UNCERTAINTY-AWARE SEQUENTIAL MONTE CARLO FOR HARD-CONSTRAINED DIFFUSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Deploying diffusion models often requires strict constraints, yet existing methods have distinct drawbacks: projection-based approaches rely on restrictive assumptions and can distort the constrained distribution, while particle methods construct intermediate potentials from point predictions of the terminal state, ignoring posterior uncertainty. Moreover, many practical constraints are non-differentiable or black-box, so gradients are unavailable. We introduce Uncertainty-Aware Sequential Monte Carlo (UA-SMC), a projection-free sampler for hard-constrained diffusion. UA-SMC directly approximates the probability of eventual constraint satisfaction and uses it as a Feynman–Kac potential. A practical time-dependent schedule captures contracting terminal uncertainty while avoiding state-dependent gradients. We analyze how approximation error in these potentials propagates to the intermediate targets and derive a finite-particle error bound. Across porous material design, rare-event sampling for nonlinear dynamics, and copyright-safe text-to-image generation, UA-SMC consistently attains high constraint satisfaction and distributional fidelity for non-differentiable and black-box constraints.

1 INTRODUCTION

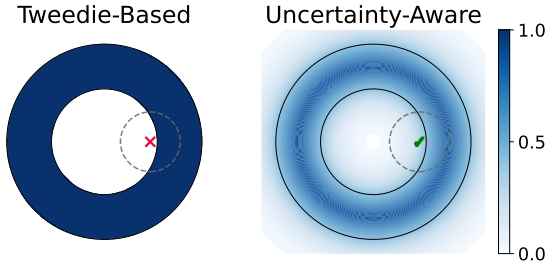


Figure 1: Contrasting Tweedie-based and uncertainty-aware resampling on non-convex annulus set.

Diffusion models (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song et al., 2020a;b) have enabled generative modeling across numerous domains spanning image synthesis (Rombach et al., 2022; Podell et al., 2024), video generation (Ho et al., 2022b;a), material and molecular design (Xie et al., 2021; Fu et al., 2024; Corso et al., 2022) and path planning (Janner et al., 2022). Yet real-world deployment often requires *constrained generation* such as content guidelines in text-to-image generation (Na et al., 2026), physical constraints in material generation (Xu et al., 2025), and collision avoidance in motion planning (Shaoul et al., 2025). These tasks can be formalized as sampling from a *hard-constrained* target distribution $\tilde{p}(\boldsymbol{x}) \propto p(\boldsymbol{x})\mathbf{1}_C(\boldsymbol{x})$, where $p(\boldsymbol{x})$ is a data distribution and $\mathbf{1}_C(\boldsymbol{x})$ is the indicator of the feasible set C. We call this *hard-constrained diffusion* when $p(\boldsymbol{x})$ is approximated with a diffusion model.

One approach to hard-constrained diffusion is projection-based, including Projected Diffusion Models (PDM) (Christopher et al., 2024) and their latent-diffusion analogue (Zampini et al., 2026), which we denote as Latent PDM (LPDM). These methods frame hard-constrained generation as constrained optimization, modifying the reverse process to iteratively project intermediate samples onto the feasible set C. Despite empirical effectiveness, they have key practical and theoretical limitations. Existing theoretical guarantees require C to be convex (Parikh & Boyd, 2014; Bubeck et al., 2015). In practice, repeated projections can accumulate samples on the boundary of C as illustrated in Figure 5.

Particle-based methods, notably Sequential Monte Carlo (SMC)-based methods (Naesseth et al., 2019), offer a projection-free alternative. Popular SMC approaches (Singhal et al., 2025; Uehara

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

et al., 2025) target reward-tilted distributions by propagating particles $\{\boldsymbol{x}_t^i\}_{i=1}^N$ through the reverse process, periodically reweighting and resampling them via intermediate potentials. For hard constraints, the exact intermediate potential is the probability of eventual feasibility given the current state $\Pr(\mathbf{X}_0 \in C \mid \mathbf{X}_t = \boldsymbol{x}_t)$. Since this exact potential is intractable, a common approach (Kim et al., 2025; Wang et al., 2026) is to instead evaluate feasibility on a point prediction of the terminal state, typically Tweedie's estimate $\hat{\boldsymbol{x}}_0(\boldsymbol{x}_t) = \mathbb{E}[\mathbf{X}_0 \mid \mathbf{X}_t = \boldsymbol{x}_t]$ (Efron, 2011). This approach yields plug-in potentials of the form $\mathbf{1}_C(\hat{\boldsymbol{x}}_0(\boldsymbol{x}_t))$, which discard uncertainty in the random terminal state. As shown in Figure 1, particles whose point estimate falls outside C can be eliminated during resampling despite retaining substantial probability of terminating feasibly.

1.1 Our Contributions

To address the limitations of projection-based methods and uncertainty-blindness of plug-in SMC potentials, we propose **Uncertainty-Aware SMC** (UA-SMC), a projection-free sampler for hard-constrained diffusion. We show that, for a broad class of inequality-based constraint sets, the exact intermediate potential is given by a conditional cumulative distribution function (CDF). UA-SMC approximates this potential with an uncertainty-calibrated sigmoid whose selectivity increases as terminal uncertainty contracts. Crucially, UA-SMC only needs to use this approximation as a Feynman-Kac potential for particle reweighting, only requiring evaluations of the constraint function V and enabling hard-constrained diffusion for non-differentiable and black-box constraints.

- We introduce UA-SMC, a projection-free hard-constrained sampler that replaces the plug-in potentials $\mathbf{1}_C(\hat{\boldsymbol{x}}_0(\boldsymbol{x}_t))$ with uncertainty-aware potentials that estimate the probability of eventual satisfaction. The practical sampler requires only evaluations of the constraint function V enabling hard-constrained generation for non-differentiable and black-box constraints. See details in Section 3.
- We quantify the error of approximating the potentials with an uncertainty-calibrated sigmoid, bound the induced total variation gap between exact and approximate intermediate targets, and derive an end-to-end finite-particle error bound for UA-SMC. See details in Section 4.
- With matched compute, UA-SMC is the only method to achieve near-perfect satisfaction with high distributional fidelity across porous material generation, rare-event sampling for nonlinear dynamics, and copyright-safe text-to-image generation, outperforming existing prominent baselines including PDM (Christopher et al., 2024; Zampini et al., 2026), FK Steering (Singhal et al., 2025), TDS (Wu et al., 2023) and Best-of-N (Karthik et al., 2023). See Section 5 for details.

1.2 Additional Related Works

Beyond the approaches discussed above, gradient-guided methods modify the reverse diffusion dynamics using differentiable likelihoods or learned terminal-feasibility surrogates to construct a guidance term (Chung et al., 2022; Guo et al., 2026). Closely related to our uncertainty motivation, Finzi et al. (2023) derive a Gaussian CDF approximation to a conditional event probability and differentiate it to construct a conditional score for guided reverse diffusion. In addition to classic SMC reweighting and resampling, TDS (Wu et al., 2023) and DAS (Kim et al., 2025) further differentiate their potentials to construct gradient-guided proposal kernels. All of these gradient-guided constructions require a differentiable surrogate of the constraint through which gradients can be propagated and do not directly apply when the constraint is available only through black-box evaluations.

UA-SMC applies to both differentiable and non-differentiable black-box constraint functions V. For constraint sets of the form $C = \{\boldsymbol{x} : V(\boldsymbol{x}) \geq \hat{V}\}$, we approximate the conditional event probability h_t and use it as a Feynman–Kac potential for particle reweighting. When V is differentiable, a Gaussian approximation motivates an oracle calibration that depends on ∇V, akin to Finzi et al. (2023). Our practical method instead replaces this oracle calibration with a time-dependent sharpness schedule that requires only evaluations of V, allowing UA-SMC to be applied when V is non-differentiable or available only as a black-box.

1.3 Organization

The remainder of the paper is organized as follows. Section 2 reviews diffusion models and SMC. Section 3 introduces UA-SMC, our uncertainty-aware approach to hard-constrained diffusion. Section 4 analyzes the uncertainty-calibrated potential approximation and establishes an end-to-end

finite-particle error bound. Section 5 benchmarks UA-SMC against projection and particle-based methods across diverse tasks.

2 BACKGROUND

This section introduces the diffusion model and SMC background needed for UA-SMC. Throughout, we use uppercase $\mathbf{X}_t$ for random variables and lowercase $\boldsymbol{x}_t$ for realizations; $\mathbf{X}_{a:b}$ and $\boldsymbol{x}_{a:b}$ denote random and realized trajectories from time a to b, respectively.

2.1 DIFFUSION MODELS

Diffusion models define complementary forward and reverse trajectories that iteratively corrupt and refine samples. Starting from $\mathbf{X}_0 \sim p_{\mathrm{data}}$, the forward process builds $\mathbf{X}_{0:T}$ with

$$q(\boldsymbol{x}_{1:T} \mid \boldsymbol{x}_0) = \prod_{t=1}^{T} q(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}), \;\; q(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}) = \mathcal{N}\Big(\boldsymbol{x}_t; \sqrt{1-\beta_t}\,\boldsymbol{x}_{t-1}, \beta_t \boldsymbol{I}\Big),$$

where $\{\beta_t\}_{t=1}^T$ is a pre-defined variance schedule. The forward process transforms p_{data} into approximately Gaussian noise $\mathbf{X}_T \sim p_T$.

The reverse process inverts this corruption. Starting from $\mathbf{X}_T \sim p_T = \mathcal{N}(\mathbf{0}, \boldsymbol{I})$, it defines $\mathbf{X}_{T:0}$ as

$$p_\theta(\boldsymbol{x}_{T:0}) = p_T(\boldsymbol{x}_T) \prod_{t=0}^{T-1} p_\theta(\boldsymbol{x}_t \mid \boldsymbol{x}_{t+1}), \;\; p_\theta(\boldsymbol{x}_t \mid \boldsymbol{x}_{t+1}) = \mathcal{N}(\boldsymbol{x}_t; \mu_\theta(\boldsymbol{x}_{t+1}, t+1), \Sigma_{t+1}(\boldsymbol{x}_{t+1})\boldsymbol{I}),$$

where μ_θ is parameterized by a neural network. This maps noise to data samples $\mathbf{X}_0 \sim p_\theta \approx p_{\mathrm{data}}$.

2.2 SEQUENTIAL MONTE CARLO

SMC approximates a sequence of distributions $\{\pi_t\}_{t=1}^T$ by propagating interacting particles, decomposing a hard sampling problem into collection of easier tasks. For trajectory $\mathbf{X}_{1:t}$, let $\pi_t(\boldsymbol{x}_{1:t}) = \gamma_t(\boldsymbol{x}_{1:t})/Z_t$, with unnormalized density γ_t and normalizing constant Z_t. Trajectories are proposed sequentially via

$$q_t(\boldsymbol{x}_{1:t}) = q_{t-1}(\boldsymbol{x}_{1:t-1})\, r_t(\boldsymbol{x}_t \mid \boldsymbol{x}_{1:t-1}),$$

where r_t extends a trajectory from $t-1$ to t.

