CONTROLLING IMAGE-SPACE DRIFT IN INFERENCE-TIME NOISE OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Inference-time alignment steers pre-trained diffusion models toward target rewards without updating model weights. Among these methods, optimizing the input noise is especially powerful: by directly modifying the generation process, it can reach rare, high-reward samples. However, this flexibility also makes reward hacking significantly worse, often producing high-scoring images that lose the structure and realism of the base model. Existing approaches try to prevent this by constraining the optimized noise, but we show this is insufficient: the denoising process can amplify small changes in noise into large differences in the generated images. Thus, we propose BIDS (Bounding Image-space Divergence via Score functions), which uses a score-based constraint to limit image-space drift during noise optimization. BIDS uses the Data Processing Inequality to relate image-space divergence to divergence over denoising trajectories, then computes a practical surrogate using only the frozen model's score function. This preserves image structure during reward optimization. Across UNet and DiT architectures, BIDS reduces distributional drift (CMMD) by over 75% and achieves up to an 80.30% human preference win rate. We further extend BIDS to pairwise preferences, enabling alignment to non-differentiable human feedback.

1 INTRODUCTION

Inference-time alignment aims to steer a frozen diffusion model toward an external reward without updating model parameters. A common strategy is to search over multiple generated trajectories, as in Best-of-N sampling, particle-based steering, or related inference-time control methods (Nakano et al., 2021; Song et al., 2023; Kim et al., 2025; Singhal et al., 2025; Singh et al., 2025). These methods can improve reward while largely preserving the base model's generation process, but their compute and memory costs grow with the number of trajectories. Furthermore, they are inherently limited by the base model's proposal distribution. In contrast, noise optimization offers a more direct single-trajectory alternative. Instead of sampling many candidates and selecting among them, it treats the stochastic variables of the diffusion sampler as optimizable inputs and backpropagates reward gradients through the denoising process (Tang et al., 2025; Guo et al., 2024). This gives the optimizer fine-grained control over the generated output while keeping the model weights frozen. In principle, optimization can reach high-reward regions that are simply too rare for sampling-based methods to find by chance.

However, the same flexibility that makes noise optimization powerful also makes it vulnerable to reward hacking (Amodei et al., 2016). Because the optimizer directly modifies the noise sequence, reward maximization pushes the denoising trajectory outside the base model's learned distribution. The resulting image achieves a high reward but loses the semantic content of the prompt. For example, optimizing an aesthetic reward for the prompt *"generate an image of a spider"* produces a high-scoring artifact that no longer preserves a coherent spider structure (Fig. 1a). Thus, the central challenge in inference-time optimization is to maximize reward while minimizing divergence from the base image distribution.

A natural solution is to impose a trust-region constraint directly in image space by penalizing divergence between the optimized and base image distributions (Ouyang et al., 2022; Rafailov et al., 2023). Unfortunately, this image-space divergence is intractable to compute exactly for diffusion models, because the final image distribution is obtained by marginalizing over the full

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

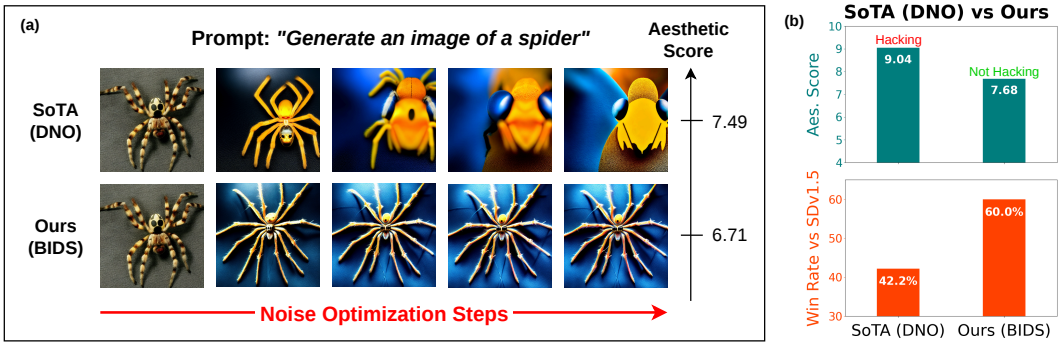

Figure 1: **BIDS reduces reward hacking by preserving the prompt's semantic structure during optimization.** **(a)** Prompt: *"generate an image of a spider"* + Aesthetic Score. Standard noise-space optimization attains high reward but generates a non-spider artifact (structural collapse). BIDS preserves the spider structure. **(b)** Trade-off: BIDS (bottom) achieves higher win rates against SDv1.5 by restricting the structural collapse seen in noise-space baselines (top).

multi-step denoising trajectory (Ho et al., 2020; Song et al., 2021b). Existing noise optimization methods therefore use a more convenient proxy: they regularize the noise sequence itself, penalizing deviations from the Gaussian prior or from the initial noise (Tang et al., 2025; Guo et al., 2024). This approach assumes that proximity in noise space implies proximity in image space. We show that this assumption is false: the denoising process can amplify small perturbations in the optimized noise sequence into large semantic changes in the output.

To address this, we focus on constraining the output distribution directly. We propose **BIDS** (**B**ounding **I**mage-space **D**ivergence via **S**core functions), an inference-time objective that preserves the efficiency of single-trajectory noise optimization while constraining image-space drift. BIDS begins with the desired image-space KL divergence between the optimized and base image distributions. Since this quantity is intractable, we apply the Data Processing Inequality to upper-bound it by the KL divergence between the corresponding trajectory distributions. We then use the frozen model's marginal score to define a computable score-based surrogate for trajectory KL. The resulting constraint requires no model fine-tuning, no auxiliary alignment network, and no numerical marginalization over image distributions.

Empirically, BIDS mitigates the reward hacking caused by direct noise optimization while retaining its inference-time flexibility (Fig. 1b). Across diverse rewards, prompts, and model architectures, BIDS improves the reward-fidelity trade-off relative to baselines. The same trajectory-divergence view also extends to discrete preferences: by treating trajectory log-likelihood ratios as implicit rewards, we derive an inference-time DPO-style objective for frozen diffusion models.

We summarize our contributions as follows:

- **Failure of noise-space regularization.** Noise optimization provides strong test-time control, yet pushes samples away from the base image distribution. Proposition 1 shows that a noise-space bound alone does not control how much the sampler amplifies a perturbation.
- **A tractable score-based image constraint (BIDS).** We apply the Data Processing Inequality to upper-bound the intractable image-space KL divergence by a trajectory KL. We use a computable score-energy surrogate from the frozen base model's marginal score, enabling distributional regularization without fine-tuning.
- **Inference-time preference optimization.** For discrete or non-differentiable rewards, we derive a DPO-style objective in which trajectory log-likelihood ratios serve as implicit rewards. This implicit formulation avoids numerical gradient estimation through the diffusion chain.
- **Empirical validation across architectures.** Across diverse UNet and DiT backbones, BIDS reduces CMMD by over 75% compared to noise-space baselines and achieves up to an 80.30% human preference win rate.

2 Related Work

Training-based and multi-trajectory alignment. Alignment methods update model parameters through reinforcement learning, preference optimization, or regression (Prabhudesai et al., 2023; Clark et al., 2024; Black et al., 2023; Fan et al., 2024; Deng et al., 2024; Wallace et al., 2024; Gu et al., 2024; Uehara et al., 2024), or train auxiliary networks for noise prediction (Zhou et al., 2025). Inference-time search instead evaluates multiple trajectories using Best-of-N (Nakano et al., 2021), Feynman-Kac steering (Song et al., 2023; Kim et al., 2025; Singhal et al., 2025), DAS (Kim et al., 2025), or GLASS Flows (Holderrieth et al., 2026). Parallel trajectories increase peak memory with trajectory count, motivating single-trajectory alternatives. Concurrent noise-optimization methods target diversity across generations (Harrington et al., 2026; Kim et al., 2026), whereas BIDS targets reward-induced image-space drift.

Single-trajectory inference-time alignment. PAG (Ahn et al., 2024), SEG (Hong, 2024), and FreeDoM (Yu et al., 2023) alter diffusion transitions rather than optimize the noise sequence. Like Classifier-Free Guidance (Ho & Salimans, 2022), these approaches are orthogonal to BIDS. Gradient-based methods (Tang et al., 2025; Eyring et al., 2024) instead optimize noise to reach rare, high-reward images. Their noise-space constraints can still permit structural collapse because denoising amplifies small noise updates. Score-space methods like Adjoint Matching (Domingo-Enrich et al., 2025) apply strict vector-matching objectives that penalize necessary angular deviations. BIDS retains the constant memory footprint of single-trajectory optimization while constraining image-space drift with our score-based KL surrogate.

3 Problem Formulation

Preliminaries. Consider a conditional diffusion model with parameters θ that generates an image x_0 in T denoising steps conditioned on a prompt c. While early methods optimized only the initial noise x_T, recent inference-time methods like Direct Noise Optimization (DNO) (Tang et al., 2025) optimize rewards by treating the full stochastic sampling process (SDE) as a differentiable mapping from noise to image. In this setting, the optimization variable $z = \{x_T, \zeta_T, \ldots, \zeta_1\}$ is a concatenated, high-dimensional sequence containing the initial noise and all noise added to intermediate denoising steps. We sample an initial noise sequence $z_{\text{init}} \sim \mathcal{N}(0, \mathbf{I})$ and initialize $z = z_{\text{init}}$. Holding z_{init} fixed, we update z via gradient ascent to maximize a differentiable target reward $r(x_0, c)$:

$$z \leftarrow z + \alpha \nabla_z r(x_0, c), \tag{1}$$

where α is the learning rate. Unconstrained optimization pushes z into low-probability regions and leads to structural collapse in generated images. This is often addressed through noise-space regularization, such as constraining noise to high-likelihood regions of the Gaussian prior or, more restrictively, to a small L_2 ball around the initial noise.

Small noise perturbations can cause large image changes. We analyze the limitations of noise-space regularization through the following constrained optimization problem:

$$\arg\max_z \mathbb{E}_{x_0 \sim p_\theta(\cdot|z,c)}\big[r(x_0, c)\big] \quad \text{s.t.} \quad \|z - z_{\text{init}}\|_2 \leq \epsilon, \tag{2}$$

which constrains optimized noise sequence z to a Euclidean ball of radius ϵ centered at z_{init}. Since z does not directly represent image semantics (Li & He, 2025), closeness in noise space does not imply closeness in image space. We demonstrate this limitation with a linear sampler, which lets us calculate exactly how noise changes affect the final image. Proposition 1 shows that even in this simple setting, small noise changes can produce large image changes.

Proposition 1 (Amplification of Noise Perturbations by a Linear Sampler). *Consider the linear sampler $x_{t-1} = Ax_t + \sigma\zeta_t$ for $t = T, \ldots, 1$, where $A \in \mathbb{R}^{d\times d}$ is the sampler's weight matrix, applied at each denoising step, and $\sigma > 0$ scales the injected noise ζ_t. For any fixed $\epsilon > 0$, changes to the full noise sequence satisfying $\|z - z_{\text{init}}\|_2 \leq \epsilon$ can produce arbitrarily large image changes across choices of A.*

Appendix A.3 provides the proof and derives a noise-space penalty equivalent to image-space KL for this linear sampler. Unlike the Euclidean distance between noise sequences, this penalty accounts

for how the sampler combines changes to the initial and injected noises. However, for nonlinear denoisers, the effect of a noise update on the final image also depends on the intermediate latents. Updating the noise changes these latents and the model's subsequent predictions. The distance between noise sequences does not measure how much these model predictions change. We therefore target changes in the generated image distribution rather than constrain noise distance alone.

