ProxySearch: Decoupled Inference-Time Scaling for Diffusion Models via Asymmetric Noise-Rank Transfer

Seungwook Kim[1] **Jongmin Lee**[2]

[1]GenGenAI, [2] Chung-Ang University

wookiekim@gengen.ai jmlee@cau.ac.kr

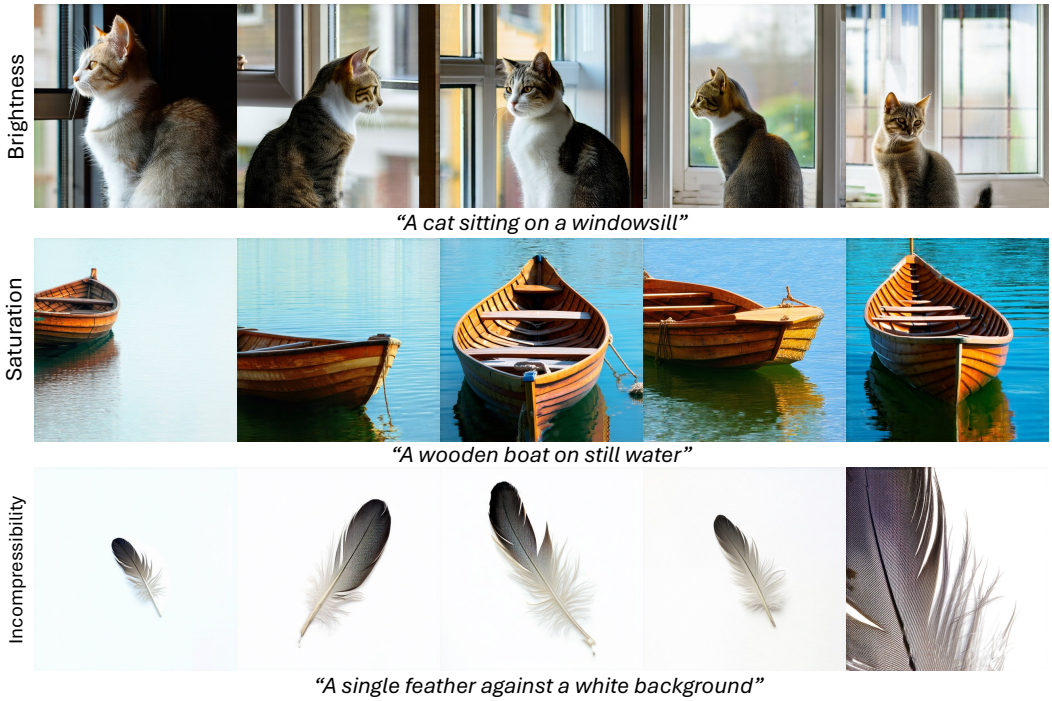

Figure 1: **PROXYSEARCH sets an image statistic by choosing the initial noise.** Each row is one SD3.5 prompt, with the target value of one statistic increasing from left to right.

Abstract

Noise search at inference time improves text-to-image diffusion but requires many multi-step teacher rollouts per query. Distilled few-step models are now routinely released, yet their potential as *proxies* inside noise search is uncharacterized. We measure proxy–teacher rank agreement across three distillation regimes and three backbones (DiT, MMDiT, UNet) and find a sharp asymmetry. On our main pool, closed-form image statistics transfer from the proxy to the teacher (Spearman ρ up to 0.80), while learned semantic scorers do not ($\rho \leq 0.12$). Paired null controls trace this rank to the first denoising step for coarse statistics and to distillation for finer ones. We use the asymmetry for noise search under a *joint constraint*: a semantic reward specifies content, and a numeric target specifies image statistics. In **PROXYSEARCH**, the proxy filters candidates by statistical match in 1 step, and the teacher returns the survivor with the highest semantic reward. No weight trades one objective against the other. On FLUX-1024, PROXYSEARCH tightens the statistical target by 21–23% at the same number of teacher trajectories, with no detected loss on the optimized reward.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

1 Introduction

Besides training better denoisers, text-to-image diffusion quality can also be improved by spending more compute at inference time, much as test-time compute is used in large language models [64]. *Noise search* draws many initial latents, renders each with the teacher denoiser, and keeps the highest-reward output [48]. Refinements re-weight candidates along the denoising trajectory [63, 72], optimize the noise directly [70], or learn a noise prior [80]. In all of them, quality grows with the number of teacher rollouts per query. Every extra candidate costs a full multi-step rollout, so search cost grows linearly with the pool size.

Yet most major teachers now ship with a distilled few-step companion (FLUX-schnell, SD3.5-Large-Turbo, SDXL-Turbo [58, 59]) that needs 1–4 instead of 28–50 steps. Search pipelines use this sibling only as a faster generator, and whether it can also judge candidates has not been tested. Inside search, a proxy helps only if it ranks candidates as the teacher would; producing similar images on average is not enough. We measure, for each verifier, whether a distilled sibling ranks a prompt's candidate noises as its teacher does. The answer depends on the verifier (Fig. 3). Eight closed-form image statistics, such as brightness and contrast, transfer from proxy to teacher (Spearman ρ up to 0.80 on our main pool), while three learned semantic rewards (CLIP [53], ImageReward [74], aesthetic [60]) stay at $\rho \leq 0.12$ on the same noises. On every pair, rank agreement on brightness is 6–16$\times$ that on CLIP score. Paired null controls show that the raw noise does not carry this rank. A single step of the *undistilled* teacher already ranks coarse statistics such as brightness, and distillation extends rank agreement to finer ones such as edge density. The distilled proxy is the only 1-step model we test that ranks all eight. So the cheap proxy can screen many candidates on image statistics, while semantics still have to be judged by the teacher.

A user may want a batch of images rendered to one brightness and contrast, or a synthetic image set balanced along a statistic [49]. Choosing the noise changes such a look (Fig. 1), but naming the look in the prompt does not set it (Fig. 5). We therefore cast search as a *joint constraint*. The prompt and a semantic reward, such as compositional accuracy or learned preference, specify content. A separate numeric target τ, such as a brightness and contrast level, specifies image statistics. Search should return an output that scores well on the semantic reward *and* hits τ, without scalarizing the two. PROXYSEARCH (Fig. 2) routes each part of the constraint to the model that ranks it. The proxy filters N candidates by statistical match at about one NFE each, and the teacher renders the top K and returns the one with the highest semantic reward. The method has two integer parameters, N and K, and no weight trading one objective against the other.

On full GenEval at FLUX-1024 (Tab. 1), PROXYSEARCH runs the same 8 teacher trajectories as teacher-only argmax ($+16\%$ denoising NFE). It tightens paired statistical ℓ_1 by 21–23$\%$ under three search rewards (paired $t = -19.3$ under GenEval) and stays on par with teacher argmax on the reward it optimizes. More teacher candidates do not produce this gain: under semantic-only selection, 16- and 64-candidate teacher pools leave ℓ_1 unchanged (Sec. 6.2). The gain therefore comes from screening on the statistic, not from a larger candidate pool. On a 100-cell pilot, a composite reward that scores all 64 teacher renders reaches a lower ℓ_1 than PROXYSEARCH, but at $6.9\times$ the total NFE (Sec. 6.3). The tightening also replicates in 50-prompt pilots on SD3.5 [14] and SDXL [51], at a GenEval cost of 2–4 pp relative to teacher-only argmax.

Our contributions are as follows:

- A characterization of proxy-to-teacher rank agreement on released distilled pairs, which separates statistics from semantics. On our main pool, closed-form image statistics transfer with Spearman ρ up to 0.80, while learned semantic rewards remain at $\rho \leq 0.12$. Paired null controls show that a single undistilled teacher step already ranks coarse statistics and that distillation extends rank agreement to finer ones (Sec. 4).
- PROXYSEARCH, a reformulation of inference-time noise search as a joint constraint on a statistical target and a user-chosen semantic reward, parameterized by two integers (N, K) with no trade-off weight to tune (Sec. 5).
- Empirical evidence that PROXYSEARCH tightens the statistical target by 21–23$\%$ in paired ℓ_1 on the full GenEval suite at FLUX-1024, at the same number of teacher trajectories as teacher-only argmax, while staying on par on the optimized reward. Pilots on SD3.5 and SDXL replicate the tightening (Sec. 6).

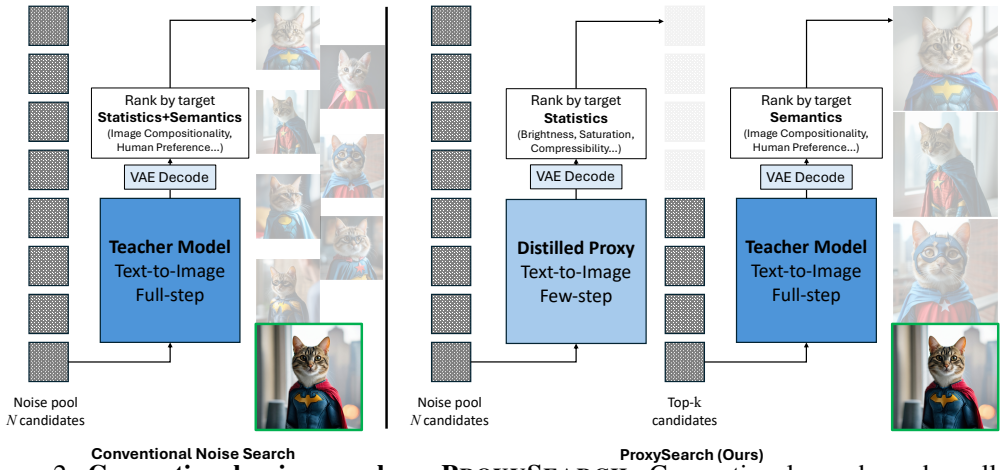

Figure 2: **Conventional noise search vs. PROXYSEARCH.** Conventional search renders all N candidates with the teacher $\mathcal{M}$ and keeps the highest-reward one. PROXYSEARCH scores the same N noises on g_{stat} with the distilled proxy $\tilde{\mathcal{M}}$ at about 1 NFE each, renders only the top K with the teacher, and returns the survivor with the highest semantic reward.

2 Related Work

Inference-time scaling for diffusion. A growing line of work treats the initial noise as a search variable. Ma et al. [48] show that random search, zero-order optimization, and path-level search all keep improving as verifier compute grows. Particle-based variants [63, 72] resample partial trajectories, gradient-based variants [15, 70, 80] optimize the latent or learn a noise prior, and initialization and trajectory-stylizing methods [1, 19, 32, 33] steer the search distribution itself. Reward-guided and restart-style sampling [2, 28, 75, 79], building on classifier-free guidance [23], similarly trades teacher NFE for higher-reward outputs. Sampling- and optimization-time techniques also improve generation beyond single images, in video [33] and text-to-3D [30, 36], and multi-view diffusion models [31, 34] generate consistent views for 3D generation. The search methods above score every candidate on teacher outputs and treat the reward as a single scalar. We instead split the reward into statistics, which transfer from a distilled proxy at little cost, and semantics, which do not.

Distillation of diffusion models. A parallel literature compresses the 28–50-step teacher schedule [24, 27, 43, 66] into 1–4 steps, through progressive distillation [57], consistency models [47, 68], rectified-flow and one-step variants [45, 46], adversarial distillation (LADD [58], which produces FLUX-schnell, and the related ADD [59] behind SD3.5-Large-Turbo and SDXL-Turbo), distribution matching [DMD/DMD2; 77, 78], and Hyper-SD [54]. This literature evaluates the distilled model as a standalone fast generator, reporting FID [22] or human preference against the teacher. To our knowledge, nobody has checked whether the proxy keeps the teacher's *rank* of a candidate-level reward across noises drawn for the same prompt; we measure this in Sec. 4.

Cheap proxies for expensive search. Pairing a cheap, biased estimator with an expensive accurate one recurs across machine learning: multi-fidelity Bayesian optimization [26], Hyperband-style allocation [16, 39], AlphaGo's value-network/rollout-policy split [62], one-shot NAS [50], speculative decoding [5, 6, 38, 69], learned process verifiers [4, 9, 42], and cascaded diffusion [25]. In each, the proxy is *designed* to work with the full evaluator. Diffusion distillation provides no such guarantee. The distilled sibling is trained to produce teacher-quality samples in few steps, not to preserve the teacher's ranking of inputs, and for the same noise the two models produce *different* images. Nor does the cascade pattern predict that rank agreement would split along statistics and semantics. We instead measure where it appears: on the statistical axes (Sec. 4).

Reward fine-tuning of diffusion. A parallel paradigm bakes reward signals into the model rather than into the sampler: by backpropagating a differentiable reward (DRaFT [8], AlignProp [52], TextCraftor [40]), by policy gradients on the denoising MDP (DPOK [17], DDPO [3]), possibly with intrinsic self-confidence rewards instead of external reward models [29], by preference optimization (Diffusion-DPO [71] and follow-ups [41, 76]), or by rejection sampling and adjoint matching [12, 13, 37]. These methods produce a single tuned model per reward. PROXYSEARCH is orthogonal: it is

training-free and leaves both checkpoints untouched. A reward-post-trained teacher still agrees with a fixed distilled proxy on statistical rank (Sec. 4).

3 Preliminaries

Diffusion sampling and noise search. A diffusion model $\mathcal{M}$ [10, 24, 27, 66, 67] maps an initial latent $z_T \sim \mathcal{N}(0, I)$ and a text prompt c to a latent $z_0 = \mathcal{M}_T^{\to 0}(z_T, c)$, decoded by a VAE to image x [56]. Inference-time noise search draws N noises $\{z_T^{(i)}\}_{i=1}^N$, renders $x^{(i)} = \mathrm{VAE}(\mathcal{M}(z_T^{(i)}, c))$, scores each with a verifier $v(\cdot)$, and returns $x^{(i^*)}$ with $i^* = \arg\max_i v(x^{(i)})$. Cost is dominated by the N teacher evaluations, each requiring 28–50 denoising steps. We count cost in neural function evaluations (NFE), one per denoising step. NFE excludes classifier-free-guidance passes, VAE decoding, text encoding, and verifier scoring, and Sec. 6.1 reports measured render time.