SMC approximates π_t by the weighted empirical measure

$$\pi_t^N(\mathrm{d}\boldsymbol{x}_{1:t}) = \sum_{i=1}^{N} w_t^{(i)} \delta_{\boldsymbol{x}_{1:t}^{(i)}}(\mathrm{d}\boldsymbol{x}_{1:t}),$$

with recursive weight update

$$\widetilde{w}_t^{(i)} = w_{t-1}^{(i)} \frac{\gamma_t(\boldsymbol{x}_{1:t}^{(i)})}{\gamma_{t-1}(\boldsymbol{x}_{1:t-1}^{(i)})\, r_t(\boldsymbol{x}_t^{(i)} \mid \boldsymbol{x}_{1:t-1}^{(i)})}, \;\; w_t^{(i)} = \widetilde{w}_t^{(i)} \Big/ \sum_{j=1}^{N} \widetilde{w}_t^{(j)}.$$

Under standard regularity conditions, $\pi_t^N \to \pi_t$ as $N \to \infty$. The framework is flexible in the choice of proposal and target.

In FK Steering, the proposal is given by the reverse diffusion transition $p_\theta(\boldsymbol{x}_t \mid \boldsymbol{x}_{t+1})$, yielding target distribution

$$\gamma_t(\boldsymbol{x}_{T:t}) = p_\theta(\boldsymbol{x}_{T:t}) \prod_{s=t}^{T} G_s(\boldsymbol{x}_{T:s}), \;\; \text{with} \;\; \prod_{s=0}^{T} G_s(\boldsymbol{x}_{T:s}) = \exp(\lambda R(\boldsymbol{x}_0)),$$

so potentials G_s upweight trajectories predicted to yield high terminal reward R. In practice G_t is built from a point estimate, e.g., the *sum* potential $G_t(\boldsymbol{x}_{T:t}) = \exp\Big(\lambda \sum_{s=t}^{T} R(\hat{\boldsymbol{x}}_0(\boldsymbol{x}_s))\Big)$, where $\hat{\boldsymbol{x}}_0(\boldsymbol{x}_s) = \mathbb{E}[\mathbf{X}_0 \mid \mathbf{X}_s = \boldsymbol{x}_s]$.

3 UA-SMC: Our Proposed Approach

In this section, we present UA-SMC. We derive the exact intermediate potentials that bridge the unconstrained reverse process and the hard-constrained target, and show they correspond to a conditional CDF. We approximate this CDF with an uncertainty-calibrated sigmoid and integrate the resulting potentials into a practical SMC algorithm.

3.1 Integrating Hard-Constraints into SMC

We consider a broad class of constraints defined by a function-threshold pair $(V, \hat{V})$:

$$\boldsymbol{x} \in C \iff V(\boldsymbol{x}) \geq \hat{V}$$

Many sets, including membership oracles, multi-objective classifier scores and simulator-based constraints can be expressed in this form. See Appendix A for details.

Let $p_\theta(\boldsymbol{x}_{T:0})$ be the reverse diffusion distribution. The exact hard-constrained terminal target is

$$\pi_0(\boldsymbol{x}_{T:0}) \propto p_\theta(\boldsymbol{x}_{T:0}) \mathbf{1}_C(\boldsymbol{x}_0).$$

Marginalizing the future denoising states gives the intermediate target

$$\begin{aligned}
\pi_t(\boldsymbol{x}_{T:t}) &\propto \int p_\theta(\boldsymbol{x}_{T:0}) \mathbf{1}_C(\boldsymbol{x}_0) \, \mathrm{d}\boldsymbol{x}_{t-1:0} \\
&= p_\theta(\boldsymbol{x}_{T:t}) \int p_\theta(\boldsymbol{x}_{t-1:0} \mid \boldsymbol{x}_t) \mathbf{1}_C(\boldsymbol{x}_0) \, \mathrm{d}\boldsymbol{x}_{t-1:0} \\
&= p_\theta(\boldsymbol{x}_{T:t}) \Pr_\theta(\mathbf{X}_0 \in C \mid \mathbf{X}_t = \boldsymbol{x}_t) \,. \qquad (1)
\end{aligned}$$

Recalling our function-threshold definition, we define

$$\pi_t(\boldsymbol{x}_{T:t}) \propto p_\theta(\boldsymbol{x}_{T:t}) h_t(\boldsymbol{x}_t), \;\; h_t(\boldsymbol{x}_t) := \Pr_\theta\Big(V(\mathbf{X}_0) \geq \hat{V} \mid \mathbf{X}_t = \boldsymbol{x}_t\Big) \,, \qquad (2)$$

where h_t is the probability of eventual feasibility conditioned on the current state. While interpretable, h_t is intractable and must be approximated.

3.2 Constructing Intermediate Target

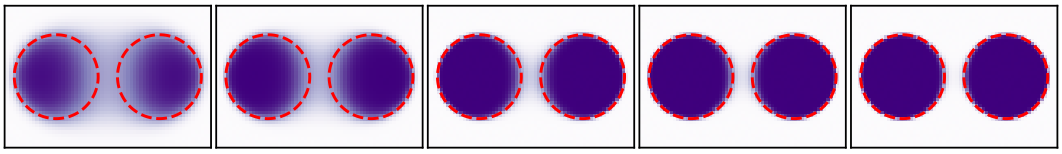


Figure 2: UA-SMC potentials for a non-convex disconnected-disks constraint under increasing sharpness. Potentials are soft when uncertain and sharp when confident.

While direct approximation of h_t in equation 2 can be performed with Monte Carlo, it is expensive and impractical. Another common approach is to replace $\mathbf{X}_0$ by Tweedie's point estimate $\hat{\boldsymbol{x}}_0(\boldsymbol{x}_t) = \mathbb{E}[\mathbf{X}_0 \mid \mathbf{X}_t = \boldsymbol{x}_t]$ (Robbins, 1964; Efron, 2011). In our $(V, \hat{V})$ setting this gives the plug-in potential

$$h_t^{\text{plug-in}}(\boldsymbol{x}_t) = \mathbf{1}_{\{\boldsymbol{x}: V(\boldsymbol{x}) \geq \hat{V}\}}(\hat{\boldsymbol{x}}_0(\boldsymbol{x}_t)).$$

However, a point estimate $\hat{\boldsymbol{x}}_0(\boldsymbol{x}_t)$ cannot reliably determine terminal feasibility because it discards uncertainty in $\mathbf{X}_0 \mid \mathbf{X}_t = \boldsymbol{x}_t$. To retain the efficiency of using $\hat{\boldsymbol{x}}_0(\boldsymbol{x}_t)$ while accounting for its uncertainty, we define the predicted margin and its error

$$r_t(\boldsymbol{x}_t) = V(\hat{\boldsymbol{x}}_0(\boldsymbol{x}_t)) - \hat{V}, \;\; \epsilon_t(\boldsymbol{x}_t) = V(\hat{\boldsymbol{x}}_0(\boldsymbol{x}_t)) - V(\mathbf{X}_0).$$

The exact potential is the conditional CDF of the error evaluated at the predicted margin:

$$h_t(\boldsymbol{x}_t) = \Pr(\epsilon_t \leq r_t \mid \mathbf{X}_t = \boldsymbol{x}_t). \qquad (3)$$

Utilizing common techniques similar to Finzi et al. (2023), we can approximate it with a Gaussian CDF. Approximating this Gaussian CDF by a sigmoid function gives a highly-accurate tractable potential

$$\tilde{h}_t^\star(\boldsymbol{x}_t) = \sigma(\alpha_t^\star(\boldsymbol{x}_t) r_t(\boldsymbol{x}_t)), \;\; \alpha_t^\star(\boldsymbol{x}_t) := c/\tau_t(\boldsymbol{x}_t), \; c > 0.$$

This has a direct interpretation. At early diffusion time, τ_t is large so $\tilde{h}_t^\star$ is soft and retains particles outside C; as denoising progresses, τ_t contracts so $\tilde{h}_t^\star$ sharpens and becomes selective when outcomes are more certain.

The oracle sharpness $\alpha_t^\star(\boldsymbol{x}_t)$ is state-dependent and requires estimating a high-dimensional conditional covariance at each step. Moreover, its construction depends on local derivative information about V, which may be unavailable when V is non-differentiable or accessible only through black-box function evaluations. We therefore amortize this state-dependent calibration with a time-dependent schedule $\{\alpha_t\}_{t=T-1}^0$ that proxies the typical contraction of uncertainty. As illustrated in Figure 2, the resulting potential

$$\tilde{h}_t(\boldsymbol{x}_t) = \sigma(\alpha_t r_t(\boldsymbol{x}_t))$$

becomes increasingly selective as sampling progresses. Crucially, this practical potential requires only evaluations of V, enabling UA-SMC to handle common constraint functions such as membership oracles, distance-based constraints, and other non-differentiable or black-box constraint functions.

In Section 4, we analyze the consequences of this design choice. Proposition 1 quantifies the error induced by the mismatch between α_t and $\alpha_t^\star$. Appendix B provides a practical procedure for selecting the time-dependent schedule and shows that, when well calibrated, it can recover much of the performance of the state-dependent oracle sharpness.

3.3 Algorithm

Building on the construction above, Algorithm 1 summarizes UA-SMC. At each reverse step it propagates particles via $p_\theta(\boldsymbol{x}_t \mid \boldsymbol{x}_{t+1})$, predicts terminal states via Tweedie's estimate, and scores feasibility with the uncertainty-aware potential presented in Section 3.2. All particle-wise operations are independent and can be parallelized. Resampling focuses computation on promising trajectories: it preserves diversity early when $\tilde{h}_t$ is soft and becomes increasingly selective as uncertainty contracts and $\tilde{h}_t$ sharpens. We use an adaptive Effective Sample Size (ESS) threshold to determine when we resample. UA-SMC avoids the convexity requirement of projection-based methods and the uncertainty blindness of plug-in SMC.

Algorithm 1 UA-SMC

Require: Particles N, steps T, model p_θ, schedule $\{\alpha_t\}_{t=T-1}^0$, function V, threshold $\hat{V}$
  Initialize $\{\boldsymbol{x}_T^n \sim \mathcal{N}(0, I), \; w_T^n = 1/N, \; h_T^n = 1\}_{n=1}^N$
  for $t = T-1, \ldots, 0$ **do**
    Sample $\tilde{\boldsymbol{x}}_t^n \sim p_\theta(\boldsymbol{x}_t \mid \boldsymbol{x}_{t+1}^n), \; n \in [N]$
    Predict $\hat{\boldsymbol{x}}_0^n = \hat{\boldsymbol{x}}_0(\tilde{\boldsymbol{x}}_t^n), \; n \in [N]$
    Score $\tilde{h}_t^n = \sigma\Big(\alpha_t(V(\hat{\boldsymbol{x}}_0^n) - \hat{V})\Big), \; n \in [N]$
    Reweight $\bar{w}_t^n = w_{t+1}^n \tilde{h}_t^n / h_{t+1}^n, \; w_t^n = \bar{w}_t^n / \sum_m \bar{w}_t^m$
    if $\mathrm{ESS}(w_t^{1:N}) < N/2$ **then**
      Resample $a_t^n \sim$ Categorical$(w_t^{1:N})$ and set $(\boldsymbol{x}_t^n, h_t^n, w_t^n) = (\tilde{\boldsymbol{x}}_t^{a_t^n}, \tilde{h}_t^{a_t^n}, 1/N)$
    else
      Set $(\boldsymbol{x}_t^n, h_t^n) = (\tilde{\boldsymbol{x}}_t^n, \tilde{h}_t^n)$
    end if
  end for
  return $\{(\boldsymbol{x}_0^n, w_0^n)\}_{n=1}^N$

4 Theoretical Results

In Section 3.2 we approximated the exact CDF potential with an uncertainty-calibrated sigmoid function. Here, we quantify the resulting potential-approximation error and show how it propagates

to the intermediate total-variation distance and the finite-particle accuracy of UA-SMC. Proofs are in Appendix C.

