BEYOND GAUSSIANITY: FROM SOURCE PROPERTIES TO MUON FOR NOISE OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Reward-guided initial-noise optimization seeks to maximize a learned reward while keeping the latent within a region that a pretrained generator maps to high-quality samples. Existing methods approximate this feasible region through regularization or projection toward Gaussian source properties. However, we find that pretrained generators tolerate substantial deviations from Gaussian entry-wise distributions while still producing realistic images across three recent models. We partly explain this tolerance through exact agreement of first- and second-order statistics after the first embedding. At the same time, controlled rearrangements can destroy or restore generation quality despite preserving the same entry values and token norms. These observations motivate a weaker spectral description of the admissible source space that is independent of the entry-wise distribution. Rather than explicitly enforcing Gaussianity, we control how reward-driven updates perturb this spectral structure. Remarkably, this control emerges simply by using Muon, whose momentum orthogonalization flattens the update spectrum. Experiments on aesthetic and text-aligned generation show that Muon reaches higher target rewards while retaining held-out scores, without an additional Gaussianity penalty or latent projection.

1 INTRODUCTION

Initial-noise optimization is now a common approach for test-time reward alignment of pretrained generators, improving a desired reward without changing the model's weights (Wallace et al., 2023; Ben-Hamu et al., 2024; Tang et al., 2025; Eyring et al., 2024). However, pushing the reward higher can degrade image quality. This naturally leads to a constrained optimization problem: maximize the reward while keeping the optimized latent within a region that the pretrained generator maps to high-quality samples. Existing approaches approximate this feasible region by preserving properties of the Gaussian training source, including its norm (Eyring et al., 2024), entry-wise statistics, and dependencies between entries (Hwang et al., 2025; Chang et al., 2026b).

We find, however, that these Gaussian-based constraints are far more restrictive than necessary: pretrained generators can produce realistic images from inputs that deviate dramatically from the standard Gaussian source. Figure 1 illustrates an extreme example. Consider a source whose entries independently take two values: $-1/\sqrt{3}$ with probability $3/4$ and $\sqrt{3}$ with probability $1/4$. It has zero mean and unit variance, but its entries follow a discrete, asymmetric two-point distribution rather than a continuous Gaussian. Nevertheless, FLUX.1-schnell (Black Forest Labs, 2024), Z-Image-Turbo (Z-Image Team et al., 2025), and FLUX.2-klein-4B (Black Forest Labs, 2025) generate realistic images from this source. Entry-wise Gaussianity is therefore not necessary for realistic generation across these recent pretrained generators.

To understand this tolerance, we distinguish the entry-wise distribution from the arrangement of entries across positions. The transformer-based generators first map latent tokens into embeddings through a learned affine transformation. Our analysis identifies which statistics match the Gaussian reference exactly and quantifies deviations across individual embedding features, token magnitudes, and relations between tokens. These results help explain why diverse entry-wise distributions satisfying broad conditions on mean, variance, token energy, and scalar tails retain generation quality. Generation is more sensitive to arrangement: placing the same token vectors in regular spatial patterns can severely degrade generation, while randomly permuting their positions restores realistic images.

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



Figure 1: Two-valued asymmetric noise generates realistic images across three generative models. Left: its probability mass and CDF (dashed: Gaussian); right: two examples from each model. Figure 13 shows images generated from Gaussian inputs using the same sampling settings.

This contrast motivates us to allow diverse shapes of the entry-wise distribution while preserving the random arrangement of the source.

Singular values connect this distinction to optimization. Viewing the latent as a matrix, we find that its singular spectrum tolerates broad changes in the entry-wise distribution but responds to certain changes in arrangement. Under standard random-matrix assumptions, independent, standardized entries from different distributions share a limiting spectral law after normalization (Marčenko & Pastur, 1967). Certain structured arrangements instead produce large spectral changes without altering the entry-wise distribution. This suggests controlling how reward-driven updates change the latent spectrum. In particular, we consider flattening the singular values of the update before it is added to the latent. A worst-case bound connects this choice to one-step changes in the latent spectrum.

Remarkably, this relaxation allows the desired spectral control to be built into the optimizer itself, rather than imposed through an explicit regularizer or projection. Muon (Jordan et al., 2024), widely used for large-scale neural network training, turns out to provide exactly this update geometry. Muon constructs the update by orthogonalizing the momentum, flattening its singular values. Existing analyses study Muon's optimization geometry, convergence, and implicit regularization in parameter training (Chen et al., 2026; Ma et al., 2026; Nguyen et al., 2026). We provide a different perspective: its orthogonalization can help preserve random arrangement by suppressing concentration in the momentum's singular spectrum before accumulation.

Our experiments on FLUX.1-schnell demonstrate Muon's effectiveness in both aesthetic and text-aligned image generation. Across both tasks, Muon reaches higher target rewards than the compared optimizers, including Gaussianity regularization methods and GPER, while retaining the initial levels of all three held-out scores. These results show that an optimizer familiar from neural network training provides a simple and effective choice for reward-guided noise optimization, with its behavior understood through the properties of the generator's source.

2 Gaussian Distribution and Reward-Guided Noise Optimization

Diffusion and flow models are trained to transport a Gaussian source distribution into the data distribution. Let $\mathbb{F}$ denote the data space and $\boldsymbol{X} \in \mathbb{R}^{N\times d}$ contain N latent tokens of dimension d. The pretrained generative model $\mathcal{M} : \mathbb{R}^{N\times d} \to \mathbb{F}$ maps the initial latent to a data sample, and the reward model $r : \mathbb{F} \to \mathbb{R}$ assigns it a scalar score. Reward-guided noise optimization searches for an

initial noise that produces a high-reward output. For a fixed prompt, its idealized objective is

$$\max_{\boldsymbol{X} \in \mathcal{A}_{\mathcal{M}}} r\big(\mathcal{M}(\boldsymbol{X})\big). \tag{1}$$

Here $\mathcal{A}_{\mathcal{M}}$ denotes the set of initial latents that the pretrained generator maps to high-quality outputs; prompt dependence is suppressed in the notation. This region is not explicitly known. In practice, we iteratively update $\boldsymbol{X}$ using reward gradients while keeping the generator fixed, enabling test-time control of the generated output without retraining (Wallace et al., 2023; Ben-Hamu et al., 2024).

However, increasing the target reward does not necessarily preserve generation quality. Without constraining $\boldsymbol{X}$ to remain in $\mathcal{A}_{\mathcal{M}}$, optimization can raise the score while making the image less realistic or removing content specified by the prompt. As shown in Figure 2, the target score increases while the chair disappears. The reward objective alone therefore provides no guarantee that the optimized noise remains suitable for generating realistic images consistent with the prompt, so such reward gains do not fulfill the quality-constrained goal in Equation 1.

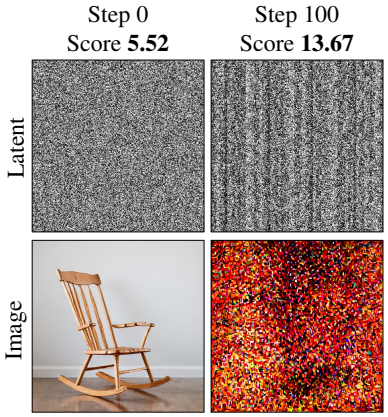


Figure 2: Quality degradation when optimizing for "a wooden rocking chair". The aesthetic score rises from 5.52 to 13.67 as the chair disappears.

Since the generator is trained with Gaussian noise, a natural approximation to the valid source region $\mathcal{A}_{\mathcal{M}}$ is to preserve properties of the Gaussian source. Previous work therefore constrains either the optimized noise or the updates applied to it. ReNO regularizes the noise norm (Eyring et al., 2024); MPGR and entropy-based noise regularization additionally control higher-order or spatial statistics (Hwang et al., 2025; Chang et al., 2026b); and GPER combines reward-gradient preconditioning with latent projection (Hwang & Sung, 2026). These methods define tractable approximations to $\mathcal{A}_{\mathcal{M}}$ using Gaussian source properties, but the training distribution does not reveal how tight these constraints need to be. We therefore first characterize how far sources can depart from Gaussianity without compromising generation, and use these properties to guide noise optimization.

3 Wider Source Distributions of Pretrained Flow Models

If a Gaussian entry-wise distribution were essential to a generator trained on Gaussian noise, replacing it with a discrete, asymmetric law would be a substantial intervention. We begin with precisely such a source:

$$\Pr\left(\xi = -\frac{1}{\sqrt{3}}\right) = \frac{3}{4}, \quad \Pr\left(\xi = \sqrt{3}\right) = \frac{1}{4}. \tag{2}$$

Its mean is zero and variance is one, but its shape is clearly far from a Gaussian distribution. Nevertheless, replacing Gaussian entries with these two values produces realistic images across FLUX.1-schnell, Z-Image-Turbo, and FLUX.2-klein-4B (Black Forest Labs, 2024; Z-Image Team et al., 2025; Black Forest Labs, 2025), without changing their weights. Figure 1 shows examples, and Table 1 compares their quality (Two-point, $p = 1/4$) with Gaussian-source generation. The study in Appendix B.5 supports this observation across prompts.

Table 1: PickScore differences from Gaussian inputs across entry-wise distributions in Figures 1 and 3. μ_3/κ_4: skewness/excess kurtosis. Shaded rows violate the conditions in Equation 3. F.1/Z/F.2: FLUX.1-schnell/Z-Image-Turbo/FLUX.2-klein. Details: Appendix B.5.

Source	Law		ΔPick ↑		
	μ_3	κ_4	F.1	Z	F.2
$\mathcal{N}(0,1)$	0.00	0.00	0.00	0.00	0.00
Rademacher	0.00	−2.00	+0.01	+0.05	+0.01
Two-point, $p = 1/3$	0.71	−1.50	+0.00	+0.03	+0.00
Two-point, $p = 1/4$	1.15	−0.67	+0.02	+0.01	+0.03
Four-point	0.32	−0.33	+0.02	+0.03	+0.03
Uniform	0.00	−1.20	−0.01	+0.03	+0.05
Arcsine	0.00	−1.50	−0.01	+0.02	+0.03
Wave	0.00	−1.07	−0.00	+0.03	+0.03
Skewed bumps	0.58	−1.31	+0.02	+0.04	+0.03
Wedge	−0.84	0.02	+0.02	+0.03	+0.03
Outside the class					
$-\Gamma(4)$ *(T,E)*	−1.00	1.49	+0.01	−0.00	+0.02
Lognormal *(T,E)*	5.80	76.35	−2.54	−3.05	−0.92
$\mathcal{N}(0.2, 1)$ *(M,T,E)*	0.00	0.00	−6.86	−5.44	−7.54

For non-Gaussian sources in DDPM (Ho et al., 2020), Theorem 5 of Liang (2025) assumes a zero-mean symmetric distribution with a density and finite second moment. Our discrete, asymmetric two-point source does not satisfy these source assumptions, motivating us to examine why it works in pretrained flow generators. Following the distinction made in MPGR (Hwang et al., 2025), we

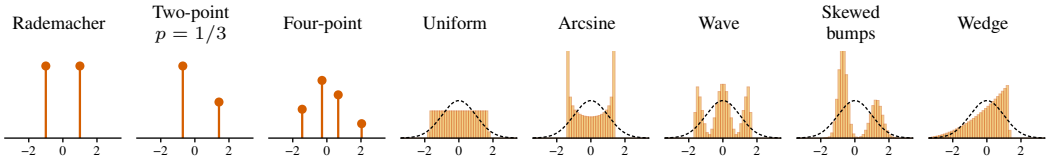


Figure 3: Eight entry-wise distributions satisfying (M), (E), and (T) in Eq. 3 yield near-Gaussian generation quality (Table 1). Dashed curves in continuous panels show the standard normal density.

examine two aspects of the source separately: its entry-wise distribution and the arrangement of its entries. For the first, we ask how much the entry-wise distribution can change; for the second, we ask what happens when the same values occupy different positions.

3.1 Entry-wise Distribution

What allows such different entry-wise distributions to work? We focus on how the transformer-based generators we study first map latent tokens into embeddings. For one input token $\mathbf{x} \in \mathbb{R}^d$, it takes the form $\mathbf{y} = \boldsymbol{W}^\top \mathbf{x} + \boldsymbol{b}$, where $\boldsymbol{W} \in \mathbb{R}^{d\times K}$ and $\boldsymbol{b} \in \mathbb{R}^K$ are frozen embedding parameters. Each embedded element is a weighted sum of entries. We therefore ask which statistics of $\mathbf{y}$ remain similar to those obtained from Gaussian inputs, even when the entry-wise distribution changes substantially.

Following this intuition, we consider broad conditions on the source mean, variance, token energy, and scalar tails. For an entry $\xi \sim F$, independent copies ξ_i, $\mathrm{S}_d = \sum_{i=1}^d \xi_i^2$, and $\mathrm{Z} \sim \mathcal{N}(0,1)$, write

$$\begin{aligned} &\text{(M)} \quad \mathbb{E}[\xi] = 0, \qquad \mathbb{E}[\xi^2] = 1, \\ &\text{(E)} \quad \Pr\big(\mathrm{S}_d \notin [L_d, U_d]\big) \le 1/N, \\ &\text{(T)} \quad \Pr\big(|\xi| > u\big) \le \Pr\big(|\mathrm{Z}| > u\big) \quad \text{for all } u \ge B. \end{aligned} \tag{3}$$

Here $[L_d, U_d]$ is the central $1-1/N$ interval of χ^2_d, and B is a fixed threshold beyond which the tail is controlled. Condition (M) sets the expected squared norm of the latent to Nd. The two tail conditions restrict extreme scalar values and permit at most one out-of-interval token in expectation; they leave the central shape of the entry-wise distribution largely unrestricted. Exact thresholds and intervals are given in Appendix B.1. The following proposition connects three aspects of the embedding to the Gaussian reference: feature statistics, token magnitude, and relations between tokens.

Proposition 1 (First-embedding statistics)**.** *Let $\mathbf{y}_t = \boldsymbol{W}^\top \mathbf{x}_t + \boldsymbol{b}$ for $t = 1, \dots, N$, with fixed $\boldsymbol{W}, \boldsymbol{b}$ and all input entries i.i.d. satisfying (M). Then*

$$\mathbb{E}[\mathbf{y}_t] = \boldsymbol{b}, \qquad \mathrm{Cov}(\mathbf{y}_t) = \boldsymbol{W}^\top \boldsymbol{W}, \qquad \mathbb{E}\big[\|\mathbf{y}_t - \boldsymbol{b}\|_2^2\big] = \|\boldsymbol{W}\|_F^2. \tag{4}$$

For distinct tokens $t \neq s$, let $\mathrm{D}_{ts} = (\mathbf{y}_t - \boldsymbol{b})^\top (\mathbf{y}_s - \boldsymbol{b})$. Their centered inner product satisfies

$$\mathbb{E}[\mathrm{D}_{ts}] = 0, \qquad \mathrm{Var}(\mathrm{D}_{ts}) = \|\boldsymbol{W}\boldsymbol{W}^\top\|_F^2. \tag{5}$$

All these quantities match those obtained from standard Gaussian inputs.

Proposition 1 shows that changing the entry-wise distribution need not change several statistics presented to the transformer: feature means and covariances, the expected squared magnitude of a centered token, and the mean and variance of centered inner products between tokens. These agreements require only matched input mean and variance, without requiring a Gaussian density. Moreover, when an embedding channel mixes sufficiently many entries without a dominant weight, its standardized distribution can be approximately Gaussian. Figure 4 illustrates this for an asymmetric three-point input in one FLUX.1-schnell channel. These agreements help explain the observed tolerance to changes in the entry-wise distribution: a substantial change in the input law can leave multiple embedding statistics unchanged or close to their Gaussian counterparts.

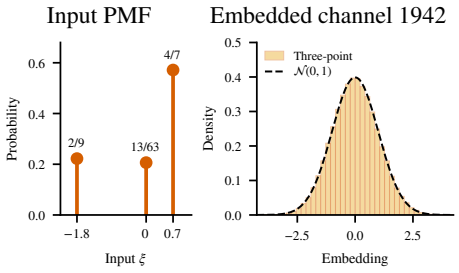


Figure 4: An asymmetric three-point input yields a nearly Gaussian standardized distribution in randomly selected channel 1942 of the FLUX.1-schnell first embedding.

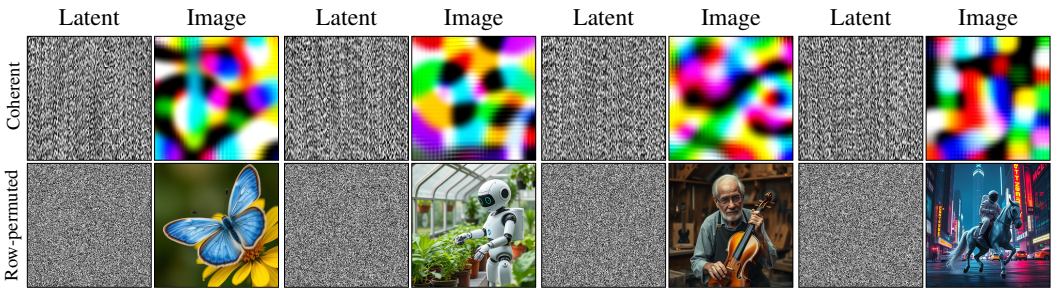


Figure 5: Token arrangement affects generation in FLUX.1-schnell. Each pair shows the input latent and its generated image. Top: arranging tokens in coherent low-DCT patterns degrades generation. Bottom: randomly permuting the same tokens restores recognizable content without changing the token values. Prompts are listed in Table 10 (entries 4, 5, 9, and 10, left to right).

Appendix B provides the proofs, higher-order corrections, and moment and CDF measurements for the two-point source across three generators. It also details the tail conditions and Gaussian approximation, including finite-population corrections.

We further test a range of entry-wise distributions on the three generators. Figure 3 shows their shapes, and Table 1 compares generation quality under the same sampling settings. The eight tested sources satisfying the conditions retain near-reference quality despite differences in continuity, symmetry, and density. The two-point example is therefore part of a broader pattern, rather than an isolated exception. Some sources that do not satisfy the conditions also work, while others clearly degrade generation. Thus, we use these conditions to organize the observed tolerance, rather than define a hard boundary for generation. The complete source measurements and controls that change the mean, scale, or tails are in Appendices B.5.2 and B.6.

3.2 Arrangement of the Entries in the Latent

The working sources above have very different entry-wise distributions, but their entries share one property: they are randomly arranged. Since a histogram does not record the positions of its values, comparisons of entry-wise distributions alone do not test this aspect of the source. We now hold the values fixed and change their placement, to examine whether an otherwise acceptable collection of entries can become unsuitable through arrangement alone.

We construct coherent arrangements that substantially degrade generation, as shown in Figure 5. To isolate the effect of position from token content, we keep the token vectors fixed and then randomly permute their rows: $\boldsymbol{X}_{\text{perm}} = \boldsymbol{P}\boldsymbol{X}$, with $\boldsymbol{P}$ a uniform permutation matrix. The model's positional assignments remain fixed. This control preserves the scalar values, each token's squared norm, and the total latent norm. A difference in generation therefore cannot be attributed to the shape of the entry-wise distribution or explained by the total norm alone.

Table 2: Changing arrangement at fixed token vectors. Metrics follow Table 1; full details are in Table 8.

Construction	Arrangement	ΔPick↑		
		F.1	Z	F.2
Low-DCT	Coherent	−9.06	−9.12	−8.76
	Permuted	−0.01	−0.23	−0.03
Hadamard	Coherent	−8.01	−7.63	−8.91
	Permuted	+0.04	+0.02	+0.00

After permutation, the generators again produce realistic images, with quality recovering toward the Gaussian reference (Table 2). The recovery is substantial, although its extent depends on the pattern and generator. The full paired measurements and further controls are provided in Appendix C. Thus, the same collection of token vectors can be suitable or damaging depending on how the vectors are assigned to positions.

These two sets of experiments distinguish the training source from the properties that support generation. Gaussian training makes Gaussian preservation a reasonable starting point, but the range of working entry-wise distributions is considerably broader. Within that range, arrangement can make the difference between successful and degraded generation, even at an identical norm. This changes the design question for noise optimization: how can we allow the observed variation in the entry-wise distribution while protecting the random arrangement of the source? We next consider a statistic that can guide the design of an update rule for this objective.

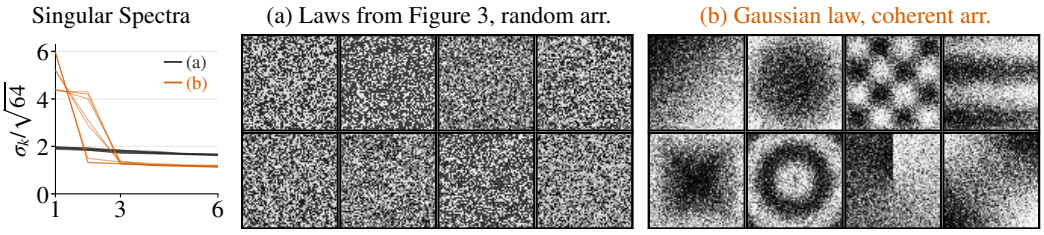


Figure 6: Leading singular values under different entry-wise distributions and arrangements. (a) Samples from the eight distributions in Figure 3 are randomly arranged into 64×64 matrices; their singular values are similar (black curves in the spectra plot). (b) The same Gaussian values are arranged into eight coherent patterns, yielding substantially different spectra (orange curves in the spectra plot).

4 From Source Properties to an Update Rule

The source experiments suggest that we need to consider two requirements together: allowing broad changes in the entry-wise distribution and responding to concentration introduced by structured arrangements. A statistic that depends on a Gaussian entry-wise density would restrict sources that already work. Instead, we consider the singular spectrum of an $m \times n$ representation of the latent, with $mn = Nd$. It describes how the entries combine across rows and columns, allowing us to study their organization beyond the scalar histogram.

4.1 Singular Values across Entry-wise Distributions

Random-matrix universality provides a useful reference. Under the standard assumptions, matrices with independent, standardized entries share the same limiting empirical singular-value distribution, even when their entry-wise distributions differ substantially (Marčenko & Pastur, 1967). Thus, a common spectral reference need not impose Gaussian entry-wise distributions. The formal statement, normalization, and assumptions are given in Appendix D.1.

Figure 6 illustrates the distinction with two finite-size comparisons. In the first, values from each of the eight working entry-wise distributions in Figure 3 are randomly arranged into 64×64 matrices. Their leading spectra are closely grouped (black), despite the visibly different entry-wise distributions. In the second, the same Gaussian values are placed into eight coherent patterns, producing substantially different spectra (orange). The first comparison illustrates tolerance to changes in the entry-wise distribution under random placement, while the second shows sensitivity to the arrangement of fixed values. Construction details are in Appendix D.4.

4.2 Designing the Update

During reward optimization, the updates themselves can carry structured patterns into the latent. This motivates processing each update before it accumulates, rather than projecting the resulting latent back to a reference spectrum. The distinction matters because a latent can satisfy spectral reference intervals while retaining harmful structure (Appendix E). We therefore seek an update rule that limits the worst-case change in the latent spectrum at a fixed update norm.

We write one step as $\boldsymbol{X}^+ = \boldsymbol{X} + \eta\boldsymbol{\Delta}$, where $\eta \geq 0$ and both matrices use the same fixed representation. A reward gradient may concentrate the update in a few singular components. At a fixed update norm, this concentration increases its largest singular value, which controls a worst-case bound on changes to the latent spectrum. We apply this criterion to momentum, which accumulates reward gradients across optimization steps.

Proposition 2 (Spectral control at a fixed update norm)**.** *For matrices $\boldsymbol{X}, \boldsymbol{\Delta} \in \mathbb{R}^{m\times n}$ and $\eta \geq 0$,*

$$\max_i \left|\sigma_i(\boldsymbol{X} + \eta\boldsymbol{\Delta}) - \sigma_i(\boldsymbol{X})\right| \leq \eta\|\boldsymbol{\Delta}\|_{\mathrm{op}}. \tag{6}$$

Among updates with $\|\boldsymbol{\Delta}\|_F = \rho > 0$ and $\mathrm{rank}(\boldsymbol{\Delta}) \leq r$, where $1 \leq r \leq \min(m, n)$, the smallest operator norm is $\rho/\sqrt{r}$. It is attained precisely when the update has r equal nonzero singular values. In particular, for a rank-r momentum with compact SVD $\boldsymbol{M} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$, the update $\boldsymbol{\Delta}_\star = (\rho/\sqrt{r})\boldsymbol{U}\boldsymbol{V}^\top$ attains this minimum.

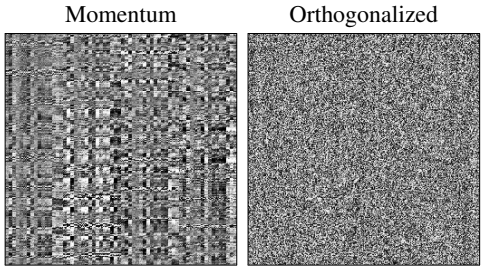


Figure 7: Visualization of orthogonalized momentum obtained during Aesthetic Score optimization. The fraction of energy captured by the top 10 singular values decreases from 77.8% to 3.9%.