Distilled proxy. A distilled variant $\tilde{\mathcal{M}}$ (FLUX-schnell, SD3.5-Large-Turbo, and SDXL-Turbo [58, 59]) shares the architecture, tokenizer, VAE, and latent space of $\mathcal{M}$ but produces an output in k_{proxy} steps. We use $k_{\text{proxy}} = 1$ unless noted, and Tab. 6 lists the setting of every run. We write $\tilde{x}^{(i)} = \mathrm{VAE}(\tilde{\mathcal{M}}(z_T^{(i)}, c))$ for the proxy image on the *same* noise and prompt.

Two verifier families. We partition the verifiers used in this paper into two families:

- **Statistics:** model-free global image statistics: brightness, contrast, saturation, colorfulness (Hasler–Süsstrunk [21]), sharpness (Laplacian variance), compressibility (JPEG size), edge density (Sobel [65]), and color temperature. Each costs $\ll$ one teacher NFE.
- **Semantics:** prompt-dependent, model-based semantic scores. The rank-transfer bank uses CLIP similarity [53], ImageReward [74], and a LAION aesthetic predictor [60]; GenEval pass [18], PickScore [35], and HPSv2 [73] serve as search objectives but are not part of the rank-transfer measurement.

Exact definitions, units, and the GenEval selection and tie-breaking rules are given in App. A.

Statistical target. A user supplies, separately from the prompt, a target vector $\tau \in \mathbb{R}^d$ over a chosen subset of d statistical axes, either directly or via a preset. We measure adherence by the negative ℓ_1 distance $g_{\text{stat}}(\ell; \tau) = -\|\ell - \tau\|_1$, where $\ell = \mathrm{stat}(x)$ is the statistics vector of image x restricted to the d axes named by τ. For example, a "bright, high-contrast" look may correspond to $\tau = (\text{brightness} = 0.72,\ \text{contrast} = 0.26)$. The target τ is a per-query specification, not a model parameter, and our main experiments use two fixed targets (Sec. 6.2).

4 Distilled proxies rank image statistics, not learned semantic scores

Protocol and the split. For each prompt and each of 32 noises, we render the proxy image $\tilde{x}^{(i)}$ ($k_{\text{proxy}} = 1$) and the teacher image $x^{(i)}$ at 512^2 and score both with 11 verifiers (8 statistics, 3 semantics). For each verifier, we then compute the Spearman correlation between the proxy and teacher rankings of the 32 noises of a prompt. We average this per-prompt ρ over 100 synthetic template prompts (App. A) on the three pairs of Tab. 5: FLUX (DiT, LADD [58]), SD3.5 (MMDiT [14], ADD [59]), and SDXL (UNet [51], ADD). Fig. 3 (a) shows the result. On our main pool, statistical verifiers transfer at $\rho \in [0.10, 0.80]$ and semantic verifiers stay at $\rho \leq 0.12$. Strength varies across axes and backbones, but on every pair rank agreement on brightness is 6–16$\times$ that on CLIP score. The proxy ranks image statistics but barely ranks semantics. Selecting candidates by proxy-side semantic scores stays far below teacher-side selection (Tab. 8). The labels "statistics" and "semantics" describe the eight statistics and three semantic scorers we measure. App. F discusses other verifiers.

Where the statistical rank comes from. We run three conditions on one shared pool of 12 GenEval prompts $\times$ 32 noises at 1024^2, so that noise j is the identical z_T for every model. Each condition is scored against the full teacher (Fig. 3 (b), Tab. 10). Statistics of the raw initial latent do not transfer at all. A single step of the undistilled teacher already preserves coarse-statistic rank (brightness $\rho = 0.68$ on FLUX, 0.91 on SD3.5). Distillation adds rank agreement on finer statistics where the 1-step

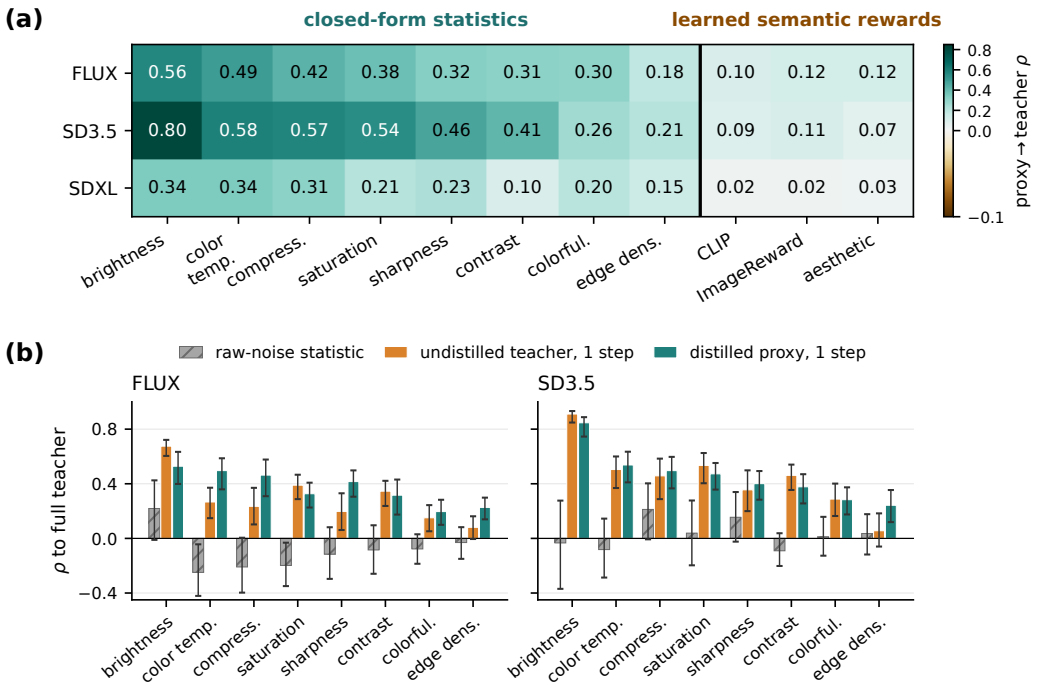

Figure 3: **Proxy–teacher rank correlation splits along statistics vs. semantics. (a)** Spearman ρ between the proxy's and the teacher's rankings of the same noises, for 11 verifiers on 3 model pairs. Statistics transfer with ρ up to 0.80, while semantics stay at $\rho \leq 0.12$. **(b)** Paired null controls on a shared noise pool. Raw-noise statistics do not transfer, one undistilled teacher step ranks coarse statistics, and the distilled proxy also ranks finer ones (most clearly on FLUX). It is the only condition significant on all eight.

teacher is weaker. On FLUX it raises color temperature $0.27 \to 0.50$, compressibility $0.23 \to 0.46$, sharpness $0.20 \to 0.42$, and edge density $0.08 \to 0.23$. On SD3.5 the gain is confined to edge density ($0.06 \to 0.24$). The distilled proxy is the only 1-step condition with significant rank agreement on all eight statistics. Two further controls indicate that the statistic a teacher produces from a given noise is largely model-specific (App. C.2). Two different teachers rendering the same z_T agree only weakly (mean $|\rho| = 0.07$ over the eight statistics), with a weak shared brightness component. A predictor trained on raw z_T recovers part of a teacher's brightness rank within a model but transfers little across models.

The first prompt-conditioned denoising step thus seeds coarse statistical rank, and distillation extends it to finer statistics. This fits the coarse-to-fine view of denoising [11, 55], although our controls do not pin the effect down to one mechanism.

Robustness of the split. On FLUX and SD3.5, a 4-step proxy does not recover semantic rank ($\rho < 0.16$) and leaves brightness and color-temperature rank unchanged, at about $3\times$ the proxy cost. On SDXL it lowers most statistical axes (App. B.3). A DMD2 score-distilled SDXL proxy [77] shows the same split as the LADD/ADD proxies (App. B). The split therefore does not seem tied to one particular distillation loss. Released teachers are also increasingly post-trained with reward RL [29, 44, 71], and a fixed proxy serves such a teacher only if the split survives the tuning. We pair one fixed proxy, a third-party distilled SD3.5-Medium model run at 1 step, with the SD3.5-Medium teacher in three versions: frozen, with a FlowGRPO text-rendering LoRA, and with a FlowGRPO PickScore LoRA. Each arm uses 100 FlowGRPO prompts $\times$ 32 noises at 1024^2, and arms are paired by prompt only (App. B.4). The text-rendering LoRA stays within about 0.06 of the frozen teacher on every statistical axis (Fig. 4). The PickScore LoRA lowers statistical ρ by roughly 0.14–0.23 on every axis except edge density ($0.133 \to 0.068$). Brightness rank agreement (0.68) remains $15\times$ that of the best semantic verifier (0.045). Of the 50 PickScore prompts, 48 lie in that LoRA's training split, and other RL objectives or LoRA settings may weaken the split further.

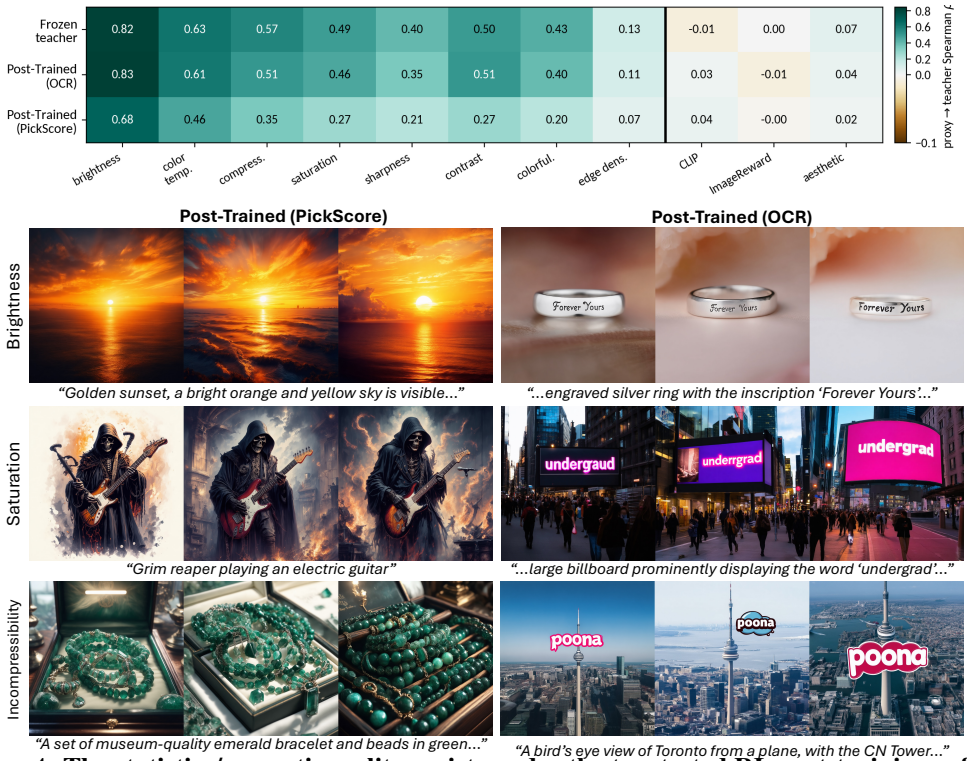

Figure 4: **The statistics/semantics split persists under the two tested RL post-trainings of the teacher.** *Top:* proxy→teacher Spearman ρ across 11 verifiers (rows). *Bottom:* qualitative samples per statistical axis (brightness, saturation, incompressibility).

5 ProxySearch

5.1 Problem setup: joint-constraint search



Figure 5: **Prompt wording alone leaves wide variance on the named statistical axis.** Five prompts (columns) that already name a statistical axis directly ("dark", "bright", "high-contrast", "saturated", "desaturated") still leave wide variance on the named axis across the 32-noise output pool.

Prompt wording alone does not pin down numeric image statistics: even prompts that name a statistical axis directly ("dark", "bright", "saturated") still vary widely on the named axis across noises (Fig. 5). Let $\mathcal{M}$ be a teacher diffusion model and $\tilde{\mathcal{M}}$ its distilled few-step proxy (Sec. 3). Given a prompt c, a user-specified semantic reward v_{sem} such as GenEval pass, PickScore, or HPSv2, and a statistical target τ (Sec. 3), the task is to select x^* from a noise pool $\{z_T^{(i)}\}_{i=1}^N$ that satisfies

both criteria. An idealized form of the task is the constrained argmax $\max_i v_{\text{sem}}(x^{(i)})$ subject to $g_{\text{stat}}(\text{stat}(x^{(i)}); \tau) \geq -\epsilon$ for a user tolerance ϵ, which PROXYSEARCH approximates with a top-K heuristic (Sec. 5.2). All reported experiments use the target-matching g_{stat} of Sec. 3. The algorithm applies to any g_{stat} for which larger values are preferred (App. A).

The two criteria are conventionally handled in one of two ways: (a) collapsed into a single scalarized reward $v_{\text{sem}}(x) + \lambda\, g_{\text{stat}}(\text{stat}(x); \tau)$ and maximized by teacher-only search, or (b) handled on the teacher alone, either by post-processing the teacher-argmax image toward τ or by filtering teacher renders on g_{stat} before the semantic argmax. Neither route uses the proxy, although Sec. 4 shows that it ranks statistics. Route (a) scores both terms on teacher renders under a tuned weight λ. Route (b) spends teacher rollouts on candidates whose statistics could have been screened cheaply, and post-processing reaches τ only by editing pixels. We instead treat the two as a *joint constraint*: statistics are screened on cheap proxy renders, and semantics are scored only on teacher renders, the images actually returned. Sec. 6 compares these baselines with PROXYSEARCH.

