

It's Never Too Late: Noise Optimization for Collapse Recovery in Trained Diffusion Models

Anne Harrington^{1*} A. Sophia Koepke^{1,2,3*} Shyamgopal Karthik²
 Trevor Darrell¹ Alexei A. Efros¹

¹ UC Berkeley ²University of Tübingen, Tübingen AI Center ³TU Munich, MCML [*Equal contribution]



Figure 1. Repeatedly sampling from text-to-image models using a fixed text prompt produces surprisingly little visual variation (top row) in both Stable Diffusion SDXL-Turbo [52] (left) and Flux.1 [schnell] [35] (right). Our approach (bottom row) directly optimizes the initial noise to recover from mode collapse, producing diverse outputs.

Abstract

Contemporary text-to-image models exhibit a surprising degree of mode collapse, as can be seen when sampling several images given the same text prompt. Previous work has attempted to address this issue by steering the model using guidance mechanisms, or by generating a large pool of candidates and refining them. In this work, we take a different direction and aim for diversity in generations via noise optimization. Specifically, we show that a simple noise optimization objective can mitigate mode collapse while preserving the fidelity of the base model. We also analyze the frequency characteristics of the noise and show that alternative noise initializations with different frequency profiles can improve both optimization and search. Our experiments demonstrate that noise optimization yields superior results in terms of generation quality and diversity. Code is available at: <https://github.com/anneharrington/divgen>

1. Introduction

Diffusion models can generate stunning images, yet, when asked to create multiple outputs given a fixed prompt, they

often produce nearly identical results over and over again across different random seeds. Figure 1 illustrates this issue, with the top row showing strikingly similar generations (e.g. of a cat). For many tasks, we need not only generation quality but also a diversity in outputs that capture the range of possible images per prompt.

At the same time, inference-time scaling has become widespread in diffusion models. The key premise of this line of work is to utilize additional compute during inference to tackle challenging problems which could not otherwise be successfully solved. In the context of diffusion models, inference-time scaling has been used with great success to improve prompt adherence [13, 40, 61] and personalization [46, 47].

Based on these insights, several inference-time approaches for improving the diversity of images generated with diffusion models have emerged. One popular approach has been to utilize guidance strategies to steer the model towards generating varied samples [11, 48, 54]. Alternatively, generating a large number of candidates and iteratively pruning them to optimize for increasing variety has recently shown success [42]. This highlights that the initial

noise inputs can play a crucial role in obtaining varied sets of generated images, if you are willing to “roll the dice” enough times. But what if, instead of just waiting for some random seed to yield a generated image with specific properties, we were able to directly optimize the input noise to satisfy desired properties [13].

In this paper, we design an end-to-end noise optimization strategy to maximize the diversity in sets of generated images, as described in Sec. 3. Specifically, we sample a batch of initial noise samples. We then directly optimize these by minimizing a pairwise similarity metric that drives samples apart. Our method outperforms prior works across multiple diffusion models and benchmarks (see Sec. 4). We demonstrate that we can flexibly select different optimization objectives that facilitate diversity in generated outputs (e.g. DINOv2 [41], LPIPS [66], DreamSim [16]). Further, we also investigate the usage of set-level diversity objectives such as Determinantal Point Processes (DPP) [12] and Vendi Score [15] and find that they are more suitable to provide increased variation backed by user studies.

In addition, we analyze how the initial noise evolves during optimization and specifically how this impacts different frequency bands (Sec. 4.1). Inspired by these observations, we explore boosting low-frequency components in the noise initialization, using pink noise, to increase output diversity. Pink noise initializations consistently improve the diversity of generated samples not only for our approach, but also the baselines we compare to for all evaluated models.

Overall, we introduce an end-to-end noise optimization framework that achieves superior output diversity compared to prior methods. Our framework supports the flexible selection of diversity optimization objectives. We find set-level diversity objectives to be most effective and identify low-frequency noise components as key drivers of diverse image generation.

2. Related Work

Inference-Time Scaling. Test-time scaling allocates additional computation during inference to solve challenging problems. Beyond scaling denoising steps in diffusion models, test-time techniques improve generation quality by finding better initial noise or refining intermediate states, often guided by pre-trained reward models. These methods fall into two categories: search-based approaches [27, 36, 59, 60] that evaluate multiple candidates, and optimization-based approaches [7, 19, 28, 40, 56, 61] that iteratively refine noise or latents through gradient descent. In the context of increasing the diversity in the outputs of the generative model, Parmar et al. [42] proposed an efficient search strategy using intermediate generations as a proxy for the final images. Differently, in this work, we demonstrate that an end-to-end noise optimization strategy along with changing the noise initialization achieves superior performance on the

quality-diversity tradeoff.

Guidance Mechanisms. Drawing from the success of classifier-free guidance (CFG) mechanisms [10, 22] in steering diffusion models towards desired objectives, several variations have been proposed to either improve the effectiveness of CFG [1, 3, 32], or reduce its computational complexity [2, 26, 33]. To increase the diversity when multiple outputs are sampled, several alternatives have been proposed [29, 54], including the usage of particle guidance [11] and DPP [31, 39]. These methods use guidance mechanism to balance the tradeoff between quality and diversity [25, 48, 49]. Unlike guidance methods that steer the model toward a particular target through modified conditioning, SliderSpace [17] encodes semantic directions of variation into LoRA [23] weights, enabling user-controlled diversity without inference-time overhead. Different to those approaches, we directly optimize the initial noise with target diversity and quality objectives.

Prompt Augmentations. Improving controllability in generation by modifying the textual conditioning input rather than the diffusion dynamics [20, 38] has also been a popular direction. These methods try to explicitly improve quality and/or diversity using LLMs to rewrite prompts for diffusion models [5, 37]. Our approach is orthogonal to these methods: while prompt refinements improve the semantic conditioning, some variations in the output space cannot be captured easily by text alone.

