

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 ³Technical University of Munich, MCML

* Equal contribution

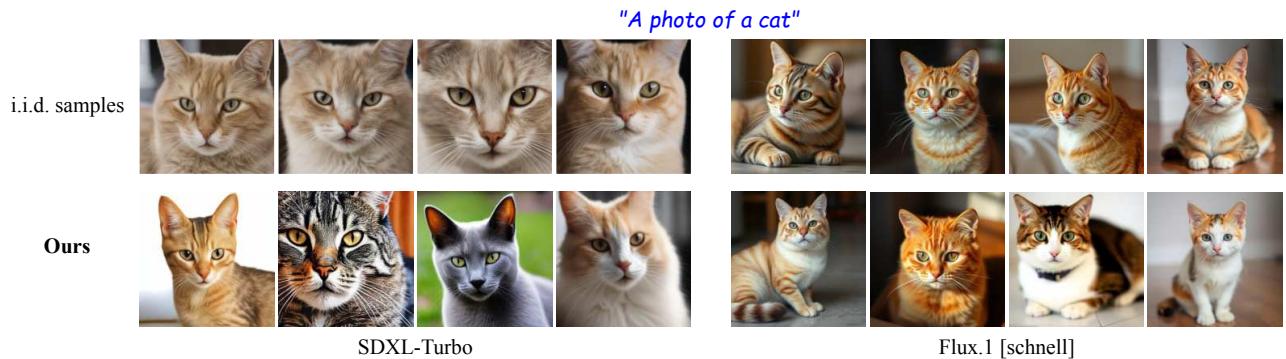


Figure 1. Repeatedly sampling from text-to-image models using a fixed text prompt produces surprisingly little visual variation (top row) in both Stable Diffusion SDXL-Turbo [56] (left) and Flux.1 [schnell] [36] (right). Our approach (bottom row) directly optimizes the initial noise to recover from mode 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. While 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 variety.

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 full 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, 43, 68] and personalization [50, 51].

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, 52, 59]. Alternatively, generating a large number of candidates and iteratively pruning them to optimize for increasing variety has recently shown success [46]. 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. Specifically, we sample a batch of initial noise samples. We are then able to directly optimize these by minimizing a pairwise similarity metric that drives samples apart. Our method outperforms prior works by large margins across multiple diffusion models and benchmarks. We demonstrate that we can flexibly select different optimization objectives that facilitate diversity in generated outputs (e.g. DINOv2 [44], LPIPS [72], 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. Inspired by these observations, we explore boosting low-frequency components in the noise initialization, using pink noise, to increase output diversity. We find that using pink noise consistently improves the diversity of generated samples not only for our approach, but also the baselines we compare to for all evaluated models.

The main elements of our contribution can be summarized as follows:

- i) We propose an end-to-end noise optimization scheme that provides superior diversity of generated outputs compared to prior methods.
- ii) Our framework allows the use and analysis of different optimization metrics to guide the model towards diverse outputs.
- iii) We analyze how noise evolves during optimization and show that boosting low-frequency components (e.g., using pink noise) consistently improves diversity across our method and baseline approaches.

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 during inference, often guided by pre-trained reward models. These methods fall into two categories: search-based approaches [27, 39, 66, 67] that evaluate multiple candidates, and optimization-based approaches [7, 18, 28, 43, 63, 68] that iteratively refine noise or latents through gradient descent. In the context of increasing the diversity in the out-

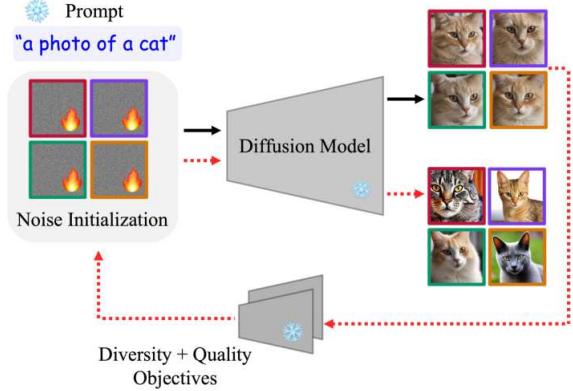


Figure 2. We optimize the noise initialization to increase visual diversity given a fixed text prompt and diffusion model. Starting from i.i.d. noise samples, we generate a set of images. Using a diversity objective (e.g. DINO dissimilarity) and optionally a quality reward (e.g. HPSv2), we update the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

puts of the generative model, Parmar et al. [46] 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 can achieve superior performance on the quality–diversity tradeoff.

Guidance Mechanisms. Drawing from the success of classifier-free guidance (CFG) mechanisms [10, 21] in steering diffusion models towards desired objectives, several variations have been proposed to either improve the effectiveness of CFG [1, 3, 33], or reduce its computational complexity [2, 26, 34]. To increase the diversity when multiple outputs are sampled, several alternatives have been proposed [30, 59], including the usage of particle guidance [11] and DPP [32, 42]. These methods use guidance mechanism to balance tradeoff between quality and diversity [24, 52, 53]. Unlike guidance mechanisms that steer the model toward a particular target through modified conditioning, we directly optimize the initial noise under a target objective to obtain the desired quality–diversity tradeoff.

Prompt Augmentations. Improving controllability in generation by modifying the textual conditioning input rather than the diffusion dynamics [19, 41] has also been a popular direction. These methods even try to explicitly improve quality and/or diversity using LLMs to rewrite prompts for diffusion models [6, 40]. 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, 18, 55, 62]. Furthermore, it

has been observed that specific noise seeds control certain global behavior [71]. However, the most popular approach is to utilize best-of-n sampling approaches [27, 31, 39, 48, 49] 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 alters the diversity of outputs.

3. Collapse Recovery in Diffusion Models

Diffusion Models. Recent generative models are based on a time-dependent formulation between a standard Gaussian distribution $z \sim p_0 = \mathcal{N}(0, \mathbf{I})$ and a data distribution $\mathbf{x}_1 \sim p_{data}$. These models define an interpolation between the initial noise $z = \mathbf{x}_0$ and the data distribution, such that

$$\mathbf{x}_t = \alpha_t \mathbf{x}_0 + \sigma_t \mathbf{x}_1, \quad (1)$$

where α_t is a decreasing and σ_t is an increasing function of $t \in [0, 1]$. Score-based diffusion [22, 25, 29, 60, 61] and flow matching [4, 37, 38] models share the observation that the process \mathbf{x}_t can be sampled dynamically using a stochastic or ordinary differential equation (SDE or ODE). The neural networks parametrizing those ODEs/SDEs are trained to learn the underlying dynamics, typically by predicting the score of the perturbed data distribution or the conditional vector field. Generating a sample then involves simulating the learned differential equation starting from $\mathbf{x}_0 \sim p_0$. The resulting generative model $g_\theta(z, c)$ ingests initial noise z and a prompt c to generate an image x .

Noise Optimization. Test-time optimization techniques aim to improve pre-trained generative models on a per-sample basis at inference. A popular gradient-based strategy is test-time noise optimization [7, 18, 28, 43, 63, 68]. Given a pre-trained generator g_θ (which could be a multi-step diffusion or flow matching model), this approach optimizes the initial noise \mathbf{x}_0 for each generation 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

$$\mathbf{x}_0^* = \arg \max_{\mathbf{x}_0} (r(g_\theta(\mathbf{x}_0)) - \text{reg}(\mathbf{x}_0)), \quad (2)$$

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 , thus ensuring the generated sample $g_\theta(\mathbf{x}_0^*)$ remains plausible. However, 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. 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 $x^{(i)} = g_\theta(\mathbf{x}_0^{(i)}, c)$. We jointly optimize the batch to meet two targets: (i) high sample-level quality via a reward $r_s(x^{(i)}, c)$ 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

$$\begin{aligned} \mathcal{L}(\mathcal{B}) = & -\frac{1}{B} \sum_{i=1}^B r_s(x^{(i)}, c) \\ & + \lambda_{\min} \frac{1}{B} \sum_{i=1}^B [\tau_s - r_s(x^{(i)}, c)]_+ \\ & + \lambda_{\text{div}} [\tau_D - v_B]_+ + \lambda_{\text{reg}} \frac{1}{B} \sum_{i=1}^B \text{reg}(\mathbf{x}_0^{(i)}), \quad (3) \end{aligned}$$

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

$$v_B = \frac{1}{P} \sum_{p=1}^P \frac{2}{B(B-1)} \sum_{1 \leq i < j \leq B} d(f_p(x^{(i)}), f_p(x^{(j)})), \quad (4)$$

with f_p a patch embedding and d a distance metric (e.g. cosine distance). Beyond pairwise distances for diversity, one can also 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 Ben-Hamu et al. [7], Samuel et al. [54, 55], we maximize the log-likelihood of r , whose unnormalized log-density is

$$K(\epsilon) = (d-1) \log \|\epsilon\| - \frac{1}{2} \|\epsilon\|^2. \quad (5)$$

Following recent works [7, 13, 54], 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 $\min_i r_s(x^{(i)}, c) \geq \tau_s$ and $v_B \geq \tau_D$ is met, or a compute budget is exhausted.

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, 57, 64]. 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 domain using a 2D Fast Fourier Transform (FFT):

$$\hat{z}^f = \text{FFT2D}(z_{\text{white}}). \quad (6)$$

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 $\frac{1}{(1+f)^{\alpha}}$:

$$\hat{z}_{\text{pink}}^f(u, v) = \hat{z}^f(u, v) \cdot \frac{1}{(1 + f_{u,v})^{\alpha}}, \quad (7)$$

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

$$\hat{z}_{\text{pink}} = \text{IFFT2D}(\hat{z}_{\text{pink}}^f), \quad (8)$$

before normalizing this to match white noise statistics: $z_{\text{pink}} = \frac{\hat{z}_{\text{pink}} - \mu}{\sigma}$, 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 a set-level variation 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. (4)), DreamSim [16], LPIPS [72], Color histogram distance, and a low-resolution pixel L2 measure that uses 32×32 features, following [65]. We also evaluate DPP [32] and Vendi [15] scores computed with a DINOv2 [CLS] kernel which has recently been shown to align well with human judgements [5]. To assess image quality and prompt alignment, we report CLIPScore [20, 47] and HPSv2 [69, 70], and provide standard deviations across test samples.

Our experiments cover popular step-distilled samplers including SDXL-Turbo [56], SANA-Sprint [9], PixArt- α -DMD [8], and Flux.1 [schnell] [36]. The full noise optimization procedure runs on a single A100 or H100 GPU. Additional details are provided in the Appendix (Sec. B).

Baselines. We compare our test-time optimization approach to sampling from i.i.d. noise, and to [46], which has been shown to outperform previous guidance-based methods [11, 52]. Following [46], we set the initial set size to 64 and select 4 diverse outputs using their default objectives.