5.2 Algorithm and cost

PROXYSEARCH routes the statistical part of the constraint to the proxy and the semantic part to the teacher (Fig. 2). Given a noise budget N and a survivor budget K:

1. **Proxy statistical filter.** For each $z_T^{(i)}$, $i = 1, \ldots, N$, render $\tilde{x}^{(i)} = \text{VAE}(\tilde{\mathcal{M}}(z_T^{(i)}, c))$ at k_{proxy} proxy steps ($k_{\text{proxy}} = 1$ unless noted). Score the statistical objective $\ell^{(i)} = g_{\text{stat}}(\text{stat}(\tilde{x}^{(i)}); \tau)$. Retain the indices $\mathcal{S} \subseteq [N]$ with the K largest $\ell^{(i)}$.
2. **Teacher semantic argmax.** For each $i \in \mathcal{S}$, render $x^{(i)} = \text{VAE}(\mathcal{M}(z_T^{(i)}, c))$ at the teacher's full step count. Return $x^{(i^*)}$ with $i^* = \arg\max_{i \in \mathcal{S}} v_{\text{sem}}(x^{(i)})$.

Teacher renders are not re-scored on g_{stat}, so the returned image carries no tolerance guarantee. The survivors are the K best proxy candidates even when τ is not attainable in the pool. PROXYSEARCH has no weight λ, but it still trades the two objectives. A smaller K/N screens harder on statistics and leaves fewer candidates for the semantic argmax. The attainability of τ in the pool and the scaling of the axes inside g_{stat} also shift the balance. The final argmax is optimal only over the survivor set $\mathcal{S}$ and the chosen v_{sem}.

The pipeline costs $N \cdot c_{\text{proxy}} + K \cdot c_{\text{teacher}}$ NFE. With a 1-step proxy, $c_{\text{proxy}}/c_{\text{teacher}}$ is $1/50$ on FLUX-1024 and SDXL (50 teacher steps) and $1/28$ on SD3.5 (28 teacher steps). At the FLUX-1024 operating point ($N = 64$, $K = 8$), PROXYSEARCH runs the same 8 teacher trajectories as an 8-candidate teacher-only argmax and adds 64 proxy NFE to the teacher's 400. This is 16% more denoising NFE. We call this matched *teacher trajectories*, not matched total compute, and Sec. 6.1 reports the measured render time.

6 Experiments

6.1 Setup

We evaluate on FLUX.1-dev / FLUX-schnell (DiT, LADD [58]), SD3.5-Large / SD3.5-Large-Turbo [14] (MMDiT, ADD), and SDXL / SDXL-Turbo [51, 59] (UNet, ADD). The compositional benchmark for image pairs is GenEval [18]. The human-preference rewards are PickScore [35] and HPSv2 [73], and CLIP, ImageReward, and the LAION aesthetic predictor serve as held-out semantic measurements. Search experiments use GenEval prompts. GE denotes GenEval scored with the official detection pipeline, reported as the per-image fraction of satisfied prompt requirements averaged over cells.[1] Details in App. A.

Compute and pairing. At matched teacher trajectories, teacher-only argmax renders N_t teacher candidates and PROXYSEARCH keeps $K = N_t$ survivors, so both run K teacher trajectories per query. PROXYSEARCH adds $N \cdot k_{\text{proxy}}$ proxy NFE: 464 vs. 400 NFE at the FLUX-1024 operating point (+16%). Measured render time grows more than NFE because each proxy render adds a VAE decode

[1] GE-driven selection uses an in-process re-implementation of the GenEval checker (Mask2Former via HF `transformers`); reported GE re-scores the selected images with the official pipeline. Selection keys and tie-breaks are in App. A.

Table 1: **PROXYSEARCH tightens statistics while matching teacher argmax on the optimized semantic reward.** FLUX-1024, $N = 64$, $K = N_t = 8$, $k_{\text{proxy}} = 1$, $n = 1106$ paired prompt–target cells (Teacher-seq rows extrapolated from a paired 100-cell subset; see text), three search objectives: GE-driven (a); PickScore-driven (b); HPSv2-driven (c). Bold marks the best entry per column within each panel (ties all bold). Paired deltas to teacher argmax are in App. D.1.

Method	statistical $\ell_1 \downarrow$	GE %↑	PickScore ↑	HPSv2 ↑	CLIP ↑	ImageReward ↑	aesthetic ↑
(a) GenEval-driven search (v_{sem}=GE)							
Teacher argmax ($N_t = 8$)	0.302	81.6	23.50	0.311	0.364	1.06	**5.35**
PROXYSEARCH ($K = 8$)	**0.233**	**82.5**	**23.52**	**0.312**	**0.365**	**1.12**	**5.35**
Teacher-seq ($N_t = 16$, 2× NFE)	0.247	**82.5**	23.49	0.307	0.364	1.03	5.34
(b) PickScore-driven search (v_{sem}=PickScore)							
Teacher argmax ($N_t = 8$)	0.297	78.9	24.14	0.322	0.379	**1.25**	**5.35**
PROXYSEARCH ($K = 8$)	**0.232**	79.2	24.14	0.322	0.378	1.24	**5.35**
Teacher-seq ($N_t = 16$, 2× NFE)	0.247	**81.2**	**24.15**	**0.323**	**0.384**	1.19	**5.35**
(c) HPSv2-driven search (v_{sem}=HPSv2)							
Teacher argmax ($N_t = 8$)	0.296	79.0	23.91	**0.329**	**0.377**	**1.27**	**5.35**
PROXYSEARCH ($K = 8$)	**0.234**	78.7	23.88	**0.329**	**0.377**	1.26	**5.35**
Teacher-seq ($N_t = 16$, 2× NFE)	0.251	**79.2**	**23.95**	0.328	**0.377**	1.24	**5.35**

and a text encoding: about $+24\%$ on FLUX (186 vs. 230 s). App. A.3 gives guidance passes and approximate render times for SD3.5 ($+28\%$) and SDXL ($+49\%$). Within a run, all methods share one noise pool indexed by $(\text{seed}, \text{prompt_id}, \text{candidate_idx})$. Teacher argmax uses the first N_t candidates of the pool that PROXYSEARCH screens, so reported differences are paired. Pools are not shared across runs, so comparisons are paired only within a run. All main results use one fixed K, and Sec. 6.4 sweeps it.

6.2 PROXYSEARCH tightens statistics at matched teacher trajectories

We run the full GenEval suite at FLUX-1024 with two statistical targets, $\tau_{\text{bright}} = (\text{brightness}, \text{contrast}) = (0.72, 0.26)$ and $\tau_{\text{dark}} = (0.22, 0.20)$. This gives $n = 1106$ paired prompt–target cells (553 prompts × 2 targets). We compare PROXYSEARCH ($N = 64$, $K = 8$, $k_{\text{proxy}} = 1$) against teacher argmax with $N_t = K$, so both methods render 8 teacher trajectories per cell. Tab. 1 reports three search objectives ($v_{\text{sem}} \in \{\text{GE, PickScore, HPSv2}\}$). Every reward other than the search objective is a held-out post-selection measurement. Tab. 1 also lists *Teacher-seq*, which spends 2× the teacher NFE: it renders $N_t = 16$ teacher candidates, keeps the 8 with the smallest teacher-side statistical ℓ_1, and argmaxes v_{sem} over them. Teacher-seq was measured on a paired 100-cell subset (50 GenEval prompts × 2 targets). Its rows are extrapolated to $n = 1106$ as the PROXYSEARCH value plus the paired Teacher-seq−PROXYSEARCH difference on that subset.

Main result. PROXYSEARCH tightens the statistic under every search objective: the paired ℓ_1 delta is -0.070 ± 0.004 SE in panel (a), -0.065 in panel (b), and -0.062 in panel (c). In panel (a), the paired t is -19.3 and the Wilcoxon signed-rank z is -18.5. On the optimized reward, PROXYSEARCH and teacher argmax are within noise of each other in every panel, and the held-out rewards behave the same way (App. D.1). The proxy filter does not depend on v_{sem}, although the final teacher argmax does.

Teacher-seq. At 2× the teacher NFE, Teacher-seq stays looser than PROXYSEARCH on the statistic in every panel and is ahead on several semantic columns in panels (b) and (c).

More teacher candidates do not substitute for the proxy. On the first 200 GenEval prompts, doubling the teacher-argmax pool to $N_t = 16$ raises GE by 1.25 pp but leaves ℓ_1 unchanged ($0.298 \rightarrow 0.296$, App. D.2). On a separate 100-cell pilot, giving teacher argmax the full 64-candidate pool that PROXYSEARCH screens also leaves ℓ_1 unchanged ($0.304 \rightarrow 0.306$). On that pilot, PROXYSEARCH reaches 0.230 at the teacher cost of 8 candidates (App. D.2). The gain comes from ranking on the statistic, not from pool size.



Figure 6: **Qualitative comparison at fixed noise seed.** Four prompts (rows) rendered by teacher argmax, composite reward, and PROXYSEARCH (ours) over the same 32-noise pool with HPSv2 as the semantic verifier. Rows 1–2 target $\tau_{\text{bright}} = (\text{brightness} = 0.72,\ \text{contrast} = 0.26)$ (bright, high contrast); rows 3–4 target $\tau_{\text{dark}} = (\text{brightness} = 0.22,\ \text{contrast} = 0.20)$ (dark, low contrast). On these four prompts, PROXYSEARCH attains the lowest statistical ℓ_1 to the target among the three methods while keeping HPSv2 and GenEval comparable; aggregate results are in Tab. 1 and Tab. 2.

Table 2: **Comparison with alternatives (GE-driven, 100-cell pilot).** FLUX at 512^2, 50 GenEval prompts $\times$ 2 targets, 64-noise pool. Teacher argmax and composite reward select among $N_t = 16$ teacher renders. PROXYSEARCH keeps $K = 4$ survivors from a 4-step proxy (800 vs. 200 teacher NFE). Random sampling is the teacher render of the first noise. Both post-process rows edit the teacher-argmax image.

Method	statistical $\ell_1 \downarrow$	GE %↑	PickScore ↑	HPSv2 ↑	CLIP ↑	ImageReward ↑	aesthetic ↑
Random sampling	0.303	81.0	23.11	0.302	0.354	0.96	**5.34**
Teacher argmax ($N_t = 16$)	0.302	87.0	23.18	0.303	0.348	0.97	**5.34**
Composite reward ($N_t = 16$, $\lambda = 0.1$)	0.205	86.5	**23.31**	**0.306**	0.352	1.02	**5.34**
Post-process shift (full)	**0.066**	84.0	22.95	0.288	0.351	0.90	5.32
Post-process shift (clamped to statistical$\approx$0.22)	0.218	**89.0**	23.19	0.301	0.348	0.96	**5.34**
PROXYSEARCH ($K = 4$, 4-step proxy)	0.230	**89.0**	23.26	**0.306**	**0.355**	**1.04**	**5.34**

6.3 Post-processing, composite reward, and human judgments

Tab. 2 compares PROXYSEARCH with post-processing and a composite reward in a smaller 100-cell pilot at 512^2. Teacher NFE is not matched there: the baselines spend 800 teacher NFE and PROXYSEARCH 200.

Post-processing. Editing the teacher-argmax image moves the statistic directly. The unclamped shift reaches statistical ℓ_1 of 0.066 but trails PROXYSEARCH on GE and on every learned semantic reward. Against the clamped shift, PROXYSEARCH ties on GE (89.0), is slightly looser on statistics (0.230 vs. 0.218), and is higher on PickScore, HPSv2, CLIP and ImageReward (aesthetic tied). The post-process operations are in App. A.

Composite reward. On the 64-candidate pilot of Sec. 6.2, a composite reward that scores all 64 teacher renders reaches a tighter statistic than PROXYSEARCH (0.143 vs. 0.230) at 6.9$\times$ the total NFE (8$\times$ the teacher NFE). PROXYSEARCH recovers about half of the composite's ℓ_1 reduction at the teacher cost of 8 candidates (App. D.2). In the pilot of Tab. 2, the composite reaches 0.205 and a higher PickScore (HPSv2 tied), while PROXYSEARCH is higher on GE, CLIP, and ImageReward. We sweep $\lambda \in \{0, 0.1, 0.5, 1, 2\}$ and report $\lambda = 0.1$. Because the GE term is a binary pass and $\ell_1 < 1$ within a pool, every $\lambda \in \{0.1, 0.5, 1, 2\}$ selects the same image on the evaluated cells, and $\lambda = 0$ reduces to teacher argmax. This holds for these runs only; with a much larger λ, the composite could give up a GE pass for a large ℓ_1 gain. Fig. 6 shows example outputs of teacher argmax, the composite reward and PROXYSEARCH.

Human judgments. In a forced-choice study (App. E), annotators saw the prompt and the requested look on pairs where PROXYSEARCH had shifted the statistic toward that look. They chose the PROXYSEARCH image over the teacher-argmax image in 71.7% of 600 judgments. This arm

measures joint content-plus-look fulfillment. Rates differ by look: 82.0–88.4% for dark, warm, and cool, 70.1% for high sharpness, and at chance for bright/high-contrast (48.1%) and low sharpness (48.2%). In a second arm, annotators were told that one image was generated with the look and the other was an ordinary photo edited afterwards, and asked which looks like the genuine unedited photograph. Against the teacher-argmax image post-processed toward the target, with the unclamped shift for bright and dark looks, they chose the PROXYSEARCH image in 84.2% of 400 judgments. Both images are generated, so this measures perceived naturalness under that framing. For non-tone looks the edit differs from Tab. 2 (App. E). When no look is named, a small pilot shows no preference either way (48.9%, 88/180).

6.4 Survivor budget and other backbones

Survivor budget. We sweep the survivor count $K \in \{2, 4, 8, 16\}$ on a 100-cell FLUX pilot (50 GenEval prompts $\times$ 2 targets, $N = 64$ proxy candidates, GenEval as the selection reward), and hold teacher argmax at $N_t = K$, so both methods run the same number of teacher trajectories except in the $K = 4$ row (Tab. 3). Statistical ℓ_1 tightens by 21.8–25.3% at every K, while GE stays level with teacher argmax ($+0.5$ to $+2.5$ pp, within noise at this size). The deployed 1-step proxy at $K = 8$ gives the same picture (-24.7% ℓ_1, $+1.0$ pp GE).

