BRIDGING CANDIDATE SELECTION AND MEAN-SHIFT GUIDANCE FOR DIFFUSION REWARD ALIGNMENT VIA LINEAR NOISE RANKING

Anonymous authors
Paper under double-blind review

ABSTRACT

Training-free reward alignment of diffusion models follows two routes. Candidate selection draws K candidate next states at each denoising step and keeps the one with the highest estimated reward, spending K neural evaluations per step. Mean-shift guidance instead moves the transition mean along a single reward gradient. We bridge these two regimes by utilizing a linear functional towards a selected direction as a proxy reward model to rank the Gaussian noises at each steps and select according to those scores. Scoring each noises only requires one inner product, bypassing the K neural network forward costs. Our analysis locates this rule precisely between the two routes. We propose a single representation that unifies the candidate selection method and the mean shift method, and view their difference only by the distribution of the strength scalar towards the reward optimizing direction. This enables us to derive exact leading order predictions of the relative reward optimization gaps between candidate selection method and mean shift method, which are validated through measurements. We further provide practical methods to compute the reward optimizing direction: a corrected direction whose benefit it predicts, and a least-squares slope, fitted from the same reward queries, that extends the rule to non-differentiable rewards. Our experiments show that our linear noise ranking matches or exceeds per-candidate selection at equal or lower neural cost and ties with strength-matched mean shift method as predicted.

1 INTRODUCTION

Pretrained diffusion models can be steered toward new objectives during generation, without fine-tuning for each task (Ho et al., 2020, Janner et al., 2022, Yu et al., 2023). How reward feedback enters the reverse process matters. Mean-shift guidance moves each transition along a reward-derived direction (Yu et al., 2023, Chung et al., 2023). It avoids scoring many alternatives, but a nonzero Gaussian mean shift has no finite worst-case density-ratio bound relative to the base transition. Noise-selection methods such as SVDD (Li et al., 2025) instead choose among draws from that transition. This finite choice admits such a bound, but evaluating every candidate can repeat expensive neural work. Terminal best-of-N delays the choice until entire trajectories have been generated (Beirami et al., 2025). We ask whether stepwise selection can keep its finite-choice advantage without paying for a neural score on every candidate.

LNR reuses one local reward direction to rank many candidate noises. Increasing the candidate count then changes inexpensive sampling and ranking work, not the number of neural evaluations needed to obtain the direction. The use of finite Gaussian choices is related to DDCM's compression framework (Ohayon et al., 2025), but our objective is reward alignment. The key question is what cheaper ranking preserves. For the basic rule that emits one selected base-transition draw, we characterize the selected-noise law and give a finite density-ratio bound even when the ranking is imperfect. We also measure the local proxy value lost relative to exact scoring of the *same* candidates. A matched mean shift can move the state equally far on average yet have a different directional variance and local proxy value. Together, these results explain the computational shortcut at the level of one reverse decision.

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

We first test the local prediction on MNIST, where we can score the same candidate pool exactly and compare measured gaps with theory. We then evaluate image generation, non-differentiable rewards, and D4RL control. This ordering separates a controlled check of the mechanism from larger-scale tests of reward and computation. Some image experiments use signed aggregation as a practical variant, which lies outside the single-winner analysis.

2 Related work

Mean-shift guidance. Pretrained diffusion models support inference-time steering through reward gradients and candidate selection (Yu et al., 2023, Chung et al., 2023, Li et al., 2025). Mean-shift methods add a reward-derived drift to the backward transition, most often the reward gradient at the Tweedie estimate (Chung et al., 2023). LAGS (Kim et al., 2026a) uses approximate guidance for reweighted sampling. DOIT (Zhu et al., 2026) uses a dynamic Doob correction to approach a specified high-reward target, and LiDAR (Kim et al., 2026b) builds the drift from lookahead samples. Both form the drift from reward values rather than reward gradients and apply it with a tuned guidance strength. Our selected noise has a mean displacement along w_t and an independent orthogonal Gaussian component. Its directional law remains non-Gaussian, so matching the mean displacement does not identify it with mean-shift guidance.

Noise selection. SVDD-PM (Li et al., 2025) scores alternatives during denoising by evaluating the reward on the Tweedie estimate of each of K candidates, whereas terminal best-of-N selects among completed outputs (Beirami et al., 2025). Our SVDD-PM baseline uses hard argmax of posterior-mean reward scores. Because hard selection returns one of N base draws, its divergence from the base grows at most logarithmically in N (Beirami et al., 2025); this certifies closeness to the base, but also limits how far a given evaluation budget can move the sample. CoDe (Singh et al., 2025) extends selection blockwise, while tree search and particle methods offer other computation–alignment trade-offs (Jain et al., 2025, Guo et al., 2025, Wu et al., 2023a, Kim et al., 2025, Singhal et al., 2025), and some, such as DAS (Kim et al., 2025), shift candidates with reward gradients before resampling. DDCM (Ohayon et al., 2025) selects among Gaussian candidates for compressed generation, and Turbo-DDCM (Vaisman et al., 2026) adds multi-entry combinations. Our noise-space construction is related, but its objective is reward alignment and our local theorem compares shared linear ranking with exact reward scoring of the same candidate pool. NoiseTilt (Hwang et al., 2026) changes reverse kernels through noise-space reward guidance and is a close experimental comparator. DNO (Tang et al., 2025) optimizes a full noise trajectory rather than making a stepwise choice. Demon (Yeh et al., 2025) also aggregates candidate noises. Its projected mixture, like our empirically normalized signed variant, is not covered by the single-winner density-ratio certificate. For basic single-winner selection, we replace K neural reward evaluations with one shared direction while retaining the finite-candidate bound.

3 Method

3.1 Preliminary: Diffusion Models

Diffusion models (Song et al., 2021a, Sohl-Dickstein et al., 2015, Song et al., 2021b) generate samples by reversing a gradual noising process. Let $t = 0, \dots, T$ index the chosen inference schedule, so generation takes T denoising steps. Network time inputs refer to the corresponding training-time labels. Starting from a clean sample $x_0 \sim P_{\text{data}}$, the forward process between adjacent schedule positions is

$$x_t = \sqrt{\alpha_t}\, x_{t-1} + \sqrt{1-\alpha_t}\, \epsilon, \qquad \epsilon \sim \mathcal{N}(0, I_d), \tag{1}$$

where α_t is the signal-retention factor and $\overline{\alpha}_t = \prod_{s=1}^{t} \alpha_s$. The denoiser $\epsilon_\theta(x_t, t)$ predicts the cumulative Gaussian noise in x_t, giving the clean-sample estimate (Efron, 2011)

$$\hat{x}_{0|t}(x_t) \;=\; \frac{x_t - \sqrt{1-\overline{\alpha}_t}\, \epsilon_\theta(x_t, t)}{\sqrt{\overline{\alpha}_t}} \;\approx\; \mathbb{E}[x_0 \mid x_t]. \tag{2}$$

For sampling, let $\mu_t = \mu_\theta(x_t, t)$ denote the mean computed by the base sampler at the current state. Starting from $x_T \sim \mathcal{N}(0, I_d)$, each reverse step draws

$$x_{t-1} \;=\; \mu_t + \sigma_t\, z_t, \qquad z_t \sim \mathcal{N}(0, I_d), \tag{3}$$

where σ_t is the standard deviation of the noise added in this reverse step. Our method selects this Gaussian noise using reward information, while keeping the base mean and noise scale unchanged. We write $P_t(\cdot \mid x_t)$ for this base transition and $p_t(\cdot \mid x_t)$ for its density at stochastic steps. For distributions, uppercase letters denote probability laws and lowercase letters their densities.

3.2 Preliminary: Soft Value-based Decoding (SVDD)

SVDD (Li et al., 2025) guides generation by comparing candidate next states. It favors candidates whose subsequent denoising is expected to produce high-reward outputs. Let $r(x_0)$ denote the clean-sample reward, including image decoding when needed, with the fixed prompt suppressed. For temperature $\tau > 0$, the soft value and the density of the ideal reward-weighted transition $q^\star_{t,\tau}$ are

$$V_t(x) = \tau \ln \mathbb{E}\big[e^{r(x_0)/\tau} \,\big|\, x_t = x\big], \qquad q^\star_{t,\tau}(y \mid x_t) \;\propto\; p_t(y \mid x_t)\, e^{V_{t-1}(y)/\tau}, \tag{4}$$

The expectation follows the base sampler from x_t to x_0. SVDD approximates this transition by drawing K_t candidate next states from the base transition,

$$y_t^{(k)} = \mu_t + \sigma_t \epsilon_t^{(k)}, \qquad \epsilon_t^{(k)} \overset{\text{iid}}{\sim} \mathcal{N}(0, I_d), \quad k = 1, \dots, K, \tag{5}$$

and sampling an index k_t^{SVDD} with the soft values $\widehat{V}_{t-1}(y_t^{(k)})$ using softmax sampling or choosing the best soft value. For the posterior-mean variant (SVDD-PM), the soft value is approximated by

$$\widehat{V}_{t-1}(y) = r\big(\hat{x}_{0|t-1}(y)\big)\,. \tag{6}$$

Below, F_t denotes the local score used to compare candidate next states. For posterior-mean scoring, $F_t(y) = \widehat{V}_{t-1}(y)$ from (6). The index t labels the transition that produces y at time $t-1$. For comparison with our linear rule, the hard-selection SVDD-PM baseline evaluates this score on every candidate and keeps

$$k_t^{\mathrm{ex}} = \arg\max_{k \le K_t} F_t(y_t^{(k)}), \qquad x_{t-1}^{\mathrm{ex}} = y_t^{(k_t^{\mathrm{ex}})}. \tag{7}$$

The rule in Equation 7 evaluates F_t K times per diffusion step, requiring $O(KT)$ diffusion and reward model forward passes, as do other selection-based methods (Li et al., 2025, Guo et al., 2025, Singh et al., 2025). In exchange, they have the advantage that every emitted state is drawn from the reliable base distribution and it is possible to keep track of the KL divergence deviation for each step (Singh et al., 2025, Beirami et al., 2025). However, the mean-shift guidance trades the opposite since the efficient reward gradient per step to shift the transition mean thereby deviating from the base distribution. To the first order, both routes move the next step next step toward the reward gradient. Indeed, the mean-shift-based methods themselves hint us about the direction towards a new intermediate reward optimization method (Kim et al., 2025, Hwang et al., 2026): candidate selection based on a reward maximizing linear direction.

3.3 Linear Noise Ranking (LNR)

Instead of evaluating the reward for each candidate, we rank the candidates by a single linear functional. At a guided step, the sampler forms a direction $w_t \in \mathbb{R}^d$ from the current state x_t and the base mean μ_t, unit-norm when nonzero and computed before the candidate noise ϵ_t is drawn, and selects

$$k_t^{\mathrm{lin}} = \arg\max_{k \le K} \langle w_t, \epsilon_t^{(k)} \rangle, \qquad x_{t-1} = y_t^{(k_t^{\mathrm{lin}})}. \tag{8}$$

Once w_t is computed, ranking K noises takes draws and inner products but no candidate-specific neural evaluations. The neural cost per guided step is the cost of computing w_t, independent of K (Algorithm 1).

The ideal soft value motivates a gradient direction, as in mean-shift guidance (Yu et al., 2023, Chung et al., 2023, Kim et al., 2026a). We instead differentiate the available candidate score F_t. For differentiable F_t, Taylor expansion at the shared transition mean gives

$$F_t\big(y_t^{(k)}\big) = F_t(\mu_t) + \sigma_t \langle g_t, \epsilon_t^{(k)} \rangle + R_{t,k}, \qquad g_t = \nabla_y F_t(y)\big|_{y=\mu_t}, \tag{9}$$

so the constant term is shared across candidates and, up to the remainder $R_{t,k}$, ranking by F_t is equivalent to ranking by $\langle g_t, \epsilon_t^{(k)} \rangle$. Therefore the natural target based on a single gradient is

Algorithm 1 LNR at one guided step (gradient direction).

Require: state x_t, base mean μ_t, scale σ_t, proxy F_t, candidate count K_t

1: $g_t \leftarrow \nabla_y F_t(y)|_{y=\mu_t}$ ▷ one forward, one backward
2: $w_t \leftarrow g_t/\|g_t\|_2$ if $g_t \neq 0$, else $w_t \leftarrow 0$ ▷ corrected or least-squares direction replaces this line
3: draw $\epsilon^{(1)}, \ldots, \epsilon^{(K_t)} \sim \mathcal{N}(0, I_d)$
4: $k^\star \leftarrow \arg\max_k \langle w_t, \epsilon^{(k)} \rangle$
5: $\epsilon^\star \leftarrow \epsilon^{(k^\star)}$ ▷ Or multiple noise aggregation (see Section 3.3).
6: **return** $x_{t-1} = \mu_t + \sigma_t \epsilon^\star$

$w_t = g_t/||g_t||$. Computing w_t with the gradient direction only requires one additional clean-sample prediction and one backward pass per diffusion step. LNR is not limited to the gradient direction. Section 5 also considers corrected and estimated directions, including directions for non-differentiable rewards.

Certainly, selecting the noise using the rule in Equation 8 is overly simplistic compared to the greedy evaluation that SVDD executes, considering the complex structure of F_t. This raises two questions. How much does LNR lose by choosing the noise from Equation 8? And how does the resulting sampler relate to mean-shift guidance with the same gradient? We answer both questions in the next Section.

4 Analysis of LNR and Mean-Shift Guidance

LNR sits between two routes: it selects among the base candidates like SVDD and follows a single gradient like guidance. We make its position exact by pricing its gap to the exact noise selection and to the mean shift guidance, and we check both prediction of the gap numerically on the MNIST (LeCun et al., 2010) dataset. Let $\Pi_t = I_d - ww^\top$ be the projector onto its orthogonal complement. Every noise vector splits as $\epsilon = zw + \Pi_t\epsilon$ with $z = \langle w, \epsilon \rangle$. Let M_K denote the maximum of K independent standard normals, with $c_K = \mathbb{E}M_K$ and $\nu_K = \mathrm{Var}M_K$. We use the classically known facts that $c_K = \sqrt{2\log K}\,(1 + o(1))$ and $\nu_K = \frac{\pi^2}{12\log K}(1 + o(1))$ (Leadbetter et al., 1983).

4.1 The price of the linear score

The linear score ranks by the first-order term of (9) and ignores $R_{t,k}$, so for some candidate sets it selects a candidate that per-candidate scoring would not. We measure this on the *same* candidate set, at the same state, by the one-step regret

$$\Delta_t = F_t\big(y^{(k_t^{\mathrm{ex}})}\big) - F_t\big(y^{(k_t^{\mathrm{lin}})}\big) \geq 0. \tag{10}$$

Write $a_t = \|g_t\|_2$, $H_t = \nabla^2 F_t(\mu_t)$, and, for the gradient direction $w = g_t/a_t$,

$$A_t = \|\Pi_t H_t w\|_2^2, \qquad B_t = \|\Pi_t H_t \Pi_t\|_F^2, \tag{11}$$

the squared coupling between the scored direction and its complement, and the squared curvature within the complement. Finally let $\gamma_K = \mathbb{E}M_K^3 - 2\,\mathbb{E}M_K$.