4 Proposed Approach

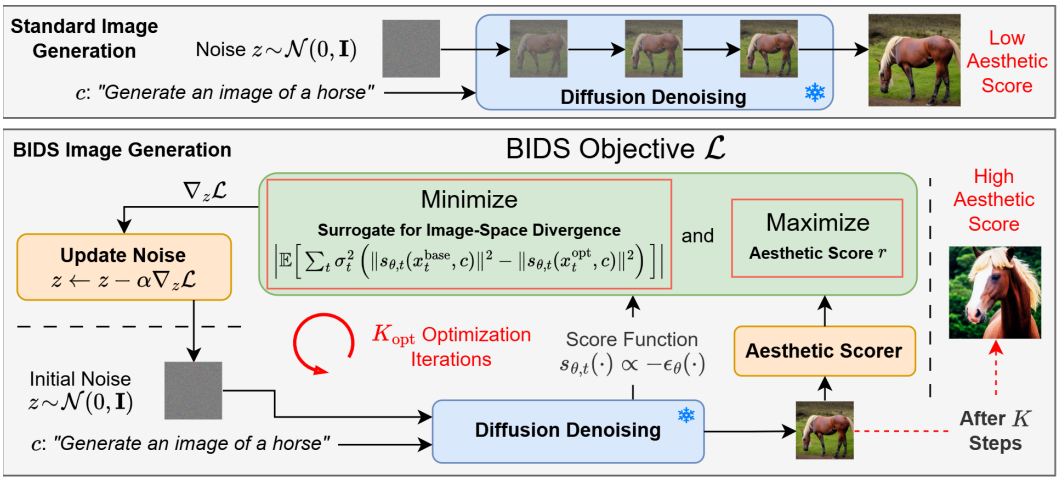

Figure 2: **BIDS enables reward optimization without structural collapse via score-based image constraints. Top:** Standard diffusion generation, where a frozen model maps initial noise to an image. **Bottom:** Noise optimization with BIDS. For a fixed prompt and frozen backbone, the algorithm evaluates the generated image against a target reward (e.g., Aesthetic Score) and performs gradient updates directly on the noise sequence. To prevent the optimized trajectory from deviating from the base model distribution and inducing structural collapse, the inference objective incorporates a tractable, score-based divergence penalty. Iterating this gradient update across successive steps yields robust inference-time alignment without updating model weights.

Trajectory KL bounds image-space divergence. Proposition 1 shows that small noise perturbations can cause large changes in the generated image. We therefore penalize the KL divergence between the optimized and base image distributions. Let $x_{0:T} = (x_0, \ldots, x_T)$ denote a denoising trajectory. Let $p^{\text{base}}(x_{0:T} \mid c)$ denote the distribution of trajectories generated from $z_{\text{init}} \sim \mathcal{N}(0, \mathbf{I})$, and let $p^{\text{opt}}(x_{0:T} \mid c)$ denote the distribution generated from the corresponding optimized noise. The corresponding image distributions, $p^{\text{base}}(x_0 \mid c)$ and $p^{\text{opt}}(x_0 \mid c)$, are their endpoint marginals. Computing the image-space KL directly is intractable because it requires marginalizing over the intermediate denoising latents $x_{1:T}$. Since x_0 is the endpoint of $x_{0:T}$, the Data Processing Inequality bounds the image-space KL by the trajectory KL:

$$d_{\text{KL}}\left[p^{\text{opt}}(x_0 \mid c) \| p^{\text{base}}(x_0 \mid c)\right] \leq d_{\text{KL}}\left[p^{\text{opt}}(x_{0:T} \mid c) \| p^{\text{base}}(x_{0:T} \mid c)\right]. \tag{3}$$

This motivates the idealized minimization objective:

$$\mathcal{L}_{\text{ideal}} := \underbrace{-\mathbb{E}_{x_0 \sim p^{\text{opt}}(\cdot|c)}[r(x_0, c)]}_{\text{Expected Reward}} + \beta \underbrace{d_{\text{KL}}\left[p^{\text{opt}}(x_{0:T} \mid c) \| p^{\text{base}}(x_{0:T} \mid c)\right]}_{\text{Trajectory KL Penalty}}, \tag{4}$$

where $\beta > 0$ controls the regularization strength. The trajectory KL penalizes optimized trajectories that are unlikely under the base model, providing a distributional constraint on reward optimization.

Score functions provide a computable surrogate for trajectory KL. Directly evaluating the trajectory KL would require evaluating the density of optimized trajectories, which is intractable. At timestep t, let x_t^{base} and x_t^{opt} denote latents before and after noise optimization, respectively. The marginal score is:

$$s_{\theta,t}(x_t, c) = \nabla_{x_t} \log p_t^\theta(x_t \mid c),$$

where p_t^θ is the frozen model's marginal density at timestep t. We obtain this score estimate from the frozen model's noise prediction, without evaluating the marginal density. At denoising step t, the model's score $s_{\theta,t}(x_t, c)$ points toward locally higher density under the base model. It provides local information about the model's distribution that noise distance alone does not provide. Noise optimization changes the denoising latents and hence the scores evaluated along the trajectory.

The use of squared norms has a direct interpretation for an isotropic Gaussian base distribution: a larger score norm $\|s_{\theta,t}(x_t, c)\|$ corresponds to a lower base-model marginal density $p_t^\theta(x_t \mid c)$. Motivated by this connection between score norm and density, we compare the squared score norms $\|s_{\theta,t}(x_t^{\text{base}}, c)\|^2$ and $\|s_{\theta,t}(x_t^{\text{opt}}, c)\|^2$ before and after optimization. We combine these comparisons across denoising steps, weighting each step by its transition noise variance. This gives the following surrogate for trajectory KL:

$$\mathcal{L}_{\text{BIDS}} = -\mathbb{E}_{z_{\text{init}}}\left[r(x_0^{\text{opt}}, c)\right] + \beta\left|\mathbb{E}_{z_{\text{init}}}\left[\underbrace{\sum_{t=0}^{T-1}\sigma_t^2\|s_{\theta,t}(x_t^{\text{base}}, c)\|^2}_{S_0} - \underbrace{\sum_{t=0}^{T-1}\sigma_t^2\|s_{\theta,t}(x_t^{\text{opt}}, c)\|^2}_{S(z,c)}\right]\right|, \quad (5)$$

where σ_t^2 is the noise variance of the transition producing x_t, and both expectations are over the initial noise draw. Here S_0 is the weighted score sum along the base trajectory, and $S(z, c)$ is the weighted score sum along the optimized trajectory. Algorithm 1 uses the paired penalty $|S_0 - S(z, c)|$, whose expectation upper-bounds $|\mathbb{E}_{z_{\text{init}}}[S_0 - S(z, c)]|$ by Jensen's inequality. It computes S_0 once and holds it fixed during optimization (Appendix A.1).

4.1 Discrete Preference Optimization via BIDS

The continuous objective defined in Eq. 5 assumes access to a differentiable reward function. However, human preference feedback and rewards such as JPEG compressibility do not provide gradients for noise optimization. Unlike DPO-based fine-tuning methods such as D3PO (Yang et al., 2024), BIDS optimizes individual generations for non-differentiable rewards while keeping the diffusion model frozen. We use the reward to rank image pairs generated during optimization and differentiate our score-based implicit reward to update their noise sequences.

Following Direct Preference Optimization (DPO) (Rafailov et al., 2023), we express an implicit reward r^* through the ratio of the optimized and base trajectory densities (Appendix A.4):

$$r^*(x_{0:T}, c) \propto \log \frac{p^{\text{opt}}(x_{0:T} \mid c)}{p^{\text{base}}(x_{0:T} \mid c)}. \quad (6)$$

Let z^w generate the preferred image along trajectory $x_{0:T}^w$, and let z^l generate the rejected image along trajectory $x_{0:T}^l$. The Bradley–Terry loss uses the difference in their implicit rewards, with σ denoting the logistic sigmoid:

$$\mathcal{L}_{\text{BIDS}} = -\mathbb{E}\Big[\log\sigma\Big(r^*(x_{0:T}^w, c) - r^*(x_{0:T}^l, c)\Big)\Big]. \quad (7)$$

The expectation is over preference pairs.

Score functions provide a computable approximation. Evaluating Eq. 6 directly is intractable. Using the same sum of squared score norms as in Eq. 5, let $S(z, c)$ denote the sum along the trajectory generated from z. For each noise sequence, we compute the reference sum S_0 from the unoptimized noise z_{init} and keep it fixed. We define the implicit-reward surrogate as:

$$\widehat{r}(z, c) := \beta\,(S_0 - S(z, c))\,. \quad (8)$$

Substituting Eq. 8 into Eq. 7 allows gradient computation ($\nabla_z\mathcal{L}$) purely via score function evaluations. This extends the Bradley–Terry objective to inference-time noise optimization using the same score-energy surrogate.

5 Experiments

Experimental Setup: Models, Baselines, Metrics. We evaluate BIDS across four diffusion backbones: standard UNet models (SDv1.5 (Rombach et al., 2022), SDXL (Podell et al., 2023)) and

single-step architectures (SDXL-Turbo (Sauer et al., 2024), PixArt-α DMD (Chen et al., 2024a)). Standard models use 100-step DDIM sampling (Song et al., 2021a) ($\eta = 1$, guidance scale 5), while single-step models use default samplers. Our main experiments use 50 optimization steps for all noise optimization methods. Appendix A.8 compares BIDS and DNO across optimization budgets and regularization strengths.

Baselines span single-trajectory noise optimization (DNO (Tang et al., 2025), InitNO (Guo et al., 2024), ReNO (Eyring et al., 2024), DyMO (Xie & Gong, 2024)), training-based alignment (DDPO (Black et al., 2023), Diffusion-DPO (Wallace et al., 2024), D3PO (Yang et al., 2024)), and Best-of-N (Nakano et al., 2021). All baseline hyperparameters follow the recommendations in their original implementations (Appendix A.1). Evaluations utilize the Animal dataset (Black et al., 2023), the Animal-Animal and Animal-Object suites (Guo et al., 2024), and HPDv2 (Wu et al., 2023a). All prompts follow the template: *"generate an image of a* [PROMPT]*"*. We prioritize head-to-head win rates (Kirstain et al., 2023; Zhang et al., 2023) over scalar rewards, as reward hacking renders raw numerical scores unreliable proxies for visual quality. Unless identified as human evaluations, win rates use GPT-4o as the judge (Appendix A.10). We report raw reward values and CLIPScores in Appendix A.6.

5.1 BIDS Reduces Reward Hacking and Drift

We compare BIDS against DNO using human-aligned reward models (Aesthetic Score (Schuhmann et al., 2022), HPSv2 (Wu et al., 2023b), PickScore (Kirstain et al., 2023)) and custom image brightness and darkness rewards to isolate reward hacking behavior.

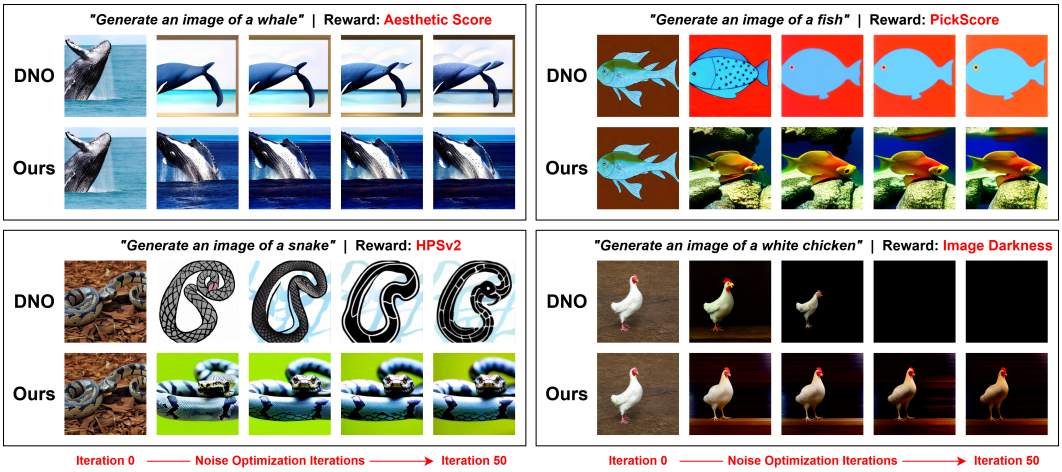

Figure 3: **BIDS mitigates structural collapse across human-aligned and synthetic rewards.** We evaluate four target objectives over 50 noise optimization steps using the SDv1.5 backbone on prompts from the Simple Animals dataset. For each prompt, DNO (top row) exhibits oversaturation artifacts, loss of detail, or complete structural collapse (e.g., the black output for Darkness). In contrast, BIDS (bottom row) enforces an image-distribution constraint, achieving high rewards while preserving the prompt's semantic structure and improving base color richness.