Proposition 1. *Under the posterior assumptions of Chung et al. (2022), the discrepancy between the exact and UA-SMC potentials satisfies*

$$|h_t(\boldsymbol{x}_t) - \tilde{h}_t(\boldsymbol{x}_t)| \leq \mathcal{E}_t^{\mathrm{Gauss}}(\boldsymbol{x}_t) + \mathcal{E}_t^{\mathrm{sig}}(\boldsymbol{x}_t).$$

Under the posterior structure common in diffusion literature,[1] Proposition 1 decomposes error into two controllable terms: $\mathcal{E}_t^{\mathrm{Gauss}}$ is the Gaussian approximation error, which decays as $t \to 0$, and $\mathcal{E}_t^{\mathrm{sig}}$ is the sigmoid approximation error to the Gaussian CDF, which is small under a well-calibrated α_t. Given this pointwise bound, we examine its impact on intermediate targets.

Theorem 1. *Let $\pi_t(\boldsymbol{x}_{T:t}) \propto p_\theta(\boldsymbol{x}_{T:t})h_t(\boldsymbol{x}_t)$ and $\tilde{\pi}_t(\boldsymbol{x}_{T:t}) \propto p_\theta(\boldsymbol{x}_{T:t})\tilde{h}_t(\boldsymbol{x}_t)$. Then*

$$\|\pi_t - \tilde{\pi}_t\|_{\mathrm{TV}} \leq \frac{\mathbb{E}_{p_\theta}[|h_t(\mathbf{X}_t) - \tilde{h}_t(\mathbf{X}_t)|]}{\mathbb{E}_{p_\theta}[h_t(\mathbf{X}_t)]}.$$

Theorem 1 shows potential error controls the total variation (TV) gap between exact and approximate intermediate targets. Combining it with standard Feynman-Kac regularity yields a finite-particle bound for UA-SMC.

Theorem 2. *Under the posterior assumptions of Chung et al. (2022) and standard regularity for Feynman-Kac particle systems (Crisan & Doucet, 2002), for each t there exists $C_t < \infty$ such that for any bounded measurable f,*

$$\mathbb{E}_N\left[|\pi_t(f) - \tilde{\pi}_t^N(f)|\right] \leq C_t \|f\|_\infty \left(\frac{1}{\sqrt{N}} + \frac{\mathbb{E}_{p_\theta}[\mathcal{E}_t^{\mathrm{Gauss}}(\mathbf{X}_t) + \mathcal{E}_t^{\mathrm{sig}}(\mathbf{X}_t)]}{\Pr_\theta(\mathbf{X}_0 \in C)}\right),$$

where $\tilde{\pi}_t^N$ is the N-particle UA-SMC empirical measure and $\mathbb{E}_N$ is over particle randomness.

Theorem 2 decomposes error into the standard $O(N^{-1/2})$ SMC error and the potential-approximation error. The latter is normalized by $\Pr_\theta(\mathbf{X}_0 \in C)$. This factor is not specific to UA-SMC; it is merely the normalization constant of π_t and reflects general difficulty of sampling rare events. We validate both terms empirically in the appendix. In particular, Appendix D.1.3 shows terminal TV grows with schedule miscalibration, consistent with the potential term; Appendix D.2.3 shows TV decays with N, consistent with the finite-particle term.

5 Numerical Experiments

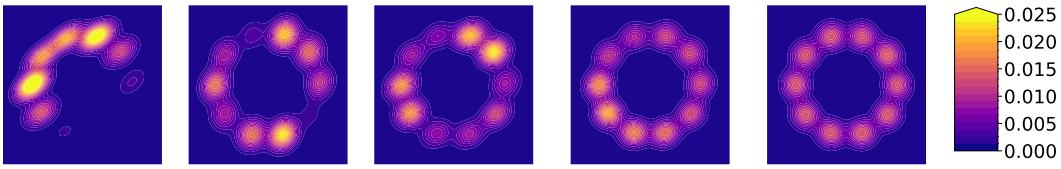
We benchmark UA-SMC against projection and particle-based methods under matched NFE budgets, evaluating (i) constraint satisfaction and (ii) distributional fidelity. We begin with an analytically tractable task, where the constrained target is available exactly and the constraint function is differentiable, allowing us to compare against gradient-based methods (Section 5.1). We then consider settings with black-box constraints: rare-event sampling for nonlinear dynamics (Section 5.2), porous material design (Section 5.3), and copyright-safe text-to-image generation (Section 5.4). Since gradients of the black-box constraint functions are unavailable, we additionally include Best-of-N (BoN) as a simple gradient-free particle baseline. Diffusion parameters, function-threshold pairs defining the constraint sets, and evaluation protocols are detailed in Appendix D.

5.1 An Analytical Example

We start with a low-dimensional synthetic task where the constrained target is analytically tractable and the constraint function is differentiable, allowing comparison with gradient-guided methods. We train a diffusion model on a ten-component Gaussian mixture

$$p(\boldsymbol{x}) = \frac{1}{10}\sum_{i=1}^{10} \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_i, \sigma^2 \boldsymbol{I}), \quad \boldsymbol{\mu}_i = 6[\cos(2\pi i/10),\ \sin(2\pi i/10)]^\top,$$

[1] Also used in (Boys et al., 2023; Song et al., 2023).



(a) PDM TV: 0.48 (b) TDS TV: 0.37 (c) FK Steering TV: 0.29 (d) UA-SMC (Ours) **TV: 0.17** (e) Analytical –

Figure 3: Sample-density heatmaps and total variation (TV) to the analytical constrained target for PDM, TDS, FK Steering, UA-SMC, and the ground truth.

with $\sigma^2 = 0.5$, and impose the non-convex annular constraint $C = \{6 \leq \|\boldsymbol{x}\| \leq 8\}$. We represent this constraint set with the differentiable function-threshold pair

$$V(\boldsymbol{x}) = (\|x\|_2^2 - 6^2)(8^2 - \|x\|_2^2) \quad \hat{V} = 0.$$

Unlike the black-box settings considered later, this function is differentiable which makes the state-dependent oracle sharpness $\alpha_t^\star(\boldsymbol{x}_t)$ computable and enables comparison with gradient-based methods.

Figure 3 compares sample-density heatmaps and TV to the exact constrained target. PDM and TDS exhibit severe mode imbalance, assigning little mass to several feasible modes, while FK Steering covers all modes but overweights a subset. UA-SMC most closely reproduces the analytical target, achieving TV of 0.17, a 41% relative reduction over FK Steering. Notably, this improvement occurs in a setting where the constraint is differentiable and TDS can exploit gradient-guided proposals, indicating that UA-SMC's gains exist in both differentiable and black-box settings. FK Steering uses a fixed temperature that applies the same selectivity regardless of terminal uncertainty, so particles that reach feasible modes early are preferentially replicated while particles still en route to other modes can be prematurely discarded. UA-SMC instead uses the state-dependent oracle sharpness $\alpha_t^\star(\boldsymbol{x}_t)$. By remaining less selective when terminal outcomes are uncertain, UA-SMC allows particles to migrate toward distinct feasible modes before resampling concentrates the population, thereby better preserving mode balance.

5.2 Trajectory Sampling

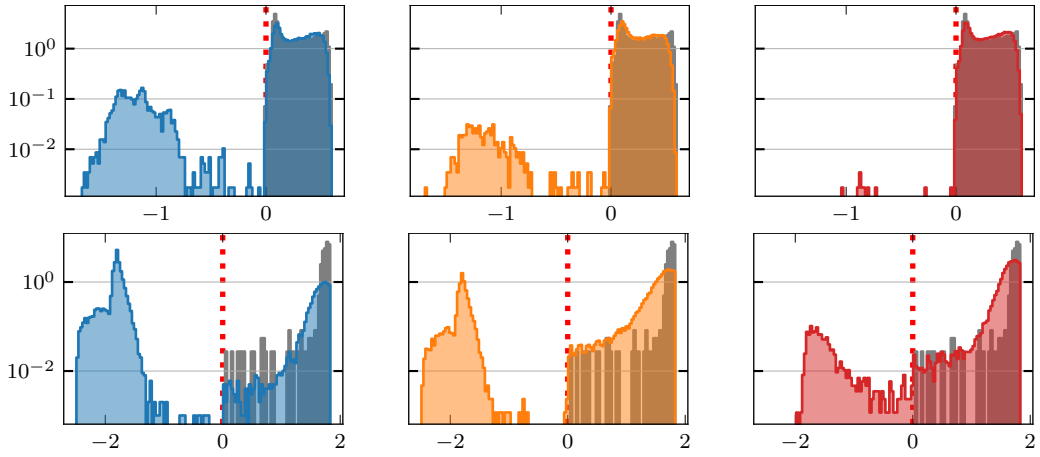
We next sample trajectories of two nonlinear dynamical systems, Lorenz '63 (L'63) (Lorenz, 1963) and FitzHugh-Nagumo (FHN) (FitzHugh, 1961; Nagumo et al., 1962), following (Finzi et al., 2023). For L'63 the constraint requires the trajectory to remain on a single attractor arm; for FHN it requires a neuronal spike. Details are available in Appendix D.2. The unconditional feasibility probabilities $\Pr(\mathbf{X}_0 \in C)$ are $\approx 1/5$ and $1/30$ respectively (Finzi et al., 2023; Huang et al., 2026), testing UA-SMC in a rare-event regime where the bound in Theorem 2 loosens.

	Constraint Satisfaction			
	BoN	FK S.	SPT	UA-SMC
L'63	0.19	0.94	0.99	**1.00**
FHN	0.04	0.35	0.81	**0.97**
	Total Variation			
L'63	0.81	0.13	0.14	**0.11**
FHN	0.98	0.80	0.64	**0.48**

Table 1: Constraint satisfaction rate (↑) and TV (↓) to the reference constraint-value distribution for trajectory sampling.

We compare UA-SMC against Best-of-N, FK Steering and Source Parallel Tempering (SPT) (Wang et al., 2026), a flow-based reward-tilting sampler. Table 1 reports satisfaction rate and TV distance between the distribution of constraint values $V(\mathbf{X}_0)$ of generated trajectories and rejection sampling trajectories generated by an ODE solver. Across all events, Best-of-N performs poorly, as expected given the rarity of the event under the unconditional distribution. On L'63, UA-SMC attains 100% satisfaction and the lowest TV. The relative performance improvement over other methods grows on the rarer FHN event, where UA-SMC achieves 97% satisfaction versus 81% for SPT and 35% for FK Steering, while also retaining the best fidelity with TV 0.48, a 25% reduction over SPT. Figure 4 shows the corresponding constraint histograms, confirming that UA-SMC best satisfies the constraints and tracks the true constraint distribution even for rare events.

5.3 Material Generation



(a) FK Steering (b) SPT (c) UA-SMC

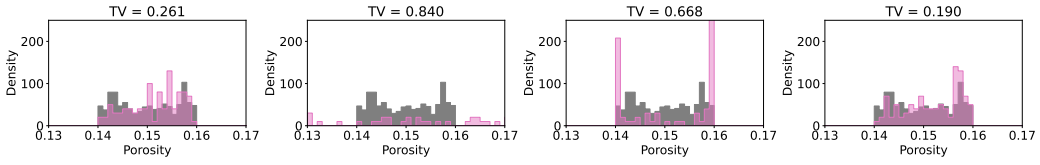
Figure 4: Histograms of constraint values $V(\boldsymbol{x}_0)$ for (a) FK Steering, (b) SPT, and (c) UA-SMC on (top) Lorenz '63 and (bottom) FitzHugh-Nagumo. Exact constraint distribution is in gray, generated constraint distributions are displayed in blue, orange, and red. UA-SMC's histograms most closely recover the constraint distribution of the rejection sample baseline.