Table 3: **K-sweep at matched teacher trajectories** ($n = 100$ cells per row: 50 prompts $\times$ 2 targets, 1024^2). ΔGE: paired difference with 95% CI from a bootstrap over prompts. *The $K = 4$ row reuses the run of Tab. 2 (512^2), whose teacher argmax uses $N_t = 16$ rather than $N_t = K$. Noise pools differ across rows; compare within a row.

K	proxy steps	Argmax GE	Ours GE	ΔGE [95% CI]	Argmax stat. ℓ_1	Ours stat. ℓ_1	Δstat. %
2	4	81.0	81.5	$+0.5$ $[-3.0, +4.5]$	0.302	0.226	-25.1
4*	4	87.0	89.0	$+2.0$ $[-2.0, +6.5]$	0.302	0.230	-23.8
8	4	87.0	89.5	$+2.5$ $[-0.5, +6.0]$	0.302	0.226	-25.3
16	4	89.0	90.5	$+1.5$ $[-3.0, +7.5]$	0.300	0.235	-21.8
8	1	88.0	89.0	$+1.0$ $[-3.5, +5.5]$	0.304	0.229	-24.7

Other backbones. With τ_{bright}, $N = 64$, $K = N_t = 8$, and $k_{\text{proxy}} = 1$ on 50 GenEval prompts shared by both backbones ($n = 50$, Tab. 4), PROXYSEARCH reduces statistical ℓ_1 from 0.204 to 0.120 (41%) on SD3.5-Large and from 0.305 to 0.267 (12%) on SDXL, while GE drops from 92.0 to 88.0 and from 83.0 to 81.0.

Table 4: GenEval and statistical ℓ_1 to $\tau = (\text{brightness} = 0.72, \text{contrast} = 0.26)$ at $N = 64$, $K = N_t = 8$, $k_{\text{proxy}} = 1$, $n = 50$ GenEval prompts per pair (the same 50 on both backbones). Statistical tightening reproduces on both backbones shown (MMDiT, UNet; DiT in Tab. 1 and Tab. 3); GenEval is 2–4 pp lower than teacher argmax on both.

Pair	Argmax GE	Ours GE	Argmax statistical ℓ_1	Ours statistical ℓ_1
SD3.5-Large-Turbo $\rightarrow$ SD3.5-Large	92.0	88.0	0.204	0.120
SDXL-Turbo $\rightarrow$ SDXL	83.0	81.0	0.305	0.267

7 Conclusion

Major diffusion teachers ship with a distilled few-step companion that inference-time search treats only as a faster generator. We show that this companion agrees with the teacher's noise ranking on closed-form image statistics but not on learned semantic rewards, and that the first denoising step seeds this rank while distillation extends it to finer statistics. PROXYSEARCH uses this split for search under a *joint constraint*: the proxy filters candidates by statistical match, and the teacher selects by semantic reward, with no trade-off weight. At the same number of teacher trajectories, it tightens the statistical target while matching teacher argmax on the optimized reward. Its benefit is limited to targets on axes the proxy ranks, and we have not yet trained a downstream model on statistic-balanced data. Rank agreement could serve as a distillation objective and as a metric reported next to FID, and the same measurement can be run for video, audio, and 3D distillation. More broadly, the distilled models released alongside each teacher are useful for inference-time scaling, not only for fast sampling.

Acknowledgments and Disclosure of Funding

This research was supported by the Scaleup TIPS grant (RS-2023-00321784: Development of Novel Generative AI Technology to Generate Domain-Specific Synthetic Data). This work was also supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2026-25598824) and the Korea Agency for Infrastructure Technology Advancement (KAIA) grant funded by the Ministry of Land, Infrastructure and Transport (RS-2026-25530752). The authors declare no competing interests.

References

[1] Lichen Bai, Shitong Shao, Zikai Zhou, Zipeng Qi, Zhiqiang Xu, Haoyi Xiong, and Zeke Xie. Zigzag diffusion sampling: Diffusion models can self-improve via self-reflection. In *International Conference on Learning Representations (ICLR)*, 2025. 3

[2] Arpit Bansal, Hong-Min Chu, Avi Schwarzschild, Soumyadip Sengupta, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Universal guidance for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2023. 3

[3] Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv preprint May 2023. Method: DDPO. 3

[4] Bradley Brown, Jordan Juravsky, Ryan Ehrlich, Ronald Clark, Quoc V. Le, Christopher Ré, and Azalia Mirhoseini. Large language monkeys: Scaling inference compute with repeated sampling. *arXiv preprint arXiv:2407.21787*, 2024. 3

[5] Tianle Cai, Yuhong Li, Zhengyang Geng, Hongwu Peng, Jason D. Lee, Deming Chen, and Tri Dao. Medusa: Simple LLM inference acceleration framework with multiple decoding heads. In *International Conference on Machine Learning (ICML)*, 2024. 3

[6] Charlie Chen, Sebastian Borgeaud, Geoffrey Irving, Jean-Baptiste Lespiau, Laurent Sifre, and John Jumper. Accelerating large language model decoding with speculative sampling. *arXiv preprint arXiv:2302.01318*, 2023. 3

[7] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2829, 2023. 18

[8] Kevin Clark, Paul Vicol, Kevin Swersky, and David J. Fleet. Directly fine-tuning diffusion models on differentiable rewards. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv preprint Sept 2023. Method: DRaFT. 3

[9] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021. 3

[10] Prafulla Dhariwal and Alex Nichol. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 4

[11] Sander Dieleman. Diffusion is spectral autoregression. Blog post, https://sander.ai/2024/09/02/spectral-autoregression.html, 2024. 5

[12] 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 *International Conference on Learning Representations (ICLR)*, 2025. 3

[13] Hanze Dong, Wei Xiong, Deepanshu Goyal, Yihan Zhang, Winnie Chow, Rui Pan, Shizhe Diao, Jipeng Zhang, Kashun Shum, and Tong Zhang. RAFT: Reward rAnked FineTuning for generative foundation model alignment. *Transactions on Machine Learning Research (TMLR)*, 2023. 3

[14] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *International Conference on Machine Learning (ICML)*, 2024. 2, 4, 7

[15] 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 *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 3

[16] Stefan Falkner, Aaron Klein, and Frank Hutter. BOHB: Robust and efficient hyperparameter optimization at scale. In *International Conference on Machine Learning (ICML)*, pages 1437–1446, 2018. 3

[17] Ying Fan, Olivia Watkins, Yuqing Du, Hao Liu, Moonkyung Ryu, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, Kangwook Lee, and Kimin Lee. DPOK: Reinforcement learning for fine-tuning text-to-image diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 3

[18] Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. GenEval: An object-focused framework for evaluating text-to-image alignment. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2023. 4, 7

[19] Xiefan Guo, Jinlin Liu, Miaomiao Cui, Jiankai Li, Hongyu Yang, and Di Huang. InitNO: Boosting text-to-image diffusion models via initial noise optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 3

[20] Anne Harrington, A. Sophia 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 (CVPR)*, 2026. 26

[21] David Hasler and Sabine E. Süsstrunk. Measuring colorfulness in natural images. In *Human Vision and Electronic Imaging VIII*, pages 87–95. SPIE, 2003. 4

[22] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Advances in Neural Information Processing Systems (NIPS)*, 2017. 3

[23] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications. 3

[24] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 3, 4

[25] Jonathan Ho, Chitwan Saharia, William Chan, David J. Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *Journal of Machine Learning Research*, 23 (47):1–33, 2022. 3

[26] Kirthevasan Kandasamy, Gautam Dasarathy, Jeff G. Schneider, and Barnabás Póczos. Multi-fidelity Bayesian optimisation with continuous approximations. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017. 3

[27] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 3, 4

[28] Tero Karras, Miika Aittala, Tuomas Kynkäänniemi, Jaakko Lehtinen, Timo Aila, and Samuli Laine. Guiding a diffusion model with a bad version of itself. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 3

[29] Seungwook Kim and Minsu Cho. Improving text-to-image generation with intrinsic self-confidence rewards. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026. 3, 5

[30] Seungwook Kim, Kejie Li, Xueqing Deng, Yichun Shi, Minsu Cho, and Peng Wang. Enhancing 3D fidelity of text-to-3D using cross-view correspondences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10649–10658, 2024. 3

[31] Seungwook Kim, Yichun Shi, Kejie Li, Minsu Cho, and Peng Wang. Multi-view image prompted multi-view diffusion for improved 3D generation. *arXiv preprint arXiv:2404.17419*, 2024. 3

[32] Sunwoo Kim, Minkyu Kim, and Dongmin Park. Test-time alignment of diffusion models without reward over-optimization. In *International Conference on Learning Representations (ICLR)*, 2025. 3

[33] Seungwook Kim, Seunghyeon Lee, and Minsu Cho. FreeAction: Training-free techniques for enhanced fidelity of trajectory-to-video generation. *arXiv preprint arXiv:2509.24241*, 2025. 3

[34] Seungwook Kim, Yichun Shi, Kejie Li, Minsu Cho, and Peng Wang. RapidMV: Leveraging spatio-angular latent space for efficient and consistent text-to-multi-view synthesis. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1674–1684, 2026. 3

[35] 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 (NeurIPS)*, 2023. 4, 7

[36] Junhong Lee, Seungwook Kim, and Minsu Cho. Harnessing the power of training-free techniques in text-to-2D generation for text-to-3D generation via score distillation sampling. *arXiv preprint arXiv:2505.19868*, 2025. 3

[37] Kimin Lee, Hao Liu, Moonkyung Ryu, Olivia Watkins, Yuqing Du, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, and Shixiang Shane Gu. Aligning text-to-image models using human feedback. *arXiv preprint arXiv:2302.12192*, 2023. 3

[38] Yaniv Leviathan, Matan Kalman, and Yossi Matias. Fast inference from transformers via speculative decoding. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 2023. 3

[39] Lisha Li, Kevin Jamieson, Giulia DeSalvo, Afshin Rostamizadeh, and Ameet Talwalkar. Hyperband: A novel bandit-based approach to hyperparameter optimization. *Journal of Machine Learning Research*, 18 (185):1–52, 2018. 3

[40] Yanyu Li, Xian Liu, Anil Kag, Ju Hu, Yerlan Idelbayev, Dhritiman Sagar, Yanzhi Wang, Sergey Tulyakov, and Jian Ren. TextCraftor: Your text encoder can be image quality controller. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7985–7995, 2024. 3

[41] Zhanhao Liang, Yuhui Yuan, Shuyang Gu, Bohan Chen, Tiankai Hang, Mingxi Cheng, Ji Li, and Liang Zheng. Aesthetic post-training diffusion models from generic preferences with step-by-step preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 3

[42] Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations (ICLR)*, 2024. 3

[43] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations (ICLR)*, 2023. 3

[44] Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-GRPO: Training flow matching models via online RL. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. 5, 20

[45] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations (ICLR)*, 2023. 3

[46] Xingchao Liu, Xiwen Zhang, Jianzhu Ma, Jian Peng, and Qiang Liu. InstaFlow: One step is enough for high-quality diffusion-based text-to-image generation. In *International Conference on Learning Representations (ICLR)*, 2024. 3

[47] Simian Luo, Yiqin Tan, Suraj Patil, Daniel Gu, Patrick von Platen, Apolinário Passos, Longbo Huang, Jian Li, and Hang Zhao. LCM-LoRA: A universal stable-diffusion acceleration module. *arXiv preprint arXiv:2311.05556*, 2023. 3

[48] Nanye Ma, Shangyuan Tong, Haolin Jia, Hexiang Hu, Yu-Chuan Su, Mingda Zhang, Xuan Yang, Yandong Li, Tommi Jaakkola, Xuhui Jia, and Saining Xie. Scaling inference time compute for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 2, 3

[49] Myeongseok Nam, Donghun Yeo, and Seungwook Kim. Training-free metrics for synthetic object detection data: A proxy for detector performance. *arXiv preprint arXiv:2606.19817*, 2026. 2

[50] Hieu Pham, Melody Y. Guan, Barret Zoph, Quoc V. Le, and Jeff Dean. Efficient neural architecture search via parameter sharing. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, 2018. 3

[51] 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. In *International Conference on Learning Representations (ICLR)*, 2024. 2, 4, 7

[52] Mihir Prabhudesai, Anirudh Goyal, Deepak Pathak, and Katerina Fragkiadaki. Aligning text-to-image diffusion models with reward backpropagation. *arXiv preprint arXiv:2310.03739*, 2023. 3

[53] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021. 2, 4, 18

[54] Yuxi Ren, Xin Xia, Yanzuo Lu, Jiacheng Zhang, Jie Wu, Pan Xie, Xing Wang, and Xuefeng Xiao. Hyper-SD: Trajectory segmented consistency model for efficient image synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 3

[55] Severi Rissanen, Markus Heinonen, and Arno Solin. Generative modelling with inverse heat dissipation. In *International Conference on Learning Representations (ICLR)*, 2023. 5

[56] 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)*, 2022. 4

[57] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *International Conference on Learning Representations (ICLR)*, 2022. 3

[58] Axel Sauer, Frederic Boesel, Tim Dockhorn, Andreas Blattmann, Patrick Esser, and Robin Rombach. Fast high-resolution image synthesis with latent adversarial diffusion distillation. In *SIGGRAPH Asia 2024 Conference Papers*, 2024. 2, 3, 4, 7

[59] Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. In *European Conference on Computer Vision (ECCV)*, 2024. 2, 3, 4, 7

[60] Christoph Schuhmann. CLIP+MLP aesthetic score predictor. https://github.com/christophschuhmann/improved-aesthetic-predictor, 2022. 2, 4, 18

[61] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5B: An open large-scale dataset for training next generation image-text models. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2022. 18

[62] David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016. 3

[63] 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 *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, 2025. Informally known as "Feynman–Kac Steering" or "FK Steering". arXiv preprint Jan 2025. 2, 3

[64] Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling parameters for reasoning. In *International Conference on Learning Representations (ICLR)*, 2025. 2

[65] Irwin Sobel. *Camera Models and Machine Perception*. PhD thesis, Stanford University, 1970. 4

[66] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations (ICLR)*, 2021. 3, 4

[67] 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 (ICLR)*, 2021. 4

[68] Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 2023. 3

[69] Mitchell Stern, Noam Shazeer, and Jakob Uszkoreit. Blockwise parallel decoding for deep autoregressive models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. 3

[70] 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 *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, 2025. arXiv preprint May 2024. Method: DNO. 2, 3

[71] 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 (CVPR)*, pages 8228–8238, 2024. 3, 5

[72] Luhuan Wu, Brian L. Trippe, Christian A. Naesseth, David M. Blei, and John P. Cunningham. Practical and asymptotically exact conditional sampling in diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 2, 3

[73] 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, 7

[74] 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. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 2, 4, 18

[75] Yilun Xu, Mingyang Deng, Xiang Cheng, Yonglong Tian, Ziming Liu, and Tommi Jaakkola. Restart sampling for improving generative processes. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 3

[76] 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 (CVPR)*, 2024. 3

[77] Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Frédo Durand, and William T. Freeman. Improved distribution matching distillation for fast image synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 3, 5, 19

[78] Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Frédo Durand, William T. Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 3

[79] 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 (ICCV)*, 2023. 3

[80] 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 (ICCV)*, 2025. arXiv preprint Nov 2024. 2, 3

Table 5: **Experimental protocol.** Teacher–proxy pairs covering DiT, MMDiT, and UNet backbones, plus a DMD2 score-distilled SDXL row used for the rank-correlation extension in App. B. Image pairs share the 11-verifier bank (8 statistics + 3 semantics). $^{\ddagger}k_{\text{proxy}} = 1$ is the operational setting used by PROXYSEARCH; raising the proxy to $k_{\text{proxy}} = 4$ leaves brightness and color-temperature rank unchanged and raises some weaker axes by up to 0.11 on FLUX and SD3.5 (App. B.3), and lowers most axes on SDXL (Tab. 7); the method ablation (Tab. 2) and the K-sweep of Tab. 3 use a 4-step proxy.