While DNO optimizes target rewards, it often fails to maintain the semantic elements dictated by the prompt. As shown in Figure 3, DNO produces severe texture artifacts and structural collapse under reward maximization (e.g., the black output for Darkness). In contrast, BIDS enforces an image-distribution constraint, preserving key visual concepts while achieving high rewards. Additional qualitative comparisons on SDv1.5 and SDXL-Turbo appear in Appendix A.5.

BIDS limits image-space drift during optimization. We analyze the distributional drift during optimization by computing a Monte Carlo estimate of our score-based surrogate for trajectory KL, averaged over all 45 prompts in the Simple Animals dataset. Appendix A.7 compares this surrogate with KDE-estimated KL during noise optimization in a controlled diffusion experiment. Figure 4a shows that DNO's estimated image-space drift increases with optimization steps, whereas BIDS

keeps estimated drift near zero while improving reward. This behavior is consistent with the structural preservation shown in Figure 3. This demonstrates that unlike noise-constrained baselines, BIDS remains robust to reward hacking even as optimization steps scale. We use 50 optimization steps in the main experiments to balance reward improvement and computational cost.

Figure 4b shows how β controls the trade-off between reward and image-space drift. Moderate regularization sharply reduces the drift surrogate while retaining reward gains. The Darkness ablation in Figure 11 (Appendix A.8) shows the complementary trade-off in prompt fidelity: stronger regularization generally preserves higher CLIPScore but reduces reward. Table 7 further shows similar CLIPScore across a wider β range, indicating that prompt alignment does not require narrow tuning. We therefore choose a moderate β for each reward–backbone setting. We also measure image-space drift independently of the score-based surrogate using CMMD between optimized and base-model images. When optimizing Aesthetic Score on SDv1.5, BIDS obtains a CMMD of 0.063 against the base model at a reward of 7.684; DNO obtains a CMMD of 0.281 at a reward of 9.044. Each method is evaluated using 900 images—20 samples for each of the 45 Simple Animals prompts—against a 900-image base-model reference.

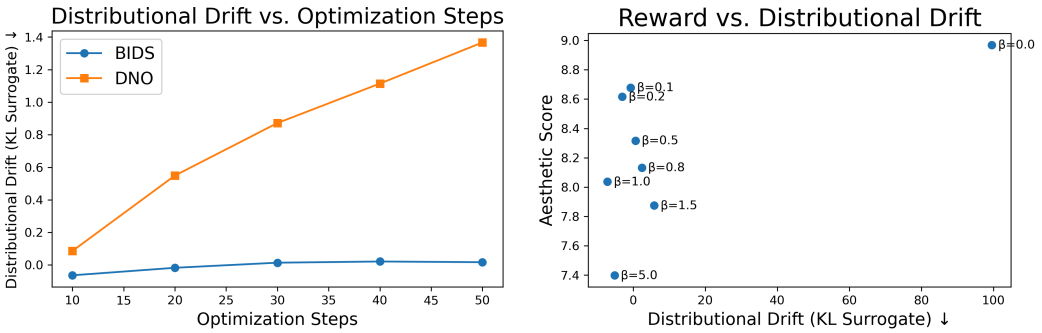

Figure 4: **BIDS improves reward while keeping image-space drift low.** We plot our score-based surrogate for trajectory KL as a proxy for distributional drift using the Simple Animals prompts. **(a)** During optimization, BIDS's drift remains near-zero while DNO's grows steadily. **(b)** Sweeping the hyperparameter β for BIDS reveals a stable trade-off, where a moderate value improves reward with negligible drift.

5.2 BIDS Improves Semantic Control and Memory Efficiency

To demonstrate the efficiency of BIDS, we compare with batched Best-of-N and steering baselines (FK Steering (Singhal et al., 2025), CoDe (Singh et al., 2025)) by matching our optimization steps to their respective wall-clock budgets. The batched sampling baselines use over 20 GB of peak GPU memory (Table 1). In contrast, BIDS maintains a constant 8.97 GB footprint while achieving win rates of 75.56% and 73.33% against FK Steering and CoDe on the image darkness reward. BIDS therefore provides competitive semantic control with lower peak memory than these batched baselines.

Table 1: **Inference-time alignment under matched wall-clock budgets.** While sampling-based steering methods require high memory overhead to generate candidate trajectories, BIDS maintains a constant 8.97 GB footprint. The memory column reports each baseline's peak GPU usage. Despite this constraint, BIDS achieves higher win rates on structural rewards (Darkness/Brightness) and remains highly competitive on Aesthetic Score.

Comparison	Aesthetic (%) ↑	Darkness (%) ↑	Brightness (%) ↑	Peak GPU Mem (GB) ↓
BIDS vs. FK Steering (16 Particles)	57.78	75.56	73.33	20.55
BIDS vs. DAS (8 Particles)	37.78	77.78	80.00	23.90
BIDS vs. CoDe ($N=20$)	55.56	73.33	77.78	24.51
BIDS vs. Best-of-N ($N=50$)	53.33	80.00	82.22	28.14

BIDS resists structural collapse relative to noise-space baselines. Table 2 summarizes the head-to-head preference evaluations. While training and sampling-based methods are included

for completeness, our primary comparison is against noise-space baselines (DNO, InitNO). BIDS achieves mean win rates above 50% across the 54 settings comprising the primary evaluation suite. Notably, BIDS maintains decisive win margins on prompts involving multiple subjects (Animal-Animal and Animal-Object). For example, BIDS achieves a 77.27% win rate against DNO on Animal-Animal prompts for the Aesthetic reward. These results support our premise: constraining image-space drift preserves greater semantic fidelity than noise-space regularization. Even against DyMO†, which uses an auxiliary semantic graph to enforce structure, BIDS wins in 8 out of 9 settings. (Additional results on fine-grained attribute preservation are in Appendix A.6).

Table 2: **BIDS outperforms weight-tuned and noise-constrained baselines on SDv1.5 across diverse prompt and reward configurations.** By limiting image-space drift, BIDS achieves higher win rates on semantic rewards (Brightness/Darkness) and maintains strong Aesthetic alignment compared to baselines restricted by noise-space priors (e.g., DNO) or global weight updates. Values represent mean ± standard deviation over five seeds.

Comparison	Animal			Animal-Animal			Animal-Object		
	Aesthetic ↑	Bright. ↑	Dark. ↑	Aesthetic ↑	Bright. ↑	Dark. ↑	Aesthetic ↑	Bright. ↑	Dark. ↑
BIDS vs DDPO	$60.00_{\pm 5.12}$	$73.33_{\pm 5.21}$	$83.33_{\pm 2.90}$	$57.58_{\pm 6.00}$	$60.61_{\pm 5.96}$	$83.33_{\pm 4.27}$	$54.86_{\pm 6.00}$	$68.06_{\pm 7.10}$	$81.25_{\pm 5.53}$
BIDS vs Diffusion-DPO	$62.22_{\pm 4.97}$	$91.11_{\pm 1.99}$	$93.33_{\pm 1.22}$	$66.67_{\pm 6.59}$	$81.82_{\pm 5.12}$	$84.85_{\pm 5.31}$	$63.19_{\pm 6.54}$	$82.64_{\pm 3.65}$	$89.58_{\pm 1.86}$
BIDS vs D3PO	$73.33_{\pm 4.61}$	$87.64_{\pm 3.72}$	$84.44_{\pm 6.55}$	$63.64_{\pm 6.29}$	$68.18_{\pm 3.65}$	$74.24_{\pm 5.35}$	$68.06_{\pm 5.35}$	$77.78_{\pm 5.31}$	$83.33_{\pm 5.79}$
BIDS vs BoN ($N=50$)	$62.00_{\pm 3.65}$	$86.66_{\pm 3.30}$	$90.33_{\pm 3.85}$	$62.12_{\pm 2.43}$	$71.21_{\pm 5.75}$	$92.42_{\pm 2.22}$	$54.86_{\pm 9.69}$	$83.33_{\pm 3.14}$	$93.06_{\pm 1.22}$
BIDS vs InitNO	$61.11_{\pm 3.37}$	$72.22_{\pm 5.75}$	$82.22_{\pm 2.90}$	$65.15_{\pm 3.98}$	$78.79_{\pm 5.07}$	$83.33_{\pm 3.30}$	$57.64_{\pm 6.40}$	$84.03_{\pm 5.35}$	$80.56_{\pm 5.75}$
BIDS vs DNO	$80.00_{\pm 2.22}$	$77.78_{\pm 6.74}$	$64.44_{\pm 7.13}$	$77.27_{\pm 4.18}$	$57.58_{\pm 5.12}$	$66.67_{\pm 5.79}$	$68.06_{\pm 1.86}$	$77.78_{\pm 5.96}$	$71.53_{\pm 5.58}$
Reference Baseline (Augmented with LLM-derived semantic graph + cross-attention)									
BIDS vs DyMO†	$53.89_{\pm 8.52}$	$68.89_{\pm 6.78}$	$68.89_{\pm 5.12}$	$56.06_{\pm 5.79}$	$80.30_{\pm 4.71}$	$87.98_{\pm 4.44}$	$46.53_{\pm 7.47}$	$84.81_{\pm 5.12}$	$87.50_{\pm 3.98}$

BIDS aligns with human preferences and transfers to diverse architectures. To validate our automated metrics, we conducted user studies with 100 anonymous participants. Evaluating subjective generative tasks exhibits naturally high variance; established studies demonstrate that expert models typically achieve only $\approx 65\%$ instance-level agreement with humans due to aesthetic subjectivity (Xu et al., 2023). Therefore, rather than relying on instance-level inter-rater reliability, we validate our proxy by demonstrating it reproduces the overall preference ranking. On SDv1.5 (optimizing Aesthetic Score on the Simple Animals dataset), participants prefer BIDS over DNO in 80.30% of comparisons (Figure 5), reflecting the GPT-4o rankings. The margin narrows on the larger SDXL backbone (optimizing HPSv2 on a 50-prompt HPDv2 subset) to a 55.75% win rate. We hypothesize that the smaller margin on SDXL reflects the greater resistance of a larger backbone to structural collapse; nevertheless, participants prefer BIDS in both evaluations.

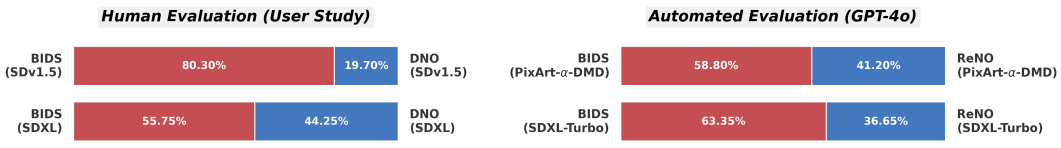

Figure 5: **BIDS achieves high win rates in human and automated preference trials.** The chart displays head-to-head win rates from focused evaluations. **Human Evaluation:** Participants prefer BIDS on both SDv1.5 (vs. DNO) and SDXL (vs. DNO). **Automated Evaluation (GPT-4o):** BIDS outperforms ReNO on efficient architectures, including one-step PixArt-α DMD and SDXL-Turbo.