Effect of Initial Noise in Generation. Several works have explored the controllability of the generation process through initial noise [13, 19, 51, 55]. It has been observed that specific noise seeds control global behavior [65]. However, the most popular approach is to utilize best-of-n sampling approaches [27, 30, 36, 44, 45] or direct noise optimization approaches. In this work, we show that directly optimizing the initial noise can be used as an effective tool to improve the diversity of generations in pre-trained diffusion models. Furthermore, we demonstrate that directly altering the frequency patterns of the initial noise itself impacts the diversity of outputs. This is motivated by an analysis of our noise optimization process and prior work demonstrating that low frequency information at initialization can enhance video diffusion [62] and determine object placement in text-to-image models [6].

3. Collapse Recovery in Diffusion Models

Preliminaries. Diffusion models interpolate between noise $z = \mathbf{x}_0 \sim \mathcal{N}(0, \mathbf{I})$ and data $\mathbf{x}_1 \sim p_{data}$, such that $\mathbf{x}_t = \alpha_t \mathbf{x}_0 + \sigma_t \mathbf{x}_1$, where α_t decreases and σ_t increases with $t \in [0, 1]$. A generative model $g_\theta(z, c)$ is obtained by simulating a learned differential equation from $\mathbf{x}_0 \sim p_0$, conditioned on a prompt c , to produce an image $\mathbf{x}$.

Test-time optimization techniques aim to improve pre-

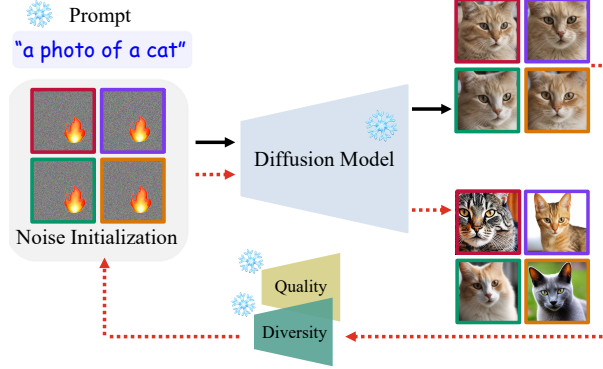


Figure 2. Noise optimization pipeline for diverse image generation. We optimize the noise initialization to increase visual diversity given a fixed text prompt and diffusion model. Starting from i.i.d. noise samples, we generate a set of images. Using a diversity objective (e.g. DINOv2 dissimilarity) and optionally a quality reward (e.g. HPSv2), we update the noise to produce output images that capture more diversity per text prompt. Our method supports optimizing over a variety of objective ensembles.

trained generative models on a per-sample basis at inference. A popular gradient-based strategy is test-time noise optimization [7, 19, 28, 40, 56, 61]. Given a pre-trained generator g_θ (e.g. a diffusion or flow matching model), this approach optimizes the initial noise $\mathbf{x}_0$ for each generated instance. The objective is to find an improved $\mathbf{x}_0^*$ that maximizes a given reward $r(g_\theta(\mathbf{x}_0))$, subject to regularization and can be formulated as

$$\mathbf{x}_0^* = \arg \max_{\mathbf{x}_0} (r(g_\theta(\mathbf{x}_0)) - \text{reg}(\mathbf{x}_0)), \quad (1)$$

where $\text{reg}(\mathbf{x}_0)$ is a regularization term designed to keep $\mathbf{x}_0^*$ within a high-density region of the prior noise distribution p_0 . These methods are designed to improve the quality of a single sample [13], as opposed to our objective of increasing the diversity in multiple generated outputs. We build on this approach for achieving this goal.

Increasing Diversity through Noise Optimization. We propose a noise optimization approach for collapse recovery as shown in Fig. 2. We start with a text-to-image diffusion model that takes as input a noise initialization and text prompt. From these inputs, we generate an output set of images. We then compute diversity and quality scores over the images using objective functions such as DINOv2[41] and image rewards such as HPSv2[63]. We use these scores to backpropagate to the initial noise, optimizing for higher output diversity without quality degradation. We iteratively optimize the noise until a user-specified compute budget or diversity/quality threshold is met. Critically, we keep the starting prompt, diffusion model, and objective/reward models frozen and only update the noise. This enables us to increase diversity without altering the model or user input.

Formally, given a prompt c , we draw a batch $\mathcal{B} = \{\mathbf{x}_0^{(i)}\}_{i=1}^B$ with $\mathbf{x}_0^{(i)} \sim \mathcal{N}(0, \mathbf{I})$ and generate $\mathbf{x}^{(i)} = g_\theta(\mathbf{x}_0^{(i)}, c)$. We jointly optimize the batch to meet two targets: (i) sample-level quality via a reward $r_s(\mathbf{x}^{(i)}, c)$ such as CLIPScore, and (ii) high batch-level diversity via a statistic v_B computed from pairwise or set-based (patch) features (e.g. using DINOv2). Let τ_s and τ_D be target thresholds for quality and diversity. We minimize a hinge-penalized diversity and quality objective

$$\begin{aligned} \mathcal{L}(\mathcal{B}) = & -\frac{1}{B} \sum_{i=1}^B r_s(\mathbf{x}^{(i)}, c) \\ & + \lambda_q \frac{1}{B} \sum_{i=1}^B [\tau_s - r_s(\mathbf{x}^{(i)}, c)]_+ \\ & + \lambda_{\text{div}} [\tau_D - v_B]_+ + \lambda_{\text{reg}} \frac{1}{B} \sum_{i=1}^B \text{reg}(\mathbf{x}_0^{(i)}), \quad (2) \end{aligned}$$