Quantitative Results. We show generation variety and image-text alignment results for text-to-image generation on GenEval [17] and on a subset of 50 random prompts per category on the T2I-CompBench [45] benchmark in Tab. 1.



Figure 3. Example images generated with SDXL-Turbo using different optimization objectives for the prompt “a photo of a teddy bear” (top row: i.i.d. samples). Additional examples are included in the Appendix (Figs. 15 and 16).

We optimize CLIPScore [20] for image-text alignment and pairwise cosine similarity scores with DINOv2 following prior work [46]. Our noise optimization demonstrates substantial improvements over i.i.d. sampled noise initializations and [46] across three different models on both benchmarks. Optimizing the noise gives direct control over the quality-diversity trade-off, 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 [46], or DINO diversity one standard deviation above [46]).

Effect of Different Set Optimization Objectives. We examine how different set-level objectives influence both the variety of generated outputs and the quality of individual samples. Results are shown in Tab. 2 along with a qualitative example in Fig. 3. 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 variation. 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 reduces diversity without harming perceptual quality. Additionally, we conduct a user study that compares the pair-

Table 1. Output diversity and image-text alignment results on GenEval and T2I-CompBench for our proposed method with the PixArt- α , SANA-Sprint-1.6B, and SDXL-Turbo models using white noise initialization. Output diversity is measured with averaged pairwise DINO, DreamSim, and LPIPS scores.

| Method | GenEval [17] | | | | T2I-CompBench [23] | | | |
|---------------------------------------|-------------------|-------------------|-------------------|-------------------|--------------------|-------------------|-------------------|-------------------|
| | DINO | DreamSim | LPIPS | CLIPScore | DINO | DreamSim | LPIPS | CLIPScore |
| PixArt-α [8] | | | | | | | | |
| i.i.d. | 0.431 \pm 0.094 | 0.182 \pm 0.080 | 0.474 \pm 0.119 | 0.326 \pm 0.030 | 0.469 \pm 0.084 | 0.188 \pm 0.069 | 0.512 \pm 0.099 | 0.326 \pm 0.027 |
| Parmar et al. [46] | 0.559 \pm 0.091 | 0.246 \pm 0.094 | 0.569 \pm 0.107 | 0.327 \pm 0.028 | 0.590 \pm 0.078 | 0.256 \pm 0.088 | 0.593 \pm 0.088 | 0.328 \pm 0.027 |
| Ours (DINO) | 0.695 \pm 0.063 | 0.335 \pm 0.107 | 0.664 \pm 0.089 | 0.337 \pm 0.026 | 0.716 \pm 0.060 | 0.331 \pm 0.102 | 0.674 \pm 0.072 | 0.335 \pm 0.023 |
| SANA-Sprint-1.6B [9] | | | | | | | | |
| i.i.d. | 0.526 \pm 0.088 | 0.229 \pm 0.075 | 0.635 \pm 0.087 | 0.336 \pm 0.032 | 0.562 \pm 0.074 | 0.252 \pm 0.078 | 0.656 \pm 0.066 | 0.334 \pm 0.029 |
| Parmar et al. [46] | 0.714 \pm 0.060 | 0.354 \pm 0.095 | 0.741 \pm 0.055 | 0.342 \pm 0.032 | 0.684 \pm 0.060 | 0.331 \pm 0.089 | 0.718 \pm 0.049 | 0.338 \pm 0.028 |
| Ours (DINO) | 0.744 \pm 0.061 | 0.438 \pm 0.099 | 0.781 \pm 0.062 | 0.335 \pm 0.030 | 0.738 \pm 0.056 | 0.437 \pm 0.105 | 0.767 \pm 0.053 | 0.330 \pm 0.029 |
| SDXL-Turbo [56] | | | | | | | | |
| i.i.d. | 0.588 \pm 0.083 | 0.249 \pm 0.089 | 0.642 \pm 0.059 | 0.335 \pm 0.031 | 0.586 \pm 0.079 | 0.244 \pm 0.077 | 0.634 \pm 0.056 | 0.332 \pm 0.029 |
| Parmar et al. [46] | 0.705 \pm 0.065 | 0.331 \pm 0.098 | 0.682 \pm 0.055 | 0.333 \pm 0.028 | 0.701 \pm 0.063 | 0.329 \pm 0.087 | 0.680 \pm 0.048 | 0.334 \pm 0.029 |
| Ours (DINO) | 0.784 \pm 0.026 | 0.411 \pm 0.102 | 0.767 \pm 0.052 | 0.349 \pm 0.029 | 0.799 \pm 0.021 | 0.424 \pm 0.085 | 0.764 \pm 0.056 | 0.351 \pm 0.027 |

Table 2. Impact of different optimization objectives for our pipeline with SDXL-Turbo on GenEval using white noise initializations. Our optimization pipeline does not hurt the overall image quality (measured by HPSv2) across different diversity objectives (the result on the metric that we optimized for is shown in brackets), despite only using a weakly weighted CLIP text-image objective as an additional reward to maintain adherence to the input prompt.

| Objective | DINO | DreamSim | LPIPS | Color | L2 | DPP | Vendi | HPSv2 | CLIPScore |
|-------------|---------------------|---------------------|---------------------|---------------------|---------------------|---------------------|---------------------|-------------------|-------------------|
| None (init) | 0.588 \pm 0.082 | 0.249 \pm 0.089 | 0.643 \pm 0.059 | 0.094 \pm 0.041 | 0.279 \pm 0.046 | 2.104 \pm 0.216 | 1.999 \pm 0.505 | 0.263 \pm 0.027 | 0.335 \pm 0.031 |
| DINO | (0.892 \pm 0.049) | 0.476 \pm 0.105 | 0.799 \pm 0.056 | 0.165 \pm 0.057 | 0.436 \pm 0.061 | 2.678 \pm 0.114 | 3.652 \pm 0.368 | 0.260 \pm 0.024 | 0.347 \pm 0.032 |
| DreamSim | 0.718 \pm 0.083 | (0.763 \pm 0.245) | 0.786 \pm 0.082 | 0.177 \pm 0.068 | 0.407 \pm 0.079 | 2.450 \pm 0.218 | 2.919 \pm 0.613 | 0.243 \pm 0.027 | 0.333 \pm 0.028 |
| LPIPS | 0.680 \pm 0.077 | 0.383 \pm 0.119 | (0.852 \pm 0.100) | 0.146 \pm 0.062 | 0.370 \pm 0.065 | 2.219 \pm 0.221 | 2.276 \pm 0.552 | 0.269 \pm 0.025 | 0.338 \pm 0.030 |
| Color | 0.661 \pm 0.076 | 0.401 \pm 0.117 | 0.726 \pm 0.069 | (0.376 \pm 0.156) | 0.408 \pm 0.080 | 2.241 \pm 0.216 | 2.330 \pm 0.552 | 0.259 \pm 0.027 | 0.346 \pm 0.032 |
| L2 | 0.684 \pm 0.065 | 0.362 \pm 0.091 | 0.768 \pm 0.056 | 0.145 \pm 0.052 | (0.492 \pm 0.081) | 2.237 \pm 0.213 | 2.318 \pm 0.538 | 0.268 \pm 0.024 | 0.335 \pm 0.033 |
| DPP | 0.787 \pm 0.043 | 0.477 \pm 0.098 | 0.778 \pm 0.054 | 0.170 \pm 0.061 | 0.444 \pm 0.058 | (2.772 \pm 0.000) | 4.000 \pm 0.001 | 0.261 \pm 0.025 | 0.368 \pm 0.035 |
| Vendi | 0.791 \pm 0.043 | 0.486 \pm 0.103 | 0.782 \pm 0.052 | 0.167 \pm 0.060 | 0.440 \pm 0.057 | 2.773 \pm 0.000 | (4.000 \pm 0.000) | 0.259 \pm 0.024 | 0.356 \pm 0.034 |

Table 3. Human preference win rates of noise optimization methods. Win rates compare the image quality resulting from each diversity objective against the baseline DINO objective, both applied to the SDXL-Turbo model.

| Diversity Objective | Win Rate (%) |
|---------------------|--------------|
| Color | 18.8 |
| L2 | 25.0 |
| LPIPS | 34.6 |
| DreamSim | 45.0 |
| DPP | 57.5 |
| Vendi | 62.5 |

wise DINO similarity metric with other diversity metrics in Tab. 3. We provide details about the user study setup in the Appendix (Sec. C.4). We observe that image sets obtained with Vendi Score [15] and DPP [12] as diversity objectives are preferred by users. The main of these set-level objectives is that they cannot be increased by simply making one

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

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, and the recent group inference method [46]. Here, we use the DINO diversity objective (corresponding to Tab. 1, Tab. 4). We consistently see increased diversity of object shapes, poses, colors and backgrounds while maintaining alignment to the input prompt.

4.1. Noise Initialization

Noise Evolution. We analyze how the optimization modifies the initial noise. In particular, we examine changes across frequency bands of the noise power spectrum, shown in Fig. 6. 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 mea-

Table 4. Output diversity and image-text alignment results on GenEval and T2I-CompBench for our proposed method and pink noise initialization with the PixArt- α , SANA-Sprint-1.6B, and SDXL-Turbo models. Output diversity is measured with averaged pairwise DINO, DreamSim, and LPIPS scores.

