	0.011	0.009

Table 5: Particle-count ablation on the Gaussian mixture with true scores, mean±std over 15 trials. The Oracle is recomputed at each M. The row at M=1024 differs from Table 1, which averages over 50 trials.

M	Oracle leak	Ours leak ↓	Oracle W_2	Ours W_2 ↓	ESS/M	Resamples
64	0.011	0.009 ± 0.014	3.6	3.5 ± 0.7	0.96	3.1
256	0.014	0.021 ± 0.028	2.5	2.1 ± 0.4	0.96	2.7
1024	0.012	0.008 ± 0.006	1.5	1.9 ± 0.6	0.97	2.2

Timestep count for DNG. With exact scores and at η=1, DNG is exact only as $T \to \infty$ (Appendix A.4). Table 4 measures how the sampler approaches that limit. Only the timestep count T varies; the method is held at the main-text configuration $(\lambda_0, \tau) = (1, 1)$ at M=1024 over 15 trials, and the Oracle is recomputed at each T as a matched floor. The leak falls monotonically from 0.053 to 0.020, and most of the gain is already reached by T=500. The gap to the matched Oracle narrows from 0.041 to 0.009 without closing, so a twentyfold serial budget still leaves the leak at about 1.8 times the Oracle value.

Particle count. Table 5 varies the particle count M with true scores over 15 trials. The Oracle is recomputed at each M, since the empirical W_2 of M particles carries a finite-sample floor that grows as M shrinks. The leak is nearly M-independent at the matched Oracle floor (the mean at M=256 is inflated by a few trials), and W_2 decreases with M as the floor does. The ESS/M stays at 0.96–0.97 with two to three resampling events at every M, so the degeneracy signature of a poor SMC path, an ESS collapse or constant resampling, is not observed. Against the closed-form posterior, the AUC of f_t near the terminal steps is 0.997, 0.989, and 0.980 for M=64, 256, and 1024, while the cosine similarity between $\nabla \log(1 - f_t)$ and $\nabla \log(1 - p(c \mid x))$ stays around 0.17–0.18 at every M: the estimate identifies the region to avoid accurately at every M without being pointwise accurate.

Comparison with Ours. Among the guidance methods in Table 1, only ours matches the Oracle leak (0.012 vs. 0.012) without inflating W_2 (1.8±0.6 vs. 1.5±0.3). This is in line with the asymptotic-exactness result in Section 4. The residual gap in W_2 relative to the Oracle is attributable to a slight inter-mode imbalance induced by resampling, visible in Figure 1 as a non-uniform allocation of particles across the nine $\neg c$-modes.

D.2 CIFAR-10 with Reference-Based Setting

This appendix gives the full specification of the CIFAR-10 experiment in Section 5.2. The complete code is available in our supplementary material.

D.2.1 Setup

Pretrained score model. We use the unconditional CIFAR-10 checkpoint from [40], distributed under the MIT license at `https://github.com/openai/improved-diffusion`. Following the instructions in that repository, the weights were obtained from `https://openaipublic.blob.core.windows.net/diffusion/march-2021/cifar10_uncond_50M_500K.pt`. This URL was confirmed accessible on 2026-05-05. Sampling uses DDPM with the respaced schedule of $T = 250$ steps, and images are normalized to $[-1, 1]$.

Avoidance conditions and references. The single-class conditions for the avoidance experiment are $c \in \{\{\text{plane}\}, \{\text{cat}\}, \{\text{frog}\}, \{\text{ship}\}\}$, and the multi-class condition *Vehicles* is $c = \{\text{plane}, \text{car}, \text{ship}, \text{truck}\}$. This gives $N = 200$ references for single-class conditions and $N = 800$ for Vehicles. The reference-selection seed is offset from the diffusion-noise seed so that the two are statistically independent.

