Inference-Time Reward Alignment of Diffusion Models via Geometric Particle Expansion

Anonymous authors
Paper under double-blind review

Abstract

Diffusion models have demonstrated strong generative performance in diverse domains, including images, molecular structures, and scientific design. Aligning pre-trained diffusion models with domain-specific objectives at inference time, without fine-tuning, has recently gained attention, particularly in settings involving non-differentiable rewards. Inspired by recent advances in Sequential Monte Carlo sampling, we propose a derivative-free alignment framework for diffusion model inference. We introduce a geometric expansion within Sequential Monte Carlo-based sampling that avoids the need for reward gradients for guidance during inference. Our approach generates additional particles via geometric noise derived from the covariance of elite particles along dominant eigendirections, followed by resampling at the same timestep to mitigate elite particle degeneracy. We demonstrate that our method consistently generates designs with improved performance and diversity when optimizing non-differentiable rewards across image, molecules, and DNA sequence generation tasks, while achieving comparable or superior reward scores for differentiable aesthetic score, without requiring reward-gradient computation.

1 Introduction

Diffusion models have emerged as a powerful class of generative models trained through approximations to the log-likelihood objective Ho et al. (2020); Song et al. (2020). Their ability to learn expressive data distributions has led to strong performance across diverse domains, including image synthesis Dhariwal & Nichol (2021), text-conditional generation Ramesh et al. (2021), and biomolecular sequence design Jing et al. (2022); Avdeyev et al. (2023). In practice, these models are often first pre-trained on large-scale datasets to capture a broad data prior and then adapted toward downstream objectives Rombach et al. (2022b). This adaptation step is especially important in scientific and engineering applications, where the goal is not only to generate realistic samples but also to discover samples that satisfy task-specific performance criteria.

A common approach to adaptation is fine-tuning, in which model parameters are updated using task-specific objectives Fan et al. (2023); Zhao et al. (2025). While effective, fine-tuning can be computationally expensive, may require repeated training for each new objective, and can distort the learned generative prior. Furthermore, fine-tuning methods based on reinforcement learning are known to be vulnerable to reward over-optimization and mode collapse Black et al. (2024); Uehara et al. (2024b). An alternative is inference-time alignment, which keeps the pre-trained model fixed and instead biases the sampling process toward high-reward outputs Kim et al. (2025); Ho & Salimans (2022). This approach is well-suited to scientific discovery, where retraining large generative models for every new reward function is often challenging and preserving the naturalness and modes of the pre-trained distribution is essential Uehara et al. (2025); Li et al. (2024a).

Despite its promise, inference-time alignment faces a central challenge: reward maximization must be balanced against distributional fidelity. Overly aggressive reward optimization can lead to mode collapse, loss of diversity, or excessive deviation from the pre-trained distribution Del Moral et al. (2006); Uehara et al. (2024b). Several approaches have been proposed for inference-time alignment in diffusion models Dai et al. (2025); Trippe et al. (2023). These methods modify the reverse diffusion trajectory using reward gradients or classifier-style guidance and therefore require differentiable rewards or surrogate models Kim et al. (2025);

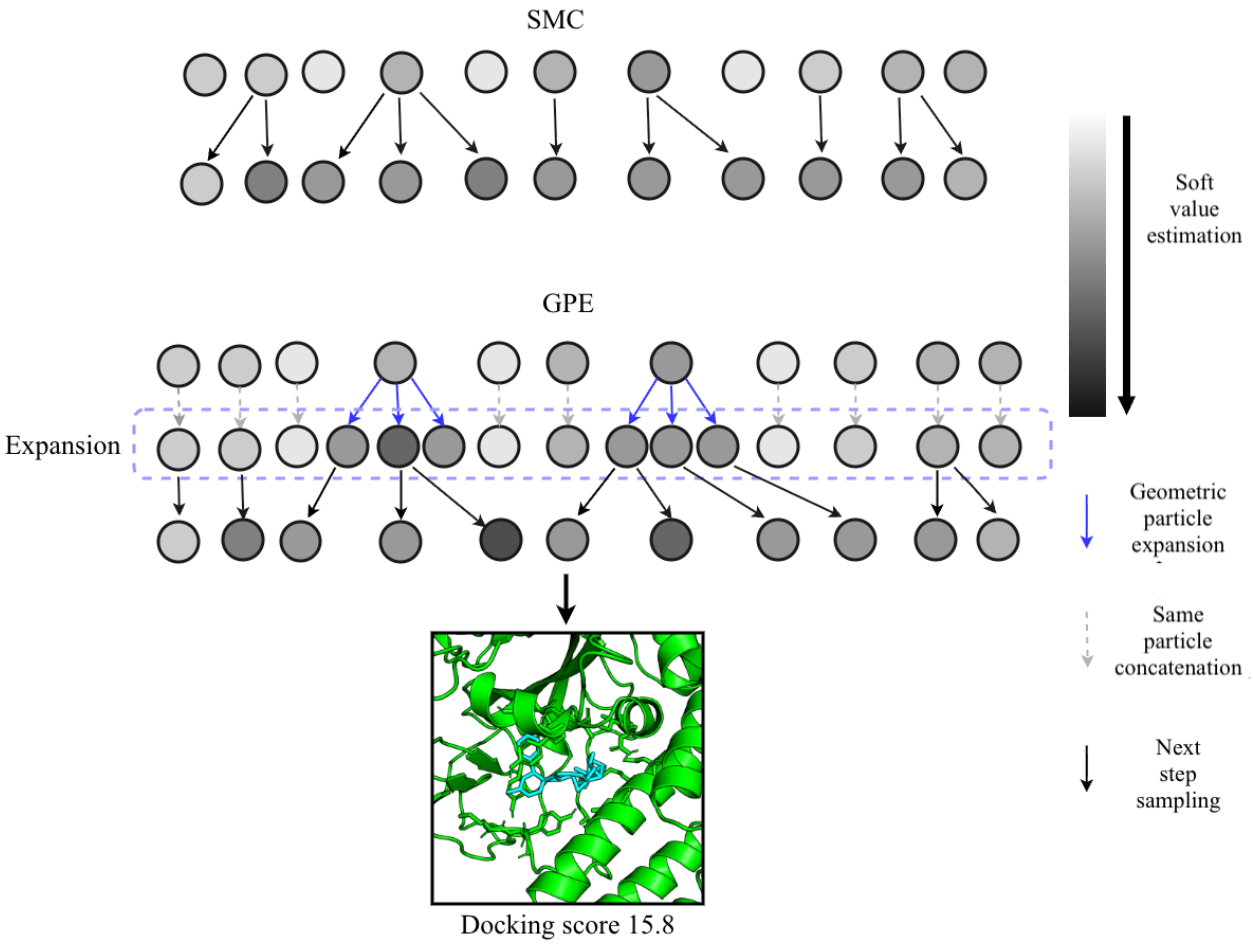


Figure 1: Illustration of a single inference-time reward-alignment step for SMC and GPE. The top panel shows vanilla Sequential Monte Carlo (SMC), where black arrows indicate one step of particle propagation. The bottom panel shows GPE, where elite particles are expanded using geometric noise before resampling. Blue arrows indicate the generation of child particles in the neighborhood of elite particles, while gray arrows indicate the transfer of particles to be concatenated with the expanded particles before the next step. Darker particles correspond to higher-reward samples and lighter particles correspond to lower-reward samples.

Dhariwal & Nichol (2021). While such assumptions are often reasonable in image synthesis and natural language processing Clark et al. (2023), many real-world scientific objectives do not provide readily available gradients Salomon-Ferrer et al. (2013). Examples include physical feasibility constraints and simulation-based performance metrics in scientific discovery Alamdari et al. (2023).

Another family relies on value functions or reward-weighted sampling to bias denoising towards high-reward candidates, often depending on the value approximation or batch-local candidate selection at each step Li et al. (2024b). Sequential Monte Carlo (SMC)-based methods provide population-level selection that combines importance weighting with global resampling, naturally preserving fidelity to a pre-trained generative model Dou & Song (2024). However, SMC is susceptible to weight degeneracy, where resampling repeatedly concentrates mass on a small subset of high-weight particles Yoon et al. (2026). This premature concentration eliminates particles with moderate to high reward along the diffusion trajectory, limiting exploration and reducing optimization effectiveness Uehara et al. (2025).

A key limitation of standard SMC in reward-guided diffusion sampling is that resampling is purely selective: it can preserve high-weight particles, but it cannot generate additional candidates in their neighborhoods. As a result, promising regions may collapse prematurely to a small number of duplicated particles, especially

under strong reward guidance or limited particle budgets. This motivates an inference-time mechanism that augments the candidate population before resampling, rather than only reweighting or selecting from the existing particles.

In this work, we address this challenge by introducing *Geometric Particle Expansion* (GPE), a derivative-free inference-time alignment method that augments SMC-guided diffusion sampling with a same-timestep particle expansion mechanism. Instead of relying on reward gradients or acceptance-rejection corrections, GPE expands the particle population around high-reward regions using the geometry of elite particles. At each reverse diffusion step, a subset of elite particles is selected according to the reward, and additional candidate particles are generated by injecting covariance-informed geometric noise aligned with the dominant eigendirections of the elite particle set. These expanded particles are then pooled with the original population and resampled at the same diffusion timestep when needed. This heuristic candidate augmentation induces exploration near promising regions while maintaining proximity to the pre-trained generative distribution. We provide theoretical guarantees showing that geometric particle expansion preserves asymptotic correctness. We demonstrate that GPE consistently improves optimization of non-differentiable rewards across a diverse range of applications. In summary, our main contributions are as follows.

1. We propose *Geometric Particle Expansion* (GPE), a same-timestep candidate augmentation method based on the dominant eigendirections of the elite particle set, followed by reward-weighted SMC-guided diffusion sampling for derivative-free inference-time reward alignment.
2. We provide a theoretical proof that the geometric particle expansion is asymptotically negligible in the resampling distribution and preserves asymptotic correctness.
3. We empirically validate GPE across non-differentiable reward optimization tasks, including controlled synthetic benchmarks, image generation, molecular optimization, biological sequence design, and further evaluate its behavior under differentiable reward settings.

2 Preliminary

2.1 Diffusion Models

Diffusion models learn to reverse a noising process Sohl-Dickstein et al. (2015); Song et al. (2020); Ho et al. (2020). Let $\mathcal{X}$ denote the sample space, which may be continuous, discrete, graph-structured, or latent, and let x_0 denote a sample drawn from the data distribution. The diffusion model introduces a sequence of latent variables $x_1, x_2, \ldots, x_T$, each representing a progressively noisier version of the data. The generative process is modeled as a latent-variable model:

$$p_\theta(x_{0:T}) = p(x_T) \prod_{t=1}^{T} p_\theta(x_{t-1} \mid x_t), \tag{1}$$

where $p(x_T) = \mathcal{N}(x_T; 0, I)$ is a simple Gaussian starting point and $p_\theta(x_{t-1} \mid x_t)$ denotes the learned reverse transitions. Since direct optimization of $\log p_\theta(x_0)$ is intractable, a forward noising process $q(x_{1:T} \mid x_0)$ is introduced as an approximate auxiliary posterior that gradually perturbs the data with Gaussian noise. The training objective is derived from a variational lower bound (ELBO) on the probability of the data Ho & Salimans (2022).