Theorem 4.1 (Leading-order scoring regret)**.** *Suppose F_t is twice continuously differentiable with $\|\nabla^2 F_t\|_{\mathrm{op}} \leq L$, and $g_t \neq 0$. Then for the rule (8) with the gradient direction (9), as $\sigma_t \to 0$ with K and d fixed,*

$$\mathbb{E}[\Delta_t \mid x_t] = \kappa_t\,\sigma_t^3 + o(\sigma_t^3), \qquad \kappa_t = \frac{1}{2a_t}\Big[\gamma_K A_t + \frac{c_K}{2} B_t\Big]. \tag{12}$$

The proof is in Appendix B.1. The expansion is conditional on a fixed state and holds as $\sigma_t \to 0$ with K and d fixed. In our schedules, σ_t^2 is typically of order $1/T$. This suggests a cubic local scaling only when the state-dependent coefficient remains controlled. A change in T also changes the visited states (a_t, A_t, B_t), so the theorem alone does not predict a trajectory-level rate.

4.2 The selection channel

Proposition 4.2 (Selection channel)**.** *Let w be a unit vector determined by x_t, let $\epsilon^{(1)}, \ldots, \epsilon^{(K)} \sim \mathcal{N}(0, I_d)$ be drawn independently afterwards, and let $k^\star$ depend only on the scores $z_k = \langle w, \epsilon^{(k)} \rangle$*

and any additional randomness independent of the candidate noises. Then

$$\epsilon^{(k^\star)} = z_{k^\star}\, w + \epsilon_\perp, \qquad \epsilon_\perp \sim \mathcal{N}(0, \Pi_t) \text{ independent of } (z_1, \dots, z_K, k^\star). \tag{13}$$

For the rule (8), $z_{k^\star} = \max_k z_k$ has the law of M_K, so the selected noise has mean $c_K w$ and covariance $\nu_K\, ww^\top + \Pi_t$.

Proof. The projections $\Pi_t \epsilon^{(k)}$ are independent of the coordinates $(z_j)_{j \le K}$ and are themselves independent $\mathcal{N}(0, \Pi_t)$ vectors. The index $k^\star$ is a function of (z_j) and of randomness independent of the candidate noises, so conditioning on $\{k^\star = k\}$ does not change the law of $\Pi_t \epsilon^{(k)}$, which is therefore $\mathcal{N}(0, \Pi_t)$ independent of (z_j) and $k^\star$. □

Selection changes only the scalar noise coordinate along w. Its mean is c_K and its variance is ν_K, while the orthogonal component remains an independent base Gaussian. A Gaussian mean shift by $c_K w$ has the same mean displacement but unit variance along w. The two laws are not identical, because the LNR's scalar noise coordinate has the non-Gaussian law of M_K. The exact channel law in Proposition 4.2 lets us compare selection with a Gaussian mean shift having the same average directional displacement. The mean-shift rule samples

$$y_{\mathrm{MS},t} = \mu_t + \alpha_t g_t + \sigma_t \epsilon_t = \mu_t + \sigma_t(\rho_t w_t + \epsilon_t),$$

where $\epsilon_t \sim \mathcal{N}(0, I_d)$, $w_t = g_t / a_t$, and $\rho_t = \alpha_t a_t / \sigma_t$. Thus the mean is shifted along the reward gradient while the full Gaussian noise is retained. This gives us intuition about the difference between mean shift and LNR: LNR's fluctuation is mostly along the orthogonal direction of w_t, whereas mean-shift has a constant variance along that direction. Another important result from the above proposition is that now it is possible to compute the regret of the mean shift method. The mean-shift scoring gap is

$$\Delta_{\mathrm{MS},t} = F_t\big(y^{(k_t^{\mathrm{ex}})}\big) - F_t\big(y_{\mathrm{MS},t}\big) = \Delta_t + F_t\big(y^{(k_t^{\mathrm{lin}})}\big) - F_t\big(y_{\mathrm{MS},t}\big). \tag{14}$$

Proposition 4.3 (Mean-shift leading order regret)**.** *Under the assumptions of Theorem 4.1, additionally suppose that F_t is three times continuously differentiable in a neighborhood of μ_t. For fixed K, d, and ρ_t, as $\sigma_t \to 0$, with $h_{\parallel,t} = w_t^\top H_t w_t$,*

$$\mathbb{E}\big[\Delta_{\mathrm{MS},t} \mid x_t\big] = \sigma_t a_t\,(c_K - \rho_t) + \frac{\sigma_t^2}{2}\, h_{\parallel,t}\Big(\mathbb{E}M_K^2 - \rho_t^2 - 1\Big) + O(\sigma_t^3). \tag{15}$$

In particular, at the matched strength $\rho_t = c_K$,

$$\mathbb{E}\big[\Delta_{\mathrm{MS},t} \mid x_t\big] = \frac{\sigma_t^2}{2}\, h_{\parallel,t}\,(\nu_K - 1) + O(\sigma_t^3), \tag{16}$$

For $K \ge 2$ and $h_{\parallel,t} \neq 0$, this gap has sign $-\operatorname{sign}(h_{\parallel,t})$ for sufficiently small σ_t, since $\nu_K < 1$.

For fixed $\rho_t > c_K$, the first-order term favors mean shift. At matched displacement $\rho_t = c_K$, the sign of the leading gap is set by $h_{\parallel,t}(\nu_K - 1)$. Positive directional curvature favors mean shift, while negative curvature favors exact selection, for sufficiently small σ_t (i.e., large T). Figure 4 shows the measured directional curvature in our MNIST experiment. Notice that the mean shift method has different leading order compared to LNR. This implies that they will have different local regret trends on large T settings.

4.3 Proof of concept: local scoring on MNIST

A controlled first check asks whether shared linear scoring loses the proxy value predicted by the local analysis. Figure 1 compares local proxy gaps on pixel-space MNIST diffusion ($d = 784$), with a smooth CNN classifier reward, $K = 64$, and 52 trajectories. The mean-shift control uses the matched displacement $\rho_t = c_K$. Solid curves are measured gaps, while dotted curves are the leading terms of Theorem 4.1 and Proposition 4.3. The horizontal coordinate $\tau = t/T$ locates a reverse step in its schedule. We observe that the predictions track the trend and fluctuations of the measured local gaps in this experiment, which reaffirms the correctness of the leading order terms from Theorem 4.1 and Proposition 4.3. Furthermore, summed local regret comparison of LNR and mean shift with matched directional strength shows that as T increases, LNR yields less local regrets compared to the mean shift, which is consistent to the fact that the measured curvature of $h_{\parallel}$ are mostly negative for the MNIST data (See Figure 4 in the Appendix). We provide further experimental results in the Appendix that successfully validates the theoretical results from this section (see Figure 4,5,6).

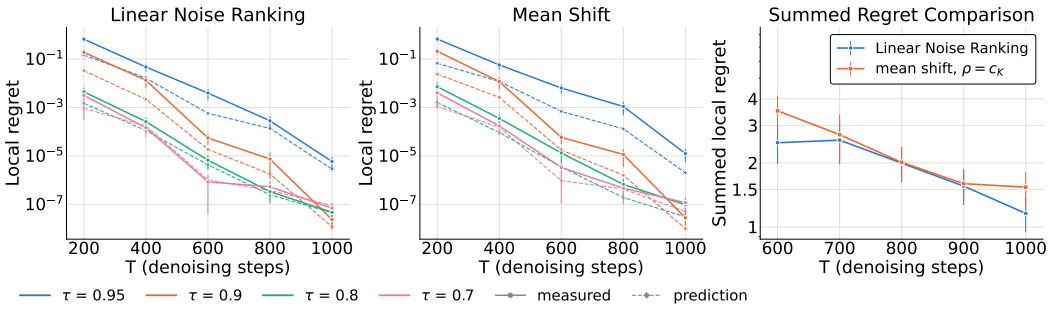


Figure 1: LNR local proxy gaps on MNIST for $K = 64$ across 52 trajectories. Solid curves are measured and dotted curves are leading-order predictions. The summed panel follows different generated states as T changes.

5 Applying LNR to Broader Reward Models

We follow the analysis from a local decision to full applications. We evaluate reward alignment on Stable Diffusion 1.5, first with differentiable PickScore and then with non-differentiable rewards, before turning to diffusion control on D4RL. The MNIST study checks the mechanism. The later studies assess reward and computation in the intended applications.

Target direction variants. Selection along the unit direction $g_t/\|g_t\|_2$ gives the conditional mean next state $\mu_t + \sigma_t c_{K_t} g_t/\|g_t\|_2$, where c_K is the expected maximum of K independent standard normals (we further explain in Section 4.2). This is a mean location, not the realized selected state. The gradient at μ_t is therefore evaluated $\sigma_t c_{K_t}$ away from that mean location. A self-consistent gradient, if one exists, would satisfy

$$v = \mathcal{T}_t(v), \qquad \mathcal{T}_t(v) := \nabla_y F_t(y)\Big|_{y=\mu_t+\sigma_t c_{K_t}\, v/\|v\|_2}. \tag{17}$$

Here v is a gradient vector, while $v/\|v\|_2$ is the unit selection direction. We use (17) to motivate a finite number of damped updates,

$$v_0 = g_t, \qquad v_{n+1} = (1-\omega)\, v_n + \omega\, \mathcal{T}_t(v_n), \qquad \omega \in (0, 1], \tag{18}$$

After P updates, we use $w_t = v_P/\|v_P\|_2$ when $v_P \neq 0$. Each update costs one additional proxy-gradient evaluation. We use at most two updates in the experiments (see Algorithm 2 in Appendix).

When F_t can be evaluated but not differentiated, we estimate a local slope $\hat{g}_t$ from scored Gaussian probes. Draw $z_1, \ldots, z_P \overset{\text{iid}}{\sim} \mathcal{N}(0, I_d)$ and evaluate $F_i = F_t(\mu_t + \sigma_t z_i)$ for each probe. We fit an affine linear model to these scores by least squares,

$$(\hat{b}_t, \hat{g}_t) \in \arg\min_{b,g} \sum_{i=1}^{P} \big(F_i - b - \langle g, z_i\rangle\big)^2 \quad \text{(minimum-norm slope)}, \qquad w_t = \hat{g}_t/\|\hat{g}_t\|_2 \quad (\hat{g}_t \neq 0). \tag{19}$$

The intercept avoids a separate evaluation at μ_t, so the fit uses exactly P score evaluations. Only the slope's orientation affects ranking. We set $w_t = 0$ if $\hat{g}_t = 0$. In the non-differentiable experiments, the scored probes are reused alongside a fresh, unscored noise pool (Algorithm 3 in Appendix).

Noise aggregation. One practical variant emits the fixed-normalization sum $M^{-1/2} \sum_{k \in S_m} \epsilon_t^{(k)}$ of the top M scores (Vaisman et al., 2026). It increases the average displacement along w_t without additional neural evaluations. For each fixed subset S, its sum is marginally standard Gaussian. The local-loss theorem below concerns $M = 1$.



Figure 2: Selected examples generated by PickScore reward optimization using our method and the baselines. Prompts and seeds are paired across methods.

5.1 PickScore steering on Image Diffusion Model

Our main image experiments steer Stable Diffusion v1.5 at 512^2 resolution with a 200-step DDIM schedule ($\eta = 1$), leaving the final deterministic step unchanged (Rombach et al., 2022, Song et al., 2021a). The reward is PickScore (Kirstain et al., 2023), a CLIP-based model trained on human preferences that scores how well an image matches its prompt. We steer it on the 50 prompts of DAS (Kim et al., 2025) with four seeds each (200 images), and report HPSv2 (Wu et al., 2023b), ImageReward (Xu et al., 2023), and CLIP cosine similarity (Radford et al., 2021) as held-out text-alignment metrics. Our rows use noise pools of size $K = 2^{14}$ and signed top-M aggregation with empirical standard-deviation normalization (Appendix A.1), with either the gradient direction or the corrected direction of Equation 18 after $n = 2$ damped updates. We compare against guidance methods (FreeDoM (Yu et al., 2023), LAGS (Kim et al., 2026a), and our SD/DDIM adaptation of NoiseTilt (Hwang et al., 2026)) and selection methods (SVDD-PM, DAS, and the gradient-free TreeG-SD (Guo et al., 2025)). We adapt TreeG-SD to SD1.5 with two paths of four destination candidates each, keeping the best two candidates per search step; it takes 35.19 s and 142.60 s per output at 50 and 200 steps on an A100-80GB.

Table 1: PickScore steering comparing with matched strength with mean shift method.

Method	Time↓	Pick	IR	TOPIQ
Base	1.8	0.2183	0.026	0.699
FreeDoM (r=3)	12.6	0.2307	0.376	0.672
DAS (N=4)	38.9	0.2354	0.459	0.680
SVDD-PM (N=2)	10.6	0.2242	0.369	0.689
SVDD-PM (N=4)	19.6	0.2279	0.354	0.674
SVDD-PM (N=8)	36.9	0.2323	0.454	0.677
LNR (linear)	11.5	0.2294	0.264	0.686
LNR (corrected)	30.9	0.2341	0.366	0.687
Mean shift (ρ matched)	11.5	0.2297	0.247	0.682

Table 1 tests the single-winner rule analyzed in Section 4 at 50 DDIM steps with $K = 4096$, on the same GPU for LNR, SVDD-PM, and the mean-shift control. Since LNR computes one direction per step regardless of K, enlarging the candidate pool adds only sampling and inner products. Linear LNR runs in 11.5 s, close to two-candidate SVDD-PM (10.6 s), yet exceeds four-candidate SVDD-PM in PickScore (0.2294 vs. 0.2279) at 59% of its time. The corrected direction reaches 0.2341 in 30.9 s, above eight-candidate SVDD-PM (0.2323 in 36.9 s) and close to DAS (0.2354 in 38.9 s).