Reference-based guidance: Safe Denoiser. We use the Safe Denoiser of Kim et al. [29] (Section 2) both as a baseline and as the proposal kernel in our SMC sampler. Its bandwidth is set once per run by the median pairwise pixel distance among the references. Following the original implementation, guidance is applied for $t > 0.08\,T = 20$ and disabled for the final steps. We sweep the repulsion strength $\eta \in \{0.0, 0.2, 0.3, 0.5, 0.7, 1.0\}$.

Online PU classifier on UNet bottleneck features. The classifier f_t is a linear logistic head on the 256-dim mean-pooled output of the UNet bottleneck block (shape $4\times4\times256$). Reference features are encoded once at $t = 0$ on the clean references. For the single-class conditions, particle features are read from the same forward pass that produces $\mu_\theta(x_t, t)$, giving $f_t(x_t) = h(\Phi_t(x_t))$ with a time-conditioned encoder Φ_t. For Vehicles, an additional UNet pass at $t = 0$ on the clipped Tweedie estimate is applied, giving $f_t(x_t) = h(\Phi_0(\hat{x}_0(x_t)))$. The head is fit by L-BFGS with warm-starting from the previous step, using the BCE surrogate of Section 5.2.

SMC parameters. We use a precomputed tempering schedule: λ_t ramps linearly from 0 to a terminal λ over the last $0.6\,T = 150$ steps. ESS-based sorted systematic resampling fires below $0.9M$, and is enabled only for $t > 15$ in single-class runs and $t > 50$ in Vehicles.

Implementation choices that depart from the abstract algorithm. We set the proposal-ratio factor r_t in (6) to 1 in the recursion of $w_{q,t}$. At $d = 3072$, this ratio has high variance, and a 128-particle filter cannot maintain a healthy ESS when it is included. The unconditional weight $w_{\emptyset,t}$ keeps the full ratio r_t at every step as in (9), and its update at resampling is the standard one of Section 3.1. This treatment, together with the BCE surrogate, deviates from the analysis in Section 4, so the procedure is a practical approximation not covered by Theorem 1; the experiments then test whether the existing negative-guidance method can be corrected by a lightweight auxiliary classifier and resampling. The classifier is trained and used only on the steps where $\lambda_t > 0$; earlier steps run the Safe Denoiser alone. The L-BFGS head uses up to 100 iterations on the first step and up to 20 iterations on later steps, so the online training cost is small.

Metrics. The generated samples carry no class labels by construction, so we evaluate them with an independently trained ResNet-18 classifier [23]. This classifier is not used during sampling and is fully decoupled from the diffusion pipeline. We start from the torchvision ResNet-18 with the `IMAGENET1K_V1` pretrained weights[2] and apply two modifications for low-resolution inputs: the first convolutional layer is replaced by a 3×3, stride-1 convolution, and the max-pooling layer is replaced by an identity mapping. The final fully connected layer is replaced by a 10-class linear head. We fine-tune all parameters for 50 epochs on the CIFAR-10 training split using SGD (learning rate 0.01, momentum 0.9, weight decay 5×10^{-4}) with cosine annealing, a batch size of 256, and data augmentation (random horizontal flipping and random 32×32 crops with four pixels of padding). Train accuracy exceeds 99%, and the checkpoint with the highest test accuracy reaches $\sim$95% Top-1 on the CIFAR-10 test split, which we consider sufficient for classifying our generated samples. The training code is included in the supplementary material.

- *Target-class rate.* The fraction of generated samples classified into c by the ResNet-18. For multi-class c, the rate counts any class in c.
- *FID.* Computed via `pytorch-fid` 0.3.0 [44] between the generated samples and the CIFAR-10 *training* images that lie in $\neg c$, so the floor is the Fréchet distance between samples from $p(\cdot \mid \neg c)$ and the $\neg c$-only training set.
- *KL.* $\mathrm{KL}(u_{\neg c} \,\|\, \hat{p}_{10})$, where $u_{\neg c}$ is the uniform distribution over non-avoided classes (mass $1/(10 - |c|)$ on each $i \notin c$, 0 on c), and $\hat{p}_{10}$ is the empirical 10-class histogram of the ResNet-18 predictions clipped at 10^{-10}.