where $[u]_+ = \max(u, 0)$. The diversity statistic aggregates global feature distances, or patch-level distances for P patches:

$$v_B = \frac{1}{P} \sum_{p=1}^P \frac{2}{B(B-1)} \sum_{1 \leq i < j \leq B} d(f_p(\mathbf{x}^{(i)}), f_p(\mathbf{x}^{(j)})), \quad (3)$$

with f_p a patch embedding and d a distance metric (e.g. cosine distance). Beyond pairwise distances for diversity, we can utilize DPP or Vendi Score on top of these pairwise similarity kernels which provide more meaningful set-level diversity metrics. To keep initial noises in high-density regions of the prior we regularize their norm. Writing $\epsilon^{(i)} \equiv \mathbf{x}_0^{(i)}$ and $r^{(i)} = \|\epsilon^{(i)}\|$, the radius r follows a χ^d law under $\mathcal{N}(0, \mathbf{I})$. Following Samuel et al. [50, 51] and Ben-Hamu et al. [7], we maximize the log-likelihood of r , whose unnormalized log-density is

$$K(\epsilon) = (d-1) \log \|\epsilon\| - \frac{1}{2} \|\epsilon\|^2. \quad (4)$$

Similar to recent works [7, 13, 50], we implement this as a penalty $\text{reg}(\mathbf{x}_0^{(i)}) = -K(\epsilon^{(i)})$, which encourages $\|\mathbf{x}_0^{(i)}\|$ to match the χ^d profile of the Gaussian prior and prevents drift to unlikely radii. We optimize $\{\mathbf{x}_0^{(i)}\}$ by backpropagating through the frozen sampler g_θ and update until the stopping criterion $\min_i r_s(\mathbf{x}^{(i)}, c) \geq \tau_s$ and $v_B \geq \tau_D$ is met, or a compute budget is exhausted.

Batched and Sequential Optimization. Our framework supports both batched and sequential optimization. In the batched setting, as in Tab. 1, we follow [42] and jointly optimize a set of 4 images per prompt. However, our approach readily scales to much larger diverse sets which is a significant advantage over batch-only methods like [42]. By generating one image at a time, each image can be optimized to differ from previous outputs. Thus, we avoid the

"A photo of a dog"



Figure 3. Our method sequentially generates large, diverse image sets. For Flux.1 [schnell], our optimization yields improved diversity of generated image sets (right) compared to i.i.d sampling (left). Additional examples are included in the Supplementary (Fig. 4).

Table 1. Output diversity and image-text alignment results on GenEval for our proposed method with SDXL-Turbo using white and pink noise initialization ($\alpha = 0.2$). Output diversity is measured with averaged pairwise DINOv2 [41], DreamSim [16], and LPIPS [66] scores. For our method, we optimize using DINOv2 for diversity and CLIP [43] for quality. Results are reported using random seed 42, incremented across prompts.

Method	White Noise				Pink Noise			
	DINO	DreamSim	LPIPS	CLIPScore	DINO	DreamSim	LPIPS	CLIPScore
i.i.d.	0.588 $\pm$ 0.083	0.249 $\pm$ 0.089	0.642 $\pm$ 0.059	0.335 $\pm$ 0.031	0.642 $\pm$ 0.068	0.305 $\pm$ 0.090	0.729 $\pm$ 0.052	0.328 $\pm$ 0.031
Parmar et al. [42]	0.705 $\pm$ 0.065	0.331 $\pm$ 0.098	0.682 $\pm$ 0.055	0.333 $\pm$ 0.028	0.749 $\pm$ 0.054	0.392 $\pm$ 0.100	0.757 $\pm$ 0.048	0.323 $\pm$ 0.028
Ours	0.784 $\pm$ 0.026	0.411 $\pm$ 0.102	0.767 $\pm$ 0.052	0.349 $\pm$ 0.029	0.786 $\pm$ 0.028	0.427 $\pm$ 0.095	0.811 $\pm$ 0.044	0.341 $\pm$ 0.029

memory overhead of simultaneously processing many candidates, enabling efficient generation of large diverse sets. We show examples of this in Fig. 3. Details on hyperparameters can be found in the Supplementary (Sec. 1.2).

Sampling Initial Noise. Diffusion models commonly initialize the denoising process with white Gaussian noise where the power spectral density is constant across all frequencies. However, natural images have a $1/f$ power spectrum: lower frequencies have more power than higher frequencies [14, 53, 57]. Motivated by this, we explore alternative noise initialization strategies that align more closely with statistical properties of natural images.

In particular, we consider *pink noise* initialization where we apply spectral filtering in the frequency domain. For this, $z_{\text{white}} \sim \mathcal{N}(0, \mathbf{I})$ is transformed to the frequency do-

main using a 2D Fast Fourier Transform (FFT):

$$\hat{z}^f = \text{FFT2D}(z_{\text{white}}). \quad (5)$$

For each frequency component at position (u, v) , we compute the radial frequency $f_{u,v} = \sqrt{u^2 + v^2}$. We then apply power-scaling by reweighing the frequency components with $\frac{1}{(1+f)^\alpha}$:

$$\hat{z}_{\text{pink}}^f(u, v) = \hat{z}^f(u, v) \cdot \frac{1}{(1 + f_{u,v})^\alpha}, \quad (6)$$

where $\alpha \in [0, 1]$ for pink noise. We then transform this back to the spatial domain by applying an inverse 2D FFT:

$$\hat{z}_{\text{pink}} = \text{IFFT2D}(\hat{z}_{\text{pink}}^f), \quad (7)$$

before normalization to match white noise statistics, $z_{\text{pink}} =$



Figure 4. Diverse image generation with Flux.1 [schnell] and SDXL-Turbo using white noise initialization (optimized with DPP [12] and HPSv2 [63] objectives). We observe that our method improves diversity in color, orientation, and style compared to i.i.d. sampling. Additional qualitative results can be found in the Supplementary (Fig. 5).