Finally, we demonstrate that BIDS extends to efficient and transformer-based architectures using the HPDv2 benchmark (Wu et al., 2023a; optimizing HPSv2). On the latent-distilled SDXL-Turbo (Sauer et al., 2024), BIDS achieves a 63.35% GPT-4o win rate against ReNO (Figure 5). We also apply BIDS to Diffusion Transformer (DiT) architectures, specifically the one-step PixArt-α DMD (Chen et al., 2024a). In this setting, BIDS achieves a 58.80% win rate against ReNO, indicating that our score-based surrogate is applicable across both UNet and DiT backbones.

5.3 BIDS Aligns Black-Box Objectives

Objectives such as human preferences or perceptual metrics are non-differentiable. We demonstrate that BIDS extends to this setting, outperforming numerical gradient baselines in quantitative benchmarks and robustness tests.

We first treat the Aesthetic Score as a black-box reward to benchmark optimization quality. Optimizing for 50 steps on Simple Animals using SDv1.5 ($\beta = 5 \times 10^{-5}$, 5 seeds), BIDS achieves a 63.56% win rate over DNO (Hybrid-2), which relies on numerical gradient estimation. Unlike numerical gradients, which become unstable or computationally expensive when the objective is discontinuous, BIDS extracts the gradient proxy purely via the model's score function. By using the reward strictly as a discrete comparator, we optimize the noise without requiring a differentiable reward model.

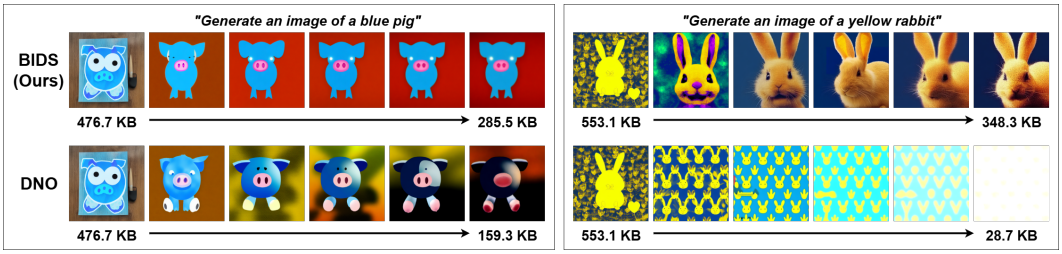

Figure 6: **BIDS preserves semantic content while optimizing a non-differentiable reward.** We evaluate optimization robustness using JPEG compressibility (file size reduction). DNO exploits this metric by severely degrading the underlying image structure to artificially minimize file size, either by collapsing the output into a featureless blob (e.g., the "blue pig") or a repeating patterned grid (e.g., the "yellow rabbit"). In contrast, BIDS generates highly compressible outputs while retaining the core semantic elements of both prompts.

We evaluate robustness on SDv1.5 using JPEG compressibility (Reich et al., 2024; Wallace, 1991), a non-differentiable objective that incentivizes reward hacking. Because minimizing file size encourages texture removal, optimizing this objective inherently degrades structural detail. As shown in Figure 6, DNO exploits this metric by degrading the image into repeating patterned grids to increase the target reward. In contrast, BIDS improves compressibility while preserving the semantic elements dictated by the prompt (e.g., the "yellow rabbit" remains visible). This demonstrates that BIDS's score-based trajectory constraint resists structural collapse, even under objectives that penalize image details.

6 Conclusion

We introduced BIDS, an inference-time objective designed to mitigate structural collapse in noise-space optimization. Building on the observation that small changes in noise can cause large changes in generated images, we shift the regularization target to the image space. The Data Processing Inequality bounds image-space divergence by trajectory divergence, motivating our score-based surrogate for preserving image structure during reward optimization. We also demonstrate that BIDS extends to discrete preference optimization, enabling the alignment of non-differentiable rewards without offline fine-tuning or numerical gradient estimation.

As an iterative method, BIDS increases generation time to improve reward while preserving image structure. It uses a single trajectory during optimization, reducing peak memory use compared with the evaluated multi-trajectory steering baselines. Structural collapse persists even in models as large as 12B parameters, showing that scaling alone does not prevent reward hacking (Appendix A.11). Appendix A.12 formulates the corresponding score-based objective for rectified-flow models using their predicted velocities.

AI Use Statement

We used generative AI tools to assist with manuscript writing, code development, mathematical derivations and proof checking, and experimental planning and analysis. GPT-4o was used for

automated image-preference evaluation. The authors take responsibility for the final manuscript and reported results.

Ethics Statement

Our human evaluation involved 100 anonymous volunteers recruited from our laboratory and university community. The task involved comparing non-sensitive generated images for prompt faithfulness and aesthetic preference. No personally identifiable information was collected, and the study was designed to be brief to respect participants' time. BIDS aims to improve the reliability of inference-time image generation by reducing reward hacking. The pretrained models can retain social and demographic biases, and generated images can be misused to create harmful or misleading content. We advocate responsible use of generative models and appropriate safeguards for their deployment.

Reproducibility Statement

Section 5 and Appendix A document the experimental settings, algorithms, evaluation procedures, and prompts, together with the assumptions and proofs supporting the theoretical results.

References

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*, pp. 1–17. Springer, 2024.

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. In *ICML 2023 Workshop on Structured Probabilistic Inference & Generative Modeling*, 2023.

Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-σ: Weak-to-strong training of diffusion transformer for 4k text-to-image generation. In *European Conference on Computer Vision*, pp. 74–91. Springer, 2024a.

Junsong Chen, Jincheng YU, Chongjian GE, Lewei Yao, Enze Xie, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. In *The Twelfth International Conference on Learning Representations*, 2024b. URL https://openreview.net/forum?id=eAKmQPe3m1.

Kevin Clark, Paul Vicol, Kevin Swersky, and David J Fleet. Directly fine-tuning diffusion models on differentiable rewards. In *The Twelfth International Conference on Learning Representations*, 2024.

Fei Deng, Qifei Wang, Wei Wei, Tingbo Hou, and Matthias Grundmann. Prdp: Proximal reward difference prediction for large-scale reward finetuning of diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7423–7433, 2024.

Carles Domingo-Enrich, Michal Drozdzal, Brian Karrer, and Ricky T. Q. Chen. Adjoint matching: Fine-tuning flow and diffusion generative models with memoryless stochastic optimal control. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=xQBRrtQM8u.

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 A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 125487–125519. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/e31bdea0a93741c2157eea705dd219eb-Paper-Conference.pdf.

Ying Fan, Olivia Watkins, Yuqing Du, Hao Liu, Moonkyung Ryu, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, Kangwook Lee, and Kimin Lee. Reinforcement learning for fine-tuning text-to-image diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024.

Yi Gu, Zhendong Wang, Yueqin Yin, Yujia Xie, and Mingyuan Zhou. Diffusion-rpo: Aligning diffusion models through relative preference optimization. *arXiv preprint arXiv:2406.06382*, 2024.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9380–9389, 2024.

Anne Harrington, A Koepke, Shyamgopal Karthik, Trevor Darrell, and Alexei A Efros. It's never too late: Noise optimization for collapse recovery in trained diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 43124–43134, 2026.

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

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf.

Peter Holderrieth, Uriel Singer, Tommi Jaakkola, Ricky T. Q. Chen, Yaron Lipman, and Brian Karrer. GLASS flows: Efficient inference for reward alignment of flow and diffusion models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=vH7OAPZ2dR.

Susung Hong. Smoothed energy guidance: Guiding diffusion models with reduced energy curvature of attention. *Advances in Neural Information Processing Systems*, 37:66743–66772, 2024.

Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *Advances in Neural Information Processing Systems*, 36:78723–78747, 2023.

Byungjun Kim, Soobin Um, and Jong Chul Ye. Diverse text-to-image generation via contrastive noise optimization. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=EVRMnAREc3.

Sunwoo Kim, Minkyu Kim, and Dongmin Park. Test-time alignment of diffusion models without reward over-optimization. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=vi3DjUhFVm.

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 *Advances in Neural Information Processing Systems*, 2023.

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

Tianhong Li and Kaiming He. Back to basics: Let denoising generative models denoise, 2025. URL https://arxiv.org/abs/2511.13720.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*, 2021.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

Mihir Prabhudesai, Anirudh Goyal, Deepak Pathak, and Katerina Fragkiadaki. Aligning text-to-image diffusion models with reward backpropagation. *arXiv preprint arXiv:2310.03739*, 2023.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=HPuSIXJaa9.

Christoph Reich, Biplob Debnath, Deep Patel, and Srimat Chakradhar. Differentiable jpeg: The devil is in the details. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 4114–4123. IEEE, January 2024. doi: 10.1109/wacv57701.2024.00408. URL http://dx.doi.org/10.1109/WACV57701.2024.00408.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, June 2022.

Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. In *European Conference on Computer Vision*, pp. 87–103. Springer, 2024.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022.

Anuj Singh, Sayak Mukherjee, Ahmad Beirami, and Hadi Jamali-Rad. Code: Blockwise control for denoising diffusion models. *arXiv preprint arXiv:2502.00968*, 2025.

Raghav Singhal, Zachary Horvitz, Ryan Teehan, Mengye Ren, Zhou Yu, Kathleen McKeown, and Rajesh Ranganath. A general framework for inference-time scaling and steering of diffusion models. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=Jp988ELppQ.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021a.

Jiaming Song, Qinsheng Zhang, Hongxu Yin, Morteza Mardani, Ming-Yu Liu, Jan Kautz, Yongxin Chen, and Arash Vahdat. Loss-guided diffusion models for plug-and-play controllable generation. In *International Conference on Machine Learning*, pp. 32483–32498. PMLR, 2023.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021b. URL https://openreview.net/forum?id=PxTIG12RRHS.

Zhiwei Tang, Jiangweizhi Peng, Jiasheng Tang, Mingyi Hong, Fan Wang, and Tsung-Hui Chang. Inference-time alignment of diffusion models with direct noise optimization. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=JpbqiD7n9r.

Masatoshi Uehara, Yulai Zhao, Tommaso Biancalani, and Sergey Levine. Understanding reinforcement learning-based fine-tuning of diffusion models: A tutorial and review. *CoRR*, abs/2407.13734, 2024. URL https://doi.org/10.48550/arXiv.2407.13734.

Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8228–8238, 2024.

Gregory K Wallace. The jpeg still picture compression standard. *Communications of the ACM*, 34(4): 30–44, 1991.

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. *CoRR*, abs/2306.09341, 2023a. URL https://doi.org/10.48550/arXiv.2306.09341.

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*, 2023b.

Xin Xie and Dong Gong. Dymo: Training-free diffusion model alignment with dynamic multi-objective scheduling. *arXiv preprint arXiv:2412.00759*, 2024.

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

Kai Yang, Jian Tao, Jiafei Lyu, Chunjiang Ge, Jiaxin Chen, Weihan Shen, Xiaolong Zhu, and Xiu Li. Using human feedback to fine-tune diffusion models without any reward model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8941–8951, 2024.

Jiwen Yu, Yinhuai Wang, Chen Zhao, Bernard Ghanem, and Jian Zhang. Freedom: Training-free energy-guided conditional diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 23174–23184, 2023.

Xinlu Zhang, Yujie Lu, Weizhi Wang, An Yan, Jun Yan, Lianke Qin, Heng Wang, Xifeng Yan, William Yang Wang, and Linda Ruth Petzold. Gpt-4v(ision) as a generalist evaluator for vision-language tasks. *CoRR*, abs/2311.01361, 2023. URL https://doi.org/10.48550/arXiv.2311.01361.