[2] `https://docs.pytorch.org/vision/stable/models/generated/torchvision.models.resnet18.html`

- *Pairwise LPIPS.* Average pairwise LPIPS [58] with the AlexNet backbone, computed on 256 random samples per configuration.

The seed pool is shared across configurations so that paired comparisons (Ours vs. the Safe Denoiser at the same η) are well-defined.

D.2.2 Results

Choice of the PU L2 regularizer. We study the sensitivity of our method to the L2 regularization strength λ_{reg} by comparing two values spanning three orders of magnitude (10^{-3} and 1.0). Table 6 reports a per-class ablation at $\lambda = 1.5$, with our method run at $\eta = 0.3$. For reference, we also list the Safe Denoiser at $\eta = 0.3$ and $\eta = 0.5$, the two operating points that bracket the knee of its rate–FID curve. The choice of λ_{reg} shifts the metrics by a small amount, with within-class variation of at most 0.9 percentage points on rate and 1.4 FID points. At either value, our method already reaches a lower target-class rate than the Safe Denoiser at $\eta = 0.3$, while paying only 2–9 FID points above it. The Safe Denoiser only matches our rate by raising η to 0.5, at which point its FID degrades by 24–52 points beyond $\eta = 0.3$. In other words, both choices of λ_{reg} yield Pareto improvements over the guidance-only baseline. Based on this, we fix $\lambda_{\text{reg}} = 10^{-3}$ for plane and cat, and $\lambda_{\text{reg}} = 1.0$ for frog, ship, and Vehicles.

Per-class sweeps. Figure 5 reports the rate-vs-quality trade-off for the cat, frog, and ship single-class conditions, using the defaults above. The qualitative picture is consistent with the plane and Vehicles results in Figure 2 in Section 5.2: our λ-sweep Pareto-dominates the Safe Denoiser at the same η on FID and KL in most configurations, and remains comparable on pairwise LPIPS. The shape of the Safe Denoiser curve clarifies why our method helps: in the small-η regime, increasing η improves the avoidance rate at modest cost to FID and KL, but beyond a knee around $\eta = 0.5$ the quality metrics degrade sharply for any further rate reduction. Our method takes a Safe Denoiser proposal in this small-η regime and adds resampling driven by the PU classifier, which extracts further rate reduction along the part of the curve where the quality cost is still mild.

ESS diagnostics. Table 7 reports the ESS of $w_{q,t}$, the number of distinct surviving ancestors, and the number of resampling events on the main configurations (η=0.3, M=128, T=250; mean±std over 80 runs each). The time-averaged ESS stays at 110–127 of 128, and resampling stays sparse (0–7 events over the 250 steps) under the conservative threshold $0.9M$, a stability that the r_t=1 substitution

Table 6: PU regularization ablation at $\lambda = 1.5$, 80 seeds per cell. Our method is run at $\eta = 0.3$, while the Safe Denoiser baseline is shown at $\eta \in \{0.3, 0.5\}$, the two values that bracket the knee of its rate–FID curve in Figure 2 and Figure 5.

Class	Method	η	Rate (%) ↓	FID ↓	KL ↓	LPIPS ↑
Plane	Safe Denoiser	0.3	4.78	20.75	0.127	0.200
	Safe Denoiser	0.5	2.95	45.96	0.279	0.175
	Ours, $\lambda_{\text{reg}} = 10^{-3}$	0.3	3.50	24.65	0.148	0.197
	Ours, $\lambda_{\text{reg}} = 1.0$	0.3	4.39	23.66	0.136	0.198
Cat	Safe Denoiser	0.3	7.16	20.63	0.150	0.207
	Safe Denoiser	0.5	6.10	44.53	0.282	0.183
	Ours, $\lambda_{\text{reg}} = 10^{-3}$	0.3	5.66	24.95	0.163	0.203
	Ours, $\lambda_{\text{reg}} = 1.0$	0.3	6.17	25.38	0.200	0.205
Frog	Safe Denoiser	0.3	2.55	41.53	0.181	0.183
	Safe Denoiser	0.5	0.87	93.59	0.503	0.162
	Ours, $\lambda_{\text{reg}} = 1.0$	0.3	1.96	49.25	0.249	0.183
	Ours, $\lambda_{\text{reg}} = 10^{-3}$	0.3	1.90	50.31	0.245	0.185
Ship	Safe Denoiser	0.3	3.63	23.68	0.130	0.194
	Safe Denoiser	0.5	1.76	55.77	0.296	0.169
	Ours, $\lambda_{\text{reg}} = 1.0$	0.3	2.52	27.44	0.133	0.196
	Ours, $\lambda_{\text{reg}} = 10^{-3}$	0.3	3.22	26.08	0.138	0.203

