POSTERIOR OVERLAP REDUCTION FOR DIVERSE TEXT-TO-IMAGE GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Text-to-image generative models can produce high-quality images yet often yield visually repetitive groups for a single prompt. Existing group diversification methods commonly compare the model's clean predictions directly or through visual embeddings. However, at an intermediate timestep, the model's conditional clean prediction estimates only the mean of a posterior over possible clean images. We argue that different posterior means can still correspond to distributions that share substantial probability mass. We call this shared mass posterior overlap and introduce Posterior Overlap Reduction (POR), a training-free method that reduces pairwise posterior overlap within a fixed group of noisy states. Under a condition on posterior variance, we derive an upper bound on overlap that decreases with symmetrized KL divergence. We obtain a model-based estimate of this divergence without evaluating posterior densities. Using this estimate, POR jointly updates the group at a single timestep early in sampling, then resumes the original sampler. Across Stable Diffusion 3 and FLUX.1-dev, POR shifts the quality–diversity frontier outward on COCO and achieves the highest Vendi Score and lowest mean pairwise similarity among the compared methods on COCO and DPG-Bench, while retaining competitive image quality and prompt alignment.

1 INTRODUCTION

Modern text-to-image generative models based on diffusion or flow matching have produced high-quality images that closely follow a text prompt (Ho et al., 2020; Rombach et al., 2022; Lipman et al., 2022; Esser et al., 2024). A single prompt can admit many plausible visual interpretations, allowing images to differ in viewpoint, layout, pose, or appearance while remaining faithful to the text (Parmar et al., 2026). In practice, however, strong guidance toward the text prompt can concentrate generation around a small number of dominant interpretations. As a result, a fixed-size group of independently generated samples often contains visually repetitive outputs (Sadat et al., 2024; Kim et al., 2025). Therefore, generating a fixed-size group that captures different visual interpretations of a prompt requires reducing visual repetition without sacrificing image quality or prompt alignment.

Existing group diversification methods commonly address this challenge through two main strategies. One strategy jointly adjusts a fixed group of candidate samples before or during sampling to encourage different visual outcomes (Corso et al., 2024; Morshed & Boddeti, 2025; Kim et al., 2025). A second strategy generates a larger pool of candidates in parallel and progressively selects a visually diverse subset during sampling (Parmar et al., 2026). Despite these procedural differences, both strategies commonly rely on the model's clean predictions as estimates of candidate visual content before sampling is complete. These methods compare clean predictions directly or through visual embeddings using distances (Parmar et al., 2026), similarities (Kim et al., 2025), or kernels (Corso et al., 2024; Morshed & Boddeti, 2025).

However, we argue that comparing clean predictions alone does not fully capture the potential for visual repetition between candidates. At an intermediate timestep, each candidate's noisy state is consistent with a range of possible clean images under the same prompt. The corresponding posterior assigns probabilities to these possible clean images. In contrast, the model's text-conditioned clean prediction estimates only the mean of the posterior (Efron, 2011; Lipman et al., 2022). In other words, different clean predictions can still correspond to posteriors that assign substantial probability mass to the same possible clean images. We call this shared probability mass *posterior overlap* and use it as a

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.

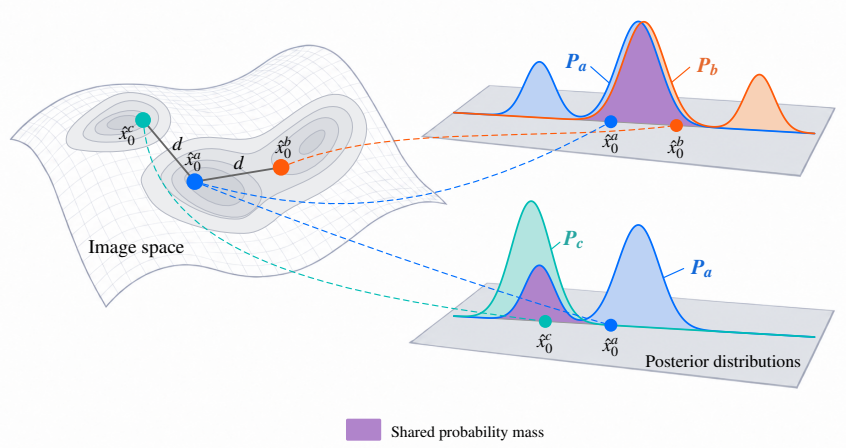


Figure 1: **A posterior view of potential visual repetition.** At a fixed timestep and prompt, the model's clean predictions estimate posterior means (left), while the corresponding posterior distributions describe possible clean images for each candidate (right). Gray contours depict a shared landscape of plausible clean images under the prompt. Different modes schematically represent different visual interpretations of the same prompt. Dashed lines connect each clean prediction to its corresponding posterior. Purple shading indicates posterior overlap, i.e., probability mass shared by the two posterior distributions. Although both pairs of clean predictions are separated by the same distance d, the overlap between P_a and P_b is greater than that between P_a and P_c. Posterior Overlap Reduction aims to reduce posterior overlap to lower the potential for visual repetition within the group.

distributional proxy for the potential for visual repetition between candidates. As Figure 1 shows, the same pairwise distance between clean predictions can correspond to substantially different amounts of posterior overlap. In this paper, we introduce Posterior Overlap Reduction (POR), a training-free method that produces less repetitive image groups by reducing pairwise posterior overlap within a fixed group of noisy states.

Since pretrained models do not provide the posterior densities needed to evaluate overlap, we introduce a tractable surrogate of the posterior overlap. When the posterior variance along the line connecting two noisy states remains comparable across the segment, we establish an upper bound on posterior overlap that decreases with the symmetrized KL divergence. At a fixed time and prompt, we can estimate symmetrized KL divergence using a scaled inner product (Karczewski et al., 2026) between noisy-state differences and differences in the model's conditional clean predictions. Based on this estimate, POR uses a group objective that favors larger estimated pairwise divergences to jointly update the intermediate states at a single timestep early in sampling before resuming the original sampler.

We demonstrate the effectiveness of the POR through two complementary evaluations. First, in a Gaussian-mixture setting where posterior overlap can be computed directly, our surrogate achieves a larger overlap reduction than a baseline based on posterior-mean comparisons. Second, we evaluate whether optimizing the posterior-overlap surrogate reduces visual repetition within generated groups on COCO (Lin et al., 2014) and the dense prompts of DPG-Bench (Hu et al., 2024). Across Stable Diffusion 3 (Esser et al., 2024) and FLUX.1-dev (Labs, 2024), POR shifts the quality–diversity frontier outward on COCO. On both benchmarks, POR achieves the highest Vendi Score and lowest mean pairwise similarity (MSS) among the compared methods, while retaining competitive image quality and prompt alignment. In summary, our contributions are:

- We propose posterior overlap, the probability mass shared by candidate posteriors, as a distributional proxy for the potential for visual repetition within a fixed group.
- We connect posterior overlap to symmetrized KL divergence to develop a computable surrogate for jointly updating a fixed group of noisy states at a single early timestep.
- We demonstrate that POR improves group diversity across Stable Diffusion 3 and FLUX.1-dev on COCO and DPG-Bench while retaining competitive image quality and prompt alignment.

2 Related Work

Inference-time Diversification. Training-free interventions can modify conditioning and guidance or favor low-density samples (Sadat et al., 2024; Kynkäänniemi et al., 2024; Um et al., 2024). Group-based approaches coordinate samples through repulsive interactions or set-level similarity criteria on particle states, clean predictions, or derived features (Corso et al., 2024; Kirchhof et al., 2024; Morshed & Boddeti, 2025). Contrastive Noise Optimization applies such coordination before sampling by optimizing initial noise through a contrastive loss on clean predictions (Kim et al., 2025). Inference-time search also seeks noise candidates that lead to higher-quality images (Ma et al., 2025), while Scaling Group Inference uses intermediate predictions to progressively select a group that balances quality and pairwise diversity (Parmar et al., 2026). A key distinction is that we use overlap between candidate posteriors as a distributional proxy for potential visual repetition, rather than relying on comparisons of clean predictions alone.

Denoising Posteriors and Information Geometry. Under Gaussian perturbation, posterior means are related to scores of the noisy marginal distributions through Tweedie's identity (Efron, 2011; Lai et al., 2025). Higher-order posterior moments have been used to quantify and visualize uncertainty about clean samples in denoising (Meng et al., 2021; Manor & Michaeli, 2024). Karczewski et al. (2026) use the exponential-family structure of denoising posteriors to study Fisher–Rao geometry and express symmetrized KL divergence in terms of natural and mean parameters. At a fixed time and prompt, the resulting scaled inner product between state differences and posterior-mean differences underlies our pairwise divergence estimate. Building on this structure, we connect symmetrized KL divergence to an upper bound on posterior overlap under a condition on posterior variance. This connection motivates a surrogate objective for group diversification that can be computed from noisy states and the model's clean predictions without evaluating posterior densities.

3 Posterior Overlap Reduction

Our goal is to reduce the potential for visual repetition within a fixed group of candidates. We first define posterior overlap as a distributional proxy for this potential, then derive a tractable surrogate for jointly updating the noisy states. Our Posterior Overlap Reduction (POR) performs these updates at a single timestep and then resumes generation with the original sampler.

3.1 Preliminaries

Intermediate states. Fix a text condition c and let $q_0(x_0 \mid c)$ denote the conditional data distribution, where $x_0 \in \mathbb{R}^d$ is an image or an autoencoder latent (Rombach et al., 2022; Esser et al., 2024). We consider Gaussian probability paths used in diffusion and flow-matching models (Ho et al., 2020; Lipman et al., 2022; Lai et al., 2025):

$$x_t = \alpha_t x_0 + \sigma_t \epsilon, \qquad \epsilon \sim \mathcal{N}(0, I). \tag{1}$$

Here ϵ is independent of x_0, and $q_t(x_t \mid x_0) = \mathcal{N}(x_t; \alpha_t x_0, \sigma_t^2 I)$ is the Gaussian perturbation kernel. We index time from data at $t = 0$ to noise at $t = T$, so generation proceeds toward decreasing t. For $T = 1$, the linear schedule $\alpha_t = 1 - t$ and $\sigma_t = t$ gives rectified-flow interpolation (Liu et al., 2022; Esser et al., 2024). For a group of N sampling trajectories, $x_t^{(i)}$ denotes the intermediate state of particle i, where $i \in \{1, \ldots, N\}$. We compare these states at the same time and under the same conditions, with $\alpha_t > 0$ and $\sigma_t > 0$.