Table 2. Output diversity (DreamSim, Vendi) and image quality (HPSv2) on GenEval using white and pink ($\alpha = 0.2$) noise initialization. We optimize with the DPP diversity objective and HPSv2 quality reward. Results are reported using random seed 42 across all prompts.

Method	White Noise			Pink Noise		
	DreamSim	Vendi	HPSv2	DreamSim	Vendi	HPSv2
SDXL-Turbo						
i.i.d.	0.262 $\pm$ 0.094	2.000 $\pm$ 0.513	0.284 $\pm$ 0.030	0.296 $\pm$ 0.091	2.136 $\pm$ 0.531	0.285 $\pm$ 0.028
Parmar et al. [42]	0.336 $\pm$ 0.100	2.769 $\pm$ 0.630	0.273 $\pm$ 0.029	0.403 $\pm$ 0.101	3.028 $\pm$ 0.594	0.273 $\pm$ 0.026
Ours	0.457 $\pm$ 0.110	4.000 $\pm$ 0.000	0.292 $\pm$ 0.025	0.474 $\pm$ 0.099	4.000 $\pm$ 0.000	0.288 $\pm$ 0.025
Flux.1 [schnell]						
i.i.d.	0.307 $\pm$ 0.100	2.013 $\pm$ 0.490	0.304 $\pm$ 0.025	0.362 $\pm$ 0.100	2.207 $\pm$ 0.502	0.297 $\pm$ 0.023
Parmar et al. [42]	0.413 $\pm$ 0.105	2.473 $\pm$ 0.554	0.296 $\pm$ 0.023	0.462 $\pm$ 0.100	2.977 $\pm$ 0.560	0.290 $\pm$ 0.024
Ours	0.446 $\pm$ 0.116	2.753 $\pm$ 0.587	0.293 $\pm$ 0.025	0.495 $\pm$ 0.097	3.038 $\pm$ 0.523	0.279 $\pm$ 0.024

$\frac{\hat{z}_{\text{pink}} - \mu}{\sigma}$, where μ and σ are the empirical mean and standard deviation.

4. Experiments

For each prompt, we sample a batch of 4 noise initializations and generate the corresponding 4 candidate images. We then compute an averaged pairwise diversity objective together with an image-level reward, and use both to optimize the initial noises so that the final set exhibits high visual diversity while preserving image quality (Fig. 2).

As diversity objectives, we consider patchwise DINOv2 (Eq. (3)), DreamSim [16], LPIPS [66], Color histogram distance, and a low-resolution pixel L2 measure that uses 32×32 features, following [58]. We also evaluate DPP [31] and Vendi [15] scores computed with a DINOv2 [CLS] kernel which has recently been shown to align well with human judgements [4]. To assess image quality and prompt alignment, we report CLIPScore [21, 43] and HPSv2 [63, 64], and provide standard deviations across test samples.

Our experiments cover popular step-distilled samplers

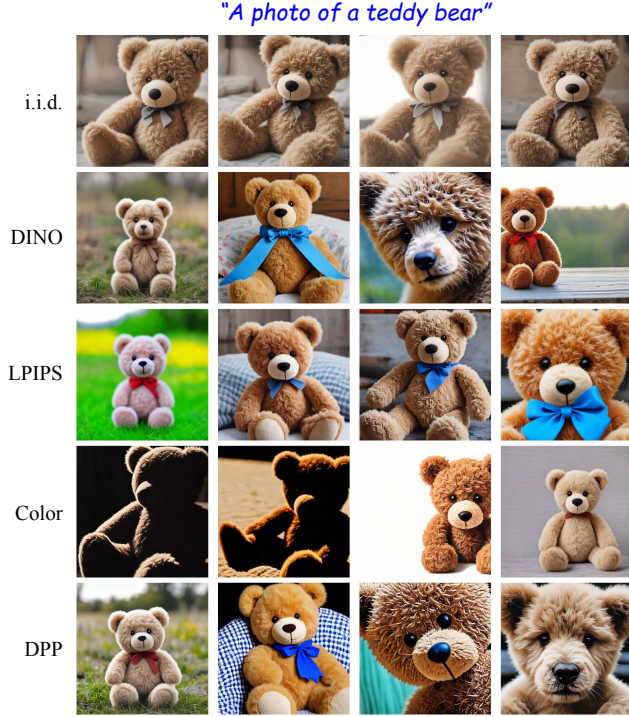


Figure 5. Different optimization objectives produce distinct image sets (SDXL-Turbo). The top row shows i.i.d. samples, and the rest show our method optimizing with DINOv2 [41], LPIPS [66], Color Histogram [58], and DPP [12] diversity objectives. Quantitative results and additional example generations can be found in Tab. 7 and Fig. 13,14 in the Supplementary.