2.2 Inference-Time Alignment

To maximize the reward for downstream objectives at inference time in diffusion models, we formulate the problem as an entropy-regularized optimization objective Uehara et al. (2024a):

$$\begin{aligned} \{p_t^\star\}_t = \arg\max_{\{p_t\}_{t=1}^T} \Big\{ & \mathbb{E}_{\{p_t\}}[r(x_0)] \\ & - \alpha \sum_{t=1}^{T} \mathbb{E}_{\{p_t\}}[\mathrm{KL}(p_t(\cdot \mid x_t) \,\|\, p_t^{\mathrm{pre}}(\cdot \mid x_t))] \Big\}. \end{aligned} \tag{2}$$

The expectation $\mathbb{E}_{\{p_t\}}[\cdot]$ is taken under the reverse-time transitions $p(x_T)\prod_{t=1}^{T} p_t(x_{t-1} \mid x_t)$, so that $\mathbb{E}_{\{p_t\}}[r(x_0)]$ corresponds to the expected terminal reward of the generated sample. This objective defines the optimal reverse diffusion process $\{p_t^\star\}_{t=1}^{T}$ that maximizes the expected reward of the generated samples while remaining close to the pre-trained model through KL regularization. Here, p_t^{pre} denotes the pre-trained reference policy. By solving this objective via the soft optimal value function, the optimal reverse transition admits the form

$$p_t^\star(x_{t-1} \mid x_t) = \frac{1}{Z(x_t)} \exp\left(\frac{v_{t-1}(x_{t-1})}{\alpha}\right) p_t^{\text{pre}}(x_{t-1} \mid x_t), \tag{3}$$

where $Z(x_t)$ is a normalization constant.

2.3 SMC-Guided Diffusion Sampling

The goal is to sample from the optimal policy in equation 3 given the pretrained distribution. Since direct sampling is intractable, we approximate this target using an empirical distribution weighted by importance sampling within SMC. At each step, the SMC generates N samples $\{x_{t-1}^{[i]}\}_{i=1}^{N}$ from a proposal distribution $q_{t-1}\left(x_{t-1}^{[i]} \mid x_t^{[i]}\right)$, with associated unnormalized importance weights

$$w_{t-1}^{[i]} = \exp\left(\frac{v_{t-1}(x_{t-1}^{[i]})}{\alpha}\right) \cdot \frac{p_{t-1}^{\text{pre}}\left(x_{t-1}^{[i]} \mid x_t^{[i]}\right)}{q_{t-1}\left(x_{t-1}^{[i]} \mid x_t^{[i]}\right)}. \tag{4}$$

To control weight degeneracy, SMC applies resampling with replacement, producing an equally weighted empirical measure supported on Dirac masses. Let δ_x denote the Dirac probability measure at x:

$$\begin{aligned} \eta_{t-1}^{N} &= \frac{1}{N}\sum_{i=1}^{N} \delta_{\tilde{x}_{t-1}^{[i]}}, \\ \tilde{x}_{t-1}^{[i]} &:= x_{t-1}^{[\zeta_i]}, \qquad \zeta_i \sim \text{Cat}\left(\left\{\frac{w_{t-1}^{[j]}}{\sum_{k=1}^{N} w_{t-1}^{[k]}}\right\}_{j=1}^{N}\right). \end{aligned} \tag{5}$$

we consider policies from the pretrained distribution as a proposal distribution in SMC sampling.

Recent SMC-based approaches have explored reward alignment for diffusion models by modifying either the sampling dynamics or the particle initialization strategy. Diffusion Alignment as Sampling (DAS) is based on tempering strategies and carefully designed proposal distributions, often requiring reward differentiation and Metropolis–Hastings acceptance corrections Kim et al. (2025). While effective, such proposal-based methods can increase sampling complexity and are not suitable for non-differentiable reward functions. Particle initialization has also been used to improve reward maximization in SMC-based diffusion sampling. Ψ-Sampler improves alignment by replacing standard Gaussian-prior initialization with reward-aware posterior initialization Yoon et al. (2026). Instead of initializing particles from the Gaussian prior, it samples initial particles from a reward-aware posterior using preconditioned Crank-Nicolson-Langlevin (pCNL) dynamics. This combines dimension-robust proposals with gradient-informed updates, improving early particle quality and alignment performance. However, because pCNL relies on gradient-informed posterior sampling, Ψ-Sampler still requires reward or surrogate gradients.

Although SMC methods are well suited to preserve fidelity to a pre-trained generative model through population-level importance weighting and global resampling, standard SMC does not directly provide an explicit exploration mechanism for reward maximization Uehara et al. (2025). In particular, increasing the sample size or relying solely on resampling does not necessarily lead to improved reward maximization or sustained diversity. Under strong reward signals, resampling tends to repeatedly select a small subset of high-weight particles, resulting in particle impoverishment and limited exploration. Therefore, existing SMC-based diffusion alignment methods require carefully designed proposal corrections, reward gradients,

or reward-aware posterior initialization, while vanilla SMC can still suffer from weight degeneracy during reward maximization. This motivates a derivative-free enhancement to SMC-based diffusion sampling, which we introduce in the next section.

3 Geometric Particle Expansion (GPE)

To enhance exploration in high-reward regions during SMC-based guidance, we introduce GPE as a candidate augmentation mechanism in the same-timestep. Figure 1 illustrates the schematic comparison between GPE and standard SMC. Additional candidate particles are generated near high-reward particles using covariance-informed geometric noise. At a given reverse timestep t, let the parent particle set be

$$\{x_t^{(i)}\}_{i=1}^N \subset \mathbb{R}^d, \qquad x_t^{(i)} \in \mathbb{R}^d,$$

with associated scalar rewards $\{r_t(x_t^{(i)})\}_{i=1}^N$. Let $k \leq N$ denote the fixed number of elite parents. Define the elite index set

$$\mathcal{E}_t \;=\; \mathrm{TopK}\Big(\{r_t(x_t^{(i)})\}_{i=1}^N,\, k\Big) \;\subset\; \{1,\dots,N\}. \tag{6}$$

Elite geometry. Let $x_t^{(i)} \in \mathbb{R}^d$. The empirical mean and covariance of the elite parents are given by

$$\begin{aligned} \mu_t &= \frac{1}{k}\sum_{i\in\mathcal{E}_t} x_t^{(i)}, \\ \widehat{\Sigma}_t &= \frac{1}{k-1}\sum_{i\in\mathcal{E}_t}(x_t^{(i)}-\mu_t)(x_t^{(i)}-\mu_t)^\top \;\in\; \mathbb{R}^{d\times d}. \end{aligned} \tag{7}$$

Intuitively, the covariance of elite particles provides a local approximation of the geometry of high-reward regions, enabling offspring generation along directions that are already supported by successful particles. Let the eigendecomposition be

$$\widehat{\Sigma}_t = U_t \Lambda_t U_t^\top, \qquad \Lambda_t = \mathrm{diag}(\lambda_{t,1},\dots,\lambda_{t,d}), \tag{8}$$

and define a numerically stabilized square root

$$\widehat{\Sigma}_t^{1/2} = U_t\,\mathrm{diag}\Big(\sqrt{\max(\lambda_{t,1},\varepsilon)},\dots,\sqrt{\max(\lambda_{t,d},\varepsilon)}\Big)\,U_t^\top. \tag{9}$$

Child generation. Let c be the number of children per elite parent, so that

$$M = k\,c, \qquad \{y_t^{(j)}\}_{j=1}^M \subset \mathbb{R}^d. \tag{10}$$

Each child $y_t^{(j)}$ is associated with an elite parent $x_t^{(p(j))}$, where $p(j) \in \mathcal{E}_t$ repeats each elite index c times. We draw independent noise variables

$$\eta_{t,\mathrm{geom}}^{(j)} \sim \mathcal{N}(0,\widehat{\Sigma}_t), \qquad \eta_{t,\mathrm{iso}}^{(j)} \sim \mathcal{N}(0,I_d), \tag{11}$$

and form the mixture

$$\eta_t^{(j)} = (1-\omega)\eta_{t,\mathrm{geom}}^{(j)} + \omega\eta_{t,\mathrm{iso}}^{(j)}, \qquad \omega \in [0,1]. \tag{12}$$

Here, $\eta_{t,\mathrm{geom}}^{(j)}$ and $\eta_{t,\mathrm{iso}}^{(j)}$ denote the geometric and isotropic noise components, respectively. In implementation, their relative contribution is controlled by the hyperparameter $\omega \in [0,1]$. Each child particle is then generated via a pCN-style correlated update

$$y_t^{(j)} = \rho_t\, x_t^{(p(j))} + \beta_t\,\eta_j, \qquad j = 1,\dots,M. \tag{13}$$

The pCN-style update preserves correlation with the parent particle and avoids overly aggressive perturbations that could move candidates far from the local diffusion manifold. The scalars $\beta_t, \rho_t \in \mathbb{R}$ are defined as follows, where T denotes the total number of reverse steps.

$$\beta_t = \beta_0 \sqrt{\frac{t+1}{T}}, \qquad \rho_t = \sqrt{1-\beta_t^2}, \tag{14}$$

Since η_j^{geom} and η_j^{iso} are independent Gaussians,

$$\eta_j \sim \mathcal{N}(0, \Sigma_{t,\text{eff}}), \qquad \Sigma_{t,\text{eff}} = (1-\omega)^2 \widehat{\Sigma}_t + \omega^2 I. \tag{15}$$

Consequently, the conditional distribution of each child satisfies

$$y_t^{(j)} \mid x_t^{(p(j))} \sim \mathcal{N}\Big(\rho_t x_t^{(p(j))},\ \beta_t^2 \Sigma_{t,\text{eff}}\Big). \tag{16}$$

The purpose of geometric expansion is not to alter the asymptotic target distribution. Rather, its effect is purely finite-sample. Under finite particle budgets, standard SMC resampling can prematurely eliminate particles located near promising reward regions before they accumulate sufficient importance weight. By generating local offspring around elite particles prior to resampling, GPE increases the probability that high-reward neighborhoods remain represented in the empirical measure. Thus, GPE acts as a finite-sample exploration mechanism while preserving asymptotic correctness. Unlike proposal-based alignment methods, GPE does not modify the reverse diffusion dynamics, the proposal distribution, or the target sequence. Instead, it augments the finite particle population available for resampling at a fixed timestep.

4 Global Resampling

At each guided step, we pool *parents* $\{x_t^{(i)}\}_{i=1}^N \subset \mathbb{R}^d$ and *children* $\{y_t^{(j)}\}_{j=1}^M \subset \mathbb{R}^d$. For $N \geq k$, we select the top-k parents and generate c children per parent, so that $M = kc$ and M is independent of N. Define the weight and the total masses

$$S_N := \sum_{i=1}^N w(x_t^{(i)}), \qquad C_M := \sum_{j=1}^M w(y_t^{(j)}). \tag{17}$$

Define the parent-only weighted empirical measure

$$\eta_N^{\text{par}} := \sum_{i=1}^N \frac{w(x_t^{(i)})}{S_N} \delta_{x_t^{(i)}}, \tag{18}$$

and the pooled parent–child weighted empirical measure

$$\eta_N^{\text{all}} := \sum_{i=1}^N \frac{w(x_t^{(i)})}{S_N + C_M} \delta_{x_t^{(i)}} + \sum_{j=1}^M \frac{w(y_t^{(j)})}{S_N + C_M} \delta_{y_t^{(j)}}. \tag{19}$$

Here, η_N^{all} is a discrete probability measure supported on the finite set $\{x_t^{(1)}, \ldots, x_t^{(N)}, y_t^{(1)}, \ldots, y_t^{(M)}\}$, assigning mass proportional to the particle weights. Conditional on the current particle population, multinomial resampling produces independent draws from the discrete distribution η_N^{all}. Algorithm 1 summarizes the proposed procedure. Appendix A provides a theoretical justification showing that, when the number of augmented candidates is fixed, their total mass vanishes in the pooled resampling distribution.