Posteriors and clean predictions. The Gaussian forward path in Eq. (1) defines a posterior over clean samples $q_t(x_0 \mid x_t, c)$ for each noisy state x_t and prompt c (Lai et al., 2025). This posterior describes the clean samples compatible with the noisy state and prompt. We define its mean as follows:

$$\mu_t(x_t, c) := \mathbb{E}[X_0 \mid X_t = x_t, c]. \tag{2}$$

Let $\hat{x}_{0,c}(x_t, t)$ denote the model's clean prediction conditioned on c, obtained directly or by converting its noise, score, or velocity prediction. Under the corresponding squared-error training objectives, the population-optimal clean prediction is $\mu_t(x_t, c)$ (Lipman et al., 2022; Lai et al., 2025). For a pretrained model, we write

$$\hat{x}_{0,c}(x_t, t) \approx \mu_t(x_t, c). \tag{3}$$

The clean prediction therefore provides a mean summary of the posterior. However, standard comparisons of these means using distances, similarities, or kernels do not explicitly account for the probability mass shared by the corresponding posteriors.

3.2 Posterior Overlap as a Group Objective

We aim to reduce visual repetition when different candidates yield similar visual outcomes within a fixed-size group. We therefore compare pairs of candidate posteriors through their *shared probability mass* over possible clean samples. We call this shared mass posterior overlap and use it as a distributional proxy for the potential for visual repetition between candidates.

Group overlap objective. For the fixed group $\mathcal{X}_t = \{x_t^{(i)}\}_{i=1}^N$, with $N \geq 2$, we denote

$$q_i^t(x_0) := q_t(x_0 \mid x_t^{(i)}, c), \tag{4}$$

and collect the posteriors in $\mathcal{Q}_t = \{q_i^t\}_{i=1}^N$. To quantify the probability mass shared by a pair, we use the overlapping coefficient (Inman & Bradley Jr, 1989):

$$\Omega_{ij}^t := \int \min\{q_i^t(x_0), q_j^t(x_0)\}\, dx_0. \tag{5}$$

This quantity lies in $[0, 1]$, equals 1 for identical posteriors, and vanishes for disjoint supports.

Within the fixed group, we aggregate this proxy through the mean pairwise posterior overlap:

$$\mathcal{R}_t(\mathcal{Q}_t) := \frac{2}{N(N-1)} \sum_{i<j} \Omega_{ij}^t. \tag{6}$$

Aggregating shared mass over pairs captures pairwise redundancy in the group's posteriors over clean samples. We discuss its connection to duplicated group coverage in Appendix A.4. Our objective is to reduce $\mathcal{R}_t$ by jointly adjusting the noisy states while keeping the group size fixed.

Evaluating posterior overlap directly requires posterior densities and integration over the clean-sample space. Since pretrained generators do not provide these densities, we develop a tractable surrogate linked to posterior overlap in Section 3.3.

3.3 A Tractable Surrogate for Posterior Overlap

The theoretical connection between posterior overlap and symmetrized KL divergence motivates our surrogate. The symmetrized KL divergence has an exact expression in terms of noisy states and posterior means. First, we establish the connection between the overlap and symmetrized KL divergence. Then, we use this expression to obtain a computable estimate of symmetrized KL divergence.

Connecting overlap to symmetrized KL divergence. Let $B(p, q) := -\log \int \sqrt{p(x)q(x)}\, dx$ denote the Bhattacharyya distance (Bhattacharyya, 1943).

Lemma 1 (Overlap bound via the Bhattacharyya coefficient)**.** *The posterior overlap in Eq.* (5) *satisfies*

$$\Omega_{ij}^t \leq \int \sqrt{q_i^t(x_0) q_j^t(x_0)}\, dx_0 = \exp[-B(q_i^t, q_j^t)]. \tag{7}$$

The bound follows by integrating the pointwise inequality $\min\{p, q\} \leq \sqrt{pq}$. Equality holds when the two posteriors are identical.

We define symmetrized KL divergence as $D_{\mathrm{sKL}}(p, q) := \frac{1}{2}[D_{\mathrm{KL}}(p\|q) + D_{\mathrm{KL}}(q\|p)]$. To connect the Bhattacharyya distance and symmetrized KL divergence, we express both as differently weighted integrals of the same directional posterior variance along the segment between the two noisy states.

At the fixed time t and prompt c, let $\Sigma_t(x_t, c) := \mathrm{Cov}(X_0 \mid X_t = x_t, c)$ denote the posterior covariance of the clean samples. For each pair, we define the state difference and the line segment between the two noisy states:

$$\Delta x_{t,ij} := x_t^{(i)} - x_t^{(j)}, \qquad x_{t,ij}(s) := x_t^{(j)} + s\Delta x_{t,ij}, \quad s \in [0, 1]. \tag{8}$$

Here s indexes position along the segment between the two noisy states at the same timestep t. The scaled directional posterior variance along this segment is

$$\begin{aligned} v_{ij}(s) &:= \frac{\alpha_t^2}{\sigma_t^4} \Delta x_{t,ij}^\top \Sigma_t(x_{t,ij}(s), c) \Delta x_{t,ij} \\ &= \frac{\alpha_t^2}{\sigma_t^4} \operatorname{Var}\big(\Delta x_{t,ij}^\top X_0 \mid X_t = x_{t,ij}(s), c\big). \end{aligned} \tag{9}$$

This quantity measures the spread of possible clean samples in the direction of the state difference, scaled by the Gaussian perturbation.

Lemma 2 (Posterior-variance representations of pairwise divergences)**.** *For the posteriors induced by Gaussian perturbation at fixed t, c, with $\alpha_t, \sigma_t > 0$:*

$$D_{\mathrm{sKL}}(q_i^t, q_j^t) = \frac{1}{2} \int_0^1 v_{ij}(s)\, ds, \tag{10}$$

$$B(q_i^t, q_j^t) = \frac{1}{2} \int_0^1 \min\{s, 1-s\} v_{ij}(s)\, ds. \tag{11}$$

Appendix A.3 proves both identities in Lemma 2. In Eqs. (10) and (11), symmetrized KL divergence weights $v_{ij}(s)$ uniformly along the segment, whereas the Bhattacharyya distance uses weights that decrease toward the endpoints ($s = 0$ and 1). We derive a lower bound on the Bhattacharyya distance in terms of symmetrized KL divergence to bound posterior overlap. The following proposition establishes such a lower bound under a common bound on the relative variation of $v_{ij}(s)$.

Proposition 1 (Overlap bound via symmetrized KL divergence)**.** *Under the setting of Lemma 2, suppose that every pair of distinct posteriors in the state region under comparison satisfies*

$$0 < m_{ij} \le v_{ij}(s) \le M_{ij} < \infty \quad (0 \le s \le 1), \qquad \frac{m_{ij}}{M_{ij}} \ge \kappa_0, \tag{12}$$

for a common constant $\kappa_0 \in (0, 1]$. Then $B(q_i^t, q_j^t) \ge \kappa_0 D_{\mathrm{sKL}}(q_i^t, q_j^t)/4$, and

$$\Omega_{ij}^t \le \exp\Big[-\frac{\kappa_0}{4} D_{\mathrm{sKL}}(q_i^t, q_j^t)\Big]. \tag{13}$$

The ratio m_{ij}/M_{ij} describes how much $v_{ij}(s)$ is allowed to vary along the segment between the two noisy states. A larger admissible κ_0 restricts this relative variation more strongly and yields a tighter upper bound on overlap. Appendix A.3 proves Proposition 1 and discusses how limited variation in directional posterior variance at high noise can support a tighter overlap upper bound. This discussion motivates the early intervention described in Section 3.4.

Estimating symmetrized KL divergence from clean predictions. The exact state–mean identity in Appendix A.2, Eq. (28), expresses symmetrized KL divergence as a scaled inner product between the endpoint state difference and the posterior-mean difference. This identity holds even for multimodal posteriors and does not require the relative variance condition in Eq. (12).

Replacing the posterior means with the pretrained model's conditional clean predictions $\hat{x}_{0,c}^{(i)}(t) := \hat{x}_{0,c}(x_t^{(i)}, t)$ from Eq. (3) gives an estimate of pairwise symmetrized KL divergence, which we use as a pairwise score:

$$\widehat{D}_{\mathrm{sKL},ij}^t = \frac{\alpha_t}{2\sigma_t^2} \Delta x_{t,ij}^\top \big[\hat{x}_{0,c}^{(i)}(t) - \hat{x}_{0,c}^{(j)}(t)\big]. \tag{14}$$

By combining noisy-state differences with conditional clean predictions, this score estimates a divergence between candidate posteriors without evaluating their densities. We discuss why we use conditional rather than classifier-free-guided clean predictions to compute the pairwise score in Appendix B.7.

3.4 Fixed-Group Optimization

Group loss. For any fixed $\kappa_0 > 0$ valid throughout the comparison region, the exponential upper bound in Eq. (13) decreases monotonically with symmetrized KL divergence. The constant κ_0

controls the scale of the bound. Determining a valid value requires information about directional posterior variance along the state segments, which the endpoint means used to compute the pairwise score do not provide. We use this monotonic relationship to construct a surrogate that decreases with each pairwise score, without estimating κ_0 or assuming $\kappa_0 = 1$. For the group objective in Eq. (6), POR minimizes the following surrogate loss at a selected timestep $t^\star$:

$$\mathcal{L}_{\mathrm{POR}}^{t^\star} = \frac{2}{N(N-1)} \sum_{i<j} \mathrm{softplus}\Big(-\widehat{D}_{\mathrm{sKL},ij}^{t^\star}\Big). \tag{15}$$

Minimizing this loss favors larger pairwise scores. The softplus form specifies the surrogate's pair weighting, giving pairs with smaller pairwise scores greater gradient weight.

State optimization. With the generator parameters, group size, and timestep $t^\star$ fixed, we take K Adam steps with learning rate ρ to jointly update the current states. At each inner step, we refresh the conditional clean predictions and differentiate Eq. (15) through the predictor with respect to the noisy states. We then resume the original sampler from the optimized group, as summarized in Algorithm 1 in Appendix B.6. We restrict optimization to one timestep to limit the added computation (Appendix B.10). We choose $t^\star$ early in sampling, as Figure 4 shows: early intervention achieves greater diversity at comparable image preference. We require $\alpha_{t^\star} > 0$, since the posterior is independent of the state at the fully noisy endpoint.