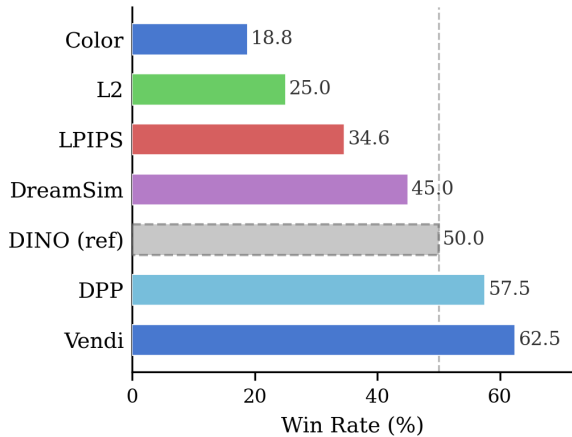


Figure 6. Human preference win rates of our method across diversity objectives for SDXL-Turbo. We use DINOv2 [41] as our reference objective and compare all other metric results against it, including Color Histogram [58], L2 distance, LPIPS [66], DreamSim [16], DPP [12], and Vendi [15]. We see that images optimized with set-level objectives (DPP, Vendi) are preferred by users.

including SDXL-Turbo [52] and a larger 10B+ parameter model Flux.1 [schnell] [35]. Additional model results for SANA-Sprint [9] and PixArt- α -DMD [8] can be found in

the Supplementary (Tab. 4, 5). These models all rely on ODE-based sampling. While SDE-based sampling could further increase diversity, we focus our work on these distilled samplers. The full noise optimization procedure runs on a single A100 or H100 GPU. Additional details are provided in the Supplementary (Sec. 2).

Baseline Comparisons. We compare our test-time optimization approach to sampling from i.i.d. noise, and to [42], which has been shown to outperform previous guidance-based methods [11, 48, 54]. Comparisons to guidance methods can be found in the Supplementary (Tab. 8). Following [42], we set the initial set size to 64 and select 4 diverse outputs using their default objectives.

We show generation diversity and image-text alignment results for text-to-image generation on GenEval [18] in Tab. 1. Here, we optimize CLIPScore [21] for image-text alignment and pairwise cosine similarity scores with DINOv2 following prior work [42]. Our noise optimization demonstrates substantial improvements over i.i.d. sampled noise initializations and [42]. Optimizing the noise gives direct control over the quality-diversity tradeoff, allowing us to flexibly balance our objectives or use additional different diversity and image quality optimization objectives. To generate the results in Tab. 1, we halted the optimization when reaching preset thresholds (CLIPScore comparable to [42], or DINO diversity one standard deviation above [42]). To evaluate diversity, we include scores for metrics that are not used in optimization and show strong generalization to unused metrics. In addition, our user study (Tab. 9 in the Supplementary) confirms that human judgments of diversity are aligned with the reported metrics.

Effect of Different Diversity Optimization Objectives.

We examine how different diversity objectives influence both the variety of generated outputs and the quality of individual samples. Qualitative examples are shown in Fig. 5. The Supplementary contains quantitative results (Tab. 8) and additional example generations (Fig. 13,14). Using SDXL-Turbo on the GenEval prompts, we compare several objectives that aim to increase visual diversity without sacrificing text-image alignment or image quality. Across all settings, our noise optimization maintains image quality while producing clear gains in visual diversity. Each objective best improves its own metric, but others improve as well, indicating that diversity in one feature space transfers to others and that our optimization increases diversity without harming perceptual quality. Additionally, we conduct a user study that compares the pairwise DINO similarity metric with other diversity metrics in Fig. 6. We provide details about the user study setup in the Supplementary (Sec. 6). We observe that image sets obtained with Vendi Score [15] and DPP [12] as diversity objectives are preferred by users. The main advantage of these set-level objectives is that they cannot be increased by simply mak-



Figure 7. Diverse image generation with SDXL-Turbo using white and pink noise initialization. Optimizing DPP [12] and HPSv2 [63], we observe that our method improves diversity in color, pose, and style compared to i.i.d. sampling. In addition, using pink noise simply at inference time without any optimization increases diversity for both i.i.d. sampling and [42]. Additional examples can be found in the Supplementary (Fig. 3).

"A picturesque autumn scene where a quaint cottage with a thatched roof sits beside a tranquil lake, surrounded by trees with leaves in vibrant shades of orange, red, and yellow. The cottage's wooden exterior is complemented by white-framed windows, and a stone chimney rises above the roofline. The lake reflects the warm fall colors, creating a mirror image of the foliage and the small structure on its calm surface."



Figure 8. Our method increases diversity on complex prompts. Optimizing noise for Flux.1 [schnell] on DPG-Bench prompts, we observe that we can increase diversity even when the prompt is highly specified.

ing one single image very different, which would boost the average pairwise score.

Quantitative Results. Informed by our user study and baseline comparison experiments, we optimize SDXL-Turbo and Flux.1 [schnell] using DPP for diversity and HPSv2 [63] for quality. We show results for these settings on GenEval in Tab. 2. We observe that we are able to completely saturate Vendi diversity scores to 4.0 and even slightly improve the HPSv2 quality score. We also see substantial gains on the DreamSim score. See Supplementary Tab. 1 for more details on hyperparameter choice.

For Flux.1 [schnell], we similarly see large gains in Vendi and DreamSim scores compared to i.i.d. sampling and [42] (Tab. 2), with a slight drop in the HPSv2 score. We observed that Flux.1 [schnell] is more sensitive to hyperparameter choice than SDXL-Turbo, particularly with balancing the quality reward. During optimization, we revert to the last latent when the HPSv2 score drops below a threshold (See Supplementary Tab. 1 for threshold values). More dynamic optimization schedules and additional image qual-

ity constraints similar to [13] could be explored in the future to push diversity higher while maintaining image quality.

Qualitative Generation Examples. Fig. 4 showcases the effectiveness of our noise optimization approach in generating images with improved variety compared to image sets generated from i.i.d.-sampled noise initializations. Here, we use the DPP diversity objective (corresponding to Tab. 2). We consistently see increased diversity of object shapes, poses, colors and backgrounds while maintaining alignment to the input prompt.

Diversity on Complex Prompts. Our method also improves diversity on complex prompts. In addition to GenEval, we evaluate our method on DPG-Bench [24] which consists of detailed prompts averaging 67 words. We optimize with Flux.1 [schnell] [34] using a DPP DINO diversity objective. We observe that complex prompts still lead to mode collapse and optimizing for diversity is effective (Fig. 8). Here, our approach also achieves substantial diversity gains (DreamSim, Vendi) while maintaining image quality (Supplementary Tab. 6).

