

Controllable Diverse Sampling through Pullback Directions

Sylvain Dehayem Kenfouo^{*1} Aaron Zweig^{*2,3} Mingxuan Zhang^{3,4} Elham Azizi^{3,4,5,6,7}

¹École Polytechnique, Institut Polytechnique de Paris

²New York Genome Center

³Irving Institute for Cancer Dynamics, Columbia University

⁴Department of Systems Biology, Columbia University

⁵Department of Computer Science, Columbia University

⁶Department of Biomedical Engineering, Columbia University

⁷Data Science Institute, Columbia University

Abstract

Conditional diffusion models can generate high-quality images that align closely with a given prompt. However, samples generated from different random seeds may remain visually similar, limiting creative exploration. Existing approaches can increase variation, but often require additional optimization or sampling overhead, while stronger perturbations may introduce distortions that degrade image quality or prompt alignment. We propose a training-free method that leverages the local pullback geometry induced by the mapping from the conditioning-embedding space to the image space to identify directions of maximal output sensitivity. By steering different sampling trajectories along these directions, we encourage distinct generations while maintaining consistency with the conditioning signal. Experiments on MS-COCO and BrushNet show the efficacy of the method in improving diversity on prompt-conditioned image generation and inpainting.

1. Introduction

Recent advances in diffusion models [12, 20, 23–25], have enabled text-to-image systems to generate visually compelling images that closely follow natural-language prompts. However, image quality and conditional alignment do not fully characterize a generative model. For a fixed condition, the model should also represent diverse valid outputs.

In practice, trajectories may concentrate around the same dominant modes of the conditional distribution, resulting in semantically redundant outputs. This limited coverage increases the cost of obtaining a diverse set of samples. Inverse problems like image inpainting[13] further constrain diversity, as only a small masked region is generated.

Many methods consider postprocessing steps that require many more gradient calculations, typically fine-tuning either the score model or the initial noise distribution to optimize a measure of diversity. Alternatively, inference-time methods based on sequential Monte Carlo can introduce diversity by drawing many samples and iteratively resampling to enforce a final distribution with high diversity. Both these methods suffer from substantially increased time complexity, either during fine-tuning or during inference. In this work, we therefore seek to answer:

* These authors contributed equally.

How can a pretrained diffusion model generate high-quality, diverse samples without fine tuning or sampling prohibitively many particles?

We propose an efficient guidance method to encourage diverse samples. Namely, for a given prompt and possible inpainting mask, we require only a constant number of Jacobian evaluations (as few as one) from which we can generate an arbitrary number of diverse samples that are conditioned by the text and/or mask. We demonstrate improved performance on diversity metrics over comparable baselines operating in the regime of no fine-tuning and cheap inference.

2. Preliminaries

Diffusion Models. We consider Denoising Diffusion Probabilistic Models (DDPM), which define a forward process that gradually corrupts samples from the data distribution with Gaussian noise [12]. Given a clean image $x_0 \sim p_{\text{data}}$, the forward Markov chain is defined as

$$q(x_{1:T} | x_0) = \prod_{t=1}^T q(x_t | x_{t-1}), \quad q(x_t | x_{t-1}) = \mathcal{N}(x_t; \sqrt{\alpha_t} x_{t-1}, (1 - \alpha_t)I), \quad (1)$$

where $\alpha_t = 1 - \beta_t$ and $\{\beta_t\}_{t=1}^T$ denotes a predefined noise schedule. We let $\bar{\alpha}_t = \prod_{s=1}^t \alpha_s$. For a sufficiently long diffusion process, x_T approaches an isotropic Gaussian distribution. Given a network ϵ_θ trained to predict the noise used to generate x_t conditioned on the condition c , the reverse process is modeled by

$$p_\theta(x_{t-1} | x_t, c) = \mathcal{N}\left(x_{t-1}; \frac{1}{\sqrt{\alpha_t}} \left(x_t - \frac{\beta_t}{\sqrt{1 - \bar{\alpha}_t}} \epsilon_\theta(x_t, t, c)\right), \sigma_t^2 I\right). \quad (2)$$

The predicted noise can then be used to estimate the original clean image at any timestep. Using Tweedie’s formula, the denoised estimate is

$$\hat{x}_{0|t}(x_t, t, c) := \mathbb{E}[x_0 | x_t, c] = \frac{1}{\sqrt{\bar{\alpha}_t}} (x_t - \sqrt{1 - \bar{\alpha}_t} \epsilon_\theta(x_t, t, c)). \quad (3)$$

It therefore defines a differentiable mapping from the current noisy state to an estimate of the final clean image, which will later be used to construct the pullback geometry.

Inpainting as an inverse problem. In image restoration problems, only a partial or corrupted measurement y of image x_0 is available. For noiseless image inpainting, let $M \in \{0, 1\}^d$ denote the mask of the observed region, then $y = M \odot x_0$. Text guidance controls the semantic content through c , whereas the mask constrains the generated sample to agree with the visible region y .

