COUPLE TO CONTROL: JOINT INITIAL NOISE DESIGN IN DIFFUSION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Diffusion models typically generate batches from independent Gaussian initial noises. We argue that independence is only one choice within a broader class of valid joint noise designs. One can instead specify a coupling: each noise remains marginally standard Gaussian, while the dependence across samples is chosen by design. This reframes initial-noise control from selecting individual seeds to designing a multi-sample gallery's joint distribution. The framework covers familiar constructions and supports direct sampling, structured initialization for noise optimization, and amortized learning of couplings. Repulsive Gaussian coupling improves gallery diversity on SD1.5, SDXL, SD3, Qwen-Image, and FLUX.1-schnell with negligible initialization overhead. On its own, it matches or outperforms several recent test-time approaches for diverse generation on several diversity metrics at lower computational cost. Because it modifies only the initial noise, it can also be layered on top of contrastive noise optimization, particle guidance, and condition-annealed sampling to further improve pixel and perceptual diversity. We further extend the same design principle to video generation, where whole-latent coupling also yields significant gains in cross-video visual diversity.

1 INTRODUCTION

Modern diffusion models have achieved remarkable photorealism, but their practical value increasingly depends on two goals: *quality*, producing realistic and prompt-faithful outputs, and *control*, following user-specified structure across generated outputs.

The quality objective has been extensively studied and engineered. Guidance mechanisms such as classifier guidance (Dhariwal & Nichol, 2021), classifier-free guidance (Ho & Salimans, 2022), and CLIP guidance (Radford et al., 2021) provide effective ways to steer samples toward high-quality, semantically aligned outputs. Control is more subtle, however, since it can reflect different user intentions. In a standard text-to-image (T2I) generation task, a user may provide a single prompt and want a diverse set of plausible images; in this case, the desired control is over the diversity of the generated gallery. A closely related goal arises in video generation, where a user may want distinct candidate clips while retaining each clip's native temporal structure. In other settings, such as e-commerce or product editing, the user may instead want to preserve a specific object while varying its surrounding context.

A natural way to control a diffusion model's output is through its initial noise. Recent studies show structured relationships between noise and output quality, layout, object position, and image perturbations (Xu et al., 2025; Ban et al., 2025; Song et al., 2025). Antithetic noise induces negatively correlated outputs (Jia et al., 2026a). Meanwhile, inference-time methods select or optimize realized noises for alignment, aesthetics, or diversity. This raises a natural question: before any per-sample optimization, can we design the joint distribution from which a gallery's initial noises are drawn?

We answer this question through a coupling framework for initial-noise design. A coupling specifies a joint distribution while preserving its marginals. Here, $(Z_1, \ldots, Z_K)$ is sampled jointly with each Z_i standard Gaussian, although the noises need not be independent. The pretrained model receives the usual single-sample input distribution, while the dependence across samples can be chosen by design. Thus coupling changes gallery-level behavior without changing any individual output's marginal distribution under a fixed generator.

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.

This view turns coupling into a flexible control layer underneath existing initial-noise methods. It naturally supports three modes of use. First, coupled noise can be used *directly* as a plug-in that replaces independent noise at the start of sampling, with no change to the pretrained model, prompt, sampler, or guidance rule. Second, coupled noise can serve as a *structured initialization for existing test-time methods*. Because it changes only the joint distribution of the starting noises, the downstream method can be applied without modifying its algorithm or hyperparameters, while potentially gaining further diversity from the coupled initialization. This allows coupling to combine its own diversity gains with the complementary benefits of existing optimization and sampling methods. Third, the coupling itself can be *amortized*: a marginal-preserving coupling can be learned once and then reused across new prompts without per-prompt optimization.