The proposition follows from the singular-value perturbation inequality and $\|\mathbf{\Delta}\|_F^2 \leq r\|\mathbf{\Delta}\|_{\text{op}}^2$; the full proof is provided in Appendix D.3. Under a fixed norm and rank budget, the proposition selects a flat update spectrum, and the bound $\rho/\sqrt{r}$ falls as more directions are equalized. Since $r \leq \min(m, n)$, the smaller matrix dimension limits the available rank. For a fixed number of entries mn, we therefore use a near-square representation to maximize $\min(m, n)$ and reduce the full-rank bound $\rho/\sqrt{\min(m, n)}$ (Chang et al., 2026a). To retain the directional information in momentum, we preserve its left and right singular vectors and equalize its nonzero singular values. The resulting direction is its polar factor, $\text{polar}(\boldsymbol{M}) = \boldsymbol{U}\boldsymbol{V}^\top$, with the common scale absorbed into the step size.

Figure 7 illustrates how orthogonalization weakens structured patterns in momentum obtained during Aesthetic optimization. The raw momentum contains visible regularity, which becomes less pronounced after singular-value flattening. This motivates processing momentum before accumulation to limit the disruption of the source's random arrangement. The operation acts on the incoming update; the latent itself is neither flattened nor projected to a reference spectrum. Proposition 2 provides a spectral criterion for this intervention through a worst-case displacement bound, without implying that every realized spectral change must be smaller.

4.3 A Source-Based Interpretation of Muon

The polar update obtained above is the momentum orthogonalization at the core of Muon (Jordan et al., 2024). Muon is familiar from neural-network training, including LLMs; here, the same operation acts on the input noise of a fixed generator. Our source analysis supplies a different motivation for this operation: the generator tolerates broad changes in the entry-wise distribution, but can be sensitive to structured arrangements. Orthogonalization addresses spectral concentration in the incoming reward momentum before it accumulates in the latent.

This interpretation emphasizes the structure introduced by each update, rather than whether the latent spectrum remains within reference intervals. Appendix E examines this distinction. A latent can retain visible structure and produce artifacts even when its singular values remain within those intervals. At nearly matched target rewards, latent singular-value clipping produces a low-quality image, with coherent structure visible in the latent. Muon instead produces a realistic image with less visible structure in the latent, even though its singular spectrum extends beyond the reference intervals (Figure 20). Across three seeds, momentum processing achieves higher Aesthetic Scores at comparable held-out scores than the tested latent projections (Figure 21). Together, these results support processing incoming structure before it accumulates, rather than relying on latent statistics.

The connection makes the source analysis actionable through a familiar optimizer: use Muon for reward-guided noise optimization (Algorithm 1), without an additional Gaussianity penalty or latent projection. Removing the polar operation recovers momentum SGD. The generator and reward model remain fixed; numerical details are provided in Appendix F.2. We next evaluate Muon's reward–quality trade-offs against existing optimizers.

Algorithm 1 Muon optimizer

$\boldsymbol{M} \leftarrow \mathbf{0}$
Iterate:
  $\boldsymbol{M} \leftarrow \beta\boldsymbol{M} + \nabla_{\boldsymbol{X}}\, r\big(\mathcal{M}(\boldsymbol{X})\big)$
  $\boldsymbol{X} \leftarrow \boldsymbol{X} + \eta\,\text{polar}(\boldsymbol{M})$

5 Experimental Results

Experimental Setup. We evaluate Muon (Jordan et al., 2024) on aesthetic and text-aligned image generation with FLUX.1-schnell (Black Forest Labs, 2024), using Aesthetic Score (Schuhmann, 2022) and PickScore (Kirstain et al., 2023) as the target rewards, respectively. For each target, the remaining three scorers among Aesthetic Score, PickScore, ImageReward (Xu et al., 2023), and HPSv2.1 (Wu et al., 2023) are held out of optimization. We compare against SGDM (Polyak, 1964; Sutskever et al., 2013), Adam (Kingma & Ba, 2015), ReNO (Eyring et al., 2024), MPGR (Hwang

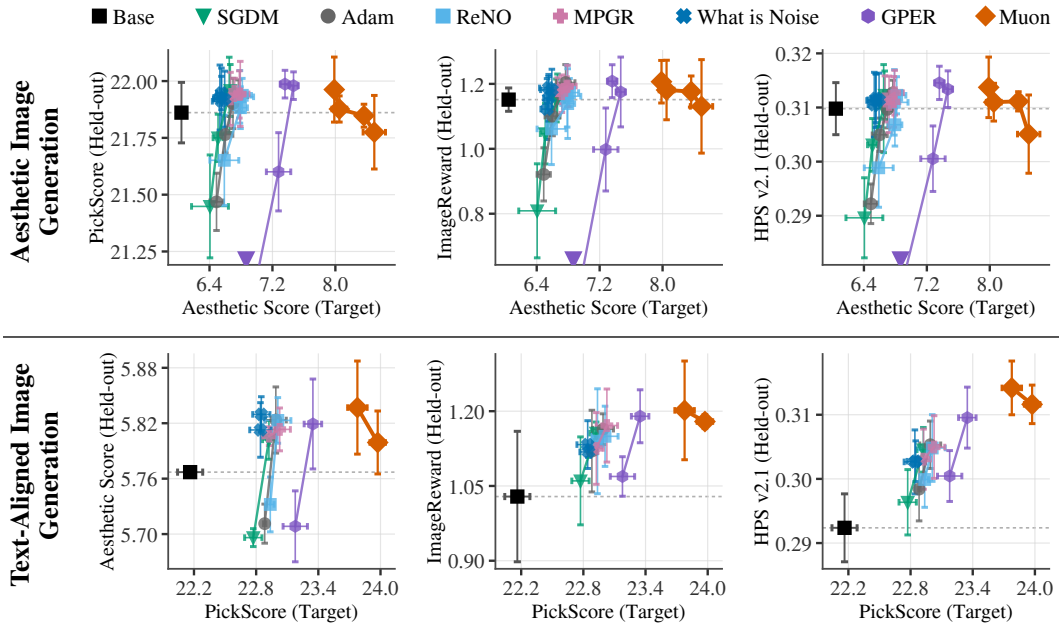


Figure 8: Quantitative results after 100 noise-optimization updates on FLUX.1-schnell. The upper row targets Aesthetic Score and the lower row targets PickScore. Points and error bars show the mean and one standard deviation across three seed-wise prompt means. Farther right and farther up are better. Dashed horizontal lines mark Base held-out scores. In the upper row, the highest-RMS GPER points fall below the displayed y-ranges and are indicated by purple downward triangles. Full values appear in Tables 11 and 12.

et al., 2025), What Is Noise (Chang et al., 2026b), and GPER (Hwang & Sung, 2026) over 100 updates and three seeds. Figure 8 shows the trade-offs obtained by varying the update RMS for each method, with the unoptimized generator as Base. The implementation details and full numerical results are provided in Appendix F. We additionally evaluate Z-Image-Turbo (Z-Image Team et al., 2025) and FLUX.2-klein-4B (Black Forest Labs, 2025) on both tasks, with their quantitative results reported in Figure 22 in the appendix.

Aesthetic Image Generation. Muon achieves higher Aesthetic Scores than the compared methods while retaining held-out scores (Figure 8, top). Among tested settings whose three held-out means are all at least their Base values, Muon reaches over 8, compared with 7.5 for GPER and at most 6.8 for the remaining baselines, from a Base score of 6. Further increasing the optimization strength raises Muon's target score to 8.5, with some reduction in held-out scores. Thus, the curves show both the larger reward range and the trade-off introduced by stronger updates, rather than only the improvement in the target reward. Full numerical comparisons are provided in Table 11.

Text-Aligned Image Generation. We observe the same advantage when optimizing PickScore (Figure 8, bottom). Under the same held-out-mean criterion, Muon reaches nearly 24, compared with 23.3 for GPER and at most around 23 for the other baselines, from 22.2 before optimization. All three held-out scores also improve over Base, and at RMS 0.004 Muon attains the highest mean of each among all tested settings (Table 12).

Qualitative Results. Figure 9 compares methods using the same prompt and initial noise within each row. For "dolphin", Muon produces a sunset scene with a dolphin leaping above the water, whereas SGDM, Adam, and ReNO produce simplified graphic depictions. The other baselines produce less aesthetically pleasing images and achieve lower Aesthetic Scores. For "three peaches and two cups", Muon depicts the requested objects in the correct numbers, whereas all baselines depict incorrect numbers of peaches or cups. These examples complement the aggregate results: replacing the noise optimizer with Muon improves reward-guided generation without a separate Gaussianity penalty or latent projection.

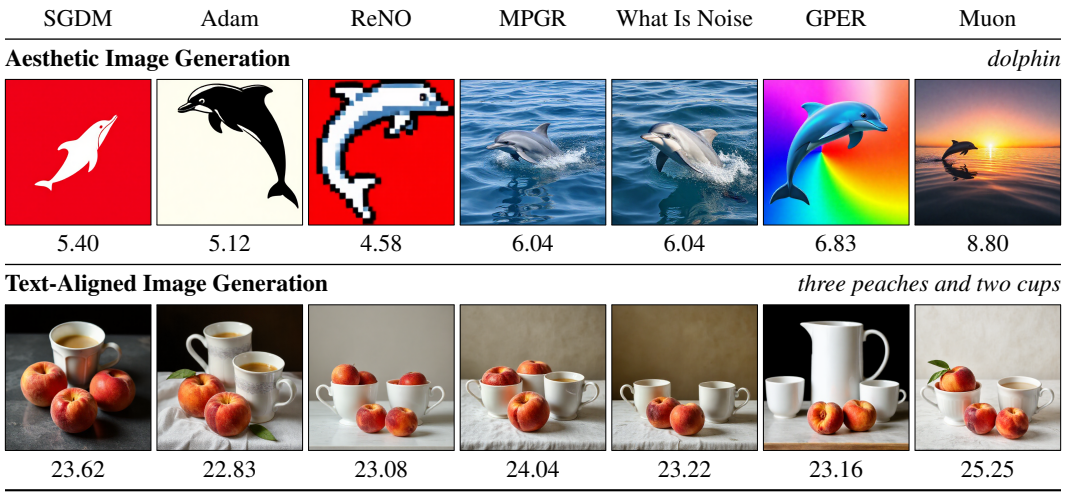


Figure 9: Qualitative examples after 100 updates on FLUX.1-schnell. Prompts appear at the upper right of each row. Numbers below the images are Aesthetic Scores (top) and PickScores (bottom).

6 Related Work

ReNO (Eyring et al., 2024) regularizes the noise norm; MPGR (Hwang et al., 2025) and What Is Noise (Chang et al., 2026b) constrain entry-wise and spatial statistics toward Gaussian references. GPER (Hwang & Sung, 2026) combines Fourier-domain gradient preconditioning with latent projection. These methods preserve properties of the Gaussian training source. Complementary work by Liang (2025) studies non-Gaussian sources in DDPM (Ho et al., 2020) and proposes a guarantee for symmetric densities with zero mean and finite second moment; Appendix A.1 examines this claim. These assumptions exclude the discrete asymmetric sources considered here. We study the tolerance of pretrained flow generators to diverse entry-wise distributions, including discrete asymmetric sources, and provide a partial explanation through embedding statistics. Controlled token rearrangements hold the values and token norms fixed, isolating positional effects that entry-wise statistics alone cannot explain.

Muon (Carlson et al., 2015; Tuddenham et al., 2022; Bernstein & Newhouse, 2024; Jordan et al., 2024) builds on spectral descent and gradient orthogonalization, applying the operation to momentum in neural-network training, including LLMs. Analyses connect Muon with decoupled weight decay to implicit spectral-norm constraints on weights (Chen et al., 2026) and study its preconditioning benefits (Ma et al., 2026). Muse (Chang et al., 2026a) shows how the matrix representation changes the polar update's geometry. These works explain orthogonalized updates in parameter optimization. We apply Muon to reward-guided initial-noise optimization with pretrained reward models and interpret its update through the generator's tolerance to entry-wise variation and sensitivity to arrangement. This perspective motivates processing momentum before accumulation without an additional Gaussianity penalty or latent projection. Spectral analysis and comparisons of latent versus update processing, including GPER, appear in Appendices D and E; Appendix A.2 expands the comparison with prior work.

7 Conclusion

Despite Gaussian training, pretrained generators can tolerate broad non-Gaussian entry-wise distributions while remaining sensitive to arrangement. Our embedding analysis helps explain this tolerance, while controlled rearrangements show that preserving the values and norms alone does not ensure successful generation. This distinction motivates controlling the structure introduced by reward-driven updates and leads to a source-based interpretation of Muon. Our experiments show improved reward–quality trade-offs in both aesthetic and text-aligned generation. These findings suggest a simple change to existing noise-optimization pipelines: replacing Adam or momentum SGD with this familiar optimizer, without an additional Gaussianity penalty or latent projection.

AI USE STATEMENT

We formulated the research questions and hypotheses, designed and directed the experiments, selected the datasets, and developed the mathematical reasoning and proof outlines. We also established the paper's narrative and its relationship to prior work. We used generative AI tools to draft sections and polish the writing based on our outlines, mathematical reasoning, and account of related work, and to assist with LaTeX tasks such as formatting tables and placing figures. All research ideas and decisions were our own. We take full responsibility for the research and the final manuscript, including all AI-assisted text and artifacts.

REPRODUCIBILITY STATEMENT

Appendices B and D provide the embedding and spectral analyses with proofs. Source construction, arrangement controls, and their evaluation protocols are described in Appendices B and C. Appendix F specifies the generators, sampling settings, rewards, prompt sets, preprocessing, optimizer configurations, proposal-update RMS, and aggregation. Appendix E defines the latent-projection and momentum-processing controls. Full reward-optimization scores, including variation across seeds, are reported in Table 9 and Appendix F.4.

ETHICS STATEMENT

This work analyzes and optimizes the initial noise of publicly released text-to-image generators with respect to public reward models. It introduces no new generative capability and no new data. Reward-guided optimization can amplify reward-model biases; the held-out metrics reported here measure transfer between reward models, not human preference or safety.

REFERENCES

Z. D. Bai and Y. Q. Yin. Limit of the smallest eigenvalue of a large dimensional sample covariance matrix. *The Annals of Probability*, 21(3):1275–1294, 1993. doi: 10.1214/aop/1176989118.

Heli Ben-Hamu, Omri Puny, Itai Gat, Brian Karrer, Uriel Singer, and Yaron Lipman. D-Flow: Differentiating through flows for controlled generation. In *International Conference on Machine Learning*, 2024.

Jeremy Bernstein and Laker Newhouse. Old optimizer, new norm: An anthology. In *NeurIPS 2024 Workshop on Optimization for Machine Learning (OPT)*, 2024.

Black Forest Labs. FLUX.1 [schnell]. Model card and reference implementation, 2024. URL `https://huggingface.co/black-forest-labs/FLUX.1-schnell`.

Black Forest Labs. FLUX.2: Frontier visual intelligence. Software and model documentation, 2025. URL `https://github.com/black-forest-labs/flux2`.

David Carlson, Volkan Cevher, and Lawrence Carin. Stochastic spectral descent for restricted boltzmann machines. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, 2015.

Franz Louis Cesista. Frequency domain Muon for convolutional neural networks: Simplified. Technical note, 2026. URL `https://leloykun.github.io/ponder/freqmuon/`.

Da Chang, Qiankun Shi, Lvgang Zhang, Di He, Yaoshuai Ma, Ganzhao Yuan, and Yongxiang Liu. Muse: Representation geometry of Muon beyond normalized momentum. *arXiv preprint arXiv:2607.14536*, 2026a.

Pascal Chang, Kai Lascheit, Jingwei Tang, Markus Gross, and Vinicius C. Azevedo. What is it like to be a noise? an entropy-based Gaussian noise regularization for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026b.

Sourav Chatterjee. A generalization of the Lindeberg principle. *The Annals of Probability*, 34(6): 2061–2076, 2006. doi: 10.1214/009117906000000575.

Lizhang Chen, Jonathan Li, and Qiang Liu. Muon optimizes under spectral norm constraints. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856.

Louis H. Y. Chen and Xiao Fang. On the error bound in a combinatorial central limit theorem. *Bernoulli*, 21(1):335–359, 2015. doi: 10.3150/13-BEJ569.

Weiye Chen, Qingen Zhu, and Qian Long. Muon-accelerated attention distillation for real-time edge synthesis via optimized latent diffusion, 2025. URL https://arxiv.org/abs/2504.08451.

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*, 2024.

Dan Friedman and Adji Bousso Dieng. The Vendi Score: A diversity evaluation metric for machine learning. *Transactions on Machine Learning Research*, 2023.

Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. DreamSim: Learning new dimensions of human visual similarity using synthetic data. In *Advances in Neural Information Processing Systems*, 2023.

Nicholas J. Higham. Singular value inequalities. Technical blog, 2021. URL https://nhigham.com/2021/05/04/singular-value-inequalities/.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, 2020.

Kaiyi Huang, Chengqi Duan, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2I-CompBench++: An enhanced and comprehensive benchmark for compositional text-to-image generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(5):3563–3579, 2025. doi: 10.1109/TPAMI.2025.3531907. URL https://doi.org/10.1109/TPAMI.2025.3531907.

Jisung Hwang and Minhyuk Sung. Gradient preconditioning for efficient and reliable reward-guided generation. In *International Conference on Machine Learning*, 2026.

Jisung Hwang, Jaihoon Kim, and Minhyuk Sung. Moment- and power-spectrum-based Gaussianity regularization for text-to-image models. In *Advances in Neural Information Processing Systems*, 2025.

Jisung Hwang, Yunhong Min, Jaihoon Kim, I-Chao Shen, and Minhyuk Sung. NoiseTilt: Noise-tilted reverse kernels for diffusion reward alignment. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2026.

Keller Jordan, Yuchen Jin, Vlado Boza, Jiacheng You, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks. Technical blog and reference implementation, 2024. URL https://kellerjordan.github.io/posts/muon/.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-Pic: An open dataset of user preferences for text-to-image generation. In *Advances in Neural Information Processing Systems*, 2023.

Stephen D. Liang. Reconsidering noise for denoising diffusion probabilistic models. In *NeurIPS 2025 Workshop on Structured Probabilistic Inference and Generative Modeling*, 2025.

Jarl Waldemar Lindeberg. Eine neue Herleitung des Exponentialgesetzes in der Wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift*, 15(1):211–225, 1922. doi: 10.1007/BF01494395.

Jianhao Ma, Yu Huang, Yuejie Chi, and Yuxin Chen. Preconditioning benefits of spectral orthogonalization in Muon. *arXiv preprint arXiv:2601.13474*, 2026.

Vladimir A. Marčenko and Leonid A. Pastur. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1(4):457–483, 1967. doi: 10.1070/SM1967v001n04ABEH001994.

Tien-Phat Nguyen, Truong Nguyen, Minh-Phuc Truong, Tuc Nguyen, James Bailey, and Trung Le. Spectral flattening is all Muon needs: How orthogonalization controls learning rate and convergence. *arXiv preprint arXiv:2605.13079*, 2026.

Iosif Pinelis. On the nonuniform Berry–Esseen bound. In Iosif Pinelis (ed.), *Inequalities and Extremal Problems in Probability and Statistics: Selected Topics*, pp. 103–138. Academic Press, 2017. doi: 10.1016/B978-0-12-809818-9.00004-5. URL https://doi.org/10.1016/B978-0-12-809818-9.00004-5.

Jim Pitman and David S. Rosenberg. Lecture 10: Setup for the central limit theorem. STAT205 lecture notes, University of California, Berkeley, 2002. URL https://www.stat.berkeley.edu/users/pitman/s205f02/lecture10.pdf. Lecturer: Jim Pitman; scribe: David S. Rosenberg.

Boris T. Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4(5):1–17, 1964. doi: 10.1016/0041-5553(64)90137-5.

Jiaxin Qing and Lexin Li. Muon-C: Operator-aligned Muon for convolutional kernels. *arXiv preprint arXiv:2609.09676*, 2026.

Mark Rudelson and Roman Vershynin. Hanson–Wright inequality and sub-gaussian concentration. *Electronic Communications in Probability*, 18(82):1–9, 2013. doi: 10.1214/ECP.v18-2865.

Christoph Schuhmann. LAION aesthetics. Technical blog, 2022. URL https://laion.ai/blog/laion-aesthetics/.

I. G. Shevtsova. An improvement of convergence rate estimates in the Lyapunov theorem. *Doklady Mathematics*, 82(3):862–864, 2010. doi: 10.1134/S1064562410060062.

Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *International Conference on Machine Learning*, 2013.

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 *International Conference on Machine Learning*, 2025.

Mark Tuddenham, Adam Prügel-Bennett, and Jonathan Hare. Orthogonalising gradients to speed up neural network optimisation. *arXiv preprint arXiv:2202.07052*, 2022.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. In Yonina C. Eldar and Gitta Kutyniok (eds.), *Compressed Sensing: Theory and Applications*, pp. 210–268. Cambridge University Press, 2012. doi: 10.1017/CBO9780511794308.006. URL https://doi.org/10.1017/CBO9780511794308.006.

Bram Wallace, Akash Gokul, Stefano Ermon, and Nikhil Naik. End-to-end diffusion latent optimization improves classifier guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

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.

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*, 2023.

Z-Image Team, Huanqia Cai, Sihan Cao, Ruoyi Du, Peng Gao, Aiming Hao, Steven Hoi, Zhaohui Hou, Shijie Huang, Dengyang Jiang, Yuming Jiang, Xin Jin, Liangchen Li, Zhen Li, Zhong-Yu Li, David Liu, Dongyang Liu, Qilong Wu, Feng Yu, Zechao Zhan, Chi Zhang, Shifeng Zhang, Ruikai Zhou, and Shilin Zhou. Z-Image: An efficient image generation foundation model with single-stream diffusion transformer. *arXiv preprint arXiv:2511.22699*, 2025.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

APPENDIX

A TECHNICAL COMPARISONS WITH PRIOR WORK

This appendix supplements Section 6 by distinguishing empirical source suitability from exact distribution recovery, and by identifying the optimizer designs used in our analysis.

A.1 NON-GAUSSIAN SOURCES AND EXACT DISTRIBUTION RECOVERY

Producing realistic images from a non-Gaussian source is a different claim from recovering the exact data distribution. Theorem 5 of Liang (2025) states that an exact-score probability-flow ODE for Gaussian-trained VP diffusion recovers the data law from any zero-mean symmetric terminal density with finite second moment. Our discrete, asymmetric source lies outside these assumptions. Even within the stated class, however, the following transport argument and counterexample show that the exact-recovery conclusion does not hold.

Let q_t be the marginal defining the exact-score velocity $v_t = f_t - \frac{1}{2}g(t)^2\nabla \log q_t$, and let μ_t be the law obtained from another terminal law r_T under the same deterministic ODE. Both densities satisfy

$$\partial_t \mu_t + \nabla\cdot(\mu_t v_t) = 0, \qquad \partial_t q_t + \nabla\cdot(q_t v_t) = 0. \tag{7}$$

For smooth densities with finite KL divergence and vanishing boundary terms, these equations give

$$\begin{aligned}\frac{d}{dt}\,\mathrm{KL}(\mu_t\|q_t) &= \int \mu_t v_t\cdot\nabla\log\frac{\mu_t}{q_t}\,dx - \int \mu_t\,\partial_t \log q_t\,dx \\ &= \int \mu_t\big(v_t\cdot\nabla\log\mu_t + \nabla\cdot v_t\big)\,dx \\ &= \int \nabla\cdot(\mu_t v_t)\,dx = 0,\end{aligned} \tag{8}$$

where we substitute $\partial_t \log q_t = -\nabla\cdot v_t - v_t\cdot\nabla\log q_t$. Equivalently, a common bijective flow preserves the KL divergence between the two input laws.

In Appendix H of the cited proof, the transition after Eq. (70) replaces $-\frac{1}{2}g(t)^2\nabla\log q_t\cdot(\nabla\log\mu_t - \nabla\log q_t)$ by $-\frac{1}{2}g(t)^2\|\nabla\log\mu_t - \nabla\log q_t\|^2$. These expressions differ, so the claimed dissipation identity in its Eq. (71) does not follow. Well-posedness and q_0 being the data law also do not imply $\mu_0 = q_0$ for another terminal law.