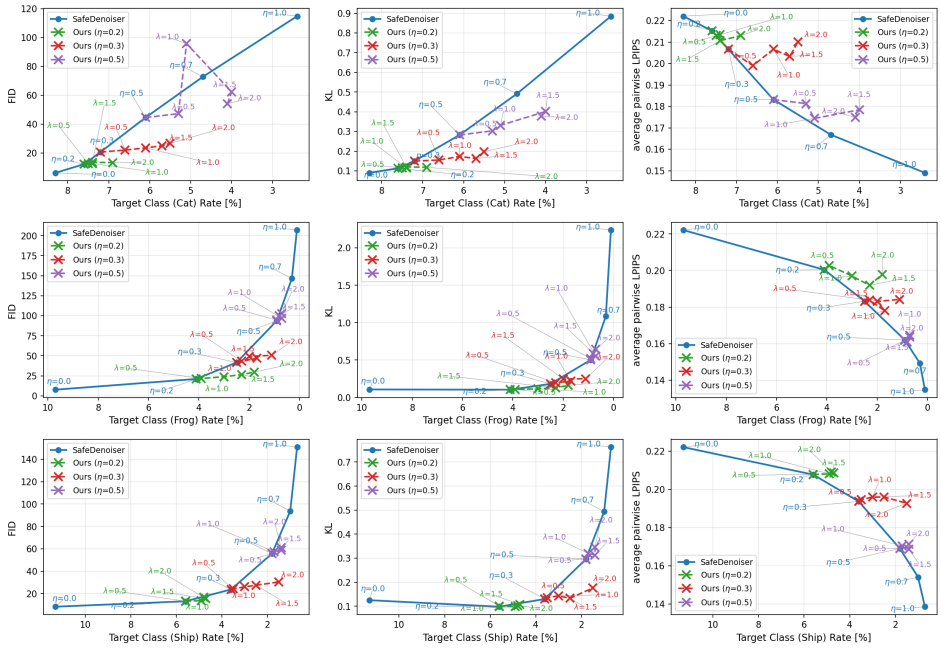


Figure 5: Avoidance rate vs. generation quality on CIFAR-10 for the cat (top), frog (middle), and ship (bottom) single-class conditions. Columns plot target-class rate against FID, KL, and average pairwise LPIPS (left to right); lower is better except for LPIPS.

Table 7: ESS diagnostics on CIFAR-10 at η=0.3, M=128, T=250, mean±std over 80 runs. Ancestors counts the distinct ancestors among the final 128 particles.

Class	λ	ESS (time avg.)	ESS (min)	Ancestors/128	Resamples
Plane	0.5	126.5 ± 0.6	59.6 ± 22.7	105.4 ± 19.0	1.3 ± 1.4
Plane	1.0	125.9 ± 0.8	43.9 ± 25.5	85.2 ± 27.0	2.9 ± 2.2
Plane	2.0	125.1 ± 0.9	37.5 ± 24.1	66.6 ± 28.2	4.7 ± 3.0
Vehicles	0.5	121.9 ± 1.8	72.5 ± 18.0	121.1 ± 9.2	0.4 ± 0.5
Vehicles	1.0	116.4 ± 2.8	33.2 ± 18.6	94.0 ± 14.1	2.3 ± 1.3
Vehicles	2.0	110.4 ± 3.1	10.1 ± 7.1	56.2 ± 18.4	6.4 ± 2.3

provides. Stronger avoidance λ or the harder Vehicles condition deepens the ESS minimum, yet even the hardest setting keeps 56 of 128 distinct ancestors, so the particles do not collapse onto a few ancestors. The ESS minima occur in the final steps, where the tilt is largest and resampling is intentionally disabled to avoid duplicating nearly final particles.