Jia et al. (2026a) study antithetic initialization in diffusion models, with a primary focus on negative output correlations and uncertainty quantification. They also introduce the K-sample equicorrelated Gaussian construction used in Algorithm 1 and demonstrate diversity gains for antithetic image pairs. Building on this foundation, we develop a framework for designing joint initial-noise distributions to control multi-sample galleries while preserving standard-Gaussian marginals. Our formulation supports subspace-dependent designs and amortized learning of reusable couplings for specified gallery objectives. We establish optimality properties of repulsive coupling for pairwise correlation and linear-feature diversity, and give an exact characterization of its effect on nonlinear feature distances. We further provide a systematic evaluation of multi-sample diversity, its dependence on gallery size and coupling strength, and its composition with existing diversity methods across image and video generation. The shared repulsive construction serves as an analytically tractable baseline within this broader design framework.

Our main contributions are as follows. First, we formulate gallery-level control through marginal-preserving initial-noise couplings, with a tractable Gaussian parameterization that supports subspace-dependent designs and amortized learning. Second, we establish the minimax pairwise-correlation property of repulsive coupling, its optimality for linear-feature diversity objectives, and an exact characterization of how Gaussian coupling affects nonlinear feature distances. Third, we evaluate coupled initialization across five image backbones and demonstrate complementary diversity gains when combined with CNO (Kim et al., 2026), Particle Guidance (Corso et al., 2024), and CADS (Sadat et al., 2024). We further investigate whole-latent coupling for video generation and structured initialization for fixed-object background generation.

2 Related Work

Initial noise: As the input to a diffusion model, initial noise has been studied both as a lens for understanding the input–output mechanism and as a lever for improving generation. Besides the mechanism studies mentioned in Section 1, optimizing initial noise has recently attracted interest for its potential to support inference-time scaling (Ma et al., 2025). These techniques have been shown to improve performance across many applications, including text-image alignment, aesthetics, diversity, and inverse problem solving; see (Guo et al., 2024; Eyring et al., 2024; Qi et al., 2024; Ben-Hamu et al., 2024; Tang et al., 2025; Zhou et al., 2025; Chihaoui et al., 2024; Wang et al., 2024; Wan et al., 2025; Jia et al., 2026b).

Coupling constructs random variables on the same probability space with prescribed marginal distributions, while allowing them to be dependent. It has played an important role in probability and statistics, especially in the study of Markov chains (Harris, 1955). Couplings are among the most fundamental tools for analyzing convergence rates of Markov chains (Levin & Peres, 2017). They are also used as algorithmic devices for designing sampling methods (Propp & Wilson, 1996; Neal, 2017), diagnosing MCMC convergence (Johnson, 1998; Biswas et al., 2019), and constructing unbiased estimators (Glynn & Rhee, 2014; Jacob et al., 2020; Deligiannidis et al., 2024). See Lindvall (2002) for coupling as a proof technique and Atchadé & Jacob (2024) for coupling-based algorithms.

Diverse sampling. Contrastive Noise Optimization (CNO) optimizes initial noises (Kim et al., 2026); Particle Guidance introduces interactions during denoising (Corso et al., 2024); and CADS varies conditioning during sampling (Sadat et al., 2024). Coupling offers a complementary choice of joint initialization. We test both standalone performance and composition while keeping each method's native configuration fixed.

3 COUPLED INITIAL NOISE

3.1 COUPLING OF RANDOM VARIABLES

Given probability distributions $p_1, \ldots, p_k$ on the same measurable space $(E, \mathcal{E})$, a coupling is a probability distribution γ on the product space E^k such that, if $(X_1, \ldots, X_k) \sim \gamma$, then each marginal satisfies $X_i \sim p_i$ for all $i = 1, \ldots, k$. In other words, a coupling describes a joint distribution whose marginals are fixed to be $p_1, p_2, \ldots, p_k$. The set of all such couplings is denoted by $\Gamma(p_1, \ldots, p_k)$.