Pair	Backbone	Distillation	k_{proxy}	teacher steps	verifiers
FLUX-schnell → FLUX-dev	DiT	LADD	$1^{\ddagger}$	50	8 stat + 3 sem
SD3.5-Large-Turbo → SD3.5-Large	MMDiT	ADD	$1^{\ddagger}$	28	8 stat + 3 sem
SDXL-Turbo → SDXL	UNet	ADD	1	50	8 stat + 3 sem
SDXL+DMD2 → SDXL	UNet	score	1	50	8 stat + 3 sem

A Experimental details

This section lists the models, protocols, compute, verifiers and prompts behind every experiment in the main text.

A.1 Model pairs and sampling

Tab. 5 lists the teacher–proxy pairs. The image pairs span three backbones (DiT, MMDiT, UNet). With the DMD2 score-distilled SDXL proxy, which enters only the rank-correlation study, they cover three distillation objectives (LADD, ADD, score). FLUX-dev runs 50 steps with guidance embedding 3.5 in a single forward pass (`true_cfg_scale` kept at 1.0). SD3.5-Large runs 28 steps and SDXL 50 steps with classifier-free guidance 4.5 and 7.5, using batch-doubled conditional and unconditional passes. The three 1-step proxies run with guidance 0. Images are scored at 1024^2 unless noted. The rank-correlation pool of Sec. 4 and its appendix variants, the high-pass control (App. C.3) and the method-ablation pilot (Tab. 2) run at 512^2.

Experiments run on NVIDIA A100-80GB GPUs with PyTorch 2.10, CUDA 12.8 and cuDNN 9. Each run fits on one GPU. The 553-prompt headline and the joint-constraint ablations were sharded across several GPUs for wall-clock time. Proxy and teacher transformers run in bf16 and the verifier heads (CLIP, ImageReward, aesthetic) in fp32.

A.2 Per-experiment protocol

Tab. 6 gives one row per experiment. Absolute GE is not comparable across tables because the prompt sets differ. On FLUX, teacher argmax reads 81.6% in Tab. 1 (553 GenEval prompts, $N_t = 8$), 93.75% in Tab. 13 (first 200 prompts in file order, only single-object, two-object and counting, $N_t = 8$), 87.0% in Tab. 2 ($N_t = 16$, 512^2) and 81.0–89.0% in Tab. 3 ($N_t = K$). GE also varies strongly by category on the headline run (teacher argmax: single-object 96.3%, position 58.5%). Comparisons are meaningful within a table, where methods share prompts, targets and the noise pool of one run.

Noises come from a helper `noise_from_id(base_seed, prompt_id, candidate_idx)` (base seed 42 for search runs) that returns the same noise for the same triple regardless of batch size. Within a run, proxy and teacher renders of candidate i use the same tensor, so every within-run comparison is paired. The helper's seed depends on Python's string hash, so a pool is reproducible across processes only when `PYTHONHASHSEED` is fixed. Several runs were launched without a fixed hash seed, and their pools cannot be regenerated: the rank-correlation runs (Tab. 7 and its appendix variants), the $N_t = 16$ arm of Tab. 13, the run behind Tab. 2, the $K = 2, 8, 16$ K-sweep runs and the three RL arms. Comparisons across these runs are paired by prompt and target, not by noise. The null controls (App. C) use a saved noise tensor.

A.3 Compute accounting

NFE counts denoising steps only (Sec. 3). It excludes classifier-free-guidance passes, VAE decoding, text encoding and verifier scoring. With a 1-step proxy, $c_{\text{proxy}}/c_{\text{teacher}}$ is $1/50$ on FLUX-1024 and SDXL and $1/28$ on SD3.5. Counting the teacher's guidance passes, it is $1/56$ on SD3.5 and $1/100$

Table 6: **Per-experiment protocol.** Teacher steps as in Tab. 5 unless listed. *Cells* are prompt–target pairs (prompts for rank-correlation studies, which use no τ). N: noises per cell (proxy candidates in search experiments). K: teacher survivors. Teacher argmax uses $N_t = K$ unless noted. In search experiments the proxy retains the top-K by g_{stat} and the teacher argmaxes the listed selection reward. *Template*: 100 synthetic template prompts (App. A.5). *pilot50*: 50 category-stratified GenEval prompts. n/r: not recorded.

Experiment	Pair(s); resolution; prompts; cells	Sampling and selection	Uncertainty
Rank-correlation studies (no τ, no selection; Spearman ρ)			
Rank transfer (Tab. 7)	FLUX, SD3.5, SDXL; 512^2; 100 template; 100/pair	$N=32$; $k_{\text{proxy}}=1$	95% Wald CI; σ; sig%
Score distillation (Tab. 7, DMD2)	SDXL+DMD2; 512^2; 100 template; 100	$N=32$; $k_{\text{proxy}}=1$	σ; sig%
Proxy step count (App. B.3)	FLUX, SD3.5 (SDXL: rank-transfer pool); 512^2; 100 template; 100/pair	$N=32$; $k_{\text{proxy}} \in \{1, 4\}$	none printed
Proxy-semantic top-1 (Tab. 8)	FLUX, SD3.5; 512^2; 100 template; 100/pair	$N \in \{1, 2, 4, 8, 16, 32\}$; $k_{\text{proxy}}=4$ (FLUX), n/r (SD3.5); proxy top-1 by CLIP, ImageReward, aesthetic	none (prompt mean)
High-pass control (App. C.3)	FLUX; 512^2; 50 hand-written; 50	$N=32$; $k_{\text{proxy}}=1$	none
Paired null controls (Tab. 10)	FLUX, SD3.5; 1024^2; 12 GenEval (stratified); 12/pair	$N=32$ (shared); $k_{\text{proxy}}=1$ (4 as reference); teacher 50/28 steps (reference), 1 step (null)	95% noise-bootstrap CI; permutation p
Shared-noise / learned predictors (App. C.2)	FLUX, SD3.5; 1024^2; 12 GenEval (stratified); 12/pair	$N=32$ (shared); $k_{\text{proxy}}=4$	noise-bootstrap CI; permutation p
RL post-training (App. B.4)	SD3.5-Medium proxy $\rightarrow$ SD3.5-Medium (frozen / text / PickScore LoRA); 1024^2; 100 FlowGRPO (50 OCR test + 50 PickScore); 100	$N=32$; $k_{\text{proxy}}=1$; teacher 28 steps	95% CI half-width over prompts
Search experiments (proxy top-K by g_{stat}, teacher argmax of the selection reward)			
Headline (Tab. 1)	FLUX; 1024^2; 553 GenEval $\times$ 2 τ; 1106	$N=64$; $K=8$ (Teacher-seq: $16 \rightarrow 8$, 100 cells measured); $k_{\text{proxy}}=1$; GE / PickScore / HPSv2	paired bootstrap 95% CI; paired t (a)
Method ablation (Tab. 2)	FLUX; 512^2; pilot50 $\times$ 2 τ; 100	$N=64$; $K=4$ ($N_t=16$); $k_{\text{proxy}}=4$; GE	none
K-sweep (Tab. 3)	FLUX; 1024^2 ($K=4$: 512^2); pilot50 $\times$ 2 τ; 100	$N=64$; $K \in \{2, 4, 8, 16\}$ ($N_t=K$; $K=4$: $N_t=16$); $k_{\text{proxy}}=4$ (1 for one $K=8$ row); GE	paired prompt-bootstrap 95% CI
Matched pool (App. D.2)	FLUX; 1024^2; pilot50 $\times$ 2 τ; 100	$N=64$; $K=8$ ($N_t \in \{8, 64\}$); $k_{\text{proxy}}=1$; GE	prompt-cluster bootstrap 95% CI
Cross-backbone (Tab. 4)	SD3.5-Large, SDXL; 1024^2; 50 GenEval (stratified, same for both) $\times$ 1 τ; 50/pair	$N=64$; $K=8$; $k_{\text{proxy}}=1$; GE	none
Teacher-pool scaling (Tab. 13)	FLUX; 1024^2; first 200 GenEval $\times$ 2 τ; 400	$N=64$; $K=8$ ($N_t \in \{8, 16\}$); $k_{\text{proxy}}=1$; GE	none
User study (App. E)	FLUX; 1024^2 (shown ≤ 512 px); GenEval $\times$ 6 looks; 144 / 97 / 180 pairs	$N=64$; $K=8$; $k_{\text{proxy}}=1$; GE; 21 items from the Tab. 2 pilot ($N_t=16$, $K=4$, 4-step proxy)	Wilson; two-way bootstrap

on SDXL. At $N=64$, $K=N_t=8$, teacher argmax and PROXYSEARCH then cost 448 and 512 passes on SD3.5 and 800 and 864 on SDXL. FLUX-dev uses a single guidance-embedded pass, so its counts equal the NFE (400 and 464).

Measured render time per cell grows more than the NFE ratio, because each proxy render carries its own VAE decode and text encoding: about +24% on FLUX (186 vs. 230 s), +28% on SD3.5 and +49% on SDXL. These timings exclude scoring. They are approximate, since the teacher-only time is apportioned from a shared teacher pool rather than timed separately.

A.4 Verifiers and selection rules

Statistics. All eight statistics are computed with numpy and PIL on the decoded image at its native resolution, without normalization: brightness (mean RGB$/255$), contrast (std of grayscale$/255$), saturation (mean HSV saturation$/255$), colorfulness (Hasler–Süsstrunk), sharpness (variance of the 3×3 Laplacian on grayscale), edge density (fraction of pixels whose Sobel magnitude exceeds its mean plus one std), color temperature (log of mean R over mean B) and compressibility (JPEG quality-95 size in KB, so a larger value means a less compressible image).

Semantic verifiers. The rank-transfer bank uses CLIP score [7, 53] (`open_clip` ViT-H-14), ImageReward [74] (`ImageReward-v1.0`) and the aesthetic predictor [60, 61] (LAION v1, CLIP ViT-L/14 backbone with a $768 \to 1024 \to 128 \to 64 \to 16 \to 1$ MLP head). The PickScore- and HPSv2-driven searches (Tab. 1, panels b and c) replace the GE search objective with PickScore (`PickScore_v1`, CLIP-H dual encoder) or HPSv2 (`hpsv2` package, v2.1) on the same $n = 1106$ paired cells.

Statistical objective. g_{stat} sums the absolute deviations over the axes named in τ with unit weights. $\tau_{\text{bright}} = (\text{brightness} = 0.72,\ \text{contrast} = 0.26)$ and $\tau_{\text{dark}} = (0.22,\ 0.20)$ are fixed presets shared by all prompts. The proxy top-K is taken by sorting the continuous proxy ℓ_1. g_{stat} can also encode axis-pushing ($\pm\ell_j$ for the j-th axis) or thresholding ($\mathbf{1}[\ell_j \geq t]$), and the algorithm applies to any g_{stat} for which larger values are preferred. All reported experiments use the target-matching form. Indicator-valued objectives can tie at the top-K cut and then need a tie-breaking rule.

Selection rules. GE-driven selection (teacher argmax and the PROXYSEARCH teacher stage alike) takes the lexicographic maximum of (in-process GenEval pass, detector confidence, smaller teacher-image ℓ_1). The confidence is the maximum detection score over all detected classes. On the stored teacher pools we checked, the first two keys were never tied, so ℓ_1 did not decide a pick. PickScore- and HPSv2-driven selection is a plain argmax. Ties go to the first candidate: the lowest noise index for teacher argmax and the smallest proxy ℓ_1 for PROXYSEARCH. HPSv2 scores tie at the maximum in 18 of the 1106 cells for teacher argmax and in 17 for PROXYSEARCH.

Post-process baselines. The unclamped shift (Tab. 2) edits the teacher-argmax image. It adds $\tau_{\text{brightness}} - \text{brightness}$ to every RGB channel (in $[0, 1]$, then clipped) and applies a PIL contrast enhancement with factor $\tau_{\text{contrast}}/\text{contrast}$ clipped to $[0.2, 3]$. The clamped variant is the pixel blend $0.68\, x_{\text{argmax}} + 0.32\, x_{\text{shift}}$ with one fixed weight for all cells (resulting statistical ℓ_1 0.218). Neither costs NFE.