Zikai Zhou, Shitong Shao, Lichen Bai, Shufei Zhang, Zhiqiang Xu, Bo Han, and Zeke Xie. Golden noise for diffusion models: A learning framework. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17688–17697, 2025.

A Appendix

A.1 Implementation Details

We demonstrate results of BIDS on multiple reward functions, including Aesthetic Score, HPSv2, PickScore, and image brightness and darkness. The first three are human-aligned, whereas the last two are useful for easily visualizing reward hacking. We set $\beta = 0.2$ for Aesthetic Score, $\beta = 0.5$ for HPSv2 and PickScore, $\beta = 1$ for brightness, and $\beta = 0.8$ for darkness. For the single-step SDXL-Turbo and PixArt-α experiments (HPSv2), we set $\beta = 5 \times 10^{-7}$. For JPEG compressibility, a non-differentiable reward, we use $\beta = 5 \times 10^{-5}$. We fix β once per reward–backbone setting, not per prompt or evaluation metric. HPSv2 and PickScore use ViT-H/14 as the backbone, and Aesthetic Score uses ViT-L/14.

For differentiable rewards, Algorithm 1 uses one paired base and optimized trajectory and penalizes the non-negative deviation $|S_0 - S|$, which measures departures in either direction from the reference score-energy. The score-energy term is normalized by 10^4; the reported β values use this scaling.

For single-step backbones, we compare the squared norms of the model predictions before and after noise optimization, using unit weight at the evaluated timestep.

We implement BIDS with AdamW and run all experiments on a single NVIDIA A100 GPU. To ensure a controlled comparison against inference-time noise-optimization baselines, we match optimization hyperparameters (optimizer, learning rate, number of optimization steps, and standard stabilizers such as gradient clipping) to the default settings recommended by each baseline for the corresponding backbone: DNO/PRNO for SDv1.5 and SDXL, and ReNO for SDXL-Turbo and PixArt. We do not reuse baseline-specific objective terms (e.g., noise-regularization components). This protocol isolates gains from the objective rather than from hyperparameter tuning. For baselines, we report results for regularized DNO (PRNO) throughout using the recommended noise regularization settings.

CMMD Evaluation. We measure distributional drift using maximum mean discrepancy between CLIP image embeddings of optimized samples and unoptimized base-model samples. Images are pooled across the same 45 Simple Animals prompts, with 20 samples per prompt in each set. Lower CMMD indicates greater similarity to the base-model reference.

User Study Details. We obtain these results from anonymous volunteers who respond using an online form. We ask these participants: *"Which of the below images would you personally prefer getting given the above prompt (based on your personal trade-off between prompt faithfulness and aesthetics)?"* We adopt this from ReNO. Given side-by-side images (prompt-matched, order-randomized), participants have the option to choose the left image, the right image, or neither/both. In the case of the last option, we assign a score of 0.5 to that image pair. Due to recruitment and budget constraints, we limit the number of participants to 100. The SDv1.5 evaluation uses the Simple Animals prompts; the SDXL evaluation uses 50 prompts from HPDv2.

A.2 ALGORITHMS

For the reader's reference, the objective for differentiable rewards is:

$$\mathcal{L}_{\text{BIDS}} = -\mathbb{E}_{z_{\text{init}}\sim\mathcal{N}(0,\mathbf{I})}\left[r(x_0^{\text{opt}}, c)\right] + \beta\left|\mathbb{E}_{z_{\text{init}}\sim\mathcal{N}(0,\mathbf{I})}[S_0 - S(z, c)]\right|. \tag{9}$$

For non-differentiable rewards, we compute reference score sums $S_0^{(1)}$ and $S_0^{(2)}$ from the unoptimized noise sequences $z_{\text{init}}^{(1)}$ and $z_{\text{init}}^{(2)}$. Both sums remain fixed during optimization. At each iteration, S_0^w and S_0^l denote the references associated with the currently preferred and rejected images. With z^w and z^l generating those images, our objective is:

$$\mathcal{L}_{\text{BIDS}}(z^w, z^l, c) = -\log\sigma\big(\beta\left[(S_0^w - S(z^w, c)) - (S_0^l - S(z^l, c))\right]\big). \tag{10}$$

Algorithm 1 BIDS (for Differentiable Rewards)

Require: Initial noise sequence $z_{\text{init}} \sim \mathcal{N}(0, \mathbf{I})$, prompt c, reward function r, sampling process (e.g., DDIM) G_θ, optimization steps K_{opt}, learning rate α, regularization strength $\beta > 0$

Initialize $z \leftarrow z_{\text{init}}$
$_, S_0 \leftarrow G_\theta(z_{\text{init}}, c)$ *// Obtain fixed $S_0 = \sum_t \sigma_t^2 \|s_{\theta,t}(x_t^{\text{base}}, c)\|^2$ (Eq. (5))*
for $k = 1, \ldots, K_{\text{opt}}$ **do**
  $x_0, S \leftarrow G_\theta(z, c)$ *// Generate x_0; obtain $S = \sum_t \sigma_t^2 \|s_{\theta,t}(x_t^{\text{opt}}, c)\|^2$ (Eq. (5))*
  $\mathcal{L}_{\text{BIDS}} \leftarrow -r(x_0, c) + \beta|S_0 - S|$ *// Compute one-trajectory objective*
  $z \leftarrow z - \alpha\nabla_z\mathcal{L}_{\text{BIDS}}$ *// Update noise*
end for
Return z

Algorithm 2 BIDS (for Non-Differentiable Rewards)

Require: Initial noise sequences $z_{\text{init}}^{(1)}, z_{\text{init}}^{(2)} \sim \mathcal{N}(0, \mathbf{I})$, prompt c, reward function r, sampling process G_θ, optimization steps K_{opt}, learning rate α, regularization hyperparameter $\beta > 0$

Initialize $z^{(1)} \leftarrow z_{\text{init}}^{(1)}$, $z^{(2)} \leftarrow z_{\text{init}}^{(2)}$
$x_0^{(i)}, S_0^{(i)} \leftarrow G_\theta(z_{\text{init}}^{(i)}, c)$ for $i \in \{1, 2\}$
Keep $S_0^{(1)}$ and $S_0^{(2)}$ fixed.
Let i_w and i_l index the preferred and rejected candidates.
for $k = 1, \ldots, K_{\text{opt}}$ **do**
  $x_0^{(1)}, S^{(1)} \leftarrow G_\theta(z^{(1)}, c)$ *// Generate images and score sums*
  $x_0^{(2)}, S^{(2)} \leftarrow G_\theta(z^{(2)}, c)$
  if $r(x_0^{(1)}, c) > r(x_0^{(2)}, c)$ **then** *// Choose preferred image; assign winner and loser*
    $x_0^w \leftarrow x_0^{(1)}, x_0^l \leftarrow x_0^{(2)}$
    $i_w \leftarrow 1, i_l \leftarrow 2$
    $S^w \leftarrow S^{(1)}, S^l \leftarrow S^{(2)}$
    $S_0^w \leftarrow S_0^{(1)}, S_0^l \leftarrow S_0^{(2)}$
  else
    $x_0^w \leftarrow x_0^{(2)}, x_0^l \leftarrow x_0^{(1)}$
    $i_w \leftarrow 2, i_l \leftarrow 1$
    $S^w \leftarrow S^{(2)}, S^l \leftarrow S^{(1)}$
    $S_0^w \leftarrow S_0^{(2)}, S_0^l \leftarrow S_0^{(1)}$
  $\mathcal{L}_{\text{BIDS}} \leftarrow -\log\sigma\big(\beta[(S_0^w - S^w) - (S_0^l - S^l)]\big)$ *// Eq. (10)*
  $z^{(i_w)} \leftarrow z^{(i_w)} - \alpha\nabla_{z^{(i_w)}}\mathcal{L}_{\text{BIDS}}$ *// Both gradients use the current loss*
  $z^{(i_l)} \leftarrow z^{(i_l)} - \alpha\nabla_{z^{(i_l)}}\mathcal{L}_{\text{BIDS}}$
end for
Return $z^{(i_w)}$

A.3 Proof of Proposition 1 and the Linear Noise-Space Reduction

Proof. To see how each noise input affects the final image, we expand the linear sampler in Proposition 1. The sampler starts from the initial noise x_T and produces the final image x_0 in T denoising steps. At each step, it multiplies the current latent $x_t \in \mathbb{R}^d$ by the matrix $A \in \mathbb{R}^{d\times d}$ and adds the injected noise $\sigma\zeta_t$, where $\zeta_t \in \mathbb{R}^d$ and $\sigma > 0$. The first two steps are

$$x_{T-1} = Ax_T + \sigma\zeta_T,$$
$$x_{T-2} = A(Ax_T + \sigma\zeta_T) + \sigma\zeta_{T-1} = A^2x_T + \sigma A\zeta_T + \sigma\zeta_{T-1}.$$

Each further step applies A to the existing terms and adds the next noise. Writing A^k for the kth matrix power, with $A^0 = \mathbf{I}$ the $d \times d$ identity, the final output is

$$x_0 = A^T x_T + \sigma \sum_{k=1}^{T} A^{k-1}\zeta_k. \tag{11}$$

Add $\delta x_T \in \mathbb{R}^d$ to the initial noise x_T and $\delta\zeta_k \in \mathbb{R}^d$ to the injected noise ζ_k at step k. Together, these changes form $\delta z = (\delta x_T, \delta\zeta_T, \ldots, \delta\zeta_1)$. Let δx_0 denote the resulting image change.

Substitute the updated noises $x_T + \delta x_T$ and $\zeta_k + \delta\zeta_k$ into Eq. 11, then subtract the original output. The terms containing the original noises cancel, leaving

$$\delta x_0 = A^T \delta x_T + \sigma \sum_{k=1}^{T} A^{k-1}\delta\zeta_k.$$

Let $\epsilon > 0$ be the maximum permitted change in the noise sequence in Eq. 2. Change only the initial noise, with $\|\delta x_T\|_2 = \epsilon$, and leave all injected noises unchanged. Then $\|\delta z\|_2 = \epsilon$, but the image changes by $A^T\delta x_T$. For $A = 2\mathbf{I}$, this image change has norm $2^T\epsilon$. Choosing larger positive multiples of $\mathbf{I}$ makes the image change arbitrarily large while the noise change remains ϵ. Thus, bounding the noise change alone gives no bound on the image change that holds across these samplers. □

From image divergence to a noise-space penalty. The same linear sampler also lets us calculate the change in the output distribution. As in the main method, draw the initial and injected noises independently from $\mathcal{N}(0, \mathbf{I})$. Let p^{base} denote the distribution of images generated from these noises. To obtain p^{opt}, add δx_T to the initial noise and $\delta\zeta_k$ to the noise injected at step k. Use the same additions for every sampled noise sequence. The output then changes by the same vector δx_0 for every sample. Eq. 11 is a linear combination of Gaussian noises, so both output distributions are Gaussian. The base mean is zero, and the optimized mean is δx_0. Adding the same vector to every output leaves the covariance unchanged.

Let $\Sigma_0 \in \mathbb{R}^{d\times d}$ denote their shared output covariance, and let $\top$ denote matrix transpose, distinct from the power T. Independence makes the covariance contributions of the noise inputs add, giving

$$\Sigma_0 = A^T(A^T)^\top + \sigma^2 \sum_{k=1}^{T} A^{k-1}(A^{k-1})^\top. \tag{12}$$

The $k = 1$ term contributes $\sigma^2\mathbf{I}$, so Σ_0 is positive definite and invertible. The noise updates shift the output mean by δx_0 without changing this covariance. Applying the KL formula for two Gaussians with the same covariance gives

$$d_{\text{KL}}[p^{\text{opt}} \| p^{\text{base}}] = \tfrac{1}{2}\delta x_0^\top \Sigma_0^{-1}\delta x_0. \tag{13}$$

Substituting the expression for δx_0 into Eq. 13 expresses image-space KL entirely in terms of changes to the initial and injected noises. Thus, in this linear setting, the image-space constraint reduces to a noise-space constraint.