A counterexample with matched moments. Take a VP process with data law $q_0 = \mathcal{N}(0, I_d)$. Then $q_t = \mathcal{N}(0, I_d)$, its exact score is $-x$, and $f_t(x) = -\frac{1}{2}\beta(t)x$, $g(t)^2 = \beta(t)$ give $v_t(x) = 0$. Choose

$$r_T = \tfrac{1}{2}\mathcal{N}(0, \tfrac{1}{2}I_d) + \tfrac{1}{2}\mathcal{N}(0, \tfrac{3}{2}I_d). \tag{9}$$

This symmetric density has zero mean and covariance I_d, but is non-Gaussian. The ODE is the identity map, so $\mu_0 = r_T \neq q_0$ and $\mathrm{KL}(\mu_0\|q_0) > 0$. Thus even matching the first two moments does not imply exact recovery. Our source study addresses which inputs retain generation quality in pretrained models, with embedding statistics helping explain their tolerance; it does not require the output law to be independent of the source distribution.

A.2 ORTHOGONALIZED UPDATES AND PRIOR OPTIMIZER DESIGNS

Muon (Jordan et al., 2024) continues earlier work on spectral descent and gradient orthogonalization (Carlson et al., 2015; Tuddenham et al., 2022; Bernstein & Newhouse, 2024). Muon has also been applied to diffusion latents for attention-distillation objectives (Chen et al., 2025). We study initial-noise optimization with pretrained reward models and interpret orthogonalization through the source's tolerance to changes in its entry-wise distribution and sensitivity to arrangement.

Muse (Chang et al., 2026a) studies how the matrix representation supplied to the polar map changes optimizer geometry, including nearest-square reshaping. We adopt that representation choice for latent momentum. The comparison with Layout Muon in Appendix E.4 asks how representation affects incoming structure and generation quality; the reshape itself is inherited from this prior design.

Muon-C (Qing & Li, 2026) applies a DFT along a convolution kernel's spatial axes and independently orthogonalizes the resulting frequency-wise channel maps. Related frequency-domain constructions are also discussed by Cesista (2026). These provide precedents for combining a transform with blockwise polar operations. Our Fourier-block control acts on latent momentum to examine whether combining Fourier and SV processing improves the trade-off in Appendix E.4. It serves this analysis of update structure.

GPER (Hwang & Sung, 2026) already uses spectral gradient preconditioning. The additional distinction between satisfying selected noise statistics and suppressing structure in incoming updates is developed with direct comparisons in Appendix E.

B SOURCE CONDITIONS AND EMBEDDING EVIDENCE

We support Section 3 in three stages: define the source conditions, explain and measure their consequences at the first embedding, and test generation quality and the limits of those conditions. We compare with Liang (2025) in Appendix A.1.

B.1 SOURCE CONDITIONS AND CONSTRUCTION

We use three conditions: matched moments (M), Gaussian-reference token energy (E), and controlled scalar tails (T). Condition (M) gives exact embedding identities; (E) and (T) control token excursions and rare scalar contributions. Entries are placed randomly. These conditions define an analytical class, rather than a necessary and sufficient criterion for successful generation.

Definition 1 (Admissible source class for analysis)**.** Let a latent consist of $N \geq 2$ tokens of dimension d, with scalar entries $\mathrm{X}_{ti} \overset{\text{iid}}{\sim} F$ in the independent-source model. Write $\mathrm{S}_d = \sum_{i=1}^{d} \mathrm{X}_i^2$ for the energy of one token, let $[L_d, U_d]$ be the central interval of the χ^2_d law that carries probability $1 - 1/N$, so that a Gaussian token leaves it with probability exactly $1/N$, and let $\mathrm{Z} \sim \mathcal{N}(0,1)$. The law F is admissible for (N, d) with tail threshold B if

$$\mathbb{E}[\mathrm{X}] = 0, \qquad \mathrm{Var}(\mathrm{X}) = 1, \tag{M}$$

$$\Pr[L_d \leq \mathrm{S}_d \leq U_d] \geq 1 - 1/N, \tag{E}$$

$$\Pr[|\mathrm{X}| > t] \leq \Pr[|\mathrm{Z}| > t] \quad \text{for every } t \geq B. \tag{T}$$

Theory and experimental constructions. We distinguish three uses of an entry-wise distribution throughout this appendix. *Independent entries* define the analytical model above and the exact embedding identities below. *Fixed-multiset placement* is used in the quantitative source experiments: we construct $M = Nd$ scalar values, standardize them to mean zero and mean square one, and place them by a uniform random permutation. Sampling one token is then sampling without replacement; Appendix B.3 gives the variance correction and Gaussian approximation. The mean and scale interventions deliberately change this normalization. *Gaussian-rank coupling* constructs the qualitative two-point examples and their paired Gaussian controls from the same random ranks, without subsequent standardization (Appendix B.5.3).

We use $B = 8$ in the source-class checks. Condition (T) is checked against the named parent law; a finite quantile approximation to an unbounded heavy-tailed law does not change that parent-law condition. For non-Gaussian finite collections, (E) is evaluated using independent draws from their empirical entry-wise distribution; this class check is distinct from the actual without-replacement placement. We describe the numerical procedure and uncertainty of these checks in Appendix B.5.

Gaussian token-energy reference. The endpoints L_d, U_d are the χ^2_d quantiles at $1/(2N)$ and $1 - 1/(2N)$. The source experiments use $(N, d) = (1024, 64)$ for FLUX.1-schnell and Z-Image-Turbo, giving $[L_d, U_d] \approx [33.1, 108.0]$, and $(4096, 128)$ for FLUX.2-klein-4B, giving $[77.4, 195.1]$. For a Gaussian latent, one token lies outside this interval in expectation. For the two-point law in Equation 2, $\mathrm{S}_d = d/3 + \frac{8}{3}\mathrm{K}_d$ with $\mathrm{K}_d \sim \mathrm{Binomial}(d, 1/4)$; its expected count is about a tenth of the Gaussian allowance at both geometries.

The joint role of (E) and (T), including a bound on energy hidden in rare entries and examples separating the two conditions, is examined with the sensitivity controls in Appendix B.6.2.

B.2 EXACT EMBEDDING STATISTICS

The first embedding preserves some Gaussian-reference statistics exactly, while other quantities depend on the entry-wise distribution or require an approximation. The overview below separates these claims for the independent-source model.

Status	Embedding quantities	Basis
Exact agreement	Feature mean and covariance; mean centered token energy; first two moments of centered inner products between distinct tokens	Independence and (M)
Distribution-dependent	Channel skewness and kurtosis; covariance fluctuations; token-energy variance	Higher source moments and embedding weights
Approximate or empirical	Standardized channel CDFs; full token-energy and inter-token distributions	Channelwise bounds and finite-dimensional measurements

The proposition below gives the exact identities and distribution-dependent corrections. Appendix B.3 treats channelwise approximation and finite-population corrections; Appendix B.4 reports the measurements.

Write $\mu_3 = \mathbb{E}[\mathrm{X}^3]$, $\kappa_4 = \mathbb{E}[\mathrm{X}^4] - 3$, $\boldsymbol{w}_k$ for the k-th column of $\boldsymbol{W}$, and $\boldsymbol{a}_k = \boldsymbol{w}_k/\|\boldsymbol{w}_k\|_2$. Set $\boldsymbol{G} = \boldsymbol{W}\boldsymbol{W}^\top/K$, $\boldsymbol{Q} = \boldsymbol{W}\boldsymbol{W}^\top$, $\boldsymbol{c} = \boldsymbol{W}\boldsymbol{b}/K$, and $\beta = \|\boldsymbol{b}\|_2^2/K$. The weight factors $\lambda(\boldsymbol{w}_k) = \sum_i a_{ki}^4$ and $s_3(\boldsymbol{w}_k) = \sum_i a_{ki}^3$ measure how the input's fourth and third moments enter the embedding.

Proposition 3 (Embedding statistics and distribution-dependent corrections)**.** *Let $\mathbf{y}_t = \boldsymbol{W}^\top\mathbf{x}_t + \boldsymbol{b}$, with fixed $\boldsymbol{W} \in \mathbb{R}^{d\times K}$ and $\boldsymbol{b} \in \mathbb{R}^K$, and let all input entries be i.i.d. and satisfy (M). Then*

$$\mathbb{E}[\mathbf{y}_t] = \boldsymbol{b}, \qquad \mathrm{Cov}(\mathbf{y}_t) = \boldsymbol{W}^\top\boldsymbol{W}, \qquad \mathbb{E}\big[\|\mathbf{y}_t - \boldsymbol{b}\|_2^2\big] = \|\boldsymbol{W}\|_F^2. \tag{10}$$

The higher-order formulas below assume the corresponding moments exist; the variance formulas involving κ_4 require a finite fourth moment. Write $w_{ki} = (\boldsymbol{w}_k)_i$.

(i) One element. *For $m \geq 3$, the m-th cumulant is $\kappa_m(\mathrm{y}_{tk}) = \kappa_m(\mathrm{X})\sum_i w_{ki}^m$; in standardized form the skewness is $\mu_3 s_3(\boldsymbol{w}_k)$ and the excess kurtosis is $\kappa_4\lambda(\boldsymbol{w}_k)$.*

(ii) Two elements. $\mathrm{Cov}(\mathrm{y}_{tk}, \mathrm{y}_{tl}) = \boldsymbol{w}_k^\top\boldsymbol{w}_l$, *and the empirical covariance over tokens, $\hat{\mathrm{C}}_{kl} = N^{-1}\sum_t(\mathrm{y}_{tk} - b_k)(\mathrm{y}_{tl} - b_l)$, satisfies $\mathbb{E}[\hat{\mathrm{C}}_{kl}] = \boldsymbol{w}_k^\top\boldsymbol{w}_l$ and*

$$N\,\mathrm{Var}(\hat{\mathrm{C}}_{kl}) = \|\boldsymbol{w}_k\|_2^2\|\boldsymbol{w}_l\|_2^2 + (\boldsymbol{w}_k^\top\boldsymbol{w}_l)^2 + \kappa_4\sum_i w_{ki}^2 w_{li}^2. \tag{11}$$

(iii) One row. *The embedded token energy $\mathrm{v}_t = K^{-1}\|\mathbf{y}_t\|_2^2 = \mathbf{x}_t^\top\boldsymbol{G}\mathbf{x}_t + 2\boldsymbol{c}^\top\mathbf{x}_t + \beta$ has $\mathbb{E}[\mathrm{v}_t] = \mathrm{tr}\,\boldsymbol{G} + \beta$ and*

$$\mathrm{Var}(\mathrm{v}_t) = 2\|\boldsymbol{G}\|_F^2 + 4\|\boldsymbol{c}\|_2^2 + \kappa_4\sum_i G_{ii}^2 + 4\mu_3\sum_i G_{ii}c_i. \tag{12}$$

(iv) Two rows. *For $t \neq s$ the centered dot product $\mathrm{D}_{ts} = (\mathbf{y}_t - \boldsymbol{b})^\top(\mathbf{y}_s - \boldsymbol{b}) = \mathbf{x}_t^\top\boldsymbol{Q}\mathbf{x}_s$ has $\mathbb{E}[\mathrm{D}_{ts}] = 0$ and $\mathbb{E}[\mathrm{D}_{ts}^2] = \|\boldsymbol{Q}\|_F^2$, with no dependence on the law. Higher moments can differ: $\mathbb{E}[\mathrm{D}_{ts}^3] = \mu_3^2\sum_{i,j}Q_{ij}^3$, and $\mathbb{E}[\mathrm{D}_{ts}^4]$ depends on κ_4.*

For $\mathbf{g} \sim \mathcal{N}(0, \boldsymbol{I}_d)$ every κ_4 and μ_3 term vanishes, so the remaining expression gives the Gaussian value of each quantity.

Proof of Propositions 1 and 3. Under (M) and independence, $\mathbb{E}[\mathbf{x}] = 0$ and $\mathbb{E}[\mathbf{x}\mathbf{x}^\top] = \boldsymbol{I}_d$, so $\mathbb{E}[\mathbf{y}] = \boldsymbol{b}$ and $\mathrm{Cov}(\mathbf{y}) = \boldsymbol{W}^\top\mathrm{Cov}(\mathbf{x})\boldsymbol{W} = \boldsymbol{W}^\top\boldsymbol{W}$; the Gaussian reference has the same input covariance. For (i), cumulants of independent summands add and $\kappa_m(w_{ki}\mathrm{X}_i) = w_{ki}^m\kappa_m(\mathrm{X})$; dividing by $\|\boldsymbol{w}_k\|_2^m$ gives the standardized forms, with $\kappa_3(\mathrm{X}) = \mu_3$ and $\kappa_4(\mathrm{X}) = \kappa_4$. For (ii), $\hat{\mathrm{C}}_{kl}$ is an average of N independent copies of $(\boldsymbol{w}_k^\top\mathbf{x})(\boldsymbol{w}_l^\top\mathbf{x})$, and for independent entries satisfying (M)

$$\begin{aligned}\mathbb{E}\big[(\boldsymbol{u}^\top\mathbf{x})^2(\boldsymbol{v}^\top\mathbf{x})^2\big] &= \mu_4\sum_i u_i^2v_i^2 + \sum_{i\neq j}\big(u_i^2v_j^2 + 2u_iv_iu_jv_j\big)\\ &= \|\boldsymbol{u}\|_2^2\|\boldsymbol{v}\|_2^2 + 2(\boldsymbol{u}^\top\boldsymbol{v})^2 + \kappa_4\sum_i u_i^2v_i^2;\end{aligned}$$

subtracting $(\boldsymbol{u}^\top \boldsymbol{v})^2$ gives Equation 11. For (iii), $\mathrm{v}_t = \mathbf{x}^\top \boldsymbol{G}\mathbf{x} + 2\boldsymbol{c}^\top \mathbf{x} + \beta$, so $\mathbb{E}[\mathrm{v}_t] = \mathrm{tr}(\boldsymbol{G}\,\mathbb{E}[\mathbf{x}\mathbf{x}^\top]) + 0 + \beta = \mathrm{tr}\,\boldsymbol{G} + \beta$. For (iv), $\mathrm{D}_{ts} = \mathbf{x}^\top \boldsymbol{Q}\mathbf{x}'$ with $\mathbf{x}, \mathbf{x}'$ independent, so $\mathbb{E}[\mathrm{D}_{ts}] = 0$ and $\mathbb{E}[(\mathbf{x}^\top \boldsymbol{Q}\mathbf{x}')^2] = \mathbb{E}[\mathbf{x}^\top \boldsymbol{Q}\,\mathbb{E}[\mathbf{x}'\mathbf{x}'^\top]\,\boldsymbol{Q}\mathbf{x}] = \mathrm{tr}(\boldsymbol{Q}^2) = \|\boldsymbol{Q}\|_F^2$, using only $\mathbb{E}[\mathbf{x}\mathbf{x}^\top] = \mathbb{E}[\mathbf{x}'\mathbf{x}'^\top] = \boldsymbol{I}_d$. Expanding the third moment leaves only terms with all three indices equal within each independent token, giving $\mathbb{E}[\mathrm{D}_{ts}^3] = \mu_3^2 \sum_{i,j} Q_{ij}^3$. For the fourth moment, conditioning on one token reduces the calculation to fourth moments of weighted sums; these involve the input variance and κ_4. To obtain Equation 12, we expand $\mathrm{v}_t - \mathbb{E}[\mathrm{v}_t] = \sum_i \{G_{ii}(\mathrm{X}_i^2 - 1) + 2c_i\mathrm{X}_i\} + 2\sum_{i<j} G_{ij}\mathrm{X}_i\mathrm{X}_j$. Distinct diagonal/linear groups are independent; every covariance involving an off-diagonal term vanishes because a mean-zero coordinate appears exactly once. Summing the remaining variances gives $(\mu_4 - 1)\sum_i G_{ii}^2 + 4\|\boldsymbol{c}\|_2^2 + 4\mu_3 \sum_i G_{ii}c_i + 4\sum_{i<j} G_{ij}^2$, which equals Equation 12 since $\mu_4 - 1 = 2 + \kappa_4$ and $2\|\boldsymbol{G}\|_F^2 = 2\sum_i G_{ii}^2 + 4\sum_{i<j} G_{ij}^2$. □

We use the same calculation for mean and scale interventions. For input mean μ and variance σ^2, $\mathbb{E}[\mathbf{y}] = \boldsymbol{b} + \mu \boldsymbol{W}^\top \mathbf{1}$ and $\mathrm{Cov}(\mathbf{y}) = \sigma^2 \boldsymbol{W}^\top \boldsymbol{W}$. A mean shift therefore adds the same vector to every embedded token.

From token energy to centered embedding energy. For i.i.d. entries satisfying (M) with excess kurtosis κ_4,

$$\mathrm{Var}(\mathrm{S}_d) = d(2 + \kappa_4), \qquad \mathrm{Var}\big(\|\mathbf{y} - \boldsymbol{b}\|_2^2\big) = 2\|\boldsymbol{Q}\|_F^2 + \kappa_4 \sum_i Q_{ii}^2, \quad \boldsymbol{Q} = \boldsymbol{W}\boldsymbol{W}^\top. \tag{13}$$

The first identity follows by summing independent squared coordinates; the second is Equation 12 with the bias removed and the normalization by K undone. For the law in Equation 2, $\mathbb{E}[\xi^4] = 7/3$ and $\kappa_4 = -2/3$. Thus $\mathrm{Var}(\mathrm{S}_d) = 4d/3$, compared with $2d$ for Gaussian entries, while the centered embedding-energy variance is reduced by $\frac{2}{3}\sum_i Q_{ii}^2$. Its mean remains $\mathrm{tr}\,\boldsymbol{Q}$. These moment identities hold for i.i.d. entries; we discuss sampling from a fixed multiset in Appendix B.3.

Matching low-order moments does not imply matching tails. The law with probability $1/6$ at each of $\pm\sqrt{3}$ and probability $2/3$ at zero has $\mu_3 = \kappa_4 = 0$. It therefore matches the displayed Gaussian covariance-fluctuation and token-energy variance formulas, and the leading skewness/kurtosis correction to a channel CDF vanishes. This does not match all higher cumulants or ensure (E). In fact, for $d = 64$, $\mathrm{S}_d = 3\mathrm{K}$ with $\mathrm{K} \sim \mathrm{Binomial}(64, 1/3)$, and $\Pr[\mathrm{S}_d < 33.1] = \Pr[\mathrm{K} \le 11] \approx 0.00316 > 1/1024$. The law satisfies (T) for $B \ge \sqrt{3}$, but fails the token-energy condition at this geometry. It remains a useful moment-matching example in Table 5.

B.3 Channelwise Gaussian approximation

Since each embedding channel is a weighted sum, classical central-limit and invariance arguments (Lindeberg, 1922; Chatterjee, 2006) motivate its Gaussian approximation. We quantify this agreement channelwise. For independent entries satisfying (M), write $\rho_3 = \mathbb{E}[|\mathrm{X}|^3]$, $\alpha(\boldsymbol{w}) = \sum_i |w_i|^3 / \|\boldsymbol{w}\|_2^3$, and $d_K(U, V) = \sup_z |\Pr(U \le z) - \Pr(V \le z)|$. The weighted Berry–Esseen inequality gives

$$d_K\left(\frac{\sum_i w_i \mathrm{X}_i}{\|\boldsymbol{w}\|_2}, \mathrm{Z}\right) \le 0.56\,\rho_3\,\alpha(\boldsymbol{w}), \qquad \mathrm{Z} \sim \mathcal{N}(0,1), \tag{14}$$

for $\boldsymbol{w} \neq 0$ and finite ρ_3 (Shevtsova, 2010; Pinelis, 2017). It follows by applying the independent-summand bound to $w_i\mathrm{X}_i/\|\boldsymbol{w}\|_2$.

As the number of mixed entries grows. For $a_i = w_i/\|\boldsymbol{w}\|_2$, we have $\alpha(\boldsymbol{w}) = \sum_i |a_i|^3 \le \max_i |a_i|$. Thus, along a sequence with uniformly bounded ρ_3 and $\max_i |a_i| \to 0$, the standardized channel converges to a Gaussian. This is the weighted central-limit mechanism: no single entry retains a nonvanishing share of the variance (Lindeberg, 1922; Pitman & Rosenberg, 2002). Under (M) and (T),

$$\rho_3 \le B + \mathbb{E}\big[|\mathrm{Z}|^3 \mathbf{1}\{|\mathrm{Z}| > B\}\big],$$

so a fixed tail threshold controls the source moment in this bound. For $|\mathrm{X}| \le B$, the bound $\rho_3 \le B$ holds directly.

Random placement of a fixed multiset. The quantitative experiments randomly permute $M = Nd$ values with mean zero and mean square one. Let π be a uniform permutation, $\mathrm{Y}_\pi = \sum_{i=1}^{d} w_i v_{\pi(i)}$, and $\rho_3 = M^{-1} \sum_j |v_j|^3$. Combinatorial CLTs treat such permutation sums directly (Chen & Fang, 2015); here we use an elementary coupling. Sampling with replacement gives the independent bound above. When conditioned on distinct indices, the ordered sample follows exactly the same law as sampling without replacement. Consequently,

$$d_K\left(\frac{\mathrm{Y}_\pi}{\|\boldsymbol{w}\|_2}, \mathrm{Z}\right) \le 0.56\,\rho_3\,\alpha(\boldsymbol{w}) + q_{M,d}, \qquad q_{M,d} = 1 - \frac{(M)_d}{M^d} \le \frac{d(d-1)}{2M}, \tag{15}$$

where $(M)_d = M(M-1)\cdots(M-d+1)$ and $1 \le d \le M$. The last term bounds the probability of an index collision. In particular, the Gaussian limit also holds for these permutation sums when the conditions above hold and $d^2/M \to 0$.

The exact finite-population moments are

$$\mathbb{E}[\mathrm{Y}_\pi] = 0, \qquad \mathrm{Var}(\mathrm{Y}_\pi) = \frac{M\|\boldsymbol{w}\|_2^2 - (\sum_i w_i)^2}{M-1}. \tag{16}$$

Indeed, distinct sampled entries have covariance $-1/(M-1)$; expanding the variance gives Equation 16. Its relative difference from $\|\boldsymbol{w}\|_2^2$ lies in $[-(d-1)/(M-1), 1/(M-1)]$, at most 9.61×10^{-4} in magnitude for $M = 65{,}536$ and $d = 64$. The empirical channel comparisons use this exact variance correction.

Finite-dimensional evidence. Our generators use fixed embedding widths, so we measure the approximation directly on every output channel in Appendix B.4. The theoretical bound describes the roles of source tails and weight mixing, while the CDFs and channelwise KS values show the actual agreement at finite dimension. For context, the largest normalized coefficient on FLUX.1-schnell reaches 0.48, and the two-point input's exact channelwise support limits range from 6.55 to 11.07 standard deviations from the mean. These support limits describe attainable values, rather than their probabilities.

Condition (T) also permits standard sub-Gaussian bounds for weighted sums and quadratic forms (Vershynin, 2012; Rudelson & Vershynin, 2013). Such bounds control tails but do not by themselves certify (E) at its fixed Gaussian quantiles; membership is checked separately. Channelwise approximation concerns one-dimensional marginals. When $\boldsymbol{W}$ has rank d, the affine embedding is injective, so discrete and continuous inputs still give distinct joint embedding laws. Generation quality is evaluated separately in Appendix B.5.

B.4 Empirical embedding evidence

We test the moment identities, channelwise approximation, and higher-order corrections on frozen embedding layers. These comparisons distinguish statistics that match exactly under the source model from distributional similarities that must be measured at finite dimension.

B.4.1 Three-generator verification of the first-embedding moments

We evaluate the frozen first affine layers of all three generators using over 2^{20} tokens per law and model, drawn from $1{,}030$ latents for FLUX.1 and Z-Image and 260 latents for FLUX.2. We measure both i.i.d. draws from the experimental finite entry-wise distribution and random arrangements of a fixed multiset. The latter use finite-population corrections for covariance, centered token energy, and the token-pair dot mean.

Table 3 reports the two-point source under fixed-multiset arrangement, using every output channel. The feature means, covariance, centered token magnitude, and token-pair second moments agree with their predictions. The largest covariance error is 1.06%, comparable to 1.07% for the i.i.d. Gaussian control at the same sample size. We observe the same pattern for the i.i.d. two-point control.

Table 3: First-embedding moment check for the fixed-multiset two-point source across three pretrained generators. Mean error is the RMS discrepancy from the predicted feature mean, divided by the predicted channel standard deviation; covariance error is the Frobenius relative discrepancy over all output channels. Energy is the measured squared norm of a centered token divided by its finite-population prediction. Dot^2 is the measured centered distinct-token dot-product second moment divided by the i.i.d. reference $\sigma^4\|\boldsymbol{W}\boldsymbol{W}^\top\|_F^2$; the fixed-multiset dot mean uses its exact finite-population correction. Intervals are 95% t intervals over 1,030 arrays (FLUX.1, Z-Image) or 260 arrays (FLUX.2).