In this experiment, we test the scientific applicability of UA-SMC for black-box physical constraints. We consider material porosity, the fraction of void space, which governs gas storage, catalytic activity, and mechanical strength of materials (Davis, 2002; Furukawa et al., 2013). We train an unconditional diffusion model on synthetic materials with porosities in $[0.1, 0.4]$ and constrain generation to $[0.14, 0.16]$, which contains only 4% of training data.

Method	Satisfaction ↑	TV ↓
BoN	16%	0.840
FK-Steering	98%	0.261
PDM	**100%**	0.668
UA-SMC	**100%**	**0.190**

Table 2: Comparison on porous material generation. Satisfaction rate and TV distance to the feasible training distribution for porosity in $[0.14, 0.16]$.

Table 2 reports satisfaction and TV distance between the porosities of 100 generated samples and the in-range training porosity distribution. Best-of-N again achieves poor constraint satisfaction as the constrained porosity range is rare under the model's unconditional sampling distribution. FK Steering reaches 98% satisfaction; their point-estimate potentials encourage feasibility but can commit early to trajectories that ultimately terminate outside of C. PDM achieves 100% satisfaction but, as shown in Figure 5 accumulates samples on the boundary of C. This is a consequence of the intermediate projection which maps samples outside the set to the nearest porosity boundary, and leads to a poor TV distance of 0.668. UA-SMC achieves 100% satisfaction with the lowest TV error of 0.190, combining exact empirical feasibility with strong agreement to the target porosity distribution. A comparison of the generated structures is available in Appendix D.3.



(a) FK Steering (b) Best-of-N (c) PDM (d) UA-SMC (Ours)

Figure 5: Porosity distributions for FK Steering, Best-of-N, PDM, and UA-SMC. Training data porosities are in gray, generated porosities are in pink. UA-SMC's histograms most closely recover the constraint distribution of the original training data.

5.4 Text-to-Image Generation

To demonstrate UA-SMC's applicability for content guidelines, we consider the task of copyright-safe generation. Diffusion models are known to memorize and reproduce individual training examples (Carlini et al., 2023), including recognizable people and trademarked content. Prior work (He

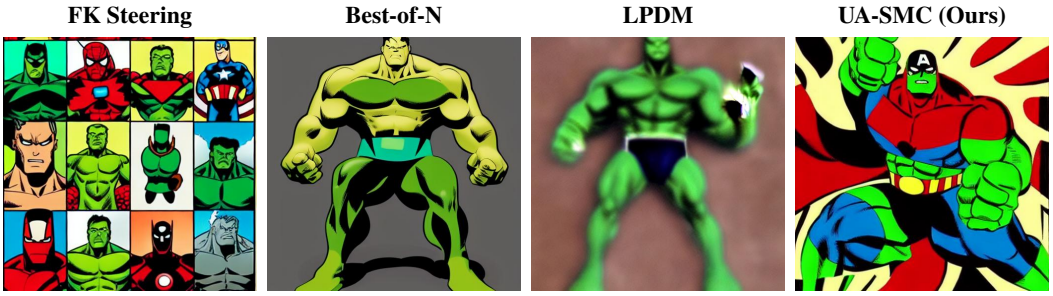


Figure 6: Comparison of generated samples from FK-Steering, Best-of-N, LPDM and UA-SMC and their classifier scores.

et al., 2025) identifies *weak anchors*: combinations of descriptive terms that elicit copyrighted content from models such as Stable Diffusion (Rombach et al., 2022).

We evaluate whether UA-SMC can prevent such generations while preserving sample quality. Specifically, we prompt Stable Diffusion using weak anchors associated with the copyrighted characters *Squirtle* from the *Pokemon* franchise and *The Hulk* from the *Marvel* franchise. We then use a pretrained model to compute a character-detection score, taking the maximum score over several textual descriptions of each character. We define an image as feasible when this score is at most 5%. We evaluate each method using its constraint-satisfaction rate and TOPIQ (Chen et al., 2024), a no-reference image-quality metric sensitive to visual distortions.

	Squirtle		Hulk	
Method	Satisfaction	TOPIQ ↑	Satisfaction	TOPIQ ↑
BoN	55%	$\underline{0.76 \pm 0.05}$	88%	$\mathbf{0.60 \pm 0.10}$
FK-Steering	86%	$\underline{0.76 \pm 0.06}$	86%	$\underline{0.59 \pm 0.09}$
LPDM	$\mathbf{98}\%$	0.64 ± 0.17	$\mathbf{99}\%$	0.54 ± 0.10
UA-SMC	$\underline{92\%}$	$\mathbf{0.77 \pm 0.04}$	$\underline{97\%}$	$\underline{0.59 \pm 0.08}$

Table 3: Constraint satisfaction rates (↑) and TOPIQ scores (↑) for text-to-image generation targeting Squirtle and the Hulk. Bold indicates the best result and underline indicates the second-best result.

Table 3 compares Best-of-N, FK Steering, LPDM, and UA-SMC. Best-of-N attains the lowest constraint satisfaction, consistent with the base model's tendency to generate the target character under weak-anchor prompts. FK Steering generally improves satisfaction, but its fixed-temperature potential applies the same selectivity throughout denoising and can prematurely concentrate particles on trajectories that ultimately violate the terminal constraint. LPDM and UA-SMC achieve the highest constraint satisfaction rates, ranging from 92-99%. However, LPDM's projection-based corrections distort sample geometry, leading to visible artifacts and around a 10-20% reduction in TOPIQ compared to UA-SMC.

Figure 6 illustrates this trade-off qualitatively. Best-of-N and FK Steering produce recognizable instances of the Hulk while retaining high visual quality. LPDM suppresses detection of the Hulk, but introduces noticeable visual distortions, consistent with its lower TOPIQ score. In contrast, UA-SMC does not generate the Hulk and preserves high image quality. Together, these results suggest that UA-SMC achieves a favorable balance between hard-constraint satisfaction and sample quality in high-dimensional text-to-image generation.

6 Concluding Remarks

This paper introduced UA-SMC, an SMC sampler that accounts for uncertainty in predicted terminal estimates through a conditional CDF interpretation of the exact intermediate potential. UA-SMC constructs uncertainty-aware feasibility potentials without requiring convexity or gradients of the constraint function. We bound the resulting approximation error and its contribution to the finite-particle sampling error. Across dynamical systems, material generation, and text-to-image generation, UA-SMC improves constraint satisfaction and distributional fidelity over projection- and SMC-based methods. Future work includes adapting this framework to other generative models (e.g. flow matching (Lipman et al., 2023; Liu et al., 2023; Albergo & Vanden-Eijnden, 2023)) and extending UA-SMC to other scientific applications, such as constrained molecular generation.

AI USE STATEMENT

In this work, we used generative AI tools for language editing and for code enhancements aimed at improving computational efficiency. All AI-assisted content was reviewed by the authors. In particular, all AI-assisted code modifications were independently verified and tested for correctness by two authors. The authors take full responsibility for the final content of this work, including all text, claims, code, and other artifacts produced with the assistance of generative AI.

ETHICS STATEMENT

UA-SMC is designed to enforce hard constraints with diffusion models. We study it for positive uses such as generating physically feasible porous materials within a narrow target porosity range, sampling rare trajectories of dynamical systems, and reducing reproduction of copyrighted characters in text-to-image generation. Our experiments use synthetic materials, academic dynamical systems, and open source text-to-image models with prompts known to elicit copyrighted characters. No human subjects or personal data are involved.

The same ability to enforce arbitrary black-box constraints could be misused to suppress legitimate content or to optimize designs for harmful purposes. Our copyright filter relies on an automated image text detector which may be biased and does not constitute a legal judgment of infringement.

REPRODUCIBILITY STATEMENT

Algorithm 1 gives the complete UA-SMC sampling procedure. Appendix D provides task-specific constraint definitions, model and sampling configurations, computational budgets, and additional implementation details for all experiments. To ensure reproducibility, all codes have been submitted as supplementary materials. Appendix B describes the procedure used to select the practical sharpness schedule. All experiments were conducted on NVIDIA RTX 3090 GPUs. Complete assumptions and proofs for the theoretical results in Section 4 are provided in Appendix C.

REFERENCES

Michael Samuel Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=li7qeBbCR1t.

Benjamin Boys, Mark Girolami, Jakiw Pidstrigach, Sebastian Reich, Alan Mosca, and O Deniz Akyildiz. Tweedie moment projected diffusions for inverse problems. *arXiv preprint arXiv:2310.06721*, 2023.

Sebastien Bubeck, Ronen Eldan, and Joseph Lehec. Finite-time analysis of projected langevin monte carlo. *Advances in Neural Information Processing Systems*, 28, 2015.

Nicolas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramer, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. In *32nd USENIX security symposium (USENIX Security 23)*, pp. 5253–5270, 2023.

Chaofeng Chen, Jiadi Mo, Jingwen Hou, Haoning Wu, Liang Liao, Wenxiu Sun, Qiong Yan, and Weisi Lin. Topiq: A top-down approach from semantics to distortions for image quality assessment. *IEEE Transactions on Image Processing*, 33:2404–2418, 2024.

Jacob K Christopher, Stephen Baek, and Ferdinando Fioretto. Constrained synthesis with projected diffusion models. *Advances in Neural Information Processing Systems*, 37:89307–89333, 2024.

Hyungjin Chung, Jeongsol Kim, Michael T Mccann, Marc L Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. *arXiv preprint arXiv:2209.14687*, 2022.

Gabriele Corso, Hannes Stärk, Bowen Jing, Regina Barzilay, and Tommi Jaakkola. Diffdock: Diffusion steps, twists, and turns for molecular docking. *arXiv preprint arXiv:2210.01776*, 2022.

Dan Crisan and Arnaud Doucet. A survey of convergence results on particle filtering methods for practitioners. *IEEE Transactions on signal processing*, 50(3):736–746, 2002.

Mark E Davis. Ordered porous materials for emerging applications. *Nature*, 417(6891):813–821, 2002.

Bradley Efron. Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106(496):1602–1614, 2011.

Mohammad Farazmand and Themistoklis P Sapsis. Extreme events: Mechanisms and prediction. *Applied Mechanics Reviews*, 71(5):050801, 2019.

Marc Anton Finzi, Anudhyan Boral, Andrew Gordon Wilson, Fei Sha, and Leonardo Zepeda-Núñez. User-defined event sampling and uncertainty quantification in diffusion models for physical dynamical systems. In *International Conference on Machine Learning*, pp. 10136–10152. PMLR, 2023.

Richard FitzHugh. Impulses and physiological states in theoretical models of nerve membrane. *Biophysical journal*, 1(6):445–466, 1961.

Xiang Fu, Tian Xie, Andrew Rosen, Tommi Jaakkola, and Jake Smith. Mofdiff: Coarse-grained diffusion for metal-organic framework design. In *International Conference on Learning Representations*, volume 2024, pp. 16510–16532, 2024.

Hiroyasu Furukawa, Kyle E Cordova, Michael O'keeffe, and Omar M Yaghi. The chemistry and applications of metal-organic frameworks. *Science*, 341(6149):1230444, 2013.