Batch-level spread. Table 8 reports the target-class rate per 128-sample batch as mean±std over the 80 seeds at η=0.3. The means equal the rates plotted in Figure 2, and the standard error of each mean is std/$\sqrt{80}$. The larger std under strong avoidance (Vehicles, λ=2.0) is consistent with the low ESS minimum in Table 7. The evaluator (test accuracy about 95%, training accuracy above 99%) follows the protocol of prior class-avoidance work [51]; all methods are scored by the same evaluator, which is independent of every generation pipeline.

Table 8: Target-class rate [%] per 128-sample batch on CIFAR-10 at η=0.3, mean±std over 80 seeds.

Method	Plane	Vehicles
Safe Denoiser	4.78 ± 1.99	27.60 ± 3.98
Ours (λ=0.5)	3.88 ± 2.31	25.86 ± 5.27
Ours (λ=1.0)	3.28 ± 2.66	19.87 ± 6.09
Ours (λ=2.0)	2.54 ± 3.01	20.93 ± 18.06

D.3 Stable Diffusion with Score-Based Setting

This subsection instantiates the score-based setting of Section 3.3 on Stable Diffusion v1.4 [42][3], a latent diffusion model that generates 512×512 images from text prompts and whose state is a 64×64 latent grid with four channels, so $d = 16{,}384$.

Task and prompts. The content prompt is `"a painting of a cafe terrace at night under a starry sky"`, which never names the avoided style. Following the negative-prompting format [25, 3], the style prompt appends `"in the style of Van Gogh"` to the content prompt, and supplies the conditional predictor $\hat{\epsilon}_\theta(\cdot, \cdot \mid c)$ required by Section 3.3. The evaluation pools and the detector calibration set use the longer suffix `"in the style of Vincent van Gogh"`.

Scores and sampling. The base model $p_{\theta,t}$ is the classifier-free-guidance combination of the content-prompt and unconditional predictions at scale 7.5, which is also the sampler that generates the content anchor. Sampling uses $T{=}50$ steps in half precision. SLD and DNG modify this same prediction, so their density without avoidance is also the content anchor; NegCFG is instead anchored on the style prediction (below). SLD [43] at its XXMAX setting below serves as the SMC proposal μ_t^{prop} for PU Steering throughout, as the Safe Denoiser does in Section D.2.

Online PU classifier. Each particle is represented by average-pooling the four latent channels of its Tweedie estimate to a 16×16 grid, giving a 1024-dimensional feature. Positives are the style-conditional Tweedie estimates and unlabeled samples are the content Tweedie estimates of the same particles, a practical variant of the construction of Section 3.3 with uniformly weighted positives and $w_{q,t}$-weighted unlabeled samples. Both reuse noise predictions that the SLD proposal already computes at every timestep, so the features cost no additional UNet evaluations. The classifier is a kernel logistic model with an RBF bandwidth set by the median heuristic scaled by 0.4, fit to the uPU risk (7) with an RKHS-norm penalty of 10^{-2} by L-BFGS at class prior $\pi{=}0.15$, refit at every timestep over the last 70% of the trajectory. Its logit is standardized across particles before the sigmoid and the resulting probability is clamped to $[10^{-3}, 1 - 10^{-3}]$. The two estimates start from the same particle, so they share the prompted content and differ mainly in the avoided style, and this contrast keeps the uPU fit stable: unlike the CIFAR-10 experiment of Section 5.2, where the small N and M forced the BCE substitution, this instantiation uses the uPU objective as written.