4 Experiments

We first compare our surrogate with a baseline based on posterior-mean comparisons in a Gaussian-mixture setting where posterior overlap can be computed directly. With identical initial groups and matched update budgets, our surrogate achieves greater overlap reduction across the evaluated noise levels (Appendix A.6). We then test whether this posterior-based criterion improves within-prompt diversity while retaining competitive image quality and prompt alignment on COCO and DPG-Bench with SD3 and FLUX.1-dev.

4.1 Experimental Setup

Datasets. Our tabular comparisons use 2,000 prompts from the COCO validation set (Lin et al., 2014) and all 1,065 prompts from DPG-Bench (Hu et al., 2024). We use the dense DPG-Bench prompts, which specify multiple objects, attributes, and relationships, to evaluate within-prompt diversity, image quality, and alignment under detailed scene descriptions.

Models and baselines. We compare POR with I.I.D. sampling, Particle Guidance (PG), Contrastive Noise Optimization (CNO), and Scaling Group Inference (SGI) on SD3 and FLUX.1-dev. PG couples particles during sampling, CNO optimizes initial noise samples using clean predictions, and SGI selects a group from a larger candidate pool. Within each backbone, the tabular comparisons use the same prompts, output group size, image resolution, and base sampling configuration wherever applicable. We generate five images per prompt for SD3 and four for FLUX.1-dev. Appendix B provides the reference configurations and implementation details.

Evaluation metrics. We evaluate text–image alignment with CLIP and image preference with PickScore (Pick), ImageReward (IR), and HPSv2. Within each prompt, Vendi summarizes group diversity, mean pairwise similarity (MSS) captures repetition, and LPIPS measures perceptual variation. We report scores averaged over prompts and compare methods at a fixed group size within each backbone. Appendix B.2 provides definitions, references, and implementation details.

4.2 Main Results

COCO. Table 1 shows that POR improves group diversity over I.I.D. sampling on both backbones, achieving the highest Vendi and lowest MSS among the compared methods. On SD3, it also achieves the highest LPIPS, extending the gains to perceptual variation. On FLUX.1-dev, SGI achieves higher LPIPS, indicating that the relative ranking depends on the diversity aspect being measured. POR combines these diversity gains with competitive image preference and the highest CLIP on both backbones, although I.I.D. sampling retains the highest PickScore and HPSv2 on FLUX.1-dev.

Table 1: **Within-prompt diversity and image quality on COCO.** Fixed reference configurations are evaluated on 2,000 validation prompts.

		Quality / Alignment				Diversity		
Model	Method	CLIP↑	Pick↑	IR↑	HPSv2↑	Vendi↑	MSS↓	LPIPS↑
SD3 (M)	I.I.D.	26.386	22.475	1.008	29.018	4.176	0.315	0.569
	PG	26.395	**22.483**	**1.011**	29.047	4.178	0.315	0.569
	CNO	26.392	22.464	1.003	28.974	4.239	0.299	0.588
	SGI	26.393	22.458	1.005	28.879	4.254	0.295	0.604
	POR (Ours)	**26.440**	22.465	1.010	**29.049**	**4.286**	**0.284**	**0.616**
FLUX.1-dev	I.I.D.	26.095	**23.020**	1.096	**30.447**	3.377	0.347	0.541
	PG	26.104	22.982	1.080	30.401	3.389	0.344	0.546
	CNO	26.056	22.980	1.072	30.334	3.444	0.322	0.570
	SGI	25.899	22.909	1.015	30.100	3.460	0.314	**0.585**
	POR (Ours)	**26.178**	23.001	**1.098**	30.261	**3.472**	**0.311**	0.569

Table 2: **Within-prompt diversity and image quality on DPG-Bench.** Fixed reference configurations are evaluated on all 1,065 dense prompts.

		Quality / Alignment				Diversity		
Model	Method	CLIP↑	Pick↑	IR↑	HPSv2↑	Vendi↑	MSS↓	LPIPS↑
SD3 (M)	I.I.D.	29.730	21.900	0.844	28.254	3.846	0.394	0.543
	PG	29.680	21.895	0.848	28.254	3.889	0.384	0.546
	CNO	29.740	21.894	0.830	28.233	3.900	0.381	0.559
	SGI	29.678	21.874	0.840	28.203	3.957	0.368	0.565
	POR (Ours)	**29.758**	**21.910**	**0.849**	**28.262**	**3.966**	**0.363**	**0.589**
FLUX.1-dev	I.I.D.	29.444	**22.380**	0.926	29.573	3.074	0.445	0.501
	PG	29.366	22.356	0.926	**29.585**	3.072	0.445	0.502
	CNO	29.378	22.357	0.915	29.506	3.136	0.425	0.528
	SGI	29.382	22.370	0.908	29.435	3.161	0.416	**0.548**
	POR (Ours)	**29.475**	22.366	**0.928**	29.462	**3.178**	**0.412**	0.532

Diversity under dense prompts. We next examine whether these gains extend to the dense prompts in DPG-Bench. The I.I.D. groups in Table 2 show greater repetition than on COCO, with lower Vendi and LPIPS and higher MSS on both backbones. POR again achieves the highest Vendi and lowest MSS, and on SD3 it leads across all reported metrics. On FLUX.1-dev, it combines these diversity gains with the highest CLIP and ImageReward, although SGI achieves higher LPIPS. These results show that POR also reduces repetition under detailed scene descriptions.

Quality–diversity tradeoff. Beyond the fixed reference configurations, we sweep each method's main diversity control to assess diversity at comparable image preference. Figure 2 reports ImageReward–Vendi and PickScore–LPIPS curves, using different preference scores and measures of image variation. For both backbones, the sweeps use 400 randomly selected COCO prompts, with five images per prompt for SD3 and four for FLUX.1-dev. Appendix B.3 lists the sweep settings.

Across the evaluated sweeps, POR achieves greater diversity at comparable image preference on both backbones. It achieves higher Vendi than the I.I.D. guidance baseline, CNO, and SGI within overlapping ImageReward ranges. PG reaches slightly higher ImageReward on SD3 at lower diversity. On FLUX.1-dev, one POR setting exceeds all displayed baseline settings in both ImageReward and Vendi. The PickScore–LPIPS curves likewise show greater perceptual variation at comparable PickScore, with higher POR learning rates increasing LPIPS at the cost of PickScore on FLUX.1-dev.

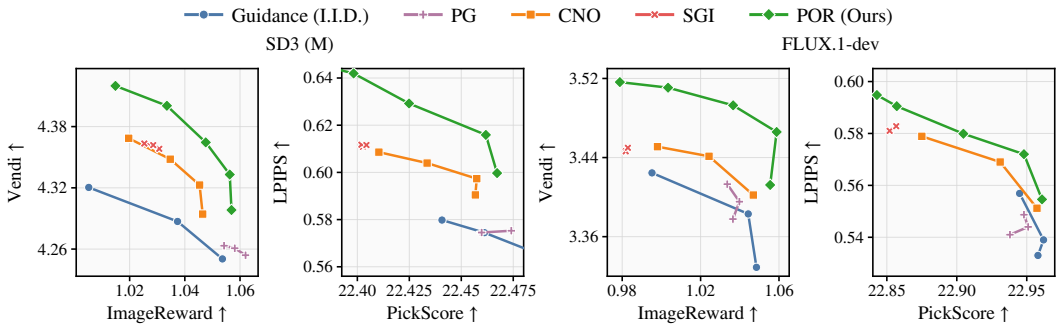


Figure 2: **Quality–diversity sweeps on SD3 and FLUX.1-dev.** For each backbone, the left panel shows ImageReward–Vendi and the right panel shows PickScore–LPIPS. Both pair image preference (horizontal axis) with diversity (vertical axis); higher is better on both axes.

Effect of the optimization objective. To isolate the effect of the optimization objective, we compare POR with Mean, a baseline that encourages separation between conditional clean predictions. Both methods use the same state-optimization procedure on SD3. Figure 3 shows that POR achieves greater diversity at comparable image preference across the three metric pairs. This comparison supports the benefit of the proposed surrogate over the Mean baseline within the same optimization framework. Appendix B.4 provides the objective definitions and sweep settings.

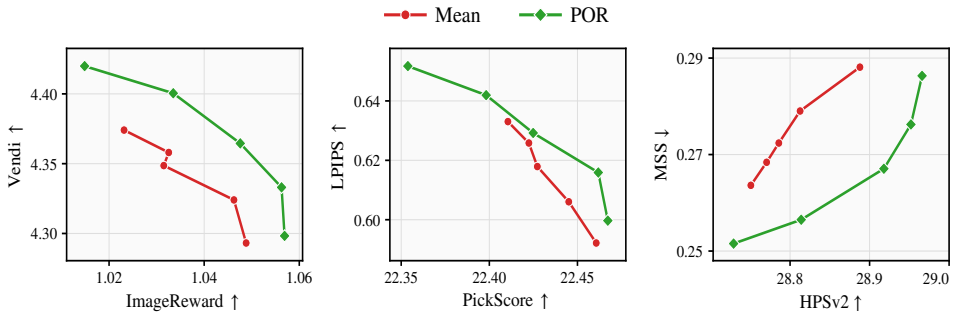


Figure 3: **Effect of the optimization objective on SD3.** Each curve shows a learning-rate sweep for Mean or POR. The panels compare ImageReward–Vendi, PickScore–LPIPS, and HPSv2–MSS from left to right. Higher Vendi and LPIPS and lower MSS indicate greater diversity.

Effect of the intervention time. On SD3, we examine how the intervention step affects the quality–diversity tradeoff. Figure 4 compares intervention steps 1, 5, 10, and 15 by sweeping the learning rate at each step; Appendix B.5 gives the sweep settings. Higher PickScore and Vendi indicate higher image preference and greater group diversity, respectively. Over the overlapping PickScore range, step 1 provides higher Vendi than the later interventions, with a larger gap toward the high-diversity end. Among the later interventions, delaying the update generally reduces Vendi at a comparable PickScore. All four curves gain diversity at the cost of lower PickScore relative to the classifier-free guidance (CFG) reference. These results support step 1 as the choice for the single intervention in our reference configuration.