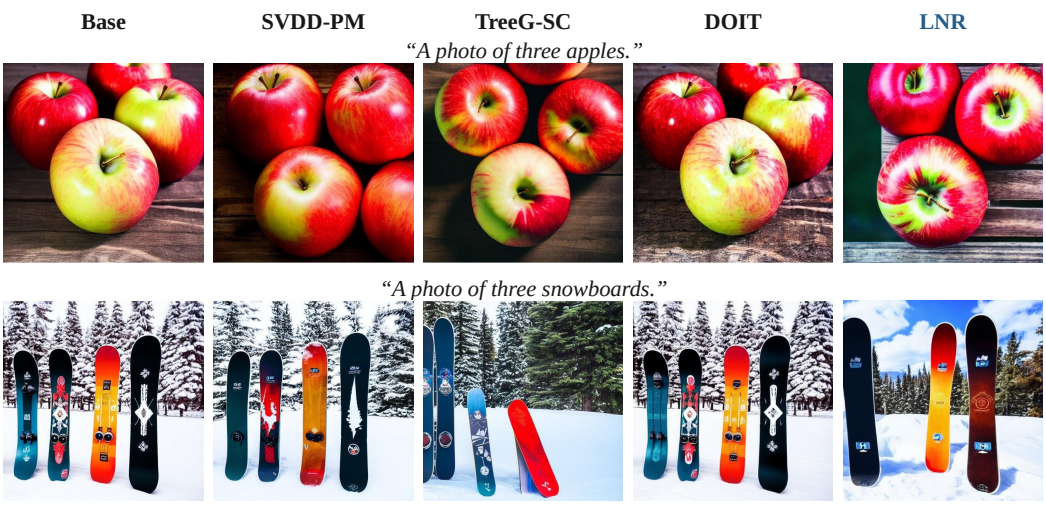


Figure 3: Selected generations for the hard counting reward model on GenEval.

Table 2: Stable Diffusion 1.5. PickScore steering.

PickScore steering

Method	Pick	HPS	IR	CLIP
SD 1.5	0.2172	0.2799	0.146	0.2573
FreeDoM	0.2423	0.2971	0.561	0.2580
LiDAR	0.2187	0.2817	0.238	0.2567
LAGS	0.2167	0.2780	−0.035	0.2542
DAS	0.2529	0.3027	**0.626**	0.2580
NoiseTilt (ρ=.03)	**0.3104**	**0.3229**	0.487	0.2586
TreeG-SD (T=50)	0.2203	0.2815	0.270	0.2596
TreeG-SD (T=200)	0.2227	0.2845	0.302	0.2589
SVDD-PM (N=4)	0.2329	0.2918	0.486	0.2577
SVDD-PM (N=8)	0.2387	0.2945	0.456	0.2577
LNR	0.2637	0.3049	0.501	0.2594
LNR (corrected)	0.2784	0.3083	0.576	**0.2619**

Non-differentiable rewards

Method	JPEG↓	Count↑	2-obj.↑
Base	108.50	0.388	0.465
DNO	39.98	0.450	0.606
SVDD-PM	41.36	0.625	0.717
TreeG-SC	36.35	0.625	0.808
DOIT	102.31	0.475	0.545
LiDAR	38.27	0.438	0.505
LNR	**33.59**	**0.700**	**0.818**

Table 3: Reward alignment on non-differentiable rewards (JPEG compressibility, object counting, two object prompts) on Stable Diffusion 1.5

Per-candidate scoring therefore buys less reward per second than a single shared direction. The mean-shift control adds $\sigma_t c_{K_t} w_t$ to a fresh Gaussian draw, matching LNR's mean displacement ($\rho_t = c_{K_t}$) but not its selected-noise distribution. It reaches 0.2297 at the same cost as linear LNR, as Proposition 4.3 predicts: at matched strength the first-order terms coincide and the two differ only at order σ_t^2. We therefore do not claim superiority over this control; rather, LNR attains the reward of gradient guidance while every emitted state remains a draw from the base transition. Held-out TOPIQ for both LNR rows stays near the base (0.686–0.687 vs. 0.699), above FreeDoM, DAS, and SVDD-PM with four or eight candidates, indicating that the reward gain does not come at the cost of image quality. Figure 2 shows paired outputs, and Appendix C.5 gives the full protocol and candidate-count sweep.

In Table 2, we use the baseline's default settings instead of the gradient strength-matched parameter. In this method, we utilize the top-M noise aggregation method. LNR raises PickScore from 0.2172 to 0.2637, and the corrected direction to 0.2784, the second highest in Table 2. Both exceed every selection baseline, including SVDD-PM with eight candidates (0.2387) and TreeG-SD (0.2227), and every guidance baseline except NoiseTilt (a concurrent work), a concurrent method evaluated on 50 images. The gains carry over to held-out metrics: the corrected row has the highest CLIP score (0.2619) and the second-highest HPS (0.3083) and ImageReward (0.576).

Table 4: Experimental results on D4RL. Higher is better. Entries are means ± standard deviations across independent runs, and the final row averages the nine task means. Linear denotes the basic shared-gradient rule.

Dataset	DOIT	DAS	SVDD-PM	LNR	LNR corrected
HalfCheetah-ME	91.4 ± 0.6	92.5 ± 0.4	93.3 ± 0.5	92.8 ± 1.1	93.0 ± 0.6
Hopper-ME	106.1 ± 1.4	107.7 ± 0.8	105.5 ± 3.4	101.2 ± 7.3	103.9 ± 5.2
Walker2d-ME	110.8 ± 0.2	109.5 ± 0.0	112.2 ± 0.1	112.1 ± 2.8	112.3 ± 2.9
HalfCheetah-M	52.0 ± 0.2	49.6 ± 0.3	48.8 ± 0.1	52.5 ± 0.3	52.7 ± 0.2
Hopper-M	79.1 ± 6.5	85.4 ± 8.7	87.9 ± 5.7	90.8 ± 3.7	90.8 ± 7.5
Walker2d-M	85.2 ± 0.1	83.6 ± 0.3	84.4 ± 0.2	85.5 ± 1.9	85.8 ± 0.9
HalfCheetah-MR	41.8 ± 0.2	44.5 ± 0.2	43.4 ± 0.5	46.5 ± 0.7	46.8 ± 0.3
Hopper-MR	98.3 ± 2.0	89.3 ± 3.4	69.7 ± 5.1	92.5 ± 3.1	93.0 ± 3.3
Walker2d-MR	65.8 ± 3.1	61.5 ± 6.2	42.8 ± 6.5	60.3 ± 7.6	63.2 ± 7.9
Average	81.2	80.4	76.4	81.6	**82.4**

5.2 Non-differentiable rewards

The right panel of Table 2 reports JPEG file size at quality 95 (Li et al., 2025) and two GenEval evaluations (Ghosh et al., 2023): exact counting and two-object accuracy. For counting, we use the black-box score $r = -|n_{\text{det}} - n_{\text{target}}|$, with n_{det} from Mask2Former (Cheng et al., 2022). Our affine least-squares direction uses $P = 20$ scored probes per guided step. The update reuses selected probes and adds fresh noises ranked by the fitted direction, with $M = 5$ for JPEG and $M = 2$ for counting (Algorithm 3). The detailed settings are reported in Table 5 of the Appendix. Our row has the lowest JPEG size (33.59 KB) among those shown. Its counting and two-object accuracies are 0.700 and 0.818, versus 0.625 and 0.808 for TreeG-SC.

5.3 D4RL control

Table 4 reports nine D4RL locomotion tasks (Fu et al., 2020), following the DOIT setup (Zhu et al., 2026) for pretrained diffusion policies and value functions. All methods return a single sample without terminal best-of-N selection. DOIT's published numbers apply best-of-N over its own samples, which accounts for the difference from its paper. DAS uses 16 particles and SVDD-PM at most 20 proposals. The linear rule averages a normalized return of 81.6, compared with 81.2 for DOIT, 80.4 for DAS, and 76.4 for SVDD-PM. The corrected row averages 82.4 but requires additional proxy-gradient evaluations. These means summarize the reported runs rather than establishing a task-wise advantage over every baseline.

6 Conclusion

We presented LNR, a shared-direction rule that bridges candidate selection and mean-shift guidance for inference-time reward alignment. It ranks fresh Gaussian noises by inner products with a direction computed once per step, so enlarging the candidate pool adds arithmetic rather than neural evaluations. For single-winner selection, the selected noise has an exact one-dimensional law with density ratio at most K, and its local proxy loss relative to exact scoring of the same candidates is only $\kappa_t \sigma_t^3$, whereas a strength-matched mean shift already departs from exact selection at order σ_t^2. Because a sampler makes T such decisions, these local effects accumulate into the final output: the summed cubic loss shrinks as the schedule is refined when κ_t stays controlled, and on MNIST the summed loss of LNR stays below that of a matched mean shift. The same local rule raises the terminal reward across image, black-box, and control tasks, with the lowest JPEG size, the highest counting and two-object accuracy, and the best D4RL average among the tested methods, while corrected and least-squares directions extend the reward–computation trade-off. Extending the local analysis to trajectory-level guarantees and to aggregated outputs is a natural next step.

AI USE STATEMENT

We used large language model tools from OpenAI and Anthropic, as assistants in developing this work. They helped refine the noise-selection formulation, formulate and scrutinize mathematical claims and proof steps, and construct diagnostic examples and counterexamples. They also assisted with experiment design, implementation and debugging, evaluation scripts, and interpretation of quantitative and qualitative results. We used these tools to search for and check references, translate and edit technical text, improve readability, and prepare tables and scientific figures. We treated AI suggestions as provisional. For material included in the paper, we compared AI-assisted mathematical steps with written derivations and numerical checks, empirical claims with saved runs and evaluation code, and citations with source publications, as appropriate. The authors made the final research and editorial decisions and take responsibility for all text, mathematics, code, figures, and results in this submission.

REFERENCES

Ahmad Beirami, Alekh Agarwal, Jonathan Berant, Alexander D'Amour, Jacob Eisenstein, Chirag Nagpal, and Ananda Theertha Suresh. Theoretical guarantees on the best-of-n alignment policy. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 3580–3602. PMLR, 2025. URL https://proceedings.mlr.press/v267/beirami25a.html.

Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1290–1299, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Cheng_Masked-Attention_Mask_Transformer_for_Universal_Image_Segmentation_CVPR_2022_paper.html.

Hyungjin Chung, Jeongsol Kim, Michael T. McCann, Marc L. Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. In *International Conference on Learning Representations*, 2023.

Bradley Efron. Tweedie's Formula and Selection Bias. *Journal of the American Statistical Association*, 106(496):1602–1614, 2011. doi: 10.1198/jasa.2011.tm11181.

Justin Fu, Aviral Kumar, Ofir Nachum, George Tucker, and Sergey Levine. D4RL: Datasets for deep data-driven reinforcement learning. *arXiv preprint arXiv:2004.07219*, 2020.

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*, volume 36, pp. 52132–52152, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/a3bf71c7c63f0c3bcb7ff67c67b1e7b1-Abstract-Datasets_and_Benchmarks.html.

Yingqing Guo, Yukang Yang, Hui Yuan, and Mengdi Wang. Training-free guidance beyond differentiability: Scalable path steering with tree search in diffusion and flow models. In *Advances in Neural Information Processing Systems*, volume 38, pp. 73343–73384, 2025. doi: 10.52202/085713-2459. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/6a14c7f9fb3f42645cfa6bd5aa446819-Abstract-Conference.html.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html.

Jisung Hwang, Yunhong Min, Jaihoon Kim, I-Chao Shen, and Minhyuk Sung. NoiseTilt: Noise-tilted reverse kernels for diffusion reward alignment. *arXiv preprint arXiv:2606.18066*, 2026. URL https://arxiv.org/abs/2606.18066.

Vineet Jain, Kusha Sareen, Mohammad Pedramfar, and Siamak Ravanbakhsh. Diffusion tree sampling: Scalable inference-time alignment of diffusion models. *NeurIPS*, 2025.

Michael Janner, Yilun Du, Joshua Tenenbaum, and Sergey Levine. Planning with diffusion for flexible behavior synthesis. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 9902–9915. PMLR, 2022. URL https://proceedings.mlr.press/v162/janner22a.html.

Heasung Kim, Taekyun Lee, Hyeji Kim, and Gustavo de Veciana. Efficient weighted sampling via score-based generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026a. URL https://openaccess.thecvf.com/content/CVPR2026/html/Kim_Efficient_Weighted_Sampling_via_Score-based_Generative_Models_CVPR_2026_paper.html.

Sunwoo Kim, Minkyu Kim, and Dongmin Park. Test-time alignment of diffusion models without reward over-optimization. In *International Conference on Learning Representations*, 2025. Spotlight. arXiv:2501.05803.

Yeongmin Kim, Donghyeok Shin, Byeonghu Na, Minsang Park, Richard Lee Kim, and Il-Chul Moon. Lookahead sample reward guidance for test-time scaling of diffusion models. *ICML*, 2026b. URL https://arxiv.org/abs/2602.03211.

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*, volume 36, pp. 36652–36663, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/73aacd8b3b05b4b503d58310b523553c-Abstract-Conference.html.

M. R. Leadbetter, Georg Lindgren, and Holger Rootzén. *Extremes and Related Properties of Random Sequences and Processes*. Springer-Verlag, New York, 1983. doi: 10.1007/978-1-4612-5449-2. URL https://link.springer.com/book/10.1007/978-1-4612-5449-2.

Yann LeCun, Corinna Cortes, Chris Burges, et al. Mnist handwritten digit database. 2010.

Xiner Li, Yulai Zhao, Chenyu Wang, Gabriele Scalia, Gokcen Eraslan, Surag Nair, Tommaso Biancalani, Shuiwang Ji, Aviv Regev, Sergey Levine, and Masatoshi Uehara. Derivative-free guidance in continuous and discrete diffusion models with soft value-based decoding. *NeurIPS*, 2025.

Guy Ohayon, Hila Manor, Tomer Michaeli, and Michael Elad. Compressed image generation with denoising diffusion codebook models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 47044–47089. PMLR, 2025.

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*, 2024. URL https://openreview.net/forum?id=di52zR8xgf.

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 *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

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

Anuj Singh, Sayak Mukherjee, Ahmad Beirami, and Hadi Jamali-Rad. CoDe: Blockwise control for denoising diffusion models. *Transactions on Machine Learning Research*, 2025. arXiv:2502.00968.

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. *ICML*, 2025.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pp. 2256–2265. pmlr, 2015.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021a. URL https://openreview.net/forum?id=St1giarCHLP.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021b.

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*, volume 267 of *Proceedings of Machine Learning Research*, pp. 58905–58930. PMLR, 2025. URL https://proceedings.mlr.press/v267/tang25h.html.

Amit Vaisman, Guy Ohayon, Hila Manor, Michael Elad, and Tomer Michaeli. Turbo-ddcm: Fast and flexible zero-shot diffusion-based image compression. In *International Conference on Learning Representations*, 2026.

Luhuan Wu, Brian Trippe, Christian Naesseth, David Blei, and John P. Cunningham. Practical and asymptotically exact conditional sampling in diffusion models. *Advances in Neural Information Processing Systems*, 36:31372–31403, 2023a.

Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023b. URL https://arxiv.org/abs/2306.09341.