3. Conditional Pullback Metric

In this section, we introduce a differential geometric framework for identifying directions in the conditioning space of diffusion models. In Euclidean spaces with a smooth map $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$, the pullback metric at $x \in \mathbb{R}^n$ evaluated at $v \in \mathbb{R}^n$ is defined as $\|J_{f(x)}v\|$. We therefore seek directions in the input space that induce proportionally larger shifts in the output space via the top right singular vectors of $J_{f(x)}$. We consider the Jacobian of the noise predictor with respect to condition:

$$J_c(x_t, t, c) = \frac{\partial \epsilon_\theta(x_t, t, c)}{\partial c}, \quad (4)$$

In the case of image inpainting, we can chain the linear masking process with Tweedie’s formula to define a Jacobian from the context to the masked portion of the image. Let $M_{edit} = 1 - M$ denote the bit-flipped mask of M , then we can define the linear operator W_M acting on image space by

$$W_M v = M_{edit} \odot v,$$

The differential of the masked clean estimate is

$$\frac{\partial (W_M \hat{x}_{0|t}(x_t, t, c))}{\partial c} = -\sqrt{\frac{1 - \bar{\alpha}_t}{\bar{\alpha}_t}} W_M J_c(x_t, t, c). \quad (5)$$

Using the conditional pullback metric, we can infer a subspace that locally produces large variability in the latent space. Explicitly, we consider span of the top K right singular vectors of Jacobian $\tilde{J}_c = W_M J_c$. Since explicitly constructing the condition Jacobian is computationally intractable, we use a sketching method to obtain these directions using only Jacobian-vector products dominant singular directions [8], for more details see Section B.1.

4. Pullback-informed prompt perturbation

With the top singular vectors, we can identify directions in the prompt embedding space along which small perturbations induce large changes in the denoising prediction and, consequently, in the predicted clean latent. We use these directions to diversify the reverse trajectories while preserving the semantic content of the original condition.

Initialization and trajectory diversification. We first sample a batch of B independent initial latent variables, $\mathcal{Z}_T = \{z_T^{(i)}\}_{i=1}^B$ with $z_T^{(i)} \sim \mathcal{N}(0, I)$, and use the same latent batch for all methods considered in our experiments. Given the base prompt c_0 , a pullback basis is then estimated from an anchor trajectory at the reference timestep t_b . The dominant eigenvectors of the condition-space metric $G_{t_b} = J_{c_0, t_b}^\top J_{c_0, t_b}$ are partitioned into mutually disjoint subspaces, one for each particle. Each particle is assigned a persistent direction obtained by projecting a fixed Gaussian field onto its corresponding subspace. The perturbations are mutually orthogonal and encourage different local variations of the conditional model. For particle i , the perturbed condition is therefore written as

$$c_i(t) = c_0 + \rho_i^* \gamma_t \Delta c_i,$$

where Δc_i denotes the prompt-relative pullback direction defined in Equation 7, γ_t is the time-dependent annealing coefficient, and ρ_i^* controls the particle-specific perturbation strength.

Scheduled perturbation. Let $S > E$ denote the perturbation window in reverse-diffusion timestep coordinates. We define the annealing schedule as a hard-sigmoid raised to a power $p > 0$, namely

$$\gamma(t; S, E) = \max \left(0, \min \left(1, \frac{t - E}{S - E} \right) \right)^p.$$

The perturbation is fully active at the beginning of the reverse process, progressively annealed between S and E , and completely removed afterward. For particle i , the condition used at timestep t is $c_{i,t} = c_0 + \rho_i^* \gamma(t; S, E, p) \Delta c_i$ where Δc_i is its assigned pullback direction.

Table 1: Generation results on MSCOCO 2017 validation captions with five generated images per caption. Stable Diffusion 1.5 and XL results use all 500 captions. Entries report mean $\pm$ standard deviation across captions. The best mean within each backbone is shown in bold.

	Alignment	Diversity		
Method	CLIP $\uparrow$	MSS $\downarrow$	Vendi $\uparrow$	DINO Mean $\downarrow$
<i>Stable Diffusion 1.5</i>				
Clean DDIM	31.3957 $\pm$ 2.4730	0.3359 $\pm$ 0.0519	4.6831 $\pm$ 0.1971	0.5945 $\pm$ 0.1638
CADS	31.2537 $\pm$ 2.4253	0.3066 $\pm$ 0.0433	4.7833 $\pm$ 0.1426	0.5469 $\pm$ 0.1628
TPSO	30.9592 $\pm$ 2.2941	0.3048 $\pm$ 0.0430	4.7838 $\pm$ 0.1464	0.5143 $\pm$ 0.1729
Pullback (ours)	31.3153 $\pm$ 2.4671	0.3194 $\pm$ 0.0501	4.7377 $\pm$ 0.1833	0.5663 $\pm$ 0.1661
<i>Stable Diffusion XL</i>				
Clean DDIM	32.0012 $\pm$ 2.7271	0.3827 $\pm$ 0.0583	4.4979 $\pm$ 0.2537	0.6899 $\pm$ 0.1571
CADS	31.7740 $\pm$ 2.5822	0.3485 $\pm$ 0.0532	4.6320 $\pm$ 0.2089	0.6374 $\pm$ 0.1667
TPSO	31.7017 $\pm$ 2.5535	0.3517 $\pm$ 0.0546	4.6141 $\pm$ 0.2192	0.6263 $\pm$ 0.1728
Pullback (ours)	31.4248 $\pm$ 2.5511	0.3196 $\pm$ 0.0470	4.7309 $\pm$ 0.1636	0.5639 $\pm$ 0.1785

Adaptive pullback refresh. The pullback basis computed at the reference timestep t_b only captures the local geometry around the current state. Since this geometry evolves along the reverse trajectory, we update the basis at P intermediate timesteps:

$$t_r^{\text{ref}} \approx S + \frac{r}{P+1}(E - S), \quad r = 1, \dots, P.$$

At each refresh timestep, we recompute the dominant eigenspace of the local pullback metric using the current latent of the anchor particle, and reproject each persistent Gaussian field η_i onto its refreshed particle-specific subspace. The previous and refreshed directions are interpolated over a short transition window to avoid abrupt changes in the conditioning trajectory 8.