| Method | Noise | GenEval [17] | | | | T2I-CompBench [23] | | | |
|---------------------------------------|-------|-------------------|-------------------|-------------------|-------------------|--------------------|-------------------|-------------------|-------------------|
| | | DINO | DreamSim | LPIPS | CLIPScore | DINO | DreamSim | LPIPS | CLIPScore |
| PixArt-α [8] | | | | | | | | | |
| i.i.d. | P | 0.533 \pm 0.088 | 0.244 \pm 0.091 | 0.604 \pm 0.116 | 0.326 \pm 0.030 | 0.558 \pm 0.077 | 0.247 \pm 0.083 | 0.626 \pm 0.095 | 0.325 \pm 0.027 |
| Parmar et al. [46] | P | 0.664 \pm 0.074 | 0.319 \pm 0.104 | 0.684 \pm 0.094 | 0.323 \pm 0.029 | 0.675 \pm 0.066 | 0.326 \pm 0.095 | 0.692 \pm 0.077 | 0.324 \pm 0.026 |
| Ours (DINO) | P | 0.764 \pm 0.039 | 0.388 \pm 0.102 | 0.750 \pm 0.067 | 0.335 \pm 0.029 | 0.770 \pm 0.046 | 0.377 \pm 0.097 | 0.748 \pm 0.063 | 0.333 \pm 0.024 |
| SANA-Sprint-1.6B [9] | | | | | | | | | |
| i.i.d. | P | 0.551 \pm 0.083 | 0.235 \pm 0.075 | 0.649 \pm 0.083 | 0.335 \pm 0.033 | 0.584 \pm 0.069 | 0.259 \pm 0.079 | 0.670 \pm 0.065 | 0.334 \pm 0.029 |
| Parmar et al. [46] | P | 0.737 \pm 0.053 | 0.369 \pm 0.093 | 0.767 \pm 0.050 | 0.341 \pm 0.032 | 0.705 \pm 0.056 | 0.346 \pm 0.090 | 0.736 \pm 0.048 | 0.335 \pm 0.028 |
| Ours (DINO) | P | 0.753 \pm 0.049 | 0.440 \pm 0.093 | 0.784 \pm 0.056 | 0.334 \pm 0.031 | 0.750 \pm 0.046 | 0.443 \pm 0.096 | 0.773 \pm 0.050 | 0.330 \pm 0.030 |
| SDXL-Turbo [56] | | | | | | | | | |
| i.i.d. | P | 0.642 \pm 0.068 | 0.305 \pm 0.090 | 0.729 \pm 0.052 | 0.328 \pm 0.031 | 0.643 \pm 0.071 | 0.303 \pm 0.080 | 0.719 \pm 0.055 | 0.326 \pm 0.028 |
| Parmar et al. [46] | P | 0.749 \pm 0.054 | 0.392 \pm 0.100 | 0.757 \pm 0.048 | 0.323 \pm 0.028 | 0.742 \pm 0.055 | 0.391 \pm 0.088 | 0.751 \pm 0.049 | 0.328 \pm 0.027 |
| Ours (DINO) | P | 0.786 \pm 0.028 | 0.427 \pm 0.095 | 0.811 \pm 0.044 | 0.341 \pm 0.029 | 0.804 \pm 0.026 | 0.440 \pm 0.084 | 0.808 \pm 0.049 | 0.344 \pm 0.026 |

Table 5. Output diversity (DreamSim, Vendi) and image quality (HPSv2) on GenEval using white noise initialization optimized with the DPP diversity objective.

| Method | DreamSim | Vendi | HPSv2 |
|-------------------------|-------------------|-------------------|-------------------|
| SDXL-Turbo | | | |
| i.i.d. | 0.249 \pm 0.089 | 1.999 \pm 0.505 | 0.263 \pm 0.027 |
| Parmar et al. [46] | 0.331 \pm 0.098 | 2.348 \pm 0.567 | 0.275 \pm 0.028 |
| Ours | 0.477 \pm 0.098 | 4.000 \pm 0.001 | 0.261 \pm 0.025 |
| Flux.1 [schnell] | | | |
| i.i.d. | 0.307 \pm 0.100 | 2.013 \pm 0.490 | 0.304 \pm 0.025 |
| Parmar et al. [46] | 0.413 \pm 0.105 | 2.473 \pm 0.554 | 0.296 \pm 0.023 |
| Ours | 0.446 \pm 0.116 | 2.753 \pm 0.587 | 0.293 \pm 0.025 |

sure 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.

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 achieving 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 qualitative results in Tab. 4, and example generations from pink noise in Fig. 4.

Interestingly, using pink noise initializations also results in higher diversity in output generations for i.i.d.-sampling and [46] while only slightly reducing the image-text alignment as measured by the CLIPScore.

4.2. Scaling Behaviors

Scaling with White / Pink Noise. A crucial aspect of inference-time scaling is to obtain the best possible improvements for the downstream task given the additional compute. We study how our approach scales in Fig. 7, where we observe that noise optimization can outperform [46] with just a few iterations. We optimized the noise for different initializations (i.e. α values in Eq. (7)), using the experimental settings from Tab. 2.

With white noise initializations, our approach requires 9 and 12 iterations to reach higher diversity scores than [46] with an initial pool size of 64 and 128 samples respectively. For $\alpha = 0.2$, our method requires only 6 / 8 iterations to outperform [46] with initial pool size 64 / 128. When comparing to [46] initialized with 64 / 128 pink noise samples, our approach with pink noise initialization ($\alpha = 0.2$) requires 12 / 15 iterations to yield more diverse images than the baseline.

Higher α values generally lead to higher diversity scores. However, the image quality decreases with noise exponents $\alpha > 0.2$ (see Fig. 9 in the Appendix). Additional rewards for image quality could easily be included in our pipeline, but would increase computational cost.

Scaling Optimization to Larger Models. Furthermore, we demonstrate the applicability of noise optimization to 10B+ parameter models such as Flux.1 [schnell] [36] in Tab. 5. SDXL-Turbo uses the setup from Tab. 2. For Flux.1 [schnell], we use 80 iterations, a DPP diversity weight $\lambda_{div} = 1.5$, learning rate of 6.0, and gradient clipping of

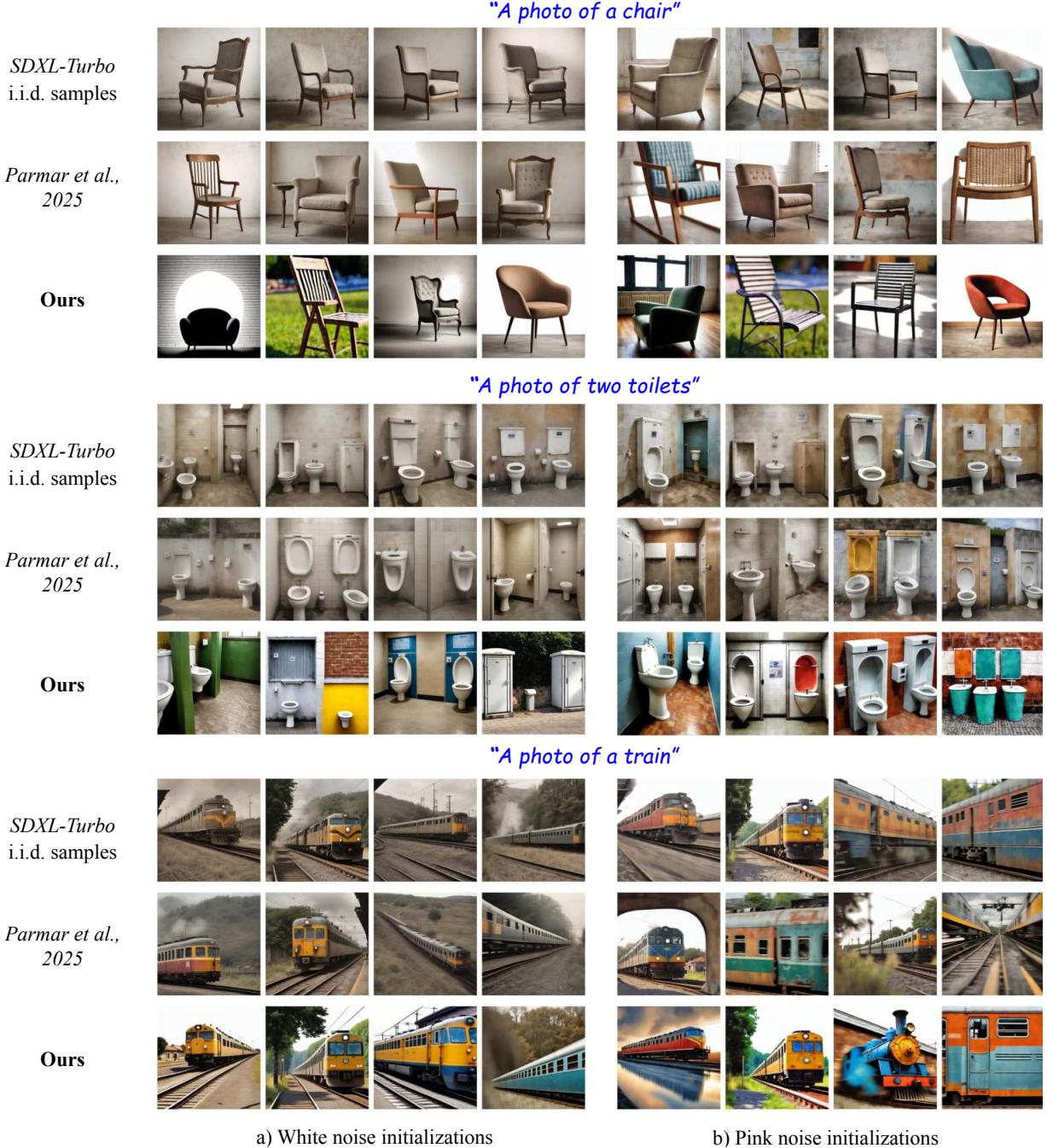


Figure 4. Image generations using our noise optimization approach for SDXL-Turbo yields improved diversity within generated image sets compared to i.i.d sampling and [46]. Pink noise initializations (b) give more diverse generations than standard white noise (a). Ours uses the DINO diversity objective (similar to Tab. 1 and Tab. 4).

0.1. During optimization, we revert to the last latent when the HPSv2 score drops below a threshold of 0.31. 1 shows two example generations with this setup. See Fig. 14 in the Appendix for further Flux.1 [schnell] generations.

Once again, we observe similar patterns as before, where the diversity measured with independent metrics (Vendi and

DreamSim) is improved. As expected, image quality decreases slightly. Additional constraints to improve image quality could be incorporated if desired, similar to [13].