5 Experiments

The primary objective of GPE is to optimize downstream rewards while preserving the naturalness of generated samples, without requiring fine-tuning. We first perform experiments on several toy multimodal

Algorithm 1 Geometric Particle Expansion in Sampling

Require: Pretrained reverse model p_θ, timestep schedule $\{t_s\}_{s=1}^{S}$, reward function R, particle count N, guidance cutoff t_{guide}, inverse temperature λ, ESS threshold ρ, local proposal parameters (K, M)

Initialize particles $\{x_{t_1}^{(i)}\}_{i=1}^{N} \sim \mathcal{N}(0, I)$
for $s = 1, \ldots, S$ **do**
  Set $(t, t') \leftarrow (t_s, t_{s+1})$, with $t' = 0$ at the final step
  Reverse diffusion step: sample $\tilde{x}_{t'}^{(i)} \sim p_\theta(\cdot \mid x_t^{(i)}; \eta)$ for all $i \in [N]$
  if $t' > t_{\text{guide}}$ **then**
    $x_{t'}^{(i)} \leftarrow \tilde{x}_{t'}^{(i)}$ for all i
    continue
  end if
  Reward step: compute $\hat{x}_0^{(i)} \leftarrow \hat{x}_0(\tilde{x}_{t'}^{(i)}, t')$ and $r^{(i)} \leftarrow R(\hat{x}_0^{(i)})$
  GPE: select the top-K particles by reward, generate M local children per selected particle, and form the candidate set $\mathcal{C}_{t'}$
  Evaluate rewards on all candidates in $\mathcal{C}_{t'}$ and set $w_j \propto \exp(\lambda r_j)$
  Compute ESS $\leftarrow 1/\sum_j w_j^2$
  if ESS $< \rho N$ **then**
    Resample $\{x_{t'}^{(i)}\}_{i=1}^{N}$ from $\mathcal{C}_{t'}$ according to $\{w_j\}$
  else
    Keep the propagated particles: $x_{t'}^{(i)} \leftarrow \tilde{x}_{t'}^{(i)}$
  end if
end for

Ensure: Final samples $\{x_0^{(i)}\}_{i=1}^{N}$

target distributions to show that GPE improves reward while maintaining proximity to the target distribution. In contrast, existing inference-time alignment and fine-tuning methods can struggle to preserve multimodal structure and remain close to the target distribution, even in simple controlled settings. Controlled synthesis benchmarks on toy multimodal datasets are built on top of DAS Kim et al. (2025). We evaluate GPE on image generation with both non-differentiable and differentiable rewards, and on molecular design and DNA sequence generation with non-differentiable rewards. For image generation, we use Stable Diffusion v1.5 Rombach et al. (2022a) and optimize aesthetic rewards, including the LAION aesthetic score Schuhmann & Beaumont (2022). For molecule generation, we use GDSS Jo et al. (2022), trained on ZINC-250k Irwin & Shoichet (2005), as the pretrained diffusion model with $T = 1000$ diffusion steps. We consider three downstream objectives: drug-likeness measured by QED, synthetic accessibility measured by SA, and docking score computed by QuickVina 2 Alhossary et al. (2015). QED and SA are computed using RDKit. All molecular reward functions are treated as non-differentiable black-box feedback. Following standard practice, we normalize SA as $(10 - \text{SA})/9$ and the docking score as $\max(-\text{DS}, 0)$, so that higher values correspond to better performance. Docking scores are evaluated with respect to Parp1 target protein. For biological sequences of DNA, we use the discrete diffusion model of Sahoo et al. (2024), trained on datasets from Gosai et al. (2023). As a downstream reward, we use the Enformer model Avsec et al. (2021) to predict enhancer activity in the HepG2 cell line. The molecule and DNA experiments are built on top of the SVDD codebase.

All experiments are conducted on a single NVIDIA A100 GPU with 40,GB of memory, except for DNA experiments, which are performed on an NVIDIA H100 GPU.

We compare GPE with soft Value-based Decoding in Diffusion models (SVDD), Sequential Monte Carlo (SMC) and best-of-N sampling. For differentiable rewards, we use Diffusion Alignment as Sampling (DAS) as a stronger inference-time alignment method than vanilla SMC. while for non-differentiable rewards, we use vanilla SMC. To ensure a fair comparison, all methods use the same number of inference steps and the same number of evaluations of the reward-function per step. Therefore, all methods are evaluated under the same total reward-query budget. For ablation studies, we use the following default GPE parameters unless otherwise specified: reward guidance scale $\lambda = 10$ ($1/\alpha = 10$), $\beta_0 = 0.15$, 4 children per elite parent, 10 elite

parents and batch size $N = 64$. For convenience, we use $\lambda = 1/\alpha$ for GPE, which makes the guidance scale more intuitive and allows integer-valued settings.

Figure 2 shows that GPE remains close to the target distribution compared to other methods, including reinforcement learning, SVDD, and DAS. This indicates that GPE preserves the structure of the sampled distribution while increasing rewards. Figure 3 further shows the performance of GPE compared to recent

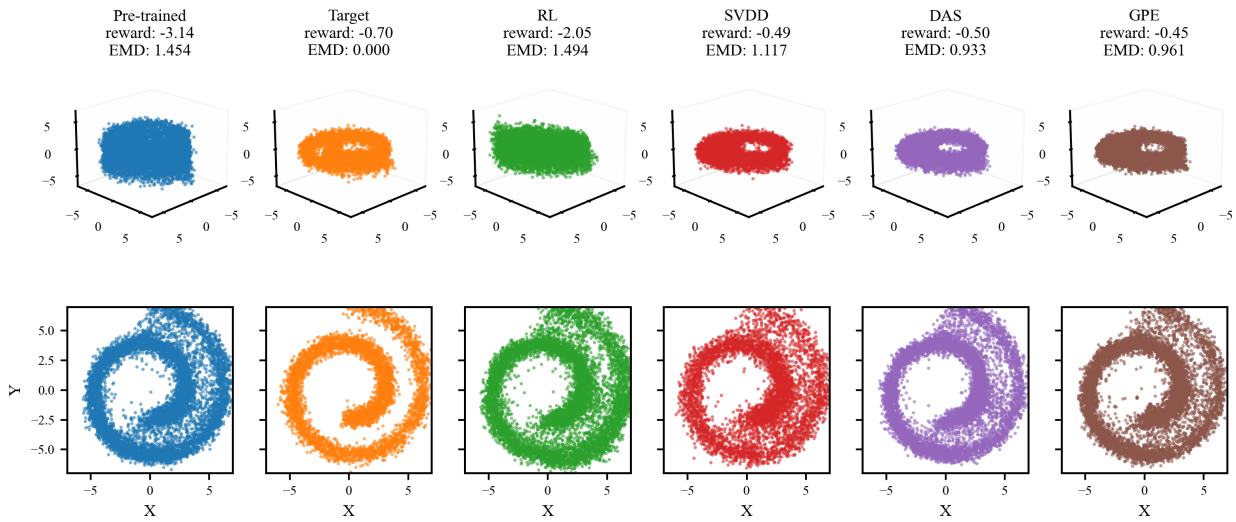


Figure 2: Comparison of reward optimization methods under a proximity constraint to the target distribution. GPE preserves proximity to the target distribution while improving reward. EMD denotes the empirical Earth Mover's Distance (Wasserstein distance) between the sample distribution of each method and the target distribution. Samples are drawn from a pre-trained model on the 3D Swiss roll, with a reward-aligned target distribution p_{tar} induced by $r(X, Y, Z) = -\frac{X^2}{100} - \frac{Y^2}{100} - Z^2$.

inference-time alignment methods for reward optimization. We observe that SMC-based inference methods, including DAS, are effective in approximating the target distribution, but high-reward particles can lose dominance and die out during the sampling process. In contrast, SVDD is effective in optimizing reward, but increasing α to push toward higher rewards can produce infeasible designs by reducing proximity to the pre-trained distribution. GPE addresses this issue by generating children in the neighborhood of elite particles, giving high-reward particles a sustained opportunity to influence the samples in subsequent steps.

Figure 4 provides an overview of representative high-reward samples generated by GPE in the image compressibility, image esthetics, and molecule generation tasks. All generated samples, except those for the aesthetic score task, outperform the corresponding samples from other methods. Table 1 compares the performance of GPE with other inference methods on the molecular generation task.

Table 1: Molecule generation alignment performance across methods in terms of reward and diversity. Function evaluations are capped at 104 per step for all methods. Higher QED, SA, and docking scores indicate better performance. Rewards are reported over the top 10% quantile of generated samples, while diversity is computed over the full sample set, with 95% confidence intervals.

Method	Molecule Drug-likeness		Molecule Synthetic Accessibility		Molecule Binding Affinity (Docking PARP1)	
	QED↑	Diversity↑	SA↑	Diversity↑	Docking Score↑	Diversity↑
Pre-trained	0.803 ± 0.007	0.8613 ± 0.1845	0.752 ± 0.005	0.8540 ± 0.2456	8.12 ± 0.5	0.8038 ± 0.2402
Best-N	0.907 ± 0.009	0.8978 ± 0.1893	0.942 ± 0.019	0.84742 ± 0.2104	10.9 ± 0.5	0.8138 ± 0.2238
SMC	0.882 ± 0.005	0.6272 ± 0.2486	0.803 ± 0.007	0.6984 ± 0.2268	10.1 ± 0.4	0.6317 ± 0.2947
SVDD	0.921 ± 0.005	0.8920 ± 0.1164	$\mathbf{1.000 \pm 0.006}$	0.8416 ± 0.2067	13.2 ± 0.3	0.8604 ± 0.1378
GPE	$\mathbf{0.947 \pm 0.003}$	0.8546 ± 0.2195	$\mathbf{1.000 \pm 0.005}$	0.8379 ± 0.2826	$\mathbf{15.2 \pm 0.2}$	0.8944 ± 0.2745