Particle-specific scale selection. The sensitivity of the condition to response map varies along the diffusion landscape. Consequently, a single perturbation scale may be too weak for some particles and too strong for others. We therefore select a particle’s specific scale ρ_i^* from a finite candidate set $\mathcal{R} = \{\rho_1, \dots, \rho_M\}$ by using the selection method in B.3.

5. Experiments & Results

For text-to-image generation, we use captions from the MSCOCO 2017 validation [14]. For image inpainting, we use the BrushBench validation set introduced with BrushNet [13]. We generate using the Stable Diffusion family of models, comparing our pullback method against baselines of DDIM, CADS [21], and TPSO [16] with standard metrics of prompt alignment and diversity. Full details are given in Section C. Tables 1 and 2 summarize the results for image generation and text-guided inpainting, respectively. Across both tasks, pullback-guided sampling consistently improves diversity over clean DDIM while maintaining a high level of alignment with the text condition. The results are particularly good for image inpainting, where the editable region constrains the possible variations and therefore makes the pullback metric a more significant initial signal.



Figure 1: Qualitative comparison on a BrushBench example with BrushNet-SDXL. Each row contains masked samples (see Figure 2 in the Appendix). Clean DDIM, CADS, and TPSO introduce local changes, whereas adaptive pullback produces more pronounced variations.

Table 2: BrushBench inpainting results for Stable Diffusion 1.5 and Stable Diffusion XL, with five images per example. Stable Diffusion 1.5 and XL results use all 500 examples. Entries report mean $\pm$ standard deviation across examples. The best result within each backbone is shown in bold.

Method	Alignment	Diversity		
	CLIP $\uparrow$	DINO $\downarrow$	MSS $\downarrow$	Vendi $\uparrow$
<i>Stable Diffusion 1.5</i>				
Clean DDIM	32.0003 $\pm$ 2.7383	0.8930 $\pm$ 0.1094	0.9643 $\pm$ 0.0323	1.1746 $\pm$ 0.1441
CADS	31.9284 $\pm$ 2.7278	0.8736 $\pm$ 0.1215	0.9594 $\pm$ 0.0367	1.1973 $\pm$ 0.1642
TPSO	31.9825 $\pm$ 2.7077	0.8807 $\pm$ 0.1174	0.9621 $\pm$ 0.0336	1.1844 $\pm$ 0.1490
Pullback (ours)	31.8796 $\pm$ 2.8055	0.8504 $\pm$ 0.1336	0.9498 $\pm$ 0.0414	1.2380 $\pm$ 0.1817
<i>Stable Diffusion XL</i>				
Clean DDIM	32.2147 $\pm$ 2.7413	0.9077 $\pm$ 0.1001	0.9701 $\pm$ 0.0285	1.1488 $\pm$ 0.1273
CADS	32.0407 $\pm$ 2.7273	0.8920 $\pm$ 0.1102	0.9656 $\pm$ 0.0308	1.1692 $\pm$ 0.1373
TPSO	32.1414 $\pm$ 2.7285	0.9014 $\pm$ 0.1043	0.9679 $\pm$ 0.0306	1.1585 $\pm$ 0.1372
Pullback (ours)	32.1108 $\pm$ 2.7274	0.8788 $\pm$ 0.1246	0.9604 $\pm$ 0.0343	1.1918 $\pm$ 0.1519

6. Conclusion

We introduced a training-free method for diversity conditional diffusion models through pullback-informed prompt perturbations. Our method identifies locally expressive directions in the conditioning space that induce the largest variation in the image space, by estimating the dominant directions of the pullback operator. These results demonstrate that the local geometry of the denoiser with respect to condition provides useful directions for diversifying generation without fine-tuning. Future work will evaluate the effectiveness of pullback directions on other constrained generation tasks.

References

- [1] Pietro Astolfi, Marlene Careil, Melissa Hall, Oscar Mañas, Matthew Muckley, Jakob Verbeek, Adriana Romero Soriano, and Michal Drozdal. Consistency-diversity-realism pareto fronts of conditional image generative models. *arXiv preprint arXiv:2406.10429*, 2024.
- [2] Yuanhao Ban, Ruochen Wang, Tianyi Zhou, Boqing Gong, Cho-Jui Hsieh, and Minhao Cheng. The crystal ball hypothesis in diffusion models: Anticipating object positions from initial noise. In *International Conference on Learning Representations*, volume 2025, pages 34384–34408, 2025.
- [3] 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, pages 22480–22507, 2024.
- [4] Riccardo De Santi, Marin Vlastelica, Ya-Ping Hsieh, Zebang Shen, Niao He, and Andreas Krause. Provable maximum entropy manifold exploration via diffusion models. *arXiv preprint arXiv:2506.15385*, 2025.
- [5] Carles Domingo i Enrich, Michal Drozdal, Brian Karrer, and Ricky TQ Chen. Adjoint matching: Fine-tuning flow and diffusion generative models with memoryless stochastic optimal control. In *International Conference on Learning Representations*, volume 2025, pages 53791–53846, 2025.
- [6] Luca Eyring, Shyamgopal Karthik, Karsten Roth, Alexey Dosovitskiy, and Zeynep Akata. Reno: Enhancing one-step text-to-image models through reward-based noise optimization. *Advances in Neural Information Processing Systems*, 37:125487–125519, 2024.
- [7] Dan Friedman and Adji Bousso Dieng. The vendi score: A diversity evaluation metric for machine learning. *arXiv preprint arXiv:2210.02410*, 2022.
- [8] Nathan Halko, Per-Gunnar Martinsson, and Joel A Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review*, 53(2):217–288, 2011.
- [9] Anne Harrington, A Koepke, Shyamgopal Karthik, Trevor Darrell, and Alexei A Efros. It’s never too late: Noise optimization for collapse recovery in trained diffusion models. *arXiv preprint arXiv:2601.00090*, 2025.
- [10] 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*, pages 7514–7528, 2021.
- [11] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.
- [12] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