Given k standard Gaussian random variables in $\mathbb{R}^d$, a coupling constructs a batch of k random variables $(Z_1, \ldots, Z_k)$ such that each Z_i remains standard Gaussian, while allowing dependence across different indices. Figure 1 provides visualizations for the case $k = 2$ and $d = 1$.

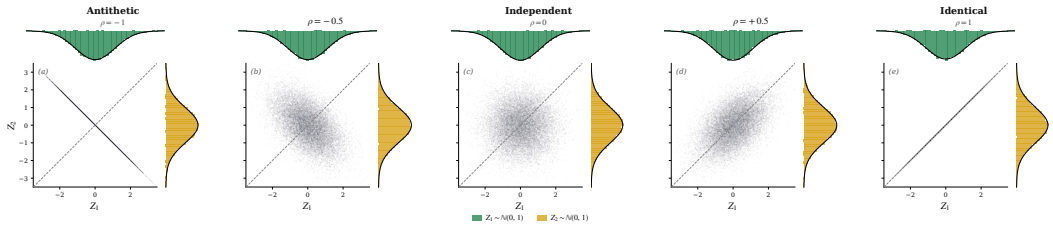

Figure 1: Five couplings of $Z_1, Z_2 \sim \mathcal{N}(0, 1)$, illustrated via 20,000 joint samples. Each panel shows the joint scatter of (Z_1, Z_2) with the marginal density of Z_1 (green, top) and Z_2 (amber, right); the dashed line marks $Y = X$. **(a) Antithetic** ($\rho = -1$): $Z_2 = -Z_1$, achieving maximal negative correlation. **(b)** $\rho = -0.5$: an intermediate negatively correlated Gaussian coupling. **(c) Independent** ($\rho = 0$): Z_1 and Z_2 are drawn independently. **(d)** $\rho = +0.5$: an intermediate positively correlated coupling. **(e) Identical** ($\rho = 1$): $Z_1 = Z_2$, all mass collapses onto the diagonal $Z_1 = Z_2$.

3.2 COUPLED GAUSSIAN NOISE

We study different couplings of the Gaussian initial noises. A few couplings useful for our paper are:

- *Identical coupling*: $Z \sim \mathcal{N}(0, I_d)$ and $Z_1 = Z_2 = \ldots = Z_k = Z$.
- *Independent coupling*: $(Z_1, \ldots, Z_k)$ are independent draws with $Z_i \sim \mathcal{N}(0, I_d)$. This serves as the standard baseline with no dependence across samples.
- *Antithetic coupling*: For $k = 2$, set $(Z_1, Z_2) = (Z, -Z)$ where $Z \sim \mathcal{N}(0, I_d)$. More generally, one can construct collections with strong negative correlations. This coupling induces strong negative dependence and is classically used for variance reduction; in diffusion generation, it can also encourage more diverse paired outputs.
- *Repulsive coupling*: Algorithm 1 generates negatively correlated Gaussian samples satisfying $\mathrm{Cov}(z_i, z_j) = -\frac{1}{k-1} I_d$, see Lemma A.1 for a formal statement and proof. It induces the same correlation between every pair of samples, and reduces to the antithetic coupling when $k = 2$.

Algorithm 1 Repulsive Coupled Initial Noise Sampling

Require: Gallery size $k \geq 2$, noise dimension d
Ensure: k initial noises that are all marginally standard Gaussian

1: Sample independent base noises $u_1, \ldots, u_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_d)$.
2: Compute the batch mean $\bar{u} \leftarrow k^{-1} \sum_{j=1}^{k} u_j$.
3: Construct repulsively coupled initial noises

$$z_i \leftarrow \sqrt{\frac{k}{k-1}} (u_i - \bar{u}), \qquad i = 1, \ldots, k.$$

4: **return** $(z_1, \ldots, z_k)$

The identical, independent, and repulsive couplings all belong to the equicorrelated Gaussian family

$$\mathrm{Cov}(z_i, z_j) = \begin{cases} I_d, & i = j, \\ cI_d, & i \neq j, \end{cases}$$

with $c = 1, 0, -1/(k-1)$ respectively. The following proposition shows equicorrelated coupling exists whenever $c \in [-1/(k-1), 1]$.