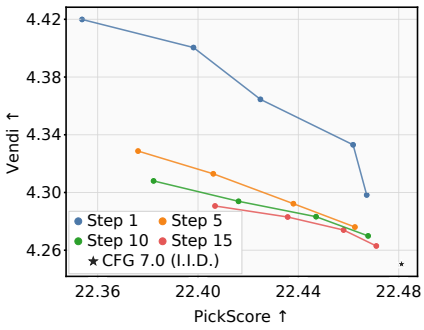


Figure 4: **Effect of the POR intervention time.** Each curve shows a learning-rate sweep at a fixed intervention step.



Figure 5: **Qualitative results on SD3.** Columns compare I.I.D. sampling, Scaling Group Inference (SGI), and POR, with three samples per method for each prompt. In the museum gallery, POR produces a rounded seated figure, a white standing statue, and a slender, angular sculpture, along with changes in windows and viewpoint. In the landscape, POR samples range from a view framed by a large foreground tree to open fields beneath dark clouds or a blue sky, while retaining the sunrise and mist. Additional qualitative results are provided in Appendix C.

5 Conclusion

In this paper, we introduced POR, a training-free method for reducing visual repetition within a fixed group of text-to-image samples. POR uses posterior overlap as a distributional proxy for potential visual repetition. The theoretical connection between posterior overlap and symmetrized KL divergence motivates a surrogate computed from noisy states and conditional clean predictions. POR uses this surrogate to jointly update the group at a single timestep early in sampling, then resumes the original sampler. On COCO and DPG-Bench, POR achieves the highest Vendi and lowest MSS among the compared methods on both SD3 and FLUX.1-dev, with competitive image quality and prompt alignment. The SD3 objective comparison further supports the benefit of our surrogate over a clean-prediction distance objective within the same state-optimization framework. Together, these results support posterior overlap as a practical criterion for group diversification under a shared prompt.

AI USE STATEMENT

We used large language models to review the manuscript's overall flow and coherence and to check whether the writing clearly conveyed our intended meaning. We also used these tools to polish sentence-level wording and grammar and improve clarity throughout the paper.

REFERENCES

Shun-ichi Amari. *Information geometry and its applications*. Springer, 2016.

Arindam Banerjee, Srujana Merugu, Inderjit S Dhillon, and Joydeep Ghosh. Clustering with bregman divergences. *Journal of machine learning research*, 6(Oct):1705–1749, 2005.

Anil Bhattacharyya. On a measure of divergence between two statistical populations defined by their probability distribution. *Bulletin of the Calcutta Mathematical Society*, 35:99–110, 1943.

Gabriele Corso, Yilun Xu, Valentin De Bortoli, Regina Barzilay, and Tommi Jaakkola. Particle guidance: non-iid diverse sampling with diffusion models. In *International Conference on Learning Representations*, volume 2024, pp. 22480–22507, 2024.

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

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

Dan Friedman and Adji Bousso Dieng. The vendi score: A diversity evaluation metric for machine learning. *arXiv preprint arXiv:2210.02410*, 2022.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pp. 7514–7528, 2021.

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

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

Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. Ella: Equip diffusion models with llm for enhanced semantic alignment. *arXiv preprint arXiv:2403.05135*, 2024.

Henry F Inman and Edwin L Bradley Jr. The overlapping coefficient as a measure of agreement between probability distributions and point estimation of the overlap of two normal densities. *Communications in Statistics-theory and Methods*, 18(10):3851–3874, 1989.

Rafał Karczewski, Markus Heinonen, Alison Pouplin, Søren Hauberg, and Vikas Garg. The spacetime of diffusion models: An information geometry perspective. In *International Conference on Learning Representations*, volume 2026, pp. 106128–106152, 2026.

Byungjun Kim, Soobin Um, and Jong Chul Ye. Diverse text-to-image generation via contrastive noise optimization. *arXiv preprint arXiv:2510.03813*, 2025.

Michael Kirchhof, James Thornton, Louis Béthune, Pierre Ablin, Eugene Ndiaye, and Marco Cuturi. Shielded diffusion: Generating novel and diverse images using sparse repellency. *arXiv preprint arXiv:2410.06025*, 2024.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in neural information processing systems*, 36:36652–36663, 2023.

Tuomas Kynkäänniemi, Miika Aittala, Tero Karras, Samuli Laine, Timo Aila, and Jaakko Lehtinen. Applying guidance in a limited interval improves sample and distribution quality in diffusion models. *Advances in Neural Information Processing Systems*, 37:122458–122483, 2024.

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

Chieh-Hsin Lai, Yang Song, Dongjun Kim, Yuki Mitsufuji, and Stefano Ermon. The principles of diffusion models. *arXiv preprint arXiv:2510.21890*, 2025.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

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

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

Nanye Ma, Shangyuan Tong, Haolin Jia, Hexiang Hu, Yu-Chuan Su, Mingda Zhang, Xuan Yang, Yandong Li, Tommi Jaakkola, Xuhui Jia, et al. Inference-time scaling for diffusion models beyond scaling denoising steps. *arXiv preprint arXiv:2501.09732*, 2025.

Hila Manor and Tomer Michaeli. On the posterior distribution in denoising: Application to uncertainty quantification. In *International Conference on Learning Representations*, volume 2024, pp. 49233–49263, 2024.

Chenlin Meng, Yang Song, Wenzhe Li, and Stefano Ermon. Estimating high order gradients of the data distribution by denoising. *Advances in Neural Information Processing Systems*, 34: 25359–25369, 2021.

Mashrur M Morshed and Vishnu Boddeti. Diverseflow: Sample-efficient diverse mode coverage in flows. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23303–23312. IEEE, 2025.

Gaurav Parmar, Or Patashnik, Daniil Ostashev, Kuan-Chieh Wang, Kfir Aberman, Srinivasa Narasimhan, and Jun-Yan Zhu. Scaling group inference for diverse and high-quality generation. In *International Conference on Learning Representations*, volume 2026, pp. 89763–89790, 2026.

Ed Pizzi, Sreya Dutta Roy, Sugosh Nagavara Ravindra, Priya Goyal, and Matthijs Douze. A self-supervised descriptor for image copy detection. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14512–14522. IEEE, 2022.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

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.

Seyedmorteza Sadat, Jakob Buhmann, Derek Bradley, Otmar Hilliges, and Romann Weber. Cads: Unleashing the diversity of diffusion models through condition-annealed sampling. In *International Conference on Learning Representations*, volume 2024, pp. 23723–23755, 2024.

Soobin Um, Suhyeon Lee, and Jong Chul Ye. Don't play favorites: Minority guidance for diffusion models. In *International Conference on Learning Representations*, volume 2024, pp. 21632–21670, 2024.

Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:15903–15935, 2023.

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

A Proofs and Additional Results

A.1 Posterior Representation and Moment Identities

Fix t and c, with $\alpha_t > 0$ and $\sigma_t > 0$. Under Gaussian perturbation, the conditional data distribution and Gaussian perturbation kernel define a joint distribution, from which Bayes' rule gives a posterior over clean samples (Lai et al., 2025):

$$q_t(x_0, x_t \mid c) = q_0(x_0 \mid c)\, q_t(x_t \mid x_0), \tag{16}$$

$$q_t(x_0 \mid x_t, c) = \frac{q_t(x_t \mid x_0) q_0(x_0 \mid c)}{q_t(x_t \mid c)}, \tag{17}$$

where $q_t(x_t \mid c)$ is the noisy marginal density. This posterior is defined by the forward joint distribution and need not coincide with the endpoint distribution obtained by continuing a chosen reverse sampler. For example, a deterministic reverse sampler can map a fixed state to a single endpoint while the posterior remains non-degenerate.

We derive the posterior representation and moment identities in the noisy state variable x_t; see also the denoising exponential-family treatment in Karczewski et al. (2026). Substituting the Gaussian perturbation kernel into Eq. (17) gives

$$\begin{aligned} q_t(x_0 \mid x_t, c) &\propto q_0(x_0 \mid c) \exp\left(-\frac{\|x_t - \alpha_t x_0\|_2^2}{2\sigma_t^2}\right) \\ &\propto q_0(x_0 \mid c) \exp\left(-\frac{\alpha_t^2}{2\sigma_t^2}\|x_0\|_2^2\right) \exp\left(\frac{\alpha_t}{\sigma_t^2} x_t^\top x_0\right). \end{aligned} \tag{18}$$

Define the common base and its state-dependent log normalizer by

$$h_t(x_0; c) := q_0(x_0 \mid c) \exp\left(-\frac{\alpha_t^2}{2\sigma_t^2}\|x_0\|_2^2\right), \tag{19}$$

$$\mathcal{A}_t(x_t) := \log \int h_t(x_0; c) \exp\left(\frac{\alpha_t}{\sigma_t^2} x_t^\top x_0\right) dx_0, \tag{20}$$

where the fixed condition c is suppressed in $\mathcal{A}_t$. The posterior is therefore

$$q_t(x_0 \mid x_t, c) = h_t(x_0; c) \exp\left\{\frac{\alpha_t}{\sigma_t^2} x_t^\top x_0 - \mathcal{A}_t(x_t)\right\}. \tag{21}$$

The base h_t is independent of x_t and need not integrate to one. This exponential tilting of the common base defines an exponential family with natural parameter $\eta = (\alpha_t/\sigma_t^2)x_t$. We derive the moment identities by differentiating with respect to the noisy state x_t.

To differentiate under the integral, the integrand and its first two state derivatives must admit an integrable dominating function on a neighborhood of the state segment. For a normalized clean-data prior, the Gaussian damping in h_t ensures this condition: on any compact state neighborhood, each polynomial factor of degree at most two times the exponential is bounded by a constant, leaving an integrable multiple of $q_0(x_0 \mid c)$. Thus the normalizer is positive and finite, the posterior second moments are finite, and $\mathcal{A}_t$ is twice continuously differentiable.