- [13] Xuan Ju, Xian Liu, Xintao Wang, Yuxuan Bian, Ying Shan, and Qiang Xu. Brushnet: A plug-and-play image inpainting model with decomposed dual-branch diffusion. In *European Conference on Computer Vision*, pages 150–168. Springer, 2024.
- [14] 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*, pages 740–755. Springer, 2014.
- [15] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11461–11471, 2022.
- [16] Debin Meng, Chen Jin, Zheng Gao, Yanran Li, Ioannis Patras, and Georgios Tzimiropoulos. Training-free generation of diverse and high-fidelity images via prompt semantic space optimization. *arXiv preprint arXiv:2511.19811*, 2025.
- [17] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.
- [18] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khali-dov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.
- [19] Yong-Hyun Park, Mingi Kwon, Jaewoong Choi, Junghyo Jo, and Youngjung Uh. Understanding the latent space of diffusion models through the lens of riemannian geometry. *Advances in Neural Information Processing Systems*, 36:24129–24142, 2023.
- [20] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.
- [21] 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, pages 23723–23755, 2024.
- [22] Chitwan Saharia, William Chan, Huiwen Chang, Chris A Lee, Jonathan Ho, Tim Salimans, David J Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. *arXiv preprint arXiv:2111.05826*, 2021.
- [23] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.
- [24] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

- [25] 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.
- [26] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3836–3847, 2023.
- [27] Guangcong Zheng, Xianpan Zhou, Xuewei Li, Zhongang Qi, Ying Shan, and Xi Li. Layout-diffusion: Controllable diffusion model for layout-to-image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22490–22499, 2023.

Appendix A. Related works

A.1. Image Diffusion Models

Conditional image generation Conditional image generation aims to synthesize realistic images that satisfy user-specified controls. These conditions may take the form of textual descriptions, reference images, class labels, segmentation maps, depth maps, or spatial layouts. Diffusion models have become a dominant framework for this task because of their strong image quality and flexibility in incorporating heterogeneous conditions. Text-conditioned latent diffusion models introduce semantic information through cross-attention, enabling high-resolution text-to-image synthesis [20], while classifier-free guidance strengthens alignment by combining conditional and unconditional denoising predictions [11]. Other approaches introduce image-to-image conditioning [22] or additional spatial controls such as edges, poses, depth maps, and layouts [26, 27]. These developments have substantially improved the controllability and fidelity of conditional image synthesis.

Image inpainting Image inpainting consists in reconstructing a missing region of an image while preserving the visible content and maintaining semantic and visual coherence across the edited region. Diffusion-based approaches address inpainting either by training models that explicitly receive the masked image and mask as conditioning inputs or by adapting the reverse process of a pretrained generative model [13, 15, 17]. These approaches aim to generate plausible content within the masked region while preserving the observed pixels and maintaining semantic and visual coherence with the unmasked region.

Diversity image generation Although diffusion models generate samples from random initial noise, stochasticity alone does not guarantee meaningful diversity under a fixed condition. Different initializations may still converge toward semantically similar outputs, particularly when strong guidance, textual constraints, or spatial observations restrict the reverse process. Recent analyses therefore emphasize a trade-off between realism, condition consistency, and distributional coverage [1], while other work shows that the initial noise can already strongly determine the spatial organization of the final image [2]. Consequently, producing multiple distinct yet equally valid completions remains challenging, especially for constrained tasks such as text-guided inpainting. Existing methods address this limitation through additional training (finetuning diversity methods), condition perturbations, or interactions between sampling trajectories.

A.2. Fine-tuning diversity methods

Fine-tuning methods enable further training towards a target reward or distributional loss. For a scalar reward, Adjoint Matching [5] backpropagates through the entire dynamics to improve the reward while staying close to the original reverse process. This has been generalized as a tool to promote diversity in the generated distribution in De Santi et al. [4]. Notably, these methods require expensive sampling through the diffusion model and are therefore no longer simulation-free.

A.3. Inference-Time Diversity Methods

Inference-time diversity methods aim to improve the coverage of the conditional distribution without retraining the underlying diffusion model. Existing approaches differ primarily in the component of the sampling process on which they intervene. Condition-space methods modify the conditioning signal supplied to the denoising network. CADS [21] perturbs the condition embedding with

Gaussian noise whose magnitude is progressively annealed throughout sampling, thereby increasing variation during the early stages of generation. TPSO [16] instead explores underutilized directions of the token-embedding space through optimized prompt perturbations, while constraining the resulting samples to preserve semantic consistency. Some family directly modifies the source of stochasticity by optimizing the initial diffusion noise [6, 9], an others instead construct a joint, non-i.i.d. distribution sample [3]. Particle Guidance introduces pairwise repulsive interactions during denoising to prevent particles from converging toward the same modes [3].