Zhengyi Guo, Wenpin Tang, and Renyuan Xu. Conditional diffusion guidance under hard constraint: A stochastic analysis approach. *arXiv preprint arXiv:2602.05533*, 2026.

Luxi He, Yangsibo Huang, Weijia Shi, Tinghao Xie, Haotian Liu, Yue Wang, Luke Zettlemoyer, Chiyuan Zhang, Danqi Chen, and Peter Henderson. Fantastic copyrighted beasts and how (not) to generate them. In *International Conference on Learning Representations*, volume 2025, pp. 84507–84534, 2025.

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

Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022a.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in neural information processing systems*, 35:8633–8646, 2022b.

Yuhao Huang, Taos Transue, Shih-Hsin Wang, William Feldman, Hong Zhang, and Bao Wang. Improving flow matching by aligning flow divergence. *arXiv preprint arXiv:2602.00869*, 2026.

Michael Janner, Yilun Du, Joshua B Tenenbaum, and Sergey Levine. Planning with diffusion for flexible behavior synthesis. *arXiv preprint arXiv:2205.09991*, 2022.

Shyamgopal Karthik, Karsten Roth, Massimiliano Mancini, and Zeynep Akata. If at first you don't succeed, try, try again: Faithful diffusion-based text-to-image generation by selection. *arXiv preprint arXiv:2305.13308*, 2023.

Sunwoo Kim, Minkyu Kim, and Dongmin Park. Test-time alignment of diffusion models without reward over-optimization. In *International Conference on Learning Representations*, volume 2025, pp. 87238–87272, 2025.

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

Xingchao Liu, Chengyue Gong, and qiang liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=XVjTT1nw5z.

Edward N Lorenz. Deterministic nonperiodic flow. *Journal of the Atmospheric Sciences*, 20(2): 130–141, 1963.

Byeonghu Na, Mina Kang, Jiseok Kwak, Minsang Park, Jiwoo Shin, SeJoon Jun, Gayoung Lee, Jin-Hwa Kim, and Il-Chul Moon. Training-free safe text embedding guidance for text-to-image diffusion models. *Advances in Neural Information Processing Systems*, 38:85984–86014, 2026.

Christian A Naesseth, Fredrik Lindsten, and Thomas B Schön. Elements of sequential monte carlo. *Foundations and Trends in Machine Learning*, 12(3):187–306, 2019.

J Nagumo, S Arimoto, and S Yoshizawa. An active pulse transmission line simulating nerve axon. *Proceedings of the IRE*, 50(10):2061–2070, 1962.

Neal Parikh and Stephen Boyd. Proximal algorithms. *Foundations and Trends in optimization*, 1(3): 127–239, 2014.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *International Conference on Learning Representations*, volume 2024, pp. 1862–1874, 2024.

Herbert Robbins. The empirical bayes approach to statistical decision problems. *The Annals of Mathematical Statistics*, 35(1):1–20, 1964.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 10674–10685. ieee, 2022.

Yorai Shaoul, Itamar Mishani, Shivam Vats, Jiaoyang Li, and Maxim Likhachev. Multi-robot motion planning with diffusion models. In *International Conference on Learning Representations*, volume 2025, pp. 95791–95811, 2025.

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. *arXiv preprint arXiv:2501.06848*, 2025.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pp. 2256–2265. pmlr, 2015.

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

Jiaming Song, Arash Vahdat, Morteza Mardani, and Jan Kautz. Pseudoinverse-guided diffusion models for inverse problems. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=9_gsMA8MRKQ.

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

Masatoshi Uehara, Yulai Zhao, Chenyu Wang, Xiner Li, Aviv Regev, Sergey Levine, and Tommaso Biancalani. Inference-time alignment in diffusion models with reward-guided generation: Tutorial and review. *arXiv preprint arXiv:2501.09685*, 2025.

Shih-Hsin Wang, Joel Keller, Taos Transue, Drake Benjamin Brown, Thomas Strohmer, and Bao Wang. Test-time guidance for flow-based generative models via parallel tempering on source distributions. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=5jGcYMUZx1.

Luhuan Wu, Brian Trippe, Christian Naesseth, David Blei, and John P Cunningham. Practical and asymptotically exact conditional sampling in diffusion models. *Advances in Neural Information Processing Systems*, 36:31372–31403, 2023.

Tian Xie, Xiang Fu, Octavian-Eugen Ganea, Regina Barzilay, and Tommi Jaakkola. Crystal diffusion variational autoencoder for periodic material generation. *arXiv preprint arXiv:2110.06197*, 2021.

Shihang Xu, Shibing Chu, Rami Mrad, Zhejun Zhang, Zhelin Li, Runxian Jiao, and Yuanping Chen. Discovery of 2d materials via symmetry-constrained diffusion model. *The Journal of Physical Chemistry C*, 129(14):6794–6802, 2025.

Stefano Zampini, Jacob K Christopher, Luca Oneto, Davide Anguita, and Ferdinando Fioretto. Training-free constrained generation with stable diffusion models. *Advances in Neural Information Processing Systems*, 38:27285–27316, 2026.

A Function-Threshold Examples

UA-SMC requires the user to specify a function-threshold pair $(V, \hat{V})$ representing the feasible set C through

$$\boldsymbol{x} \in C \iff V(\boldsymbol{x}) \geq \hat{V}.$$

The practical UA-SMC algorithm (Algorithm 1) only evaluates V and therefore does not require V to be equipped with accessible gradients. This representation encompasses a broad class of black-box and structured constraints.

- **Membership oracle.** Any set C can be represented using a binary membership oracle. One may therefore construct a function-threshold pair as

$$V(\boldsymbol{x}) = \begin{cases} 1, & \boldsymbol{x} \in C, \\ 0, & \boldsymbol{x} \notin C, \end{cases} \qquad \hat{V} = \frac{1}{2}.$$

- **Multi-objective classifier scores.** Let $s_j(\boldsymbol{x})$ denote the output of a classifier for property j, and suppose each score is required to lie within an acceptable interval $[\ell_j, u_j]$, for $j = 1, \ldots, m$. The feasible set is

$$C = \{\boldsymbol{x} : \ell_j \leq s_j(\boldsymbol{x}) \leq u_j, \quad j = 1, \ldots, m\}.$$

  A corresponding function-threshold pair is

$$V(\boldsymbol{x}) = \min_{j=1,\ldots,m} \{s_j(\boldsymbol{x}) - \ell_j,\ u_j - s_j(\boldsymbol{x})\}, \qquad \hat{V} = 0.$$

- **Simulator-based constraint metric.** Let $S(\boldsymbol{x})$ denote a black-box simulator or evaluation procedure that returns a scalar constraint metric, and suppose feasibility requires this metric to exceed a prescribed threshold κ:

$$C = \{\boldsymbol{x} : S(\boldsymbol{x}) \geq \kappa\}.$$

  Then a corresponding function-threshold pair is

$$V(\boldsymbol{x}) = S(\boldsymbol{x}) - \kappa, \qquad \hat{V} = 0.$$

These examples illustrate that UA-SMC applies to a broad set of constraint functions. In particular, it remains applicable when gradients of V are undefined, unavailable, or prohibitively expensive, as can occur for membership oracles, multi-objective classifier-based constraints, and black-box simulators.

B Heuristic Schedule

B.1 Sharpness Sweep

When the state-dependent uncertainty is impractical to compute, we construct an approximate time-dependent sharpness using the following schedule:

$$\alpha_t = \alpha_{\min} + \left(1 - \frac{t}{T}\right)^p (\alpha_{\max} - \alpha_{\min}).$$

This schedule increases from $\alpha_{\min}$ at the beginning of denoising to $\alpha_{\max}$ at the terminal step, while p controls how rapidly the sharpness increases. We select candidate $\alpha_{\min}$, $\alpha_{\max}$, and p through a grid search, evaluating each configuration according to constraint satisfaction and the task-specific fidelity metric. Although the schedule is no longer state dependent, it captures the overall contraction of uncertainty during denoising and performs well empirically across our experiments.

B.2 Comparison to oracle

Section 3.2 replaces the time- and state-dependent oracle sharpness $\alpha_t^\star(\boldsymbol{x}_t)$ with a time-only schedule $\{\alpha_t\}_{t=T-1}^0$. Except in Section 5.1 where $\alpha_t^\star$ is available in closed form, all experiments use the heuristic schedule selected from the sweep in Appendix B.1. The analytical experiment therefore

provides a controlled setting for measuring the performance gap between the practical amortized schedule and the oracle sharpness.

We run a sweep over the following parameters

$$\alpha_{\min} \in \{0, 5, 10\}, \qquad \alpha_{\max} \in \{100, 500, 1000\}, \qquad p \in \{0, 5, 10\}.$$

for six seeds. Across the sweep, the best-performing schedule was $(\alpha_{\min}, \alpha_{\max}, p) = (0, 100, 10)$, achieving TV 0.182 ± 0.004. This closely approaches the oracle state-dependent calibration, which achieves a TV of 0.174, while improving substantially over FK Steering's TV of 0.29. These results show that the practical schedule can recover much of the performance of the oracle sharpness, justifying its use for later experiments.

C Proofs of Theoretical Results

In this section, we provide proofs for the theoretical results presented in Section 4.

Proposition 2. *Following, Chung et al. (2022); Song et al. (2023); Boys et al. (2023), suppose that*

$$\mathbf{X}_0 \mid \mathbf{X}_t = \boldsymbol{x}_t \sim \mathcal{N}(\hat{\boldsymbol{x}}_0(\boldsymbol{x}_t), \Sigma_t(\boldsymbol{x}_t)),$$

and let $V : \mathbb{R}^d \to \mathbb{R}$ be affine. Then

$$\epsilon_t \mid \mathbf{X}_t = \boldsymbol{x}_t \sim \mathcal{N}(0, \tau_t^2(\boldsymbol{x}_t)),$$

where

$$\tau_t^2(\boldsymbol{x}_t) = \nabla V(\hat{\boldsymbol{x}}_0(\boldsymbol{x}_t))^\top \Sigma_t(\boldsymbol{x}_t) \nabla V(\hat{\boldsymbol{x}}_0(\boldsymbol{x}_t)).$$

Proof. Fix $\boldsymbol{x}_t$ and write

$$\mu_t = \hat{\boldsymbol{x}}_0(\boldsymbol{x}_t), \qquad \Sigma_t = \Sigma_t(\boldsymbol{x}_t).$$

Since V is affine,

$$V(\mathbf{X}_0) = V(\mu_t) + \nabla V(\mu_t)^\top (\mathbf{X}_0 - \mu_t).$$

Hence, by the definition of ϵ_t,

$$\epsilon_t = -\nabla V(\mu_t)^\top (\mathbf{X}_0 - \mu_t).$$

Conditional on $\mathbf{X}_t = \boldsymbol{x}_t$,

$$\mathbf{X}_0 - \mu_t \sim \mathcal{N}(0, \Sigma_t).$$

Therefore, $\epsilon_t \mid \mathbf{X}_t = \boldsymbol{x}_t$ is a zero-mean Gaussian with variance

$$\mathrm{Var}(\epsilon_t \mid \mathbf{X}_t = \boldsymbol{x}_t) = \nabla V(\mu_t)^\top \Sigma_t \nabla V(\mu_t) = \tau_t^2(\boldsymbol{x}_t).$$