SMC control. As in Section D.2, we set the proposal-ratio factor r_t in (6) to 1. The tempering increment is chosen by bisection to hold the effective sample size at $0.5M$, capped at $1.5\lambda/T$ per timestep and at the terminal λ. Systematic resampling fires when the effective sample size falls below $0.7M$, only within the middle window between 30% and 70% of the reverse steps, at a cadence of five steps (four for the $M{=}128$, $\lambda{=}20$ configuration), and never at the final timestep, so each run returns weighted particles that can be read either unweighted as plain images or weighted as the SMC estimate. Log-weights are clipped to $[-2, 2]$ around their mean.

Baselines. SLD [43] runs at three settings on the same guidance scale. MAX is its strongest published preset HYP-MAX (no warm-up, safety scale 5000, threshold 1, momentum scale 0.5, momentum decay 0.7), which leaves the style rate at 0.334 against the floor value 0.238. We therefore add two settings beyond the published presets: XXMAX raises the safety scale to 20000 and the momentum scale to 0.7, and X3MAX further raises the momentum scale to 1 and the momentum decay to 0.8. For NegCFG, preliminary sweeps on the guidance combination did not lower the style rate, the content guidance at scale 7.5 dominating the style term, so it is run in the negative-prompting form $\hat{\epsilon}_{\mathrm{style}} + s\,(\hat{\epsilon}_{\mathrm{content}} - \hat{\epsilon}_{\mathrm{style}})$ with $s = \eta + 1$ and $\eta \in \{0.5, 1, 1.5, 2, 3, 4, 5, 6.5\}$. DNG [34] runs the guided prediction with the recursion (4) in the form of its released implementation, at temperature $\tau{=}0.1$, offset $\delta{=}0.02$, squared distances summed over the $16{,}384$ latent dimensions, an initial posterior estimate of 0.05, and the estimate clamped to $[10^{-6}, 0.99]$, swept over $\eta \in \{100, 150, 180, 200, 220, 240, 260, 280, 300\}$.

Detector, reference, and floor. The style detector is a logistic head on the CSD style embedding [47][4], a 768-dimensional descriptor of artistic style, fit by L-BFGS with an ℓ_2 penalty of 10^{-2} on

[3] `https://huggingface.co/CompVis/stable-diffusion-v1-4` (CreativeML OpenRAIL-M license).

[4] `https://huggingface.co/tomg-group-umd/CSD-ViT-L` (CC BY 4.0 license).



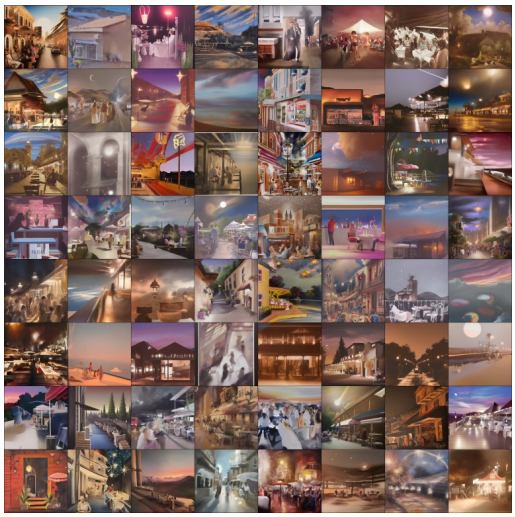
(a) Avoided style



(b) PU Steering



(c) SLD



(d) NegCFG

Figure 6: Randomly selected 8×8 image grids, from Table 9: PU Steering at M=128, λ=16 read as the equal-weight output after resampling, SLD at XXMAX, and NegCFG at η=3, its lowest-FID strength.

a calibration set whose positives are Van Gogh renderings of the content and whose negatives are Rembrandt renderings and plain content, all from seed blocks disjoint from every evaluation pool. The reference set keeps the 2048 images least like the avoided style, the bottom quarter by detector score of an 8192-image content pool. The oracle floor regenerates an independent 8192-image pool with disjoint seeds and applies the same filter. FID uses the Inception-2048 features of `pytorch-fid` [44], KID [6] is the unbiased cubic-polynomial-kernel estimate averaged over 100 random subsets of 1024, and Style-FD is the Fréchet distance in the CSD embedding space. The reported rate is the mean of the detector probability over the evaluated images, not a thresholded fraction. For the weighted rows, all three distances instead use closed-form importance-weighted moments over the full set, without subsampling. Because the oracle floor is itself a filtered content pool, its rate 0.238 is the residual style rate of that distribution; it is a reference point rather than a lower bound.