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*, volume 36, pp. 15903–15935, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/33646ef0ed554145eab65f6250fab0c9-Abstract-Conference.html.

Po-Hung Yeh, Kuang-Huei Lee, and Jun-Cheng Chen. Training-free diffusion model alignment with sampling demons. In *International Conference on Learning Representations*, 2025. arXiv:2410.05760.

Jiwen Yu, Yinhuai Wang, Chen Zhao, Bernard Ghanem, and Jian Zhang. FreeDoM: Training-free energy-guided conditional diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 23174–23184, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Yu_FreeDoM_Training-Free_Energy-Guided_Conditional_Diffusion_Model_ICCV_2023_paper.html.

Qijie Zhu, Zeqi Ye, Han Liu, Zhaoran Wang, and Minshuo Chen. Training-free adaptation of diffusion models via doob's h-transform. *ICML*, 2026.

Appendix

A Hyperparameters and Implementation Details

In this section, we present the hyperparameter used for running our method as well as other baselines. We show detailed algorithm of the corrected direction (Algorithm 2) and the direction for non-differentiable reward models (Algorithm 3). For the non-differentiable reward models, we detail the

Algorithm 2 Corrected direction (one guided step)

Require: state x_t, base mean μ_t, noise scale σ_t, initial candidate count K, damping $\omega \in (0, 1]$, number of refinements n, small scalar value $\delta = 10^{-10}$

$v_0 \leftarrow \nabla_y F_t(y)\big|_{y=\mu_t}$; $K_{\text{site}} \leftarrow K$ ▷ gradient at the transition mean
for $j = 0, \dots, n-1$ **do**
  $f_j \leftarrow \nabla_y F_t\big(\mu_t + \sigma_t c_{K_{\text{site}}}\, v_j/\|v_j\|_2\big) - v_j$ ▷ residual at the landing point
  if $j = 0$ **then**
    $v_{j+1} \leftarrow v_j + \omega f_j$ ▷ damped update
  else
    $\theta \leftarrow \text{clip}_{[-2,2]}\big(\langle f_j, f_j - f_{j-1}\rangle/(\|f_j - f_{j-1}\|_2^2 + \delta\})\big)$ ▷ Anderson(1)
    $v_{j+1} \leftarrow v_j + \omega f_j - \theta\big[(v_j - v_{j-1}) + \omega(f_j - f_{j-1})\big]$
  end if
end for
If successive update norms fail to decrease, repeat from v_0 with $K_{\text{site}} \leftarrow \max\{2, \lfloor\sqrt{K}\rfloor\}$; if the retry fails, set $v_n \leftarrow v_0$
$w_t \leftarrow v_n/\|v_n\|_2$
Draw $\epsilon^{(1)}, \dots, \epsilon^{(K_{\text{site}})}$ i.i.d. from $\mathcal{N}(0, I_d)$
$k^\star \leftarrow \arg\max_k \langle w_t, \epsilon^{(k)}\rangle$
return $x_{t-1} \leftarrow \mu_t + \sigma_t\, \epsilon^{(k^\star)}$

Algorithm 3 LNR with a non-differentiable reward

Require: score F_t, $P \geq M \geq 1$ scored probes, $K \geq M_2 \geq 0$ fresh noises, EMA weight $\beta \in [0, 1)$, fresh-noise weight $\lambda \geq 0$

$x_T \sim \mathcal{N}(0, I)$ and $\bar{g} \leftarrow 0$
for $t = T, \dots, 1$ **do**
  compute the base mean μ_t and noise scale σ_t from x_t
  draw $\epsilon_1, \dots, \epsilon_P \sim \mathcal{N}(0, I)$
  for $i = 1, \dots, P$ **do**
    $F_i \leftarrow F_t(\mu_t + \sigma_t \epsilon_i)$ ▷ one score evaluation
  end for
  obtain $\hat{g}_t$ as the minimum-norm slope in (19)
  $\bar{g} \leftarrow \beta\bar{g} + (1 - \beta)\hat{g}_t$ ▷ set $\beta = 0$ to disable averaging
  let I_t index the M largest $\langle \bar{g}, \epsilon_i\rangle$, $1 \leq i \leq P$
  $s_1 \leftarrow \sum_{i \in I_t} \epsilon_i$ ▷ reuse scored probes
  draw $\xi_1, \dots, \xi_K \overset{\text{iid}}{\sim} \mathcal{N}(0, I_d)$
  let J_t index the M_2 largest $\langle \bar{g}, \xi_j\rangle$, $1 \leq j \leq K$
  $s_2 \leftarrow \sum_{j \in J_t} \xi_j$ ▷ no extra score evaluations
  $x_{t-1} \leftarrow \mu_t + \sigma_t\,(s_1 + \lambda s_2)/\sqrt{M + \lambda^2 M_2}$
end for
return $\text{Dec}(x_0)$

A.1 Implementation details and scope

Schedule, mean, and gradient. The sampler uses the DDIM variance with $\eta = 1$ and the corresponding training-time labels. One classifier-free-guidance forward computes the base mean. That

T=50 sampling steps					
Method	Setting	NFE	Decodes	Reward evals	Time (s)
Base (no guidance)	–	100	1	0	2
SVDD-PM	P=20	2,060	981	980	80
TreeG-SC	A=2, K=10	2,160	983	982	81
Doob h-transform	M=20, γ=0.5, τ=0.5	100 + 96 bwd	961	960	72
DNO	50 it., B=4, μ=0.01, lr = 0.01	30,100 + 5,000 bwd	301	251	381†
LiDAR	K=50, λ=5000, s=12.5	600	51	50	8‡
LiDAR, matched budget	K=980, λ=5000, s=12.5	9,900	981	980	111†
LNR	P=20, M=2	2,060	981	980	81

Table 5: Hyperparameter settings for the non-differentiable reward model steering experiments.

mean is then treated as an independent input for the reward gradient. A second forward at the next training-time label predicts a clean latent, which is decoded and scored. For images, r includes VAE decoding and the fixed-prompt image reward, so differentiation still passes through the decoder. All smoothness assumptions concern the full composite proxy F_t. Differentiation is with respect to the mean, not x_t, and does not pass through the preceding sampling update. The proxy is a predicted clean-image reward rather than an exact expected future reward.

B PROOFS

B.1 PROOF OF THEOREM 4.1: THE EXACT LEADING ORDER

This appendix proves the asymptotic statement of Theorem 4.1. Section B.1.4 below treats a perturbed scoring direction. Only the stated hypotheses are used: F twice continuously differentiable with $\|\nabla^2 F\|_{\mathrm{op}} \le L$ everywhere, and $g \neq 0$. If $L = 0$, then F is affine and $\Delta = 0$. In the rest of the proof, take $L > 0$.

A simplistic overview of the proof is first bounding Δ (Equation 25) using index-wise differences D_i's, and then showing that $\mathbb{E}[(D_2)_+] = \kappa\sigma^3 + o(\sigma^3)$ and that the other terms are $o(\sigma^3)$ as well.

B.1.1 SETTING AND NOTATION

Fix the step and condition on x_t; all expectations below are over the candidate noises only, and we drop the index t. Write μ, $\sigma > 0$, F, and

$$\begin{aligned} g &= \nabla F(\mu) \neq 0, \qquad a = \|g\|_2, \qquad w = g/a, \\ H &= \nabla^2 F(\mu), \qquad \Pi = I_d - ww^\top. \end{aligned} \tag{20}$$

Set

$$b = \Pi H w, \qquad A = \|b\|_2^2, \qquad H_\perp = \Pi H \Pi, \qquad B = \|H_\perp\|_F^2, \qquad h_\parallel = w^\top H w, \tag{21}$$

so that $A, B \le$ powers of L and $|h_\parallel| \le L$. The candidates are $\epsilon^{(1)}, \dots, \epsilon^{(K)} \sim \mathcal{N}(0, I_d)$ i.i.d. with $K \ge 2$ fixed ($K = 1$ gives $\Delta \equiv 0$ and $\gamma_1 = c_1 = 0$, so the theorem is trivial), and

$$z_k = \langle w, \epsilon^{(k)} \rangle, \qquad x_k = \Pi \epsilon^{(k)}, \qquad y^{(k)} = \mu + \sigma \epsilon^{(k)}. \tag{22}$$

The rule (8) with the direction (9) selects $J = \arg\max_k z_k$ (ties broken by index; ties have probability zero), and

$$\Delta = \max_{k \le K} F\big(y^{(k)}\big) - F\big(y^{(J)}\big) \ \ge 0. \tag{23}$$

Let $z_{(1)} > z_{(2)} > \cdots > z_{(K)}$ be the ordered scores and $I_1, \dots, I_K$ the corresponding indices, and write

$$D_j = F\big(y^{(I_j)}\big) - F\big(y^{(I_1)}\big), \qquad j = 2, \dots, K, \tag{24}$$

so that $\Delta = \max\big(0, \max_{2 \le j \le K} D_j\big)$ and

$$(D_2)_+ \ \le \ \Delta \ \le \ (D_2)_+ + \sum_{j=3}^{K} (D_j)_+. \tag{25}$$

Finally, M_K denotes the maximum of K independent standard normals, $c_K = \mathbb{E}M_K$, $\gamma_K = \mathbb{E}M_K^3 - 2\,\mathbb{E}M_K$, and $N = \max_k \|\epsilon^{(k)}\|_2^2$; N has moments of all orders.

Since $F \in C^2$ with $\|\nabla^2 F\|_{\mathrm{op}} \le L$, Taylor's theorem with integral remainder gives, for every $h \in \mathbb{R}^d$,

$$F(\mu + h) = F(\mu) + \langle g, h \rangle + \tfrac{1}{2} h^\top H h + r(h), \qquad r(h) = \int_0^1 (1 - s)\, h^\top \big[\nabla^2 F(\mu + sh) - H\big] h \, ds, \tag{26}$$

so that, with $\omega(s) = \sup_{\|v\|_2 \le s} \|\nabla^2 F(\mu + v) - H\|_{\mathrm{op}}$,

$$|r(h)| \ \le \ \tfrac{1}{2} \|h\|_2^2 \, \omega(\|h\|_2), \qquad \omega \text{ nondecreasing}, \qquad \omega(s) \le 2L \ \forall s, \qquad \omega(s) \to 0 \text{ as } s \downarrow 0, \tag{27}$$

the last limit because $\nabla^2 F$ is continuous, hence uniformly continuous on the compact ball $\bar{B}(\mu, 1)$. Dropping the quadratic term in (26) instead gives the first-order form

$$\big|F(\mu + h) - F(\mu) - \langle g, h \rangle\big| \ \le \ \tfrac{L}{2} \|h\|_2^2. \tag{28}$$

B.1.2 PREPARATORY LEMMAS

Lemma B.1 (Top-two channel). *The pair $(z_{(1)}, z_{(2)})$ has density*

$$\begin{aligned} &p_K(m,g)\,dm\,dg \quad \textit{with} \\ &p_K(m,g) = K(K-1)\,\phi(m)\,\phi(m-g)\,\Phi(m-g)^{K-2}, \qquad g>0. \end{aligned} \tag{29}$$

and, independently of $(z_1,\dots,z_K)$, the pair $(X_1,X_2) := (x_{I_1}, x_{I_2})$ consists of two independent $\mathcal{N}(0,\Pi)$ vectors.

Proof. The density of the top two order statistics of K i.i.d. standard normals is $K(K-1)\phi(m_1)\phi(m_2)\Phi(m_2)^{K-2}$ on $\{m_1 > m_2\}$ (K choices for the maximal index, $K-1$ for the second, the remaining $K-2$ scores below m_2); substituting $m_1 = m$, $m_2 = m-g$ gives (29). For the second claim, the array $(x_k)_{k\le K}$ is independent of the array $(z_k)_{k\le K}$, with x_k i.i.d. $\mathcal{N}(0,\Pi)$, and (I_1,I_2) is a measurable function of $(z_k)_k$ alone. Hence for bounded measurable f,h,

$$\begin{aligned} \mathbb{E}\big[f(x_{I_1},x_{I_2})\,h(z_1,\dots,z_K)\big] &= \sum_{i_1\neq i_2} \mathbb{E}\big[f(x_{i_1},x_{i_2})\big] \\ &\quad\times \mathbb{E}\big[h(z)\,\mathbf{1}\{(I_1,I_2)=(i_1,i_2)\}\big] \\ &= \mathbb{E}\big[f(X_1',X_2')\big]\,\mathbb{E}\big[h(z)\big]. \end{aligned} \tag{30}$$

where (X_1',X_2') are independent $\mathcal{N}(0,\Pi)$, since $\mathbb{E}[f(x_{i_1},x_{i_2})]$ does not depend on the distinct pair (i_1,i_2). □

Lemma B.2 (Tie functional and Stein identity). *For continuously differentiable p with p, p' of polynomial growth, define*

$$\Gamma_K[p] = K(K-1)\int_{\mathbb{R}} p(m)\,\phi(m)^2\,\Phi(m)^{K-2}\,dm. \tag{31}$$

Then

$$\Gamma_K[p] = \mathbb{E}\big[M_K\,p(M_K)\big] - \mathbb{E}\big[p'(M_K)\big]. \tag{32}$$

In particular,

$$\Gamma_K[1] = c_K, \qquad \Gamma_K[m] = \mathbb{E}M_K^2 - 1, \qquad \Gamma_K[m^2] = \mathbb{E}M_K^3 - 2\,\mathbb{E}M_K = \gamma_K. \tag{33}$$

Proof. The density of M_K is $K\phi(m)\Phi(m)^{K-1}$, and $\frac{d}{dm}\big[\phi(m)\Phi(m)^{K-1}\big] = -m\phi(m)\Phi(m)^{K-1} + (K-1)\phi(m)^2\Phi(m)^{K-2}$. Multiplying by $K\,p(m)$ and integrating,

$$K\int_{\mathbb{R}} p(m)\,\frac{d}{dm}\big[\phi\,\Phi^{K-1}\big](m)\,dm = -\mathbb{E}\big[M_K p(M_K)\big] + \Gamma_K[p]. \tag{34}$$

Integrating the left side by parts, the boundary terms vanish by Gaussian decay against polynomial growth, and the left side equals $-K\int p'\,\phi\,\Phi^{K-1} = -\mathbb{E}[p'(M_K)]$, which is (32). Taking $p=1$, $p=m$, $p=m^2$ gives (33). □

Lemma B.3 (A priori bounds). *For every realization of the candidates, for every $\sigma>0$:*

$$\Delta \;\le\; \tfrac{L}{2}\,\sigma^2\big(\|\epsilon^{(J^*)}\|_2^2 + \|\epsilon^{(J)}\|_2^2\big) \;\le\; L\,\sigma^2 N, \tag{35}$$

where J^* is any maximizer of $F(y^{(k)})$. Moreover, for every $j\ge 2$,

$$\begin{aligned} D_j > 0 \implies z_{(1)} - z_{(j)} &< \frac{L\sigma}{2a}\big(\|\epsilon^{(I_j)}\|_2^2 + \|\epsilon^{(I_1)}\|_2^2\big) \\ &\le \frac{L\sigma N}{a}, \\ (D_j)_+ &\le L\sigma^2 N. \end{aligned} \tag{36}$$