Model	Mean error / SD	Cov. error	Energy ratio [95% CI]	Dot^2 ratio [95% CI]
FLUX.1-schnell	0.00101	0.485%	1.0003 [0.9999, 1.0008]	0.9987 [0.9949, 1.0025]
Z-Image-Turbo	0.00094	0.681%	1.0000 [0.9998, 1.0001]	0.9993 [0.9964, 1.0023]
FLUX.2-klein-4B	0.00120	1.060%	1.0000 [0.9999, 1.0001]	1.0026 [0.9997, 1.0054]

B.4.2 Source and embedding channel CDFs

Main-text illustration. Figure 4 uses an asymmetric three-point law on $\{-1.8, 0, 0.7\}$ with probabilities $\{2/9, 13/63, 4/7\}$. It has mean zero and variance one, satisfies (T) at $B = 8$, and has token-energy excursion probability $2.17 \times 10^{-4} < 1/1024$ at $(N, d) = (1024, 64)$, satisfying (E). We select FLUX.1-schnell channel 1942 uniformly at random before drawing the inputs. For 250,000 i.i.d. tokens per law, we subtract the channel bias and divide by its weight norm. The empirical KS distance to the standard Gaussian is 0.00295 for the three-point input and 0.00126 for an independently sampled Gaussian control. This is a single-channel illustration; the following comparisons examine the two-point source across models and all output channels.

Cross-model and all-channel checks. Figure 10 compares the two-point input CDF with a prespecified embedding channel from each model. The channels are selected by the median rank of the frozen-weight statistic $\alpha(\boldsymbol{w}_k)$, independently of the source samples. After centering and scaling by their predicted moments, their CDFs are much closer to Gaussian.

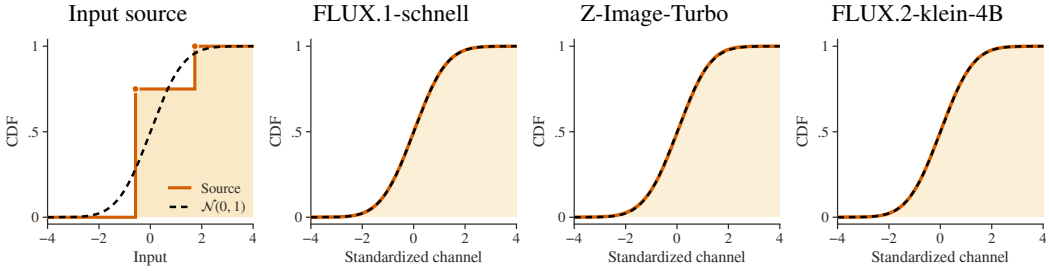


Figure 10: The fixed two-point input has a stepped CDF, but a prespecified first-embedding channel on each generator is much closer to the standard Gaussian CDF (black dashed). Orange shows the empirical CDF from more than 2^{20} tokens per model; the output channels are 1511, 376, and 1680 in the displayed order. Table 4 reports the distribution of residual discrepancies across all channels.

To measure the full range of channel behavior, we compute the empirical KS distance to Φ for every output channel: 3,072, 3,840, and 3,072 in model order. KS is the sample counterpart of d_K; a value of 0.01 means a maximum CDF separation of one percentage point. Fixed-multiset outputs are standardized using Equation 16. Table 4 reports the channelwise median, 99th percentile, and observed maximum, using an i.i.d. Gaussian control as the finite-sampling reference.

For the two-point law, input KS is 0.4681, while the embedding-channel median is at most 0.0036 across the three models. Figure 11 shows the residual variation across channels. Across all tested sources satisfying (M), (E), and (T), including the boundary probes, the largest observed channelwise KS is 0.0260. The shaded exponential and lognormal controls fail (E) and (T). The lognormal control produces larger discrepancies after embedding despite being closer to Gaussian at the input. Thus, input CDF distance alone does not predict the quality of this approximation.

Table 4: Channelwise KS to Φ ($\downarrow$) over all 3,072 or 3,840 output channels, with at least 2^{20} tokens per channel. Input KS describes the scalar source. The first group passes (M), (E), and (T) at both tested geometries; the shaded controls fail (E) and (T). Boundary probes were selected near the (E) limit at $(1024, 64)$. Gaussian i.i.d. provides the sampling floor; max is the observed maximum, not a bound.

		FLUX.1			Z-Image			FLUX.2		
Source law	Input KS	med	p99	max	med	p99	max	med	p99	max
Sources satisfying (M), (E), and (T)										
Gaussian (i.i.d. reference)	—	0.0008	0.0015	0.0020	0.0008	0.0016	0.0020	0.0008	0.0015	0.0020
Gaussian quantile multiset	$< 10^{-4}$	0.0008	0.0016	0.0020	0.0008	0.0015	0.0022	0.0009	0.0017	0.0022
Two-point $\{-1/\sqrt{3}, +\sqrt{3}\}$	0.4681	0.0031	0.0120	0.0190	0.0036	0.0124	0.0221	0.0020	0.0072	0.0108
Five-level Gaussian quantization	0.1303	0.0015	0.0025	0.0029	0.0015	0.0024	0.0040	0.0011	0.0020	0.0028
Three-point boundary ($a = 1.67$)	0.3207	0.0008	0.0016	0.0021	0.0009	0.0016	0.0020	0.0009	0.0016	0.0019
Two-point boundary ($p = 0.218$)	0.4832	0.0036	0.0141	0.0222	0.0041	0.0144	0.0260	0.0023	0.0084	0.0128
Outside the source class: (E) and (T) fail										
Centered exponential	0.1587	0.0077	0.0190	0.0282	0.0083	0.0198	0.0269	0.0046	0.0119	0.0171
Centered lognormal	0.2287	0.0338	0.0491	0.0635	0.0349	0.0500	0.0594	0.0227	0.0333	0.0424

Input KS for the two shaded controls is computed from their population CDFs.

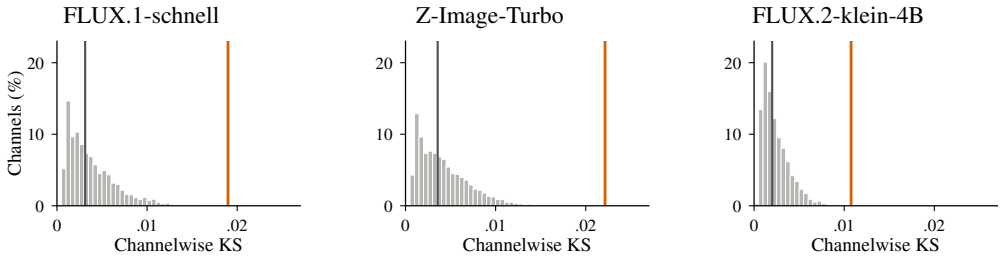


Figure 11: Channelwise KS to Φ for the fixed two-point source, over every first-embedding output channel. Gray and orange lines mark the median and observed maximum, respectively. All panels use the same axes; each channel has more than 2^{20} tokens.

Source-class boundary probes. The three-point and two-point probes approach the (E) limit at $(N, d) = (1024, 64)$. They are selected by convolving the squared-atom law of each rounded-and-standardized finite multiset, scanning $a = 1.20, 1.21, \ldots, 6.00$ and $p = 0.020, 0.022, \ldots, 0.500$. The selected laws are listed in Table 4. We use these probes from each family to test the approximation near the condition's boundary at the generators' actual embedding dimensions.

B.4.3 Token magnitudes and inter-token distributions

Figure 12 compares full empirical distributions of centered token energy and token-pair inner products. The two-point curves remain close to the Gaussian-quantile controls across all three models. The lognormal control instead has a much broader token energy distribution even though its mean is matched. This comparison distinguishes the exact moment identities from the distributional agreement observed for the working two-point source.

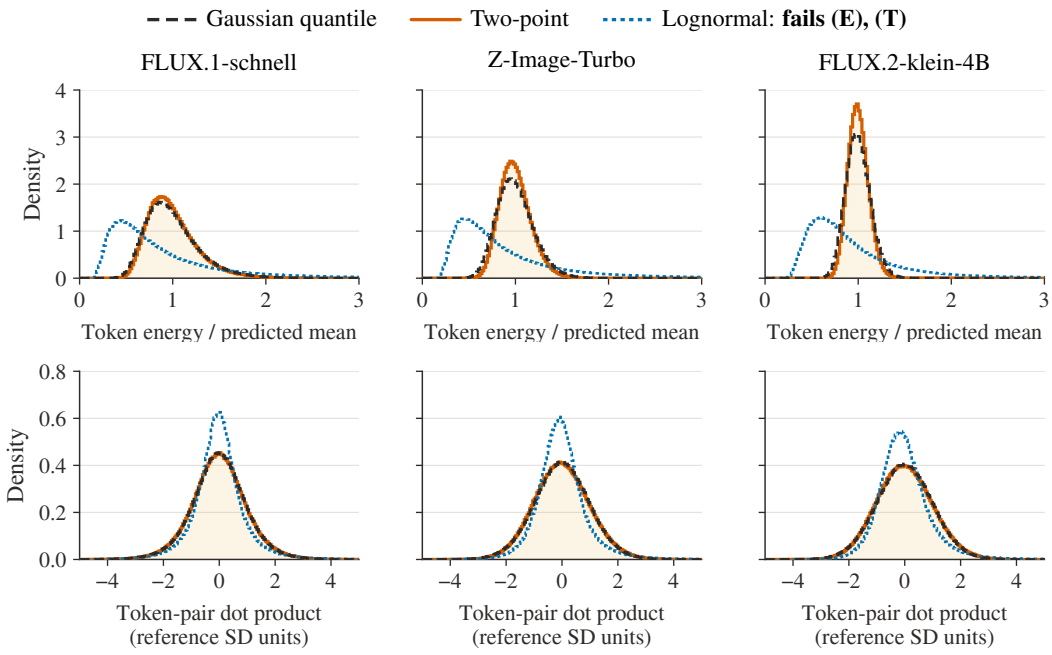


Figure 12: Centered embedding statistics beyond their means. The two-point source satisfies (M), (E), and (T); the lognormal comparison satisfies (M) but **fails (E) and (T)**. Top: token squared norm divided by its finite-multiset prediction. Bottom: distinct-token inner products centered by their finite-multiset predicted mean and divided by the i.i.d. reference standard deviation. Each curve uses over one million tokens or token pairs from fixed-multiset arrangements. Histograms are normalized by all samples, including those outside the displayed ranges. The two-point source remains close to the Gaussian-quantile control, while the lognormal source, which lies outside the class, has a different profile.

B.4.4 Higher-order statistics: predictions and measurements

Table 5 compares the predicted and measured embedding kurtosis and token-energy variance. We use the frozen first token-embedding layer of FLUX.1-schnell, including its bias, with 64 independent latents of $N = 1024$ tokens per law, $d = 64$, and all 3,072 output channels. Inputs are i.i.d. We compute the predictions using the moments and weight factors in Proposition 3, with $\mathrm{v}_t = K^{-1}\|\mathbf{y}_t\|_2^2$.

Across the bounded sources, the measurements follow the corrections that depend on the entry-wise distribution. Input kurtosis is strongly reduced at the channel level, while differences in token-energy variance remain visible. The shaded controls outside the class illustrate both moment matching without (E) and larger higher-order deviations.

Table 5: Higher-order embedding statistics on FLUX.1-schnell: predicted (pred.) and measured (meas.). Shaded rows violate the source conditions.

			Element ex. kurtosis		$10^4\,\mathrm{Var}(\mathrm{v}_t)$	
Law	μ_3	κ_4	pred.	meas.	pred.	meas.
Gaussian	$+0.00$	$+0.00$	$+0.000$	-0.007	14.74	14.63
Rademacher	$+0.00$	-2.00	-0.081	-0.085	9.26	9.32
Uniform	$+0.00$	-1.20	-0.049	-0.054	11.45	11.37
Two-point, $p = 1/4$	$+1.15$	-0.67	-0.027	-0.033	12.87	12.67
Outside the source class: (E) or (T) fails						
Three-point, $\kappa_4 = 0$	$+0.00$	-0.00	-0.000	-0.004	14.74	14.83
Centered exponential	$+2.00$	$+6.00$	$+0.244$	$+0.235$	31.09	30.76
Lognormal	$+6.18$	$+110.94$	$+4.510$	$+4.016$	318.19	357.07

B.5 GENERATION QUALITY AND DIVERSITY

B.5.1 SOURCE EVALUATION PROTOCOL

We evaluate 83 source interventions on each of three pretrained generators, using ten prompts and thirty images for each condition–prompt pair. The interventions comprise 35 entry-wise distributions, 13 mean shifts, nine scale changes, 15 arrangement or coherent-dose controls, and 11 token-energy controls. We hold the generator weights, prompt, and sampling settings fixed while changing the source. The model settings and prompt list are in Appendix F.1 and Table 10.

For each condition, we subtract each prompt's $1{,}000$-image Gaussian reference mean from its 30-image condition mean and average the ten paired differences. The 95% interval is the mean plus or minus 2.262 prompt-level standard errors (t_9), conditional on the estimated reference means. We also report $n_{1\%}$, the number of the 300 condition images below their prompt-specific Gaussian first PickScore percentile. These estimated thresholds have shared uncertainty within each prompt. These counts record how many scores fall below the reference threshold; intervals containing zero and counts near the nominal 1% rate do not establish equivalence.

Table 6 reports PickScore and ImageReward differences and PickScore lower-tail counts for the entry-wise distributions. Figures 14 and 15 show the sensitivity sweeps; arrangement controls are interpreted in Appendix C. The Gaussian quantile control is distinct from the independently sampled Gaussian reference.

In the scalar-law comparison, μ_3 and κ_4 are centered, standardized moments of the realized multiset. We check (M) using its exact mean and variance, and (T) using the parent law at $B = 8$, as defined in Appendix B.1. The (E) check uses the analytic Gaussian CDF for Gaussian, mean, and scale controls. For other laws, it uses independent draws from the finite multiset with a fixed seed and a 99.9% Wilson interval wholly to one side of $1/N$. Checks are performed at both model geometries specified in Appendix B.1; generation instead permutes the fixed multiset without replacement.

Table 6 groups the 15 entry-wise distributions satisfying all conditions before the remaining laws, with the Gaussian control first. Within these groups, sources are ordered by the number of conditions satisfied and the normalized token-energy excursion rate $\max_{(N,d)} N\hat{p}_E$. Token-arrangement and token-energy interventions are not counted as additional entry-wise distributions.

Source definitions and the main table. Table 1 contains the sources in Figures 3 and 1, a Gaussian reference, and selected outside-class controls. Its Gaussian reference differences are zero by definition. Quantitative scalar constructions are standardized after forming their values, except for the deliberate mean and scale interventions. Their skewness and excess kurtosis therefore describe shape, rather than deviations in mean or scale.

Rademacher puts equal mass on $\{-1, +1\}$. A two-point law is named by the probability p of its upper atom, which fixes the pair: *Two-point,* $p = 1/3$ is $\{-1/\sqrt{2}, +\sqrt{2}\}$ and *Two-point,* $p = 1/4$ is $\{-1/\sqrt{3}, +\sqrt{3}\}$, the law of Equation 2; we use the same naming convention for the out-of-class *Two-point,* $p = 0.02$. *Four-point* puts weights $2 : 4 : 3 : 1$ on the atoms -1.4, -0.3, $+0.6$, $+1.9$ before standardization. *Uniform* and *Arcsine* are those laws on a bounded interval. *Wave* has density proportional to $1 + 0.9\cos(2\pi x)$ on $[-1, 1]$; *Skewed bumps* is the mixture $2.4\mathcal{N}(-0.55, 0.22^2) + \mathcal{N}(1.15, 0.30^2)$ restricted to $[-1.5, 2.2]$; *Wedge* has density proportional to $(x + 3.2)^2$ on $[-2.6, 2.6]$. The three shaded rows are the reflected $\Gamma(4)$, the standardized lognormal, and the Gaussian shifted to mean $+0.20$. Table 6 reports the scalar-law responses, and Figure 14 shows the mean-shift sweep.

B.5.2 ENTRY-WISE DISTRIBUTION COVERAGE AND GENERATION QUALITY

The full 35-law comparison in Table 6 extends the main table across discrete, continuous, symmetric, and asymmetric sources. Many bounded laws retain small reward changes and low-tail counts near the Gaussian reference. Sources outside the class include both those that work and those that degrade generation, with degradation depending on the generator. Appendix B.6 examines the mean, scale, and token-energy interventions; arrangement and coherent-dose results remain in Appendix C.

The table separates sources satisfying all three conditions from those violating (E) or (T), with (M) fixed, and retains every measured response, including working sources outside the class.

Table 6: Generation under 35 entry-wise distributions satisfying (M): zero mean and unit variance. Groups distinguish sources satisfying all conditions from those violating (E) or (T); checks use each model's geometry. μ_3/κ_4 denote skewness/excess kurtosis. Entry-wise distribution plots show FLUX.1 source mass/density on $[-4, 4]$ (triangles: off-range atoms; dashed black: Gaussian density for continuous laws). ΔPick/ΔIR are changes from the Gaussian reference (mean $\pm$ 95% t_9 interval half-width across ten prompts). $n_{1\%}$ counts images below the prompt-specific Gaussian first PickScore percentile, out of 300. F.1/Z/F.2 denote FLUX.1-schnell/Z-Image-Turbo/FLUX.2-klein.

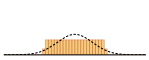
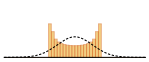
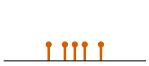
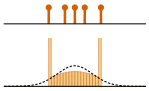
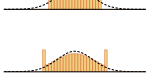
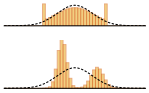
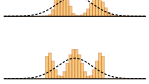
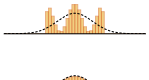
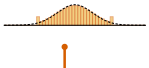
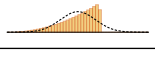
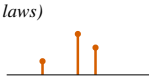
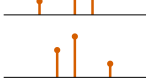
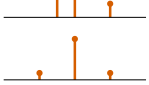
Source	Entry-wise distribution	Model	(M)	(E)	(T)	μ_3	κ_4	ΔPick↑	ΔIR↑	$n_{1\%}$ ↓
All (M), (E), (T) satisfied (15 laws)										
Gaussian quantile control		F.1	✓	✓	✓	0.00	0.00	-0.03 ± 0.08	-0.002 ± 0.018	6
		Z	✓	✓	✓	0.00	0.00	$+0.03 \pm 0.03$	$+0.020 \pm 0.016$	1
		F.2	✓	✓	✓	0.00	0.00	$+0.04 \pm 0.02$	$+0.004 \pm 0.023$	3
Two-point, $p = 1/3$		F.1	✓	✓	✓	0.71	−1.50	0.00 ± 0.06	$+0.008 \pm 0.022$	5
		Z	✓	✓	✓	0.71	−1.50	$+0.03 \pm 0.05$	$+0.017 \pm 0.024$	0
		F.2	✓	✓	✓	0.71	−1.50	0.00 ± 0.04	-0.002 ± 0.017	6
Rademacher		F.1	✓	✓	✓	0.00	−2.00	$+0.01 \pm 0.07$	$+0.016 \pm 0.021$	1
		Z	✓	✓	✓	0.00	−2.00	$+0.05 \pm 0.04$	$+0.015 \pm 0.010$	1
		F.2	✓	✓	✓	0.00	−2.00	$+0.01 \pm 0.05$	-0.005 ± 0.015	2
Uniform		F.1	✓	✓	✓	0.00	−1.20	-0.01 ± 0.09	-0.003 ± 0.022	5
		Z	✓	✓	✓	0.00	−1.20	$+0.03 \pm 0.04$	$+0.023 \pm 0.016$	2
		F.2	✓	✓	✓	0.00	−1.20	$+0.05 \pm 0.05$	-0.003 ± 0.017	3
Arcsine		F.1	✓	✓	✓	0.00	−1.50	-0.01 ± 0.08	-0.002 ± 0.020	4
		Z	✓	✓	✓	0.00	−1.50	$+0.02 \pm 0.04$	$+0.018 \pm 0.014$	1
		F.2	✓	✓	✓	0.00	−1.50	$+0.03 \pm 0.06$	-0.008 ± 0.013	3
Five-level quantization		F.1	✓	✓	✓	0.00	−1.05	-0.01 ± 0.08	-0.009 ± 0.023	4
		Z	✓	✓	✓	0.00	−1.05	$+0.04 \pm 0.04$	$+0.024 \pm 0.018$	1
		F.2	✓	✓	✓	0.00	−1.05	$+0.03 \pm 0.04$	$+0.003 \pm 0.015$	3
Gaussian clipped at 1.0		F.1	✓	✓	✓	0.00	−1.39	0.00 ± 0.08	-0.003 ± 0.019	6
		Z	✓	✓	✓	0.00	−1.39	$+0.04 \pm 0.03$	$+0.024 \pm 0.013$	0
		F.2	✓	✓	✓	0.00	−1.39	$+0.03 \pm 0.06$	-0.006 ± 0.020	4
Gaussian clipped at 1.5		F.1	✓	✓	✓	0.00	−0.96	-0.05 ± 0.09	-0.014 ± 0.025	3
		Z	✓	✓	✓	0.00	−0.96	$+0.03 \pm 0.04$	$+0.028 \pm 0.016$	1
		F.2	✓	✓	✓	0.00	−0.96	$+0.03 \pm 0.03$	-0.004 ± 0.020	5
Skewed bumps		F.1	✓	✓	✓	0.58	−1.31	$+0.02 \pm 0.07$	$+0.008 \pm 0.021$	0
		Z	✓	✓	✓	0.58	−1.31	$+0.04 \pm 0.04$	$+0.016 \pm 0.018$	0
		F.2	✓	✓	✓	0.58	−1.31	$+0.03 \pm 0.06$	$+0.001 \pm 0.014$	2
Wave		F.1	✓	✓	✓	0.00	−1.07	0.00 ± 0.07	$+0.003 \pm 0.012$	2
		Z	✓	✓	✓	0.00	−1.07	$+0.03 \pm 0.05$	$+0.021 \pm 0.020$	1
		F.2	✓	✓	✓	0.00	−1.07	$+0.03 \pm 0.05$	$+0.001 \pm 0.020$	4
Gaussian clipped at 2.0		F.1	✓	✓	✓	0.00	−0.55	0.00 ± 0.10	0.000 ± 0.019	5
		Z	✓	✓	✓	0.00	−0.55	$+0.04 \pm 0.04$	$+0.020 \pm 0.011$	3
		F.2	✓	✓	✓	0.00	−0.55	$+0.04 \pm 0.02$	$+0.005 \pm 0.023$	3
Two-point, $p = 1/4$		F.1	✓	✓	✓	1.15	−0.67	$+0.02 \pm 0.04$	$+0.009 \pm 0.021$	2
		Z	✓	✓	✓	1.15	−0.67	$+0.01 \pm 0.05$	$+0.021 \pm 0.027$	2
		F.2	✓	✓	✓	1.15	−0.67	$+0.03 \pm 0.07$	$+0.011 \pm 0.013$	2
Triangular		F.1	✓	✓	✓	0.00	−0.60	-0.01 ± 0.09	-0.001 ± 0.017	5
		Z	✓	✓	✓	0.00	−0.60	$+0.04 \pm 0.03$	$+0.021 \pm 0.014$	0
		F.2	✓	✓	✓	0.00	−0.60	$+0.04 \pm 0.02$	$+0.002 \pm 0.024$	4
Four-point		F.1	✓	✓	✓	0.32	−0.33	$+0.02 \pm 0.08$	-0.001 ± 0.023	2
		Z	✓	✓	✓	0.32	−0.33	$+0.03 \pm 0.05$	$+0.018 \pm 0.010$	1
		F.2	✓	✓	✓	0.32	−0.33	$+0.03 \pm 0.04$	$+0.001 \pm 0.019$	1
Wedge		F.1	✓	✓	✓	−0.84	0.02	$+0.02 \pm 0.05$	0.000 ± 0.011	2
		Z	✓	✓	✓	−0.84	0.02	$+0.03 \pm 0.03$	$+0.013 \pm 0.033$	0
		F.2	✓	✓	✓	−0.84	0.02	$+0.03 \pm 0.07$	$+0.011 \pm 0.020$	3
At least one condition fails (20 laws)										
Three-point, $\mu_3 = -1$		F.1	✓	×	✓	−1.00	0.00	0.00 ± 0.07	$+0.001 \pm 0.019$	4
		Z	✓	×	✓	−1.00	0.00	$+0.01 \pm 0.04$	$+0.006 \pm 0.019$	3
		F.2	✓	×	✓	−1.00	0.00	$+0.01 \pm 0.06$	$+0.002 \pm 0.020$	2
Three-point, $\mu_3 = +1$		F.1	✓	×	✓	1.00	0.00	$+0.01 \pm 0.07$	-0.005 ± 0.024	7
		Z	✓	×	✓	1.00	0.00	$+0.04 \pm 0.05$	$+0.012 \pm 0.010$	2
		F.2	✓	×	✓	1.00	0.00	-0.04 ± 0.06	-0.004 ± 0.018	6
Three-point, $a = 2$		F.1	✓	×	✓	0.00	1.00	0.00 ± 0.08	$+0.003 \pm 0.021$	3
		Z	✓	×	✓	0.00	1.00	$+0.03 \pm 0.03$	$+0.015 \pm 0.027$	0
		F.2	✓	×	✓	0.00	1.00	$+0.04 \pm 0.06$	$+0.001 \pm 0.015$	0
Three-point, $a = 3$		F.1	✓	×	✓	0.00	6.00	$+0.02 \pm 0.08$	-0.030 ± 0.037	5
		Z	✓	×	✓	0.00	6.00	-0.18 ± 0.33	-0.026 ± 0.055	31
		F.2	✓	×	✓	0.00	6.00	$+0.02 \pm 0.06$	$+0.002 \pm 0.017$	2
Three-point, $a = 4$		F.1	✓	×	✓	0.00	13.00	-0.42 ± 0.31	-0.116 ± 0.171	37
		Z	✓	×	✓	0.00	13.00	-0.94 ± 0.74	-0.141 ± 0.232	141
		F.2	✓	×	✓	0.00	13.00	-0.02 ± 0.06	$+0.005 \pm 0.019$	4
Two-point, $p = 0.05$		F.1	✓	×	✓	4.13	15.05	-0.95 ± 0.59	-0.255 ± 0.452	100
		Z	✓	×	✓	4.13	15.05	-1.30 ± 0.91	-0.290 ± 0.456	180
		F.2	✓	×	✓	4.13	15.05	-0.16 ± 0.18	-0.018 ± 0.041	21