Define $Z_t(x_t) = \exp[\mathcal{A}_t(x_t)]$. Differentiating its logarithm gives

$$\begin{aligned} \nabla_{x_t} \mathcal{A}_t(x_t) &= \frac{\nabla_{x_t} Z_t(x_t)}{Z_t(x_t)} = \frac{\alpha_t}{\sigma_t^2} \frac{\int x_0 h_t(x_0; c) \exp\left(\frac{\alpha_t}{\sigma_t^2} x_t^\top x_0\right) dx_0}{Z_t(x_t)} \\ &= \frac{\alpha_t}{\sigma_t^2} \mu_t(x_t, c). \end{aligned} \tag{22}$$

Differentiating again expresses the Hessian as a scaled posterior covariance:

$$\begin{aligned} \nabla_{x_t}^2 \mathcal{A}_t(x_t) &= \frac{\nabla_{x_t}^2 Z_t(x_t)}{Z_t(x_t)} - \frac{\nabla_{x_t} Z_t(x_t) \nabla_{x_t} Z_t(x_t)^\top}{Z_t(x_t)^2} \\ &= \frac{\alpha_t^2}{\sigma_t^4} \left[\mathbb{E}[x_0 x_0^\top \mid x_t, c] - \mu_t(x_t, c)\mu_t(x_t, c)^\top\right] \\ &= \frac{\alpha_t^2}{\sigma_t^4} \Sigma_t(x_t, c). \end{aligned} \tag{23}$$

The log normalizer differs from the log density of the noisy marginal:

$$\log q_t(x_t \mid c) = \mathcal{A}_t(x_t) - \frac{\|x_t\|^2}{2\sigma_t^2} - \frac{d}{2}\log(2\pi\sigma_t^2). \tag{24}$$

Differentiating Eq. (24) and using Eq. (22) gives Tweedie's formula (Efron, 2011; Lai et al., 2025),

$$\mu_t(x_t, c) = \frac{x_t + \sigma_t^2 \nabla_{x_t} \log q_t(x_t \mid c)}{\alpha_t}. \tag{25}$$

A.2 Symmetrized KL Divergence from States and Posterior Means

The standard exponential-family KL identity (Banerjee et al., 2005; Amari, 2016) specializes to the scaled inner product between state differences and posterior-mean differences below; the corresponding denoising-family identity is also given by Karczewski et al. (2026). To see the cancellation directly, let $\Delta x_{t,ij} = x_t^{(i)} - x_t^{(j)}$. On the common support of the two posteriors, Eq. (21) gives

$$\log \frac{q_i^t(x_0)}{q_j^t(x_0)} = \frac{\alpha_t}{\sigma_t^2}\Delta x_{t,ij}^\top x_0 - \mathcal{A}_t(x_t^{(i)}) + \mathcal{A}_t(x_t^{(j)}). \tag{26}$$

Taking expectations in the two directions, with $\mu_t^{(i)} = \mu_t(x_t^{(i)}, c)$, yields

$$\begin{aligned} D_{\mathrm{KL}}(q_i^t \| q_j^t) &= \frac{\alpha_t}{\sigma_t^2}\Delta x_{t,ij}^\top \mu_t^{(i)} - \mathcal{A}_t(x_t^{(i)}) + \mathcal{A}_t(x_t^{(j)}), \\ D_{\mathrm{KL}}(q_j^t \| q_i^t) &= -\frac{\alpha_t}{\sigma_t^2}\Delta x_{t,ij}^\top \mu_t^{(j)} - \mathcal{A}_t(x_t^{(j)}) + \mathcal{A}_t(x_t^{(i)}). \end{aligned} \tag{27}$$

Adding the two expressions cancels the log normalizers. Dividing by two and applying Eq. (22) gives

$$\begin{aligned} D_{\mathrm{sKL}}(q_i^t, q_j^t) &= \frac{\alpha_t}{2\sigma_t^2}\Delta x_{t,ij}^\top \big[\mu_t^{(i)} - \mu_t^{(j)}\big] \\ &= \frac{1}{2}\Delta x_{t,ij}^\top \big[\nabla \mathcal{A}_t(x_t^{(i)}) - \nabla \mathcal{A}_t(x_t^{(j)})\big]. \end{aligned} \tag{28}$$

This state–mean identity is exact for multimodal posteriors as well and does not require the relative variance condition. The factor $1/2$ follows from our definition of symmetrized KL divergence as half the sum of the two directional KL divergences.

A.3 Proof of the Variance Representations and the Overlap Bound

Proof of Lemma 2. Substituting Eq. (21) into the Bhattacharyya coefficient and integrating over x_0 gives

$$\begin{aligned} \int \sqrt{q_i^t(x_0) q_j^t(x_0)}\, dx_0 &= \exp\left[-\frac{\mathcal{A}_t(x_t^{(i)}) + \mathcal{A}_t(x_t^{(j)})}{2}\right] \\ &\quad \times \int h_t(x_0; c) \exp\left[\frac{\alpha_t}{\sigma_t^2}\left(\frac{x_t^{(i)} + x_t^{(j)}}{2}\right)^\top x_0\right] dx_0 \\ &= \exp\left[\mathcal{A}_t\left(\frac{x_t^{(i)} + x_t^{(j)}}{2}\right) - \frac{\mathcal{A}_t(x_t^{(i)}) + \mathcal{A}_t(x_t^{(j)})}{2}\right]. \end{aligned} \tag{29}$$

The inner integral is the normalizer at the midpoint state. Taking the negative logarithm of the full Bhattacharyya coefficient yields

$$B(q_i^t, q_j^t) = \frac{\mathcal{A}_t(x_t^{(i)}) + \mathcal{A}_t(x_t^{(j)})}{2} - \mathcal{A}_t\left(\frac{x_t^{(i)} + x_t^{(j)}}{2}\right). \tag{30}$$

On the state segment $x_{t,ij}(s)$ in Eq. (8), define $f(s) = \mathcal{A}_t(x_{t,ij}(s))$. The chain rule and the moment identities in Appendix A.1 give

$$f'(s) = \Delta x_{t,ij}^\top \nabla \mathcal{A}_t(x_{t,ij}(s)) = \frac{\alpha_t}{\sigma_t^2} \Delta x_{t,ij}^\top \mu_t(x_{t,ij}(s), c),$$

$$f''(s) = \Delta x_{t,ij}^\top \nabla^2 \mathcal{A}_t(x_{t,ij}(s)) \Delta x_{t,ij} \tag{31}$$

$$= \frac{\alpha_t^2}{\sigma_t^4} \Delta x_{t,ij}^\top \Sigma_t(x_{t,ij}(s), c) \Delta x_{t,ij} = v_{ij}(s) \geq 0. \tag{32}$$

This restricts a function of the noisy state to a line segment; it assumes neither a one-dimensional posterior nor a local quadratic approximation. The posterior at an interior state is obtained by conditioning on that state, rather than by averaging the endpoint densities.

Equation (28) and the fundamental theorem of calculus yield

$$D_{\mathrm{sKL}}(q_i^t, q_j^t) = \frac{1}{2}[f'(1) - f'(0)] = \frac{1}{2} \int_0^1 f''(s)\, ds \tag{33}$$

$$= \frac{1}{2} \int_0^1 v_{ij}(s)\, ds. \tag{34}$$

For the midpoint gap $B = (f(0) + f(1))/2 - f(1/2)$, integration by parts on each half of the segment gives

$$\int_0^{1/2} s f''(s)\, ds = f(0) - f(1/2) + \tfrac{1}{2} f'(1/2),$$

$$\int_{1/2}^1 (1-s) f''(s)\, ds = f(1) - f(1/2) - \tfrac{1}{2} f'(1/2). \tag{35}$$

Adding cancels the midpoint derivative terms. Multiplying by $1/2$ therefore gives

$$B(q_i^t, q_j^t) = \frac{f(0) + f(1)}{2} - f(1/2)$$

$$= \frac{1}{2} \int_0^1 \min\{s, 1-s\}\, v_{ij}(s)\, ds. \tag{36}$$

This proves both identities in Lemma 2. The triangular weight follows from the midpoint gap and weights the variance at each segment position.

Proof of Proposition 1. By Lemma 1,

$$\Omega_{ij}^t \leq \int \sqrt{q_i^t(x_0) q_j^t(x_0)}\, dx_0 = \exp[-B(q_i^t, q_j^t)]. \tag{37}$$

Now impose the relative variance condition in Eq. (12) on nonidentical posterior pairs. Since $\int_0^1 \min\{s, 1-s\}\, ds = 1/4$, the two exact representations in Lemma 2 give

$$\frac{m_{ij}}{2} \leq D_{\mathrm{sKL}}(q_i^t, q_j^t) \leq \frac{M_{ij}}{2}, \qquad \frac{m_{ij}}{8} \leq B(q_i^t, q_j^t) \leq \frac{M_{ij}}{8}. \tag{38}$$

Consequently,

$$\frac{m_{ij}}{4M_{ij}} D_{\mathrm{sKL}}(q_i^t, q_j^t) \leq B(q_i^t, q_j^t) \leq \frac{M_{ij}}{4m_{ij}} D_{\mathrm{sKL}}(q_i^t, q_j^t). \tag{39}$$

The lower inequality and $m_{ij}/M_{ij} \geq \kappa_0$ imply $B(q_i^t, q_j^t) \geq (\kappa_0/4) D_{\mathrm{sKL}}(q_i^t, q_j^t)$. Combining this with Eq. (37) yields

$$\Omega_{ij}^t \leq \exp[-B(q_i^t, q_j^t)] \leq \exp\Big[-\frac{\kappa_0}{4} D_{\mathrm{sKL}}(q_i^t, q_j^t)\Big],$$

which is Eq. (13).

Identical posteriors have $D_{\mathrm{sKL}} = B = 0$ and $\Omega = 1$, so the exact representations reduce to $0 = 0$ and the overlap bound holds with equality, without a positive lower variance bound. The common κ_0 must apply to the nonidentical pairs throughout the state region under comparison, although their m_{ij} and M_{ij} may differ. Within this range, increasing exact symmetrized KL divergence decreases the common upper bound on overlap.

Relative variation in posterior variance at high noise. The condition in Proposition 1 concerns how much directional posterior variance changes along the segment between two noisy states. Under Gaussian perturbation, the state difference enters the posterior through the scaled difference $(\alpha_t/\sigma_t^2)\Delta x_{t,ij}$. At high noise, this scaled difference can be small for comparable state separations. If posterior covariance responds smoothly to this change, the directional variance may change little along the segment. If this variance changes throughout the segment by only a small fraction of its positive value at one endpoint, the bounds m_{ij} and M_{ij} can be chosen with a ratio close to one. This can support a larger admissible κ_0 and a tighter overlap upper bound, motivating the early intervention in Section 3.4.