Proof. By (28) applied to both terms, $\Delta = \sigma a(z_{J^*} - z_{(1)}) + \text{rem} \le \text{rem}$, since $z_{J^*}\le z_{(1)}$, and the remainder is bounded by the right side of (35). The same expansion gives $D_j \le -\sigma a(z_{(1)} - z_{(j)}) + \frac{L}{2}\sigma^2(\|\epsilon^{(I_j)}\|_2^2 + \|\epsilon^{(I_1)}\|_2^2)$, and (36) follows. □

Lemma B.4 (Auxiliary probability bounds). *(a) For every $\delta > 0$, $\Pr\big(z_{(1)} - z_{(3)} \le \delta\big) \le K^3\delta^2$. (b) For $\tau \ge 4d$, $\Pr(N \ge \tau) \le K\, e^{-\tau/8}$. (c) Let $\tau_\sigma = \log^2(1/\sigma)$ and $\mathcal{G}_\sigma = \{N \le \tau_\sigma\}$. Then for every $p \ge 1$ there is $\sigma_0(d, K, L, p) > 0$ with $\mathbb{E}[\Delta\, \mathbf{1}_{\mathcal{G}_\sigma^c}] \le \sigma^p$ for all $\sigma \le \sigma_0$.*

Proof. (a) On the event, with i the maximal index there are distinct $j, l \ne i$ with $|z_j - z_i| \le \delta$ and $|z_l - z_i| \le \delta$. For a fixed ordered triple of distinct indices, conditioning on z_i and using $\sup \phi = 1/\sqrt{2\pi}$,

$$\Pr\big(|z_j - z_i| \le \delta,\ |z_l - z_i| \le \delta\big) \le \big(2\delta/\sqrt{2\pi}\big)^2 \le \delta^2, \tag{37}$$

and a union bound over at most K^3 triples gives (a). (b) For one codeword, the chi-square Chernoff bound with parameter $1/4$ gives $\Pr(\|\epsilon\|_2^2 \ge \tau) \le 2^{d/2} e^{-\tau/4} \le e^{-\tau/8}$ for $\tau \ge 4d \ge 4d \log 2$; a union bound over K codewords gives (b). (c) By (35) and Cauchy–Schwarz,

$$\mathbb{E}\big[\Delta\, \mathbf{1}_{\mathcal{G}_\sigma^c}\big] \le L\sigma^2 \big(\mathbb{E}N^2\big)^{1/2} \Pr(\mathcal{G}_\sigma^c)^{1/2} \le L\sigma^2 \big(Kd(d+2)\big)^{1/2} K^{1/2} e^{-\tau_\sigma/16}, \tag{38}$$

valid once $\tau_\sigma \ge 4d$, and $e^{-\tau_\sigma/16} = \sigma^{\log(1/\sigma)/16}$ is eventually smaller than any fixed power of σ. □

Lemma B.5 (Reduction to the runner-up). *There is $C = C(K, L, a, d)$ such that for all small σ,*

$$0 \ \le\ \mathbb{E}[\Delta] - \mathbb{E}\big[(D_2)_+\big] \ \le\ C\, \sigma^4 \log^6(1/\sigma). \tag{39}$$

Proof. The lower bound is (25). For the upper bound, by (25) it suffices to bound $\sum_{j \ge 3} \mathbb{E}[(D_j)_+]$. Split on $\mathcal{G}_\sigma$. On $\mathcal{G}_\sigma$, (36) gives, for $j \ge 3$,

$$(D_j)_+ \ \le\ L\sigma^2 \tau_\sigma\, \mathbf{1}\Big\{z_{(1)} - z_{(3)} \le \frac{L\sigma\tau_\sigma}{a}\Big\}, \tag{40}$$

using $z_{(1)} - z_{(3)} \le z_{(1)} - z_{(j)}$. Summing over j and applying Lemma B.4(a) with $\delta = L\sigma\tau_\sigma/a$,

$$\mathbb{E}\Big[\sum_{j \ge 3} (D_j)_+\, \mathbf{1}_{\mathcal{G}_\sigma}\Big] \ \le\ K\, L\sigma^2\tau_\sigma \cdot K^3 \Big(\frac{L\sigma\tau_\sigma}{a}\Big)^2 = \frac{K^4 L^3}{a^2}\, \sigma^4 \tau_\sigma^3. \tag{41}$$

Off $\mathcal{G}_\sigma$, $\sum_{j \ge 3} (D_j)_+ \le K\Delta$ pointwise is not needed; simply $\sum_{j \ge 3} (D_j)_+ \le KL\sigma^2 N$ by (36), and the argument of Lemma B.4(c) bounds its expectation on $\mathcal{G}_\sigma^c$ by σ^5 for small σ. Collecting terms gives (39) with $\tau_\sigma^3 = \log^6(1/\sigma)$. □

B.1.3 The runner-up asymptotics

Proposition B.6. *As $\sigma \to 0$,*

$$\mathbb{E}\big[(D_2)_+\big] = \kappa\, \sigma^3 + o(\sigma^3), \qquad \kappa = \frac{1}{2a}\Big[\gamma_K A + \frac{c_K}{2} B\Big]. \tag{42}$$

Proof. Step 1: disintegration. By Lemma B.1, conditionally on $(z_{(1)}, z_{(2)}) = (m, m - g)$ the runner-up difference is

$$\Psi_\sigma(m, g; X_1, X_2) = F\big(\mu + \sigma[(m - g)w + X_2]\big) - F\big(\mu + \sigma[mw + X_1]\big), \tag{43}$$

with (X_1, X_2) i.i.d. $\mathcal{N}(0, \Pi)$ independent of (m, g), so by Tonelli

$$\mathbb{E}\big[(D_2)_+\big] = \int_{\mathbb{R}} \int_0^\infty \mathbb{E}_X\big[(\Psi_\sigma(m, g))_+\big]\, p_K(m, g)\, dg\, dm. \tag{44}$$

Step 2: the far region $g \ge 1$ contributes $O(\sigma^4)$. Let $S = \|mw + X_1\|_2^2 + \|(m - g)w + X_2\|_2^2$. By (28), $\Psi_\sigma \le -\sigma a g + \frac{L}{2}\sigma^2 S$, so on $\{g \ge 1\}$, $\Psi_\sigma > 0$ forces $S > 2a/(L\sigma)$, and then $(\Psi_\sigma)_+ \le \frac{L}{2}\sigma^2 S$. Hence, by Markov's inequality applied conditionally and integrated over (44),

$$\mathbb{E}\big[(D_2)_+ \mathbf{1}\{g \ge 1\}\big] \le \frac{L}{2}\sigma^2\, \mathbb{E}\Big[S\, \mathbf{1}\Big\{S > \frac{2a}{L\sigma}\Big\}\Big] \le \frac{L}{2}\sigma^2 \Big(\frac{L\sigma}{2a}\Big)^2 \mathbb{E}[S^3] = \frac{L^3 \mathbb{E}[S^3]}{8a^2}\, \sigma^4, \tag{45}$$

and $\mathbb{E}[S^3] < \infty$ since $z_{(1)}, z_{(2)}$ and the $\|X_i\|_2$ have moments of all orders.

Step 3: the near region, rescaled. On $\{g<1\}$ substitute $g=\sigma u/a$, $u\in(0,a/\sigma)$:

$$\mathbb{E}\big[(D_2)_+\mathbf{1}\{g<1\}\big]=\frac{\sigma^3}{a}\int_{\mathbb{R}}\int_0^{a/\sigma}\mathbb{E}_X\big[\sigma^{-2}(\Psi_\sigma)_+\big]\,p_K\big(m,\sigma u/a\big)\,du\,dm. \tag{46}$$

Expanding both terms of Ψ_σ with (26) and collecting (writing $s=\sigma u/a$),

$$\begin{aligned}\sigma^{-2}\Psi_\sigma = W-u-\frac{\sigma u}{a}\Big[\Big(m-\frac{\sigma u}{2a}\Big)h_\parallel+\langle b,X_2\rangle\Big]\\ +\sigma^{-2}\big[r\big(\sigma[(m-s)w+X_2]\big)-r\big(\sigma[mw+X_1]\big)\big].\end{aligned} \tag{47}$$

where the tie statistic is

$$W=W_m(X_1,X_2)=m\,\langle b,\,X_2-X_1\rangle+\tfrac12\big(X_2^\top H_\perp X_2-X_1^\top H_\perp X_1\big). \tag{48}$$

Indeed, with $\epsilon_1=mw+X_1$, $\epsilon_2=(m-s)w+X_2$ and $q_i=\epsilon_i^\top H\epsilon_i=z_i^2h_\parallel+2z_i\langle b,X_i\rangle+X_i^\top H_\perp X_i$, a direct computation gives $\frac12(q_2-q_1)=W-s\big[(m-\frac{s}{2})h_\parallel+\langle b,X_2\rangle\big]$, and the linear parts contribute $-\sigma^2u$. Denote the two error terms in (47) by e_1 and e_2; by (27) and $s\le1$,

$$\begin{aligned}&|e_1|\le\frac{\sigma u}{a}\,\Theta,\qquad \Theta:=(|m|+1)L+\|b\|_2\,\|X_2\|_2,\\ &|e_2|\le L\,S'',\qquad S'':=3m^2+2+\|X_1\|_2^2+\|X_2\|_2^2.\end{aligned} \tag{49}$$

and moreover, for fixed (m,u,X_1,X_2), $e_1\to0$ and $e_2\le\frac12 S''\,\omega\big(\sigma\sqrt{S''}\big)\to0$ as $\sigma\to0$.

Step 4: the large-Θ part of (46) is negligible. Split the inner expectation on $\{\sigma\Theta\le a/2\}$ and its complement. On $u<a/\sigma$, (47)–(49) give the crude bound $\sigma^{-2}(\Psi_\sigma)_+\le|W|+\Theta+LS''$, so by Cauchy–Schwarz there is a polynomial $c(m)$ with

$$\mathbb{E}_X\big[\sigma^{-2}(\Psi_\sigma)_+\,\mathbf{1}\{\sigma\Theta>a/2\}\big]\le c(m)\,\Pr\big(\Theta>a/(2\sigma)\big)^{1/2}. \tag{50}$$

If $|m|\le a/(4L\sigma)$ and $\sigma\le a/(8L)$, then $(|m|+1)L\le3a/(8\sigma)$. For $b\neq0$, $\Pr(\Theta>a/(2\sigma))\le\Pr\big(\|X_2\|_2>a/(8\sigma\|b\|_2)\big)\le e^{-c/\sigma^2}$ for small σ. If $b=0$, this probability is zero. For $|m|>a/(4L\sigma)$, the factor $\phi(m)$ in $p_K(m,\cdot)\le K^2\phi(m)/\sqrt{2\pi}$ is itself e^{-c/σ^2}. Since the u-range has length a/σ, the total contribution of this part to (46) is $O\big(\sigma^2e^{-c/\sigma^2}\big)\cdot\text{poly}=o(\sigma^p)$ for every p.

Step 5: dominated convergence on the main part. On $\{\sigma\Theta\le a/2\}$ we have $|e_1|\le u/2$, hence by (47),

$$\sigma^{-2}(\Psi_\sigma)_+\,\mathbf{1}\{\sigma\Theta\le a/2\}\;\le\;\big(V_m-u/2\big)_+,\qquad V_m:=|W_m|+L\,S''\;\ge0. \tag{51}$$

For fixed (m,u): as $\sigma\to0$ the indicator tends to 1 a.s., the errors tend to 0, and $x\mapsto(x)_+$ is 1-Lipschitz, so by dominated convergence in X (dominating function V_m, integrable),

$$\mathbb{E}_X\big[\sigma^{-2}(\Psi_\sigma)_+\,\mathbf{1}\{\sigma\Theta\le a/2\}\big]\;\longrightarrow\;\mathbb{E}_X\big[(W_m-u)_+\big]. \tag{52}$$

Furthermore $p_K(m,\sigma u/a)\to K(K-1)\phi(m)^2\Phi(m)^{K-2}$ pointwise, and uniformly in (g,σ),

$$p_K(m,g)\;\le\;\frac{K(K-1)}{\sqrt{2\pi}}\,\phi(m). \tag{53}$$

By (51) and (53), the integrand of the main part of (46) is dominated by $G(m,u)=\frac{K^2}{\sqrt{2\pi}}\phi(m)\,\mathbb{E}_X[(V_m-u/2)_+]$, and

$$\int_{\mathbb{R}}\int_0^\infty G(m,u)\,du\,dm=\frac{K^2}{\sqrt{2\pi}}\int_{\mathbb{R}}\phi(m)\,\mathbb{E}\big[V_m^2\big]\,dm<\infty, \tag{54}$$

using $\int_0^\infty(v-u/2)_+\,du=v^2$ for $v\ge0$ and $\mathbb{E}[V_m^2]\le c(1+m^4)$. Dominated convergence in (m,u) therefore gives

$$\begin{aligned}\frac{1}{\sigma^3}\,\mathbb{E}\big[(D_2)_+\mathbf{1}\{g<1\}\big]\;\longrightarrow\;\frac1a\int_{\mathbb{R}}K(K-1)\phi(m)^2\Phi(m)^{K-2}\\ \times\int_0^\infty\mathbb{E}\big[(W_m-u)_+\big]\,du\,dm.\end{aligned} \tag{55}$$

Step 6: evaluation of the limit. By Tonelli and $\int_0^\infty (x-u)_+\,du = x_+^2/2$,

$$\int_0^\infty \mathbb{E}\big[(W_m-u)_+\big]\,du = \tfrac{1}{2}\,\mathbb{E}\big[(W_m)_+^2\big]. \tag{56}$$

Under the swap $(X_1,X_2)\mapsto(X_2,X_1)$, whose law is invariant, $W_m\mapsto -W_m$; hence W_m is symmetric and $\mathbb{E}[(W_m)_+^2]=\frac{1}{2}\mathbb{E}[W_m^2]$. The two parts of (48) are uncorrelated (their product is an odd polynomial of a centered Gaussian), and, using $\Pi b=b$, $H_\perp\Pi=H_\perp$, and $\mathrm{Var}(x^\top Mx)=2\,\mathrm{tr}\big((M\Sigma)^2\big)$ for $x\sim\mathcal{N}(0,\Sigma)$,

$$\mathbb{E}\big[W_m^2\big] = m^2\cdot 2\,b^\top\Pi b + \tfrac{1}{4}\cdot 2\mathrm{Var}\big(X^\top H_\perp X\big) = 2m^2A+B. \tag{57}$$

Substituting into (55) and using Lemma B.2,

$$\begin{aligned}\lim_{\sigma\to0}\frac{\mathbb{E}[(D_2)_+\mathbf{1}\{g<1\}]}{\sigma^3} &= \frac{1}{4a}\,\Gamma_K\big[2m^2A+B\big] = \frac{1}{4a}\big(2A\gamma_K+Bc_K\big)\\ &= \frac{1}{2a}\Big[\gamma_K A+\frac{c_K}{2}B\Big].\end{aligned} \tag{58}$$

Together with Step 2 and Step 4 this proves (42). □

Proof of Theorem 4.1 (asymptotic part). Combine Lemma B.5 and Proposition B.6:

$$\mathbb{E}[\Delta] = \mathbb{E}[(D_2)_+] + O\big(\sigma^4\log^6(1/\sigma)\big) = \kappa\sigma^3 + o(\sigma^3). \tag{59}$$

□

B.1.4 Perturbed scoring directions

The rule uses the direction $w_\sigma\propto w+(\sigma/a)\psi$ with a fixed $\psi\perp w$ (a component of ψ along w only rescales w_σ and leaves the argmax unchanged). Its scores are, up to the common positive factor,

$$s_k = z_k+\frac{\sigma}{a}\,\langle\psi,x_k\rangle, \qquad J_\psi=\arg\max_{k\le K} s_k. \tag{60}$$

Write $\Delta^\psi=\max_k F(y^{(k)})-F(y^{(J_\psi)})$.