Table 6 continued

Source	Entry-wise distribution	Model	(M)	(E)	(T)	μ_3	κ_4	ΔPick↑	ΔIR↑	$n_{1\%}$ ↓
Three-point, $a = 6$		F.1	✓	×	✓	0.00	33.01	-3.50 ± 0.83	-1.248 ± 0.879	295
		Z	✓	×	✓	0.00	33.01	-2.99 ± 0.97	-0.884 ± 0.679	290
		F.2	✓	×	✓	0.00	33.00	-0.15 ± 0.10	-0.039 ± 0.035	14
Two-point, $p = 0.02$		F.1	✓	×	✓	6.86	45.01	-4.03 ± 0.60	-1.564 ± 0.812	300
		Z	✓	×	✓	6.86	45.01	-4.12 ± 0.95	-1.398 ± 0.788	300
		F.2	✓	×	✓	6.86	45.02	-1.40 ± 0.54	-0.103 ± 0.181	184
Gaussian body + ± 10 spikes		F.1	✓	×	×	0.00	0.29	-0.05 ± 0.09	-0.006 ± 0.014	7
		Z	✓	×	×	0.00	0.29	0.00 ± 0.04	$+0.016 \pm 0.017$	0
		F.2	✓	×	×	0.00	0.04	$+0.04 \pm 0.02$	-0.001 ± 0.023	5
Gaussian body + ± 20 spikes		F.1	✓	×	×	0.00	4.81	-0.21 ± 0.13	-0.056 ± 0.100	15
		Z	✓	×	×	0.00	4.81	-0.31 ± 0.19	-0.045 ± 0.094	54
		F.2	✓	×	×	0.00	0.60	$+0.03 \pm 0.02$	$+0.001 \pm 0.022$	4
$\Gamma(4)$		F.1	✓	×	×	1.00	1.49	$+0.03 \pm 0.07$	-0.007 ± 0.031	6
		Z	✓	×	×	1.00	1.49	$+0.02 \pm 0.03$	$+0.008 \pm 0.016$	1
		F.2	✓	×	×	1.00	1.50	$+0.03 \pm 0.03$	-0.003 ± 0.019	3
Reflected $\Gamma(4)$		F.1	✓	×	×	-1.00	1.49	$+0.01 \pm 0.03$	0.000 ± 0.012	1
		Z	✓	×	×	-1.00	1.49	0.00 ± 0.04	$+0.014 \pm 0.030$	3
		F.2	✓	×	×	-1.00	1.50	$+0.02 \pm 0.08$	$+0.018 \pm 0.024$	2
Student t_7		F.1	✓	×	×	0.00	1.88	0.00 ± 0.05	-0.004 ± 0.020	4
		Z	✓	×	×	0.00	1.88	-0.01 ± 0.05	$+0.012 \pm 0.018$	4
		F.2	✓	×	×	0.00	1.95	$+0.03 \pm 0.02$	$+0.007 \pm 0.021$	1
Laplace		F.1	✓	×	×	0.00	2.98	-0.04 ± 0.06	-0.025 ± 0.027	8
		Z	✓	×	×	0.00	2.98	-0.02 ± 0.09	$+0.013 \pm 0.019$	6
		F.2	✓	×	×	0.00	3.00	$+0.02 \pm 0.03$	$+0.013 \pm 0.022$	4
Student t_5		F.1	✓	×	×	0.00	4.47	-0.05 ± 0.08	-0.027 ± 0.041	3
		Z	✓	×	×	0.00	4.47	-0.28 ± 0.26	-0.029 ± 0.053	44
		F.2	✓	×	×	0.00	4.98	$+0.02 \pm 0.03$	$+0.008 \pm 0.021$	1
Centered exponential		F.1	✓	×	×	2.00	5.96	-0.07 ± 0.09	-0.017 ± 0.033	7
		Z	✓	×	×	2.00	5.96	-0.35 ± 0.39	-0.039 ± 0.070	53
		F.2	✓	×	×	2.00	5.99	0.00 ± 0.07	-0.003 ± 0.021	3
Reflected exponential		F.1	✓	×	×	-2.00	5.96	-0.19 ± 0.11	-0.054 ± 0.054	13
		Z	✓	×	×	-2.00	5.96	-0.49 ± 0.50	-0.049 ± 0.128	64
		F.2	✓	×	×	-2.00	5.99	-0.01 ± 0.09	$+0.026 \pm 0.021$	3
Student t_4		F.1	✓	×	×	0.00	10.35	-0.43 ± 0.31	-0.085 ± 0.138	44
		Z	✓	×	×	0.00	10.35	-0.98 ± 0.54	-0.120 ± 0.159	159
		F.2	✓	×	×	0.00	13.40	0.00 ± 0.06	$+0.003 \pm 0.019$	6
Student t_3		F.1	✓	×	×	0.00	52.69	-1.42 ± 0.58	-0.243 ± 0.243	173
		Z	✓	×	×	0.00	52.69	-2.16 ± 0.78	-0.322 ± 0.255	249
		F.2	✓	×	×	0.00	108.55	-0.20 ± 0.15	-0.022 ± 0.041	29
Lognormal		F.1	✓	×	×	5.80	76.35	-2.54 ± 0.86	-0.709 ± 0.693	257
		Z	✓	×	×	5.80	76.35	-3.05 ± 0.90	-0.720 ± 0.515	297
		F.2	✓	×	×	6.03	91.35	-0.92 ± 0.41	-0.080 ± 0.116	151

B.5.3 GAUSSIAN CONTROLS FOR THE TWO-VALUED SOURCE

Construction of Figure 1. Figure 1 shows two of three generated prompts using the Gaussian ranks of one fixed seed to construct the source. We use the entry-wise distribution in Equation 2 for all three models; the overlaid realized mass and CDF curves agree to three decimals. The constructed tensor enters the first affine layer without centering, rescaling, or clipping. No rejection or image selection is applied to these qualitative source examples. The paired Gaussian controls below use the same ranks and sampling settings.

In the quantitative comparison, the two-point source places 0.7% of images below the Gaussian first PickScore percentile on each generator (Table 6). The following paired Gaussian images complement those prompt-level measurements.



Figure 13: Gaussian controls paired with Figure 1. Left: pooled input-element density and empirical CDF (dashed: standard Gaussian); right: two examples from each model at matched prompts and seed.

B.5.4 Output diversity

Generation scores alone do not test whether a source preserves output diversity. We compare the two-valued source with a paired Gaussian control using ten prompts and one hundred images per prompt per source on each of the three generators: $6{,}000$ images in total, with the same one hundred seeds used for both sources.

All statistics are computed within each prompt. The mean pairwise distance is the average over its $\binom{100}{2}$ image pairs. DreamSim (Fu et al., 2023) uses its pretrained preprocessing; LPIPS (Zhang et al., 2018) uses AlexNet on RGB tensors rescaled to $[-1, 1]$. For each distance matrix D, we also compute the Vendi score from $K = 1 - D$ (Friedman & Dieng, 2023). DreamSim gives a Gram matrix by construction. All 60 LPIPS $1 - D$ matrices are positive semidefinite, with a minimum eigenvalue of 0.084.

Table 7 shows small changes in both pairwise distance and effective diversity. Mean pairwise distances change by at most 1.2% across the two metrics and three generators. Paired t_9 statistics use the ten prompts; no comparison survives a 5% Bonferroni correction over the twelve reported comparisons. These measurements support retained output diversity in the tested setting.

Table 7: Output diversity of the two-valued source compared with its paired Gaussian control, ten prompts × one hundred images per source on each generator. Pairwise distances and Vendi scores are computed within each prompt; changes and t_9 statistics are paired across the ten prompts. Diversity scores are comparable at fixed sample size and metric backbone.

		mean pairwise distance				Vendi score			
Backbone	Generator	Gauss.	two-pt	change	t_9	Gauss.	two-pt	change	t_9
DreamSim	FLUX.1-schnell	0.1720	0.1727	+0.4%	+0.40	3.09	3.10	+0.5%	+0.46
	Z-Image-Turbo	0.0978	0.0989	+1.2%	+0.45	1.92	1.93	+0.3%	+0.26
	FLUX.2-klein	0.1247	0.1243	−0.3%	−0.29	2.36	2.36	−0.1%	−0.12
LPIPS	FLUX.1-schnell	0.6350	0.6304	−0.7%	−2.12	32.78	32.17	−1.9%	−2.26
	Z-Image-Turbo	0.4840	0.4839	−0.0%	−0.05	16.29	16.11	−1.1%	−1.22
	FLUX.2-klein	0.5986	0.5934	−0.9%	−2.10	29.47	29.00	−1.6%	−1.98

B.6 Sensitivity and limits of the source conditions

B.6.1 Mean and scale controls

We change one moment at a time while keeping the Gaussian shape fixed: $\mathrm{X} = \mathrm{G} + \mu$ or $\mathrm{X} = s\mathrm{G}$. Figure 14 shows both reward differences and lower-tail fractions for the full sweeps, using the protocol in Appendix B.5. A small change in expected norm can have a large effect: on FLUX.1-schnell, $\mu = 0.10$ adds only 1% to expected token energy but yields $269/300$ low-tail images, whereas $s = 1.15$ adds 32% and yields $32/300$.

We further compare the two signs of the mean shift at the same expected squared norm. Both $\mu = -0.10$ and $+0.10$ have $\mathbb{E}[\mathrm{X}^2] = 1.01$, yet Z-Image-Turbo has $7/300$ versus $237/300$ low-tail images. The signed mean axis preserves this asymmetry. At the first embedding, a mean shift adds the same vector $\mu \boldsymbol{W}^\top \mathbf{1}$ to every token, whereas scaling stretches each token's existing centered embedding. Relative expected token-energy changes are μ^2 and $s^2 - 1$, respectively; they do not specify how the generator responds to the intervention. Appendix B.2 gives the exact embedding identities.

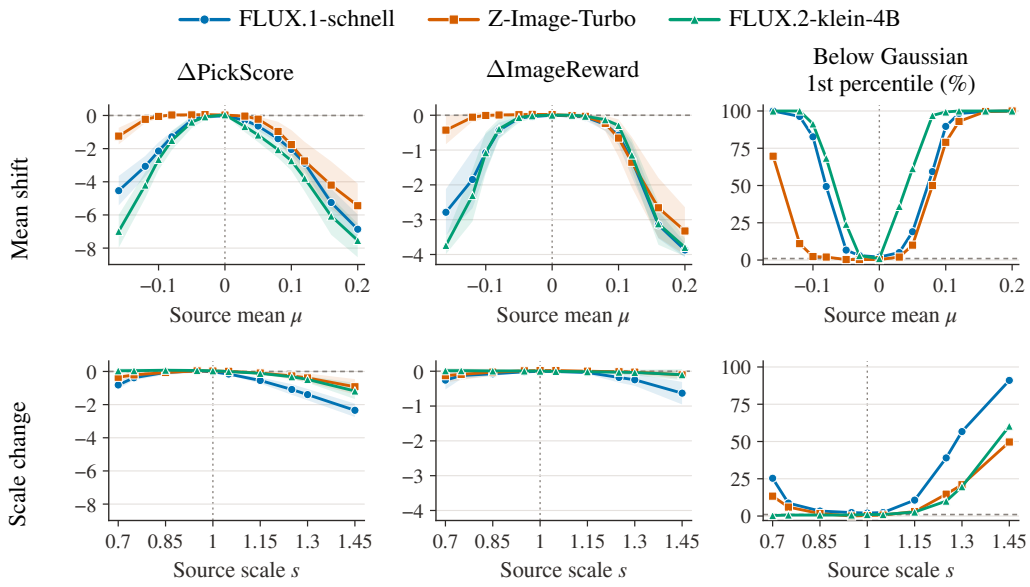


Figure 14: Sensitivity to Gaussian mean (top) and scale (bottom) across three pretrained generators. Scores are changes from the prompt-specific Gaussian reference; shading shows 95% t_9 intervals across ten prompts. The right column reports the fraction below the prompt-specific Gaussian first PickScore percentile; the dashed line marks the nominal 1% rate. Points at $\mu = 0$ and $s = 1$ are the measured Gaussian-quantile control. Lines connect tested conditions.

B.6.2 Scalar bounds and token-energy excursions

A bounded entry-wise distribution can still produce large token-energy excursions. Consider the three-point law placing mass $1/(2a^2)$ at each of $\pm a$ and the rest at zero. It satisfies (M), and for $a = 6$ it also satisfies (T) at $B = 8$. Its token energy is $\mathrm{S}_d = a^2\mathrm{K}_d$, where $\mathrm{K}_d \sim \mathrm{Binomial}(d, 1/a^2)$. At $d = 64$ and $N = 1024$, token energy falls below L_d with zero spikes and exceeds the exact U_d (which rounds to 108.0) with three or more. Thus $\Pr[\mathrm{S}_d \notin [L_d, U_d]] \approx 0.43$, over four hundred times the (E) budget, despite $|\mathrm{X}| \le 6$.

The generation results also distinguish the magnitude of individual entries from concentration within a token. On FLUX.1-schnell, Gaussian-body sources with isolated ± 20 entries yield $15/300$ low-tail images, whereas the bounded sparse $a = 6$ law yields $295/300$ (Table 6). This contrast illustrates why a bound on individual entries alone is insufficient; condition (E) addresses their combined magnitude within a token.

Why both auxiliary conditions are used. Condition (E) bounds the probability of a token-energy excursion. Condition (T) additionally limits the energy carried by rare scalar values. Combining these two conditions gives the following bound.

Proposition 4 (Token and scalar tails jointly control hidden energy)**.** *Let F satisfy (M), (E), and (T), and let $\mathcal{E} = \{\mathrm{S}_d \notin [L_d, U_d]\}$. Write $\varepsilon_B = \mathbb{E}[(\mathrm{Z}^2 - B^2)_+]$ for the Gaussian excess energy beyond the scalar threshold. Then*

$$\mathbb{E}\big[\mathrm{S}_d \mathbf{1}_{\mathcal{E}}\big] \le dB^2 \Pr[\mathcal{E}] + d\varepsilon_B \le dB^2/N + d\varepsilon_B. \tag{17}$$

The first term decreases with N; the second is a Gaussian-tail remainder at the chosen B.

Proof. We decompose $\mathrm{S}_d = A_d + R_d$, where $A_d = \sum_i \min(\mathrm{X}_i^2, B^2) \le dB^2$ and $R_d = \sum_i (\mathrm{X}_i^2 - B^2)_+$. By (T) and tail integration, $\mathbb{E}[R_d] \le d\mathbb{E}[(\mathrm{Z}^2 - B^2)_+] = d\varepsilon_B$. Consequently, $\mathbb{E}[\mathrm{S}_d \mathbf{1}_{\mathcal{E}}] \le dB^2 \Pr[\mathcal{E}] + \mathbb{E}[R_d]$, and (E) gives the final bound. □

Neither auxiliary condition implies the other. The bounded sparse law above satisfies (M) and (T) but violates (E).

Conversely, set $q = 10^{-8}$, $c = \sqrt{0.9/(1-q)}$, and $\sigma = \sqrt{0.1/q}$. Give each of $\pm c$ probability $(1-q)/2$, and draw σZ with probability q. This symmetric law satisfies (M), with the rare component contributing 0.1 of the variance. If a token contains no rare draw, its energy is $0.9d/(1-q)$, inside the reference interval at both experimental geometries. Hence the excursion probability is at most $dq < 1/N$, so (E) holds. However, $\Pr[|X| > 8] \approx q$ greatly exceeds the standard Gaussian tail; (T) fails. At $N = 4096, d = 128$, a whole latent misses the rare component with probability $(1-q)^{Nd} \approx 0.995$ and then has mean square about 0.9. This illustrates why an excursion-count budget alone does not control the variance hidden in rare entries.

B.6.3 Token-energy interventions outside the class

We randomly select Gaussian token rows and set them to zero, or rescale them to squared norm 8 or 160. Figure 15 reports the resulting lower-tail fractions, using the same 300-image protocol as the scalar-law comparison.

The response depends on the intervention and generator. At 30% modified tokens, zeroing is most damaging on FLUX.1-schnell, while energy 160 has the largest effect on Z-Image-Turbo. For FLUX.2, energy 160 is inside its Gaussian reference interval, whereas it exceeds the upper endpoint for the other models. The same absolute token energy therefore represents a different intervention relative to each model's reference.

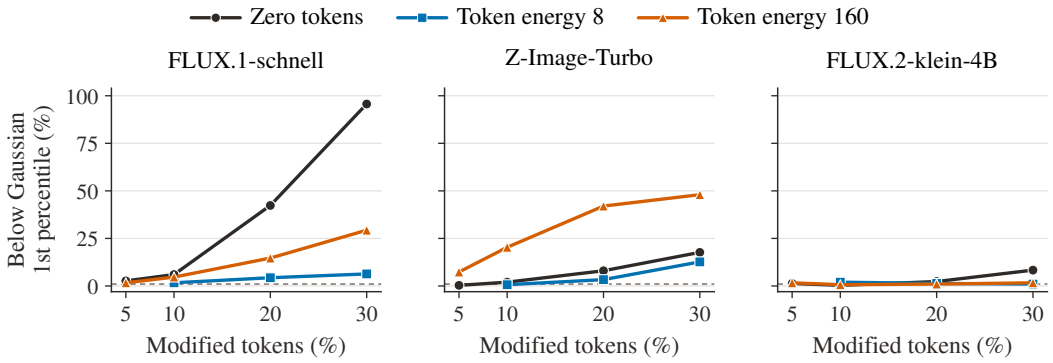


Figure 15: Outside-class interventions on selected Gaussian token rows. The vertical axis is the fraction of 300 images below the prompt-specific Gaussian first PickScore percentile; the dashed line marks the nominal 1% rate. Lines connect measured fractions of modified tokens. Zero and energy-8 tokens lie below the Gaussian token-energy interval. Energy 160 lies outside this interval for FLUX.1 and Z-Image, but inside it for FLUX.2.

What the source class establishes. The reflected $\Gamma(4)$ control retains near-reference generation despite violating the conditions, while the rare-atom two-point source with $p = 0.02$ degrades generation. Together with the mean and scale interventions, these results show tolerance to broad changes in the entry-wise distribution without reducing source suitability to either total norm or class membership. Arrangement controls are analyzed in Appendix C.

Figure 16 shows saved images side by side using the same seed and two fixed prompts. It complements the scalar-law comparison and the mean/scale sweeps without selecting examples by reward score.

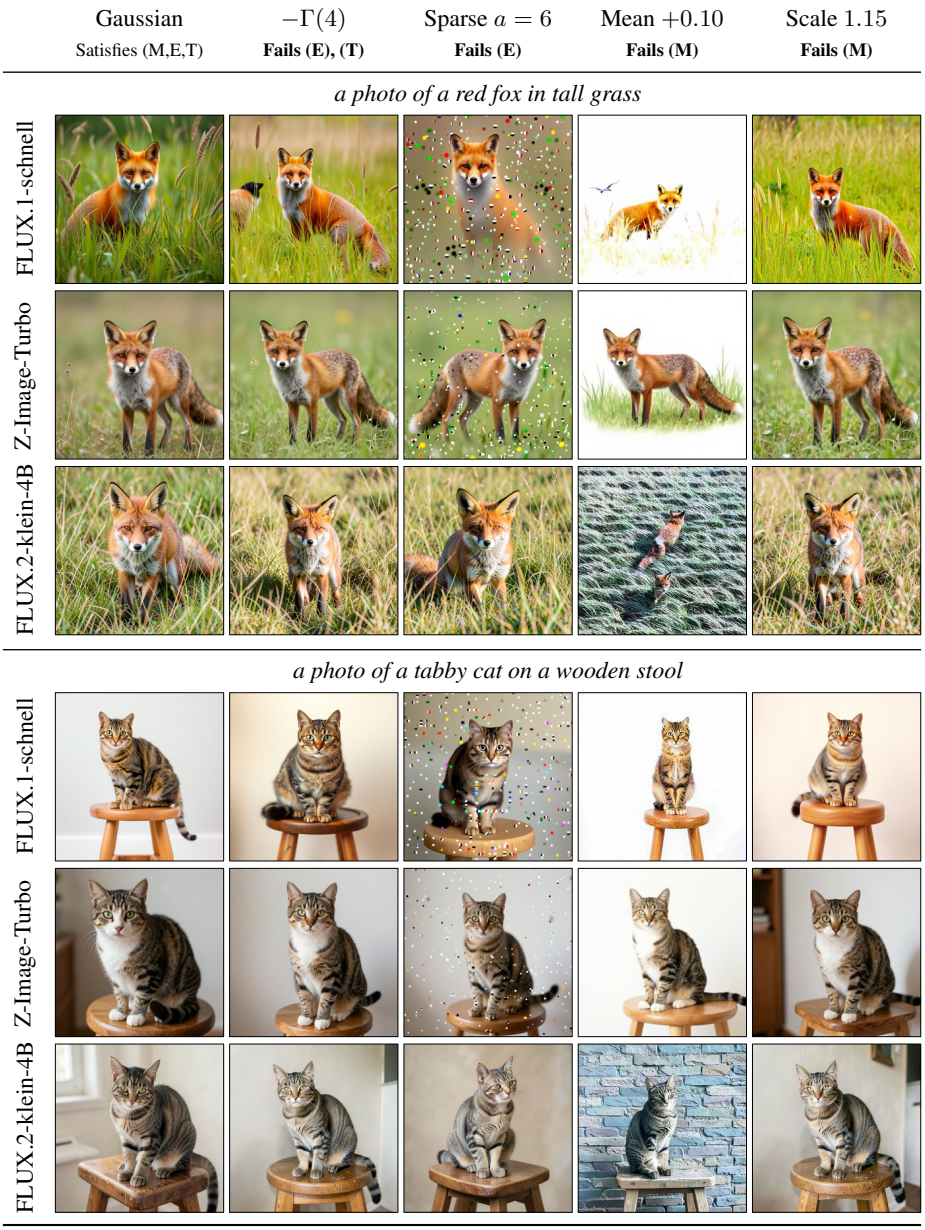


Figure 16: Paired source interventions at the same prompt and seed. The prompts are the first two in the fixed ten-prompt protocol, selected without inspecting reward scores or images. Only the Gaussian reference satisfies all three conditions. The reflected Gamma law fails (E) and (T), the sparse law fails (E), and the mean-shift and scale controls fail (M). The reflected Gamma law nevertheless gives near-reference generation quality. Quantitative comparisons are in Table 6 and Figure 14.

C ARRANGEMENT AND GENERATION

Appendix B shows that substantial changes in the entry-wise distribution can preserve generation under random placement. We now ask whether the same values can become unsuitable through their arrangement. Our paired controls keep complete token vectors fixed and change only their assignment to positions, isolating placement from token content. We first test designed spatial patterns, then use repeated Gaussian tokens and partial coherent mixtures to test how broadly the effect occurs.

C.1 CONTROLLED ARRANGEMENTS

For each structured source, we form a paired control by uniformly permuting complete rows while keeping the generator's positional assignments fixed. Both inputs therefore contain the same scalar values and token vectors, with identical token norms and total norm. The comparison changes which token occupies each position. These sources have deliberately introduced dependencies; they are not additional i.i.d. sources of the kind studied in Appendix B.

We use three families of spatial patterns: smooth waves (low-DCT), rapid oscillations (high-DCT), and block patterns (Hadamard). Each feature column is arranged over the $h \times h$ token grid, where $N = h^2$; $(h, d) = (32, 64)$ for FLUX.1-schnell and Z-Image-Turbo, and $(64, 128)$ for FLUX.2-klein-4B. Low-DCT and high-DCT use the d nonconstant two-dimensional DCT-II modes (u, v) with the smallest or largest $u^2 + v^2$, respectively. Hadamard uses the first d nonconstant columns of the Sylvester Hadamard matrix of order N. We normalize the basis columns, mix them with a seeded $d \times d$ Gaussian coefficient matrix, center the result, whiten its feature Gram matrix (eigenvalue floor 10^{-12}), and rescale it to unit RMS. These steps standardize feature covariance and overall scale before we form the paired control.