A.4 Relation Between Pairwise Overlap and Duplicated Group Coverage

For a fixed group size, mean pairwise posterior overlap also controls a set-level measure of duplicated coverage, up to constants depending only on N. Define

$$\mathcal{C}_t^{\text{distinct}} := \int \max_i q_i^t(x_0)\, dx_0, \qquad \mathcal{R}_t^{\text{cov}} := N - \mathcal{C}_t^{\text{distinct}}. \tag{40}$$

At each x_0, sort the posterior values as $a_1 \geq a_2 \geq \cdots \geq a_N \geq 0$. The pointwise duplicated mass is

$$\sum_{i=1}^{N} q_i^t - \max_i q_i^t = \sum_{k=2}^{N} a_k, \tag{41}$$

whereas the pointwise sum of pairwise minima is

$$\sum_{i<j} \min\{q_i^t, q_j^t\} = \sum_{k=2}^{N} (k-1) a_k. \tag{42}$$

Since $1 \leq k-1 \leq N-1$ for $k = 2, \ldots, N$,

$$\mathcal{R}_t^{\text{cov}} \leq \sum_{i<j} \Omega_{ij}^t \leq (N-1)\mathcal{R}_t^{\text{cov}}. \tag{43}$$

Using Eq. (6), the middle term equals $\binom{N}{2}\mathcal{R}_t(\mathcal{Q}_t)$. For fixed group size, pairwise posterior redundancy and duplicated group coverage control each other up to constants depending only on N; they are not identical for $N > 2$ because higher-order intersections are counted differently.

A.5 Matched Mean Distances in a Fixed Posterior Family

To complement the schematic in Figure 1, we consider a one-dimensional, equally weighted three-component Gaussian-mixture clean-data prior with means $(-3, 0, 3)$ and common standard deviation 0.08. At the selected noise level, we use $\alpha_t = 0.45$ and $\sigma_t = 0.55$. The posterior densities and means are available analytically. We choose intermediate states by numerically inverting the posterior-mean function and evaluate overlap by numerical quadrature. Table 3 reports two pairs with the same mean distance. For the first pair, both posteriors place substantial mass on the central mode; for the second, they concentrate on different modes. Their overlaps differ despite the common clean-data prior, time, and mean distance. This comparison matches distances between means, not the individual means themselves.

Table 3: An illustrative fixed-time posterior comparison. Equal mean distances coexist with different shared probability masses.

Posterior means	Mean distance	Posterior overlap
$(-1, 1)$	2.000	0.634
$(-3, -1)$	2.000	0.339

A.6 Controlled Gaussian-Mixture Optimization

We compare actual overlap reduction using the proposed surrogate and a baseline based on posterior-mean comparisons, with identical initial groups and state-update budgets. The Gaussian-mixture setting provides directly computable posterior means and overlap, allowing us to evaluate the objectives without approximation error from a learned predictor.

Distribution and initialization. The clean-data distribution is a two-dimensional Gaussian mixture, $q_0(x_0) = \sum_{r=1}^{3} w_r \mathcal{N}(x_0; m_r, \mathrm{diag}(s_r^2))$, with parameters in Table 4. We initialize $1{,}024$ independent groups of $N = 4$ candidates by drawing $x_0 \sim q_0$ and $\epsilon \sim \mathcal{N}(0, I)$ independently and setting $x_t = \alpha_t x_0 + \sigma_t \epsilon$. All sampled groups are retained, and all candidates share the same clean-data prior and path coefficients. Across noise levels, we reuse the same clean and standard-normal noise draws.

Table 4: Gaussian-mixture parameters; s_r contains coordinate-wise standard deviations.

Component r	Weight w_r	Mean m_r	Standard deviations s_r
1	0.40	$(-4.5, 0)$	$(1.0, 1.0)$
2	0.45	$(0, 0)$	$(0.5, 0.3)$
3	0.15	$(4.5, 0)$	$(6.0, 6.0)$

Objectives and optimization. The Mean baseline compares posterior means using squared Euclidean distance. At a fixed time, let $\mu_i = \mu_t(x_t^{(i)})$ be the exact posterior mean and $D_{ij} = D_{\mathrm{sKL}}(q_i^t, q_j^t)$ be computed by the state–mean identity in Eq. (28). We minimize

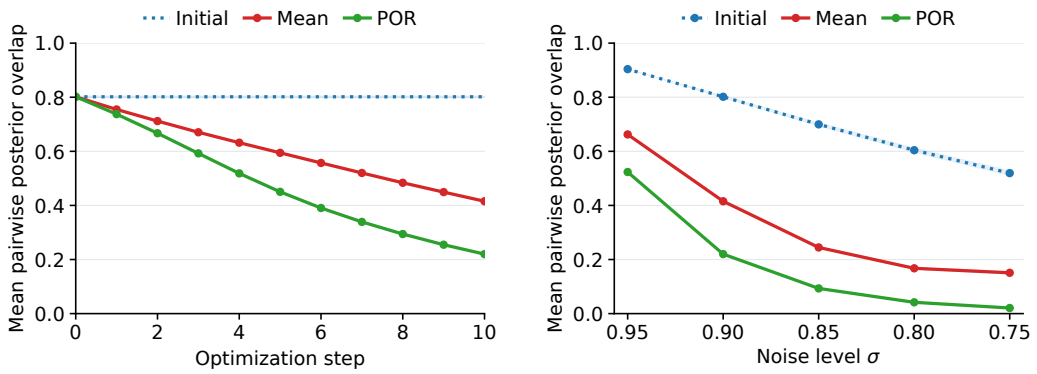
$$\begin{aligned} \mathcal{L}_{\mathrm{Mean}} &= \frac{1}{6} \sum_{i<j} \mathrm{softplus}\big(-\|\mu_i - \mu_j\|_2^2\big)\,, \\ \mathcal{L}_{\mathrm{POR}} &= \frac{1}{6} \sum_{i<j} \mathrm{softplus}(-D_{ij}). \end{aligned} \tag{44}$$

Both losses use unit temperature and differentiate through the exact posterior-mean function, without a learned predictor. To match the state-update budget across objectives with different gradient magnitudes, we jointly update the four states by normalized gradient descent: each group's negative gradient is rescaled to root-mean-square (RMS) magnitude 0.2 across its eight coordinates. Ten updates give both methods a cumulative RMS path length of 2.0.

Overlap evaluation. We compute overlap exactly up to numerical precision. The shared base in Eq. (21) makes the posterior log-density ratio affine in x_0, so overlap reduces to Gaussian-mixture halfspace probabilities evaluated with univariate Gaussian cumulative distribution functions. We average over each group's six pairs and then over groups. Shaded bands show pointwise 95% normal-approximation confidence intervals, using groups as the independent units. In Figure 6, green and red solid curves denote POR and Mean, respectively; the blue dotted Initial curve shows overlap before optimization.

Overlap during optimization. Figure 6(a) fixes $(\alpha_t, \sigma_t) = (0.1, 0.9)$. From the common initial overlap of 0.802, Mean and POR reach 0.594 and 0.450 after five updates, and 0.415 and 0.220 after ten updates, respectively. The surrogate achieves a larger overlap reduction within the same update budget in this setting.

Comparison across noise levels. Figure 6(b) shows overlap after ten updates at $\sigma_t \in \{0.95, 0.90, 0.85, 0.80, 0.75\}$, using $\alpha_t = 1 - \sigma_t$ and displaying σ_t in decreasing order from left to right. POR yields lower final overlap than Mean at all five evaluated levels. At $\sigma_t = 0.95$, initial, Mean, and POR overlaps are 0.904, 0.662, and 0.523; at $\sigma_t = 0.75$, they are 0.519, 0.151, and 0.021. These results support the surrogate's effectiveness relative to the Mean baseline under the evaluated conditions.



(a) Optimization steps

(b) Noise levels

Figure 6: Mean pairwise posterior overlap over 1,024 groups of four candidates: (a) during optimization at $\sigma_t = 0.9$; (b) after ten updates across noise levels with $\alpha_t = 1 - \sigma_t$.

B Implementation Details

B.1 Experimental Protocol and Reference Configurations

The tabular comparisons use 2,000 prompts from the COCO validation set and all 1,065 prompts from DPG-Bench. We generate five images per prompt with SD3 and four with FLUX.1-dev. Within each backbone, all compared methods use the same prompt set, output group size, image resolution, and base sampling configuration whenever applicable. The tabular comparisons use the fixed reference configurations in Table 5, with each backbone's sampler guidance scale shared across methods. SGI uses a candidate pool of $M = 32$.

SD3 applies classifier-free guidance by combining conditional and unconditional predictions, whereas FLUX.1-dev receives a distilled guidance scale through a guidance embedding. We use sampler guidance scale as a common term for these controls.

Table 5: Fixed reference configurations for the SD3 and FLUX.1-dev tabular comparisons.

Method	Control parameter	SD3 (M)	FLUX.1-dev
I.I.D.	sampler guidance scale	7.0	3.5
PG	coefficient	30	30
CNO	learning rate	0.001	0.001
SGI	λ	1.5	1.5
POR	learning rate	0.01	0.01

B.2 Evaluation Metrics

We evaluate image quality and alignment per image and diversity within the group generated for each prompt. Per-image scores are averaged within each group, and the resulting prompt-level scores are averaged over prompts. The diversity measures below are computed for each group and then averaged over prompts.

Image quality and prompt alignment. CLIP-based alignment (Radford et al., 2021; Hessel et al., 2021) measures the similarity between normalized image and text embeddings. We report $100 \max(\cos(e_{\text{image}}, e_{\text{text}}), 0)$, where the factor of 100 sets the reporting scale. PickScore (Kirstain et al., 2023) predicts image preference from a prompt–image pair; we average its raw scores rather than probabilities normalized across the generated group. ImageReward (Xu et al., 2023) supplies a

complementary scalar reward learned from human preference feedback. Human Preference Score v2 (HPSv2) (Wu et al., 2023) also predicts human preference using a scoring model trained on pairwise choices. Higher scores are preferred for these four measures.

Group diversity. Vendi Score (Friedman & Dieng, 2022) summarizes the effective diversity encoded by a group's image-similarity matrix. For N generated images, let f_i be the normalized SSCD image descriptor (Pizzi et al., 2022) and let $S_{ij} = f_i^\top f_j$. If $\lambda_1, \ldots, \lambda_N$ are the eigenvalues of S/N, then

$$\mathrm{Vendi}(S) = \exp\left(-\sum_{\ell=1}^{N} \lambda_\ell \log \lambda_\ell\right), \qquad \mathrm{MSS}(S) = \frac{2}{N(N-1)} \sum_{i<j} S_{ij}, \tag{45}$$

with $0 \log 0 = 0$. Vendi is larger when the group contains less redundant image representations, whereas MSS directly averages their pairwise similarity and is lower for more diverse groups. Since the Vendi scale depends on group size, we compare methods within each backbone at a fixed N.