Localization. From $s_{J_\psi}\ge s_{I_1}$,

$$z_{(1)}-z_{J_\psi} \;\le\; \frac{2\sigma}{a}\,\max_k|\langle\psi,x_k\rangle| \;\le\; \frac{2\sigma\|\psi\|_2}{a}\,N^{1/2}, \tag{61}$$

so the analogue of Lemma B.3 holds with the constant enlarged by $\|\psi\|_2$: pointwise $\Delta^\psi\le\sigma a(z_{(1)}-z_{J_\psi})+L\sigma^2N\le C_\psi\sigma^2(1+N)$, and the winner J_ψ, as well as any F-maximizer, lies among the top-two score ranks except on the event $\{z_{(1)}-z_{(3)}\le C_\psi\sigma\tau_\sigma/a\}$ once restricted to $\mathcal{G}_\sigma$. Lemma B.4 and the argument of Lemma B.5 then reduce the claim, at cost $O(\sigma^4\log^6(1/\sigma))$, to the contribution of the top two ranks,

$$\Delta_{12}^{\psi} := (\Psi_\sigma)_+\,\mathbf{1}\{s_{I_2}\le s_{I_1}\}+(-\Psi_\sigma)_+\,\mathbf{1}\{s_{I_2}>s_{I_1}\}, \tag{62}$$

which records the loss when the ψ-rule keeps rank one (first term: it loses if rank two was better) or promotes rank two (second term: it loses if rank one was better).

The disagreement indicator is exact in the rescaled variables. With $g=\sigma u/a$ as before,

$$s_{I_2}-s_{I_1} = -g+\frac{\sigma}{a}\langle\psi,X_2-X_1\rangle = \frac{\sigma}{a}\,(P-u), \qquad P:=\langle\psi,\,X_2-X_1\rangle, \tag{63}$$

so $\{s_{I_2}>s_{I_1}\}=\{P>u\}$ with no error term. The far region and the large-Θ region are handled exactly as in Steps 2 and 4 of Proposition B.6 (the extra term in (61) only enlarges constants). On the main part, (52) holds for each branch of (62), with limits $(W_m-u)_+\mathbf{1}\{P\le u\}$ and $(u-W_m)_+\mathbf{1}\{P>u\}$ respectively; the second branch is dominated, on $\{\sigma\Theta\le a/2,\ u<a/\sigma\}$, by $(u+V_m)\mathbf{1}\{P>u\}$, whose du-integral has expectation bounded by a polynomial in m times

a Gaussian factor, since P is $\mathcal{N}(0, 2\|\psi\|_2^2)$; dominated convergence applies as before. The limit of $\sigma^{-3}\mathbb{E}[\Delta^\psi]$ is therefore

$$\begin{aligned}&\frac{1}{a}\int_{\mathbb{R}} K(K-1)\phi(m)^2\Phi(m)^{K-2}\,\mathbb{E}\big[T_m\big]\,dm,\\ &T_m := \int_0^\infty \Big[(W_m-u)_+\mathbf{1}\{P\le u\}\\ &\qquad\qquad + (u-W_m)_+\mathbf{1}\{P>u\}\Big]du.\end{aligned} \tag{64}$$

Evaluation of $\mathbb{E}[T_m]$. Using $(x)_+ - (-x)_+ = x$,

$$\begin{aligned}T_m &= \int_0^\infty (W_m-u)_+\,du + \int_0^{P_+}\big[(u-W_m)_+ - (W_m-u)_+\big]\,du\\ &= \frac{(W_m)_+^2}{2} + \frac{P_+^2}{2} - W_m P_+.\end{aligned} \tag{65}$$

Take expectations term by term, conditionally on m. First, $\mathbb{E}[(W_m)_+^2] = \frac{1}{2}\mathbb{E}[W_m^2] = m^2A + \frac{B}{2}$ by (57) and symmetry. Second, $P \sim \mathcal{N}(0, 2\|\psi\|_2^2)$ is symmetric, so $\mathbb{E}[P_+^2] = \frac{1}{2}\mathbb{E}[P^2] = \|\psi\|_2^2$. Third, split $W_m = W^{\mathrm{g}} + W^{\mathrm{q}}$ into its Gaussian part $W^{\mathrm{g}} = m\langle b, X_2 - X_1\rangle$ and quadratic part W^{q}. The swap $(X_1, X_2) \mapsto (X_2, X_1)$ sends $(W^{\mathrm{q}}, P) \mapsto (-W^{\mathrm{q}}, -P)$ and preserves the law, so $\mathbb{E}[W^{\mathrm{q}}P_+] = -\mathbb{E}[W^{\mathrm{q}}P_-]$, while $\mathbb{E}[W^{\mathrm{q}}P] = 0$ (an odd Gaussian moment); adding, $\mathbb{E}[W^{\mathrm{q}}P_+] = 0$. For the jointly Gaussian pair (W^{g}, P), Stein's lemma with $f(P) = P_+$, $f' = \mathbf{1}\{P > 0\}$ gives

$$\mathbb{E}\big[W^{\mathrm{g}}P_+\big] = \mathrm{Cov}\big(W^{\mathrm{g}}, P\big)\,\mathbb{E}\big[\mathbf{1}\{P>0\}\big] = 2m\,\langle b,\psi\rangle\cdot\tfrac{1}{2} = m\,\langle b,\psi\rangle. \tag{66}$$

Collecting,

$$\mathbb{E}\big[T_m\big] = \frac{m^2A}{2} + \frac{B}{4} + \frac{\|\psi\|_2^2}{2} - m\,\langle b,\psi\rangle = \frac{\|m\,b-\psi\|_2^2}{2} + \frac{B}{4}. \tag{67}$$

Integration. Substituting (67) into (64) and applying Lemma B.2 coordinatewise,

$$\begin{aligned}\kappa(\psi) &= \frac{1}{2a}\Big[\Gamma_K\big[\|m\,b-\psi\|_2^2\big] + \frac{c_K}{2}B\Big]\\ &= \frac{1}{2a}\Big[A\,\gamma_K - 2\langle b,\psi\rangle\big(\mathbb{E}M_K^2 - 1\big)\\ &\qquad + c_K\|\psi\|_2^2 + \frac{c_K}{2}B\Big].\end{aligned} \tag{68}$$

Completing the square with $\bar{m}_K = (\mathbb{E}M_K^2 - 1)/c_K$,

$$\kappa(\psi) = \frac{1}{2a}\Big[c_K\,\|\psi - \bar{m}_K\,b\|_2^2 + \big(\gamma_K - c_K\bar{m}_K^2\big)A + \frac{c_K}{2}B\Big], \tag{69}$$

where $b = \Pi H w$. Setting $\psi = 0$ recovers (12), and $\psi = c_K\,b$ gives the coupling coefficient $\Gamma_K[(m-c_K)^2] = \mathbb{E}\big[M_K(M_K-c_K)^2\big] = \tilde{\gamma}_K$. □

B.2 The scoring gap of a matched mean shift

This appendix derives the leading-order scoring gap of a mean-shift emission charged against exact scoring of a candidate set it does not use. The quantity is the one measured by the mean-shift rows of the regret experiment; unlike the regret of a rule that outputs a candidate it is signed, and its expansion isolates the longitudinal variance mismatch between the two families.

B.2.1 Setting

Fix the step, condition on x_t, and drop the index t. As in Appendix B.1, write $g = \nabla F(\mu) \neq 0$, $a = \|g\|_2$, $w = g/a$, $H = \nabla^2 F(\mu)$, $\Pi = I_d - ww^\top$, $h_\parallel = w^\top H w$, and $H_\perp = \Pi H \Pi$. Retain the assumptions of Theorem 4.1: $F \in C^2(\mathbb{R}^d)$, $\|\nabla^2 F(y)\|_{\mathrm{op}} \le L$ for every y, and $g \neq 0$. In addition, assume that F is three times continuously differentiable in a neighborhood of μ. Choose $r_0 > 0$ so that $\|\nabla^3 F\|_{\mathrm{op}} \le L_3$ on $\bar{B}(\mu, r_0)$. The state, proxy, K, d, and ρ are fixed in the limit $\sigma \to 0$.

Two independent sources of randomness:

$$\epsilon^{(1)}, \dots, \epsilon^{(K)} \overset{\text{iid}}{\sim} \mathcal{N}(0, I_d), \qquad \epsilon \sim \mathcal{N}(0, I_d) \text{ independent of the candidates.} \tag{70}$$

The comparator selects the highest-scoring candidate,

$$J^{\text{ex}} = \arg\max_{k \le K} F\big(\mu + \sigma\epsilon^{(k)}\big), \tag{71}$$

and the mean-shift rule at strength $\rho \ge 0$ emits

$$y_S = \mu + \sigma\,\epsilon_S, \qquad \epsilon_S = \rho\, w + \epsilon. \tag{72}$$

The scoring gap is

$$\Delta_{\text{FD}} = F\big(\mu + \sigma\epsilon^{(J^{\text{ex}})}\big) - F\big(y_S\big), \tag{73}$$

a signed quantity: y_S is not one of the candidates, and its score can exceed the best candidate score.

Proposition B.7 (Mean-shift scoring gap)**.** *Under the assumptions above, as $\sigma \to 0$ with K, d, and ρ fixed,*

$$\mathbb{E}\big[\Delta_{\text{FD}} \mid x_t\big] = \sigma a\,(c_K - \rho) + \frac{\sigma^2}{2}\, h_{\|}\Big(\mathbb{E}M_K^2 - \rho^2 - 1\Big) + O(\sigma^3). \tag{74}$$

In particular, at the matched strength $\rho = c_K$,

$$\mathbb{E}\big[\Delta_{\text{FD}} \mid x_t\big] = \frac{\sigma^2}{2}\, h_{\|}\,(\nu_K - 1) + O(\sigma^3), \tag{75}$$

For $K \ge 2$ and $h_{\|} \ne 0$, this gap has sign $-\operatorname{sign}(h_{\|})$ for sufficiently small σ, since $\nu_K < 1$.

B.2.2 Proof

Step 1: reduce exact scoring to linear selection. Let $J = \arg\max_k \langle w, \epsilon^{(k)}\rangle$ be the linear winner, and write $y^{\text{lin}} = \mu + \sigma\epsilon^{(J)}$ and $y^{\text{ex}} = \mu + \sigma\epsilon^{(J^{\text{ex}})}$. Theorem 4.1 gives

$$\mathbb{E}\big[F(y^{\text{ex}}) - F(y^{\text{lin}})\big] = \kappa\sigma^3 + o(\sigma^3) = O(\sigma^3). \tag{76}$$

Consequently,

$$\mathbb{E}\Delta_{\text{FD}} = \mathbb{E}\big[F(y^{\text{lin}}) - F(y_S)\big] + O(\sigma^3). \tag{77}$$

Only the moments of the linear winner and the mean shift are needed.

Step 2: control the expected Taylor remainder. Write

$$R(h) = F(\mu + h) - F(\mu) - \langle g, h\rangle - \tfrac{1}{2}h^\top H h.$$

The local third-derivative bound gives $|R(h)| \le L_3\|h\|_2^3/6$ for $\|h\|_2 \le r_0$. The global Hessian bound also gives $|R(h)| \le L\|h\|_2^2$ for every h. For either $\eta = \epsilon^{(J)}$ or $\eta = \epsilon_S = \rho w + \epsilon$,

$$\begin{aligned}\mathbb{E}|R(\sigma\eta)| &\le \frac{L_3}{6}\sigma^3\mathbb{E}\|\eta\|_2^3 + L\sigma^2\mathbb{E}\big[\|\eta\|_2^2\mathbf{1}_{\{\|\eta\|_2 > r_0/\sigma\}}\big] \\ &= O(\sigma^3).\end{aligned} \tag{78}$$

Indeed, $\|\epsilon^{(J)}\|_2 \le \max_{k\le K}\|\epsilon^{(k)}\|_2$, and ϵ_S is a Gaussian with a fixed mean. Both have Gaussian tails for fixed K, d, and ρ, so the tail term decays faster than any fixed power of σ. Thus, for these two choices of η,

$$\mathbb{E}\big[F(\mu + \sigma\eta)\big] = F(\mu) + \sigma\mathbb{E}\langle g, \eta\rangle + \frac{\sigma^2}{2}\mathbb{E}\big[\eta^\top H\eta\big] + O(\sigma^3). \tag{79}$$

Step 3: compute the two sets of moments. Write $q(\eta) = \eta^\top H\eta$ and $\epsilon = Zw + \epsilon_\perp$, where $Z \sim \mathcal{N}(0,1)$ and $\epsilon_\perp \sim \mathcal{N}(0, \Pi)$ are independent. Then $\epsilon_S = (\rho + Z)w + \epsilon_\perp$, so

$$\mathbb{E}\langle g, \epsilon_S\rangle = a\rho, \qquad \mathbb{E}q(\epsilon_S) = h_{\|}(\rho^2 + 1) + \operatorname{tr}(H_\perp). \tag{80}$$

The cross term vanishes by independence and centering, and $\mathbb{E}[\epsilon_\perp^\top H\epsilon_\perp] = \operatorname{tr}(H\Pi) = \operatorname{tr}(H_\perp)$. By Proposition 4.2, the linear winner satisfies $\epsilon^{(J)} = M_K w + \xi$ in distribution, with $\xi \sim \mathcal{N}(0, \Pi)$ independent of M_K. Hence

$$\mathbb{E}\langle g, \epsilon^{(J)}\rangle = ac_K, \qquad \mathbb{E}q(\epsilon^{(J)}) = h_{\|}\mathbb{E}M_K^2 + \operatorname{tr}(H_\perp). \tag{81}$$

Step 4: collect the terms. Apply (79) to the two terms in (77) and substitute (80) and (81). The transverse traces cancel between the linear winner and the mean shift, giving

$$\mathbb{E}\Delta_{\mathrm{FD}} = \sigma a(c_K - \rho) + \frac{\sigma^2}{2} h_{\|}\left(\mathbb{E}M_K^2 - \rho^2 - 1\right) + O(\sigma^3). \tag{82}$$

At $\rho = c_K$, the identity $\mathbb{E}M_K^2 - c_K^2 = \nu_K$ gives (75). For $K \geq 2$ and $h_{\|} \neq 0$, $\nu_K < 1$ makes the nonzero quadratic term determine the sign for sufficiently small σ. □

C Ablations and additional experiments

The following studies supplement the main comparisons with MNIST diagnostics, qualitative examples, and backbone-specific ablations. The signed and unsigned 50-step sweeps use different selection rules and are reported separately.