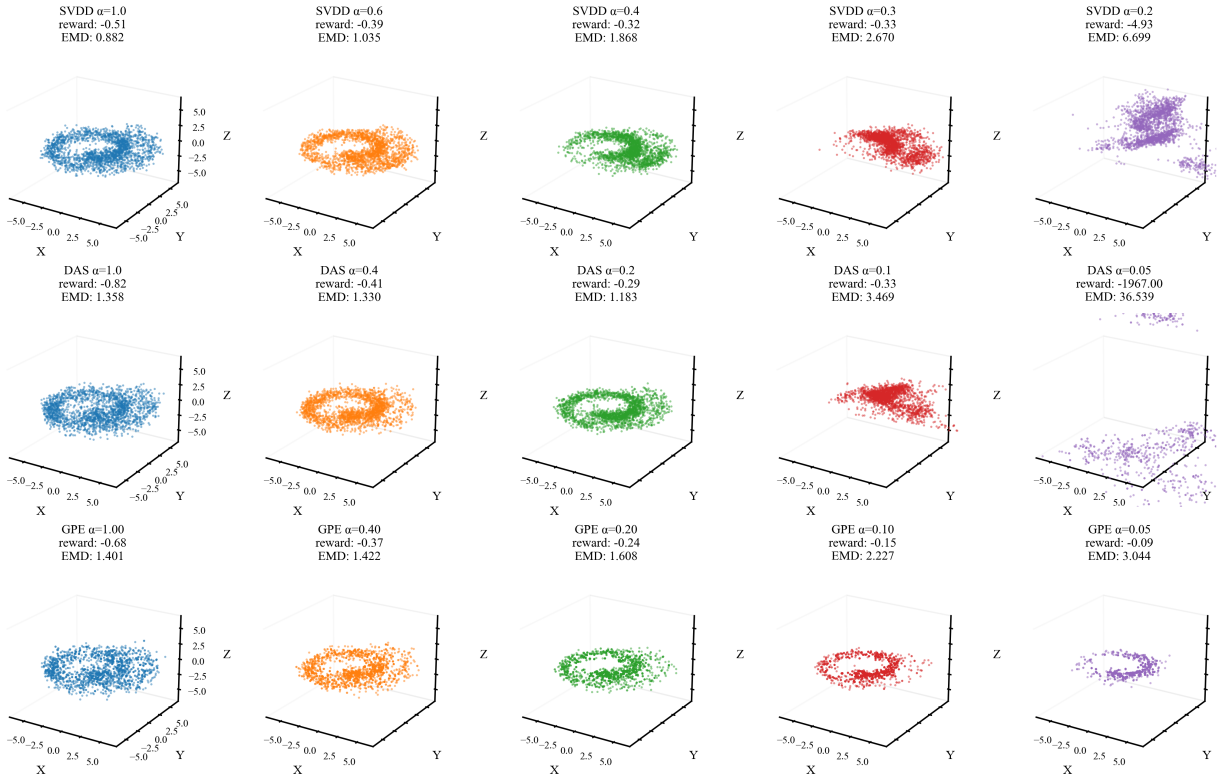


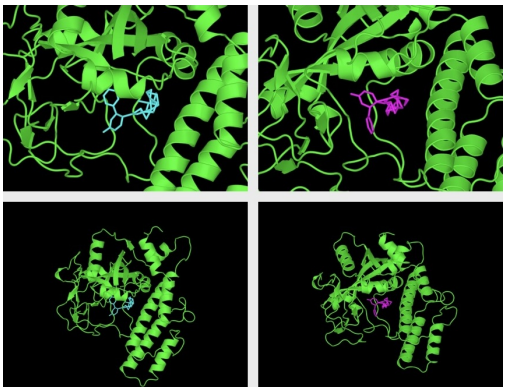
Figure 3: Comparison of inference-time alignment methods under different temperatures. Decreasing α beyond a certain point for SVDD and DAS results in samples diverging from the target distribution, while GPE samples stay close to the target distribution and GPE outperforms RL, SVDD, and DAS in reward maximization. Decreasing α in the GPE method yields higher rewards while remaining closer to the target distribution. Particles outside the 3D space are also shown to demonstrate that other sampling methods diverge from the pretrained model as α decreases, while GPE increases particle overlap and reduces diversity while improving reward, rather than producing invalid designs.

Figure 5 compares GPE with denoising diffusion policy optimization (DDPO) on the image compressibility task. GPE not only achieves stronger reward optimization but also preserves high image quality under high guidance strength and strong reward-maximization pressure. Examples of high quality images generated by incompressibility reward are shown in appendix B and Figure 18.

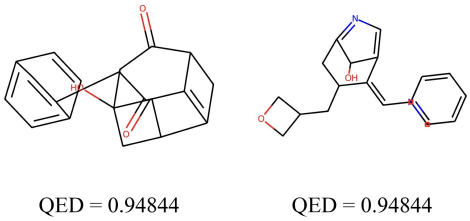
Figure 6 compares GPE with other methods in molecular generation and DNA sequence generation by reporting the reward distributions for drug-likeness, measured by QED, and enhancer activity in the HepG2 cell line. As noted earlier, molecule generation uses the GDSS diffusion model with $T = 1000$ diffusion steps, while DNA enhancer generation uses a discrete diffusion model with $T = 128$ diffusion steps.

6 Conclusion

We introduced *Geometric Particle Expansion* (GPE), a derivative-free inference-time alignment method for diffusion models. GPE augments SMC-guided sampling by expanding particles around high-reward elites using covariance-informed geometric noise. This allows the sampler to sustain exploration near promising regions without fine-tuning the pretrained model or requiring reward gradients. We showed that fixed-size particle expansion is asymptotically negligible in the pooled resampling distribution, preserving the asymptotic correctness of the parent SMC procedure. Empirically, GPE improves reward optimization while maintaining diversity across synthetic benchmarks, image generation, molecular generation, and DNA sequence generation. These results suggest that local geometric rejuvenation of elite particles provides a simple and broadly



(a) Molecule generated by GPE with a high PARP1 docking reward of 15.8.



(b) Molecules generated by GPE with high QED rewards.

(c) Image generated by GPE with a high aesthetic reward for the prompt "horse" (score: 6.96).

(d) Image generated by GPE under the compressibility objective (39 KB).

Figure 4: Representative high-reward samples generated by GPE across different applications, including molecular docking, molecular drug-likeness optimization, image aesthetic optimization, and image compressibility. In (a), the surfaces and ribbons represent the target protein, while the generated small molecule is shown at the center of the binding pocket.

applicable mechanism for reducing particle impoverishment in reward-guided diffusion sampling, particularly when rewards are black-box or non-differentiable.

7 Impact Statement

GPE enables inference-time optimization of pretrained diffusion models under black-box reward functions without fine-tuning or reward-gradient computation. This makes the method relevant to scientific discovery, engineering optimization, operational decision-support system, molecular design, and biological sequence generation, where objectives are often simulation-based, discrete, or non-differentiable. At the same time, reward functions should be treated as proxies rather than definitive measures of validity. High-scoring generated candidates, especially in the scientific and biomedical domains, require downstream domain-specific validation. The method also inherits the risks of reward misspecification: poorly designed rewards can

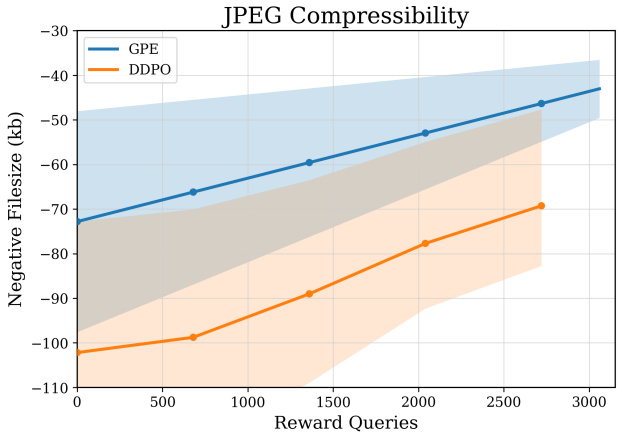
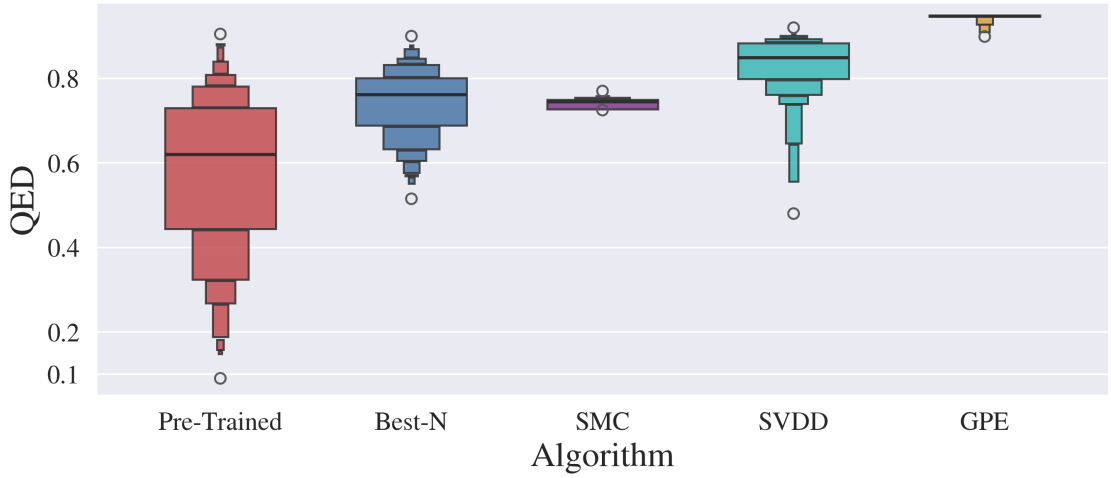
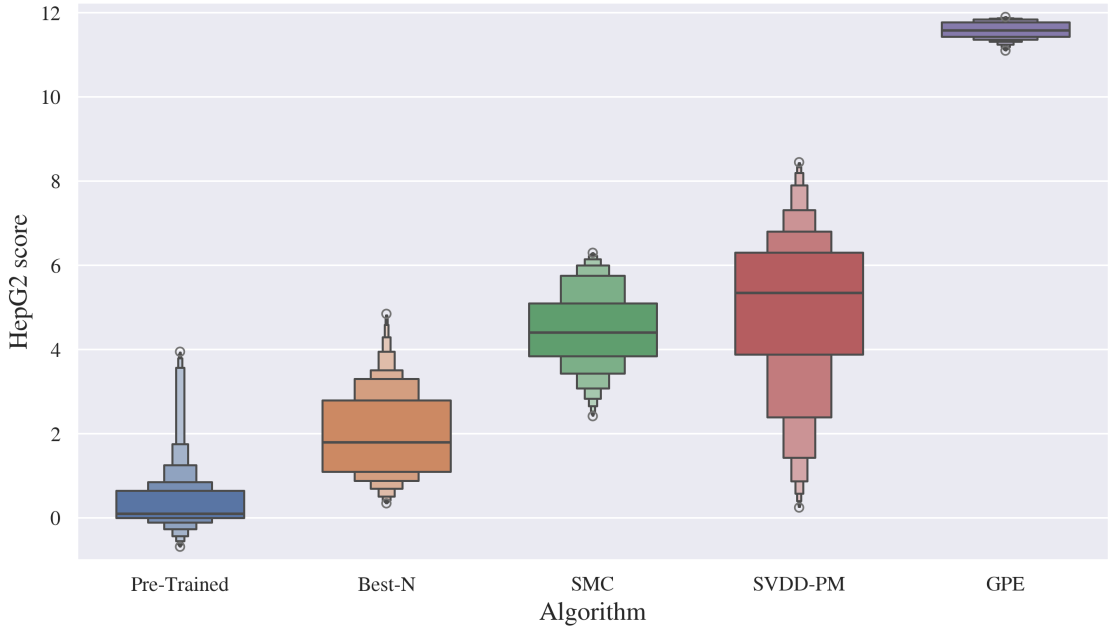


Figure 5: Comparison of GPE and DDPO on the image compressibility task. GPE achieves higher compressibility reward under the same reward-query budget.

encourage unrealistic, biased, or undesirable samples. Future work should study uncertainty-aware reward design, human-in-the-loop validation, and safeguards for high-stakes deployment.



(a) Reward distribution for molecular generation under QED reward guidance.



(b) Reward distribution for HepG2 of DNA Enhancers.

Figure 6: Comparison of reward distributions across methods for molecular QED optimization and HepG2 DNA enhancer activity. GPE consistently shifts the generated samples toward higher-reward regions, which demonstrates strong inference-time reward alignment across both domains. We use SVDD-PM, which relies on a posterior-mean approximation of the soft value function, because it outperformed the other SVDD variant.

References

Sarah Alamdari, Nitya Thakkar, Rianne Van Den Berg, Neil Tenenholtz, Robert Strome, Alan M Moses, Alex X Lu, Nicolò Fusi, Ava P Amini, and Kevin K Yang. Protein generation with evolutionary diffusion: sequence is all you need. *BioRxiv*, pp. 2023–09, 2023.