Sequential Generation. In our experiments so far, we generated sets of 4 images following [46]. However, our approach readily scales to much larger diverse sets which is

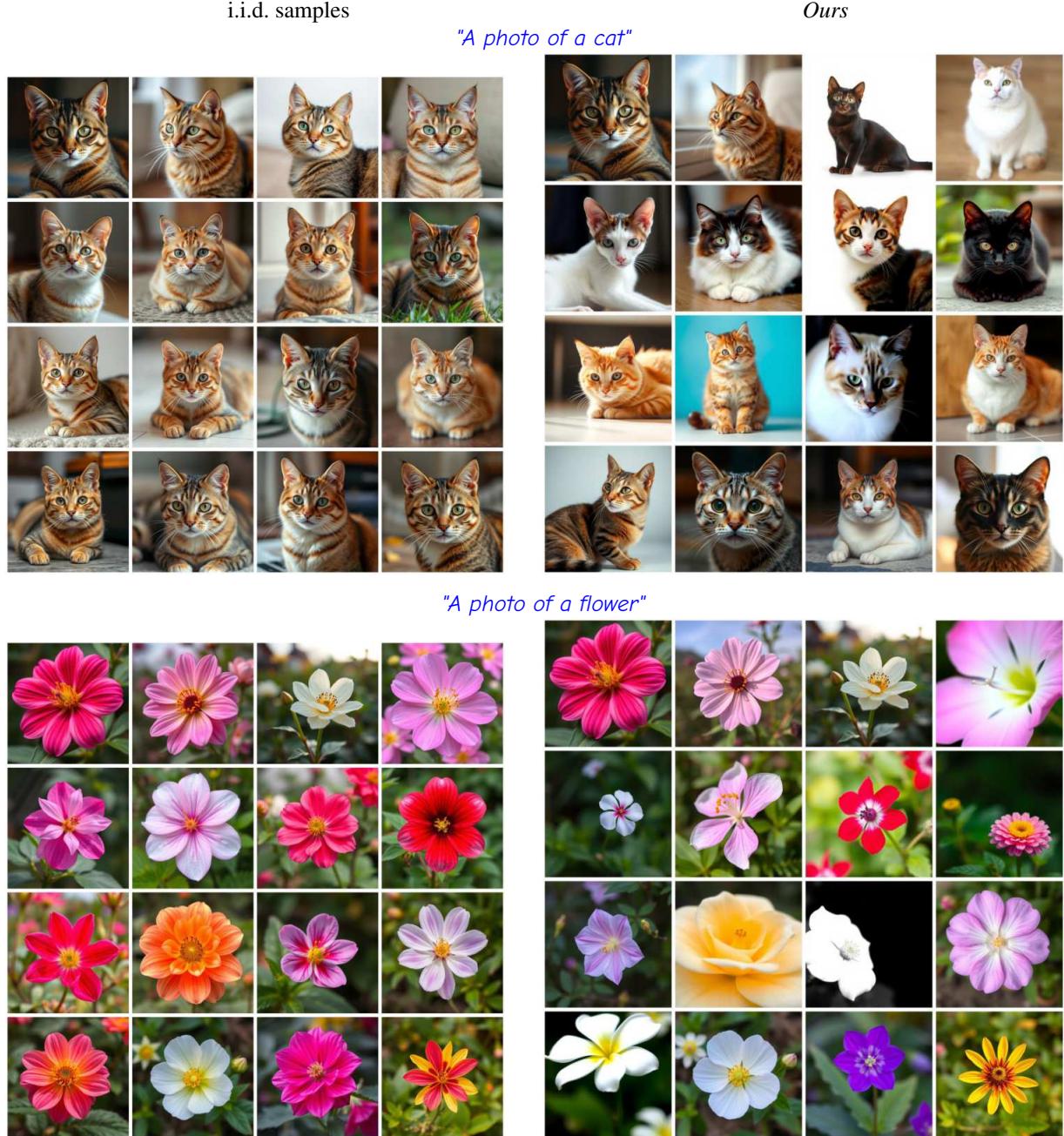


Figure 5. Sequential image generations using our noise optimization approach for Flux.1 [schnell] yields improved diversity of generated image sets compared to i.i.d sampling. Our approach scales to large image sets by sequentially generating diverse images.

a significant advantage over batch methods. By generating one image at a time, each image can be optimized to differ from previous outputs. We avoid the memory overhead of simultaneously processing many candidates, enabling efficient generation of large diverse sets. We show examples of this in Fig. 5. Here, we use 25 iterations, a learning rate of 3.0, $\lambda_{div} = 15$ for the DPP diversity objective, $\lambda_q = 1$ for a HPSv2 quality reward, and gradient clipping of 0.15.

5. Conclusion

In this work, we investigated the critical impact of initial noise on the variation in 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 effec-

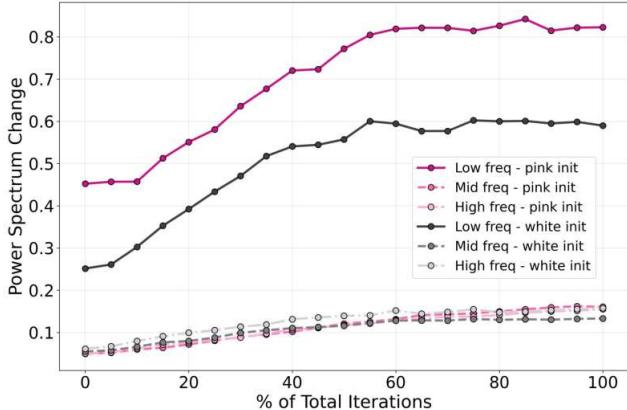


Figure 6. Noise change in different bins in the power spectrum of the noise through optimization iterations, showing that the largest changes occur for low frequencies (low freq corresponds to the lowest third of the spectrum).

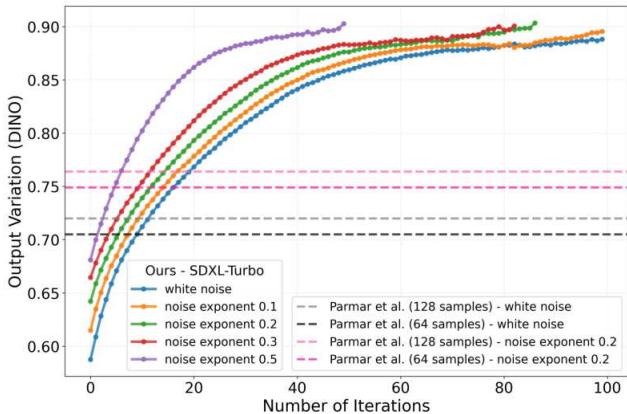


Figure 7. Output variation across optimization iterations for SDXL-Turbo with different noise initializations on GenEval. Higher noise exponents produce greater diversity. Dashed lines are baseline scores from [46] for white noise (gray/black) and pink noise with exponent 0.2 (pink tones) using 64 and 128 samples. Our approach reaches higher diversity (output variation) than [46], requiring only relatively few iterations to outperform [46].

tive strategy of using pink noise initializations, which consistently enhance the variety of outputs across models and baselines. Our experiments demonstrate that our approach significantly outperforms prior methods, offering a general solution for generating diverse images.

Acknowledgements: This work was in part supported by the BMFTR (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 *European Conference on Computer Vision*, 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] Michael S Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. In *ICLR*, 2023. [3](#)
- [5] 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. [4](#)
- [6] 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](#)
- [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. [4, 5, 6, 12, 16](#)
- [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. [4, 5, 6, 12, 16](#)
- [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, 4](#)
- [12] Mohamed Elfeki, Camille Couprie, Morgane Riviere, and Mohamed Elhoseiny. Gdpp: Learning diverse generations using determinantal point processes. In *ICML*, 2019. [2, 5](#)
- [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. In *NeurIPS*, 2024. [1, 2, 3, 7, 12](#)
- [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):2379–2394, 1987. 3
- [15] Dan Friedman and Adji Bousoo Dieng. The vendi score: A diversity evaluation metric for machine learning. *arXiv preprint arXiv:2210.02410*, 2022. 2, 4, 5, 12
- [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, 12
- [17] Dhruba Ghosh, Hanna Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. In *NeurIPS*, 2023. 4, 5, 6, 13, 16
- [18] 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
- [19] 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
- [20] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *EMNLP*, 2021. 4
- [21] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 2
- [22] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 3
- [23] Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xi-hui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. In *NeurIPS*, 2023. 5, 6
- [24] Tariq Berrada Ifriqi, Adriana Romero-Soriano, Michal Drozdzal, Jakob Verbeek, and Karteek Alahari. Entropy rectifying guidance for diffusion and flow models. In *NeurIPS*, 2025. 2
- [25] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *NeurIPS*, 2022. 3
- [26] Tero Karras, Miika Aittala, Tuomas Kynkänniemi, 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, 3
- [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] Diederik P. Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. In *NeurIPS*, 2021. 3
- [30] 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
- [31] 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. In *NeurIPS*, 2023. 3
- [32] Alex Kulesza, Ben Taskar, et al. Determinantal point processes for machine learning. *Foundations and Trends® in Machine Learning*, 5(2–3):123–286, 2012. 2, 4, 12
- [33] Mingi Kwon, Jaeseok Jeong, Yi Ting Hsiao, Youngjung Uh, et al. Tcfg: Tangential damping classifier-free guidance. In *CVPR*, 2025. 2
- [34] Tuomas Kynkänniemi, 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
- [35] Black Forest Labs. Flux. <https://github.com/black-forest-labs/flux>, 2024. 12
- [36] Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space, 2025. 1, 4, 6, 12, 17
- [37] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *ICLR*, 2023. 3
- [38] 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. 3
- [39] 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, 3
- [40] Oscar Mañas, Pietro Astolfi, Melissa Hall, Candace Ross, Jack Urbanek, Adina Williams, Aishwarya Agrawal, Adriana Romero-Soriano, and Michal Drozdzal. Improving text-to-image consistency via automatic prompt optimization. *arXiv preprint arXiv:2403.17804*, 2024. 2
- [41] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6038–6047, 2023. 2
- [42] Mashrur M Morshed and Vishnu Boddeti. Diverseflow: Sample-efficient diverse mode coverage in flows. In *CVPR*, 2025. 2
- [43] Zachary Novack, Julian McAuley, Taylor Berg-Kirkpatrick, and Nicholas J. Bryan. Ditto: Diffusion inference-time t-optimization for music generation, 2024. 1, 2, 3
- [44] Maxime Oquab, Timothée Darcret, 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, 12

- [45] Dong Huk Park, Samaneh Azadi, Xihui Liu, Trevor Darrell, and Anna Rohrbach. Benchmark for compositional text-to-image synthesis. In *NeurIPS Datasets and Benchmarks Track*, 2021. 4, 13
- [46] Gaurav Parmar, Or Patashnik, Daniil Ostanhev, 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, 4, 5, 6, 7, 9, 12, 13, 16, 18, 19, 20
- [47] 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, 12
- [48] 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. 3
- [49] Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems*, 32, 2019. 3
- [50] 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
- [51] 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
- [52] 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, 4
- [53] Seyedmorteza Sadat, Otmar Hilliges, and Romann M Weber. Eliminating oversaturation and artifacts of high guidance scales in diffusion models. In *ICLR*, 2024. 2
- [54] Dvir Samuel, Rami Ben-Ari, Nir Darshan, Haggai Maron, and Gal Chechik. Norm-guided latent space exploration for text-to-image generation. In *NeurIPS*, 2023. 3
- [55] 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
- [56] Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. *arXiv preprint arXiv:2311.17042*, 2023. 1, 4, 5, 6, 12, 16
- [57] Eero P Simoncelli and Bruno A Olshausen. Natural image statistics and neural representation. *Annual review of neuroscience*, 24(1):1193–1216, 2001. 3
- [58] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 12
- [59] Jaskirat Singh, Lindsey Li, Weijia Shi, Ranjay Krishna, Jejin 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
- [60] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021. 3
- [61] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *ICLR*, 2021. 3
- [62] 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
- [63] 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
- [64] Antonio Torralba and Aude Oliva. Statistics of natural image categories. *Network: computation in neural systems*, 14(3):391, 2003. 3
- [65] Antonio Torralba, Rob Fergus, and William T Freeman. 80 million tiny images: A large data set for nonparametric object and scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 30(11):1958–1970, 2008. 4, 12
- [66] 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, 2025. 2
- [67] Masatoshi Uehara, Yulai Zhao, Chenyu Wang, Xiner Li, Aviv Regev, Sergey Levine, and Tommaso Biancalani. Inference-time alignment in diffusion models with reward-guided generation: Tutorial and review, 2025. 2
- [68] Bram Wallace, Akash Gokul, Stefano Ermon, and Nikhil Naik. End-to-end diffusion latent optimization improves classifier guidance. In *ICCV*, 2023. 1, 2, 3
- [69] 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. 4, 12
- [70] Xiaoshi Wu, Keqiang Sun, Feng Zhu, Rui Zhao, and Hongsheng Li. Better aligning text-to-image models with human preference. In *ICCV*, 2023. 4
- [71] 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. 3
- [72] 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, 12

A. Implementation Details

A.1. Optimization Objectives and Metrics

Output diversity. We use multiple diversity objectives that aim at generating a set of diverse images from diffusion models. In the following, we first describe the pairwise diversity metrics that we used.