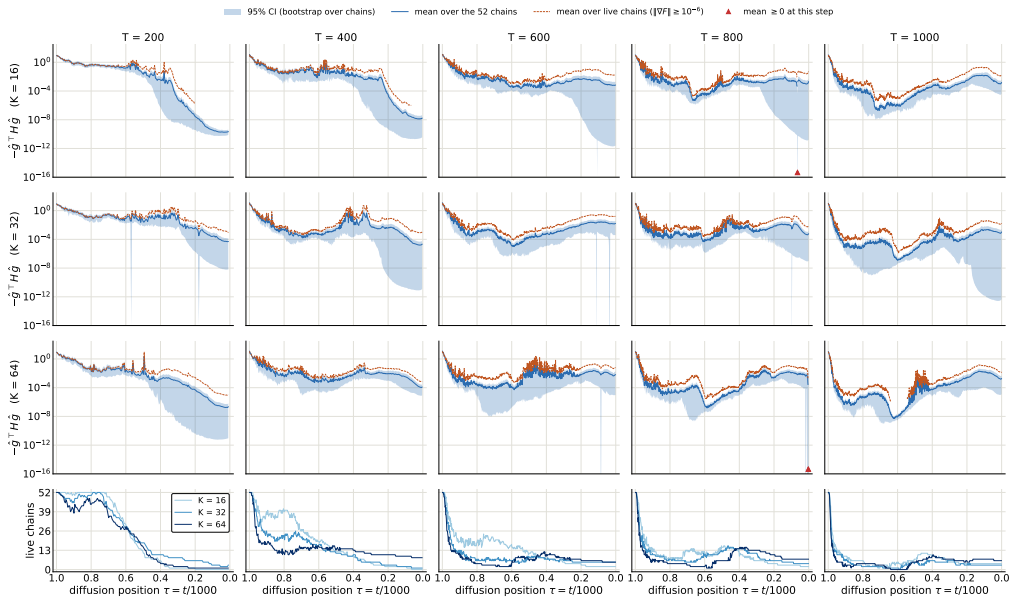


Figure 4: $h_{||}$ measured on MNIST.

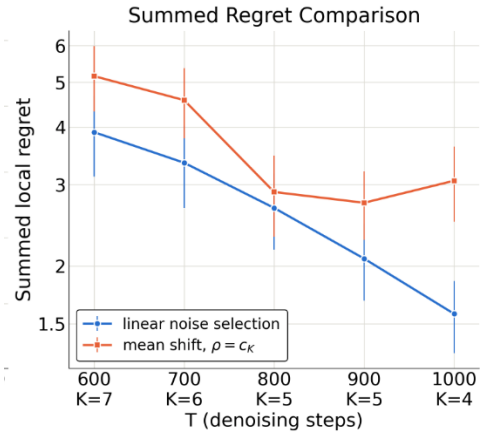


Figure 5: Summed regret comparison with $\sqrt{T}c_K = \text{const.}$.

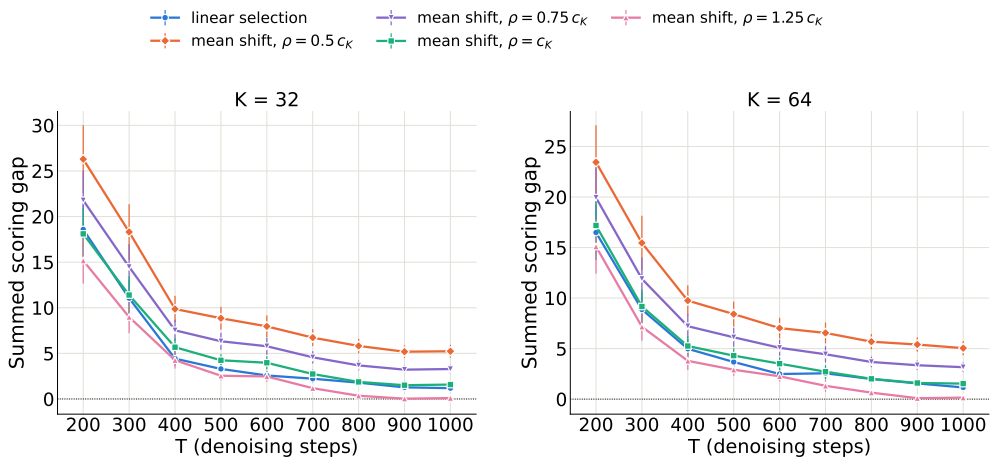


Figure 6: Summed regret comparison on different strength parameter of mean shift method.

	M_2=0	M_2=1			
	–	λ=0.5, β=0	λ=0.5, β=0.5	λ=1, β=0	λ=1, β=0.5
M=2	0.600	0.588	**0.662**	0.613	0.625
M=3	0.613	0.662	**0.700**	0.675	0.613

Table 6: Effect of the additional candidate noises on GenEval counting (official accuracy, all 80 prompts, one seed), P=20 candidates per step, T=50. M_2=0 is plain top-M recombination (λ and β have no effect without these additional candidates). For reference on the same images: SVDD-PM 0.625, TreeG-SC 0.625, no guidance 0.388.

C.1 QUALITATIVE EXAMPLES

Figures 7 and 8 show example outputs from SVDD-PM and our signed top-4 variant. Both methods use Stable Diffusion 1.5 with 200 denoising steps. SVDD-PM evaluates eight candidates at each stochastic step. Our signed top-4 variant selects from 16,384 noises every second step. The numbers printed below the images are PickScore-v1 logits. For the same image-text embedding pair, each logit is 98.8645 times its unscaled cosine. The quantitative tables report cosine scores.

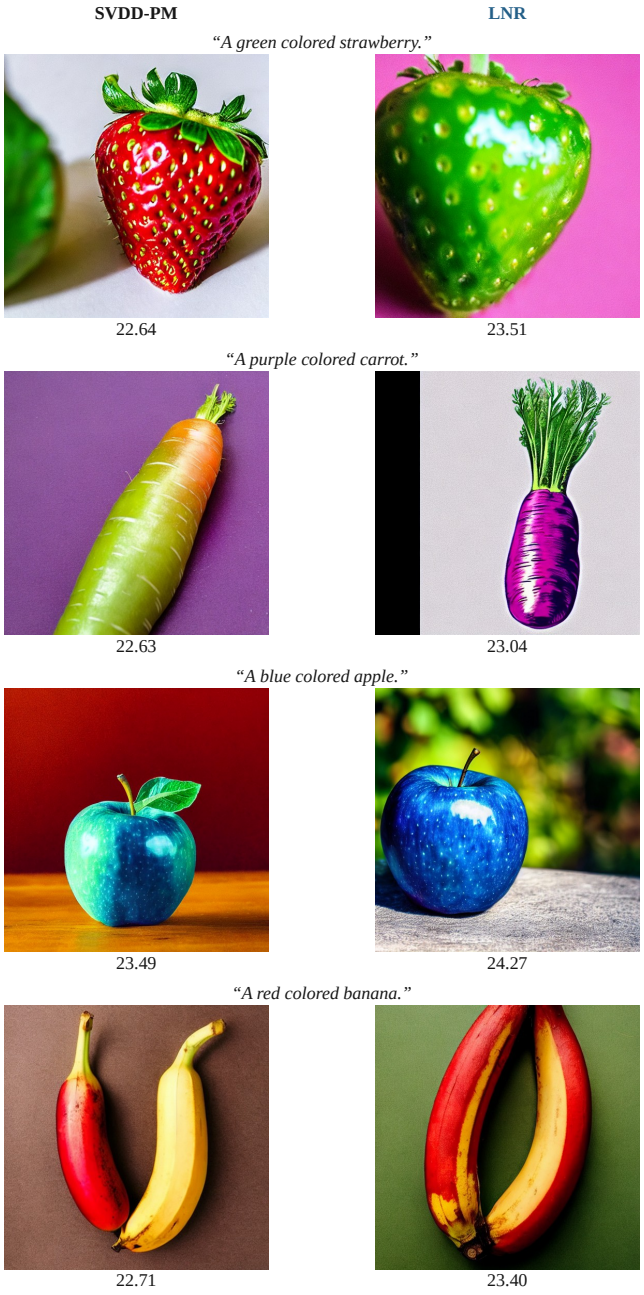


Figure 7: Example outputs from SVDD-PM and our signed top-4 variant. Printed scores are PickScore-v1 logits.

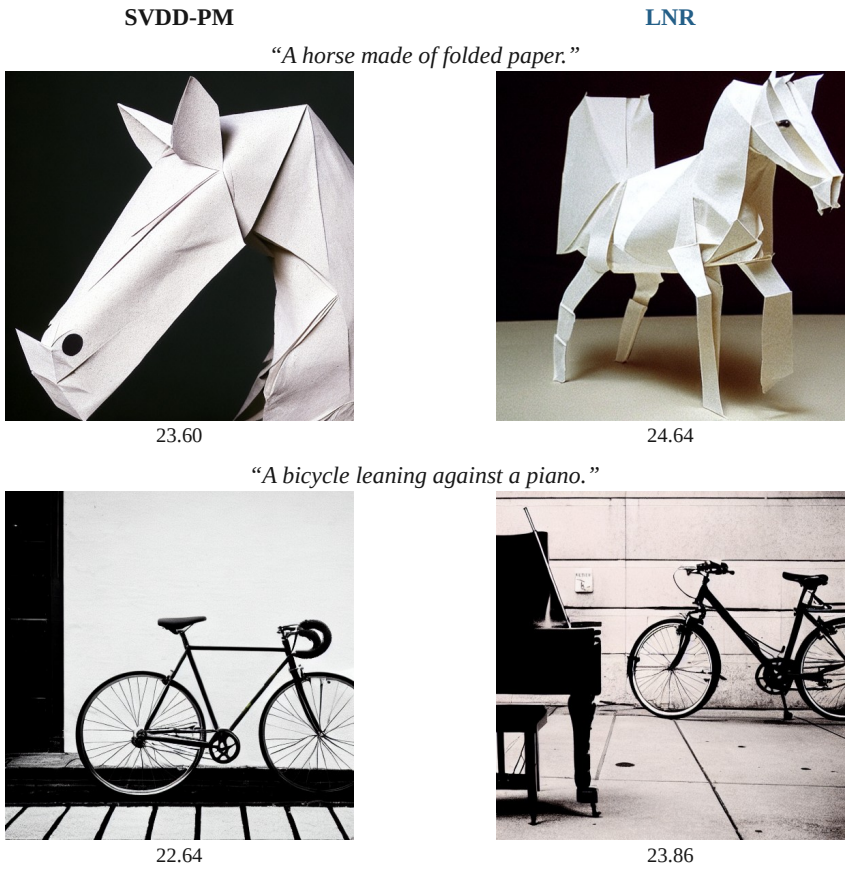


Figure 8: Additional example outputs from SVDD-PM and our signed top-4 variant. Printed scores are PickScore-v1 logits.

C.2 Full metrics from the pilot

Empirically normalized signed variants. The signed rows use the empirical normalization described in Appendix A.1. We evaluate this variant empirically and do not apply the Gaussian finite-selection certificate to these rows.

Table 7: Full metrics for the pilot comparison. PickScore is converted from the recorded logits to the cosine scale used in Table 2. Uncertainty intervals are not reported for this ten-prompt evaluation with one output per prompt and method. Signed rows use empirical-standard-deviation normalization and are not covered by the Gaussian certificate.

Method	Pick cos.↑	HPS↑	IR↑	MUSIQ↑	ARNIQA↑	TOPIQ↑
Base	0.2230	.280	−.20	69.7	.681	.625
LNR (argmax, every step)	0.2469	.302	+.76	70.0	.704	.623
LNR (top-16, stride 2)	0.2549	.306	+.35	71.2	.704	.618
LNR (signed top-16, stride 2)	0.2557	.305	+.75	71.2	.698	.637
LNR (signed top-30, stride 2)	0.2598	.307	+.31	70.9	.699	.648
LNR (top-16, every step)	0.2674	.303	+.33	69.8	.645	.603
SVDD-PM ($K = 4$)	0.2391	.296	+.38	72.2	.704	.631
SVDD-PM ($K = 8$)	0.2409	.303	+.76	71.4	.706	.640
DAS (native)	0.2518	.309	+.40	72.0	.713	.607
DAS (half temperature)	0.2686	.320	+1.02	70.0	.672	.561

C.3 SDXL AND FLUX EXPERIMENTS

These completed comparisons use 50 shared prompts, 50 denoising steps, and 1024×1024 images. We report mean quality scores and median generation times. PickScore is reported on the cosine scale. FreeDoM provides a mean-shift guidance baseline, and DAS uses mean-shifted proposals with particle resampling. The FLUX experiments use FLUX.1-dev.

C.3.1 SDXL

SDXL (Podell et al., 2024) uses guidance scale 5.0. Table 8 compares methods at the recorded settings. LNR rows use $K = 16{,}384$; the corrected row uses two refinement updates.

Table 8: SDXL method comparison. Median time includes prompt conditioning.