C.2 GENERATION RESULTS ACROSS MODELS

Table 8 follows the three-model protocol in Appendix B.5: ten prompts and thirty images per condition–prompt pair, with a separate $1{,}000$-image Gaussian reference for each prompt. We report PickScore and ImageReward differences with prompt-level intervals, together with $n_{1\%}$, the number of the 300 images below their prompt's Gaussian first-percentile PickScore.

Coherent low-DCT and Hadamard sources substantially degrade generation on all three models, placing all 300 images below the reference percentile. High-DCT also degrades generation, although FLUX.2-klein-4B is less affected than the other two models. Permuting the same tokens largely restores the scores across these constructions. Recovery varies with the pattern and model: for example, permuted low-DCT still places $37/300$ Z-Image-Turbo images below the percentile. The large changes between each coherent source and its permuted control arise despite identical token content and norms.

Qualitative examples. Figure 5 illustrates the low-DCT comparison on FLUX.1-schnell for prompts 4, 5, 9, and 10 in Table 10. These examples were selected for recognizable subjects and visible detail after permutation, and a clear contrast between paired generations. The portrait and butterfly share one seed selected from 32 candidates; the robot and astronaut each use a seed selected from eight paired candidates. Table 8 reports the full ten-prompt comparison.

C.3 DUPLICATION AND DOSE CONTROLS

Beyond designed basis patterns. To test whether the arrangement effect requires a designed basis pattern, we repeat Gaussian tokens instead. We sample $N/2$ token vectors and use each twice, placing the copies either at consecutive positions in the token order (adjacent) or at random positions (dispersed). Both sources contain exactly the same token vectors. Adjacent placement lowers the scores on all three models, whereas dispersed copies remain close to the Gaussian reference (Table 8). Thus the large degradation is associated with where the repeated tokens are placed, beyond the effect of duplication alone.

Partial coherent contributions. The preceding controls impose structure throughout the source. We next ask whether the effect persists when only a partial coherent component is mixed into Gaussian

noise. Given a Gaussian latent $\mathbf{G}$ and the feature-whitened low-DCT source $\boldsymbol{C}$, we use

$$\boldsymbol{X}_\lambda = \sqrt{1-\lambda}\,\mathbf{G} + \sqrt{\lambda}\,\boldsymbol{C}, \qquad \lambda \in \{0.10, 0.25, 0.50, 0.75\}. \tag{18}$$

Here λ specifies the nominal coherent contribution. Degradation appears already at the smallest tested dose, with 27 to 160 of the 300 images below the reference percentile, and becomes more pronounced as the contribution increases. At the three larger doses, we also permute the rows of the mixed source; the scores then return close to the Gaussian reference. The comparison shows that arrangement matters even for a partial coherent contribution.

Table 8: Arrangement, duplication, and dose controls (15 per model). ΔPick/ΔIR: change from the Gaussian reference (mean ± 95% t_9 half-width over ten prompts); $n_{1\%}$: images out of 300 below the prompt's Gaussian first-percentile PickScore (about 3 by chance; 300 means all).

Intervention	Model	ΔPick↑	ΔIR↑	$n_{1\%}\downarrow$
Low-DCT: coherent	F.1	-9.06 ± 0.60	-3.978 ± 0.133	300
	Z	-9.12 ± 0.68	-3.874 ± 0.226	300
	F.2	-8.76 ± 0.95	-3.853 ± 0.258	300
Low-DCT: permuted	F.1	-0.01 ± 0.07	$+0.004 \pm 0.017$	9
	Z	-0.23 ± 0.21	-0.032 ± 0.050	37
	F.2	-0.03 ± 0.06	$+0.003 \pm 0.030$	1
High-DCT: coherent	F.1	-7.12 ± 0.86	-3.815 ± 0.225	300
	Z	-8.12 ± 0.84	-3.849 ± 0.225	300
	F.2	-2.95 ± 0.99	-1.378 ± 0.847	264
High-DCT: permuted	F.1	$+0.12 \pm 0.11$	$+0.004 \pm 0.022$	1
	Z	$+0.01 \pm 0.07$	$+0.005 \pm 0.030$	3
	F.2	0.00 ± 0.07	-0.004 ± 0.016	5
Hadamard: coherent	F.1	-8.01 ± 0.58	-3.969 ± 0.130	300
	Z	-7.63 ± 0.70	-3.876 ± 0.228	300
	F.2	-8.91 ± 1.02	-3.855 ± 0.258	300
Hadamard: permuted	F.1	$+0.04 \pm 0.05$	$+0.019 \pm 0.018$	2
	Z	$+0.02 \pm 0.04$	$+0.001 \pm 0.015$	5
	F.2	0.00 ± 0.06	$+0.004 \pm 0.010$	2
Duplicates: adjacent	F.1	-2.79 ± 0.90	-1.229 ± 0.835	259
	Z	-2.54 ± 1.30	-1.000 ± 0.829	266
	F.2	-0.67 ± 0.26	-0.091 ± 0.082	107
Duplicates: dispersed	F.1	-0.12 ± 0.07	-0.012 ± 0.014	4
	Z	$+0.03 \pm 0.05$	$+0.006 \pm 0.020$	2
	F.2	-0.02 ± 0.05	-0.007 ± 0.028	3
Low-DCT dose 0.10	F.1	-0.87 ± 0.36	-0.198 ± 0.150	85
	Z	-0.14 ± 0.12	-0.025 ± 0.062	27
	F.2	-1.08 ± 0.50	-0.233 ± 0.156	160
Low-DCT dose 0.25	F.1	-3.89 ± 0.63	-2.042 ± 0.724	296
	Z	-2.73 ± 1.06	-1.118 ± 0.770	278
	F.2	-4.42 ± 1.07	-1.777 ± 0.722	288
Low-DCT dose 0.25: permuted	F.1	$+0.02 \pm 0.07$	$+0.010 \pm 0.013$	5
	Z	-0.02 ± 0.04	$+0.015 \pm 0.020$	6
	F.2	-0.01 ± 0.08	$+0.009 \pm 0.018$	2
Low-DCT dose 0.50	F.1	-7.14 ± 0.71	-3.839 ± 0.184	300
	Z	-6.69 ± 0.98	-3.423 ± 0.358	300
	F.2	-7.46 ± 0.94	-3.600 ± 0.300	300
Low-DCT dose 0.50: permuted	F.1	$+0.05 \pm 0.05$	$+0.017 \pm 0.014$	1
	Z	-0.02 ± 0.03	$+0.002 \pm 0.020$	3
	F.2	$+0.01 \pm 0.07$	$+0.003 \pm 0.013$	4
Low-DCT dose 0.75	F.1	-8.56 ± 0.57	-3.976 ± 0.132	300
	Z	-8.51 ± 0.69	-3.839 ± 0.211	300
	F.2	-8.44 ± 0.96	-3.847 ± 0.256	300
Low-DCT dose 0.75: permuted	F.1	$+0.01 \pm 0.09$	-0.007 ± 0.017	3
	Z	-0.04 ± 0.06	-0.004 ± 0.023	6
	F.2	-0.02 ± 0.07	$+0.011 \pm 0.023$	2

C.4 What arrangement adds to the source analysis

Tolerance to broad changes in the entry-wise distribution does not imply tolerance to arbitrary arrangements. Across these controls, the same token vectors produce substantially different generation quality when assigned to different positions. The shape of the entry-wise distribution and total norm alone therefore cannot explain source suitability. This motivates studying arrangement-sensitive statistics that allow broad variation in the entry-wise distribution, and how they can guide updates in noise optimization.

D Spectral control across entry-wise distributions

The argument in Section 4 has two steps: independent sources can share a spectral reference despite different entry-wise distributions, and flattening an update minimizes a worst-case bound on spectral change. We establish these steps and distinguish their roles. The common reference describes the source; the flat spectrum is a choice for the update added to it.

D.1 A spectral reference across entry-wise distributions

For an $N \times d$ matrix with $N \geq d$, the squared singular values divided by N are the eigenvalues of $\boldsymbol{X}^\top \boldsymbol{X}/N$. Their empirical distribution records how energy is spread across singular directions. The following classical result explains why this distribution can provide a reference beyond Gaussian inputs.

Theorem 1 (Common limiting spectrum of independent entries)**.** *Let $(\xi_{ij})_{i,j\geq 1}$ be an infinite array of i.i.d. real variables with a fixed law, $\mathbb{E}[\xi_{ij}] = 0$, and $\mathbb{E}[\xi_{ij}^2] = 1$. Let $\mathbf{X}_{N,d} = (\xi_{ij})_{1\leq i\leq N,\, 1\leq j\leq d}$, where $N, d \to \infty$, $d \leq N$, and $d/N \to \gamma \in (0, 1]$. Then*

$$\mu_{N,d} := \frac{1}{d}\sum_{i=1}^{d} \delta_{\sigma_i^2(\mathbf{X}_{N,d})/N} \implies \mathrm{MP}_\gamma \quad \textit{almost surely}, \tag{19}$$

where δ_a is a unit point mass and MP_γ is the Marchenko–Pastur law (Marčenko & Pastur, 1967), supported on $[(1-\sqrt{\gamma})^2, (1+\sqrt{\gamma})^2]$. If $\mathbb{E}[\xi_{ij}^4] < \infty$, the largest and smallest singular values of $\mathbf{X}_{N,d}/\sqrt{N}$ also converge almost surely to $1+\sqrt{\gamma}$ and $1-\sqrt{\gamma}$, respectively (Bai & Yin, 1993).

The limiting law depends on the aspect ratio, not on the particular standardized entry-wise distribution. For a square matrix, the normalized singular values have limiting density $\pi^{-1}\sqrt{4-s^2}$ on $0 \leq s \leq 2$. Thus Gaussian, Rademacher, uniform, and asymmetric standardized laws can share a spectral reference. Finite variance suffices for the empirical law; the finite fourth moment supplies the stated control of its extreme singular values.

This is an asymptotic statement about independent entries. It does not identify finite-size rank-wise confidence intervals across different entry-wise distributions, nor apply directly to coherent arrangements or latents produced by reward optimization. Figure 6 therefore provides a separate finite-size illustration; its construction and numerical checks are given in Appendix D.4.

D.2 What singular values measure, and what they leave out

The common reference is useful because singular values measure more than the scalar histogram. In the original token–feature layout, they constrain total energy and the largest aggregate response along a feature direction.

Proposition 5 (Norm, mean, and directional bounds)**.** *For $\boldsymbol{X} \in \mathbb{R}^{N\times d}$, let $\bar{x} = (Nd)^{-1}\mathbf{1}_N^\top \boldsymbol{X}\mathbf{1}_d$ and let $\boldsymbol{x}_t$ be its rows. Then*

$$\begin{aligned} \|\boldsymbol{X}\|_F^2 &= \sum_i \sigma_i(\boldsymbol{X})^2, & |\bar{x}| &\leq \frac{\sigma_1(\boldsymbol{X})}{\sqrt{Nd}}, \\ \max_{\|\boldsymbol{a}\|_2=1} \|\boldsymbol{X}\boldsymbol{a}\|_2 &= \sigma_1(\boldsymbol{X}), & \max_t \|\boldsymbol{x}_t\|_2 &\leq \sigma_1(\boldsymbol{X}). \end{aligned} \tag{20}$$

Proof. The first identity follows from the SVD. For the mean, $|\mathbf{1}_N^\top \boldsymbol{X} \mathbf{1}_d| \le \sqrt{N}\,\|\boldsymbol{X}\|_{\mathrm{op}}\sqrt{d}$. The variational definition of the operator norm gives the third identity, and $\|\boldsymbol{x}_t\|_2 = \|\boldsymbol{X}^\top \boldsymbol{e}_t\|_2 \le \|\boldsymbol{X}\|_{\mathrm{op}}$ gives the final bound. □

For a unit feature direction $\boldsymbol{a}$, $\|\boldsymbol{X}\boldsymbol{a}\|_2^2$ is the energy of that direction across tokens. A large leading singular value therefore permits strong directional concentration. A mean shift adds the rank-one component $\mu \mathbf{1}_N \mathbf{1}_d^\top$, whereas multiplying the source by a scalar multiplies every singular value by its absolute value. A statistic that tolerates changes in the entry-wise distribution can thus remain sensitive to scale and certain forms of concentration.

The spectrum still leaves arrangement information unmeasured. Complete row permutations preserve it but can change generation (Appendix E.1). Nor do these bounds imply the source conditions (M), (E), and (T): for example, the row-norm bound is not a token-energy tail guarantee. The spectrum can guide update design without completely characterizing a suitable source.

D.3 From a spectral reference to a flat update

For $\boldsymbol{X}^+ = \boldsymbol{X} + \eta\boldsymbol{\Delta}$, fix the representation used to form the update. Let $T : \mathbb{R}^{N\times d} \to \mathbb{R}^{p\times q}$ be a linear bijection preserving the Frobenius inner product, with $pq = Nd$; a reshape is one example. The represented update is $\eta T(\boldsymbol{\Delta})$.

Theorem 2 (Minimum spectral-displacement bound at fixed energy)**.** *For any $\boldsymbol{X}, \boldsymbol{\Delta} \in \mathbb{R}^{N\times d}$ and $\eta \ge 0$,*

$$\max_i |\sigma_i(T(\boldsymbol{X} + \eta\boldsymbol{\Delta})) - \sigma_i(T(\boldsymbol{X}))| \le \eta \|T(\boldsymbol{\Delta})\|_{\mathrm{op}}. \tag{21}$$

Fix $\rho > 0$ and a rank budget $1 \le r \le \min(p, q)$. Then

$$\min_{\|\boldsymbol{\Delta}\|_F = \rho,\, \mathrm{rank}(T(\boldsymbol{\Delta})) \le r} \|T(\boldsymbol{\Delta})\|_{\mathrm{op}} = \frac{\rho}{\sqrt{r}}. \tag{22}$$

The minimum is attained exactly when $T(\boldsymbol{\Delta})$ has rank r and r equal nonzero singular values. In particular, if the represented momentum has compact rank-r SVD $T(\boldsymbol{M}) = \boldsymbol{U}_r \boldsymbol{\Sigma}_r \boldsymbol{V}_r^\top$, then

$$\boldsymbol{\Delta}_{\mathrm{M}} = T^{-1}\left(\frac{\rho}{\sqrt{r}} \boldsymbol{U}_r \boldsymbol{V}_r^\top\right), \qquad \|\boldsymbol{\Delta}_{\mathrm{M}}\|_F = \rho, \qquad \|T(\boldsymbol{\Delta}_{\mathrm{M}})\|_{\mathrm{op}} = \frac{\rho}{\sqrt{r}}. \tag{23}$$

Proof. The singular-value perturbation inequality (Higham, 2021), applied to $T(\boldsymbol{X})$ and $T(\boldsymbol{X}) + \eta T(\boldsymbol{\Delta})$, gives Equation 21. For every feasible update, the isometry of T gives

$$\rho^2 = \sum_i \sigma_i(T(\boldsymbol{\Delta}))^2 \le r\, \|T(\boldsymbol{\Delta})\|_{\mathrm{op}}^2. \tag{24}$$

Equality requires exactly r nonzero singular values, all equal to $\rho/\sqrt{r}$. The columns of $\boldsymbol{U}_r$ and $\boldsymbol{V}_r$ are orthonormal, so $\boldsymbol{U}_r \boldsymbol{V}_r^\top$ has r unit singular values and attains equality after scaling. This proves both the minimum and its equality condition. □

Taking T to be the identity gives Proposition 2. The criterion selects flat amplitudes but leaves the singular vectors undetermined. Retaining those of the momentum yields Muon's scaled polar factor. This flat profile controls the update's worst-case effect; it is not the non-flat random-matrix reference for the source.

Representation and the implemented rank convention. At a fixed entry count and Frobenius update norm, a near-square reshape maximizes the available rank $s = \min(p, q)$, reducing the full-rank bound $\eta\rho/\sqrt{s}$. For example, refolding 1024×64 to 256×256 raises s from 64 to 256 and halves this bound. The comparison concerns spectra in their respective representations; it does not imply a tighter bound on the original token–feature spectrum.

Equation 23 uses the r nonzero directions of a compact SVD. Our implementation uses all s directions of the reduced SVD, including a null-space completion for rank-deficient momentum (Appendix F.2). In exact arithmetic, $\boldsymbol{U}_s \boldsymbol{V}_s^\top$ has s unit singular values; scaling by $\rho/\sqrt{s}$ attains Equation 22 with budget s. It coincides with the compact factor for full-rank momentum, but uses additional directions otherwise.

What the bound establishes. The perturbation result is deterministic, so it remains valid when the update depends on the current latent. Minimizing its right-hand side does not minimize every realized spectral displacement. For example, from $\boldsymbol{X} = \mathrm{diag}(2, 1)$, the unit-Frobenius update $\boldsymbol{\Delta}_1 = \left(\begin{smallmatrix} 0 & 1 \\ 0 & 0 \end{smallmatrix}\right)$ changes its singular values by $O(\eta^2)$, whereas the flat update $\boldsymbol{\Delta}_2 = \boldsymbol{I}_2/\sqrt{2}$ changes both by $\eta/\sqrt{2}$. The latter has the smaller operator norm, but the former has the smaller realized displacement for sufficiently small η.

Nor does a one-step bound certify that an optimization trajectory stays inside Gaussian reference intervals or preserves image quality. It supplies a criterion for the update profile; Appendix E tests its consequences for arrangement and generation. The related spectral-norm steepest-descent view (Bernstein & Newhouse, 2024) concerns optimization geometry. Here the same polar operation is motivated by the perturbation entering a pretrained generator's source.

D.4 Finite-size construction and verification

Figure 6 separates two interventions at 64×64: changing the entry-wise distribution under random placement, and changing the arrangement of one fixed Gaussian-shaped multiset. Every displayed matrix is recentered and scaled to empirical variance one; spectra are reported as $\sigma_i/\sqrt{64}$.

Random and coherent matrices. For each of the eight laws in Figure 3, we select $4{,}096$ equally spaced empirical quantiles from its experimental value pool, standardize this multiset, and apply a fixed-seed random permutation. For the coherent matrices, we instead use a single shared multiset of Gaussian midpoint quantiles $\Phi^{-1}((i - 1/2)/4096)$, $i = 1, \ldots, 4096$, standardized in the same way. We assign the sorted values according to the ranks of $[s - \bar{s}]/\,\mathrm{sd}(s) + 0.9\epsilon$, where s is a spatial score field and ϵ is fixed-seed i.i.d. Gaussian noise. The eight fields are a diagonal ramp, squared radius, checker, wave, square rings, radial waves, angle, and saddle. All eight coherent matrices therefore have identical scalar values and squared Frobenius norm 4096; only their arrangement changes. The display applies Φ in grayscale, while spectral calculations use the original entry values.

Reference intervals and measured separation. We estimate a pointwise central 99.99% interval for each ordered singular value from $500{,}000$ i.i.d. standard-normal matrices. Unlike the displayed matrices, these reference draws are not individually recentered or variance-normalized. Each estimated tail contains approximately 25 observations, and the intervals are pointwise rather than a simultaneous 99.99% region over all ranks.

All 64 ranks of each of the eight randomly arranged matrices lie inside these reference intervals. In every coherent matrix, the first ten ranks lie outside; 53–57 of the 64 ranks do so overall. Their normalized largest singular values range from 4.337 to 5.992. These are descriptive comparisons with the Gaussian reference, not coverage guarantees for the fixed-multiset construction, whose entries are dependent.

As a separate independent-entries check, we draw $3{,}000$ matrices per law by sampling with replacement from its experimental value pool. Their median-spectrum RMSE from the Gaussian median ranges from 0.00037 to 0.00748. This supports the observed proximity of the median profiles, without asserting equality of finite-size interval widths. Together, the two comparisons illustrate the distinction used in the main text: a spectral reference can tolerate substantial changes in the entry-wise distribution while responding to some arrangements of fixed values.

D.5 A complementary Fourier reference

The Fourier controls in Appendix E address frequency allocation. Their reference also extends beyond Gaussian entry-wise distributions under independence.

Proposition 6 (Fourier reference under independent entries)**.** *Let the entries of $\mathbf{X} \in \mathbb{R}^{N \times d}$ be i.i.d. from a fixed law with mean zero, variance one, and finite third absolute moment. Fix d, a real unit vector $\boldsymbol{a} \in \mathbb{R}^d$, and a finite set of k distinct positive frequency indices $f_1, \ldots, f_k$. For sufficiently large N these lie below $N/2$. With $z_t = \mathbf{x}_t^\top \boldsymbol{a}$, define the orthonormal DFT coefficients*

$$\widehat{z}_f = \frac{1}{\sqrt{N}} \sum_{t=0}^{N-1} z_t \exp(-2\pi \mathrm{i} f t/N). \tag{25}$$

As $N \to \infty$, $(\widehat{z}_{f_1}, \ldots, \widehat{z}_{f_k})$ converges jointly to independent complex Gaussian variables whose real and imaginary parts are independent $\mathcal{N}(0, 1/2)$. Consequently,

$$\frac{1}{k}\sum_{j=1}^{k} |\widehat{z}_{f_j}|^2 \implies \frac{\chi^2_{2k}}{2k}. \tag{26}$$

Equivalently, the limit is a Gamma variable with shape k and rate k. DC and, for even N, Nyquist are real coefficients and are excluded.

Proof. The z_t are independent with mean zero, variance one, and finite third absolute moment. Any fixed nonzero real linear combination of the selected real and imaginary DFT parts is a sum $\sum_t w_{N,t} z_t$, with $\max_t |w_{N,t}| = O(N^{-1/2})$. DFT orthogonality gives variance $1/2$ for each part and zero cross-covariances. Thus the combination has a positive constant variance and $\sum_t \mathbb{E}[|w_{N,t} z_t|^3] = O(N^{-1/2})$. The Lyapunov CLT and the Cramér–Wold argument yield the joint Gaussian limit. Each squared complex modulus is then unit-mean exponential; summing k independent terms gives the stated energy law. □

For Gaussian entries, the independence and energy law hold exactly at finite N for the selected non-DC, non-Nyquist coefficients. For general entry-wise distributions, the proposition concerns a fixed finite collection; it does not establish Gaussian rank-wise coverage for an increasing set of all bands. The FFT layout, DC/Nyquist handling, and finite Gaussian references used by the implemented controls are specified in Appendix E.4.

E WHY PROCESS UPDATES BEFORE ACCUMULATION?

Appendix D motivates flattening an update through a bound on spectral displacement. GPER (Hwang & Sung, 2026) motivates spectral gradient projection through constraints designed to characterize white Gaussian noise. Here we examine what matching such statistics explains about generation. We distinguish two roles of spectral statistics: identifying a suitable latent and processing an incoming update. Selected profiles can satisfy Gaussian references, or even be flat, while allowing harmful arrangements. Yet enforcing a profile on each update can weaken structure before it accumulates. This distinction provides an explanation for why spectral processing can help even when the resulting latent does not retain the reference values.

E.1 FLAT PROFILES DO NOT DEFINE A SUITABLE SOURCE

Singular values and Fourier statistics respond to different properties of an arrangement while allowing broad variation in the entry-wise distribution. Fourier coefficients also have a common Gaussian limit under independent, standardized entries and the moment conditions in Proposition 6. The mean and variance set the DC component and power scale, while tails and higher moments affect the finite-size approximation. These references describe selected statistics of random inputs; they do not specify every aspect of their arrangement.

The token-permutation experiment gives a simple example. Let $\boldsymbol{X}$ have its original token–feature layout, $\boldsymbol{P}$ permute complete rows, and $H(\boldsymbol{X}) = \boldsymbol{X}\boldsymbol{W} + \mathbf{1}\boldsymbol{b}^\top$ be the tokenwise affine embedding. Since $\boldsymbol{P}^\top\boldsymbol{P} = \boldsymbol{I}$ and $\boldsymbol{P}\mathbf{1} = \mathbf{1}$,

$$H(\boldsymbol{P}\boldsymbol{X}) = \boldsymbol{P}H(\boldsymbol{X}), \qquad (\boldsymbol{P}\boldsymbol{X})^\top(\boldsymbol{P}\boldsymbol{X}) = \boldsymbol{X}^\top\boldsymbol{X}. \tag{27}$$

Thus the scalar and token multisets, feature Gram matrix, singular spectrum, and each embedding channel's empirical distribution are unchanged. Yet generation quality differs substantially between the coherent and permuted inputs in Table 8. This spectral invariance concerns the original token–feature matrix (1024×64 for FLUX.1 in our setting); it is not guaranteed after the near-square reshape used by Muon.

The coherent-dose controls in Appendix C illustrate the difference between this spectrum and a token-Fourier summary. On FLUX.1-schnell, the lowest of eight token-Fourier bands carries energy shares 0.167, 0.229, 0.333, and 0.437 at coherent doses $\lambda = 0.10$, 0.25, 0.50, and 0.75, respectively. At the three larger doses, permutation brings the share back to approximately 0.125 while leaving the token–feature singular spectrum unchanged. The two summaries therefore respond to different aspects of arrangement.