LPIPS (Zhang et al., 2018) measures perceptual distance between image pairs. We use the average over all $N(N-1)/2$ distinct pairs in a prompt group, with a higher value indicating greater perceptual variation. These measures describe diversity among final generated images and complement the quality and alignment scores.

Metric implementations. We use CLIP ViT-L/14, PickScore v1 with its CLIP ViT-H/14 processor, ImageReward-v1.0, and AlexNet LPIPS v0.1. CLIP and PickScore use their repository-provided processors, and ImageReward uses its package preprocessing. SSCD descriptors use the `sscd_disc_mixup` checkpoint with RGB images resized to 320×320 by bicubic interpolation and normalized with the ImageNet channel means and standard deviations. LPIPS inputs are resized to 256×256 by bicubic interpolation and scaled to $[-1, 1]$.

B.3 Quality–Diversity Sweeps

For both backbones, the sweeps in Figure 2 use 400 randomly selected COCO prompts, with five images per prompt for SD3 and four for FLUX.1-dev. POR uses conditional clean predictions and differentiates the objective through the predictor with respect to the particle states, as described in Section 3. The predictor parameters remain frozen during these updates.

Each sweep varies one primary control parameter per method while keeping the remaining settings fixed. SGI uses $M = 32$ candidates for both backbones, as in the tabular comparisons. Table 6 lists the SD3 sweep values.

Table 6: SD3 parameter sweeps used for the COCO quality–diversity tradeoff.

Method	Swept parameter	Values
I.I.D.	CFG scale	$\{4, 5, 7\}$
PG	coefficient	$\{30, 60, 200\}$
CNO	learning rate	$\{0.001, 0.0015, 0.002, 0.0025\}$
SGI	weight factor λ	$\{1.5, 3, 5\}$
POR	learning rate	$\{0.005, 0.01, 0.015, 0.02, 0.025\}$

Table 7 lists the control parameters for the displayed FLUX.1-dev sweep in Figure 2. CNO and POR vary their learning rates, while PG varies its repulsion coefficient and SGI its diversity weight.

B.4 Optimization-Objective Comparison on SD3

Comparison setup. Figure 3 compares POR with a Mean baseline using the same 400 COCO prompts and five images per prompt as the SD3 sweep in Appendix B.3. For each prompt, both methods start from the same noise and use the same sampler, intervention time, optimizer, and number of state updates. Both follow the SD3 reference settings in Table 9: 28 sampling steps with CFG 7.0

Table 7: FLUX.1-dev parameter sweeps used for the displayed quality–diversity tradeoff.

Method	Swept parameter	Values
I.I.D.	distilled guidance scale	$\{2, 3.5, 5\}$
PG	coefficient	$\{30, 60, 200\}$
CNO	learning rate	$\{0.0005, 0.001, 0.002\}$
SGI	weight factor λ	$\{1.5, 3\}$
POR	learning rate	$\{0.005, 0.01, 0.015, 0.02, 0.025\}$

and five Adam updates at intervention step 1. Only the pairwise comparison score is replaced, and each objective is evaluated over its own learning-rate sweep.

Mean baseline. At the selected timestep $t^\star$, let $\hat{x}_i = \hat{x}_{0,c}(x_{t^\star}^{(i)}, t^\star)$ denote the conditional clean prediction for candidate i. The Mean baseline replaces POR's estimated pairwise symmetrized KL divergence with the squared Euclidean distance between these predictions:

$$M_{ij} = \|\hat{x}_i - \hat{x}_j\|_2^2. \tag{46}$$

Both variants use a decreasing softplus loss averaged over candidate pairs and jointly update the noisy states while keeping the pretrained generator frozen. Thus, Mean encourages separation between clean predictions, whereas POR uses both the noisy-state difference and the clean-prediction difference through Eq. (14). The clean predictions are refreshed at every optimization step for both objectives.

Learning-rate sweeps. Mean uses learning rates $\{0.001, 0.002, 0.003, 0.004, 0.005\}$, and POR uses $\{0.005, 0.01, 0.015, 0.02, 0.025\}$. We compare the resulting quality–diversity curves rather than matching numerical learning rates, since the two objectives have different gradient scales. Figure 3 reports ImageReward–Vendi, PickScore–LPIPS, and HPSv2–MSS, using the metric definitions and aggregation in Appendix B.2.

B.5 Intervention-Time Sweep

Figure 4 compares intervention steps 1, 5, 10, and 15. The common learning-rate sweep $\{0.005, 0.01, 0.015, 0.02\}$ is used at every intervention time. We additionally evaluate learning rate 0.025 at step 1 to extend the high-diversity end of the early-intervention curve. All remaining POR optimization and sampling settings are held fixed.

The implementation indexes sampling steps from zero and applies POR before the solver update at the selected index. Thus, intervention at step 1 follows one ordinary sampling update. At index k, we set σ_t to the scheduler's noise level at that index and use $\alpha_t = 1 - \sigma_t$ for the flow parameterization.

Table 8: POR learning-rate sweeps used for the intervention-time comparison on SD3.

Intervention step	Learning-rate sweep
1	$\{0.005, 0.01, 0.015, 0.02, 0.025\}$
5	$\{0.005, 0.01, 0.015, 0.02\}$
10	$\{0.005, 0.01, 0.015, 0.02\}$
15	$\{0.005, 0.01, 0.015, 0.02\}$

B.6 POR at One Selected Sampling Step

Algorithm 1 summarizes the state-update procedure using the state notation x_t from Section 3.1. The pretrained generator is frozen, and only the particle states at one selected early sampling step are updated. Each inner iteration refreshes the conditional model prediction at the current latent, constructs the conditional clean prediction, computes estimates of pairwise symmetrized KL

divergence in Eq. (14), and takes one latent-optimization step. Sampling then resumes with the original sampler.

Algorithm 1 Posterior Overlap Reduction at a Selected Timestep

Require: Sampler S, condition c, $N \geq 2$, $t^\star$, $K > 0$, learning rate ρ

$X_T = \{x_T^{(i)}\}_{i=1}^N, \quad x_T^{(i)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I)$
$X_{t^\star} \leftarrow S(X_T, c)$
for $k = 1, \ldots, K$ **do**
  $\hat{x}_i \leftarrow \hat{x}_{0,c}(x_{t^\star}^{(i)}, t^\star), \quad i = 1, \ldots, N$ ▷ Conditional clean predictions
  $d_{ij} \leftarrow \frac{\alpha_{t^\star}}{2\sigma_{t^\star}^2} \langle x_{t^\star}^{(i)} - x_{t^\star}^{(j)}, \hat{x}_i - \hat{x}_j \rangle, \quad i < j$ ▷ Pairwise score
  $\mathcal{L} \leftarrow \frac{2}{N(N-1)} \sum_{i<j} \mathrm{softplus}(-d_{ij})$
  $X_{t^\star} \leftarrow \mathrm{Adam}(X_{t^\star}, \nabla_{X_{t^\star}} \mathcal{L}; \rho)$
end for
$X_0 \leftarrow S(X_{t^\star}, c)$
return X_0

B.7 Clean Predictions Used in the POR Objective

For latent diffusion or flow implementations, we denote the current particle state by z_t and the corresponding clean latent by z_0. Let f_c and f_u denote conditional and unconditional predictions in the same model parameterization. Classifier-free guidance (Ho & Salimans, 2022) combines them as

$$f_{\text{cfg}} = f_u + w(f_c - f_u), \tag{47}$$

with guidance scale w. For the SD3 flow parameterization, define

$$\hat{z}_{0,c} = z_t - \sigma_t v_c, \qquad \hat{z}_{0,u} = z_t - \sigma_t v_u. \tag{48}$$

Then

$$\hat{z}_{0,\text{cfg}} = \hat{z}_{0,u} + w(\hat{z}_{0,c} - \hat{z}_{0,u}) = \hat{z}_{0,c} + (w-1)(\hat{z}_{0,c} - \hat{z}_{0,u}). \tag{49}$$

For $w > 1$, the CFG-combined clean prediction is an affine extrapolation beyond the conditional prediction rather than, in general, the mean of the reference conditional posterior $q_t(z_0 \mid z_t, c)$. Across two particles,

$$\Delta\hat{z}_{0,\text{cfg}} = \Delta\hat{z}_{0,c} + (w-1)\Delta(\hat{z}_{0,c} - \hat{z}_{0,u}), \tag{50}$$

so a CFG-based pairwise score mixes posterior-mean differences with guidance-induced changes in the clean predictions. We use the conditional branch to estimate the posterior-mean differences in Eq. (14), while CFG remains part of the base sampler's update rule.

FLUX.1-dev. For FLUX.1-dev, we use the prompt-conditioned velocity predictor with its guidance input set to 1.0. We form the clean prediction as $\hat{z}_{0,c} = z_t - \sigma_t v_c$ in packed latent coordinates, while the base sampler uses distilled guidance scale 3.5. FLUX.1-dev receives this scale through a guidance embedding. We use the clean prediction computed with guidance input 1.0 as an approximate conditional posterior-mean estimate in Eq. (14).

B.8 Full-Gradient Latent Optimization

The pretrained generator parameters remain frozen during POR. The conditional clean prediction remains differentiable with respect to the current particle state. For the SD3 flow parameterization,

$$\hat{z}_{0,c}(z, t) = z - \sigma_t v_c(z, t, c), \tag{51}$$

so its input Jacobian is

$$\frac{\partial \hat{z}_{0,c}}{\partial z} = I - \sigma_t J_{v_c}(z). \tag{52}$$

Each inner POR step differentiates the POR objective in Eq. (15) through the conditional predictor with respect to the particle states. No gradient is accumulated for the pretrained model parameters; only the particle states are optimized. The same differentiation rule applies to the clean prediction in the packed latent coordinates of FLUX.1-dev.

Numerical precision. The latent variables and POR loss are evaluated in FP32 in the reference implementation, even when the frozen predictor's forward pass uses reduced precision.

B.9 POR Reference Settings

Table 9 records the POR-specific reference settings for SD3 and FLUX.1-dev. The sampler guidance scale controls generation; the conditional prediction used in the POR objective is constructed as described in Appendix B.7.

Table 9: POR-specific reference settings for SD3 and FLUX.1-dev.