DINO. This diversity objective and metric uses DINOv2 [44] patch features to measure perceptual diversity as defined in Eq. (4). Specifically, we compute the pairwise cosine distances (i.e. d is the cosine distance) between patch features in different images. Lower values indicate similar images, and values closer to 1 represent higher diversity. We also refer to this metric as “Output variation (DINO)”.

DreamSim. We use pairwise DreamSim dissimilarity scores obtained with a DINO ViT-B/16 backbone that was trained to align with human perception [16]. Lower values indicate similar images, whereas values closer to 1 correspond to more diversity in the outputs.

LPIPS. We use LPIPS [72] to quantify the dissimilarity between a pair of images with a VGG [58] backbone. Specifically, LPIPS computes a weighted sum of perceptual similarities across the outputs of all five convolutional blocks of VGG16. Values close to 0 indicate similar images, whereas values closer to 1 indicate higher diversity.

Color Histogram. We consider the pairwise color histogram distance between images. In particular, we calculate color histograms for each channel considering 32 bins. We use soft histograms with Gaussian kernels to ensure that this operation is differentiable. We then measure the pairwise L2 distance between the resulting color histograms of two images, and normalize this such that the final score is in the range $[0, 1]$.

L2. Inspired by the image similarity used in [65], we use a low-resolution L2 distance between pairs of images. In particular, we resize the generated images to 32×32 and compute the L2 distance between the resulting 3072-dimensional vectors representing each image. We normalize this score to be in the range $[0, 1]$. Higher values correspond to higher diversity.

In addition to the above described averaged pairwise diversity objectives, we consider two set-based metrics.

DPP. We normalize the DINOv2 [CLS] token embeddings \bar{f}_i for each image $x^{(i)}$. The normalized embeddings are used to construct a similarity kernel matrix $K_s = \bar{F}\bar{F}^T$ where $\bar{F} = [\bar{f}_1, \bar{f}_2, \dots, \bar{f}_N]^T$, and N the number of images. The kernel is symmetrized as $K_{sym} = (K_s + K_s^T)/2$ and augmented with $K \leftarrow K_{sym} + \epsilon I$ where $\epsilon = 10^{-6}$. The Determinantal Point Process (DPP) score [32] is then computed as the log-determinant:

$$\mathcal{D}\text{DPP} = \log \det(I + K). \quad (9)$$

This score ranges between $[0, \log(16)]$ for a set of four images, with 0 indicating that all images are identical, and 2.77 stating that all images in the set are maximally diverse.

Vendi. Starting with the same similarity kernel K as in DPP, we compute its eigenvalue decomposition to obtain $\lambda_1, \lambda_2, \dots, \lambda_N$. These eigenvalues are normalized to form a probability distribution $p_i = \lambda_i / \sum_{j=1}^N \lambda_j$. The Vendi score [15] is defined as the exponential of the Shannon entropy of this distribution:

$$\mathcal{D}\text{Vendi} = \exp \left(- \sum_{i=1}^N p_i \log(p_i + \delta) \right), \quad (10)$$

where $\delta = 10^{-12}$ to prevent numerical issues. This score is between $[1, 4]$ for a set of four images, measuring the effective number of diverse images in the set. A score of 1 signifies that all images are effectively similar, and 4 shows that each image in the set is unique.

Image Quality. We measure image quality using CLIP-Score and a human preference score.

CLIPScore. Similar to [13], we use a reward model that pushes the optimization process to preserve image quality and prompt relevance. Specifically, we use a pretrained CLIP [47] ViT-B/32 model. [46] also used this model to ensure image quality and prompt following.

HPSv2. We use the HPSv2 [69] metric as additional (evaluation) metric to confirm that using different diversity objectives does not result in significant degradation of image quality (see Tab. 2). It is based on a CLIP [47] ViT-H/14 backbone.

A.2. Hyperparameter Choices

We use the SDXL-Turbo [56], SANA-Sprint [9], PixArt- α -DMD [8], and Flux.1 [schnell] [36] models in our experiments. For **i.i.d. samples**, we randomly sample input noise and generate a set of four images in a model’s default configuration without altering the four initial noises.

Parmar et al. [46]. We apply [46] to the SDXL-Turbo, SANA-Sprint, PixArt- α , and Flux.1 [schnell] models. We use the default parameters that were used for Flux.1 [schnell] [35, 36] in [46], since this setting is closest to our setup with one-step / few-step models. However, for SDXL-Turbo and PixArt, we use image resolutions of 512×512 , 768×768 for SANA-Sprint, and 512×512 for Flux.1 [schnell].

Ours. For all models and experiments, we set the regularization parameter $\lambda_{reg} = 0.01$ (see Eq. (3)). To obtain the results in Tabs. 1 and 4 and Fig. 4, we optimize for 100 iterations with the DINO and CLIPScore objectives until

reaching the mean CLIPScore and one standard deviation above the DINO diversity score obtained with [46]. Unless otherwise mentioned, we use a learning rate of 10.0 and gradient clipping of 0.1.

SDXL-Turbo. We generate images of resolution 512×512 . For the results in Tabs. 1 and 4 and Fig. 4, we use $\lambda_q = 50$ and $\lambda_{div} = 80$.

For the experiments that compare different diversity objectives (Tab. 2, qualitative results in Fig. 15, Fig. 16, and analyses in Sec. C.1, Sec. C.3), we set $\lambda_q = 10$ and $\lambda_{div} = 50$ and do 100 optimization iterations for the DINO, DPP and Vendi objectives. For the Color Histogram, LPIPS, and the L2 objective, we set $\lambda_{div} = 60$ and optimize for at most 60 iterations. For DreamSim, we set $\lambda_{div} = 70$ and use up to 50 optimization iterations.

Early stopping terminates optimization once the diversity objective surpasses a predetermined threshold. For the DreamSim, LPIPS, and DINO diversity objectives, we set the threshold to 0.9. For the L2 objective, we stop optimizing when the respective scores reach a value that more than doubles the initial value for i.i.d.-sampled noise initializations. The DPP, Vendi, and the Color histogram objectives end optimization when quadrupling their initial values. The different choices of stopping criteria and weights λ_{div} and λ_q arise from the fact that the corresponding metrics have different ranges and initial values.

PixArt- α . We generate images of resolution 512×512 . Similar to the SDXL-Turbo settings, we use $\lambda_q = 50$ and $\lambda_{div} = 80$.

SANA-Sprint. We generate images of resolution 768×768 . We set $\lambda_q = 10$ and $\lambda_{div} = 25$.

A.3. Datasets

GenEval [17] is a text-to-image generation benchmark that evaluates models across 553 diverse prompts requiring understanding of complex compositional relationships. Unless mentioned otherwise, we report results across all prompts in the dataset.

T2I-CompBench [45] tests compositional understanding in text-to-image models across eight distinct categories: color, shape, texture, spatial relationships, non-spatial attributes, complex compositions, 3D spatial reasoning, and numeracy. We select 50 random prompts per category, resulting in a set of 400 prompts.

B. Computational Cost

We measure the time per iteration on a single A100 80GB GPU in Tab. 6. Numbers reported are an average over 100 iterations with the error reported over three different seeds. It takes less than 15 iterations to reach similar levels of diversity as Parmar et al. [46] on GenEval [17] (see Fig. 7).

Table 6. Time per iteration of our proposed optimization approach. We report time on a single A100 80GB in seconds using the same optimization objectives as Tab. 5.

| Model | Time per Iteration |
|------------------|--------------------|
| SDXL-Turbo | 0.345 ± 0.004 |
| Flux.1 [schnell] | 1.092 ± 0.008 |

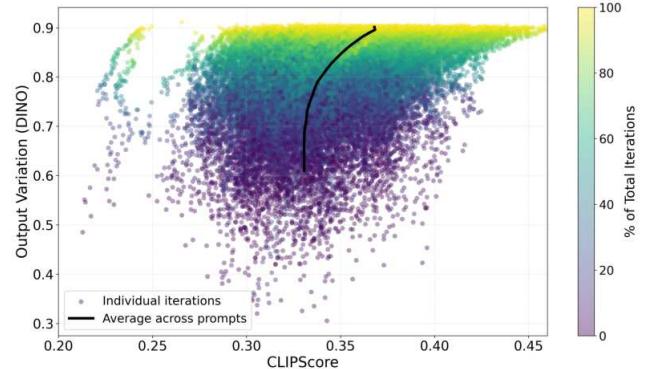


Figure 8. Scatter plot of CLIPScore and DINO diversity during optimization for SDXL-Turbo with white noise initialization on GenEval. Points are colored by iteration progress. The averaged trajectory (black) shows joint improvements in image quality and diversity, demonstrating that our method overcomes the quality–diversity trade-off.

C. Additional Experimental Results

C.1. Quality-Diversity Relationship

The scatter plot in Fig. 8 illustrates the relationship between image quality (measured by CLIPScore) and output diversity (DINO) throughout the optimization process for the white noise configuration on the GenEval dataset. The plot corresponds to the setup used for Fig. 7. Note that early stopping terminated optimization after 100 iterations or when the DINO diversity objective reached a threshold of 0.9.

Each point in the plot represents a single iteration across all prompts, colored by the percentage of total iterations completed (darker points indicate early iterations, lighter points indicate later stages). The black line shows the averaged trajectory across all prompts, revealing that both CLIPScore and DINO diversity increase jointly during optimization. This demonstrates that our approach overcomes the quality–diversity tradeoff described in [46]. Our improved output variation does not come at the expense of prompt alignment.