Appendix B. Pullback-Informed Prompt Perturbation

Let $C_0^+ \in \mathbb{R}^{L \times d_c}$ denote the positive token condition, where d_c is the token-embedding dimension. After vectorization, the condition belongs to $\mathbb{R}^D$, with $D = Ld_c$.

B.1. Matrix-free estimation of expressive directions

We first follow the clean CFG-DDIM trajectory of one reference particle to a basis-estimation timestep t_b , obtaining $z_{t_b}^{\text{ref}}$. At the fixed state $(z_{t_b}^{\text{ref}}, t_b, C_0^+)$, we seek directions in the conditioning space that induce the largest local variation in the denoiser output. For a unit direction v , this variation can be measured by

$$\frac{1}{\varepsilon^2} \|f_b(C_0^+ + \varepsilon v) - f_b(C_0^+)\|_2^2, \quad \|v\|_2 = 1,$$

where $f_b(C) = \epsilon_\theta(z_{t_b}^{\text{ref}}, t_b, C)$. Using the first-order approximation $f_b(C_0^+ + \varepsilon v) - f_b(C_0^+) \simeq \varepsilon J_b v$, the problem of finding the K most expressive orthonormal directions becomes

$$V^* = \arg \max_{V^\top V = I_K} \|J_b V\|_F^2 = \arg \max_{V^\top V = I_K} \text{tr}(V^\top J_b^\top J_b V). \quad (6)$$

Therefore, by the Rayleigh–Ritz theorem, the columns of V^* are the K dominant eigenvectors of

$$G_b = J_b^\top J_b, \quad J_b = \left. \frac{\partial f_b}{\partial C} \right|_{C=C_0^+}.$$

Equivalently, they are the dominant right singular vectors of J_b . We estimate these directions without explicitly constructing J_b .

$$J_b v \approx \frac{f_b(C_0^+ + \varepsilon v) - f_b(C_0^+ - \varepsilon v)}{2\varepsilon}, \quad G_b v = J_b^\top (J_b v),$$

where the second operation is evaluated with a vector–Jacobian product.

To approximate the K dominant eigenvectors, we use randomized block subspace iteration followed by Rayleigh–Ritz extraction [8]. We sample $\Omega \in \mathbb{R}^{D \times K}$ with Gaussian entries and set

$$Q_0 = \text{QR}(\Omega), \quad Q_{r+1} = \text{QR}(G_b Q_r), \quad r = 0, \dots, R-1.$$

We then construct

$$B = \frac{1}{2} \left(Q_R^\top G_b Q_R + (Q_R^\top G_b Q_R)^\top \right),$$

diagonalize $B = U \Lambda U^\top$, sort the eigenvalues in decreasing order, and lift the resulting vectors back to condition space:

$$V = Q_R U = [v_1, \dots, v_K].$$

B.2. Fixed disjoint perturbation directions

The basis indices are divided into balanced, mutually disjoint sets $\mathcal{I}_1, \dots, \mathcal{I}_N$. For each particle, a Gaussian field $\xi_i \sim \mathcal{N}(0, I_D)$ is sampled once and kept fixed. Its projection onto the assigned pullback subspace is

$$\bar{d}_i = \sum_{k \in \mathcal{I}_i} \langle \xi_i, v_k \rangle v_k, \quad d_i = \frac{\bar{d}_i}{\|\bar{d}_i\|_2}.$$

To express the perturbation relative to the magnitude of the original condition, each real token is scaled by its clean embedding norm:

$$\Delta c_i = \|C_0^+\|_2 d_i, \quad (7)$$

where $\|C_0^+\|_2$ denote the norms of the token embeddings.

Adaptive pullback refresh For the refresh, the pullback basis and the corresponding directions are recomputed at intermediate denoising steps. Let d_i^{old} and d_i^{new} denote the directions before and after recomputation. To avoid an abrupt change, we interpolate between them over $L = 2$ steps:

$$d_i^{(\ell)} = \frac{(1 - \beta_\ell) d_i^{\text{old}} + \beta_\ell d_i^{\text{new}}}{\|(1 - \beta_\ell) d_i^{\text{old}} + \beta_\ell d_i^{\text{new}}\|_2}, \quad \beta_\ell = \frac{\ell}{2}, \quad \ell \in \{1, 2\}. \quad (8)$$

B.3. Particle-specific magnitude selection

We select a particle’s specific magnitude from a finite candidate set $\mathcal{R} = \{\rho_1, \dots, \rho_M\}$ at a specific timestep t_p before the final sampling run.

For each particle i and candidate magnitude ρ_m , we compute the predicted clean image $\hat{x}_{i,m}$ at t_p . We additionally use $\rho_0 = 0$ to obtain the unperturbed reference $\hat{x}_{i,0}$. Semantic diversity between two candidates is measured using the cosine distance between their DINO representations:

$$d_{\text{DINO}}(x, x') = 1 - \cos(\phi_{\text{DINO}}(x), \phi_{\text{DINO}}(x')), \quad (9)$$

where ϕ_{DINO} is a frozen DINO image encoder. Prompt alignment is measured by

$$s_{\text{CLIP}}(x, C) = \cos(\phi_{\text{CLIP}}^I(x), \phi_{\text{CLIP}}^T(C)), \quad (10)$$

where ϕ_{CLIP}^I and ϕ_{CLIP}^T are the frozen CLIP image and text encoders, respectively.