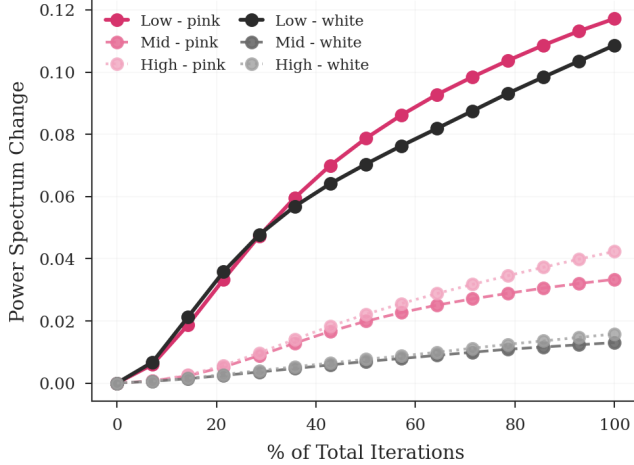


Figure 9. Noise change across different power spectrum bins throughout optimization iterations with SDXL-Turbo averaged over all prompts of GenEval [18]. While optimizing the starting noise, we observe the largest changes occur in the lowest third of the spectrum (low freq). This trend holds for both white and pink noise initializations.

4.1. Noise Initialization

Noise Evolution. We analyze how the optimization modifies the initial noise, which is white Gaussian to start. In particular, we examine changes across frequency bands of the noise power spectrum, shown in Fig. 9. We compute the spectrum via a Fourier Transform on the raw noise latents and track how it evolves over the course of optimization. For interpretability, we divide the spectrum into three equally sized frequency bins and measure the change in each bin relative to the initial noise.

We observe that the majority of the change occurs in the lowest frequency bin, corresponding to the bottom third of the spectrum. Low-frequency components show noticeably larger shifts than mid- or high-frequency components. This indicates that the optimization primarily acts on the low-frequency structure of the noise, with higher frequencies remaining relatively stable throughout the process. Interestingly, we also observe that the final optimized noise remains Gaussian (i.e. follows a standard normal distribution with mean 0 and standard deviation 1), but it is no longer spectrally white due to shifts in the low frequencies.

Pink Noise Initialization. As the majority of noise changes across iterations occur in the low-frequency range, we explore pink noise initializations as they are more likely to cover different regions of the noise space in terms of low noise frequencies which appears to be critical for generating diverse images. The $1/f$ frequency distribution inherent in pink noise allocates greater power to lower frequencies, aligning well with the observed optimization dynamics. The increased diversity in generated images from pink noise initializations is confirmed by our quantitative results

in Tab. 1 and Tab. 2, and example generations from pink noise in Fig. 7. Interestingly, using pink noise initializations also results in higher diversity in output generations for i.i.d.-sampling and [42] while only slightly reducing the image-text alignment as measured by HPSv2. See Tab. 1 in the Supplementary for details on optimization hyperparameters. For additional qualitatively examples, see Supplementary Fig. 2 and 3.

4.2. Scaling Behavior with White and Pink Noise

A crucial aspect of inference-time scaling is to obtain the best possible improvements for the downstream task given additional compute. We observe that our noise optimization approach can outperform [42] in terms of image diversity with just a few iterations (Fig. 1 in the Supplementary). We optimized the noise for different initializations (i.e. α values in Eq. (6)), using the experimental settings from Tab. 1.

With white noise initializations, our approach requires 9 and 12 iterations to reach higher diversity scores than [42] with an initial pool size of 64 and 128 samples respectively. For $\alpha = 0.2$, we require only 6 / 8 iterations to outperform [42] with initial pool size 64 / 128. Our approach with pink noise initialization ($\alpha = 0.2$) requires 12 / 15 iterations to yield more diverse images than [42] with similar initialization. For SDXL-Turbo, our approach takes 0.345 s per iteration (with DINO, CLIP objectives) on a single A100 GPU. That means in the pink noise setting, our method takes 2.07 s in total vs. 11.20 s for [42]. See Supplementary Sec. 2 for additional details on computational cost.

Higher α values generally lead to higher diversity scores. However, the image quality decreases with noise exponents $\alpha > 0.2$ (see Fig. 11 in the Supplementary).

5. Conclusion

In this work, we investigated the critical impact of initial noise on the diversity of diffusion model outputs. We proposed an end-to-end noise optimization approach for maximizing variation across generated samples which allows the flexible selection of diversity optimization objectives. Our noise evolution analysis further inspired a simple yet effective strategy of using pink noise initializations, which consistently enhances the variety of outputs across models and baselines. Our experiments demonstrate that our approach offers a general solution for generating diverse images that significantly outperforms prior methods.

Acknowledgements: This work was in part supported by the BMFT (FKZ: 16IS24060), the DFG (SFB 1233, project number: 276693517), and NSF IIS-2403305. AH was supported by NDSEG Graduate Fellowship. This research utilized compute resources at the Tübingen Machine Learning Cloud.