Amr Alhossary, Stephanus Daniel Handoko, Yuguang Mu, and Chee-Keong Kwoh. Fast, accurate, and reliable molecular docking with quickvina 2. *Bioinformatics*, 31(13):2214–2216, 2015.

Pavel Avdeyev, Chenlai Shi, Yuhao Tan, Kseniia Dudnyk, and Jian Zhou. Dirichlet diffusion score model for biological sequence generation. In *International Conference on Machine Learning*, pp. 1276–1301. PMLR, 2023.

Žiga Avsec, Vikram Agarwal, Daniel Visentin, Joseph R Ledsam, Agnieszka Grabska-Barwinska, Kyle R Taylor, Yannis Assael, John Jumper, Pushmeet Kohli, and David R Kelley. Effective gene expression prediction from sequence by integrating long-range interactions. *Nature methods*, 18(10):1196–1203, 2021.

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=YCWjhGrJFD`.

Kevin Clark, Paul Vicol, Kevin Swersky, and David J. Fleet. Directly fine-tuning diffusion models on differentiable rewards. *ArXiv*, abs/2309.17400, 2023. URL `https://api.semanticscholar.org/CorpusID:263310957`.

Longquan Dai, Xiaolu Wei, He Wang, Shaomeng Wang, and Jinhui Tang. Aligning text-to-image diffusion models to human preference by classification. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL `https://openreview.net/forum?id=YvypxK4kut`.

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, 2006.

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

Zehao Dou and Yang Song. Diffusion posterior sampling for linear inverse problem solving: A filtering perspective. In *The Twelfth International Conference on Learning Representations*, 2024.

Ying Fan, Olivia Watkins, Yuqing Du, Hao Liu, Moonkyung Ryu, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, Kangwook Lee, and Kimin Lee. Dpok: Reinforcement learning for fine-tuning text-to-image diffusion models. *Advances in Neural Information Processing Systems*, 36:79858–79885, 2023.

Sager J Gosai, Rodrigo I Castro, Natalia Fuentes, John C Butts, Susan Kales, Ramil R Noche, Kousuke Mouri, Pardis C Sabeti, Steven K Reilly, and Ryan Tewhey. Machine-guided design of synthetic cell type-specific cis-regulatory elements. *bioRxiv*, 2023.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance, 2022. URL `https://arxiv.org/abs/2207.12598`.

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

John J Irwin and Brian K Shoichet. Zinc- a free database of commercially available compounds for virtual screening. *Journal of chemical information and modeling*, 45(1):177–182, 2005.

Bowen Jing, Gabriele Corso, Jeffrey Chang, Regina Barzilay, and Tommi S. Jaakkola. Torsional diffusion for molecular conformer generation. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL `https://openreview.net/forum?id=w6fj2r62r_H`.

Jaehyeong Jo, Seul Lee, and Sung Ju Hwang. Score-based generative modeling of graphs via the system of stochastic differential equations. In *International conference on machine learning*, pp. 10362–10383. PMLR, 2022.

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`.

Xiner Li, Yulai Zhao, Chenyu Wang, Gabriele Scalia, Gökcen Eraslan, Surag Nair, Tommaso Biancalani, Shuiwang Ji, Aviv Regev, Sergey Levine, and Masatoshi Uehara. Derivative-free guidance in continuous and discrete diffusion models with soft value-based decoding. In *NeurIPS 2024 Workshop on AI for New Drug Modalities*, 2024a. URL `https://openreview.net/forum?id=GvgQIIBdfe`.

Xiner Li, Yulai Zhao, Chenyu Wang, Gabriele Scalia, Gokcen Eraslan, Surag Nair, Tommaso Biancalani, Shuiwang Ji, Aviv Regev, Sergey Levine, et al. Derivative-free guidance in continuous and discrete diffusion models with soft value-based decoding. *arXiv preprint arXiv:2408.08252*, 2024b.

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.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pp. 8821–8831. Pmlr, 2021.

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*, pp. 10684–10695, 2022a.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022b. URL `https://arxiv.org/abs/2112.10752`.

Subham S Sahoo, Marianne Arriola, Yair Schiff, Aaron Gokaslan, Edgar Marroquin, Justin T Chiu, Alexander Rush, and Volodymyr Kuleshov. Simple and effective masked diffusion language models. *Advances in Neural Information Processing Systems*, 37:130136–130184, 2024.

Romelia Salomon-Ferrer, David A Case, and Ross C Walker. An overview of the amber biomolecular simulation package. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 3(2):198–210, 2013.

Christoph Schuhmann and Romain Beaumont. Laion aesthetic predictor, 2022.

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.

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

Brian L. Trippe, Luhuan Wu, Christian A. Naesseth, David Blei, and John Patrick Cunningham. Practical and asymptotically exact conditional sampling in diffusion models. In *ICML 2023 Workshop on Structured Probabilistic Inference & Generative Modeling*, 2023. URL `https://openreview.net/forum?id=r9s3Gbxz7g`.

Masatoshi Uehara, Yulai Zhao, Kevin Black, Ehsan Hajiramezanali, Gabriele Scalia, Nathaniel Lee Diamant, Alex M Tseng, Tommaso Biancalani, and Sergey Levine. Fine-tuning of continuous-time diffusion models as entropy-regularized control. *arXiv preprint arXiv:2402.15194*, 2024a.

Masatoshi Uehara, Yulai Zhao, Kevin Black, Ehsan Hajiramezanali, Gabriele Scalia, Nathaniel Lee Diamant, Alex M Tseng, Sergey Levine, and Tommaso Biancalani. Feedback efficient online fine-tuning of diffusion models. *arXiv preprint arXiv:2402.16359*, 2024b.

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, 2025. URL https://arxiv.org/abs/2501.09685.

TaeHoon Yoon, Yunhong Min, Kyeongmin Yeo, and Minhyuk Sung. ψ-sampler: Initial particle sampling for SMC-based inference-time reward alignment in score models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=slVqJAI5sT.

Yulai Zhao, Masatoshi Uehara, Gabriele Scalia, Sunyuan Kung, Tommaso Biancalani, Sergey Levine, and Ehsan Hajiramezanali. Adding conditional control to diffusion models with reinforcement learning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=svp1EBA6hA.

A Proofs

This section demonstrates that adding M particles and resampling from the resulting empirical measure yields an asymptotically negligible deviation.

Theorem 1. *(Expanded candidate particles are asymptotically negligible in the resampling distribution.) For a fixed M,*

$$\left\|\eta_N^{\text{all}} - \eta_N^{\text{par}}\right\|_{\text{TV}} \xrightarrow[N\to\infty]{\mathbb{P}} 0, \qquad \left\|\eta_N^{\text{all}} - \eta_N^{\text{par}}\right\|_{\text{TV}} = O_{\mathbb{P}}(1/N). \tag{20}$$

The result follows directly from Lemma 1 and Lemma 2. It shows that while children may facilitate exploration in high reward proximity, their asymptotic proportion vanishes, and their contribution to the empirical measure is negligible.

Lemma 1 (Deterministic bound)**.** *For any bounded measurable φ with $\|\varphi\|_\infty := \sup_z |\varphi(z)| < \infty$,*

$$\left|\eta_N^{\text{all}}(\varphi) - \eta_N^{\text{par}}(\varphi)\right| \le 2\|\varphi\|_\infty \frac{C_M}{S_N}. \tag{21}$$

Proof. Following the standard normalized weighted empirical measure notation used in SMC Naesseth et al. (2019), we write

$$\eta_N^{\text{all}}(\varphi) = \frac{\sum_{i=1}^N w(x_t^{(i)})\varphi(x_t^{(i)}) + \sum_{j=1}^M w(y_t^{(j)})\varphi(y_t^{(j)})}{S_N + C_M}, \tag{22}$$

$$\eta_N^{\text{par}}(\varphi) = \frac{\sum_{i=1}^N w(x_t^{(i)})\varphi(x_t^{(i)})}{S_N}. \tag{23}$$

Subtracting equation 23 from equation 22 and rearranging yields

$$\begin{aligned}\eta_N^{\text{all}}(\varphi) - \eta_N^{\text{par}}(\varphi) = &\frac{\sum_{j=1}^M w(y_t^{(j)})\varphi(y_t^{(j)})}{S_N + C_M} \\ &- \frac{C_M}{S_N + C_M} \cdot \frac{\sum_{i=1}^N w(x_t^{(i)})\varphi(x_t^{(i)})}{S_N}.\end{aligned} \tag{24}$$

Using $|\varphi| \le \|\varphi\|_\infty$ gives

$$\left|\frac{\sum_{j=1}^M w(y_t^{(j)})\varphi(y_t^{(j)})}{S_N + C_M}\right| \le \|\varphi\|_\infty \frac{C_M}{S_N + C_M}, \tag{25}$$

and similarly

$$\left|\frac{C_M}{S_N + C_M} \cdot \frac{\sum_{i=1}^N w(x_t^{(i)})\varphi(x_t^{(i)})}{S_N}\right| \le \|\varphi\|_\infty \frac{C_M}{S_N + C_M}. \tag{26}$$

Therefore,

$$\left|\eta_N^{\text{all}}(\varphi) - \eta_N^{\text{par}}(\varphi)\right| \le 2\|\varphi\|_\infty \frac{C_M}{S_N + C_M} \le 2\|\varphi\|_\infty \frac{C_M}{S_N}. \tag{27}$$

Lemma 2 (Probabilistic bound)**.** *Under the following assumptions:*

(A1) *The parent particle weights have a finite positive mean:*

$$0 < \mu := \mathbb{E}[w(x)] < \infty. \tag{28}$$

(A2) *The children particle weights have uniformly bounded first moments, conditional on the parent particles, and there exists a constant $k < \infty$ such that, for all $j \le M$,*

$$\mathbb{E}\Big[w(y_t^{(j)}) \mid \{x_t^{(i)}\}_{i=1}^N\Big] \le k, \quad \textit{a.s., uniformly in } N. \tag{29}$$

Then, for fixed M,

$$\frac{C_M}{S_N} \xrightarrow[N\to\infty]{\mathbb{P}} 0, \qquad \frac{C_M}{S_N} = O_{\mathbb{P}}\left(\frac{1}{N}\right). \tag{30}$$

Proof. By the weak law of large numbers,

$$\frac{S_N}{N} = \frac{1}{N}\sum_{i=1}^{N} w(x_t^{(i)}) \xrightarrow{\mathbb{P}} \mu, \tag{31}$$

hence for any $\delta \in (0,1)$,

$$\mathbb{P}(S_N \geq (1-\delta)\mu N) \to 1. \tag{32}$$

Moreover, by conditional expectation and equation 29,

$$\mathbb{E}[C_M \mid \{x_t^{(i)}\}_{i=1}^N] = \sum_{j=1}^{M} \mathbb{E}[w(y_t^{(j)}) \mid \{x_t^{(i)}\}_{i=1}^N] \leq Mk, \tag{33}$$

Hence, for any $a > 0$,

$$\begin{aligned} \mathbb{P}(C_M \geq a) &= \mathbb{E}\Big[\mathbb{P}\Big(C_M \geq a \mid \{x_t^{(i)}\}_{i=1}^N\Big)\Big] \\ &\leq \mathbb{E}\left[\frac{\mathbb{E}\Big[C_M \mid \{x_t^{(i)}\}_{i=1}^N\Big]}{a}\right] \;\leq\; \frac{Mk}{a}. \end{aligned} \tag{34}$$

This shows that $C_M = O_{\mathbb{P}}(1)$. On the high probability event in equation 32, we have

$$\frac{C_M}{S_N} \leq \frac{C_M}{(1-\delta)\mu N} = O_{\mathbb{P}}(1/N), \tag{35}$$

which proves equation 30.