Setting	SD3 (M)	FLUX.1-dev
Sampling steps	28	28
Sampler guidance scale	7.0	3.5
Intervention step index	1	1
Inner optimization steps	5	5
Group size	5	4
Optimizer	Adam	Adam
Learning rate	0.01	0.01
Model precision	FP16	BF16
Latent optimization precision	FP32	FP32

B.10 Inference Efficiency

Table 10 compares the wall-clock time and peak allocated GPU memory needed to generate one group with SD3. We use the reference configuration in Table 9: five images per group, 28 sampling steps, and CFG 7.0. POR performs five latent-optimization updates at intervention step 1 with learning rate 0.01, and SGI uses $M = 32$ candidates, matching the main COCO comparison.

Across eager and compiled execution, POR takes 1.21–1.22$\times$ the time of I.I.D. sampling and 11.1–12.3% more time than CNO, while requiring 26.7–29.9% less time than SGI. Its reported peak allocated GPU memory is 24.533 GiB, the same as I.I.D. sampling and PG in both execution modes. Compiled timings exclude compilation and warm-up overhead.

Table 10: **Inference efficiency on SD3.** Each group contains five images, generated with 28 sampling steps and CFG 7.0. Time is reported as mean $\pm$ sample standard deviation over 10 measured runs, excluding two warm-up runs. SGI uses $M = 32$ candidates; POR uses five optimization updates at step 1. Peak memory is the maximum allocated GPU memory observed across the measured runs. Compiled timings also exclude compilation overhead.

Method	Execution	Time (s/group)↓	Peak memory (GiB)↓
I.I.D.	Eager	11.522 ± 0.079	24.533
PG	Eager	11.638 ± 0.021	24.533
CNO	Eager	12.555 ± 0.054	24.564
SGI ($M = 32$)	Eager	19.037 ± 0.163	25.446
POR (Ours)	Eager	13.950 ± 0.025	24.533
I.I.D.	Compiled	9.505 ± 0.025	24.533
PG	Compiled	9.528 ± 0.026	24.533
CNO	Compiled	10.311 ± 0.027	24.564
SGI ($M = 32$)	Compiled	16.532 ± 0.356	25.445
POR (Ours)	Compiled	11.583 ± 0.036	24.533

C Additional Qualitative Results

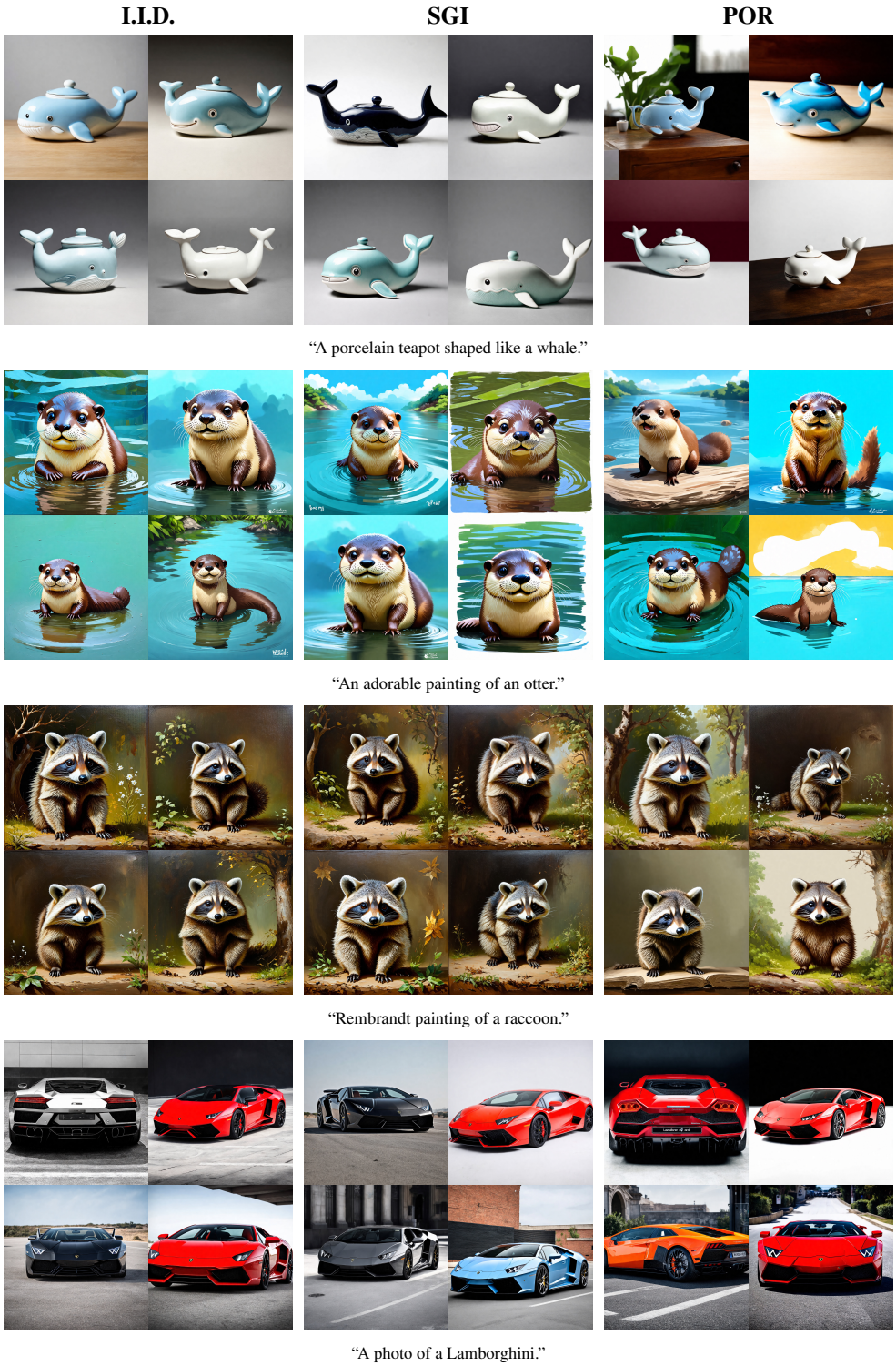


Figure 7: **Additional visual variation within SD3 groups.** Columns show independent sampling (I.I.D.), Scaling Group Inference (SGI), and POR, with four displayed samples per method for each listed prompt. The whale-shaped teapot, otter, raccoon, and car examples illustrate differences in object appearance, framing, and background.

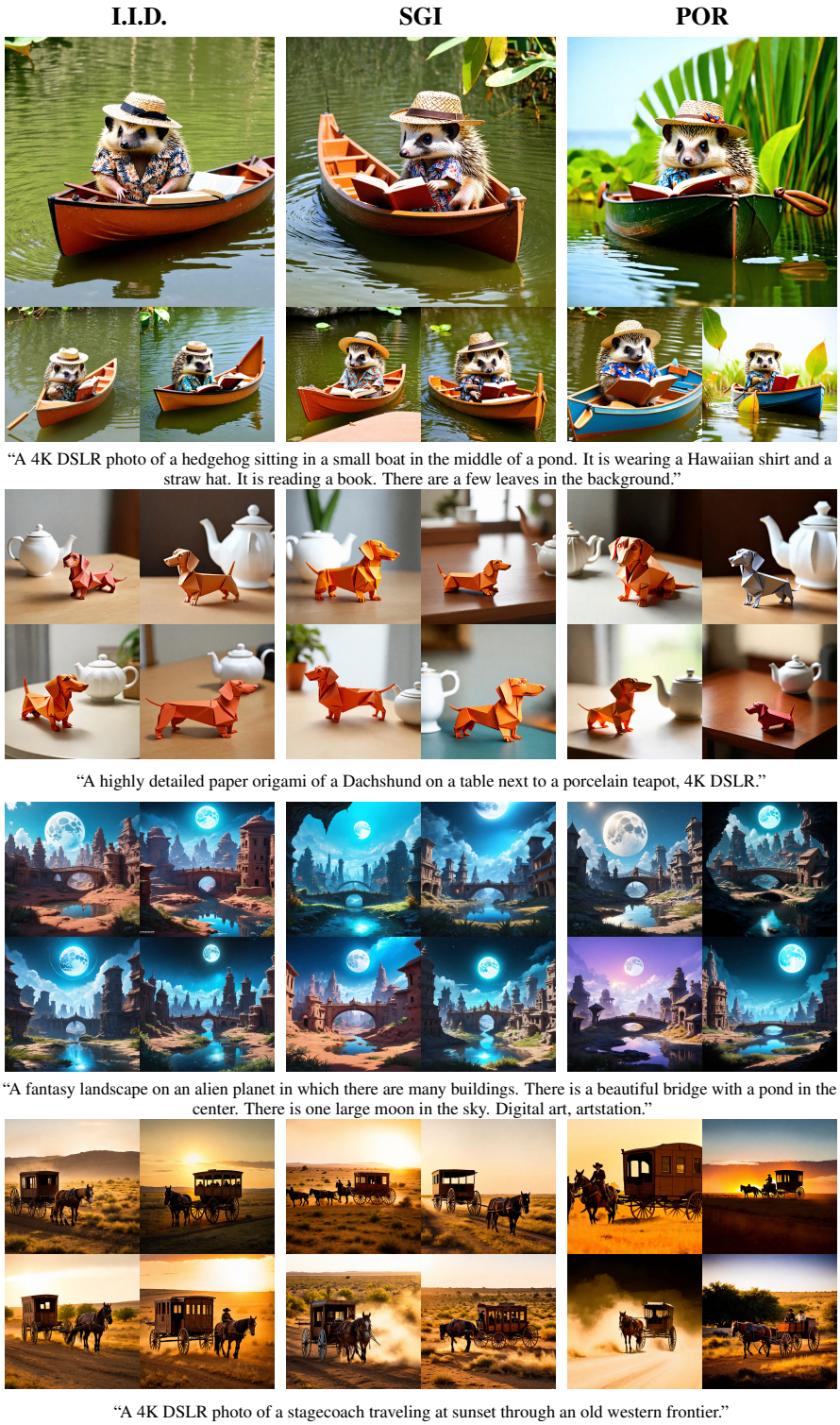


Figure 8: **Visual variation under detailed prompts on SD3.** Columns show independent sampling (I.I.D.), Scaling Group Inference (SGI), and POR. The boat scene displays three samples per method, and each remaining prompt displays four. The examples illustrate variation in framing, object arrangement, scene composition, and illumination under the listed prompts.