The particle-specific magnitudes are jointly selected by maximizing the average pairwise DINO distance while limiting the loss of prompt alignment relative to the unperturbed prediction:

$$\begin{aligned} (m_1^*, \dots, m_B^*) \in \arg \max_{\substack{m_1, \dots, m_B \\ m_i \in \{1, \dots, M\}}} & \frac{2}{B(B-1)} \sum_{i < j} d_{\text{DINO}}(\hat{x}_{i,m_i}, \hat{x}_{j,m_j}) \\ \text{subject to} & \quad s_{\text{CLIP}}(\hat{x}_{i,m_i}, C) \geq s_{\text{CLIP}}(\hat{x}_{i,0}, C) - \delta_{\text{CLIP}}, \quad i = 1, \dots, B. \end{aligned} \quad (11)$$

The selected magnitudes are $\rho_i^* = \rho_{m_i^*}$.

B.4. Scheduled reverse sampling

After selecting ρ_i^* , sampling is restarted from the original initial latents. The perturbation envelope is

$$\alpha(t) = \begin{cases} 1, & t \geq T_S, \\ \left(\frac{t - T_E}{T_S - T_E} \right)^p, & T_E < t < T_S, \\ 0, & t \leq T_E. \end{cases}$$

At each reverse step, the particle-specific positive condition is

$$C_{i,t}^+ = C_0^+ + \alpha(t)\rho_i^* \Delta_{C_{i,t}} \quad C_{i,t}^- = C_0^-.$$

The resulting pair is used in the standard CFG prediction, followed by one DDIM update.

Appendix C. Experimental Details

Tasks and datasets For complete text-to-image generation, we use captions from the MSCOCO 2017 validation [14]. For image inpainting, we use the BrushBench validation set introduced with BrushNet [13]. Each inpainting example contains a source image, a binary edit mask, and a text condition describing the expected content. We evaluate 500 conditions from each benchmark and generate five images per condition, resulting in 2,500 generated images for every method on each task.

Configuration Complete generation is performed with Stable Diffusion 1.5. The inpainting evaluation uses the corresponding Stable Diffusion 1.5 and XL BrushNet backbone [13]. We use 50 DDIM sampling steps, classifier-free guidance with scale 7.5, and $\eta = 0$. We follow the publicly available default configurations of the corresponding models.

For every condition, all compared methods receive exactly the same initial latent particles. All experiments use fixed random seeds to ensure fair and model-consistent comparisons.

Compared methods We compare the following sampling procedures:

- **Clean DDIM**: standard DDIM sampling without condition perturbation;
- **CADS**: condition annealing with isotropic Gaussian perturbations [21];
- **TPSO**: token-prompt embedding space optimization applied during sampling [16];
- **Pullback**: our method, which perturbs the condition along dominant pullback directions and uses particle-specific perturbation magnitudes.

C.1. Evaluation Metrics

The evaluation measures three complementary properties: semantic alignment, perceptual diversity, and set-level diversity.

Text-image alignment. Text-image consistency is evaluated using the CLIP Score [10], defined as the cosine similarity between CLIP embeddings of the generated image and its corresponding prompt. Higher values indicate stronger semantic alignment with the prompt.

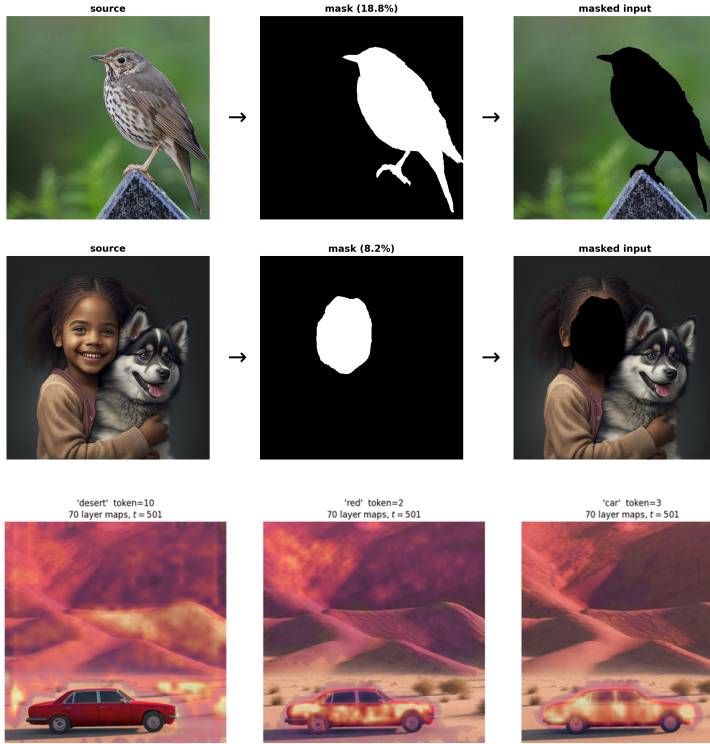


Figure 2: Illustration of the text-guided inpainting task. Given a text prompt, the model regenerates the masked region while preserving the unmasked region.

The prompts are “a small bird perched on top of a rock” (top) and “a young girl holding a husky dog on a black background” (bottom).

Prompt.

“a red car with tinted windows driving through the desert, photo-realistic, 8k”

Figure 3: Token-dependent cross-attention at DDIM timestep $t = 499$. Each visualization averages 28 U-Net cross-attention maps over attention heads and layers. Warmer regions receive greater relative attention for the corresponding token.