B Further Experiments

To evaluate GPE on multimodal target distributions, we compare inference across different numbers of modes and guidance scales against RL, SVDD, and DAS in Figures 7, 8, and 9. RL often concentrates samples around fewer modes, indicating that KL regularization alone is insufficient to preserve coverage of the full target distribution. Under strong guidance for reward maximization, SVDD and DAS also deviate from the target distribution. In contrast, GPE preserves proximity to the target distribution while generating high-reward particles.

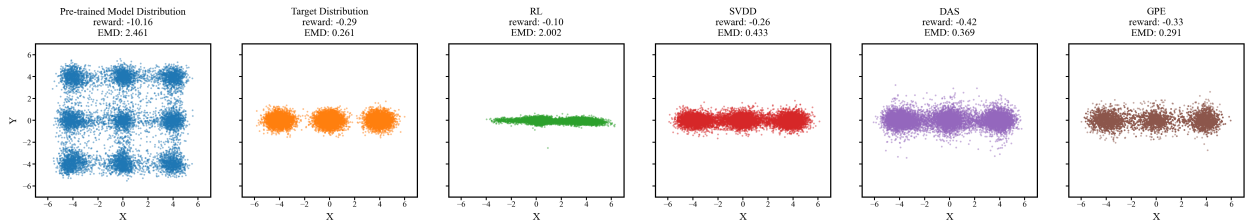


Figure 7: GPE outperforms all sampling and fine-tuning methods for reward optimization under a constraint of staying close to the target distribution for a mixture of Gaussians. GPE not only increases rewards effectively but also shows all modalities equally sampled, with modes completely separated similar to the target distribution, while SVDD, DAS, and RL produce particles that are more connected and have higher EMD. The reward is computed using $r(X,Y) = -X^2/100 - Y^2$, and EMD denotes the sample-based estimate of the Earth Mover's Distance. The target distribution shows nonzero EMD because the EMD is computed using a subset of particles (500 out of 10,000).

We further study the sensitivity of GPE to its main inference-time hyperparameters on the QED molecular generation task. Figure 10 reports ablations over the reward guidance scale λ, the child-expansion scale β_0, the number of children generated per elite particle, and the ESS resampling threshold. These parameters

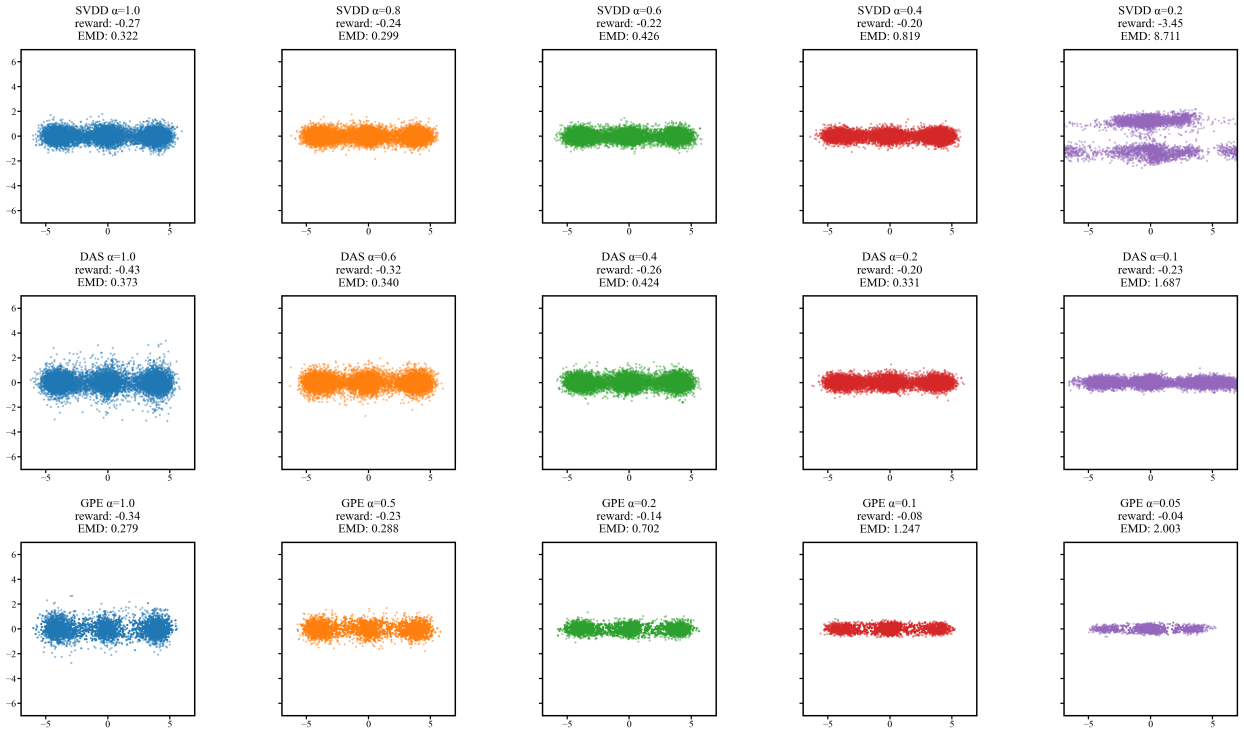


Figure 8: GPE excels at optimizing rewards while maintaining proximity to the target distribution. As α decreases, SVDD and DAS samples increasingly diverge from the target distribution and fail to increase the reward further, whereas GPE maintains proximity to the target distribution (low EMD) while continuing to drive samples toward high-reward regions. Lower α values for SVDD and DAS are omitted, as further decreases cause particles to leave the domain, yielding empty plots. Note that the GPE guidance parameter α is equivalent to $1/\lambda$.

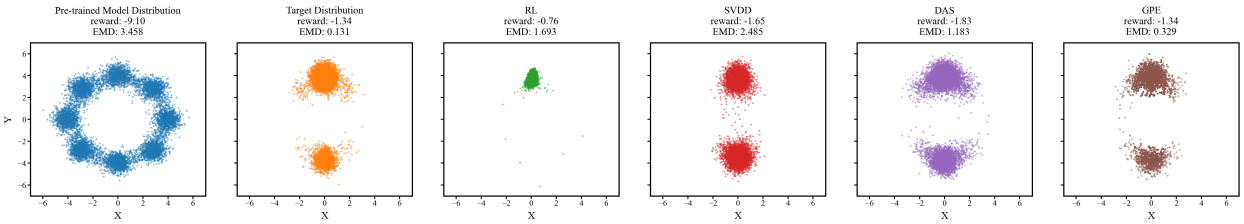


Figure 9: GPE excels at sampling all targeted modes while maintaining close proximity to the target distribution (low EMD) and increasing reward. The RL fine-tuned model achieves high rewards but collapses to a single mode. SVDD captures multiple modes but fails to represent the full distribution, leading to higher EMD. DAS improves reward over SVDD with better mode coverage, whereas GPE remains closest to the target distribution, even capturing the tails particles with higher reward than other sampling methods.

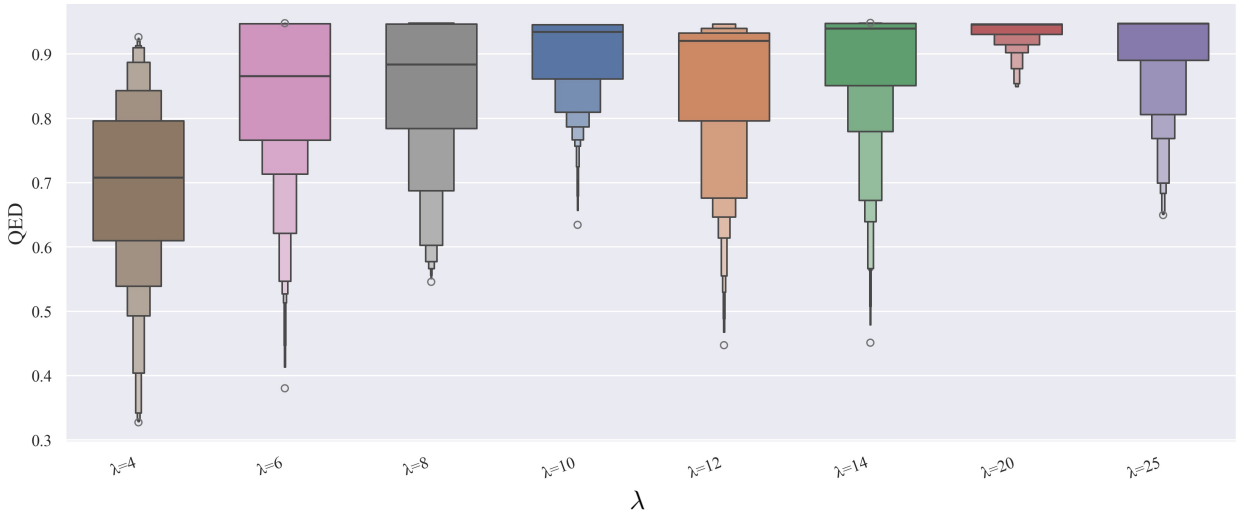
control the strength of reward alignment, the local exploration radius around elite particles, the amount of same-step particle expansion, and the frequency of resampling, respectively. Overall, the results show that increasing the guidance scale generally shifts the QED distribution toward higher values. This improvement is not strictly monotonic, but $\lambda = 10$ and above provide strong guidance that increases the maximum QED value as well as the mean and minimum rewards. We later analyze the effects of these hyperparameters in terms of the trade-off between reward maximization and diversity.

Diffusion decoding contains relatively sparse information in the early denoising steps and increasingly dense information as sampling approaches the final stages. Therefore, later denoising steps are more critical for

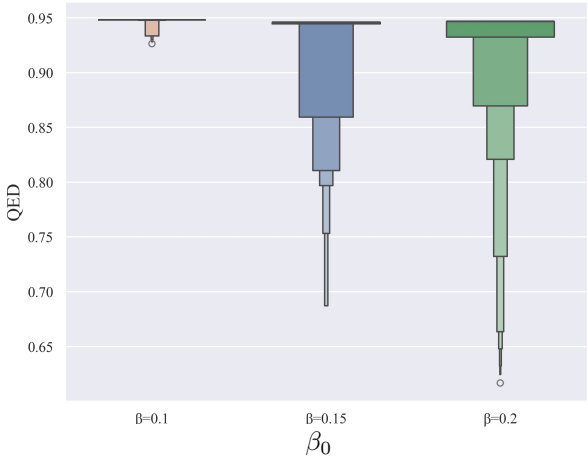
reward guidance. We evaluate this effect on GPE for the QED molecular generation task by varying the reward cutoff, which defines the denoising timestep at which reward-guided resampling begins. A smaller cutoff applies guidance earlier in the denoising trajectory, while a larger cutoff restricts guidance to later denoising steps. Figure 11 shows that later-stage guidance improves reward, but applying guidance too close to the final steps reduces performance because too few denoising steps remain for particle expansion to influence the samples effectively. We further evaluate the sensitivity of GPE on the HepG2 enhancer generation task by varying the reward guidance scale λ, the child-expansion scale β_0, and the ESS resampling threshold. The results are summarized in Figure 12. The reward-guidance scale is domain-specific. For the HepG2 optimization task, values up to $\lambda = 14$ improve the reward, whereas increasing λ beyond this point reduces the rewards of the generated samples.