A.5 Prompt banks

GenEval. The headline uses the full 553-prompt GenEval evaluation set $\times$ 2 targets (1106 paired cells). The same cells feed all three panels of Tab. 1. The 100-cell FLUX pilots (Tabs. 2, 3 and App. D.2) use pilot50, 50 GenEval prompts sampled with stratification by category (seed 42), $\times$ 2 targets. Tab. 4 uses a second stratified 50-prompt sample (41 prompts shared with pilot50) and one target. The Teacher-seq measurement uses a third 50-prompt GenEval sample.

Rank-correlation template. The rank-transfer pool of Sec. 4 and its appendix variants use 100 synthetic template prompts $\times$ 32 noises at 512^2. Five templates are filled from small vocabularies of colors, objects, materials, locations, actions and counts (seed 42), for example "a bird colored 3 with book 2". None is a GenEval prompt, and the same 100 prompts are used for every pair.

FlowGRPO. The RL study (App. B.4) uses 100 FlowGRPO prompts (`github.com/yifan123/flow_grpo`): the first 50 prompts of its OCR test split, for example "A close-up of a sleek smartwatch on a wrist, the screen displaying *Step Goal Achieved* with a celebratory animation, [...]", and 50 PickScore prompts, for example "gold tip pyramid in the night, extremely detailed, rain, stars".

Table 7: **Per-verifier proxy–teacher Spearman ρ.** $n = 100$ template prompts $\times$ 32 noises at 512^2, 1-step proxy vs. full teacher (50 steps for FLUX and SDXL, 28 for SD3.5). ρ: mean per-prompt Spearman. σ: its standard deviation across prompts. sig%: proportion of prompts at $p < 0.05$ on the within-prompt Spearman test (chance is 5%). The 95% Wald CI on each FLUX, SD3.5 and SDXL-Turbo mean has half-width 0.03–0.04. DMD2: the SDXL-Base UNet swapped with the official 1-step score-distilled DMD2 weights [77] (`dmd2_sdxl_1step_unet_fp16.bin`), run at 1 step with guidance scale 0.0. Within each pair every statistical verifier transfers more strongly than every semantic verifier. Across pairs the ranges overlap at the low end (SDXL-Turbo contrast 0.095 vs. FLUX ImageReward 0.116).

	FLUX (LADD)			**SD3.5** (ADD)			**SDXL-Turbo** (ADD)			**SDXL-DMD2** (score)		
Verifier	ρ	σ	sig%	ρ	σ	sig%	ρ	σ	sig%	ρ	σ	sig%
statistics (8 verifiers)												
brightness	0.557	0.227	83	0.804	0.168	99	0.337	0.216	53	0.553	0.131	96
color_temperature	0.486	0.201	81	0.584	0.175	91	0.336	0.195	51	0.306	0.174	42
compressibility	0.421	0.193	67	0.567	0.196	88	0.310	0.181	45	0.239	0.196	30
saturation	0.380	0.211	57	0.538	0.176	82	0.211	0.203	24	0.086	0.195	9
sharpness	0.322	0.198	46	0.455	0.201	72	0.234	0.174	24	0.190	0.172	19
contrast	0.314	0.191	47	0.410	0.194	62	0.095	0.197	11	0.159	0.164	14
colorfulness	0.303	0.198	37	0.263	0.228	39	0.196	0.193	23	0.088	0.170	9
edge_density	0.183	0.197	20	0.210	0.201	27	0.145	0.204	17	0.079	0.199	11
semantics (3 verifiers)												
CLIP	0.095	0.185	12	0.085	0.204	12	0.021	0.170	3	0.022	0.185	6
ImageReward	0.116	0.181	13	0.114	0.204	16	0.016	0.162	4	−0.006	0.163	4
aesthetic	0.118	0.188	13	0.069	0.179	6	0.025	0.175	2	0.013	0.172	4

B Rank transfer: full results

This section gives the numbers behind Fig. 3 (a) and the robustness checks of the statistics/semantics split (Sec. 4 and Sec. 4).

B.1 Per-verifier rank correlation

Tab. 7 reports the rank correlations for all four pairs. Three patterns hold for every pair: statistical ρ is positive on every verifier, semantic ρ is at most 0.118, and brightness is the strongest statistical axis. SDXL-Turbo carries the weakest statistical signal of the ADD/LADD pairs (best sig% 53 for brightness, vs. 99 on SD3.5). It differs from the other proxies in both backbone (UNet) and distillation recipe, so we do not attribute this to either. Its statistics/semantics separation is still the largest ($\approx 16\times$: brightness 0.337 vs. CLIP 0.021).

The DMD2 score-distilled proxy shows the same within-pair split as the LADD and ADD proxies (brightness 0.553 vs. CLIP 0.022), but not the statistical profile of SDXL-Turbo. Statistical rank transfers in rough magnitude order, from brightness (strongest) to edge density (0.079, weakest), which transfers only marginally.

B.2 Proxy-side semantic selection

This subsection tests whether the proxy can select on semantics despite the low semantic ρ of Tab. 7. In a separate run on the same 100 template prompts $\times$ 32 noises, for each $N \in \{1, 2, 4, 8, 16, 32\}$ we take the first N candidates and record the teacher-semantic score of the candidate the proxy ranks first (*proxy*). We compare it with the maximum teacher-semantic score over the same N (*oracle*), the value an unrestricted teacher search would attain. On CLIP and aesthetic the proxy pick stays close to its $N = 1$ value. ImageReward proxy pick rises with N (FLUX $0.239 \to 0.460$, SD3.5 $-0.005 \to 0.338$) but stays well below the oracle, which grows monotonically with N. Semantic gains in this setting come from teacher rollouts.

Table 8: **Proxy-semantic top-1 selection vs. the teacher-semantic oracle on the same N candidates.** Prompt-mean teacher-semantic score. 100 template prompts $\times$ 32 noises at 512^2. Proxy: 4 steps on FLUX, step count not recorded on SD3.5. Teacher: 50 steps (FLUX), 28 steps (SD3.5). At $N=32$ the oracle-minus-proxy gap on FLUX is $+0.058$ CLIP, $+0.077$ aesthetic and $+0.773$ ImageReward. SD3.5 follows the same pattern.

		$N=1$		$N=8$		$N=32$	
Pair	semantic verifier	proxy	oracle	proxy	oracle	proxy	oracle
FLUX	CLIP	0.294	0.294	0.294	0.335	0.298	0.356
FLUX	ImageReward	0.239	0.239	0.318	0.926	0.460	1.233
FLUX	aesthetic	5.377	5.377	5.378	5.420	5.376	5.452
SD3.5	CLIP	0.334	0.334	0.343	0.380	0.346	0.397
SD3.5	ImageReward	-0.005	-0.005	0.252	0.933	0.338	1.194
SD3.5	aesthetic	5.275	5.275	5.284	5.342	5.299	5.373

B.3 Proxy step count

We test whether more proxy steps recover semantic rank transfer. For FLUX-schnell→dev and SD3.5-Large-Turbo→Large we re-run the rank-correlation protocol at $k_{\text{proxy}} \in \{1, 4\}$ (100 template prompts $\times$ 32 noises at 512^2, all 11 verifiers). This run draws its own noises (App. A), so its $k_{\text{proxy}}=1$ values differ slightly from Tab. 7, for example SD3.5 ImageReward 0.136 here vs. 0.114 there. A few values exceed the 0.80 and 0.12 bounds quoted for that pool. The SDXL panel of Fig. 7 uses the pool of Tab. 7.

On FLUX and SD3.5, brightness and color temperature are flat from $k_{\text{proxy}}=1$ to 4 (FLUX brightness $0.559 \to 0.559$, SD3.5 brightness $0.807 \to 0.807$). Several weaker statistical axes gain significantly at $k_{\text{proxy}}=4$ (paired Wilcoxon over prompts): SD3.5 edge density $+0.106$, SD3.5 contrast, colorfulness and sharpness about $+0.04$, and FLUX sharpness $+0.04$. Semantic ρ stays below 0.16 on every verifier (FLUX CLIP $0.078 \to 0.111$, ImageReward $0.099 \to 0.104$, aesthetic $0.113 \to 0.118$; SD3.5 CLIP $0.114 \to 0.138$, ImageReward $0.136 \to 0.159$, aesthetic $0.052 \to 0.126$). The SD3.5 aesthetic ($p=0.002$) and FLUX CLIP ($p=0.046$) gains are nominally significant. The 4-step proxy is not a usable semantic ranker.

On SDXL, a 4-step SDXL-Turbo proxy lowers most statistical axes (brightness $0.337 \to 0.272$, color temperature $0.336 \to 0.270$). Contrast and edge density rise slightly ($0.095 \to 0.118$, $0.145 \to 0.171$), and the semantic verifiers stay near zero (CLIP 0.021, aesthetic 0.006 at $k_{\text{proxy}}=4$).

The 1-step and 4-step proxy renders rank candidates almost identically on brightness ($\rho \geq 0.99$ on FLUX and SD3.5), less closely on the other statistics (down to $\rho=0.42$), and less closely still on semantics ($\rho \in [0.22, 0.70]$). A 4-step render costs about $3\times$ a 1-step render (2.05 vs. 0.69 s on FLUX-1024). The deployed 1-step proxy thus gives up 0.02–0.11 in ρ on several weaker axes for a third of the proxy cost. This is a rank-level comparison. The K-sweep (Tab. 3) reports one 1-step search row on a separate pool.

B.4 Teacher RL post-training

This subsection gives the protocol and per-cell numbers behind the RL result of Sec. 4. The proxy is held fixed across conditions. It is the third-party distilled checkpoint `tensorart/stable-diffusion-3.5-medium-turbo`, run at 1 step with guidance 1.5. Its model card recommends 8 steps and does not state the distillation method. The teacher is SD3.5-Medium (28 steps, guidance 4.5) in three conditions: frozen, with the FlowGRPO text-rendering LoRA (`jieliu/SD3.5M-FlowGRPO-Text`), and with the FlowGRPO PickScore-reward LoRA (`jieliu/SD3.5M-FlowGRPO-PickScore`) [44]. Each condition uses $n=100$ FlowGRPO prompts $\times$ 32 noises at 1024^2 (App. A.5). Of the 50 PickScore prompts, 48 are in the PickScore LoRA's training split, so that arm is evaluated largely on prompts its teacher was tuned on. Each arm draws its own noises, so arms are paired by prompt only. Fig. 4 shows the per-verifier ρ under each teacher (top) and percentile traversals along three statistical axes (bottom). Tab. 9 reports per-cell 95% CI half-widths, $1.96\,\sigma/\sqrt{n}$ across the 100 prompts with σ the sample standard deviation.

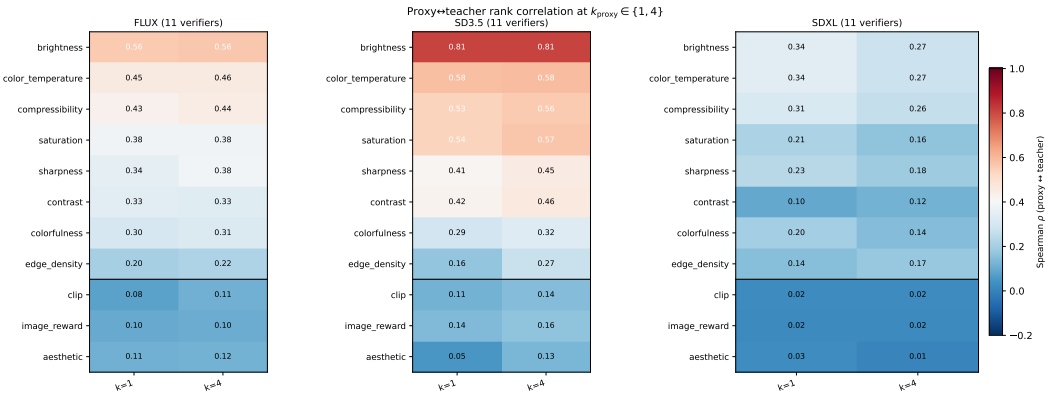

Figure 7: **A 4-step proxy does not recover semantic rank.** Per-verifier Spearman ρ between the proxy at $k_{\text{proxy}} \in \{1, 4\}$ and the full teacher, 100 template prompts $\times$ 32 noises at 512^2. Panels: FLUX-schnell→dev, SD3.5-Large-Turbo→Large and SDXL-Turbo→SDXL. Rows: eight statistics (top) and three semantic verifiers (bottom). The FLUX and SD3.5 panels come from a separate run with its own noises. The SDXL panel uses the pool of Tab. 7. Semantic ρ stays below 0.16 at $k_{\text{proxy}} = 4$ on every verifier and pair.

Table 9: **Proxy→teacher Spearman ρ under three teacher conditions** ($\rho \pm 95\%$ CI half-width). The text-rendering LoRA stays within about 0.06 of the frozen teacher on every statistical axis. The PickScore-reward LoRA lowers statistical ρ by roughly 0.14–0.23 on every axis except edge density ($0.133 \to 0.068$), and brightness (0.684) still exceeds the best semantic verifier (CLIP, 0.045) by about $15\times$.

Verifier	Frozen teacher	Text-rendering LoRA	PickScore-reward LoRA
statistics (8 verifiers)			
brightness	$+0.823 \pm 0.039$	$+0.829 \pm 0.035$	$+0.684 \pm 0.052$
color_temperature	$+0.628 \pm 0.050$	$+0.615 \pm 0.049$	$+0.460 \pm 0.052$
compressibility	$+0.570 \pm 0.044$	$+0.509 \pm 0.050$	$+0.347 \pm 0.048$
saturation	$+0.491 \pm 0.048$	$+0.458 \pm 0.048$	$+0.271 \pm 0.050$
sharpness	$+0.400 \pm 0.051$	$+0.350 \pm 0.052$	$+0.207 \pm 0.044$
contrast	$+0.496 \pm 0.056$	$+0.508 \pm 0.047$	$+0.273 \pm 0.054$
colorfulness	$+0.432 \pm 0.046$	$+0.396 \pm 0.047$	$+0.203 \pm 0.050$
edge density	$+0.133 \pm 0.051$	$+0.113 \pm 0.055$	$+0.068 \pm 0.040$
semantics (3 verifiers)			
CLIP score	-0.010 ± 0.036	$+0.025 \pm 0.037$	$+0.045 \pm 0.041$
ImageReward	$+0.001 \pm 0.041$	-0.009 ± 0.039	-0.003 ± 0.033
aesthetic	$+0.072 \pm 0.038$	$+0.043 \pm 0.042$	$+0.020 \pm 0.039$

C Where statistical rank comes from: null controls

This section separates what the raw noise, one teacher step and distillation each contribute to the statistical rank (Fig. 3 (b), Sec. 4). All controls use one saved pool of 32 noise tensors and 12 GenEval prompts stratified by category, at 1024^2. Noise j is the identical z_T for every model and condition. Because all prompts share the 32 noises, the noise is the independent unit: CIs resample noise indices (the same resample for every prompt), and two-sided permutation tests permute them ($B = 2000$ each).