□

Proposition 3. *Let $\epsilon_t^{(1)}$ denote the first-order linearization of ϵ_t for a nonlinear constraint function V. For every t, the discrepancy between the exact conditional CDF of ϵ_t and its first-order Gaussian approximation is uniformly bounded over $z \in \mathbb{R}$ as*

$$\sup_{z \in \mathbb{R}} |\Pr(\epsilon_t \le z \mid \mathbf{X}_t = \boldsymbol{x}_t) - \Pr(\epsilon_t^{(1)} \le z \mid \mathbf{X}_t = \boldsymbol{x}_t)| \le \mathcal{E}_t^{Gauss}$$

where $\mathcal{E}_t^{Gauss} = \sqrt{\frac{\tau_t}{2\pi}} + \Pr\left(\frac{|R_t(\boldsymbol{x}_t)|}{\tau_t^{3/2}} > 1 \mid \mathbf{X}_t = \boldsymbol{x}_t \right) \to 0$ as $t \to 0$

Proof. To simplify notation, let A and B denote the events

$$A = \{\epsilon_t \le z\}, \;\; B = \{\epsilon_t^{(1)} \le z\}.$$

For arbitrary event E, define the conditional probability measure

$$\mathbb{P}_{\boldsymbol{x}_t}(E) := \Pr(E \mid \mathbf{X}_t = \boldsymbol{x}_t).$$

Since $\mathbb{P}_{\boldsymbol{x}_t}(A), \mathbb{P}_{\boldsymbol{x}_t}(B) \ge 0$, their absolute difference satisfies

$$|\mathbb{P}_{\boldsymbol{x}_t}(A) - \mathbb{P}_{\boldsymbol{x}_t}(B)| \le \max\{\mathbb{P}_{\boldsymbol{x}_t}(A \cap B^C), \mathbb{P}_{\boldsymbol{x}_t}(B \cap A^C)\}$$

Fix $u > 0$ and define the good-remainder event

$$\mathcal{G}_u = \{|R_t| \le u\}, \qquad R_t = \epsilon_t - \epsilon_t^{(1)}.$$

Since $\mathcal{G}_u$ and $\mathcal{G}_u^c$ partition the sample space,

$$\mathbb{P}_{\boldsymbol{x}_t}(A \cap B^c) = \mathbb{P}_{\boldsymbol{x}_t}(A \cap B^c \cap \mathcal{G}_u) + \mathbb{P}_{\boldsymbol{x}_t}(A \cap B^c \cap \mathcal{G}_u^c).$$

Now enforce the event A and substitute the decomposition $\epsilon_t = \epsilon_t^{(1)} + R_t$:

$$\epsilon_t \le z \implies \epsilon_t^{(1)} + R_t \le z \implies \epsilon_t^{(1)} \le z - R_t.$$

Additionally enforcing the event $\mathcal{G}_u$, we obtain

$$\epsilon_t^{(1)} \le z - R_t \le z + |R_t| \le z + u.$$

Finally, enforcing the event B^c gives

$$\epsilon_t^{(1)} > z.$$

Combining these inequalities yields the event inclusion

$$\{A \cap B^c \cap \mathcal{G}_u\} \subseteq \left\{z < \epsilon_t^{(1)} \le z + u\right\}.$$

The second event also admits a simple inclusion

$$\{A \cap B^C \cap \mathcal{G}_u^C\} \subseteq \{\mathcal{G}_u^C\},$$

yielding the following bound

$$\mathbb{P}_{\boldsymbol{x}_t}(A \cap B^C) \le \mathbb{P}_{\boldsymbol{x}_t}(z \le \epsilon_t^1 \le z + u) + \mathbb{P}_{\boldsymbol{x}_t}(|R_t| > u)$$

Using the same argument one finds

$$\mathbb{P}_{\boldsymbol{x}_t}(B \cap A^C) \le \mathbb{P}_{\boldsymbol{x}_t}(z - u \le \epsilon_t^1 \le z) + \mathbb{P}_{\boldsymbol{x}_t}(|R_t| > u)$$

Therefore the absolute discrepancy satisfies

$$|\mathbb{P}_{\boldsymbol{x}_t}(A) - \mathbb{P}_{\boldsymbol{x}_t}(B)| \le \max\{\mathbb{P}_{\boldsymbol{x}_t}(z \le \epsilon_t^1 \le z + u), \mathbb{P}_{\boldsymbol{x}_t}(z - u \le \epsilon_t^1 \le z)\} + \mathbb{P}_{\boldsymbol{x}_t}(|R_t| > u)$$

Recall that $\epsilon_t^1 \mid \mathbf{X}_t = \boldsymbol{x}_t$ is a Gaussian random variable with bounded density

$$f_Y(s) = \frac{1}{\tau_t\sqrt{2\pi}} \exp\left(-\frac{s^2}{2\tau_t^2}\right) \le \frac{1}{\tau_t\sqrt{2\pi}}$$

Therefore, $\mathbb{P}_{\boldsymbol{x}_t}(z \le \epsilon_t^1 \le z + u)$ is bounded

$$\mathbb{P}_{\boldsymbol{x}_t}(z \le Y \le z + u) = \int_z^{z+u} f_Y(s)ds \le \int_z^{z+u} \frac{1}{\tau\sqrt{2\pi}}ds = \frac{u}{\tau\sqrt{2\pi}}$$

The same argument can be used to show

$$\mathbb{P}_{\boldsymbol{x}_t}(z - u \le Y \le z) \le \frac{u}{\tau\sqrt{2\pi}}.$$

Therefore, for $u > 0$

$$|\mathbb{P}_{\boldsymbol{x}_t}(A) - \mathbb{P}_{\boldsymbol{x}_t}(B)| \le \frac{u}{\tau\sqrt{2\pi}} + \mathbb{P}_{\boldsymbol{x}_t}(|R_t| > u)$$

In particular, let $u = \tau_t^{3/2}$ and observe the following

$$\begin{aligned}|\Pr(\epsilon_t \le z \mid \mathbf{X}_t = \boldsymbol{x}_t) - \Pr(\epsilon_t^1 \le z \mid \mathbf{X}_t = \boldsymbol{x}_t)| &= |\mathbb{P}_{\boldsymbol{x}_t}(A) - \mathbb{P}_{\boldsymbol{x}_t}(B)| \\ &\le \sqrt{\frac{\tau_t}{2\pi}} + \Pr\left(\frac{|R_t|}{\tau_t^{3/2}} > 1 \mid \mathbf{X}_t = \boldsymbol{x}_t\right)\end{aligned}$$

As $t \to 0$, the posterior over $\mathbf{X}_0$ given $\mathbf{X}_t$ contracts around $\hat{\boldsymbol{x}}_0(\boldsymbol{x}_t)$ and $\Delta_t \to 0$. Since $|R_t| = \mathcal{O}_p(\Delta_t^2)$ and $\tau_t^{3/2} = \Theta(\Delta_t^{3/2})$ then $\mathcal{E}_t^{\text{Gauss}}(\boldsymbol{x}_t) \to 0$.

□

Proposition 4. *Let $\Phi\left(\frac{z}{\tau_t}\right)$ denote the CDF of the first-order Gaussian approximation ϵ_t^1 and $\sigma(cz)$ denote its sigmoid approximation. The error between these CDFs satisfies*

$$\left|\Phi\left(\frac{z}{\tau_t}\right) - \sigma(\alpha_t z)\right| \leq \mathcal{E}_t^{sig}(z)$$

where

$$\mathcal{E}_t^{sig}(z) = \epsilon_{sig}(c) + \frac{|z|}{4}|\alpha_t^* - \alpha_t| \quad \epsilon_{sig}(c) = \sup_{z\in\mathbb{R}} |\Phi(z) - \sigma(cz)|$$

Proof. By the triangle inequality

$$\left|\Phi\left(\frac{z}{\tau_t}\right) - \sigma(\alpha_t z)\right| \leq \left|\Phi\left(\frac{z}{\tau_t}\right) - \sigma(\alpha_t^* z)\right| + |\sigma(\alpha_t^* z) - \sigma(\alpha_t z)|$$

Let $\alpha_t^\star = c/\tau_t(\boldsymbol{x}_t)$ for some $c > 0$. Then the intrinsic error of approximating the Gaussian CDF by a scaled sigmoid is bounded by

$$\left|\Phi\left(\frac{z}{\tau_t}\right) - \sigma(\alpha_t^\star z)\right| \leq \epsilon_{\text{sig}}(c), \qquad \epsilon_{\text{sig}}(c) := \sup_{u\in\mathbb{R}} |\Phi(u) - \sigma(cu)|\,.$$

To bound the second term, we invoke that the sigmoid is Lipschitz with constant $1/4$

$$|\sigma(\alpha_t^* z) - \sigma(\alpha_t z)| \leq \frac{1}{4}|\alpha_t^* z - \alpha_t z| \leq \frac{|z|}{4}|\alpha_t^* - \alpha_t|$$

Combining results and defining $\mathcal{E}_t^{sig}$ as written below

$$\mathcal{E}_t^{sig}(z) = \epsilon_{sig}(c) + \frac{|z|}{4}|\alpha_t^* - \alpha_t|,$$

we arrive at the desired conclusion. □

Proposition 1. *Under the posterior assumptions of Chung et al. (2022), the discrepancy between the exact and UA-SMC potentials satisfies*

$$|h_t(\boldsymbol{x}_t) - \tilde{h}_t(\boldsymbol{x}_t)| \leq \mathcal{E}_t^{\text{Gauss}}(\boldsymbol{x}_t) + \mathcal{E}_t^{\text{sig}}(\boldsymbol{x}_t).$$

Proof. Insert the conditional CDF of the first-order approximation and apply the triangle inequality:

$$\begin{aligned}|h_t(\boldsymbol{x}_t) - \tilde{h}_t(\boldsymbol{x}_t)| \leq & \left|\Pr(\epsilon_t \leq r_t \mid \mathbf{X}_t = \boldsymbol{x}_t) - \Pr(\epsilon_t^{(1)} \leq r_t \mid \mathbf{X}_t = \boldsymbol{x}_t)\right| \\ & + \left|\Phi\left(\frac{r_t}{\tau_t}\right) - \sigma(\alpha_t r_t)\right|.\end{aligned}$$

Applying Propositions 3 and 4 and evaluating at $z = r_t$ gives the desired result.