We also evaluate how GPE hyperparameters affect diversity of samples. Figure 13 shows the variation in molecular diversity with respect to the reward guidance scale λ, the pCN child-expansion scale β_0, the number of children per elite particle, and the ESS resampling threshold. Together with Figure 10, these results illustrate the trade-off between reward optimization and molecular diversity. Figure 10a shows that increasing the guidance scale λ generally improves reward values, but it also tends to reduce diversity. A closer inspection of the toy experiments in Figures 3 and 8 makes this observation more intuitive and shows that, for GPE, increasing the guidance strength produces higher-reward samples while keeping them concentrated near the target distribution. As a result, the particles become more similar and tend to overlap in the high-reward region. In contrast, the other methods do not exhibit this overlap and instead diverge from the target distribution. The results suggest that $\lambda \in [8, 12]$ provides a favorable balance between reward maximization and diversity preservation. Nevertheless, the preferred configuration remains application-dependent, as different use cases may prioritize reward attainment and diversity preservation differently.

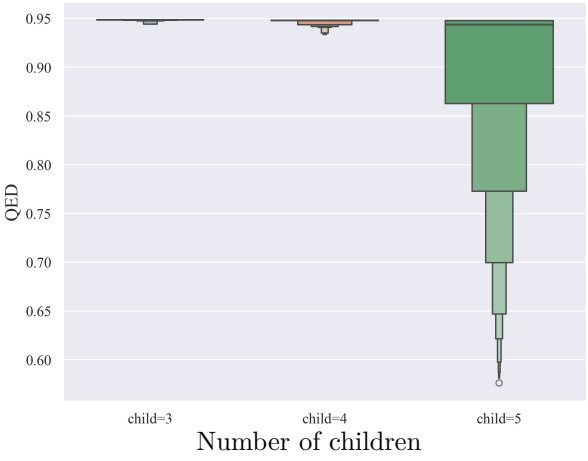
To show the impact of isotropic noise in GPE, we provide Figure 14, which compares two GPE configurations for aesthetic-reward guidance using the prompt *"snail"*. In both cases, the reward model, prompt, and sampling settings are kept fixed, and the only difference is the mixing coefficient ω used to generate child particles around high-reward parents. When $\omega = 0$, no isotropic noise is injected, so the noise is purely geometric and the particle evolution is determined by the covariance structure estimated from the current high-reward samples. As observed in the figure, while this can improve exploitation early in the guidance steps, it restricts exploration if the estimated covariance is narrow, leading to early concentration around a limited region. In contrast, using $\omega = 0.3$, chosen as an empirically effective parameter, introduces an isotropic component and therefore creates a mixture of geometric and full-space exploratory noise. This helps maintain diversity while still exploiting the local geometry of high-reward samples, resulting in more effective guidance.



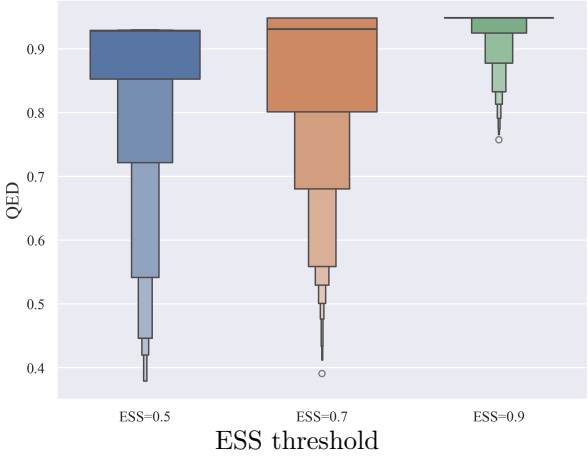
(a) Effect of reward guidance scale λ.



(b) Effect of child-expansion scale β_0.



(c) Effect of the number of children per elite particle.



(d) Effect of the ESS resampling threshold.

Figure 10: Hyperparameter sensitivity analysis of GPE on the QED molecular generation task. We vary the reward guidance scale λ, the child-expansion scale β_0, the number of children generated per elite particle, and the effective sample size (ESS) resampling threshold.

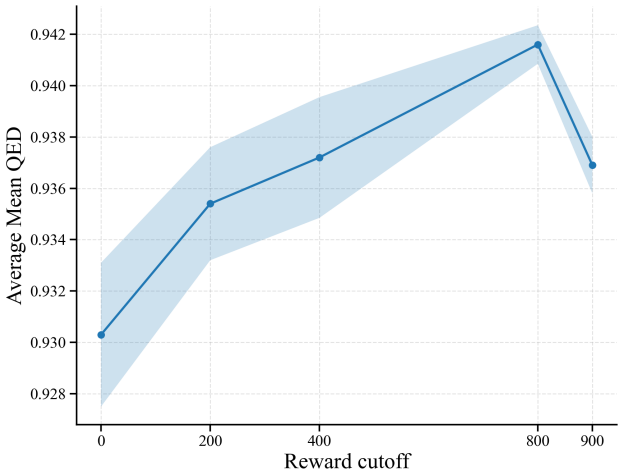
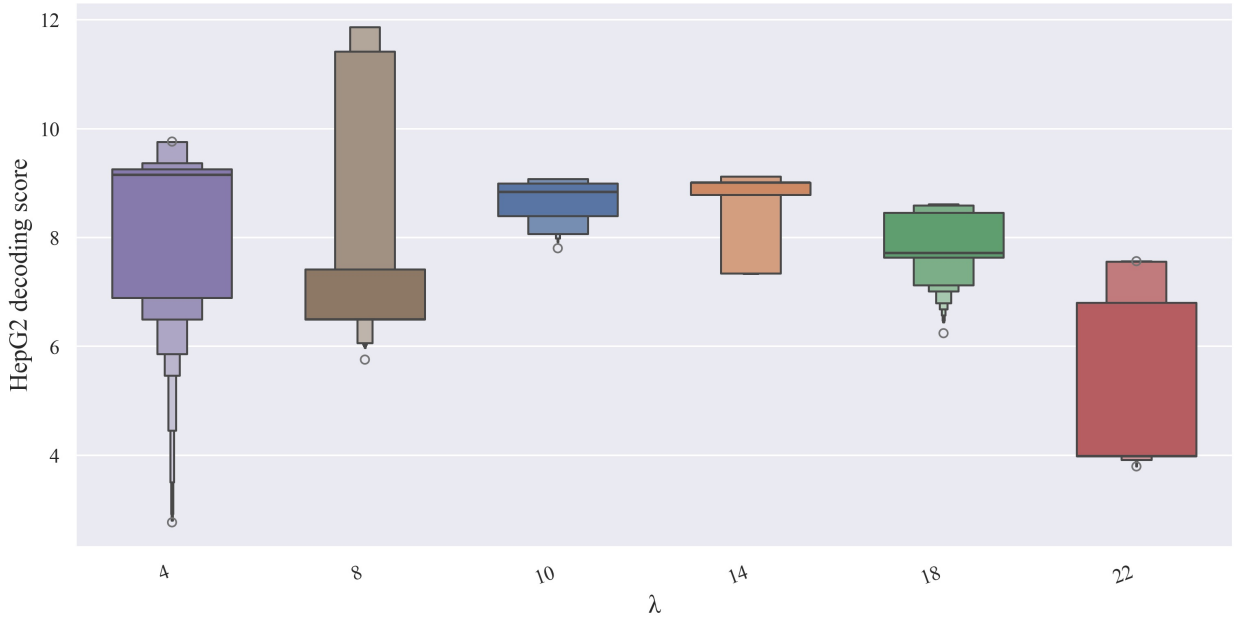
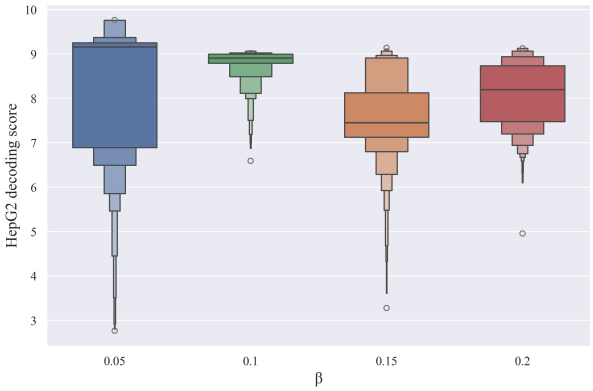


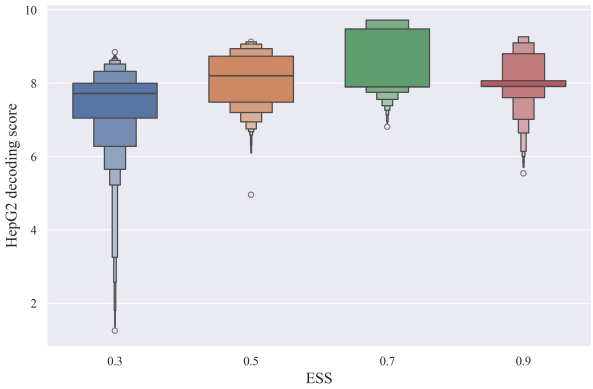
Figure 11: Effect of the reward cutoff on GPE performance for the QED molecular generation task. The reward cutoff determines when reward-guided resampling begins during denoising.



(a) Effect of reward guidance scale λ.

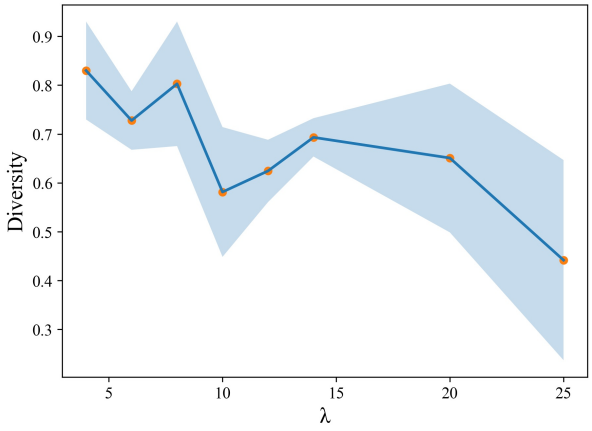


(b) Effect of child-expansion scale β_0.

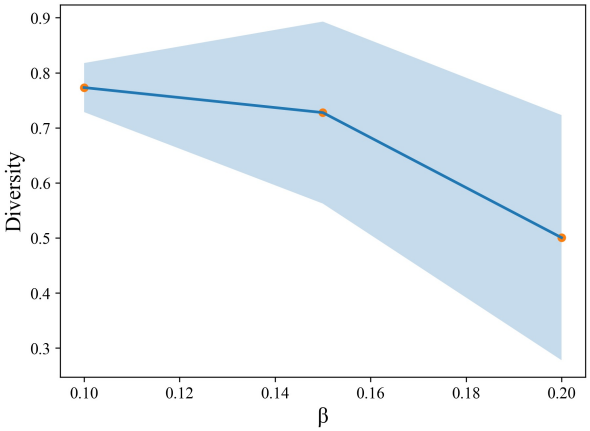


(c) Effect of ESS resampling threshold.

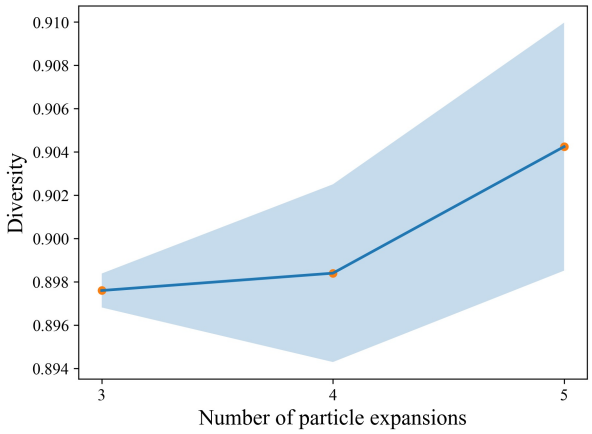
Figure 12: Histograms of samples generated by GPE for the HepG2 biological sequence generation task under different hyperparameter settings. We vary the reward-guidance scale λ, the child-expansion scale β_0, and the ESS resampling threshold ρ.