C.2. Noise Evolution Analysis

Here, we provide further analysis of the change in noise latents across iterations. In Fig. 12, we show the average noise

"A photo of a bear"

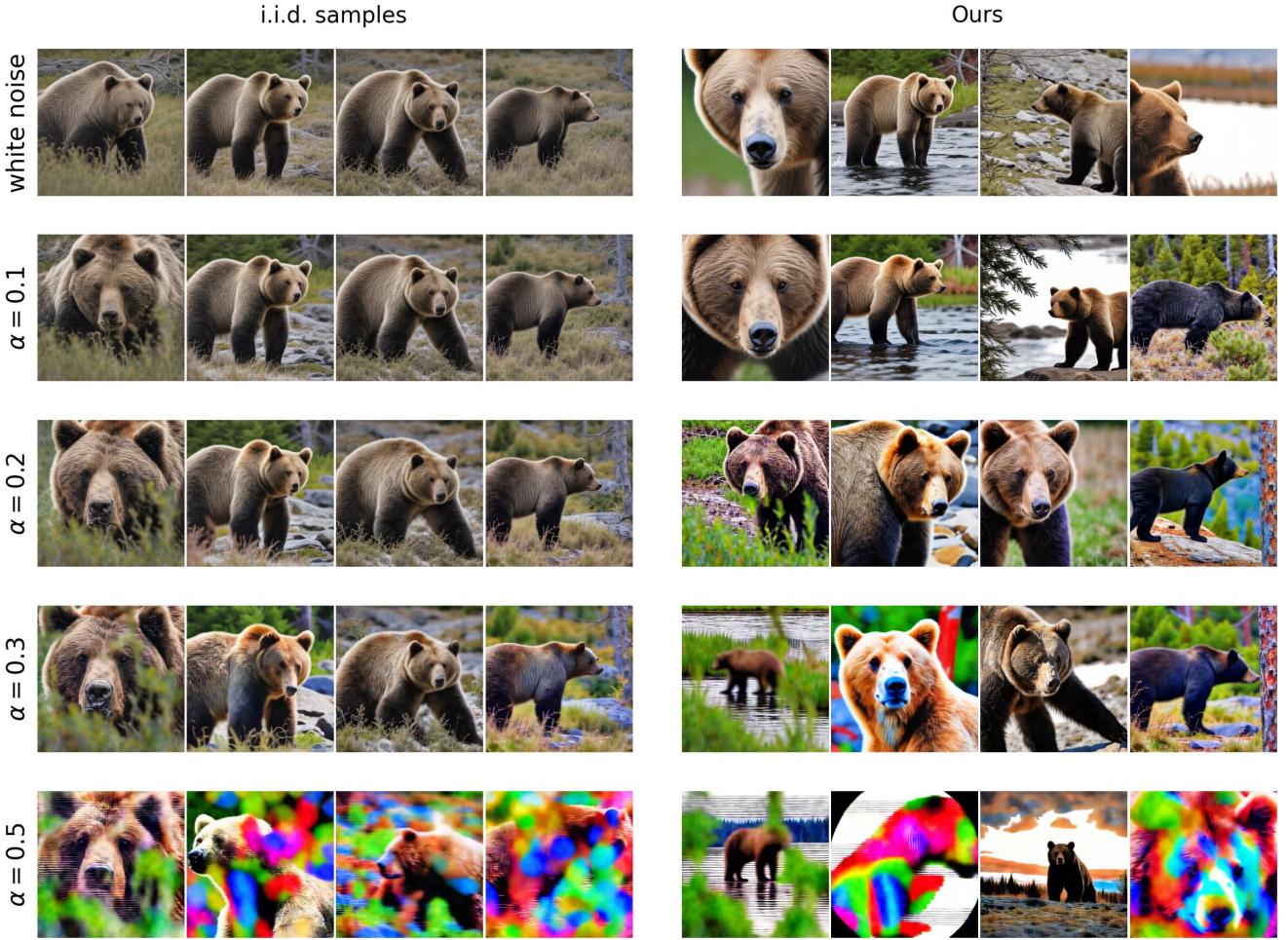


Figure 9. Effect of noise exponent values on image generation. Each row compares i.i.d. samples from initial noise (left) with our outputs (right) for different α values. Results were obtained with SDXL-Turbo and noise optimization using DINO diversity and CLIPScore.

change on the raw noise signal, measured by the L2 norm. The shaded regions around the lines indicate the standard deviation, showing the variability in noise change across different samples. We observe that the L2 norm increases steadily over iterations for white noise initializations.

The average norm change for white noise initializations is slightly lower for pink noise compared to white noise (Fig. 12). This confirms that using pink noise as initialization is favorable for our optimization.

We also analyse the spatial change in noise, both overall and decomposed into frequency bands Figs. 10 and 11 for SDXL-Turbo. The first column in Fig. 10 shows the images produced from randomly sampled white noise initializations. Subsequent columns show the intermediate outputs, with the final column displaying the images after optimization. For each iteration, we also visualize a heatmap of the noise change, computed as the averaged L2 difference

between the current latent and its initial value. Early in the process the heatmaps remain dark, indicating minimal deviation from the original noise. As optimization proceeds, brighter regions emerge in areas where the noise undergoes substantial modification. These regions align with the parts of the image that change the most (e.g. altered bird species or rearranged branches).

Furthermore, we visualize the noise evolution decomposed into frequency bands in Fig. 11. This visualization demonstrates that the low frequency components of the noise are being modified most significantly during the optimization process.

Noise Delta Computation. For each optimization step t , let $\mathbf{z}_t \in \mathbb{R}^{N \times C \times H \times W}$ be the noise. We define the noise change as $\Delta \mathbf{z}_t = \mathbf{z}_t - \mathbf{z}_{t-1}$, with \mathbf{z}_0 the initial noise. To

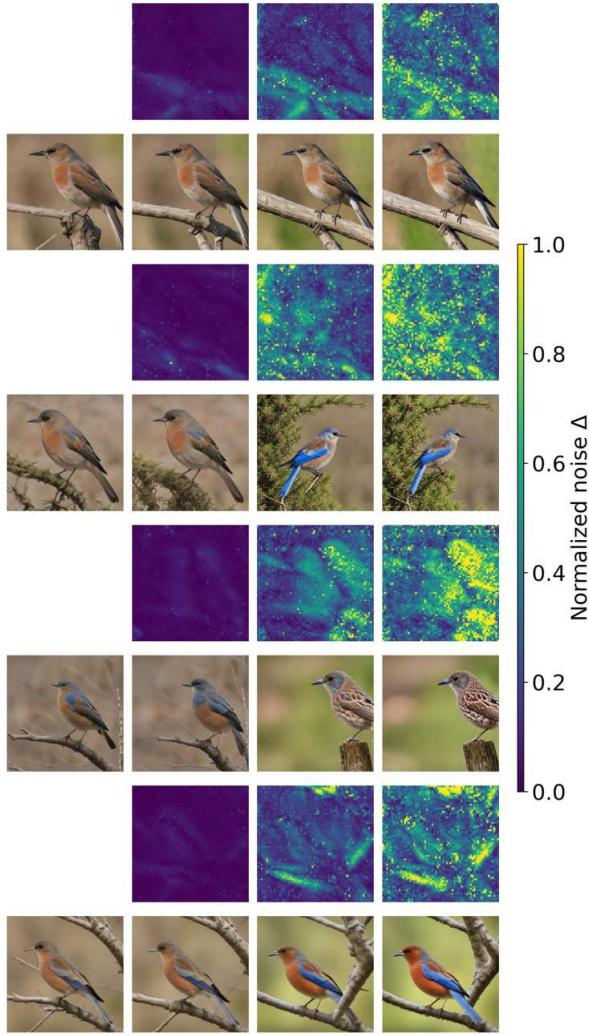


Figure 10. Noise evolution across optimization iterations for a set of four images. We show spatial heatmaps with the averaged L2 difference between the current noise latent and the initial white noise along with the corresponding generated image. Images were generated with SDXL-Turbo and the prompt: “A photo of a bird”.

visualize how the noise changes spatially, we compute

$$M_t(h, w) = \sqrt{\sum_{c=1}^C (\Delta z_t)_{c,h,w}^2}. \quad (11)$$

This results in a heatmap $M_t \in \mathbb{R}^{H \times W}$ showing the noise change at each location.

Frequency Band Decomposition. We decompose M_t into three frequency bands. For this, we compute the 2D FFT:

$$\mathcal{F}t(u, v) = \mathcal{F}\{M_t\}, \quad P_t(u, v) = |\mathcal{F}t(u, v)|^2,$$

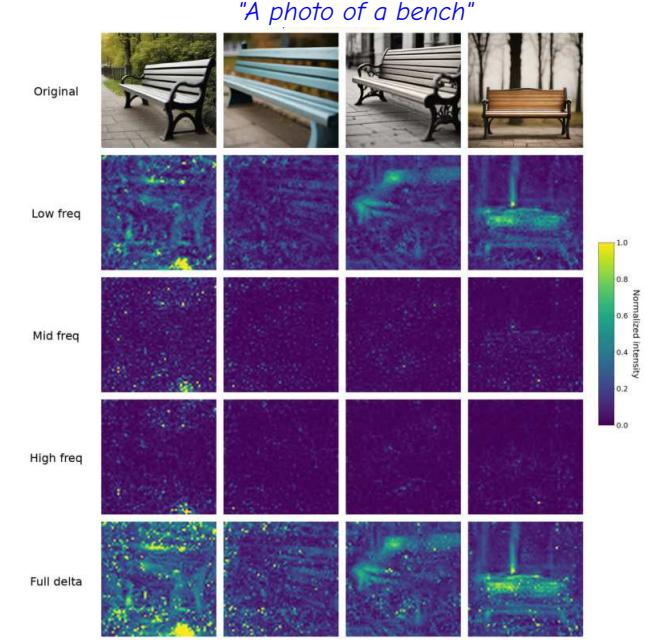


Figure 11. Example showing how the noise changes across optimization iterations in different frequency bands for SDXL-Turbo with white noise initialization and DINO diversity objective. We see that most of the change happens in the lowest third of the frequencies.

where (u, v) are frequency coordinates. The radial distance from the zero-frequency center is

$$r(u, v) = \sqrt{(u - u_c)^2 + (v - v_c)^2}, \quad (12)$$

and we define three frequency bins:

- Low: $[0, r_{\max}/3]$,
- Mid: $[r_{\max}/3, 2r_{\max}/3]$,
- High: $[2r_{\max}/3, r_{\max}]$,

for $r_{\max} = \sqrt{u_c^2 + v_c^2}$.