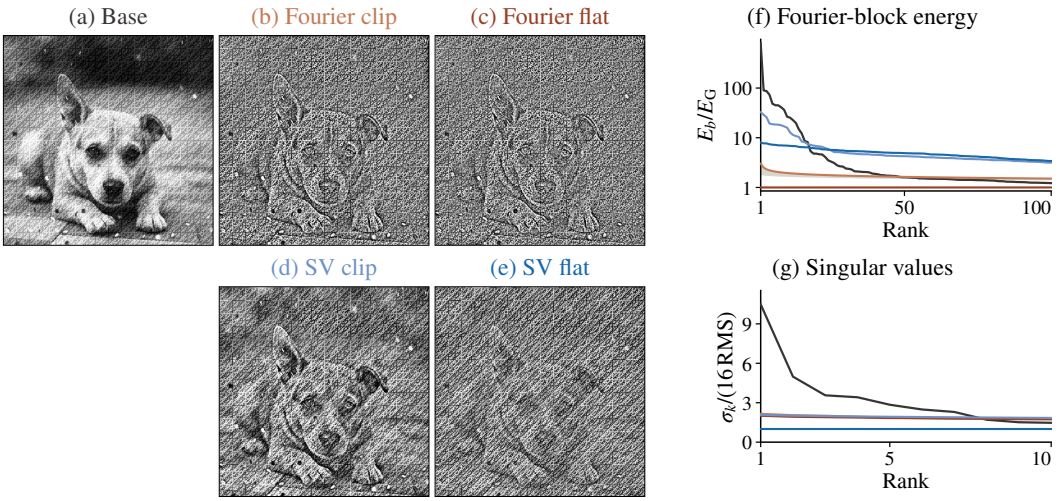


Figure 17: Clipping and flattening leave visible structure. All panels transform the same structured field with a Gaussian entry-wise distribution at equal RMS and use a common normal-CDF view. Curve colors match the panel titles; shading marks pointwise central 99.99% Gaussian reference intervals. All ranks are processed, with the leading 100 Fourier ranks and ten singular values shown.

Figure 17 examines a structured field with a Gaussian entry-wise distribution. Here the Fourier controls rescale band amplitudes while preserving within-band directions and phases; the optimization controls use GPER's within-band projection (Appendix E.4). Fourier and SV clipping bring their respective ordered statistics into Gaussian reference intervals, while flattening equalizes the corresponding amplitudes. Both operations weaken visible patterns, but neither removes them completely. Fourier and SV controls also change each other's profiles, although they leave different residual patterns. Even the stronger flattening operation therefore leaves information that the selected profile does not describe.

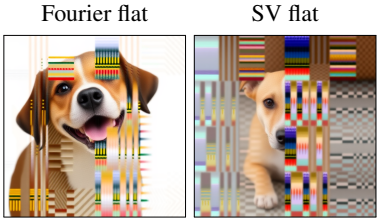


Figure 18: Flat profiles do not guarantee a suitable source. The flattened fields in Figure 17 produce artifacts when used as sources for four-step FLUX.1-schnell generation with prompt "dog".

Figure 18 uses the two flattened fields as sources to FLUX.1-schnell. Both generations have conspicuous artifacts. Thus flattening a statistical profile is insufficient to map a tensor to a source on which the generator works normally. The distinction matters for latent constraints: correcting the measured values can leave an unsuitable arrangement intact.

E.2 Processing incoming structure

During reward-guided optimization, gradients and momentum carry patterns induced by the generator and reward model. Repeatedly adding these vectors can build structure in the latent while its selected statistics remain within reference intervals. Processing the incoming vector changes what is added at each step, including when the current latent already satisfies those constraints.

Figure 19 shows this intervention on an actual momentum snapshot. We apply each operator to the same momentum, standardize the result, and use it as the entire source to generate an image with the prompt "dog". The original momentum contains strong regular patterns and produces an unrecognizable image. Near-square SV flattening and Fourier-block SV flattening weaken these patterns and recover clearer dog shapes. This diagnostic reveals how processing changes the structure carried by a vector. In optimization, only a scaled update is added to the current latent; the update itself need not be a complete working source.

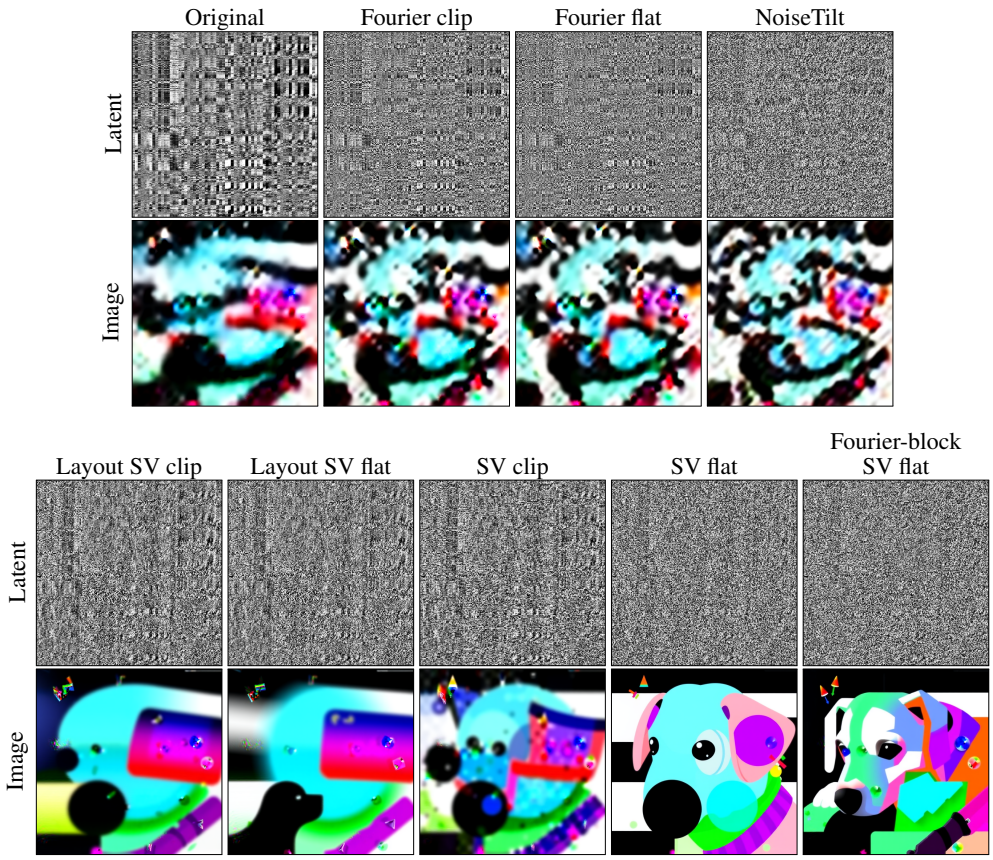


Figure 19: Suppressing structure in a momentum snapshot. Each tensor is obtained by processing the same momentum from Aesthetic Score optimization, then standardized and used as the entire source with prompt "dog". Near-square SV flattening and Fourier-block SV flattening weaken visible patterns and recover clearer dog shapes. SV flat is the transform used by Muon.

Figure 20 shows the consequence during optimization from the same Gaussian initialization. At nearly identical Aesthetic Scores, latent SV clipping keeps all 256 ordered singular values inside their reference intervals to numerical tolerance, yet visible patterns remain in the latent and the image becomes an abstract artifact. Muon instead processes momentum before each update and retains a realistic dog, even though its latent singular values exceed those same intervals. The state with the reference-compliant spectrum is therefore the one with worse generation in this comparison.

Together with Figure 19, this contrast supports an explanation based on suppressing structure before accumulation. Latent projection permits arrangements that satisfy the measured values, whereas flattening momentum weakens the patterns entering through each update. Muon's benefit here cannot be explained simply by keeping the latent spectrum within Gaussian reference intervals. Its use of the spectrum is effective as an operation on incoming updates, even though the same statistic is incomplete as a criterion for source suitability.

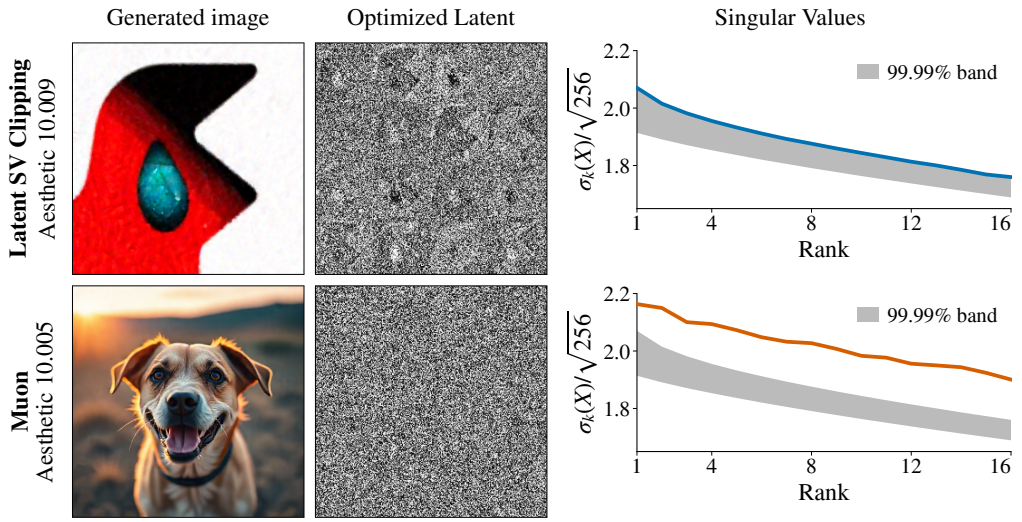


Figure 20: Latent spectral constraints do not guarantee realism. At matched Aesthetic Scores, latent SV clipping stays within the reference intervals to numerical tolerance but retains visible latent structure and produces an artifact. Muon processes momentum before each update and retains a realistic dog despite exceeding those intervals. The prompt and Gaussian start are shared. Latent panels tile 16 spatial channels; curves show the leading 16 singular values of the 256×256 reshape, divided by $\sqrt{256}$ without RMS normalization. Gray bands are central 99.99% Gaussian intervals.

E.3 Where should we intervene?

We compare the two intervention locations across 45 prompts and three paired seeds, using shared Gaussian starts and 100 updates on FLUX.1-schnell. Latent controls project the source after an SGDM step; momentum controls process the direction before it is added to the source. Figure 21 reports their trade-offs between the target Aesthetic Score and held-out PickScore. Operator definitions follow in Appendix E.4; optimizer settings and the proposal-RMS convention are given in Appendix F.2.

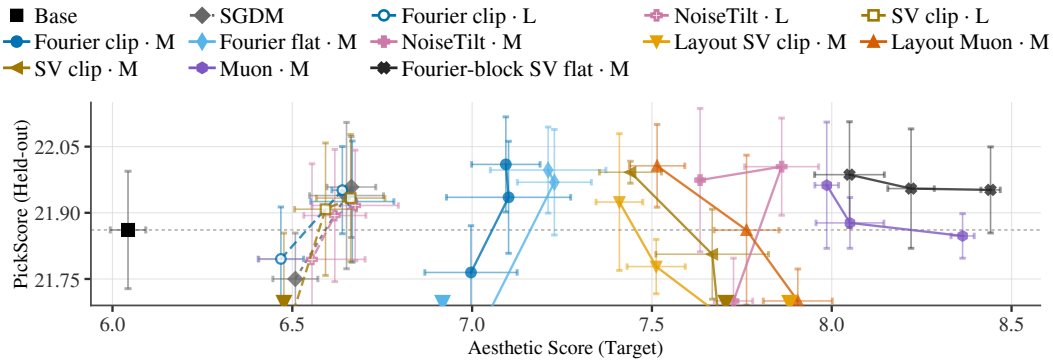


Figure 21: Momentum processing reaches higher Aesthetic Scores at comparable held-out PickScore than the tested latent projections. FLUX.1-schnell results after 100 updates show three-seed means with one SD; lines connect RMS settings. L denotes latent projection and M denotes momentum processing. The dashed line marks Base; downward triangles indicate means below the plotted range. Full scores are in Table 9.

The tested latent projections remain close to SGDM, whereas each tested form of momentum processing reaches a higher target reward at comparable PickScore. This separation also holds for reference clipping: the same type of statistical constraint becomes useful when applied to incoming momentum. Together with Figure 20, these results support suppressing structure before accumulation rather than relying on constraints on the resulting latent.

GPER (Hwang & Sung, 2026) already demonstrates the benefit of gradient projection and relates it to noise-aligned directions. Its construction motivates the spectral constraints as a characterization of white Gaussian noise. Our analysis separates the usefulness of this type of processing from the ability of the selected statistics to identify a suitable source. Figures 18 and 20 show why spectral compliance alone is insufficient, while Figure 21 supports using those statistics to weaken incoming structure. This provides a common interpretation of Fourier preconditioning and SV flattening: incomplete source diagnostics can still define useful update operators. The remaining question is how these operators should modify momentum.

E.4 How should momentum be modified?

We now read Figure 21 within the momentum controls, comparing how strictly they constrain a profile and which representation they use. The NoiseTilt controls apply its whitening operator (Hwang et al., 2026), with $\alpha = 10^{-4}$, to the latent or momentum; they do not use its full reverse-kernel sampler.

Clipping and flattening. Clipping uses pointwise central 99.99% Gaussian intervals for each ordered statistic. For a descending normalized profile t_i, we set

$$\widehat{t}_i = \mathrm{clip}(\lambda t_i; \ell_i, u_i), \qquad \sum_i \widehat{t}_i^{\,p} = C, \tag{28}$$

where $[\ell_i, u_i]$ is the rank-wise interval and λ matches the prescribed norm. Singular values are normalized by $\sqrt{m}$ times the reference RMS for an $m \times n$ matrix, with $p = 2$ and $C = n$; reconstruction retains the singular vectors. Fourier controls group GPER's compact real FFT of the flattened tensor into $2{,}048$ bands of 16 complex coefficients. Their mean-normalized energies use $p = 1$, $C = 2{,}048$, and rank-wise bounds from $\chi^2_{32}/32$ order statistics. The clipped energies are restored to the original band order. Momentum clipping preserves incoming RMS; latent clipping targets unit RMS after each SGDM step. These are pointwise reference intervals, not a simultaneous confidence region.

Flattening instead equalizes the selected profile. Fourier clip and Fourier flat share GPER's within-band projection, using clipped and equal band energies, respectively. Figure 21 shows a better trade-off from flattening in both SV representations, with a smaller improvement for Fourier processing. Removing the unequal amplitudes still allowed by the Gaussian intervals is therefore useful in these comparisons.

Matrix representation. Layout Muon applies SV flattening directly to the 1024×64 token–feature matrix, whereas Muon uses its 256×256 reshape, following Muse (Chang et al., 2026a) and Section 4.2. Their clipping counterparts are Layout SV clip and SV clip, using fixed-RMS and i.i.d. Gaussian matrix references, respectively. Near-square Muon reaches a higher useful reward range than Layout Muon and the Fourier-only controls. Thus the choice of representation matters even when the same flattening operation is used.

Combining Fourier and SV processing. The Fourier-block variant applies an orthogonal real DFT representation T_{sr} along the token axis. It lists real nonnegative-frequency coordinates followed by imaginary positive, non-Nyquist coordinates, with $\sqrt{2}$ scaling outside DC and Nyquist. We split $T_{\mathrm{sr}}\boldsymbol{M}_t$ into sixteen consecutive 64×64 blocks $\boldsymbol{B}_b$ and set

$$\boldsymbol{D}_t = T_{\mathrm{sr}}^{\top}\,\mathrm{stack}_{b=1}^{16}\big(8\,\mathrm{polar}(\boldsymbol{B}_b)\big), \qquad \mathrm{RMS}(\boldsymbol{D}_t) = 1. \tag{29}$$

These full-feature blocks differ from GPER's bands. For full-column-rank $\boldsymbol{M}$, orthogonal $\boldsymbol{T}$ satisfies $\boldsymbol{T}^{\top}\,\mathrm{polar}(\boldsymbol{T}\boldsymbol{M}) = \mathrm{polar}(\boldsymbol{M})$, so a Fourier transform followed by one global polar map cancels; the block partition is essential. Related transform–block–polar methods are discussed in Appendix A.2.

Fourier-block processing further extends the frontier in Figure 21. At matched RMS η, its updates and Layout Muon both satisfy $\boldsymbol{\Delta}^{\top}\boldsymbol{\Delta} = N\eta^2\boldsymbol{I}_d$ for full-rank polar factors, and hence have the same token–feature singular values. Their different outcomes show that this spectrum alone does not explain the trade-off. Changing representations or combining operators changes which patterns are suppressed, rather than simply increasing control of one profile.

An open question. The weaker patterns and clearer dog content in Figure 19 broadly agree with the optimization trends, providing a visual clue to the effect of processing. This agreement suggests studying what structure an operator removes, beyond whether its output satisfies a chosen statistical profile.

The results favor flattening over clipping and show further benefits from changing representations and combining Fourier and SV processing. They do not yet determine the best form or strength of modification. Weak processing can allow harmful patterns to accumulate, while stronger processing may also remove useful reward-gradient information. We have not established where this balance lies or how it varies across models, tasks, and optimization states. Finding a principle for this choice remains future work. Near-square Muon provides a simple practical choice among established optimizers; the Fourier-block comparison motivates further study of how an update should be processed.

E.5 Full Quantitative Results in Figure 21

Table 9 reports all 36 operating points in Figure 21, including Aesthetic Score and the three held-out rewards with seed-wise variation.

Table 9: FLUX.1-schnell Aesthetic-guided control operating points at update 100 (45 prompts, three seeds). Entries are mean ± one standard deviation across seed-wise prompt means. Group headings identify where the operator acts; every row uses a Gaussian start and step 0 is shared. RMS is the nominal proposal-update RMS, before any post-update latent projection.

Method	RMS	Aesthetic (target)	PickScore	ImageReward	HPSv2.1
Base (step 0)	–	6.043 ± 0.049	21.861 ± 0.133	1.152 ± 0.036	0.3098 ± 0.0048
A. Reference	*SGDM update only*				
SGDM	0.00025	6.664 ± 0.068	21.958 ± 0.114	1.163 ± 0.068	0.3123 ± 0.0045
	0.0005	6.651 ± 0.103	21.939 ± 0.166	1.149 ± 0.066	0.3098 ± 0.0081
	0.001	6.509 ± 0.063	21.750 ± 0.104	1.047 ± 0.093	0.3031 ± 0.0049
B. Latent projection	*SGDM step ⟶ project latent*				
Fourier clip	0.00025	6.667 ± 0.116	21.926 ± 0.137	1.157 ± 0.080	0.3118 ± 0.0052
	0.0005	6.639 ± 0.029	21.951 ± 0.099	1.180 ± 0.054	0.3112 ± 0.0041
	0.001	6.468 ± 0.063	21.795 ± 0.118	1.072 ± 0.111	0.3041 ± 0.0058
NoiseTilt	0.00025	6.675 ± 0.120	21.917 ± 0.125	1.186 ± 0.104	0.3112 ± 0.0057
	0.0005	6.618 ± 0.086	21.894 ± 0.150	1.150 ± 0.147	0.3087 ± 0.0083
	0.001	6.555 ± 0.148	21.795 ± 0.217	1.048 ± 0.156	0.3020 ± 0.0088
SV clip	0.00025	6.662 ± 0.095	21.933 ± 0.144	1.203 ± 0.026	0.3113 ± 0.0054
	0.0005	6.592 ± 0.086	21.908 ± 0.150	1.086 ± 0.140	0.3079 ± 0.0081
	0.001	6.477 ± 0.113	21.606 ± 0.248	0.966 ± 0.138	0.2981 ± 0.0066
C. Momentum modification	*modify momentum ⟶ fixed-RMS latent step*				
Fourier clip	0.001	7.094 ± 0.095	22.010 ± 0.108	1.237 ± 0.055	0.3138 ± 0.0039
	0.002	7.101 ± 0.173	21.935 ± 0.127	1.211 ± 0.096	0.3107 ± 0.0047
	0.003	6.997 ± 0.128	21.765 ± 0.106	1.064 ± 0.088	0.3050 ± 0.0040
Fourier flat	0.001	7.211 ± 0.161	21.997 ± 0.098	1.249 ± 0.047	0.3142 ± 0.0045
	0.0025	7.229 ± 0.103	21.969 ± 0.120	1.178 ± 0.055	0.3124 ± 0.0038
	0.005	6.918 ± 0.196	21.465 ± 0.201	0.838 ± 0.173	0.2915 ± 0.0094
NoiseTilt	0.0025	7.634 ± 0.005	21.974 ± 0.163	1.215 ± 0.098	0.3132 ± 0.0068
	0.005	7.861 ± 0.103	22.005 ± 0.110	1.223 ± 0.037	0.3145 ± 0.0040
	0.0075	7.727 ± 0.053	21.700 ± 0.097	1.004 ± 0.165	0.3037 ± 0.0059
Layout SV clip	0.0025	7.410 ± 0.065	21.924 ± 0.155	1.187 ± 0.149	0.3117 ± 0.0058
	0.005	7.512 ± 0.081	21.778 ± 0.062	1.168 ± 0.008	0.3067 ± 0.0007
	0.0075	7.883 ± 0.158	21.549 ± 0.074	0.992 ± 0.066	0.2973 ± 0.0038
Layout Muon	0.0025	7.514 ± 0.077	22.006 ± 0.094	1.205 ± 0.075	0.3133 ± 0.0036
	0.005	7.763 ± 0.090	21.861 ± 0.170	1.155 ± 0.161	0.3096 ± 0.0083
	0.0075	7.906 ± 0.096	21.700 ± 0.073	1.040 ± 0.088	0.3028 ± 0.0018
SV clip	0.0025	7.440 ± 0.086	21.992 ± 0.025	1.215 ± 0.050	0.3136 ± 0.0012
	0.005	7.668 ± 0.157	21.806 ± 0.102	1.169 ± 0.092	0.3071 ± 0.0042
	0.0075	7.705 ± 0.073	21.500 ± 0.183	1.012 ± 0.121	0.2992 ± 0.0053
Muon	0.005	7.986 ± 0.033	21.963 ± 0.143	1.207 ± 0.066	0.3137 ± 0.0056
	0.0075	8.051 ± 0.094	21.877 ± 0.058	1.181 ± 0.092	0.3110 ± 0.0035
	0.01	8.364 ± 0.032	21.848 ± 0.051	1.176 ± 0.048	0.3111 ± 0.0018
Fourier-block SV flat	0.005	8.049 ± 0.097	21.986 ± 0.120	1.236 ± 0.056	0.3145 ± 0.0037
	0.0075	8.221 ± 0.065	21.955 ± 0.135	1.215 ± 0.116	0.3139 ± 0.0066
	0.01	8.442 ± 0.027	21.952 ± 0.098	1.231 ± 0.022	0.3143 ± 0.0035

F Complete experimental results and implementation

This section gives the shared experimental settings and optimizer implementation, followed by the complete results.

F.1 Models, rewards, prompts, and evaluation

Generators and sampling. We use pretrained FLUX.1-schnell (Black Forest Labs, 2024), Z-Image-Turbo (Z-Image Team et al., 2025), and FLUX.2-klein-4B (Black Forest Labs, 2025). Source and arrangement studies use resolutions 512^2, 512^2, and 1024^2, with four, eight, and four sampling steps, respectively. Their token–feature layouts are 1024×64, 1024×64, and 4096×128, in the transformer's token order. Reward optimization and evaluation use one Euler step for FLUX.1-schnell, eight steps for Z-Image-Turbo, and four for FLUX.2-klein-4B. Models use `bfloat16`; optimized latents use `float32`.

Rewards and aggregation. Aesthetic Score (Schuhmann, 2022) and PickScore (Kirstain et al., 2023) are the target rewards. For each target, the other three scorers among these two, ImageReward (Xu et al., 2023), and HPSv2.1 (Wu et al., 2023) are held out of the gradient. Aesthetic optimization uses 45 animal prompts; PickScore optimization uses 100 fixed prompts from T2I-CompBench++ (Huang et al., 2025). We use three random seeds. For each prompt and seed, all methods start from the same initial latent. We first average scores over prompts separately for each seed, then report the mean and sample standard deviation of these three averages. Results are evaluated at update 100 for FLUX.1, 50 for Z-Image, and 30 for both FLUX.2 targets. Base denotes the unoptimized generator at update 0.

Scorers use bilinear resizing to 224×224 without antialiasing or center cropping, followed by their respective normalization. Target rewards are evaluated on floating-point RGB outputs; held-out rewards are evaluated on the corresponding 8-bit RGB images. PickScore uses the frozen model's learned scaling of cosine similarity between normalized text and image embeddings. We report these logits without candidate-set softmax normalization.