□

Theorem 1. *Let $\pi_t(\boldsymbol{x}_{T:t}) \propto p_\theta(\boldsymbol{x}_{T:t})h_t(\boldsymbol{x}_t)$ and $\tilde{\pi}_t(\boldsymbol{x}_{T:t}) \propto p_\theta(\boldsymbol{x}_{T:t})\tilde{h}_t(\boldsymbol{x}_t)$. Then*

$$\|\pi_t - \tilde{\pi}_t\|_{\text{TV}} \leq \frac{\mathbb{E}_{p_\theta}[|h_t(\mathbf{X}_t) - \tilde{h}_t(\mathbf{X}_t)|]}{\mathbb{E}_{p_\theta}[h_t(\mathbf{X}_t)]}.$$

Proof. Denote the path measure of the reverse diffusion model

$$Q_t(d\boldsymbol{x}_{T:t}) = p_\theta(\boldsymbol{x}_{T:t})d\boldsymbol{x}_{T:t}$$

We can express our intermediate-time targets as

$$\pi_t(\boldsymbol{x}_{T:t}) = \frac{Q_t(d\boldsymbol{x}_{T:t})h_t(\boldsymbol{x}_t)}{Z_t} \qquad \tilde{\pi}_t(\boldsymbol{x}_{T:t}) = \frac{Q_t(d\boldsymbol{x}_{T:t})\tilde{h}_t(\boldsymbol{x}_t)}{\tilde{Z}_t}$$

with the normalization constants

$$Z_t = \mathbb{E}_{Q_t(d\boldsymbol{x}_{T:t})}[h_t(\boldsymbol{x}_t)] \qquad \tilde{Z}_t = \mathbb{E}_{Q_t(d\boldsymbol{x}_{T:t})}[\tilde{h}_t(\boldsymbol{x}_t)]$$

By the tower property

$$Z_t = \mathbb{E}_{Q_t}[h_t(\mathbf{X}_t)] = \Pr_\theta(\mathbf{X}_0 \in C)$$

We can bound the total variation between measures

$$2\|\pi_t(\boldsymbol{x}_{T:t}) - \tilde{\pi}_t(\boldsymbol{x}_{T:t})\|_{TV} = \int \left| \frac{h_t(\boldsymbol{x}_t)}{Z_t} - \frac{\tilde{h}_t(\boldsymbol{x}_t)}{\tilde{Z}_t} \right| Q_t(d\boldsymbol{x}_{T:t})$$

With some algebra, we can re-express this

$$\begin{aligned} 2\|\pi_t(\boldsymbol{x}_{T:t}) - \tilde{\pi}_t(\boldsymbol{x}_{T:t})\|_{TV} &\leq \int \left| \frac{h_t(\boldsymbol{x}_t) - \tilde{h}_t(\boldsymbol{x}_t)}{Z_t} \right| + \left| \tilde{h}_t(\boldsymbol{x}_t) \left(\frac{1}{Z_t} - \frac{1}{\tilde{Z}_t} \right) \right| Q_t(d\boldsymbol{x}_{T:t}) \\ &\leq \int \left| \frac{h_t(\boldsymbol{x}_t) - \tilde{h}_t(\boldsymbol{x}_t)}{Z_t} \right| Q_t(d\boldsymbol{x}_{T:t}) + \int \left| \tilde{h}_t(\boldsymbol{x}_t) \left(\frac{1}{Z_t} - \frac{1}{\tilde{Z}_t} \right) \right| Q_t(d\boldsymbol{x}_{T:t}) \end{aligned}$$

The first term is proportional to the expected error in potentials

$$\int \left| \frac{h_t(\boldsymbol{x}_t) - \tilde{h}_t(\boldsymbol{x}_t)}{Z_t} \right| Q_t(d\boldsymbol{x}_{T:t}) = \frac{1}{Z_t} \mathbb{E}_{Q_t} |h_t(\boldsymbol{x}_t) - \tilde{h}_t(\boldsymbol{x}_t)|$$

The second term bounded by a term proportional to expectation of the error in potentials

$$\begin{aligned} \int \left| \tilde{h}_t(\boldsymbol{x}_t) \left(\frac{1}{Z_t} - \frac{1}{\tilde{Z}_t} \right) \right| Q_t(d\boldsymbol{x}_{T:t}) &= \left| \frac{1}{Z_t} - \frac{1}{\tilde{Z}_t} \right| \int \tilde{h}_t(\boldsymbol{x}_t) Q_t(d\boldsymbol{x}_{T:t}) \\ &= \left| \frac{1}{Z_t} - \frac{1}{\tilde{Z}_t} \right| \tilde{Z}_t = \frac{|Z_t - \tilde{Z}_t|}{Z_t} \leq \frac{1}{Z_t} \mathbb{E}_{Q_t} |h_t(\boldsymbol{x}_t) - \tilde{h}_t(\boldsymbol{x}_t)| \end{aligned}$$

Combining results above, we find the total variation between arbitrary marginals is bounded by the error in the potentials

$$2\|\pi_t - \tilde{\pi}_t\|_{TV} \leq \left(\frac{2}{Z_t} \right) \mathbb{E}_{Q_t} |h_t(\boldsymbol{x}_t) - \tilde{h}_t(\boldsymbol{x}_t)| \implies \|\pi_t - \tilde{\pi}_t\|_{TV} \leq \frac{\mathbb{E}_{Q_t} |h_t(\boldsymbol{x}_t) - \tilde{h}_t(\boldsymbol{x}_t)|}{\Pr_Q(\mathbf{X}_0 \in C)}$$

Substituting in reverse diffusion for Q_t gives the desired result. □

Theorem 2. *Under the posterior assumptions of Chung et al. (2022) and standard regularity for Feynman-Kac particle systems (Crisan & Doucet, 2002), for each t there exists $C_t < \infty$ such that for any bounded measurable f,*

$$\mathbb{E}_N \left[|\pi_t(f) - \tilde{\pi}_t^N(f)| \right] \leq C_t \|f\|_\infty \left(\frac{1}{\sqrt{N}} + \frac{\mathbb{E}_{p_\theta}[\mathcal{E}_t^{\text{Gauss}}(\mathbf{X}_t) + \mathcal{E}_t^{\text{sig}}(\mathbf{X}_t)]}{\Pr_\theta(\mathbf{X}_0 \in C)} \right),$$

where $\tilde{\pi}_t^N$ is the N-particle UA-SMC empirical measure and $\mathbb{E}_N$ is over particle randomness.

Proof. Let $\mathbb{E}_N$ denote expectation with respect to the randomness of the particle system. By the triangle inequality,

$$\mathbb{E}_N \left[\left| \pi_t(f) - \tilde{\pi}_t^N(f) \right| \right] \leq \mathbb{E}_N \left| \pi_t(f) - \tilde{\pi}_t(f) \right| + \mathbb{E}_N \left[\left| \tilde{\pi}_t(f) - \tilde{\pi}_t^N(f) \right| \right].$$

Since π_t and $\tilde{\pi}_t$ are deterministic measures,

$$\begin{aligned} \mathbb{E}_N \left| \pi_t(f) - \tilde{\pi}_t(f) \right| = \left| \pi_t(f) - \tilde{\pi}_t(f) \right| &\leq 2\|f\|_\infty \|\pi_t - \tilde{\pi}_t\|_{\text{TV}} \\ &\leq \frac{2\|f\|_\infty}{\Pr_\theta(\mathbf{X}_0 \in C)} \mathbb{E}_{p_\theta(\mathbf{X}_{T:t})} \left[|h_t(\mathbf{X}_t) - \tilde{h}_t(\mathbf{X}_t)| \right] \\ &\leq \frac{2\|f\|_\infty}{\Pr_\theta(\mathbf{X}_0 \in C)} \mathbb{E}_{p_\theta(\mathbf{X}_{T:t})} \left[\mathcal{E}_t^{\text{Gauss}}(\mathbf{X}_t) + \mathcal{E}_t^{\text{sig}}(\mathbf{X}_t) \right], \end{aligned}$$

where the second inequality follows from Theorem 1, and the third follows from Proposition 1.

By the standard fixed-time mean-square convergence result for particle approximations Theorem 2 Crisan & Doucet (2002), there exists a constant $K_t < \infty$, independent of N and f, such that

$$\mathbb{E}_N\left[\left|\tilde{\pi}_t^N(f) - \tilde{\pi}_t(f)\right|^2\right] \leq \frac{K_t \|f\|_\infty^2}{N}.$$

An application of Jensen's inequality gives

$$\mathbb{E}_N\left[\left|\tilde{\pi}_t^N(f) - \tilde{\pi}_t(f)\right|\right] \leq \frac{\sqrt{K_t}\|f\|_\infty}{\sqrt{N}}.$$

Combining the preceding inequalities, there exists a constant $C_t < \infty$, independent of N, such that

$$\mathbb{E}_N\left[\left|\pi_t(f) - \tilde{\pi}_t^N(f)\right|\right] \leq C_t \|f\|_\infty \left(\frac{1}{\sqrt{N}} + \frac{\mathbb{E}_{p_\theta(\mathbf{x}_{T:t})}\left[\mathcal{E}_t^{\text{Gauss}} + \mathcal{E}_t^{\text{sig}}\right]}{\Pr_\theta(\mathbf{X}_0 \in C)}\right).$$

□

D Experimental Details and Additional Results

In this section, for each experimental setting presented in Section 5 we present implementation details and additional results.

D.1 An Analytical Example

This section provides implementation details and additional results for the analytical experiment in Section 5.1. All methods use the same pretrained MLP-based diffusion model and are evaluated under a matched computational budget. Below, we describe the function-threshold pair, sampling configurations, and a sharpness-ablation study.

D.1.1 Function-Threshold Details

For the annular constraint set $C = \{6 \leq \|x\| \leq 8\}$, we used the following function-threshold pair

$$V(\boldsymbol{x}) = (\|x\|_2^2 - 6^2)(8^2 - \|x\|_2^2) \quad \hat{V} = 0.$$

Because this setting is low-dimensional, and the function is differentiable, we used the exact α_t^* sharpness, defined as

$$\alpha_t^* = \frac{c}{\sqrt{\nabla V(\hat{\boldsymbol{x}}_0(\boldsymbol{x}_t))^T \Sigma_t(\boldsymbol{x}_t) \nabla V(\hat{\boldsymbol{x}}_0(\boldsymbol{x}_t))}}$$

where following Finzi et al. (2023), $\Sigma_t(\boldsymbol{x}_t)$ is computed as follows

$$\Sigma_t(\boldsymbol{x}_t) = \frac{\sigma_t^2}{s_t^2}(\boldsymbol{I} + \sigma_t^2 \nabla^2 \log p(\boldsymbol{x}_t)).$$

D.1.2 Computational Comparison

To ensure a fair comparison, we allocate each method the same total budget of 1,000 diffusion model evaluations. For an SMC method with N particles and T reverse-diffusion steps, each particle requires one diffusion model evaluation per step, giving a total budget of TN for UA-SMC, FK-Steering and TDS. For PDM with L diffusion levels and M Langevin updates per level, each update requires one diffusion model evaluation, giving a total budget of LM. Table 4 therefore presents our computational budget, $T = 100$ steps and $N = 10$ particles for the SMC methods, and $L = 10$ diffusion levels with $M = 100$ Langevin updates per level for PDM.

Category	Property	Value
Sampling (SMC)	Denoising Steps T	100
	Number of Particles N	10
Sampling (PDM)	Diffusion Levels L	10
	Langevin Steps M	100

Table 4: Inference hyperparameters for the analytical experiment.

D.1.3 Sharpness Ablation Study

We study how misspecification of the sharpness parameter α_t affects the accuracy of UA-SMC. As shown in Section 4, the approximation error depends on the sharpness mismatch $|\alpha_t - \alpha_t^*|$, where α_t^* denotes the oracle sharpness.



Figure 7: Total variation distance as a function of the sharpness miscalibration factor λ. The oracle sharpness α_t^* corresponds to $\lambda = 1$.

To isolate the effect of this mismatch, we consider the analytical mixture-of-Gaussians experiment from Section 5.1, for which α_t^* is available in closed form. We introduce a global miscalibration factor $\lambda > 0$ and define the family of sharpness schedules

$$\alpha_t(\lambda) = \lambda \alpha_t^*, \qquad \forall t. \tag{4}$$

Thus, $\lambda = 1$ recovers the oracle schedule used in Section 5.1, while deviations from one introduce a mismatch of

$$|\alpha_t(\lambda) - \alpha_t^*| = |\lambda - 1|\,|\alpha_t^*|. \tag{5}$$

Figure 7 reports the TV distance between the empirical distribution of the generated samples and the analytical target distribution. The distance is minimized at the calibrated value $\lambda = 1$ and increases as λ moves away from one. This behavior supports the theoretical prediction that larger sharpness mismatch leads to greater approximation error.