A.4 Derivation of DPO Objective

Let $R(x_{0:T}, c)$ be a trajectory reward and let $p^{\text{base}}(x_{0:T} \mid c)$ be the base trajectory distribution defined in Section 4. Define the normalizing constant $Z(c) = \mathbb{E}_{x_{0:T}\sim p^{\text{base}}(\cdot|c)}[\exp(R(x_{0:T}, c)/\beta)]$

and assume $0 < Z(c) < \infty$. The optimal trajectory distribution p^* for the KL-regularized objective has the following form:

$$p^*(x_{0:T} \mid c) = \frac{1}{Z(c)} p^{\text{base}}(x_{0:T} \mid c) \exp\left(\frac{R(x_{0:T}, c)}{\beta}\right). \tag{14}$$

Rearranging gives

$$R(x_{0:T}, c) = \beta \log Z(c) + \beta \log \frac{p^*(x_{0:T} \mid c)}{p^{\text{base}}(x_{0:T} \mid c)}. \tag{15}$$

The additive constant cancels in a pairwise comparison. Assume that the optimized distribution p^{opt} is absolutely continuous with respect to p^{base}, with finite log-density ratios on the compared trajectories. Define

$$r^*(x_{0:T}, c) := \beta \log \frac{p^{\text{opt}}(x_{0:T} \mid c)}{p^{\text{base}}(x_{0:T} \mid c)}. \tag{16}$$

For a preferred trajectory $x^w_{0:T}$ and rejected trajectory $x^l_{0:T}$, the Bradley–Terry loss, averaged over preference pairs, is

$$\mathcal{L}_{\text{BT}} = -\mathbb{E}\left[\log \sigma\left(r^*(x^w_{0:T}, c) - r^*(x^l_{0:T}, c)\right)\right]. \tag{17}$$

The log-density ratio is intractable, so we use $\widehat{r}(z, c) = \beta(S_0 - S(z, c))$ as the practical implicit-reward surrogate, where S is defined in Eq. 5. For each noise sequence, we compute the reference sum S_0 from the unoptimized noise z_{init} and keep it fixed. S^w_0 and S^l_0 denote the references for the currently preferred and rejected sequences (Appendix A.2). For the noise sequences z^w and z^l generating the preferred and rejected images, the resulting pairwise objective is

$$\mathcal{L}_{\text{BIDS}}(z^w, z^l, c) = -\log \sigma\left(\beta\left[(S^w_0 - S(z^w, c)) - (S^l_0 - S(z^l, c))\right]\right). \tag{18}$$

A.5 Additional Qualitative Results

Brightness Reward Hacking in BIDS vs. DNO

Here, we show reward hacking on image brightness, prompting the diffusion model with *"generate an image of a black* [ANIMAL]*"*. The contrast between the prompt and the reward provides a clear visualization of reward hacking.

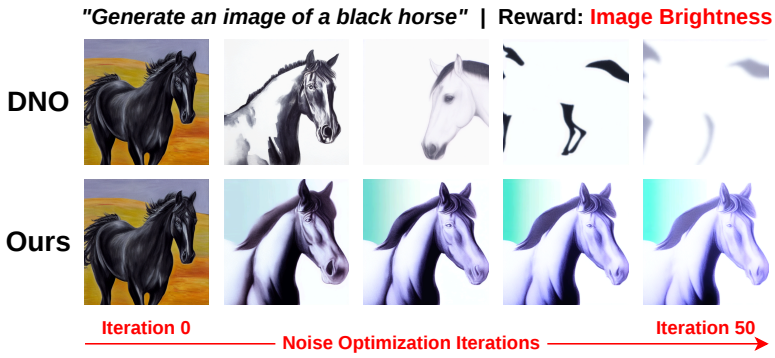

Figure 7: **BIDS vs DNO in reward hacking**. On the Image Brightness reward, we demonstrate that BIDS is able to effectively mitigate reward hacking and generate better, more realistic images while maintaining prompt fidelity, when compared to the state-of-the-art baseline on the same seed. In the top row, after 50 optimization steps, DNO completely hacks the brightness reward and generates an image that is overly white and unrealistic. In contrast, in the bottom row, BIDS is able to mitigate reward hacking and produce much better images, while aligning with the target reward.

Qualitative Results: BIDS vs. ReNO on SDXL-Turbo

In Figure 8, we compare our method (middle) with Base SDXL-Turbo (left) and ReNO (right). We use the HPSv2 objective. Under equivalent sampling settings, BIDS tends to produce cleaner, more literal interpretations with stable composition and style consistency (e.g., large-scale scene layout, totem-like geometry, and concept blending), whereas ReNO often introduces extra flourishes or alternative readings. These examples match our quantitative trends that BIDS delivers stronger alignment than noise-only regularization.

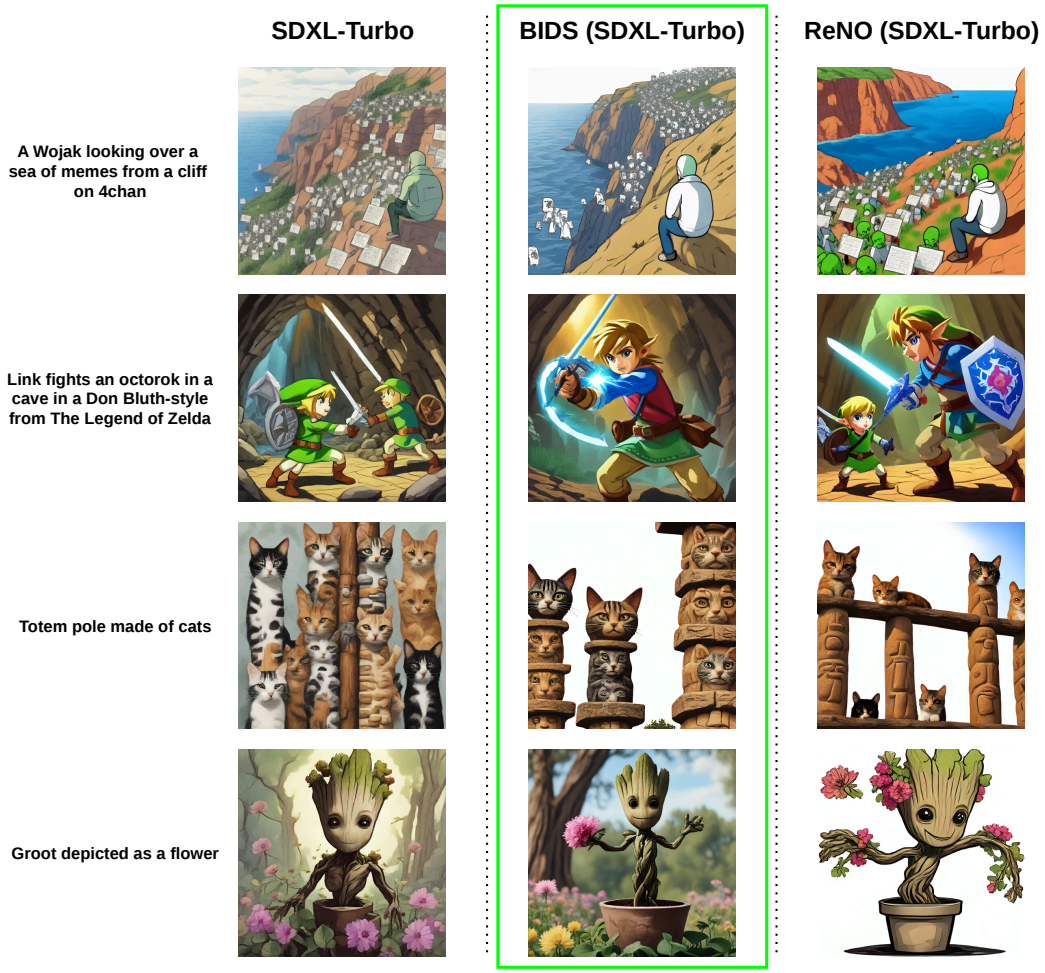

Figure 8: **Qualitative comparison on SDXL-Turbo (left: Base, middle: BIDS, right: ReNO), optimized for HPSv2.** Prompts (top to bottom): "Wojak looking over a sea of memes from a cliff on 4chan," "Link fights an octorok in a cave in a Don Bluth style," "Totem pole made of cats," "Groot depicted as a flower." Across these cases, BIDS offers crisper compositions and more literal, prompt-faithful structure (e.g., coherent large-scene layout, pole-like stacking, integrated concept depiction), while ReNO reflects plausible but looser interpretations. Images are shown at equal inference settings for both methods.

A.6 Additional Quantitative Results

In this section, we present additional results comparing BIDS to baselines.

Win Rates of All Methods Against SDv1.5

In Table 3 below, we present the win rates of several methods including BIDS against SDv1.5 using Animal prompts. The results demonstrate that our method achieves the highest win rate in three rewards: Aesthetic Score, brightness, and darkness. We also represent this in the bar plot diagram in Figure 9. BIDS consistently outperforms baselines in win rates.

Table 3: **Win rates of all methods against SDv1.5.** The table reports the percentage of pairwise comparisons between all methods (including BIDS) vs SDv1.5 on three objectives: *Aesthetic Score, Brightness,* and *Darkness* with GPT-4o as the judge. We sample Best-of-N (BoN) with base SDv1.5. Across all objectives, our method consistently achieves higher win rates than SDv1.5. These results demonstrate that our method not only outperforms previous baselines in terms of win rates over various target objectives, but is also able to generate better images while more effectively optimizing the target reward functions.

Method (vs SDv1.5)	Aesthetic ↑	Brightness ↑	Darkness ↑
DDPO	51.11	80.00	71.11
Diffusion-DPO	48.89	77.77	66.67
D3PO	53.33	71.22	71.11
BoN ($N = 30$)	51.11	80.11	62.22
BoN ($N = 40$)	55.55	77.77	63.33
BoN ($N = 50$)	55.55	75.55	61.11
DNO	42.22	75.56	66.67
BIDS (Ours)	**60.00**	**91.11**	**88.89**

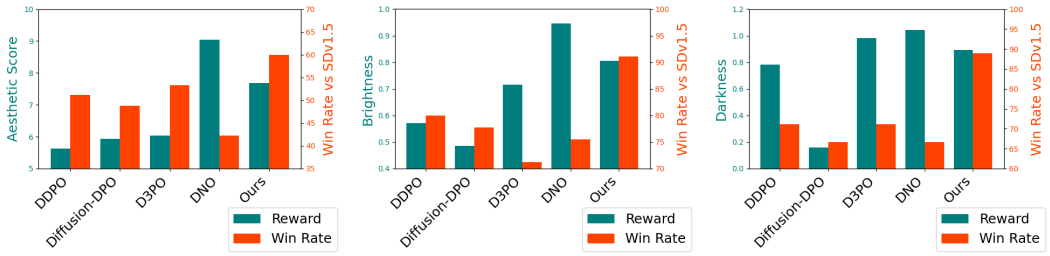

Figure 9: **Win rates (vs. SDv1.5) and average rewards across methods.** We evaluate our method, BIDS, on Aesthetic Score (left), brightness (middle), and darkness (right), comparing against DDPO, Diffusion-DPO, D3PO, and DNO. BIDS consistently outperforms baselines in win rates despite lower average rewards. Notably, higher rewards can indicate overoptimization, resulting in lower win rates.

Overall Rewards Across Methods and Reward Functions. Table 4 presents the mean rewards of images generated by different methods (including BIDS). Rewards are averaged across all images generated from the Animal prompts.