Source evaluation. Source and arrangement studies use the ten prompts in Table 10, thirty images per condition–prompt pair, and a separate $1{,}000$-image Gaussian reference per model–prompt pair. Source construction and evaluation are described in Appendix B; arrangement controls are defined in Appendix C. The lower-tail count $n_{1\%}$ counts images below their prompt-specific empirical first PickScore percentile under the Gaussian reference.

Table 10: Fixed prompts for the source and arrangement studies.

	Prompt	Subject	Style
1	*a photo of a red fox in tall grass*	animal	plain
2	*a photo of a tabby cat on a wooden stool*	animal	plain
3	*a photo of a ceramic teapot on a linen cloth*	object	plain
4	*a macro photograph of a blue butterfly on a yellow flower*	animal	photographic
5	*a small friendly white robot watering plants in a glass greenhouse*	object	rendered
6	*an aerial photograph of a winding turquoise river through an autumn forest*	scene	photographic
7	*a photograph of a snow-covered mountain peak under a clear blue sky*	scene	photographic
8	*a watercolor painting of a lighthouse on a cliff at sunset*	scene	painted
9	*a cinematic portrait photograph of an elderly violin maker in a wooden workshop*	person	photographic
10	*an astronaut riding a horse through a neon-lit city at night*	composite	rendered

F.2 Optimizers and implementation

Optimizer settings. SGDM (Polyak, 1964; Sutskever et al., 2013) and the momentum controls use $\beta = 0.9$ without Nesterov acceleration. Adam (Kingma & Ba, 2015) uses $(\beta_1, \beta_2) = (0.9, 0.999)$ and $\epsilon = 10^{-8}$. Each baseline's regularizer and base optimizer follow its original paper or released code. ReNO (Eyring et al., 2024) uses its norm regularizer with the released weight 0.01 and

Nesterov momentum 0.9. MPGR (Hwang et al., 2025) uses moment orders $\{1, 2\}$, frequency blocks of 16, a Fourier-magnitude target of 0.875, spectral-loss weight 25, and regularization coefficient 2.0 applied to a gradient normalized to the reward-gradient RMS, with Nesterov momentum 0.9. What Is Noise (Chang et al., 2026b) uses its Bethe–Kikuchi regularizer at three pyramid levels with 128 bins, the relative gradient weight 460 reported in its supplement, and plain SGD. GPER (Hwang & Sung, 2026) follows its released implementation, normalizing and projecting the reward gradient before Adam and projecting the latent after each update; we leave the initial latent unprojected so that every method starts from the same latent.

We do not adopt each paper's step size, iteration count, or gradient clipping. All methods instead share the protocol of this section: a fixed proposal-update RMS swept per method, the same number of updates per model, and a reward-gradient RMS cap of 0.03, the default of GPER's released code, applied before regularizer gradients are added.

Muon representation and scaling. Muon (Jordan et al., 2024) uses the nearest-square representation selected in Section 4.2, following Muse (Chang et al., 2026a). FLUX.1 and Z-Image refold 1024×64 to 256×256; FLUX.2 refolds 4096×128 to 1024×512. We compute $\boldsymbol{U}\boldsymbol{V}^\top$ from the reduced SVD in `float32` and restore the original layout.

We control the update size by its root mean square (RMS). For a processed direction $\boldsymbol{D}_t$, we use $\mathrm{RMS}(\boldsymbol{D}_t) = \|\boldsymbol{D}_t\|_F / \sqrt{Nd}$ and propose $\boldsymbol{X}_t + \eta \boldsymbol{D}_t / \mathrm{RMS}(\boldsymbol{D}_t)$. Thus the reported RMS η refers to the proposed update; a subsequent latent projection can change the final displacement. We vary η separately for each method; Appendix F.4 reports every tested value.

F.3 Results on Additional Generators

Figure 22 extends the optimizer comparison to Z-Image-Turbo and FLUX.2-klein-4B on both target rewards. We evaluate at update 50 for Z-Image and update 30 for FLUX.2, using the prompt sets, scorers, and aggregation in Appendix F.1.

Z-Image-Turbo. Muon attains the highest tested target mean on both tasks. For PickScore, it reaches 22.83 while all three held-out means exceed Base, compared with 22.61 for GPER under the same criterion. For Aesthetic Score, the curve extends to 7.10, with reductions in held-out scores as optimization becomes stronger.

FLUX.2-klein-4B. Muon reaches an Aesthetic Score of 8.65 while all three held-out means remain above Base. When optimizing PickScore, it reaches 31.42 with all three held-out means above Base, compared with at most 29.49 among the tested baselines.

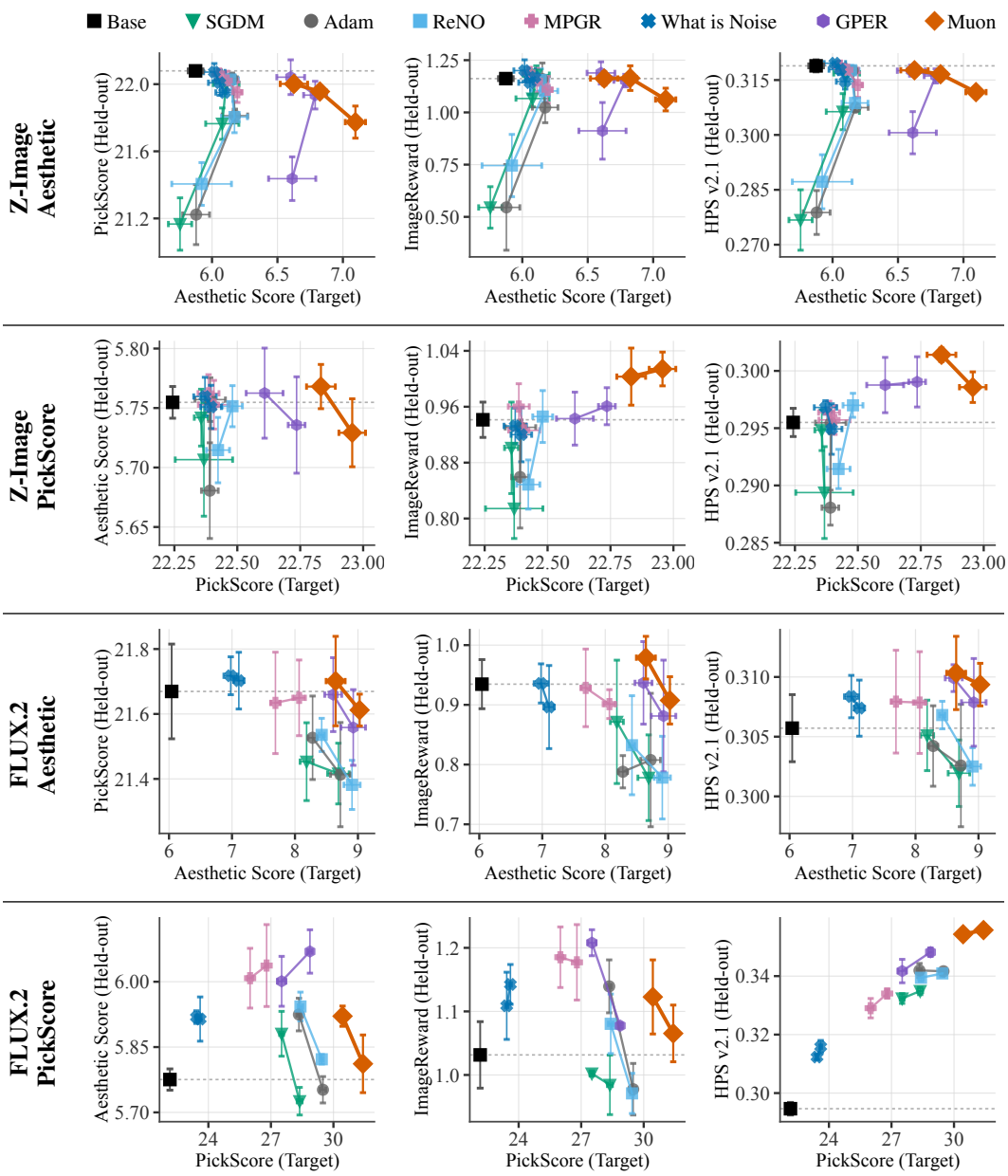


Figure 22: Reward–held-out-score trade-offs on Z-Image-Turbo at update 50 (top two rows) and FLUX.2-klein-4B at update 30 (bottom two rows). Within each model, Aesthetic precedes PickScore. Farther right and farther up are better. Points show three-seed means with one seed standard deviation; lines connect the tested RMS settings. Dashed lines mark each task's shared step-0 held-out score. All plotted values appear in Tables 13, 14, 15, and 16.

F.4 Full Quantitative Results

Tables 11 and 12 list all FLUX.1-schnell operating points shown in Figure 8. Tables 13 and 14 provide the Z-Image-Turbo results, and Tables 15 and 16 provide the FLUX.2-klein-4B results shown in Figure 22. Each table includes the shared step-0 scores and every tested RMS setting, with the target and all three held-out scores reported as mean ± one standard deviation across three seed-wise prompt means.

Table 11: All Aesthetic-guided operating points plotted in Figure 8 at update 100 (45 prompts, three seeds). Each score is mean ± one standard deviation across seed-wise prompt means; the step-0 row is shared by all methods.

Method	RMS	Aesthetic (target)	PickScore	ImageReward	HPSv2.1
Step 0	–	6.043 ± 0.049	21.861 ± 0.133	1.152 ± 0.036	0.3098 ± 0.0048
SGDM	0.00025	6.664 ± 0.068	21.958 ± 0.114	1.163 ± 0.068	0.3123 ± 0.0045
	0.0005	6.651 ± 0.103	21.939 ± 0.166	1.149 ± 0.066	0.3098 ± 0.0081
	0.001	6.509 ± 0.063	21.750 ± 0.104	1.047 ± 0.093	0.3031 ± 0.0049
	0.0015	6.406 ± 0.234	21.448 ± 0.226	0.809 ± 0.145	0.2896 ± 0.0074
Adam	0.00025	6.767 ± 0.119	21.935 ± 0.070	1.204 ± 0.055	0.3127 ± 0.0023
	0.0005	6.699 ± 0.058	21.901 ± 0.111	1.169 ± 0.015	0.3097 ± 0.0034
	0.001	6.601 ± 0.095	21.765 ± 0.081	1.101 ± 0.061	0.3050 ± 0.0031
	0.0015	6.489 ± 0.084	21.468 ± 0.126	0.921 ± 0.082	0.2922 ± 0.0037
ReNO	0.00025	6.787 ± 0.174	21.931 ± 0.116	1.162 ± 0.095	0.3110 ± 0.0061
	0.0005	6.817 ± 0.116	21.937 ± 0.074	1.173 ± 0.043	0.3123 ± 0.0034
	0.001	6.794 ± 0.088	21.883 ± 0.092	1.139 ± 0.107	0.3068 ± 0.0039
	0.0015	6.588 ± 0.184	21.651 ± 0.199	1.061 ± 0.110	0.2989 ± 0.0073
MPGR	0.00025	6.667 ± 0.114	21.922 ± 0.117	1.179 ± 0.040	0.3116 ± 0.0043
	0.0005	6.791 ± 0.110	21.944 ± 0.143	1.193 ± 0.064	0.3127 ± 0.0041
	0.001	6.779 ± 0.077	21.941 ± 0.104	1.191 ± 0.031	0.3110 ± 0.0047
	0.0015	6.719 ± 0.069	21.956 ± 0.057	1.170 ± 0.058	0.3106 ± 0.0052
What is Noise	0.00025	6.552 ± 0.116	21.911 ± 0.146	1.188 ± 0.035	0.3113 ± 0.0050
	0.0005	6.593 ± 0.143	21.920 ± 0.123	1.182 ± 0.063	0.3114 ± 0.0049
	0.001	6.552 ± 0.064	21.947 ± 0.073	1.156 ± 0.073	0.3117 ± 0.0045
	0.0015	6.532 ± 0.078	21.928 ± 0.143	1.122 ± 0.094	0.3105 ± 0.0060
GPER	0.001	7.359 ± 0.061	21.987 ± 0.061	1.208 ± 0.051	0.3145 ± 0.0031
	0.0025	7.468 ± 0.059	21.980 ± 0.061	1.175 ± 0.108	0.3134 ± 0.0034
	0.005	7.278 ± 0.155	21.601 ± 0.172	0.998 ± 0.128	0.3005 ± 0.0061
	0.0075	6.866 ± 0.162	20.913 ± 0.109	0.475 ± 0.107	0.2762 ± 0.0061
Muon	0.005	7.986 ± 0.033	21.963 ± 0.143	1.207 ± 0.066	0.3137 ± 0.0056
	0.0075	8.051 ± 0.094	21.877 ± 0.058	1.181 ± 0.092	0.3110 ± 0.0035
	0.01	8.364 ± 0.032	21.848 ± 0.051	1.176 ± 0.048	0.3111 ± 0.0018
	0.0125	8.495 ± 0.142	21.775 ± 0.162	1.131 ± 0.144	0.3051 ± 0.0072

Table 12: All PickScore-guided operating points plotted in Figure 8 at update 100 (100 prompts, three seeds). Each score is mean ± one standard deviation across seed-wise prompt means; the step-0 row is shared by all methods.

Method	RMS	PickScore (target)	Aesthetic	ImageReward	HPSv2.1
Step 0	–	22.163 ± 0.121	5.767 ± 0.005	1.029 ± 0.131	0.2924 ± 0.0053
SGDM	0.00025	22.923 ± 0.061	5.802 ± 0.021	1.156 ± 0.022	0.3045 ± 0.0035
	0.001	22.772 ± 0.083	5.696 ± 0.010	1.060 ± 0.088	0.2964 ± 0.0051
Adam	0.00025	22.990 ± 0.101	5.823 ± 0.036	1.164 ± 0.031	0.3053 ± 0.0037
	0.001	22.882 ± 0.012	5.711 ± 0.021	1.120 ± 0.082	0.2984 ± 0.0050
ReNO	0.00025	23.009 ± 0.126	5.823 ± 0.025	1.150 ± 0.060	0.3048 ± 0.0052
	0.001	22.936 ± 0.036	5.732 ± 0.030	1.140 ± 0.105	0.2999 ± 0.0043
MPGR	0.00025	22.925 ± 0.067	5.806 ± 0.010	1.125 ± 0.072	0.3031 ± 0.0047
	0.001	23.027 ± 0.103	5.813 ± 0.023	1.171 ± 0.073	0.3050 ± 0.0049
What is Noise	0.00025	22.842 ± 0.104	5.813 ± 0.030	1.133 ± 0.048	0.3026 ± 0.0050
	0.001	22.850 ± 0.083	5.830 ± 0.019	1.118 ± 0.011	0.3028 ± 0.0032
GPER	0.002	23.348 ± 0.084	5.819 ± 0.049	1.190 ± 0.053	0.3096 ± 0.0048
	0.004	23.177 ± 0.118	5.708 ± 0.038	1.069 ± 0.040	0.3004 ± 0.0040
Muon	0.004	23.777 ± 0.094	5.837 ± 0.050	1.202 ± 0.099	0.3142 ± 0.0042
	0.008	23.973 ± 0.075	5.799 ± 0.034	1.179 ± 0.014	0.3116 ± 0.0030

Table 13: Complete Z-Image-Turbo Aesthetic-guided operating points at update 50 (45 prompts, three seeds). Each entry is mean ± one standard deviation across seed-wise prompt means; step 0 is shared by all methods.

Method	RMS	Aesthetic (target)	PickScore	ImageReward	HPSv2.1
Step 0	–	5.874 ± 0.056	22.079 ± 0.028	1.162 ± 0.017	0.3189 ± 0.0017
Adam	0.001	6.153 ± 0.041	22.034 ± 0.033	1.159 ± 0.077	0.3176 ± 0.0008
	0.0025	6.175 ± 0.098	21.812 ± 0.039	1.024 ± 0.073	0.3075 ± 0.0010
	0.005	5.879 ± 0.103	21.223 ± 0.179	0.545 ± 0.205	0.2788 ± 0.0060
SGDM	0.001	6.108 ± 0.099	22.002 ± 0.003	1.180 ± 0.043	0.3168 ± 0.0006
	0.0025	6.078 ± 0.127	21.765 ± 0.093	1.066 ± 0.040	0.3064 ± 0.0049
	0.005	5.756 ± 0.089	21.167 ± 0.156	0.545 ± 0.099	0.2767 ± 0.0083
ReNO	0.001	6.141 ± 0.061	22.028 ± 0.032	1.155 ± 0.053	0.3177 ± 0.0013
	0.0025	6.172 ± 0.099	21.804 ± 0.093	1.103 ± 0.051	0.3087 ± 0.0014
	0.005	5.921 ± 0.228	21.405 ± 0.127	0.746 ± 0.149	0.2872 ± 0.0074
MPGR	0.001	6.055 ± 0.051	22.063 ± 0.023	1.192 ± 0.024	0.3188 ± 0.0014
	0.0025	6.113 ± 0.084	22.022 ± 0.056	1.152 ± 0.058	0.3178 ± 0.0003
	0.005	6.194 ± 0.033	21.952 ± 0.061	1.109 ± 0.064	0.3136 ± 0.0034
What is Noise	0.001	6.018 ± 0.082	22.072 ± 0.051	1.200 ± 0.052	0.3196 ± 0.0004
	0.0025	6.059 ± 0.086	22.008 ± 0.075	1.143 ± 0.057	0.3179 ± 0.0014
	0.005	6.096 ± 0.067	21.951 ± 0.006	1.169 ± 0.033	0.3146 ± 0.0021
GPER	0.0025	6.601 ± 0.106	22.042 ± 0.104	1.189 ± 0.053	0.3176 ± 0.0015
	0.005	6.788 ± 0.088	21.935 ± 0.083	1.152 ± 0.030	0.3161 ± 0.0019
	0.01	6.613 ± 0.180	21.437 ± 0.130	0.912 ± 0.136	0.3006 ± 0.0058
Muon	0.0025	6.626 ± 0.105	22.002 ± 0.035	1.162 ± 0.032	0.3177 ± 0.0015
	0.005	6.825 ± 0.047	21.956 ± 0.021	1.164 ± 0.059	0.3166 ± 0.0007
	0.01	7.096 ± 0.077	21.774 ± 0.096	1.062 ± 0.054	0.3117 ± 0.0013

Table 14: Complete Z-Image-Turbo PickScore-guided operating points at update 50 (100 prompts, three seeds). Each entry is mean ± one standard deviation across seed-wise prompt means; step 0 is shared by all methods.

Method	RMS	PickScore (target)	Aesthetic	ImageReward	HPSv2.1
Step 0	–	22.243 ± 0.017	5.755 ± 0.013	0.941 ± 0.025	0.2955 ± 0.0012
Adam	0.001	22.392 ± 0.061	5.757 ± 0.018	0.931 ± 0.010	0.2955 ± 0.0009
	0.0025	22.391 ± 0.033	5.681 ± 0.040	0.859 ± 0.073	0.2881 ± 0.0015
SGDM	0.001	22.357 ± 0.027	5.742 ± 0.024	0.901 ± 0.065	0.2948 ± 0.0018
	0.0025	22.367 ± 0.114	5.707 ± 0.048	0.814 ± 0.043	0.2894 ± 0.0040
ReNO	0.001	22.481 ± 0.040	5.752 ± 0.017	0.946 ± 0.037	0.2970 ± 0.0010
	0.0025	22.423 ± 0.046	5.715 ± 0.027	0.849 ± 0.035	0.2914 ± 0.0017
MPGR	0.001	22.386 ± 0.042	5.762 ± 0.016	0.960 ± 0.033	0.2964 ± 0.0010
	0.0025	22.407 ± 0.062	5.753 ± 0.020	0.926 ± 0.009	0.2957 ± 0.0013
What is Noise	0.001	22.372 ± 0.045	5.760 ± 0.016	0.933 ± 0.025	0.2968 ± 0.0006
	0.0025	22.396 ± 0.041	5.751 ± 0.018	0.920 ± 0.039	0.2949 ± 0.0022
GPER	0.0025	22.609 ± 0.073	5.763 ± 0.038	0.943 ± 0.038	0.2988 ± 0.0024
	0.005	22.736 ± 0.033	5.736 ± 0.040	0.961 ± 0.026	0.2990 ± 0.0022
Muon	0.005	22.832 ± 0.057	5.768 ± 0.019	1.003 ± 0.041	0.3014 ± 0.0006
	0.01	22.957 ± 0.053	5.729 ± 0.029	1.014 ± 0.024	0.2986 ± 0.0013

Table 15: Complete FLUX.2-klein-4B Aesthetic-guided operating points at update 30 (45 prompts, three seeds). Each entry is mean ± one standard deviation across seed-wise prompt means; step 0 is shared by all methods.

Method	RMS	Aesthetic (target)	PickScore	ImageReward	HPSv2.1
Step 0	–	6.042 ± 0.052	21.669 ± 0.146	0.935 ± 0.041	0.3057 ± 0.0028
Adam	0.005	8.280 ± 0.073	21.527 ± 0.129	0.788 ± 0.027	0.3042 ± 0.0034
	0.0075	8.721 ± 0.156	21.413 ± 0.160	0.808 ± 0.112	0.3026 ± 0.0051
SGDM	0.005	8.185 ± 0.103	21.453 ± 0.120	0.872 ± 0.103	0.3051 ± 0.0030
	0.0075	8.690 ± 0.171	21.416 ± 0.094	0.778 ± 0.072	0.3019 ± 0.0028
ReNO	0.005	8.424 ± 0.041	21.536 ± 0.051	0.832 ± 0.083	0.3068 ± 0.0012
	0.0075	8.906 ± 0.128	21.382 ± 0.076	0.778 ± 0.069	0.3025 ± 0.0016
MPGR	0.005	7.690 ± 0.073	21.634 ± 0.156	0.928 ± 0.065	0.3079 ± 0.0043
	0.0075	8.066 ± 0.100	21.650 ± 0.117	0.901 ± 0.024	0.3079 ± 0.0043
What is Noise	0.005	7.108 ± 0.083	21.703 ± 0.087	0.896 ± 0.070	0.3074 ± 0.0023
	0.0075	6.980 ± 0.109	21.718 ± 0.058	0.936 ± 0.033	0.3084 ± 0.0018
GPER	0.005	8.607 ± 0.135	21.660 ± 0.114	0.937 ± 0.069	0.3099 ± 0.0011
	0.0075	8.926 ± 0.195	21.558 ± 0.117	0.881 ± 0.094	0.3079 ± 0.0037
Muon	0.005	8.646 ± 0.151	21.701 ± 0.138	0.979 ± 0.036	0.3103 ± 0.0031
	0.0075	9.027 ± 0.060	21.612 ± 0.049	0.907 ± 0.040	0.3094 ± 0.0018

Table 16: Complete FLUX.2-klein-4B PickScore-guided operating points at update 30 (100 prompts, three seeds). Each entry is mean ± one standard deviation across seed-wise prompt means; step 0 is shared by all methods.

Method	RMS	PickScore (target)	Aesthetic	ImageReward	HPSv2.1
Step 0	–	22.155 ± 0.073	5.775 ± 0.024	1.032 ± 0.052	0.2947 ± 0.0022
Adam	0.03	28.344 ± 0.096	5.925 ± 0.037	1.139 ± 0.042	0.3419 ± 0.0024
	0.04	29.495 ± 0.137	5.752 ± 0.030	0.978 ± 0.041	0.3417 ± 0.0007
SGDM	0.03	27.512 ± 0.092	5.881 ± 0.051	1.002 ± 0.004	0.3323 ± 0.0017
	0.04	28.378 ± 0.119	5.726 ± 0.032	0.984 ± 0.047	0.3349 ± 0.0012
ReNO	0.03	28.417 ± 0.110	5.943 ± 0.033	1.080 ± 0.047	0.3394 ± 0.0012
	0.04	29.440 ± 0.076	5.822 ± 0.011	0.971 ± 0.032	0.3409 ± 0.0011
MPGR	0.03	26.002 ± 0.118	6.008 ± 0.069	1.185 ± 0.048	0.3290 ± 0.0033
	0.04	26.795 ± 0.060	6.037 ± 0.094	1.177 ± 0.059	0.3340 ± 0.0015
What is Noise	0.03	23.428 ± 0.052	5.919 ± 0.015	1.109 ± 0.053	0.3123 ± 0.0010
	0.04	23.605 ± 0.050	5.914 ± 0.051	1.142 ± 0.031	0.3159 ± 0.0021
GPER	0.04	27.515 ± 0.201	6.001 ± 0.057	1.208 ± 0.020	0.3417 ± 0.0040
	0.08	28.873 ± 0.104	6.069 ± 0.050	1.078 ± 0.004	0.3482 ± 0.0015
Muon	0.06	30.437 ± 0.146	5.921 ± 0.023	1.123 ± 0.058	0.3543 ± 0.0016
	0.08	31.416 ± 0.123	5.812 ± 0.066	1.065 ± 0.045	0.3557 ± 0.0019