D.2 Dynamical Systems

This section provides experimental details and additional results for the dynamical systems experiment in Section 5.2. The experimental setup closely follows those in (Finzi et al., 2023; Huang et al., 2026; Wang et al., 2026). All methods use the same pretrained UNet-based diffusion model and are evaluated under a similar computational budget. Below, we describe the function-threshold pair, and sampling configurations. For UA-SMC, we performed the sweep described in Section B.1 to generate α_t schedule $(\alpha_{\min}, \alpha_{\max}, p) = (0, 40, 0)$.

D.2.1 Dynamical Systems and Function-Threshold Details

Lorenz '63 The dynamical system introduced in (Lorenz, 1963) is given by

$$\begin{bmatrix}\dot{x}_1\\ \dot{x}_2\\ \dot{x}_3\end{bmatrix} = F(\boldsymbol{x}) = \begin{bmatrix}\sigma(x_2 - x_1)\\ x_1(\rho - x_3) - x_2\\ x_1 x_2 - \beta x_3\end{bmatrix}$$

where chaotic behavior is induced when $\sigma = 10$, $\rho = 28$, and $\beta = 8/3$. Following Finzi et al. (2023), we instead use the scaled system $f(\boldsymbol{x}) = F(20\boldsymbol{x})/20$ to ensure that $x_i(\tau) \in [-3, 3]$ for all dynamics time τ and $i \in [3]$. The event of a trajectory staying on one of the chaotic attractor arms is encoded by the constraint $V(\boldsymbol{x}) = 0.6 - \|\mathcal{F}[\boldsymbol{x} - \bar{\boldsymbol{x}}]\|_1$ with $\hat{V} = 0$ where $\mathcal{F}$ is the Fourier transform over τ, $\|\cdot\|_1$ is the 1-norm over frequency magnitudes and system dimension, and $\bar{\boldsymbol{x}}$ is the mean of $\boldsymbol{x}(\tau)$ over τ.

FitzHugh-Nagumo The dynamical system introduced in (FitzHugh, 1961; Nagumo et al., 1962) models an excitable neuron and is given by

$$\dot{x}_i = x_i(a_i - x_i)(x_i - 1) - y_i + k\sum_{j=1}^{d} A_{ij}(x_j - x_i)$$

$$\dot{y}_i = b_i x_i - c_i y_i$$

for $i \in [2]$. The parameter values used are from Farazmand & Sapsis (2019); Finzi et al. (2023); Huang et al. (2026): $a_1 = a_2 = -0.025794$, $b_1 = 0.0065$, $b_2 = 0.0135$, $c_1 = c_2 = 0.2$, $k = 0.128$, and $A_{ij} = 1 - \delta_{ij}$ where δ_{ij} is the Kronecker delta. The neuron-spiking event is defined by the constraint $V(\boldsymbol{x}) = \max_\tau(x_1(\tau) + x_2(\tau))/2 - 2.5$ with $\hat{V} = 0$.

D.2.2 Computational Comparison

We allocate each method the same total budget of 800 diffusion model evaluations. For a particle-based method with N particles and T reverse-diffusion steps, each particle requires one diffusion model evaluation per step, giving a total budget of TN for UA-SMC, FK-Steering and Best-of-N. SPT additionally performs S source-exploration updates for every chain at each diffusion level. Its total computational budget is therefore STN, a factor of S larger than those of FK Steering and UA-SMC. Table 5 summarizes these configurations. Although SPT has more function evaluations, this comparison matches the number of diffusion levels and parallel trajectories and allows SPT its source-exploration computation.

Category	Property	Value
Sampling (BoN, FK S, UA-SMC)	Diffusion steps T	100
	Parallel trajectories N	8
Sampling (SPT)	Diffusion Steps T	100
	Source-exploration steps S	8
	Trajectores N (L'63)	4
	Trajectories N (FHN)	8

Table 5: Inference hyperparameters for the dynamical-system experiments.

D.2.3 Particle Ablation Study

We study how the number of particles N affects the performance of UA-SMC. As shown in Section 4, the finite-particle error decays as the number of particles N increases. Figure 8 shows that, for both Lorenz '63 and FitzHugh-Nagumo, the TV distance between the distribution of constraint values $V(\mathbf{X}_0)$ of generated trajectories and rejection sampling trajectories generated by an ODE solver decreases as N increases. This TV distance behavior empirically demonstrates the theoretical result that a larger N reduces the TV distance to the target distribution. We also see that constraint satisfaction rate increases with N.

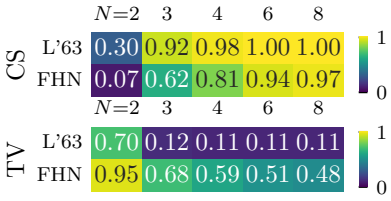


Figure 8: Constraint satisfaction (top) and TV distance (bottom) of UA-SMC as the number of particles N varies.

D.3 Material Generation

This section provides implementation details and additional results for the material generation experiment in Section 5.3, a scientific task with a black-box physical constraint. All methods use the same pretrained U-Net-based diffusion model and are evaluated under a matched computational budget. Below, we describe the function-threshold pair and sampling configurations, followed by a qualitative comparison of the generated samples. For UA-SMC, we performed the sweep described in Section B.1 to generate α_t schedule $(\alpha_{\min}, \alpha_{\max}, p) = (0, 750, 1/3)$.

D.3.1 Function-Threshold Details

For the porosity interval $[0.14, 0.16]$, we used the following function-threshold pair

$$V(\boldsymbol{x}) = -|P(\boldsymbol{x}) - 0.15| \qquad \hat{V} = -0.01$$

where P computes the porosity of the generated sample $\boldsymbol{x}$. A non-zero threshold $\hat{V} = -0.01$ sets a 1% acceptable range about 0.15, allowing for a porosity range of $[0.14, 0.16]$.

D.3.2 Computational Comparison

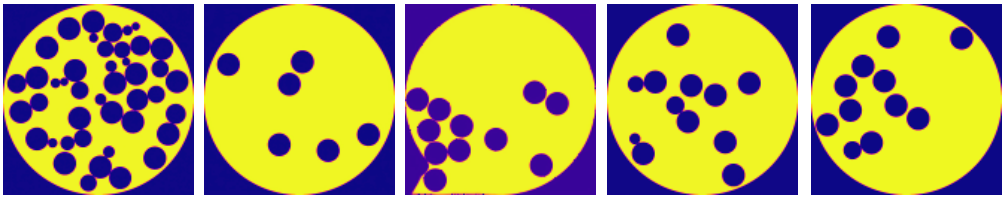
To ensure a fair comparison, we allocate each method the same total budget of 10,000 diffusion model evaluations. For an SMC method with N particles and T reverse-diffusion steps, each particle requires one diffusion model evaluation per step, giving a total budget of TN. For PDM with L diffusion levels and M Langevin updates per level, each update requires one diffusion model evaluation, giving a total budget of LM. Table 6 therefore presents

Category	Property	Value
Sampling (SMC)	Denoising Steps T	1,000
	Number of Particles N	10
Sampling (PDM)	Diffusion Levels L	10
	Langevin Steps M	1,000

Table 6: Hyperparameters for material generation.

our computational budget, $T = 1000$ steps and $K = 10$ particles for the SMC methods, and $L = 10$ diffusion levels with $M = 1000$ Langevin updates per level for PDM.

D.3.3 Generated Sample Comparison



(a) FK Steering Porosity: 42.5% (b) Best-of-N Porosity: 8.4% (c) PDM Porosity: 14.0% (d) UA-SMC (Ours) Porosity: 14.8% (e) Training Sample Porosity: 15.3%

Figure 9: A comparison of generated structures and their porosity. A constraint-satisfying training sample is provided as well.

Figure 9 compares the generated structures for Best-of-N, PDM, FK Steering and UA-SMC. We observe that FK Steering and Best-of-N generate physically valid samples outside of the constraint distribution. We observe that PDM generates porosity-satisfying samples, but displays warped geometry. These observed sample artifacts are the consequence of projecting intermediate samples; the top-k projection algorithm, taken from (Christopher et al., 2024), can break sample geometry to satisfy the constraint. We observe that UA-SMC's sample is geometrically valid and satisfies the constraint.

D.4 Text-To-Image

This section provides implementation details and additional results for the experiment in Section 5.4. All methods use Stable Diffusion v1.4 with the same weak-anchor prompts and DDIM sampling configuration. For *The Hulk*, all methods use the weak-anchor prompt *Marvel, green, strong, superhero, giant, powerful, massive, angry, muscular, cartoon*, while for *Squirtle*, they use *Pokémon, small, blue, popular, tough, evolution, Blastoise, anime, tail, cute, cartoon*. Both weak-anchor prompts are taken from He et al. (2025).

Below, we specify the function-threshold pairs and sampling configurations and present a qualitative comparison of the samples generated by each method. For UA-SMC, we performed the sweep described in Section B.1 to generate α_t schedule $(\alpha_{\min}, \alpha_{\max}, p) = (5, 100, 2)$.

D.4.1 Function-Threshold

We use a pretrained SigLIP2 model to define the character-absence constraint. For target character c, let $\mathcal{D}_c$ denote its set of text descriptions and let $\ell(\boldsymbol{x}, d)$ be the corresponding image-text logit. We define the character score as

$$\Phi_c(\boldsymbol{x}) = \max_{d \in \mathcal{D}_c} \ell(\boldsymbol{x}, d).$$

Character	**Descriptions $\mathcal{D}_c$**
Squirtle	"Squirtle"; "the Pokémon Squirtle"; "Squirtle from the Pokémon franchise"; "Squirtle, the small blue turtle Pokémon"
Hulk	"Hulk"; "the Incredible Hulk"; "the Marvel superhero Hulk"; "an animated image of Hulk"

Table 7: Character descriptions used by the SigLIP detector. Logits are aggregated by taking the maximum over descriptions.

The function-threshold pair is then given by

$$V_c(\boldsymbol{x}) = \ell_{\text{thr}} - \Phi_c(\boldsymbol{x}), \qquad \hat{V} = 0,$$

where $\ell_{\text{thr}} = \text{logit}(0.05)$ is the detector threshold. Thus, $V_c(\boldsymbol{x}) \geq 0$ when none of the descriptions detects the target character above the prescribed threshold. We note the inherent dependence of all methods on the classifier as it defines C; this detector may not align with human detection. Table 7 lists the descriptions used for each character.

D.4.2 Computational Comparison

Category	Property	Value
Base model	Model	SD v1.4
Shared	CFG scale	7.5
	Sampler	DDIM ($\eta = 1$)
SMC	Diffusion steps T	50
	Particles K	10
LPDM	Diffusion levels L	50
	Proximal updates P	10

Table 8: Hyperparameters for copyright-safe generation.

To ensure a fair comparison, we allocate each method the same maximum budget of 500 diffusion model evaluations. For a particle-based method with N particles and T reverse-diffusion steps, each particle requires one diffusion model evaluation per step, giving a total budget of TN for UA-SMC, FK-Steering and Best-of-N. For LPDM with L diffusion levels and P proximal updates per level, each update requires one diffusion model evaluation, giving a total budget of LP. Table 8 therefore presents our computational budget, $T = 50$ steps and $N = 10$ particles for the SMC methods, and $L = 50$ diffusion levels with $P = 10$ proximal updates per level for LPDM. We chose this high number of diffusion levels because under the $LP = 500$ budget, preliminary tuning showed that allocating computation across more diffusion levels performed better than concentrating more proximal updates at a smaller number of levels.