C.1 Paired null controls

Tab. 10 compares three 1-step conditions against the full teacher: the global mean of the raw latent, the undistilled teacher truncated to one step (teacher guidance), and the distilled proxy. The 4-step proxy is listed for reference. Re-rendering the full teacher for two prompts (64 images per backbone) matches the stored teacher statistics at Spearman 0.96–1.00 on every statistic.

Table 10: **Paired null controls** (Fig. 3 (b)). Mean per-prompt Spearman ρ against the full teacher (50 FLUX / 28 SD3.5 steps), 12 GenEval prompts $\times$ 32 shared noises at 1024^2. *Raw*: global mean of z_T. *T-1*: undistilled teacher, 1 step (teacher guidance). *P-1* / *P-4*: distilled proxy, 1 / 4 steps. Bold: 95% CI from a bootstrap over noises excludes 0 ($B = 2000$, not corrected for the eight statistics).

	FLUX				SD3.5			
Statistic	Raw	T-1	P-1	P-4	Raw	T-1	P-1	P-4
brightness	0.22	**0.68**	**0.53**	**0.53**	−0.03	**0.91**	**0.85**	**0.86**
color temp.	**−0.25**	**0.27**	**0.50**	**0.50**	−0.08	**0.50**	**0.54**	**0.60**
compressibility	−0.21	**0.23**	**0.46**	**0.46**	0.21	**0.46**	**0.50**	**0.51**
saturation	**−0.20**	**0.39**	**0.33**	**0.32**	0.04	**0.53**	**0.47**	**0.51**
sharpness	−0.12	**0.20**	**0.42**	**0.38**	0.15	**0.36**	**0.40**	**0.40**
contrast	−0.08	**0.35**	**0.32**	**0.30**	−0.09	**0.46**	**0.38**	**0.40**
colorfulness	−0.08	**0.15**	**0.20**	**0.20**	0.01	**0.29**	**0.28**	**0.33**
edge density	−0.03	0.08	**0.23**	**0.14**	0.04	0.06	**0.24**	**0.30**

The raw-latent mean shows no positive transfer. On FLUX it is weakly anti-correlated with color temperature and saturation. The 1-step teacher ranks the coarse statistics as well as or better than the 1-step proxy (brightness, contrast and saturation on both backbones). Distillation contributes on the finer statistics. On FLUX the proxy roughly doubles the 1-step teacher's ρ on color temperature, compressibility and sharpness, and it is significant on edge density where the teacher is not. On SD3.5 the gain is confined to edge density. The distilled proxy is the only 1-step condition significant on all eight statistics, and its 1-step and 4-step variants agree closely (App. B.3).

The CIs are not corrected for testing eight statistics. The 1-step teacher is also not a drop-in proxy: with classifier-free guidance it needs two forward passes per candidate on SD3.5, and it does not rank the finer statistics. Whether it could replace the proxy on coarse axes when no distilled sibling exists is untested.

C.2 Shared noise across models and learned noise predictors

These controls ask whether the screened statistic is carried by the noise itself. They pass the same pool through both the FLUX and the SD3.5 pairs, with proxies at $k_{\text{proxy}} = 4$ and the eight statistics only.

Cross-architecture agreement. Within each pair, the 4-step proxy ranks the teacher's statistics on these noises ($p < 0.01$ on all eight statistics, column P-4 of Tab. 10). Across architectures, the FLUX and SD3.5 *teachers* rendering the identical z_T agree only weakly: mean $|\rho| = 0.07$ over the eight statistics and $|\rho| \leq 0.07$ on seven of them. They share a weak brightness component ($\rho = 0.26$, 95% noise-bootstrap CI $[0.08, 0.40]$, uncorrected for the eight statistics tested). The two proxies share a similar brightness component ($\rho = 0.35$). The statistic a model produces from a given noise is therefore largely, though not entirely, specific to that model.

Learned noise predictors. To test whether a learned function of the raw noise could replace the proxy, we fit a ridge regressor and a small MLP for each prompt, statistic and model. Their inputs are 38 hand-crafted features of z_T: per-channel means and standard deviations, global moments and spectral summaries, with no prompt conditioning. We use 5-fold cross-validation over the 32 noises with teacher statistics as labels. Within a model, the ridge recovers part of the teacher's brightness rank on held-out noises (FLUX 0.39, SD3.5 0.55). Mean $|\rho|$ over the eight statistics is 0.18 on FLUX and 0.15 on SD3.5 for the ridge, and 0.20 and 0.17 for the MLP. Raw-noise features thus carry some signal once a predictor is trained on a model's own teacher outputs. Across models, a predictor fit on one teacher and evaluated on held-out noises of the other transfers brightness at only 0.15–0.18, not distinguishable from zero at the noise level, and every other statistic stays near zero. Such a predictor needs teacher renders of the target model as labels. The distilled proxy does not.

Scope. These controls use a 4-step proxy at 1024^2 on 12 prompts, unlike the 1-step, 512^2, 100-prompt rank-transfer study. The learned-predictor result is exploratory (fixed hyperparameters, no model selection). We do not compare a learned predictor with the proxy as a screening rule inside PROXYSEARCH.

C.3 High-pass control

An alternative account of the split is that the statistical verifiers read a fixed low-frequency band that proxy and teacher share. If a verifier depended only on frequencies below a cutoff, high-pass filtering both images above that cutoff would send the verifier to a constant and its proxy–teacher rank correlation to zero. We render a separate FLUX pool (50 hand-written prompts $\times$ 32 noises at 512^2, 1-step FLUX-schnell proxy and 50-step FLUX-dev teacher on the same noises). Before scoring, both images pass through an ideal radial high-pass filter computed via FFT in image space. It zeroes the coefficients within $1/8$ of the corner radius (0.177 of the per-axis Nyquist frequency, 2.5% of coefficients including DC), and the signed residual is clipped to $[0, 255]$ and quantized to 8 bits. The high-frequency statistics still transfer after filtering (compressibility 0.42, sharpness 0.29, edge density 0.20), as expected for statistics that read high-frequency content by construction. The low-band statistics cannot be read the same way. Removing DC makes the residual zero-mean, and after clipping about 63% of output values are zero. Filtered "brightness" then tracks the amplitude of the high-frequency residual rather than image brightness (across 120 FLUX renders, Spearman 0.97 with residual RMS and -0.08 with the original brightness). The same rectification affects color temperature, colorfulness, saturation and contrast, and the semantic verifiers see out-of-distribution residual images. The control therefore neither supports nor rules out a low-band explanation for the low-band statistics.

D Search results: uncertainty, controls and other backbones

This section gives the uncertainty on Tab. 1, shows that the statistical gain comes from ranking rather than pool size, and reports the cross-backbone pilot (Sec. 6).

D.1 Paired-bootstrap CIs for the headline result

Tab. 11 reports the ProxySearch$-$teacher-argmax delta for every metric of Tab. 1. The statistical ℓ_1 delta excludes zero by a wide margin in all three search objectives (CI upper bounds ≤ -0.056). In panel (a), the paired t statistic is -19.3 and the Wilcoxon signed-rank z is -18.5. The CI on the optimized reward includes zero in every panel, with lower bounds of -0.5 pp GE, -0.03 PickScore and -0.001 HPSv2. In panel (c) the HPSv2 upper bound rounds to 0.000. The held-out GE CIs of panels (b) and (c) also include zero.

Two held-out deltas have nominal CIs that exclude zero: ImageReward in panel (a) and aesthetic in panel (c). Neither verifier is optimized by any stage. The CIs are not corrected for the metrics and panels tested. The aesthetic delta ($+0.004$ on a mean near 5.35) is negligible. The ImageReward delta also holds under a prompt-cluster bootstrap, but the median per-cell difference is 0 (signed-rank $p = 0.088$), so it shows no semantic cost rather than a clear gain.

Table 11: **Paired-bootstrap 95% CIs on ProxySearch$-$teacher-argmax deltas for Tab. 1.** 10,000 resamples, $n = 1106$ paired cells per panel. Bold: the CI excludes zero.

	(a) GE-driven		(b) PickScore-driven		(c) HPSv2-driven	
Metric	Δ	95% CI	Δ	95% CI	Δ	95% CI
stat. ℓ_1	$\mathbf{-0.070}$	$[\mathbf{-0.077}, \mathbf{-0.063}]$	$\mathbf{-0.065}$	$[\mathbf{-0.072}, \mathbf{-0.059}]$	$\mathbf{-0.062}$	$[\mathbf{-0.068}, \mathbf{-0.056}]$
GE (pp)	$+0.9$	$[-0.5, +2.3]$	$+0.3$	$[-1.2, +1.7]$	-0.32	$[-1.81, +1.18]$
PickScore	$+0.02$	$[-0.03, +0.06]$	-0.01	$[-0.03, +0.02]$	-0.02	$[-0.06, +0.01]$
HPSv2	$+0.000$	$[-0.001, +0.001]$	$+0.000$	$[-0.001, +0.001]$	-0.000	$[-0.001, +0.000]$
CLIP	$+0.001$	$[-0.001, +0.003]$	-0.001	$[-0.002, +0.001]$	$+0.000$	$[-0.001, +0.002]$
ImageReward	$\mathbf{+0.06}$	$[\mathbf{+0.02}, \mathbf{+0.10}]$	-0.01	$[-0.03, +0.02]$	-0.01	$[-0.03, +0.02]$
aesthetic	$+0.000$	$[-0.003, +0.003]$	$+0.001$	$[-0.001, +0.004]$	$\mathbf{+0.004}$	$[\mathbf{+0.002}, \mathbf{+0.006}]$

D.2 More teacher candidates do not tighten the statistic

Matched candidate pool. The headline fixes the number of teacher trajectories, so ProxySearch screens 64 candidates while teacher argmax renders 8. To separate pool size from ranking on the statistic, we give the teacher-side baselines the same 64-candidate pool. All rows of Tab. 12 use the

same 100 cells (pilot50 $\times$ $\{\tau_{\text{bright}}, \tau_{\text{dark}}\}$) and one 64-noise pool at FLUX-1024. Teacher argmax at $N_t = 8$ uses the first 8 noises of that pool, and PROXYSEARCH uses $N = 64$, $K = 8$, $k_{\text{proxy}} = 1$.

Enlarging the teacher pool from 8 to 64 does not tighten the statistic under semantic-only selection ($0.304 \to 0.306$, paired difference $+0.002$, 95% CI $[-0.006, +0.010]$). The statistical gain comes from ranking on the statistic, not from pool size. At equal teacher trajectories on the same pool, PROXYSEARCH tightens ℓ_1 by 25% relative to teacher argmax (paired difference -0.075, 95% CI $[-0.098, -0.053]$). With the full 64-candidate teacher pool, the composite reaches a tighter statistic than PROXYSEARCH (0.143 vs. 0.230) at $6.9\times$ the total NFE ($8\times$ the teacher NFE). PROXYSEARCH recovers about half of the composite's ℓ_1 reduction at the teacher cost of $N_t = 8$.

Table 12: **Matched candidate pool** (FLUX-1024, 100 cells, one shared 64-noise pool). GE differences among the $N_t = 64$ rows and PROXYSEARCH are within noise. On these cells every $\lambda \in \{0.1, 0.5, 1, 2\}$ selects the same composite image. Selection used the in-process GenEval reimplementation, and GE re-scores the selected images with the official pipeline.

Method	teacher candidates	total NFE (teacher)	stat. ℓ_1	GE (%)
Teacher argmax	8	400 (400)	0.304	88.0
Teacher argmax	64	3200 (3200)	0.306	89.0
Composite reward ($\lambda = 0.1$)	64	3200 (3200)	**0.143**	89.0
PROXYSEARCH ($N = 64$, $K = 8$)	8	**464** (400)	0.230	89.0

Teacher-pool scaling. We also rerun teacher argmax at $N_t = 16$ on the first 200 GenEval prompts in file order (80 single-object, 99 two-object and 21 counting prompts, $n = 400$ cells). The $N_t = 8$ and PROXYSEARCH rows of Tab. 13 are the headline run restricted to these cells. The $N_t = 16$ arm is a separate run with an independent noise draw, so it is paired with the other rows by prompt and target but not by noise. Doubling the teacher pool raises GE by 1.25 pp ($93.75\% \to 95.0\%$) but leaves statistical ℓ_1 unchanged ($0.298 \to 0.296$) at twice the teacher NFE ($400 \to 800$). PROXYSEARCH at $K = 8$ matches teacher argmax at $N_t = 8$ on GE (93.75%) and reaches ℓ_1 0.226 at 464 NFE, 42% fewer than the doubled baseline.

Table 13: **Teacher-pool scaling** (FLUX-1024, first 200 GenEval prompts $\times$ 2 targets, $n = 400$ cells). The $N_t = 16$ row uses an independent noise draw. GE is higher than in Tab. 1 (a) because these prompts contain only the single-object, two-object and counting categories.

Method	GE acc. (%)	stat. ℓ_1	total NFE
Teacher argmax (N_t=8)	93.75	0.298	400
Teacher argmax (N_t=16)	**95.00**	0.296	800
PROXYSEARCH (K=8, **1-step proxy**)	93.75	**0.226**	**464**

E User study

This section gives the design, results and ethics disclosures of the forced-choice study behind the human judgments of Sec. 6.3. The study has two main arms (A and B) and a separate no-look pilot (C), with no learned reward model in the loop. Annotators were unpaid lab volunteers. Tab. 14 summarizes the three arms.