References

- [1] Donghoon Ahn, Hyoungwon Cho, Jaewon Min, Wooseok Jang, Jungwoo Kim, SeonHwa Kim, Hyun Hee Park, Kyong Hwan Jin, and Seungryong Kim. Self-rectifying diffusion sampling with perturbed-attention guidance. In *ECCV*, 2024. 2
- [2] Donghoon Ahn, Jiwon Kang, Sanghyun Lee, Jaewon Min, Minjae Kim, Wooseok Jang, Hyoungwon Cho, Sayak Paul, SeonHwa Kim, Eunju Cha, et al. A noise is worth diffusion guidance. *arXiv preprint arXiv:2412.03895*, 2024. 2
- [3] Donghoon Ahn, Jiwon Kang, Sanghyun Lee, Minjae Kim, Jaewon Min, Wooseok Jang, Sangwu Lee, Sayak Paul, Susung Hong, and Seungryong Kim. Fine-grained perturbation guidance via attention head selection. *arXiv preprint arXiv:2506.10978*, 2025. 2
- [4] Isabela Albuquerque, Ira Ktena, Olivia Wiles, Ivana Kajić, Amal Rannen-Triki, Cristina Vasconcelos, and Aida Nematzadeh. Benchmarking diversity in image generation via attribute-conditional human evaluation. *arXiv preprint arXiv:2511.10547*, 2025. 5
- [5] Kumar Ashutosh, Yossi Gandelsman, Xinlei Chen, Ishan Misra, and Rohit Girdhar. LLMs can see and hear without any training. *arXiv preprint arXiv:2501.18096*, 2025. 2
- [6] 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. *arXiv preprint arXiv:2406.01970*, 2024. 2
- [7] Heli Ben-Hamu, Omri Puny, Itai Gat, Brian Karrer, Uriel Singer, and Yaron Lipman. D-flow: Differentiating through flows for controlled generation. In *ICML*, 2024. 2, 3
- [8] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, et al. Pixart-alpha: Fast training of diffusion transformer for photorealistic text-to-image synthesis. In *ICLR*, 2024. 6
- [9] Junsong Chen, Shuchen Xue, Yuyang Zhao, Jincheng Yu, Sayak Paul, Junyu Chen, Han Cai, Enze Xie, and Song Han. Sana-sprint: One-step diffusion with continuous-time consistency distillation. *arXiv preprint arXiv:2503.09641*, 2025. 6
- [10] Hyungjin Chung, Jeongsol Kim, Geon Yeong Park, Hyelin Nam, and Jong Chul Ye. Cfg++: Manifold-constrained classifier free guidance for diffusion models. *arXiv preprint arXiv:2406.08070*, 2024. 2
- [11] Gabriele Corso, Yilun Xu, Valentin De Bortoli, Regina Barzilay, and Tommi Jaakkola. Particle guidance: non-iid diverse sampling with diffusion models. *arXiv preprint arXiv:2310.13102*, 2023. 1, 2, 6
- [12] Mohamed Elfeki, Camille Couprie, Morgane Riviere, and Mohamed Elhoseiny. Gdpp: Learning diverse generations using determinantal point processes. In *ICML*, 2019. 2, 5, 6, 7
- [13] Luca Eyring, Shyamgopal Karthik, Karsten Roth, Alexey Dosovitskiy, and Zeynep Akata. Reno: Enhancing one-step text-to-image models through reward-based noise optimization. *NeurIPS*, 2024. 1, 2, 3, 7
- [14] David J Field. Relations between the statistics of natural images and the response properties of cortical cells. *Journal of the Optical Society of America A*, 4(12), 1987. 4
- [15] Dan Friedman and Adji Bousso Dieng. The vendi score: A diversity evaluation metric for machine learning. *arXiv preprint arXiv:2210.02410*, 2022. 2, 5, 6
- [16] Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. *arXiv preprint arXiv:2306.09344*, 2023. 2, 4, 5, 6
- [17] Rohit Gandikota, Zongze Wu, Richard Zhang, David Bau, Eli Shechtman, and Nick Kolkin. Sliderspace: Decomposing the visual capabilities of diffusion models. In *ICCV*, 2025. 2
- [18] Dhruva Ghosh, Hanna Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *NeurIPS*, 2023. 6, 8
- [19] Xiefan Guo, Jinlin Liu, Miaomiao Cui, Jiankai Li, Hongyu Yang, and Di Huang. Initno: Boosting text-to-image diffusion models via initial noise optimization. In *CVPR*, 2024. 2, 3
- [20] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 2
- [21] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *EMNLP*, 2021. 5, 6
- [22] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 2
- [23] Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *ICLR*, 2022. 2
- [24] 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. 7
- [25] Tariq Berrada Ifriqi, Adriana Romero-Soriano, Michal Drozdal, Jakob Verbeek, and Karteek Alahari. Entropy rectifying guidance for diffusion and flow models. *NeurIPS*, 2025. 2
- [26] Tero Karras, Miika Aittala, Tuomas Kynkäänniemi, Jaakko Lehtinen, Timo Aila, and Samuli Laine. Guiding a diffusion model with a bad version of itself. *NeurIPS*, 2024. 2
- [27] Shyamgopal Karthik, Karsten Roth, Massimiliano Mancini, and Zeynep Akata. If at first you don't succeed, try, try again: Faithful diffusion-based text-to-image generation by selection. *arXiv preprint arXiv:2305.13308*, 2023. 2
- [28] Korrawe Karunratanakul, Konpat Preechakul, Emre Aksan, Thabo Beeler, Supasorn Suwajanakorn, and Siyu Tang. Optimizing diffusion noise can serve as universal motion priors. In *CVPR*, 2024. 2, 3
- [29] 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. 2