Proposition 3.1 (Equicorrelated Gaussian coupling and its minimax property)**.** *Let $k \geq 2$ and $d \geq 1$. For a scalar $c \in \mathbb{R}$, there exists a Gaussian coupling of k standard Gaussian vectors in $\mathbb{R}^d$ satisfying*

$$\mathrm{Cov}(z_i, z_j) = \begin{cases} I_d, & i = j, \\ cI_d, & i \neq j, \end{cases}$$

if and only if $-1/(k-1) \leq c \leq 1$. Moreover, the endpoint $c = -1/(k-1)$ is the most repulsive possible coupling in the following minimax sense. Let Γ_k denote the set of all couplings of k standard Gaussian vectors in $\mathbb{R}^d$. Then

$$\inf_{\gamma \in \Gamma_k} \max_{1 \leq i < j \leq k} \frac{1}{d} \mathbb{E}_\gamma \langle z_i, z_j \rangle = -\frac{1}{k-1}.$$

The equicorrelated Gaussian coupling with $c = -1/(k-1)$ attains this value.

Proposition 3.1 shows that the repulsive coupling is the most spread-out joint design for k initial noises in the minimax pairwise-correlation sense. This property makes it especially useful for the diverse generation experiments studied later. For linear feature maps within the Gaussian coupling family, it also minimizes average RBF similarity; see Appendix A.3.

We also study what coupling can and cannot change, and compare its effects with i.i.d. noise.

Proposition 3.2 (What coupling can and cannot change)**.** *Fix a prompt p and a deterministic generator G. Let $(Z_1, \ldots, Z_K)$ be any coupling such that $Z_i \sim \mathcal{N}(0, I_d)$ for each i, and define $X_i = G(p, Z_i)$. Then each X_i has the same marginal distribution as the output generated from an independent standard Gaussian noise $Z \sim \mathcal{N}(0, I_d)$. Consequently, for any integrable single-output score q, $\mathbb{E}\left[K^{-1} \sum_{i=1}^K q(p, X_i)\right] = \mathbb{E}_{Z \sim \mathcal{N}(0, I_d)} \left[q(p, G(p, Z))\right]$.*

Thus, changing the coupling cannot change the expected value of an averaged single-output metric, such as average image quality or average prompt alignment, as long as the generator and prompt are fixed. It can, however, change any metric that depends on two or more samples jointly, such as pairwise distance, similarity, diversity, or other gallery-level objectives. To understand the direction of this gallery-level effect, we give a quantitative result in Theorem A.3. Informally,

$$\text{effect of coupling} \approx \sum_{i<j} R_{ij}\, \mathbb{E}_{\mathrm{iid}} \left[\mathrm{tr}\big(\nabla^2_{ij} H_p\big)\right],$$

where H_p is the metric of interest. The formal statement and proof is deferred to the appendix. The proof interpolates between independent and coupled Gaussians along $R_t = (1-t)I_K + tR$ and applies Gaussian integration by parts.

Local linear interpretation. The effect of repulsive coupling can also be seen from a first-order approximation of the generator. Fix a prompt p, and let G_p denote the mapping from the initial noise to the generated image. Around a typical noise value, write $G_p(z) \approx a_p + J_p z$, where J_p is the local Jacobian. Suppose Z_i and Z_j are marginally standard Gaussian and satisfy $\mathrm{Cov}(Z_i, Z_j) = cI$. Then

$$\mathbb{E}\|G_p(Z_i) - G_p(Z_j)\|_2^2 \approx \mathbb{E}\|J_p(Z_i - Z_j)\|_2^2 = 2(1-c)\|J_p\|_F^2.$$

Under this local approximation, making c more negative increases the expected pairwise separation. Independent sampling corresponds to $c = 0$, while the repulsive equicorrelated coupling uses $c = -1/(K-1)$, giving $\mathbb{E}\|G_p(Z_i) - G_p(Z_j)\|_2^2 \approx 2\frac{K}{K-1}\|J_p\|_F^2$. This gives the local factor $K/(K-1)$ relative to independence. Appendix A.7 gives an exact nonlinear characterization: repulsion increases squared feature distance when the weighted odd-order Hermite energy exceeds the even-order energy. Section 5 provides an empirical examination.