Stimuli. All images are FLUX.1-dev teacher renders at 1024^2, shown side by side after downscaling to at most 512 px (JPEG). Left/right placement was randomized per pair. The PROXYSEARCH image is the GenEval argmax over teacher renders of the top 8 of 64 one-step FLUX-schnell candidates by statistical ℓ_1 (464 NFE). The teacher-argmax image is the GenEval argmax over 8 teacher candidates (400 NFE). A few items come from the 512^2 pilot of Tab. 2, where teacher compute is not matched (Arm A 4/144, Arm B 2/97, Arm C 15/180). Prompts are GenEval prompts under six requested looks with fixed numeric targets: "bright, high-contrast" (τ_{bright}), "dark, moody" (τ_{dark}), warm and cool (color temperature 0.5 and -0.1), and high and low sharpness (Laplacian variance 1000 and 200). The warm, cool and sharpness items come from separate 50-prompt runs at the headline operating point that appear in no other table.

Table 14: **User-study arms.** Rate: share of judgments choosing the PROXYSEARCH image. Two-way: annotator×item bootstrap, which accounts for repeated judgments. Arm C has one judgment per pair.

Arm	Question	Annotators	Pairs	Judgments	Rate (%)	Wilson 95% CI	Two-way 95% CI
A (look named)	which image better fulfils the request	20	144	600	71.7	[67.9, 75.1]	[63.7, 79.1]
B (look named)	which image looks like the genuine, unedited photograph	20 (same)	97	400	84.2	[80.4, 87.5]	[76.2, 91.1]
C (no look)	which image is preferred overall as a photograph	5	180	180	48.9 (88/180)	[41.7, 56.1]	n/a

Arm A (request fulfillment). Annotators read the prompt and the named look (e.g., *A user asked for "{prompt}", with this look: dark and moody, with low-key lighting. Which image better fulfils the request?*) and judged the look and overall image quality together. We kept only pairs in which, on the displayed images, the PROXYSEARCH image is shifted further in the requested direction than the teacher-argmax image by a fixed margin (brightness ≥ 0.04, color temperature ≥ 0.06, or a sharpness ratio ≥ 1.25). All 144 pairs (105 prompts) satisfy this, so results are conditional on PROXYSEARCH having shifted the statistic. Neither image is required to reach τ. Both images pass the in-process GenEval check equally often (117 of 144 each). Twenty annotators gave 600 judgments (2–5 per pair). The arm measures joint content-plus-look fulfillment, not semantic superiority. Rates differ strongly by look: dark 88.4%, warm 85.7%, cool 82.0%, high sharpness 70.1%, bright/high-contrast 48.1% and low sharpness 48.2%. For high sharpness, the two-way bootstrap interval reaches down to 49.1%, and a two-way cluster-robust interval is $[54.0, 86.3]$. We have no explanation for the two looks at chance. In both, the two images rarely both reach the target.

Arm B (perceived naturalness). Annotators were told that both images had the described look, that one was generated that way and the other was an ordinary photo edited afterwards, and were asked which looks like the genuine, unedited photograph. In fact both images are generated. The comparator is the teacher-argmax image post-processed toward the target: the unclamped brightness/contrast shift of Tab. 2 for bright and dark looks, a red/blue gain for warm and cool looks, and an unsharp mask or Gaussian blur for the sharpness looks. The last three operations appear nowhere else in the paper. Pairs were only required to be comparable on the targeted statistic, within a fixed tolerance. On tone and color the post-processed image is usually the closer of the two to τ. The same twenty annotators, with Arm A and Arm B items interleaved in individually shuffled orders, gave 400 judgments on 97 pairs. Because both images are generated, the arm measures perceived photographic naturalness under a "one image was edited" framing, not realism against real photographs. Participants were not debriefed about this framing afterwards. The effect is not significant for low sharpness (5 pairs).

Arm C (no-look pilot). In a separate pilot, five annotators each judged 36 distinct pairs (180 judgments, one per pair). They were asked which image they prefer overall as a photograph. No look was named, no prompt was shown and no manipulation check was applied. Arm C raters saw the same intro page as Arms A and B, which mentions requested looks and edited images, so the pilot is not fully unprimed. The pilot is dominated by bright and dark items (147 of 180 pairs). Rates differed across the five disjoint annotator blocks ($p = 0.007$), and with disjoint item sets this cannot be attributed to annotators rather than items. In this small pilot, neither method is preferred overall.

Procedure and exclusions. Studies ran on a self-hosted web page. Answers were given by click or arrow key, with no tie option and no time limit. The intro page stated that some questions ask which image better fulfills a requested look, others which image is the genuine, unedited photograph, and that there are no right or wrong answers. All three arms used this intro page. Median response time was about 2 s per pair. The manipulation check was introduced after an initial batch of 168 responses had been collected on an earlier stimulus set. That set was regenerated, and those responses are archived and excluded. Excluding the 15 final judgments on items that the same annotator ID had seen in the earlier batch leaves the Arm A and B rates essentially unchanged (71.65% and 84.09%). No final response was dropped.

F Limitations

When PROXYSEARCH does not help. PROXYSEARCH needs a statistical target on axes the proxy ranks. We expect little benefit (i) on weakly transferring axes such as edge density, or after teacher reward post-training that attenuates transfer; (ii) without a statistical target, since the proxy is not a usable semantic ranker (Tab. 8); (iii) for targets that no noise in the pool can reach, since the filter

guarantees no tolerance; (iv) when proxy renders collapse across noises [20]; and (v) when N is close to K, where the method reduces to teacher argmax.

Scope of the evidence. "Statistics" and "semantics" are descriptive labels for the eight statistics and three semantic scorers we measure, not a typology, and a new verifier must be measured before it is assigned to either family. The proxy is a cheap ranker, not a faithful one: its rank agreement peaks at 0.80 (Tab. 7), which limits how tightly screening can hit the target. Beyond the FLUX headline, our comparisons are single-run pilots, mostly without CIs, and matched teacher trajectories are not matched total compute (App. A.3). The null controls do not pin down a single causal mechanism. Finally, the workflows that motivate the task are motivation only: we have not trained a downstream model on statistic-balanced data.

G Broader impact

PROXYSEARCH is a training-free inference-time selection method that routes statistical screening to a distilled proxy and semantic reward optimization to the teacher. It is orthogonal to the underlying generative model and inherits the standard text-to-image diffusion risk profile. On the positive side, it exposes a numeric statistical target, supplied separately from the prompt, as an explicit user control without retraining the released generator. Two considerations are specific to its deployment. First, inference-time selection concentrates whatever bias is encoded in the semantic verifier. Compositional detectors and learned preference models inherit their training-data priors, so the choice of detector is an explicit deployment decision, not a transparent quality signal. Second, the proxy filter could reduce the cost of adversarial selection against noise-domain watermarks and provenance signals whose detection statistic is ranked faithfully by the proxy. We do not evaluate such schemes here. They should be evaluated under proxy-screened search, not only against direct-sampling baselines.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and Sec. 1 state (i) the empirical statistics/semantics proxy–teacher rank-agreement asymmetry and (ii) PROXYSEARCH, the joint-constraint search procedure that exploits it to hit a user-supplied statistical target while remaining competitive on the optimized semantic reward, together with the headline FLUX-1024 results. Empirical claims are backed by Sec. 4 (transfer asymmetry) and Sec. 6.2 (headline result); the compute accounting (matched teacher trajectories, $N_t = K = 8$, with the proxy adding 64 NFE on top of the teacher's 400) is given in Sec. 6.1, and per-backbone variation (smaller statistical tightening on SDXL, 12% vs. 41% on SD3.5, and GE 2–4 pp below teacher argmax on both off-FLUX pilots) is reported in Tab. 4 and App. F.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: App. F discusses statistics/semantics as descriptive labels rather than an a-priori typology, the imperfect proxy–teacher alignment (Spearman ρ peaking at 0.80 on the three main image pairs, Tab. 7) that limits how tightly a proxy-screened search can hit the statistical target, and the regimes in which PROXYSEARCH offers little benefit (App. F): weakly transferring axes, no statistical target, infeasible targets, proxy collapse, and small N. The tested distillation regimes (LADD, ADD, score) are enumerated in Sec. 4, and the weaker SDXL result is reported in Tab. 4.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.

- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper is empirical; no theorems are proven. Sec. 4 states an empirical asymmetry, not a theoretical result.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: App. A documents hardware, per-pair sampling settings, the noise-seeding helper and which runs' noise pools cannot be regenerated (pairing holds within each run), pool sizes, prompt sources, statistic definitions, GenEval selection and tie-breaking rules, and the post-process baselines; Tab. 6 gives the per-experiment protocol. Sec. 6.2 gives the joint-search hyperparameters ($N = 64$, $K = 8$, $k_{\text{proxy}} = 1$).

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example

(a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
(b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: Code is not released with this submission. All datasets used (GenEval, FlowGRPO OCR and PickScore prompt sets, LAION-5B/Aesthetics) and all models (FLUX-dev/schnell, SD3.5-Large/Turbo, SD3.5-Medium, the third-party `tensorart/stable-diffusion-3.5-medium-turbo` proxy, the FlowGRPO text-rendering and PickScore LoRAs, SDXL/SDXL-Turbo, DMD2-SDXL, CLIP, ImageReward, PickScore, HPSv2, Mask2Former) are publicly available, and App. A documents the sampling, seeding, and prompt protocol; we plan to release the code post-acceptance.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: App. A reports per-pair sampling settings (denoising steps, guidance scales), seeds, pool sizes, and prompt sources; Sec. 6.2 and Sec. 6.1 cover the joint-search hyperparameters and the compute accounting (matched teacher trajectories, $N_t = K$, with the proxy's additional $N \cdot k_{\text{proxy}}$ NFE reported separately); App. A tabulates the per-experiment protocol (Tab. 6).

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Tab. 7 (App. B) reports the per-prompt mean ρ with its prompt-level σ and significance fraction, and the 95% Wald CI half-width for the three ADD/LADD pairs; the headline statistical tightening is reported with a paired t statistic ($t=-19.3$, -0.070 ± 0.004 SE; Wilcoxon signed-rank $z=-18.5$), and Tab. 11 gives paired-bootstrap 95% CIs (10,000 resamples over the $n=1106$ paired cells) on the PROXYSEARCH−teacher-argmax delta for every metric of Tab. 1; App. B.4 tabulates per-cell 95% CI half-widths ($1.96\,\sigma/\sqrt{n}$ over prompts) for the RL post-training condition; the paired null and shared-noise controls report noise-bootstrap CIs and permutation p-values, treating the shared noise as the independent unit; the matched-pool control (App. D.2) uses a prompt-cluster bootstrap; and the user study (App. E) reports Wilson 95% CIs plus a two-way annotator×item bootstrap for the clustered judgments. Tab. 3 reports paired prompt-bootstrap 95% CIs on ΔGE; the other smaller-scale controls (Tabs. 2, 4, 13 and the high-pass control of App. C.3) report point estimates without uncertainty.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: App. A reports the compute platform (NVIDIA A100-80GB), per-pair sampling settings, and the pool sizes that determine per-experiment cost. All experiments fit on a single A100-80GB; the 553-prompt headline scaleup and the joint-constraint ablations are sharded across multiple A100s for wall-clock convenience. Sec. 6.1 reports measured per-cell render times (e.g., FLUX-1024: about 186 s for teacher argmax and 230 s for PROXYSEARCH, excluding scoring).

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: The work uses publicly released models and benchmarks and is not deployed beyond research evaluation. It includes a small forced-choice user study on generated images (App. E); no personal data beyond an annotator number and timestamps was recorded. One arm used a cover story (annotators were told one image was an edited photograph, while both were generated), which the appendix discloses. We have read and conform to the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The broader-impact statement (final appendix section, App. G) notes the positive use (an explicit statistical-appearance control without retraining the generator) and two risks specific to PROXYSEARCH's deployment as a training-free inference-time procedure operating on existing public diffusion models: concentration of semantic-verifier bias, and potentially cheaper adversarial selection against noise-domain watermarks and provenance signals.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper releases no new model checkpoints, datasets, or generation systems. PROXYSEARCH is a training-free inference-time procedure operating on existing public diffusion models.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All used models (FLUX-dev/schnell, SD3.5-Large/Turbo, SD3.5-Medium, the third-party `tensorart/stable-diffusion-3.5-medium-turbo` proxy, the FlowGRPO text-rendering and PickScore LoRAs, SDXL/SDXL-Turbo, DMD2-SDXL, CLIP, OpenCLIP, ImageReward, PickScore, HPSv2, and the mmdet and HF Mask2Former detectors used for GenEval) and benchmarks (GenEval, FlowGRPO OCR and PickScore prompt sets, LAION-5B/Aesthetics) are cited and used under their respective public licenses; Tab. 5 and the references list provide attribution.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: No new datasets, model checkpoints, or benchmarks are released with this submission. The reported method (PROXYSEARCH) is an inference-time procedure with no new asset artifact.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: App. E describes the forced-choice user study (twenty annotators in Arms A and B, five in the Arm C pilot; 1180 judgments) and gives the question wording of each arm, the content of the intro page shown in all three arms, the stimulus construction, the manipulation check, and the cover story used in Arm B. Annotators were unpaid lab volunteers; we did not retain a screenshot of the study page, so the appendix gives the question wording instead.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [No]

    Justification: Participants viewed pairs of generated images of GenEval prompts and clicked a choice; no personal data beyond an annotator number and timestamps was recorded, and we consider the risk minimal. Arm B told participants that one image had been edited, while both were generated (App. E). Annotators were unpaid lab volunteers and were not debriefed about the Arm B framing. The study was not submitted for IRB (or equivalent) review.

    Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs are not used as components of the method. PROXYSEARCH operates on diffusion proxy/teacher pairs and closed-form image statistics; no LLM is in the method pipeline.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.