For each bin $b \in \{\text{low, mid, high}\}$, we apply a band-pass mask to the power spectrum:

$$P_t^{(b)}(u, v) = P_t(u, v) \cdot \mathcal{M}_b(u, v), \quad (13)$$

and scale the original FFT to preserve phase:

$$\mathcal{F}_t^{(b)}(u, v) = \mathcal{F}_t(u, v) \cdot \sqrt{\frac{P_t^{(b)}(u, v)}{P_t(u, v) + \epsilon}}, \quad \epsilon = 10^{-10}. \quad (14)$$

The spatial representation is obtained via the inverse FFT:

$$M_t^{(b)}(h, w) = \left| \mathcal{F}^{-1}\{\mathcal{F}_t^{(b)}\} \right|. \quad (15)$$

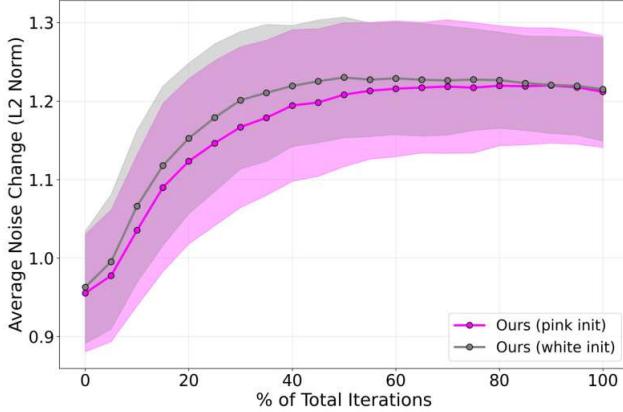


Figure 12. Noise change across iterations on raw noise signal measured as the L2 norm between subsequent iterations. White noise initialization results in slightly higher overall noise change across iterations than pink noise initialization.

We then normalize, so the frequency bands sum to the full magnitude:

$$\tilde{M}_t^{(b)}(h, w) = M_t^{(b)}(h, w) \cdot \frac{M_t(h, w)}{\sum_{b'} M_t^{(b')}(h, w) + \epsilon}. \quad (16)$$

This ensures $\sum_{b'} \tilde{M}_t^{(b')} = M_t$ at each pixel.

Visual Observations. Our noise evolution videos confirm that most noise change happens in the low frequency components. These changes directly correspond to spatial changes in the generations throughout the optimization steps. This observation along with the fact that natural images have a 1/f power spectrum inspires our exploration of noise initializations with stronger low-frequency components (e.g. pink noise).

C.3. Pink Noise Example Generations

Higher α values (see Eq. (7)) generally lead to higher diversity scores. However, the image quality decreases with high noise exponents (see generations for $\alpha = 0.3$ and $\alpha = 0.5$ in Fig. 9 which have patchy artefacts). Note, that we use CLIPScore as the only image quality reward during optimization for Tab. 4. However, additional rewards for image quality could easily be included in our pipeline, but would naturally increase computational cost.

In our experiments, we use a noise exponent of 0.2 (referred to as pink noise), which provides substantial gains in sample diversity and reduces the number of required iterations, while preserving image quality.

C.4. User Study

We conduct a human user preference study to determine which methods produce more diverse outputs, similar to

Table 7. Output diversity results on the single-object subset of GenEval for our proposed approach with the PixArt- α , SANA-Sprint-1.6B, and SDXL-Turbo models using white noise initialization. Output diversity is measured with averaged pairwise DINO, DreamSim, and LPIPS scores.

| Method | DINO | DreamSim | LPIPS |
|---------------------------------------|-------------------|-------------------|-------------------|
| PixArt-α [8] | | | |
| i.i.d. | 0.382 ± 0.093 | 0.160 ± 0.078 | 0.460 ± 0.126 |
| Parmar et al. [46] | 0.520 ± 0.093 | 0.227 ± 0.094 | 0.563 ± 0.116 |
| Ours | 0.731 ± 0.077 | 0.370 ± 0.117 | 0.691 ± 0.096 |
| SANA-Sprint-1.6B [9] | | | |
| i.i.d. | 0.494 ± 0.091 | 0.219 ± 0.081 | 0.631 ± 0.070 |
| Parmar et al. [46] | 0.695 ± 0.061 | 0.363 ± 0.112 | 0.733 ± 0.052 |
| Ours | 0.752 ± 0.065 | 0.485 ± 0.109 | 0.795 ± 0.058 |
| SDXL-Turbo [56] | | | |
| i.i.d. | 0.529 ± 0.077 | 0.218 ± 0.089 | 0.611 ± 0.058 |
| Parmar et al. [46] | 0.667 ± 0.069 | 0.320 ± 0.118 | 0.661 ± 0.053 |
| Ours | 0.808 ± 0.047 | 0.450 ± 0.131 | 0.768 ± 0.046 |

Table 8. Human preference win rates from a user study for our method against i.i.d. sampling and Parmar et al. [46] for PixArt- α [8], SANA-Sprint-1.6B [9], and SDXL-Turbo [56].

| Method | Win % vs i.i.d. | Win % vs [46] |
|----------------------|-----------------|---------------|
| PixArt- α [8] | 90.00 | 77.50 |
| SANA-Sprint-1.6B [9] | 85.00 | 66.25 |
| SDXL-Turbo [56] | 88.75 | 91.25 |

Parmar et al. [46]. We compare our method to baselines such as i.i.d. sampling and Parmar et al. [46], as well as across different target diversity objectives.

During the study, we show participants a 2x2 grid of images generated from our method and a comparison. We ask the user to select “which grid of images has higher variety?”. For each pairing, we collect 10 user preferences to determine a per prompt win rate. User data is anonymized and crowd-sourced.

We run trials across all single object prompts in the GenEval benchmark [17] (prompts 1 to 80). For reference, we also report diversity scores for this subset in Tab. 7. We count the number of wins across trials for each model to compute a final overall win percentage. In the results in Tab. 8, we observe that our method shows the highest win rate across all three models.

In addition, we compared our method across different diversity objectives (see Tab. 3).

C.5. Failure Cases

Despite the effectiveness of our optimization approach, several failure modes can be observed. We visualize these in Fig. 13. When using DreamSim, the optimization some-

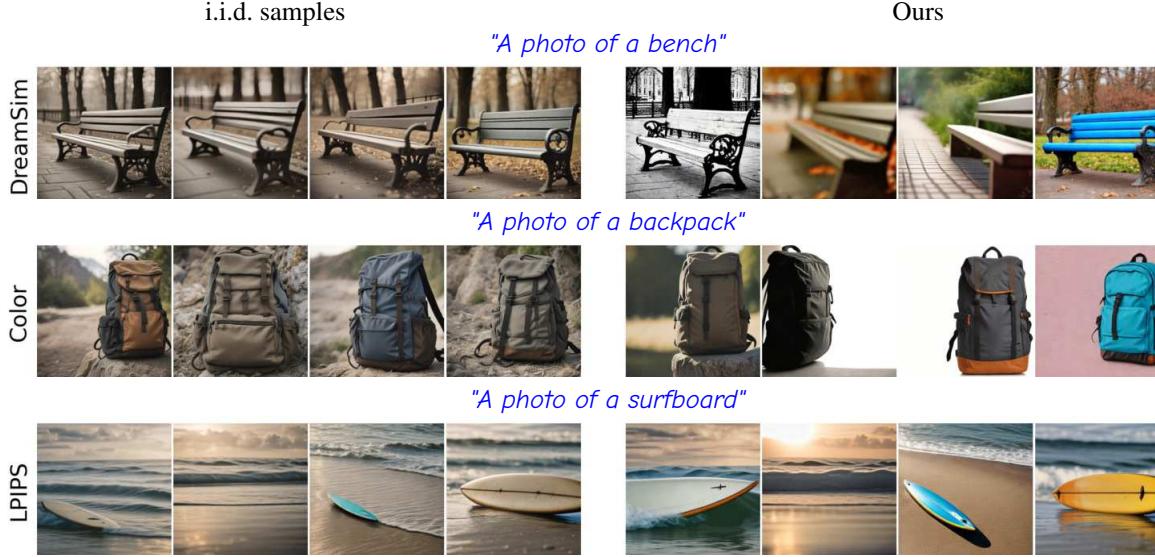


Figure 13. Failure cases of our method for different optimization objectives (SDXL-Turbo). Top row: Removing fine details through blurring one image increases perceptual distance without introducing meaningful diversity. Middle row: Overly simple compositions (e.g. plain backgrounds) lead to high color diversity scores as different solid colors maximize L2 color histogram distance effectively. Bottom row: LPIPS optimization fails to recover semantic content that is missing in the generation from the initial noise.

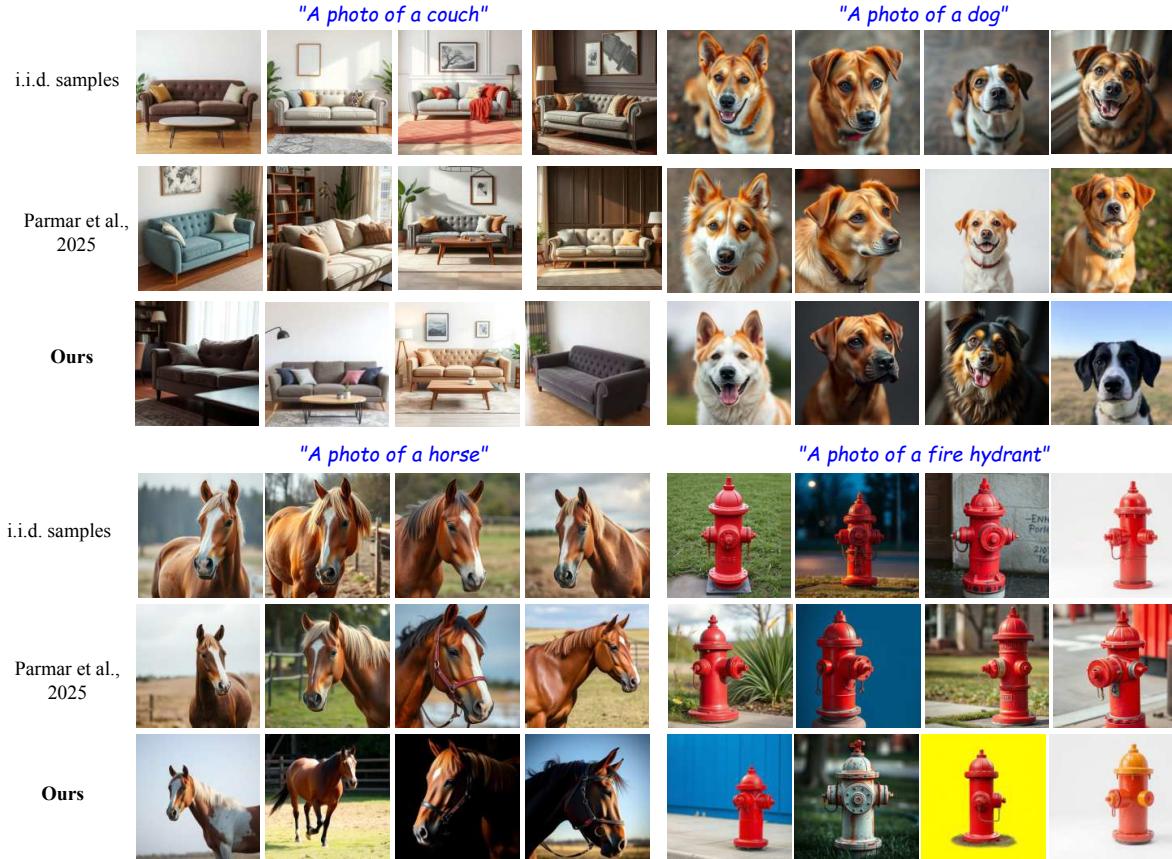


Figure 14. Image generations applying our method to Flux.1 [schnell] [36]. We achieve greater visual diversity compared to baselines while maintaining image quality.