Table 4: **Overall rewards across methods and reward functions.** The above table compares our method with other baselines across various reward metrics and runtime. These rewards include Aesthetic Score, HPSv2, PickScore, brightness, and darkness. The leftmost column contains all methods. The second column (runtime) lists the approximate time needed to fine-tune or optimize each method (for a single image). We note DDPO, Diffusion-DPO, and D3PO take several hours due to fine-tuning. Though our method does not generate images with the highest rewards, we observe higher rewards compared to SDv1.5 after just five minutes of optimization with less reward hacking as discussed in the main paper. Inference times are measured using a single A100 GPU.

Method	Training Time	Inference Time	Aesthetic	HPSv2	PickScore	Brightness	Darkness
SDv1.5	-	0.04 min	5.367	0.278	20.390	-0.123	0.087
DDPO	12 h	-	5.623	0.307	20.120	0.570	0.782
Diffusion-DPO	2 h	-	5.930	0.302	21.086	0.485	0.159
D3PO	24 h	-	6.027	0.297	20.030	0.715	0.982
BoN ($N=30$)	-	0.9 min	6.044	0.318	21.404	0.006	0.329
BoN ($N=40$)	-	1.31 min	6.080	0.320	21.427	0.094	0.339
BoN ($N=50$)	-	1.64 min	6.111	0.321	21.365	0.105	0.353
InitNO	-	0.5 min	5.259	0.286	20.120	-0.041	0.113
DyMO	-	0.85 min	5.720	0.322	21.220	0.320	0.008
DNO	-	5 min	9.044	0.287	25.568	0.945	1.044
BIDS (Ours)	-	5 min	7.684	0.339	21.161	0.805	0.894

CLIPScores on the Simple Animals dataset. We optimize several key methods (including fine-tuning methods) for Aesthetic Score on the Simple Animals dataset, evaluating CLIPScore as a prompt fidelity check. We observe that strong reward gains may come with weaker prompt adherence. Our method, BIDS, achieves the highest CLIPScore while remaining competitive on the actual target reward. We emphasize that *these results should not be considered in a vacuum, and win rate is a more holistic metric.*

Table 5: **Comparison of CLIPScores according to method used, optimizing for Aesthetic Score.** We report CLIPScores, measuring prompt fidelity, and Aesthetic Score for various methods. BIDS achieves the best CLIPScore with comparable Aesthetic Score, whereas DNO achieves substantially higher Aesthetic Scores at the cost of prompt adherence.

Method	CLIPScore	Aesthetic Score
DDPO	24.898	5.623
Diffusion-DPO	25.340	5.930
D3PO	23.298	6.027
BoN ($N = 50$)	24.070	6.111
InitNO	24.555	5.259
DyMO	25.397	5.720
DNO	22.959	9.044
BIDS (Ours)	25.964	7.684

Hybrid Feasibility (BIDS + FK Steering). We implemented a simple hybrid combining BIDS with FK Steering to test complementarity. Concretely, we run multiple denoising trajectories in parallel and, at periodic intervals, select the candidate that maximizes the BIDS objective. The backbone remains frozen and configurations are fixed. This check is not wall-clock matched and serves only as a feasibility probe. On Simple Animals (45 prompts), optimizing for Aesthetic Score with $\beta = 0.1$, **BIDS with FK Steering achieves a 55.56% win rate over FK Steering alone (same backbone/GPU).**

T2I-CompBench (Compositional Fidelity). As an additional compositionality check (separate from reward-hacking mitigation), we report T2I-CompBench (Huang et al., 2023) scores for PixArt-α (DMD) (Chen et al., 2024b;a) in Table 6. Both ReNO and BIDS improve over the unaligned backbone across all subsets. ReNO achieves the strongest scores on Color/Shape/Texture/Spatial, while BIDS is close on these axes and attains the best performance on the more challenging Non-Spatial and Complex subsets (0.3211 and 0.3466). We view these results as a complementary diagnostic of compositional fidelity; our main conclusions rely on preference-based win rates and reward-hacking stress tests that directly match the paper's objective.

Table 6: **T2I-CompBench compositional fidelity on PixArt-α (DMD).** Higher is better. We report subset scores for Color, Shape, Texture, Spatial, Non-Spatial, and Complex prompts. Both ReNO and BIDS improve over the unaligned backbone across all subsets; BIDS attains the best scores on the challenging Non-Spatial and Complex splits.

Method	**Attribute Binding**			**Object Relationship**		**Complex ↑**
	Color ↑	Shape ↑	Texture ↑	Spatial ↑	Non-Spatial ↑	
PixArt-α (DMD)	0.3782	0.3419	0.4691	0.1787	0.3060	0.2913
ReNO	0.4694	0.4601	0.5917	0.2175	0.3114	0.3078
BIDS (Ours)	0.4550	0.4415	0.5763	0.2123	**0.3211**	**0.3466**

A.7 Empirical Validation of the Score-Energy Surrogate

To validate how our score-energy regularizer (Eq. 5) relates empirically to KL divergence, we conduct a controlled low-dimensional experiment, since estimating KL in high-dimensional image spaces is intractable. We train a 1D UNet diffusion model (dimension 32, no downsampling) on a standard Gaussian $\mathcal{N}(0, \mathbf{I})$ for 3000 steps using Adam with a learning rate of 10^{-3}. We then optimize the noise over 50 steps with a learning rate of 0.1 to shift the generated distribution toward a target mean of $\mu = 2.5$. At each step, we compute the "True KL" using Kernel Density Estimation on 5000 samples and compare it with our score-energy surrogate. While KDE is usable in this 1D setting, it is infeasible in high-dimensional image spaces due to the curse of dimensionality.

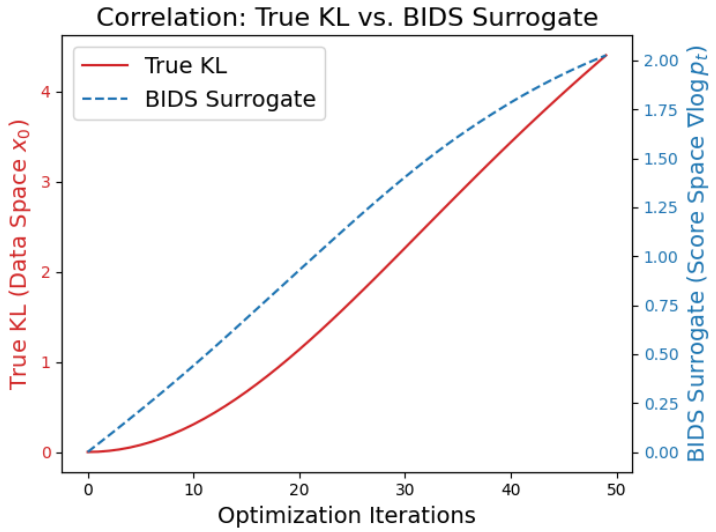

Figure 10: **Empirical validation of the BIDS surrogate on a 1D diffusion toy problem.** We compare the data-space "True" KL divergence (red, estimated via KDE) with our score-energy surrogate (blue) over 50 noise-optimization steps. Their strong monotonic agreement supports the surrogate as a practical proxy for distributional drift in this controlled setting.

As shown in Figure 10, the score-energy surrogate (dashed blue) and the "True KL" (solid red) are strongly monotonic over the optimization run. Their absolute scales differ because KL integrates log-density ratios ($\log \frac{p}{q}$), whereas our surrogate sums squared score norms ($\|\nabla \log p\|^2$). This controlled experiment validates the intended use of the surrogate as an empirical proxy for distributional drift.

A.8 Ablation Analysis

Impact of β on CLIPScores and Rewards

We justify our choice of hyperparameter β by evaluating results across five different seeds, plotting the mean curves for CLIPScore and reward against optimization steps. CLIPScore is a metric which scores how well an image aligns with a given prompt. For our results, we use CLIPScore with a ViT-L/14 vision backbone and optimize for image darkness. As shown in Figure 11 increasing β in our method leads to lower rewards but higher CLIPScores.

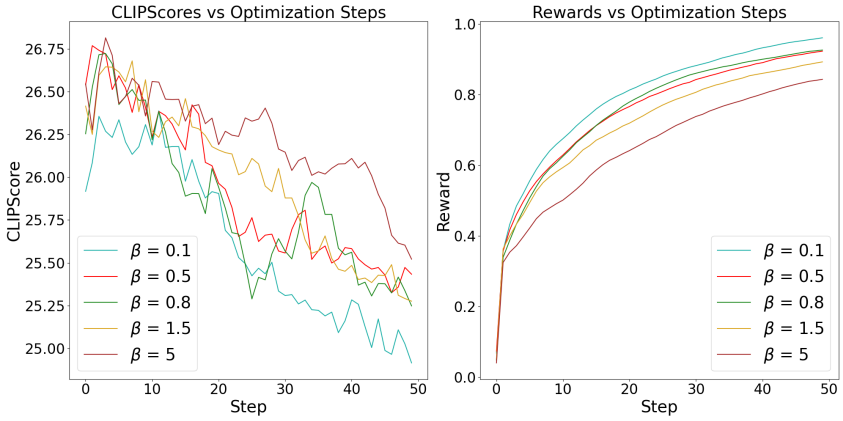

Figure 11: **Effect of different β on CLIPScores (left) and rewards (right).** We obtain CLIPScores and rewards for all entries in the animals list in Appendix A.9. Curves depict the average result of five different seeds. The curves depict that lower β exhibits higher rewards but lower CLIPScore. Higher β tends to inadequately optimize for the reward.

Optimization Budget. We additionally compare BIDS and DNO across optimization budgets on the 45 Simple Animals prompts using SDv1.5, seed 0, and 50 DDIM denoising steps per image. Figure 12 reports mean CLIPScore and Darkness reward throughout optimization, with markers at 5, 10, 20, 30, and 50 steps. BIDS uses $\beta = 0.8$ and DNO uses $\gamma = 1$. BIDS continues increasing reward with a smaller decline in CLIPScore than DNO.

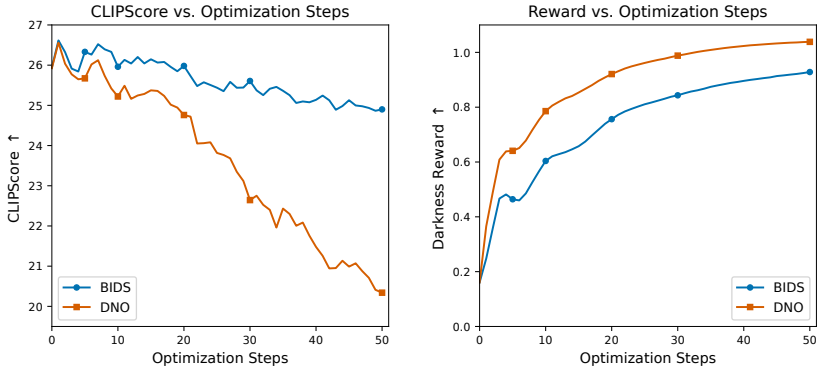

Figure 12: **Prompt alignment and reward across optimization budgets.** Mean CLIPScore (left) and Darkness reward (right) over 45 Simple Animals prompts on SDv1.5, using one seed and 50 DDIM denoising steps per image. Markers identify 5, 10, 20, 30, and 50 optimization steps. BIDS ($\beta = 0.8$) retains higher CLIPScore as optimization continues, while DNO ($\gamma = 1$) reaches higher reward.

Regularization Strength. Table 7 compares the two methods across regularization strengths at 50 optimization steps, using the same prompts, seed, and sampler as Figure 12. For BIDS, CLIPScore

remains between 24.64 and 25.00 across the tested β values, while larger β reduces reward. Increasing DNO's γ also reduces reward and raises CLIPScore. We fix β per reward–backbone setting as described in Appendix A.1.

Table 7: **Reward and prompt alignment across regularization strengths.** Mean Darkness reward and CLIPScore over 45 Simple Animals prompts on SDv1.5, using one seed, 50 optimization steps, and 50 DDIM denoising steps. BIDS maintains similar CLIPScore across the tested β values while larger β reduces reward.