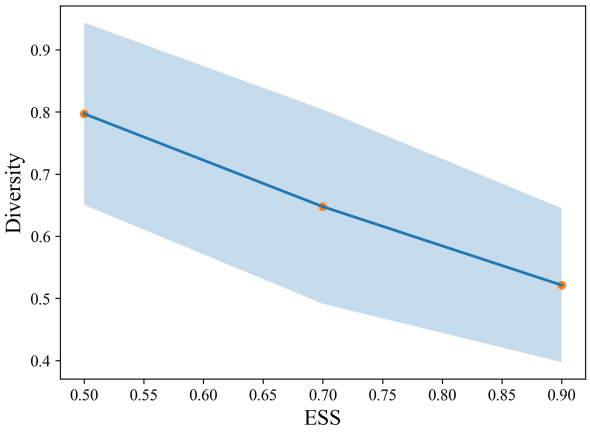
(a) Effect of reward guidance scale λ



(b) Effect of child-expansion scale β_0



(c) Effect of children per elite particle



(d) Effect of ESS resampling threshold

Figure 13: Ablation study of GPE on molecular diversity during reward-aligned molecule generation. We vary the reward guidance scale λ, the pCN child-expansion scale β_0, the number of children generated per elite particle, and the ESS resampling threshold ρ.

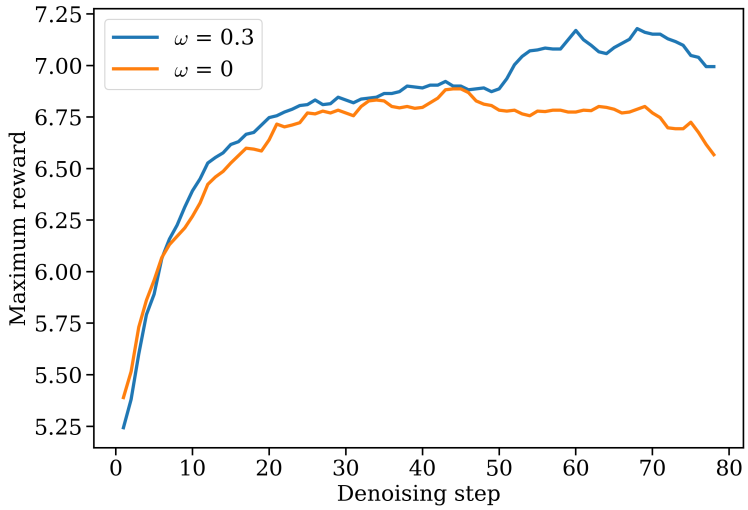


Figure 14: Effect of combining geometric and isotropic noise in GPE for image generation under an aesthetic reward with the prompt *"snail"*. The two cases differ only in the noise-mixing coefficient ω used when generating child particles around high-reward parents. For $\omega = 0$, the noise is purely geometric, while $\omega = 0.3$ introduces an empirically effective isotropic component. The isotropic component helps preserve exploration and prevents early-step concentration of particles while retaining the local geometric structure of high-reward samples.

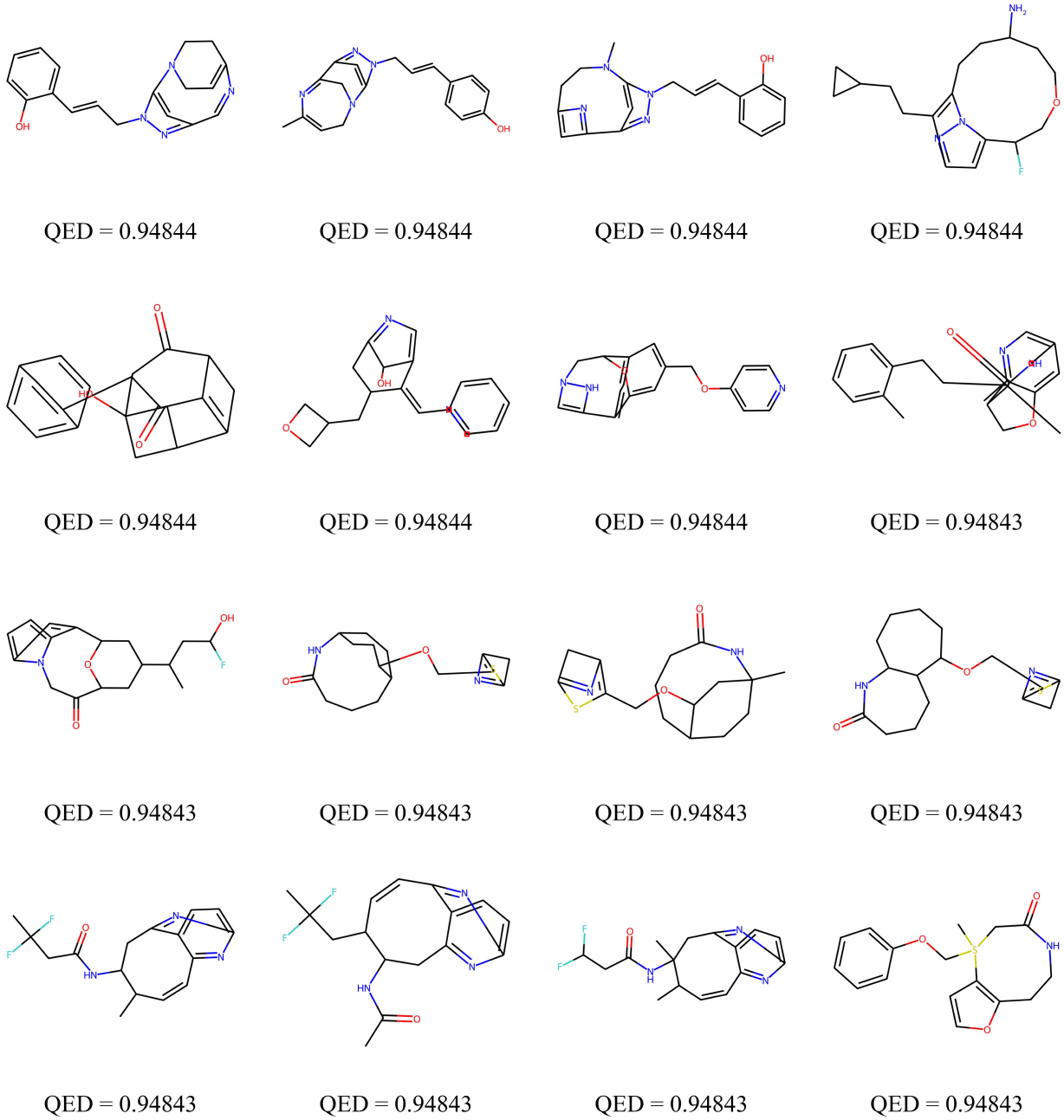


Figure 15: Examples of molecules generated by GPE with high reward QED

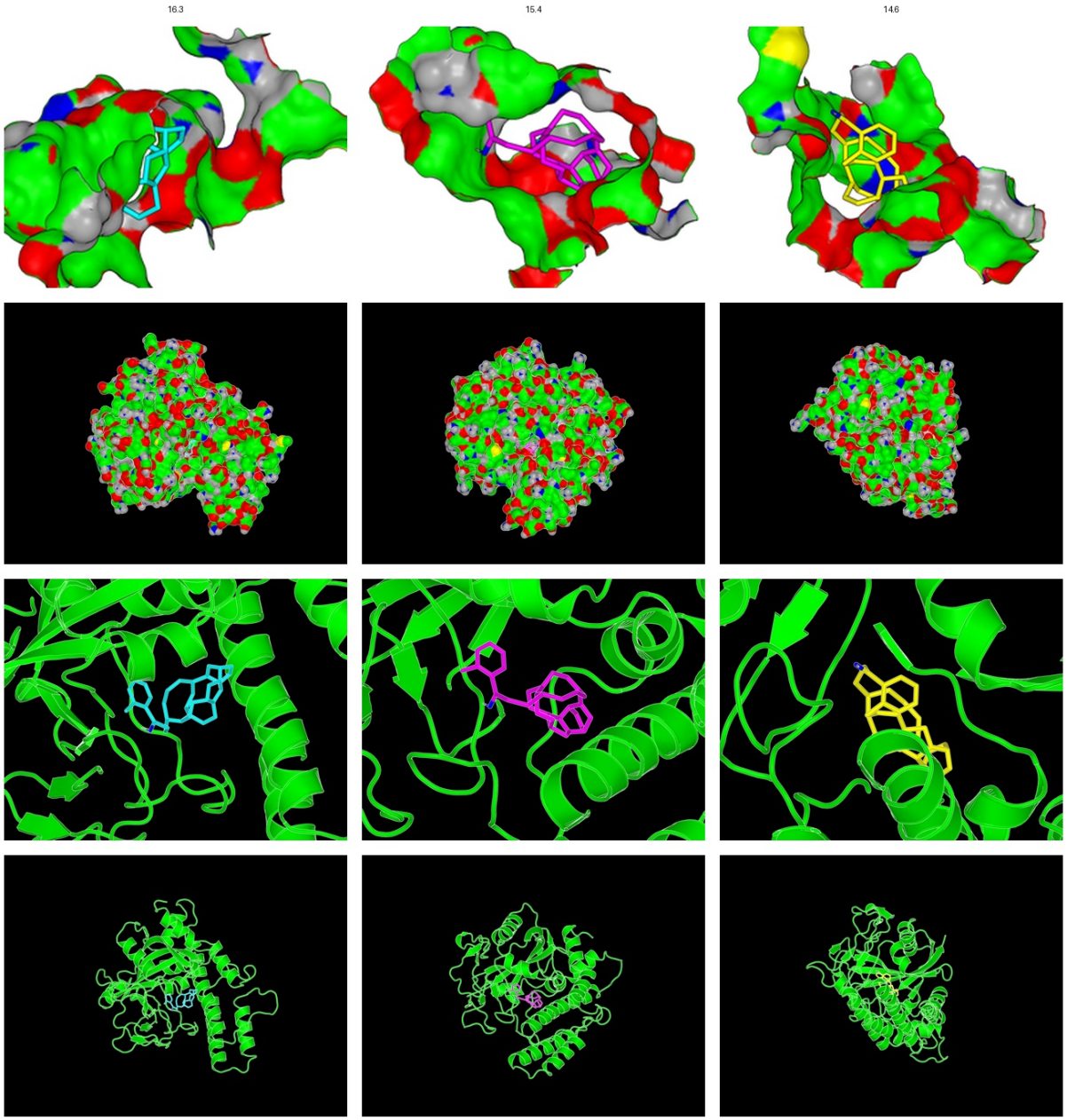


Figure 16: Examples of molecules generated by GPE with high reward in the PARP1 docking task, where the reward is defined as $\max(-\mathrm{DS}, 0)$.

Figure 17: Examples of generated images with high aesthetic scores. Higher scores indicate better aesthetic quality. The prompt and corresponding aesthetic score are shown below each image. All images were generated using at most 4875 aesthetic reward queries.

(a) Horse (43 KB)

(b) Cheetah (63 KB)

(c) Rabbit (39 KB)

(d) Panda (48 KB)

Figure 18: Examples of high-quality images achieving high JPEG compressibility. The prompt and resulting JPEG size are shown below each image; lower JPEG size corresponds to a higher compressibility reward.