3.3 Parametrizing Gaussian Couplings

The constructions above are particular choices within a broader design space of couplings. Rather than specifying the dependence structure by hand, we can parametrize a large class of valid noise couplings and then optimize over the resulting parameters. The following theorem gives a matrix parametrization for Gaussian couplings.

Theorem 3.3 (Matrix parametrization of Gaussian noise couplings)**.** *Let $u_1, \dots, u_r \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_d)$, and let $A \in \mathbb{R}^{M\times r}$ be a matrix with unit Euclidean norm on each row, i.e., $\sum_{\ell=1}^{r} A_{m\ell}^2 = 1$ for each m. Writing $Z = [z_1, \dots, z_M]^\top$ $U = [u_1, \dots, u_r]^\top$, and set $Z = AU$. Then each $z_m \sim \mathcal{N}(0, I_d)$. Hence $(z_1, \dots, z_M)$ defines a valid coupling of M standard Gaussian noise vectors in $\mathbb{R}^d$. Moreover, for every pair i, j,*

$$\mathrm{Cov}(z_i, z_j) = \left(\sum_{\ell=1}^{r} A_{i\ell} A_{j\ell}\right) I_d.$$

Conversely, suppose $(z_1, \dots, z_M)$ is a centered jointly Gaussian coupling such that each $z_m \sim \mathcal{N}(0, I_d)$ and $\mathrm{Cov}(z_i, z_j) = \rho_{ij} I_d$. Let $R = (\rho_{ij})_{i,j=1}^{M}$. If $\mathrm{rank}(R) \le r$, then there exists a matrix $A \in \mathbb{R}^{M\times r}$ with unit-norm rows such that $R = AA^\top$. With this choice of A, the construction above has the same joint distribution as $(z_1, \dots, z_M)$.

The parameterization reduces a $(Kd) \times (Kd)$ covariance design to a $K \times K$ sample-level matrix when the dependence is shared across coordinates. Independence uses $A = I_K$; repulsion uses $A = \sqrt{K/(K-1)}(I_K - K^{-1}\mathbf{1}\mathbf{1}^\top)$.

Coupling on a subspace: We can also couple the noises selectively on a subspace of interest, while using a simple default coupling on its complement. Let $V \subset \mathbb{R}^d$ be a subspace with orthogonal projection P_V, and let $P_{V^\perp} = I - P_V$. The subspace V may represent a color channel, a low-dimensional image statistic, a semantic direction, or a region associated with an object. We write

$$z_i = P_V z_i + P_{V^\perp} z_i.$$

On V, the coupling can be chosen to reflect the intended control: identical coupling for preservation, repulsive coupling for diversity, or a learnable coupling matrix A for optimization. On $V^\perp$, we keep a simple fixed coupling, for example an equicorrelated coupling or the independent coupling. Drawing the two components independently on the orthogonal subspaces, with standard Gaussian marginals on their respective subspaces, ensures that each full noise vector satisfies $z_i \sim \mathcal{N}(0, I_d)$. This parametrization lets us optimize the dependence structure only in the coordinates that are most relevant to the target behavior, while keeping a simple fixed coupling on the remaining directions.