Qualitative comparison. Figure 6 shows randomly selected images from each sampler. Grid (a) is the avoided style, generated with the style prompt: most images follow the composition of *Café Terrace at Night*, with the yellow awning and the swirling starry sky. PU Steering in grid (b) keeps the prompted scene while spreading over a variety of painting styles, a minority of its images still

resembling the avoided style. SLD in grid (c) stays close to the avoided style in a visibly larger fraction of images, and NegCFG in grid (d) loses the saturated palette and shows washed-out colors and distortions, removing the style at the cost of the scene.

Results. Table 9 reports every configuration and Figure 7 plots the resulting frontier. PU Steering at $M{=}128$, $\lambda{=}16$ reduces the rate of the equal-weight output to 0.235, essentially the floor value, while attaining the lowest FID (28.10 against the floor's 18.52) and KID (0.0054 against 0.0002) among the equal-weight outputs. NegCFG reaches rates down to 0.222, but its FID never falls below 82.75, worse than the content anchor at 60.98, and its Style-FD stays at or above 0.897: it removes the style by destroying the content. DNG traces a different curve, its FID falling to a minimum of 44.88 at $\eta{=}200$ and rising again to 93.17 at $\eta{=}300$ while the rate never falls below 0.277. SLD leaves the rate at 0.334 at its strongest published preset, and the amplified settings reach 0.296 with FID between 41.82 and 43.70. Raising the particle count from $M{=}64$ to $M{=}128$ at $\lambda{=}16$ lowers the rate from 0.276 to 0.235, the FID from 33.46 to 28.10 and the KID from 0.0105 to 0.0054, while Style-FD rises slightly from 0.086 to 0.095. The mean terminal ESS stays between $0.47M$ and $0.51M$ across configurations. The last block disables resampling and switches the tempering to a linear ramp to keep the run stable; the weighted estimate then stays accurate while the plain images degrade toward the proposal, which is why the main protocol reads the resampled particles unweighted.

Sample-size stability. The 2048 images per configuration are far fewer than the $30{,}000$ on which FID is usually evaluated for text-to-image models [42]. The independent repetition of the $M{=}64$, $\lambda{=}16$ configuration in Table 9 gives a rate of 0.280 against 0.276 and an FID of 33.41 against 33.46, with the remaining metrics similarly close, so the reported values are stable across seeds at this sample size.

E Broader Impacts

This paper presents work whose goal is to advance the statistical foundations of negative guidance for diffusion models. The same mechanism can support safety-related applications of pretrained generative models. It can also in principle be repurposed to amplify rather than suppress an undesired condition. Such repurposing does not exceed what is already afforded by the underlying pretrained model and existing guidance methods, and our work does not increase the generative capability of the base model. We release no new pretrained model or dataset, and use only publicly released assets.

Table 9: Stable Diffusion v1.4 avoiding Van Gogh style, 2048 images per configuration, lower is better throughout. Gray rows are anchors rather than competing methods. For PU Steering, "u." evaluates the plain images and "w." the SMC-weighted estimate, ESS$/M$ is the mean per-run terminal effective sample size, and λ is the terminal tempering strength. Bold marks the best FID, KID, and Style-FD among the methods outside the no-resampling ablation. Representative baseline strengths are listed; Figure 7 plots every configuration of the sweeps.