Method	Time (s)↓	Pick↑	HPSv2↑	IR↑	TOPIQ↑
Base	7.8	0.22820	0.288	0.779	0.639
FreeDoM	77.3	0.23526	0.295	0.901	0.657
DAS ($N = 4$)	341.6	0.24016	0.297	0.873	0.679
NoiseTilt ($\rho = 0.3$)	76.7	0.25930	0.287	-0.483	0.691
SVDD-PM ($N = 8$)	173.8	0.23668	0.294	0.900	0.660
LNR ($M = 6$)	71.7	0.23792	0.296	0.823	0.664
LNR (corrected, $M = 6$)	199.6	0.24247	0.299	0.804	0.657
LNR ($M = 8$)	71.8	0.23838	0.296	0.859	0.666

C.3.2 SDXL: COMPLETED COMPONENT ABLATIONS

The prompt set, model, reward, and generation settings match Table 8. Table 9 varies one selection parameter at a time. The K sweep fixes $M = 8$, and the M sweep fixes $K = 16{,}384$.

Table 9: SDXL LNR ablations. Time includes prompt conditioning.

K sweep ($M = 8$)			M sweep ($K = 16{,}384$)		
K	Pick	Time (s)	M	Pick	Time (s)
1,024	0.23677	71.4	1	0.23169	71.8
4,096	0.23849	71.5	4	0.23641	71.7
16,384	0.23838	71.8	8	0.23838	71.8
65,536	0.23930	73.2	16	0.23941	71.9
			32	0.24030	71.7
			64	0.24027	71.8

LNR with $M = 32$ reaches PickScore 0.24030 in 71.7 s, compared with 0.23668 in 173.8 s for eight-candidate SVDD-PM. Increasing M to 64 does not improve PickScore and lowers ImageReward from 0.785 to 0.421.

The two settings in Table 10 change both K and M, so they are separate from the one-factor sweeps. The one-update corrected setting at $K = 16{,}384$, $M = 8$ appears only in Table 11.

Table 10: SDXL joint settings ($K = 65{,}536$, $M = 32$).

Method	Pick↑	HPSv2↑	IR↑	TOPIQ↑
LNR	0.24281	0.298	0.650	0.641
LNR, 2 correction updates	0.25147	0.302	0.679	0.631

Table 11: SDXL correction and selection frequency. Times use the same measurement protocol.

Setting	Time (s)↓	Pick↑	IR↑	TOPIQ↑
Per-candidate scoring				
SVDD-PM ($N = 2$)	49.1	0.23123	0.799	0.659
SVDD-PM ($N = 4$)	90.6	0.23533	0.884	0.655
SVDD-PM ($N = 8$)	173.8	0.23668	0.900	0.660
LNR, $K = 16{,}384$ *and* $M = 8$				
Linear	71.8	0.23838	0.859	0.666
Corrected, 1 update	135.8	0.24298	0.892	0.668
Corrected, 2 updates	199.9	0.24455	0.890	0.652
Linear, every second step	40.5	0.23310	0.804	0.659
Corrected, 1 update, every second step	73.0	0.23530	0.851	0.658

At fixed K and M, one and two correction updates raise PickScore from 0.23838 to 0.24298 and 0.24455, with additional computation. Guiding every second step instead reduces runtime, but also reduces PickScore. Update counts denote the configured maximum, with the original stopping rule retained.

C.3.3 FLUX: MATCHED COMPARISON AND ABLATIONS

The completed rerun uses the same 50 prompts, 50 steps, and guidance scale 3.5. Every row below uses the same deterministic reward-backward implementation, including its CPU resizing gradient in the measured time. Prompt conditioning is excluded. FreeDoM and DAS retain the flow adaptations described above.

Table 12: FLUX method comparison under the matched numerical protocol.

Setting	Time (s)↓	Pick↑	HPSv2↑	IR↑	TOPIQ↑
Base	30.5	0.23252	0.298	0.878	0.679
FreeDoM	339.9	0.23389	0.297	0.905	0.648
DAS ($N = 4$)	1047.7	0.23330	0.299	0.966	0.703
NoiseTilt ($\rho = 0.3$)	280.1	0.27273	0.319	0.955	0.658
SVDD-PM ($N = 2$)	110.3	0.23259	0.298	0.867	0.677
SVDD-PM ($N = 4$)	189.9	0.23278	0.299	0.867	0.680
SVDD-PM ($N = 8$)	349.8	0.23259	0.298	0.884	0.676
SVDD-PM ($N = 16$)	670.5	0.23330	0.299	0.872	0.680
LNR, $K = 16{,}384$ *and* $M = 512$					
Linear	299.3	0.24011	0.302	0.877	0.673
Corrected, 1 update	558.6	0.24335	0.303	0.896	0.669
Corrected, 2 updates	816.3	0.24500	0.305	0.901	0.672

At $M = 512$, two correction updates raise PickScore from 0.24011 to 0.24500, while median time rises from 299.3 to 816.3 s. NoiseTilt has the highest PickScore in this comparison.

Table 13 reports the one-factor FLUX sweeps under this matched protocol. The K sweep fixes $M = 512$, and the M sweep fixes $K = 16{,}384$. Both use LNR without correction.

Table 13: FLUX LNR ablations under the matched protocol.

K sweep ($M = 512$)			M sweep ($K = 16{,}384$)		
K	Pick	Time (s)	M	Pick	Time (s)
1,024	0.23701	289.7	1	0.23313	293.0
4,096	0.23936	291.6	8	0.23432	292.9
16,384	0.24011	299.3	32	0.23589	293.3
65,536	0.24144	328.7	128	0.23777	294.5
			512	0.24011	299.3
			1,024	0.24209	304.3
			2,048	0.24473	317.0

At $M = 2{,}048$, LNR reaches PickScore 0.24473 in 317.0 s, versus 0.23259 in 349.8 s for eight-candidate SVDD-PM.

Neural work versus elapsed time. Every uncorrected LNR setting for FLUX above uses 99 denoiser forward passes and 49 backward passes per image. Raising K from $1{,}024$ to $65{,}536$ at fixed $M = 512$ leaves these counts unchanged, while runtime rises from 289.7 to 328.7 s. Thus neural scoring is independent of K, but total runtime is not. At fixed $K = 16{,}384$ and $M = 512$, one and two correction updates increase the counts to 148/98 and 197/147 forward/backward passes, respectively.

C.4 Stable Diffusion at 50 steps

We vary the candidate-pool size on Stable Diffusion 1.5 while fixing $M = 1$. Each setting uses 50 DDIM steps, guidance scale 5.0, and one 512×512 image for each of the same 50 prompts. Both variants retain the SD implementation's signed selection and empirical noise normalization. The corrected variant uses up to two damped direction updates, with the other correction settings unchanged. At $K = 4, 8, 16, 64$, correction raises PickScore at every tested setting, but median time rises from 13.8 to 37.2 s per image. At $K = 64$, LNR and corrected LNR score 0.22490 and 0.22858, respectively. Table 14 gives SVDD-PM references under the same prompts and generation settings. This signed, empirically normalized pilot uses a different selection rule from the unsigned single-winner experiment in Figure 9.

Table 14: SVDD-PM references for the signed-selection experiment.

Candidates	Time (s)↓	Pick↑	IR↑	TOPIQ↑
$N = 16$	96.5	0.23361	0.367	0.658
$N = 64$	377.1	0.23799	0.445	0.667

Both SVDD-PM settings obtain higher PickScore than these small-pool signed settings, at median times of 96.5 and 377.1 s. The pilot is reported separately from the 200-step comparison and the unsigned single-winner study.

LAGS at 50 steps. We evaluate LAGS with $c = 10$ on the same 50 prompts, using four seeds per prompt for 200 images in total. The model, resolution, guidance scale, and DDIM settings are unchanged. LAGS uses FP32 for its finite-difference guidance, whereas the signed-selection sweep above uses BF16 autocast. Its timing is therefore not a matched-precision comparison. LAGS obtains a PickScore of 0.21839, HPSv2 of 0.279, ImageReward of 0.021, and TOPIQ of 0.699, with a median generation time of 25.4 s per image on an A100-80GB.

C.5 Reward alignment and computational cost at 50 steps

C.5.1 Comparison with the image-generation baselines

Protocol and evaluation. We test how much candidate selection improves reward under a common 50-step schedule, and how much computation that improvement requires. Every method uses Stable Diffusion 1.5 at 512×512, DDIM with $\eta = 1$, guidance scale 5, and a deterministic final step. We use DAS's 50 prompts, with four samples per prompt and all 200 outputs included in each mean. As in Table 2, we optimize PickScore-v1 and report its unscaled cosine similarity. HPSv2, ImageReward (IR), and CLIP-L/14 cosine measure held-out preference or alignment. MUSIQ, ARNIQA, and TOPIQ measure image quality without the prompt. Evaluation uses the saved final images.

Scoring and selection. SVDD-PM predicts a clean image and evaluates its reward separately for each candidate next state. Our linear rule differentiates the same proxy once at the transition mean, through the denoiser, decoder, and reward model. Unless varied, both our rows use $K = 4096$ and unsigned single-winner selection ($m = 1$) at all 49 stochastic steps, rather than the top-M aggregation in the earlier tables. We sample the linear winner directly from its exact distribution. The corrected variant adds two gradient refinements with damping $1/2$, without Anderson acceleration or restarts.

Baselines. We compare against the unsteered model, SVDD-PM with four or eight candidates, and gradient or particle guidance. FreeDoM uses three repeat updates in its time-travel window. DAS uses four particles, exponential tempering, an effective-sample-size threshold of $N/2$, and reward temperature 10^{-4}. LAGS uses $c = 10$ without APG. For the SD/DDIM adaptation of NoiseTilt, we report three noise-tilt strengths. These are common-backbone comparisons, not reproductions of each method's original benchmark.

Table 15: Reward alignment on SD 1.5 at 50 steps. All rows use the same 200-image evaluation set. Higher is better for every metric.

Method	PickScore	HPSv2	IR	CLIP cos.	MUSIQ	ARNIQA	TOPIQ
Unsteered							
Stable Diffusion 1.5	0.2183	0.2792	0.026	0.2538	73.1	0.737	0.699
Mean-shift and particle guidance							
FreeDoM ($r = 3$)	0.2307	0.2893	0.376	0.2560	72.5	0.735	0.672
LAGS† ($c = 10$)	0.2182	0.2792	0.020	0.2538	73.1	0.737	0.698
DAS ($N = 4$)	0.2354	0.2927	0.459	0.2577	73.4	0.735	0.680
NoiseTilt ($\rho = 0.3$)	0.2508	0.2824	-0.828	0.2349	65.2	0.674	0.530
NoiseTilt ($\rho = 0.1$)	0.2757	0.3058	0.046	0.2519	67.8	0.697	0.542
NoiseTilt ($\rho = 0.03$)	0.2690	0.3069	0.386	0.2560	71.0	0.722	0.623
Candidate selection							
SVDD-PM ($N = 4$)	0.2279	0.2869	0.354	0.2583	73.2	0.731	0.674
SVDD-PM ($N = 8$)	0.2323	0.2897	0.454	0.2594	73.4	0.733	0.677
LNR	0.2294	0.2876	0.264	0.2539	72.8	0.729	0.686
LNR (corrected)	0.2341	0.2896	0.366	0.2576	73.0	0.730	0.687

†LAGS uses the common reward proxy. Reproduction of its original benchmark remains unverified. NoiseTilt is an SD/DDIM adaptation.

C.5.2 Computation and selection ablations

Time is the median sampling latency including prompt encoding and final decoding, excluding cached reward text embeddings, image saving, and post-hoc evaluation. These are observed median times, not hardware-normalized across methods.

Figure 9 compares candidate-count sweeps on the same 200 images. Our unsigned linear selector varies K, while SVDD-PM varies N. Median latency stays between 11.5 and 11.6 s across the tested K values, while SVDD-PM takes 10.6, 19.6, and 36.9 s at $N = 2, 4, 8$, respectively. All points in this sweep use the same GPU.

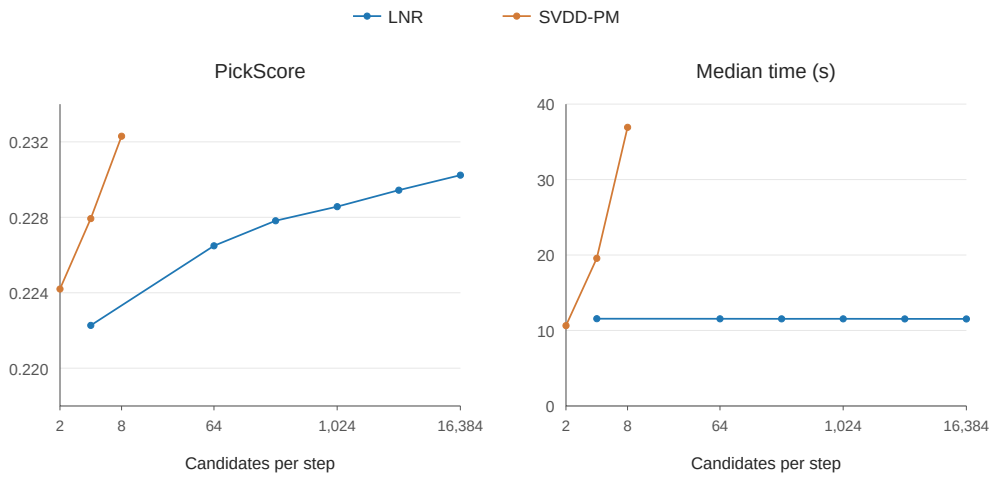


Figure 9: PickScore and median sampling time versus candidate count on the same 200 images and GPU. Our unsigned selector varies K, while SVDD-PM varies N.

More candidates at nearly constant latency. Figure 9 shows PickScore rising from 0.2223 to 0.2302 as K increases, while median latency stays within 11.5–11.6 s. At $K = 4096$, linear selection takes 11.5 rather than 19.6 s on the same GPU, with PickScore 0.2294 versus 0.2279. The paired 95% interval for this difference is $[0.00036, 0.00267]$. Flat latency across K reflects the direct winner sampler.

Direction and refinement. Stride 2 applies LNR on alternate stochastic steps and leaves the other steps unsteered. Forward reuse computes the reward gradient through the current-step denoiser prediction at x_t, in place of a second prediction at the transition mean. Random direction still computes the reward gradient but ranks candidates with an independent random unit vector. Normalized mean shift uses the same u_t as linear LNR but samples $\mu_t + \sigma_t(\epsilon_t + c_{K_t} u_t)$, where ϵ_t is fresh standard Gaussian noise. It matches the selected noise's mean directional displacement but not its distribution.

Two corrections reach PickScore 0.2341 in 30.9 s. Stride 2 and forward reuse take 6.7 and 9.8 s, respectively. Random direction returns nearly to the base score. Normalized mean shift reaches 0.2297, versus 0.2294 for linear selection at the same median time. This experiment does not establish an advantage over the mean-shift control.

Held-out quality. Against four-candidate SVDD-PM, linear selection has higher TOPIQ (0.686 versus 0.674), but lower IR (0.264 versus 0.354) and CLIP cosine (0.2539 versus 0.2583). The result supports a reward–computation trade-off, not uniform superiority across metrics.