Semantic diversity. We extract DINOv2 image representations [18] and compute the mean pairwise cosine similarity within each set of samples: Lower mean similarity indicates greater overall semantic diversity.

Set-level diversity. We report the Mean Similarity Score (MSS) and Vendi Score [7] for per-prompt (intra-set) diversity, both computed using SSCD features following [21]. Lower MSS and higher Vendi indicates greater within-condition diversity.

Appendix D. Further Experimental Details

D.1. Feature-space and conditioning-space pullback directions

Park et al. [19] use the pullback metric to identify locally sensitive directions in the noisy latent space of a diffusion model. Their construction relies on the observation that the intermediate representations of the denoising model encode semantic and spatial information about the generated image. Let

$$f_t(\cdot; c) : \mathcal{X}_t \rightarrow \mathcal{H}_t \quad (12)$$

denote the mapping from the noisy latent x_t to a selected intermediate U-Net feature representation $h_t = f_t(x_t; c)$ under condition c . Its Jacobian and induced pullback metric are

$$J_x = \frac{\partial f_t(x_t; c)}{\partial x_t}, \quad G_x = J_x^\top J_x. \quad (13)$$

The dominant eigenvectors $\{v_k\}$ of G_x , identify input directions that produce large local variations in the feature representation.

For direction v_k and strength s , we construct the norm-matched latent perturbation

$$\delta x_{t,k}(s) = s \frac{\|x_t\|_2}{\|v_k\|_2} v_k. \quad (14)$$

Following the x -space guidance formulation of Park et al. [19], the perturbed latent is

$$\tilde{x}_{t,k}(s) = x_t + \gamma [\epsilon_\theta(x_t + \delta x_{t,k}(s), t, c) - \epsilon_\theta(x_t, t, c)], \quad (15)$$

where γ controls the magnitude of the noise-prediction response.

We compare this construction with the conditioning-space pullback. For a fixed latent x_t , we consider the condition-response mapping

$$g_t(C) = \epsilon_\theta(x_t, t, C), \quad (16)$$

with Jacobian and pullback metric

$$J_C = \frac{\partial \epsilon_\theta(x_t, t, C)}{\partial C}, \quad G_C = J_C^\top J_C. \quad (17)$$

The dominant eigenvectors $\{d_k\}$ of G_C identify condition directions that induce large changes in the predicted noise.

For both constructions, we compute a rank-4 approximation of the pullback metric at the same latent and retain the four dominant directions. We keep others parameters same for both. For the latent-feature pullback, we evaluate

$$s \in \{0.005, 0.015, 0.02, 0.025, 0.03, 0.04, 0.05\}, \quad (18)$$

with $\gamma = 5$. For the conditioning pullback, we evaluate

$$s \in \{0.1, 0.25, 0.5, 1.0, 1.5, 2.0, 2.5\}. \quad (19)$$

The numerical values of s are not directly comparable across the two methods because the perturbations operate in different spaces with different sensitivity scales. We therefore select a specific range of values that clearly reveals the progressive effect of the perturbation on the generated image. The experiment compares how the output evolves within each method as the perturbation strength increases.

Qualitative observations. The latent-feature pullback primarily produces localized variations and does not induce substantial semantic diversity. Increasing the perturbation strength does not necessarily amplify these variations in a meaningful way; instead, it progressively introduces high-contrast contours, geometric distortions, and other off-manifold artifacts. In contrast, the conditioning-space pullback produces broader semantic transformations that affect a larger portion of the image. These transformations remain direction-dependent across the strength sweep, with different pullback directions controlling distinct semantic attributes.