Method	Config	Rate ↓	FID ↓	KID ↓	Style-FD ↓	ESS$/M$
Oracle floor	2048 vs. 2048	0.238	18.52	0.0002	0.009	—
Content (no avoidance)		0.409	60.98	0.0552	0.213	—
Prompted Van Gogh		0.645	164.32	0.1982	0.661	—
SLD	MAX	0.334	43.08	0.0294	0.114	—
SLD	XXMAX	0.317	41.82	0.0263	0.118	—
SLD	X3MAX	0.296	43.70	0.0284	0.180	—
NegCFG	η=0.5	0.279	112.58	0.1074	0.939	—
NegCFG	η=2	0.222	85.56	0.0678	0.900	—
NegCFG	η=3	0.224	82.75	0.0608	0.897	—
NegCFG	η=6.5	0.280	97.35	0.0775	0.992	—
DNG	η=100	0.382	56.87	0.0500	0.192	—
DNG	η=200	0.303	44.88	0.0273	0.145	—
DNG	η=260	0.277	63.84	0.0411	0.243	—
DNG	η=300	0.289	93.17	0.0737	0.366	—
PU Steering (u.)	M=128, λ=16	0.235	**28.10**	0.0054	0.095	—
(w.)		0.228	38.68	**0.0047**	0.110	0.47
PU Steering (u.)	M=128, λ=20	0.243	31.07	0.0056	0.085	—
(w.)		0.245	43.25	0.0072	0.092	0.51
PU Steering (u.)	M=128, λ=24	0.253	29.54	0.0058	**0.082**	—
(w.)		0.239	39.95	0.0060	0.111	0.50
PU Steering (u.)	M=64, λ=16	0.276	33.46	0.0105	0.086	—
(w.)		0.283	43.27	0.0115	0.096	0.48
PU Steering (u.)	M=64, λ=16 (repeat)	0.280	33.41	0.0106	0.090	—
(w.)		0.281	43.42	0.0116	0.094	0.49
PU Steering, no rs. (u.)	M=64, λ=8	0.314	40.41	0.0243	0.114	—
(w.)		0.272	39.23	0.0104	0.103	0.50
PU Steering, no rs. (u.)	M=64, λ=12	0.319	41.76	0.0256	0.122	—
(w.)		0.268	38.08	0.0090	0.106	0.51

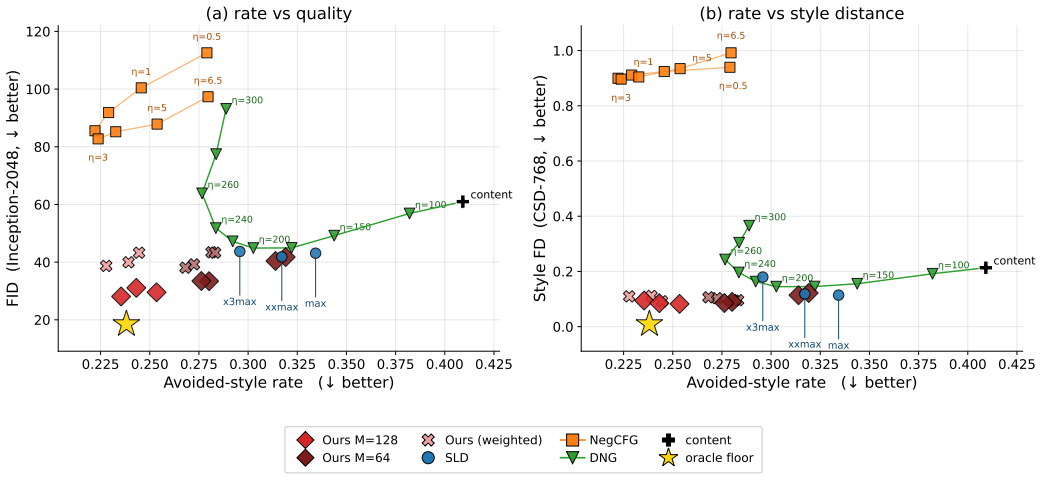


Figure 7: Stable Diffusion, avoided-style rate against FID (left) and Style-FD (right); lower left is better in both panels. Every configuration of Table 9 is plotted, including the two no-resampling points, which are the rightmost M=64 markers.