Method	Setting	Darkness Reward ↑	CLIPScore ↑
BIDS	$\beta = 0.1$	0.9517	24.6376
	$\beta = 0.5$	0.9389	24.7578
	$\beta = 0.8$	0.9283	24.9000
	$\beta = 1.5$	0.9151	24.6539
	$\beta = 5$	0.8809	24.9956
	$\beta = 20$	0.8331	24.7822
DNO	$\gamma = 0$	1.3130	14.7509
	$\gamma = 0.1$	1.2514	16.4081
	$\gamma = 1$	1.0384	20.3404
	$\gamma = 10$	0.9664	23.8907

Early Stopping Across Regularization Strengths. We test whether stopping DNO earlier can match BIDS's darkness reward while meeting the same minimum mean CLIPScore. Figure 13 allows both methods to choose their stopping step and regularization strength.

We run each method with every regularization value listed in Table 7: β for BIDS and γ for DNO. We record darkness reward and CLIPScore both before optimization and after each optimization step. For each regularization value and step count, we compute the mean darkness reward over the 45 prompts and, separately, the mean CLIPScore over those same prompts. The regularization value and step count are the same for all 45 prompts, not chosen separately for each prompt.

Table 7 reports only the scores after 50 noise-optimization steps. Figure 13 also includes the scores recorded at earlier steps, when mean CLIPScore can be higher. The left panel compares results after 1, 2, ..., 20 noise-optimization steps; the right panel compares results after 1, 2, ..., 50 noise-optimization steps.

In Figure 13, the horizontal axis specifies the minimum mean CLIPScore required for the comparison. For each minimum mean CLIPScore, we first exclude recorded results below that minimum. Among the remaining results, we select and plot the highest mean darkness reward separately for BIDS and DNO. Repeating this filtering and selection across minimum mean CLIPScores produces the full curves. Thus, each curve shows the highest recorded reward the method achieves while meeting each prompt-alignment requirement, with early stopping allowed. The filtering is used only to construct the figure from recorded scores; it does not impose a CLIPScore constraint during noise optimization.

For example, consider a minimum mean CLIPScore of 26 in the 50-step comparison. The selected BIDS result uses $\beta = 1.5$ and 29 steps. Its mean darkness reward is 0.8161 and mean CLIPScore is 26.12. The selected DNO result uses $\gamma = 10$ and 26 steps. Its mean darkness reward is 0.7427 and mean CLIPScore is 26.17. Thus, BIDS achieves reward–alignment combinations that the tested DNO configurations do not match, even when DNO can choose its regularization strength and stop early. DNO achieves higher rewards when a lower mean CLIPScore is allowed.

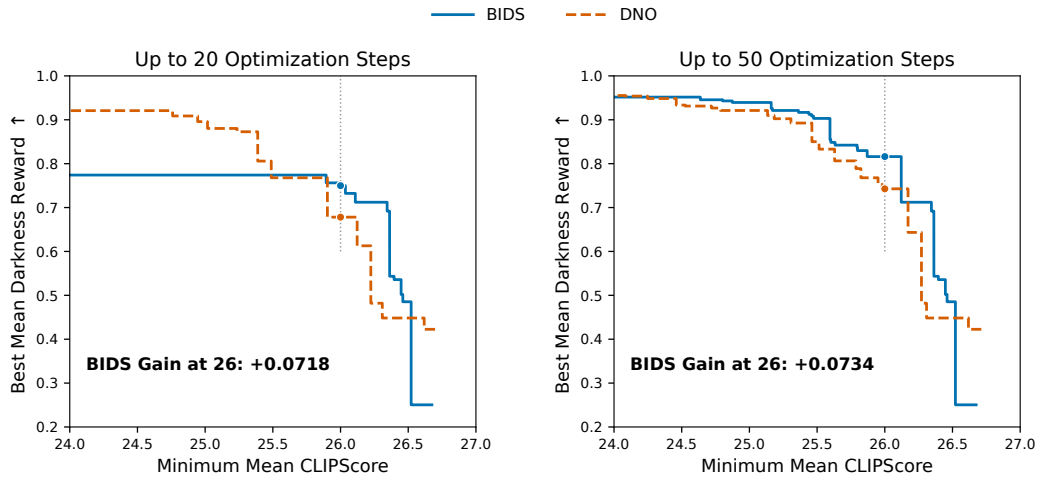

Figure 13: **BIDS can achieve higher darkness reward than DNO at the same minimum mean CLIPScore, even when both methods can stop early.** The marked points show the comparison at a minimum mean CLIPScore of 26. For each minimum mean CLIPScore, we exclude results below that minimum and plot each method's highest remaining mean darkness reward. Abrupt drops occur when a higher minimum CLIPScore excludes the previously selected result, requiring a different regularization setting or stopping step. We compare all regularization values in Table 7. For each value, we consider scores recorded both before optimization and after each of the first 20 (left) or 50 (right) noise-optimization steps. Each regularization value and step count applies to all 45 prompts. We use the same single seed as Figure 12.

A.9 Animal Prompts List

We use the following set of Animal DDPO prompts.

cat	dog	horse	monkey	rabbit	zebra	spider	bird	sheep
deer	cow	goat	lion	tiger	bear	raccoon	fox	wolf
lizard	beetle	ant	butterfly	fish	shark	whale	dolphin	squirrel
mouse	rat	snake	turtle	frog	chicken	duck	goose	bee
pig	turkey	fly	llama	camel	bat	gorilla	hedgehog	kangaroo

A.10 GPT-4o Evaluation

We compute pairwise win rates using GPT-4o as a fixed judge. For each prompt, we present two images in randomized left/right order and request a single binary choice (1 or 2) at temperature 0. We use a single judging template across all experiments; the only placeholders are the text prompt and the target reward specification, which includes a short natural-language *reward definition* describing what constitutes higher reward.

System Prompt

You are a helpful, impartial, and precise assistant that follows the user's instructions and outputs exactly one character: 1 or 2.

User Prompt Template

You are judging Image 1 vs Image 2.

Text prompt: `<PROMPT>`
Target reward: `<TARGET_REWARD>`
Reward definition: `<REWARD_DEFINITION>`

Decision rule:

1. **Prompt adherence first:** Decide whether each image matches the prompt's content *and intended style*. If only one matches, select it.
2. **If both match:** select the image that better satisfies the target reward, *without sacrificing prompt adherence*.
3. **Anti-hacking check (any reward):** do not reward degenerate solutions that maximize the reward by collapsing content (e.g., near-solid white/black, blank/washed-out, extreme artifacts), unless the prompt explicitly calls for it.
4. **Penalize unrequested defects:** text/watermarks/signatures, severe distortions, broken anatomy/geometry, heavy banding, extreme oversharpen halos, or noisy texture spam.

Output: exactly one character, `1` or `2`.

Reward definitions. For the experiments in this paper, we instantiate `<REWARD_DEFINITION>` with short descriptions (e.g., *Aesthetic:* overall visual appeal consistent with intended style; *Brightness:* higher luminance while keeping the subject recognizable; *Darkness:* lower luminance while keeping the subject recognizable).

A.11 Structural Collapse Persists at Scale

A natural question regarding inference-time alignment is whether structural collapse is simply an artifact of limited model capacity. Under this assumption, a sufficiently large and expressive base model might possess a robust enough internal prior to resist reward-induced degradation without requiring explicit trajectory constraints.

To test this hypothesis, we evaluate unconstrained noise optimization on FLUX-schnell Labs (2024), a state-of-the-art 12-billion parameter generation model. We optimize the initial latents to maximize the HPDv2 reward. As demonstrated in Figure 14, scaling model capacity to 12B parameters does not prevent structural collapse. As the number of optimization steps increases, the unconstrained trajectory systematically strips away volumetric textures and complex shading. To artificially inflate the scalar reward score, the optimizer degrades high-fidelity subjects into flat, high-contrast abstractions.

Unconstrained noise optimization causes structural collapse even on the 12B FLUX-schnell model. Increasing model size alone is therefore insufficient to prevent structural collapse during noise optimization.

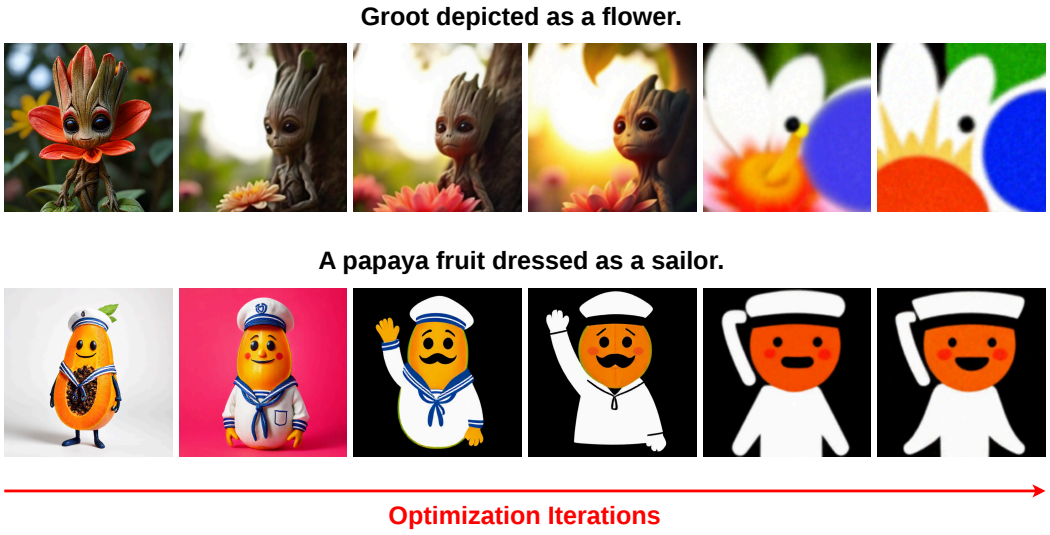

Figure 14: **Massive capacity does not prevent structural collapse.** We evaluate unconstrained noise optimization on the 12B FLUX-schnell model to maximize the HPDv2 reward. As optimization steps increase (left to right), the trajectory strips away volumetric textures and shading, degrading high-fidelity subjects into flat abstractions. This visual degradation demonstrates that the failure mode observed in smaller models persists at scale.

A.12 Theoretical Extension to Rectified Flow

We extend the score-based objective to rectified-flow models by expressing the marginal score in terms of the predicted velocity.

In Rectified Flow, the standard interpolation is $x_t = (1-t)x_0 + t\epsilon$, where $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ is independent of x_0 conditional on the prompt c, and $t \in (0, 1]$. Let $v_t(x, c) = \mathbb{E}[\epsilon - x_0 \mid x_t = x, c]$ denote the conditional mean velocity. For the interpolation marginal $p_t^\theta(x \mid c)$, the marginal score is

$$s_{\theta,t}(x, c) = \nabla_x \log p_t^\theta(x \mid c) = -\frac{x + (1-t)v_t(x, c)}{t}.$$

We estimate the marginal score by substituting the learned velocity v_θ into this identity.

For K sampling times t_k with weights λ_k, use the base latents $x_{t_k}^{\text{base}}$ and optimized latents $x_{t_k}^{\text{opt}}$ defined in Section 4. For one paired base and optimized trajectory, the practical objective is:

$$\mathcal{L}_{\text{RF-BIDS}}(z, c) = -r(x_0^{\text{opt}}, c) + \beta \left| \sum_{k=1}^{K} \lambda_k \left(\|s_{\theta,t_k}(x_{t_k}^{\text{base}}, c)\|^2 - \|s_{\theta,t_k}(x_{t_k}^{\text{opt}}, c)\|^2 \right) \right|.$$

The estimated marginal score can therefore be used in the same score-based objective.