- [30] 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. *NeurIPS*, 2023. 2
- [31] Alex Kulesza, Ben Taskar, et al. Determinantal point processes for machine learning. *Foundations and Trends in Machine Learning*, 2012. 2, 5
- [32] Mingi Kwon, Jaeseok Jeong, Yi Ting Hsiao, Youngjung Uh, et al. Tcfg: Tangential damping classifier-free guidance. In *CVPR*, 2025. 2
- [33] 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. *NeurIPS*, 2024. 2
- [34] Black Forest Labs. Flux. <https://github.com/black-forest-labs/flux>, 2024. 7
- [35] Black Forest Labs. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025. 1, 6
- [36] 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. 2
- [37] Oscar Mañas, Pietro Astolfi, Melissa Hall, Candace Ross, Jack Urbanek, Adina Williams, Aishwarya Agrawal, Adriana Romero-Soriano, and Michal Drozdal. Improving text-to-image consistency via automatic prompt optimization. *arXiv preprint arXiv:2403.17804*, 2024. 2
- [38] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *CVPR*, 2023. 2
- [39] Mashrur M Morshed and Vishnu Boddeti. Diverseflow: Sample-efficient diverse mode coverage in flows. In *CVPR*, 2025. 2
- [40] Zachary Novack, Julian McAuley, Taylor Berg-Kirkpatrick, and Nicholas J Bryan. Ditto: Diffusion inference-time t-optimization for music generation. *arXiv preprint arXiv:2401.12179*, 2024. 1, 2, 3
- [41] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 2, 3, 4, 6
- [42] 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. *arXiv preprint arXiv:2508.15773*, 2025. 1, 2, 3, 4, 5, 6, 7, 8
- [43] 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 *ICML*, 2021. 4, 5
- [44] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *ICML*, 2021. 2
- [45] Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. *NeurIPS*, 2019. 2
- [46] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *CVPR*, 2023. 1
- [47] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Wei Wei, Tingbo Hou, Yael Pritch, Neal Wadhwa, Michael Rubinstein, and Kfir Aberman. Hyperdreambooth: Hypernetworks for fast personalization of text-to-image models. In *CVPR*, 2024. 1
- [48] Seyedmorteza Sadat, Jakob Buhmann, Derek Bradley, Otmar Hilliges, and Romann M Weber. Cads: Unleashing the diversity of diffusion models through condition-annealed sampling. *arXiv preprint arXiv:2310.17347*, 2023. 1, 2, 6
- [49] Seyedmorteza Sadat, Otmar Hilliges, and Romann M Weber. Eliminating oversaturation and artifacts of high guidance scales in diffusion models. In *ICLR*, 2024. 2
- [50] Dvir Samuel, Rami Ben-Ari, Nir Darshan, Haggai Maron, and Gal Chechik. Norm-guided latent space exploration for text-to-image generation. *NeurIPS*, 2023. 3
- [51] Dvir Samuel, Rami Ben-Ari, Simon Raviv, Nir Darshan, and Gal Chechik. Generating images of rare concepts using pre-trained diffusion models. In *AAAI*, 2024. 2, 3
- [52] Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. *arXiv preprint arXiv:2311.17042*, 2023. 1, 6
- [53] Eero P Simoncelli and Bruno A Olshausen. Natural image statistics and neural representation. *Annual review of neuroscience*, 2001. 4
- [54] Jaskirat Singh, Lindsey Li, Weijia Shi, Ranjay Krishna, Yejin Choi, Pang Wei Koh, Michael F Cohen, Stephen Gould, Liang Zheng, and Luke Zettlemoyer. Negative token merging: Image-based adversarial feature guidance. *arXiv preprint arXiv:2412.01339*, 2024. 1, 2, 6
- [55] Aravindan Sundaram, Ujjayan Pal, Abhimanyu Chauhan, Aishwarya Agarwal, and Srikrishna Karanam. Cocono: Attention contrast-and-complete for initial noise optimization in text-to-image synthesis. *arXiv preprint arXiv:2411.16783*, 2024. 2
- [56] Zhiwei Tang, Jiangweizhi Peng, Jiasheng Tang, Mingyi Hong, Fan Wang, and Tsung-Hui Chang. Inference-time alignment of diffusion models with direct noise optimization. *arXiv preprint arXiv:2405.18881*, 2024. 2, 3
- [57] Antonio Torralba and Aude Oliva. Statistics of natural image categories. *Network: computation in neural systems*, 2003. 4
- [58] Antonio Torralba, Rob Fergus, and William T Freeman. 80 million tiny images: A large data set for nonparametric object and scene recognition. *TPAMI*, 2008. 5, 6
- [59] Masatoshi Uehara, Xingyu Su, Yulai Zhao, Xiner Li, Aviv Regev, Shuiwang Ji, Sergey Levine, and Tommaso Biancalani. Reward-guided iterative refinement in diffusion models at test-time with applications to protein and dna design. *arXiv preprint arXiv:2502.14944*, 2025. 2

- [60] Masatoshi Uehara, Yulai Zhao, Chenyu Wang, Xiner Li, Aviv Regev, Sergey Levine, and Tommaso Biancalani. Inference-time alignment in diffusion models with reward-guided generation: Tutorial and review. *arXiv preprint arXiv:2501.09685*, 2025. [2](#)
- [61] Bram Wallace, Akash Gokul, Stefano Ermon, and Nikhil Naik. End-to-end diffusion latent optimization improves classifier guidance. In *ICCV*, 2023. [1](#), [2](#), [3](#)
- [62] Tianxing Wu, Chenyang Si, Yuming Jiang, Ziqi Huang, and Ziwei Liu. Freeinit: Bridging initialization gap in video diffusion models. In *ECCV*, 2024. [2](#)
- [63] 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. [3](#), [5](#), [7](#)
- [64] Xiaoshi Wu, Keqiang Sun, Feng Zhu, Rui Zhao, and Hongsheng Li. Better aligning text-to-image models with human preference. In *ICCV*, 2023. [5](#)
- [65] Katherine Xu, Lingzhi Zhang, and Jianbo Shi. Good seed makes a good crop: Discovering secret seeds in text-to-image diffusion models. In *WACV*, 2025. [2](#)
- [66] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. [2](#), [4](#), [5](#), [6](#)