times produces blurry images as the method exploits perceptual distance which can remove high-frequency details (top row). Color histogram diversity tends to encourage plain backgrounds since uniform color regions efficiently maximize histogram L2 distances. LPIPS diversity exhibits a critical limitation: it does not recover semantic content missing from the initial noise visualization (e.g., if a surfboard is not generated at first, it remains absent), as LPIPS diversifies existing perceptual features rather than introducing new semantic elements. This could be recovered with a larger weighting of image quality and prompt adherence rewards in the optimization process.

C.6. Qualitative Results for Flux.1 [schnell]

Finally we test our optimization on a larger model, Flux.1 [schnell]. Using the best diversity objective from our ablations DPP, we generate results in Fig. 14. Compared to i.i.d. sampling and the default settings from [46], we observe greater output diversity across multiple prompts, particularly in terms of object color, orientation, lighting, and also different backgrounds and positioning.

C.7. Qualitative Results for Different Objectives

We show generation results that compare different diversity objectives in Fig. 15 and Fig. 16. We can observe that our approach yields more diverse image output sets compared to [46] and generations from i.i.d.-sampled noise initializations across different diversity objectives. All generations are obtained from white noise initializations using the SDXL-Turbo model.

"A photo of a horse"

"A photo of a fire hydrant"

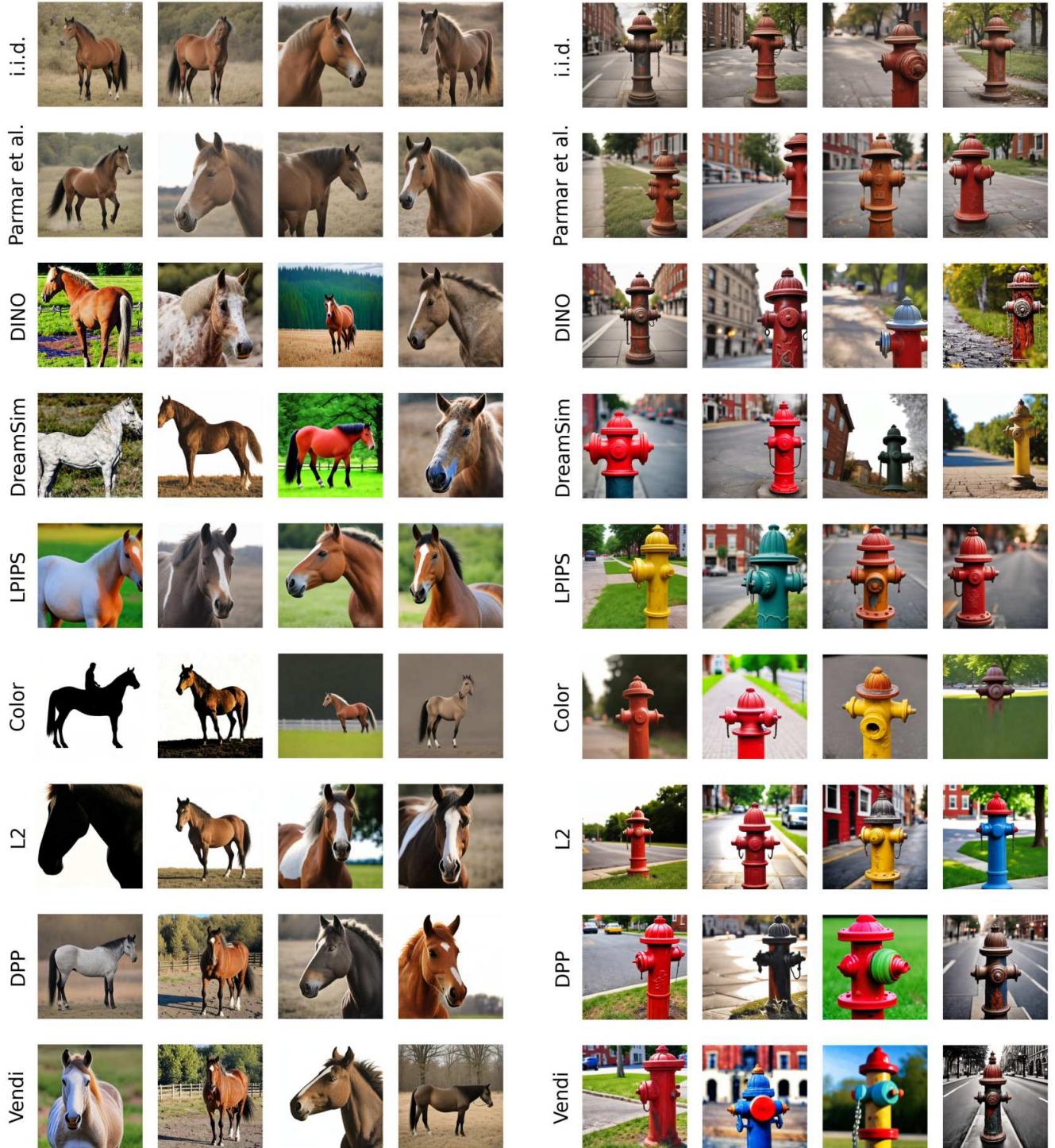


Figure 15. Impact of diversity objectives on the resulting noise optimization and image generations compared to i.i.d sampled noise initialization and the search method proposed by Parmar et al. [46]. Our approach results in more varied generations in terms of object pose, appearance, colors, and backgrounds (e.g. different horse breeds in different surroundings, and fire hydrants in different colors).

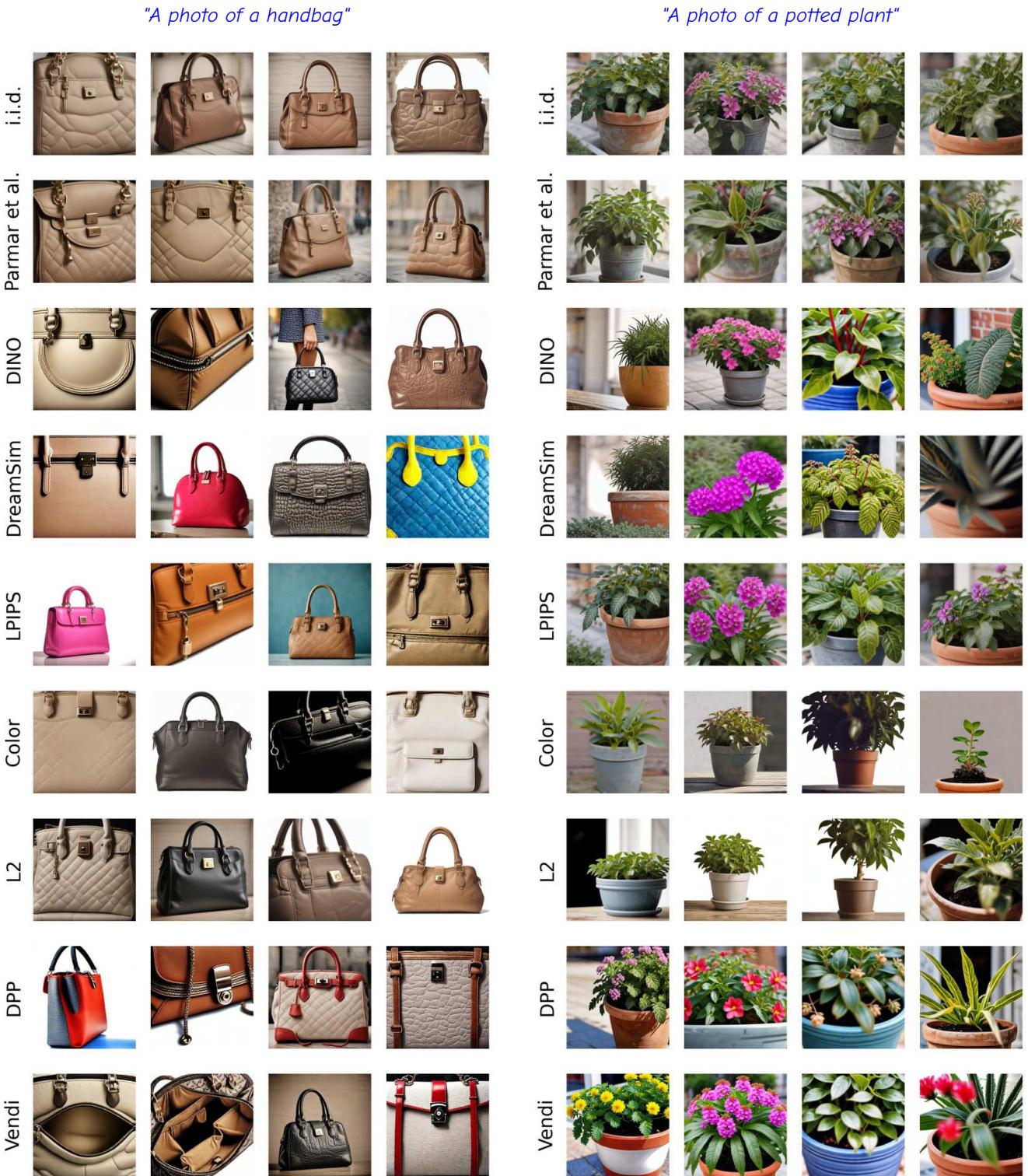


Figure 16. Impact of diversity objectives on the resulting noise optimization and image generations compared to i.i.d sampled noise initialization and the search method proposed by Parmar et al. [46]. The generated handbags and potted plants show larger variation in terms of handbag types and colors, and plant species.