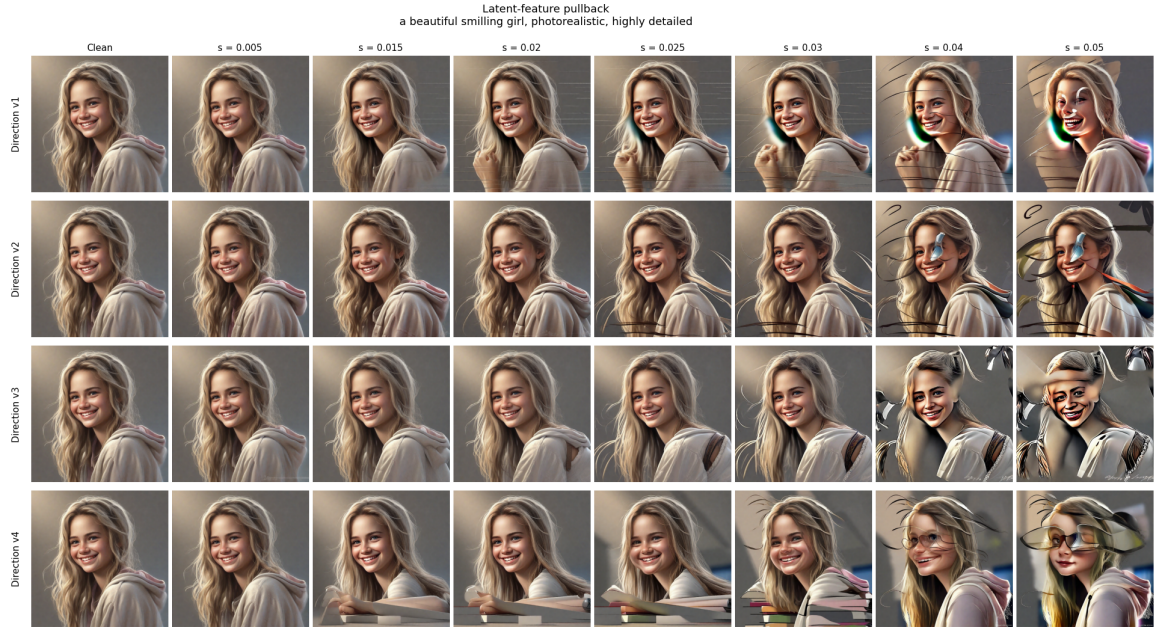


Figure 4: Effect of the perturbation strength along latent-feature pullback directions. Rows correspond to the four dominant directions $v_1, \dots, v_4$ of the pullback metric $G_x = J_x^\top J_x$, where J_x is the Jacobian of the selected U-Net feature representation with respect to the noisy latent x_t . Columns show the clean sample followed by increasing strengths. Moderate strengths produce predominantly localized, whereas larger strengths introduce visible distortions.

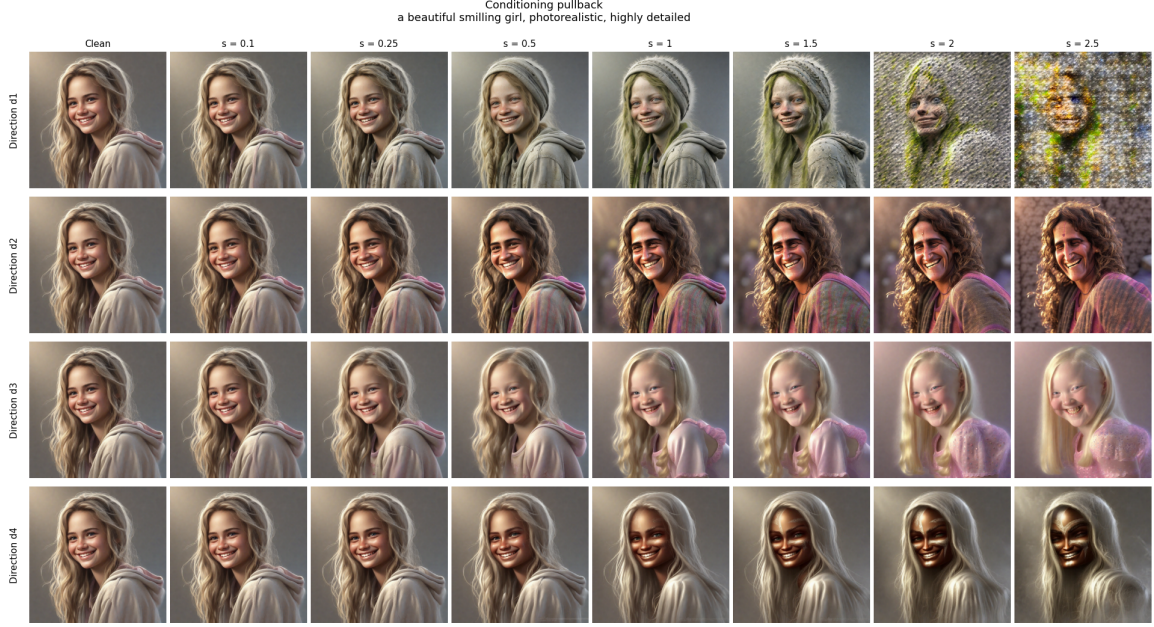


Figure 5: Effect of the perturbation strength along conditioning-space pullback directions. Rows correspond to the four dominant directions of the condition pullback metric $G_C = J_C^\top J_C$, where $J_C = \partial \epsilon_\theta(x_t, t, C) / \partial C$. Columns show the clean sample followed by increasing strengths. The different directions produce distinct global semantic transformations, while large strengths can reduce image quality and fidelity.

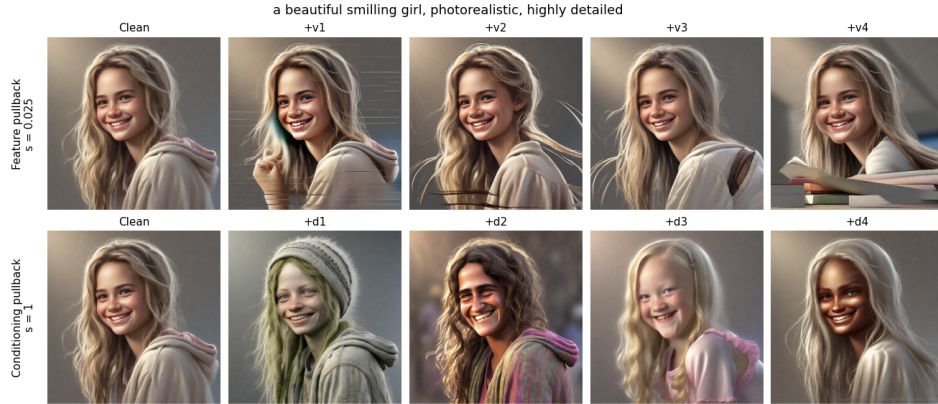


Figure 6: Qualitative comparison between latent-feature and conditioning-space pullback directions computed from the same anchor. The top row shows the clean sample and the four dominant latent-feature directions at $s = 0.025$, using an x -space guidance scale of $\gamma = 5$. The bottom row shows the clean sample and the four dominant conditioning directions at $s = 1$. In this example, the latent-feature directions primarily affect localized appearance, pose, and nearby image structure, whereas the conditioning-space directions induce broader semantic changes. All results use the same prompt, initial noise, and reverse-process parameters.