3.4 Coupling-Based Noise Design and Optimization

Couplings can be used directly, optimized at the sample level, or learned at the distribution level. Direct repulsive sampling requires only centering and rescaling the initial noises. For test-time noise optimization, sample $z_{1:K}^{(0)} \sim \gamma$ and apply the existing objective

$$z_{1:K}^{\star} \in \arg\max_{z_1,\dots,z_K} \mathcal{R}_p\big(G(p, z_1), \dots, G(p, z_K)\big). \tag{1}$$

Here coupling supplies the initialization; subsequent optimization need not preserve Gaussian marginals. Alternatively, learn a reusable coupling itself:

$$\gamma^{\star} \in \arg\max_{\gamma\in\Gamma} \mathbb{E}_{p\sim\mathcal{P}} \mathbb{E}_{z_{1:K}\sim\gamma} \left[\mathcal{R}_p\big(G(p, z_1), \dots, G(p, z_K)\big)\right], \tag{2}$$

where Γ is the set of couplings with standard-Gaussian marginals. Row-normalized matrices from Theorem 3.3 provide a tractable family. The learned distribution can then be sampled on new prompts without test-time optimization. Appendix B.1.5 reports learned diversity couplings, and Appendix B.2 illustrates structured brightness relationships.

4 Applications

For deterministic samplers, write the output as $G(p, z)$. This includes DDIM, probability-flow ODEs, and flow-matching models (Song et al., 2021a; Lipman et al., 2023). For stochastic sam-

plers, include all injected noises in the input to G: coupling across galleries independently at each time step preserves each chain's original random-input law (Appendix D.4).

4.1 Diverse Generation

An important use case of coupled initial noise is diverse generation. Given a prompt p, a generator G, and a desired gallery size K, the goal is to generate K images that are diverse within the batch, while each image still maintains fidelity, visual quality, and prompt alignment. As discussed in Section 1, independently sampled noises can produce redundant outputs (Ban et al., 2025).

Our idea is to use the most spread-out initial-noise batch to encourage diverse outputs. Therefore, we sample $(z_1, \dots, z_K)$ from the repulsive Gaussian coupling in Algorithm 1, so that each $z_i \sim \mathcal{N}(0, I_d)$, while for $i \neq j$, $\mathrm{Cov}(z_i, z_j) = -\frac{1}{K-1} I_d$. We then generate the gallery by $x_i = G(p, z_i)$ for each i. Since each marginal noise remains standard Gaussian, each image is generated from the usual input distribution of the pretrained diffusion model. The only change is the dependence structure across the batch. Thus, repulsive coupling can increase gallery-level variation without modifying the prompt, model, sampler, guidance rule, or denoising procedure.

This gives a training-free and optimization-free method for diverse generation. It can also be combined with test-time diversity optimization methods by using the repulsive coupling as the initialization instead of independent Gaussian noise. We evaluate both uses in Section 5.1.

4.2 Background Generation with a Fixed Object

Another useful application is controlled background generation while preserving a specified foreground object. This task also illustrates a second use of coupled initial noise: the coupling supplies a structured initialization that can be directly combined with noise optimization, rather than replacing optimization altogether. Given an input image $x^\star$, a foreground mask $m \in \{0,1\}^{H \times W}$, and optionally a prompt p, the goal is to produce a gallery with diverse backgrounds while keeping the masked foreground faithful to $x^\star$ (see Figure 5 for illustration). This setting involves three desiderata: 1. the foreground object should be preserved, 2. the background should vary across the gallery, and 3. the resulting images should remain natural, without visible artifacts at the object-background boundary.

Coupling the foreground noise identically does not by itself guarantee exact foreground preservation after denoising. We therefore view the task as an inpainting problem: the foreground is the known region to be preserved, while the background is the region to be resampled. Starting from the coupled initialization $z_i^{(0)}$, we refine only the foreground component by optimizing the initial noise, following the diffusion inverse-problem strategy of Wang et al. (2024); Jia et al. (2026b).

$$\min_{z_1,\dots,z_K} \sum_{i=1}^{K} \left\| m \odot \left(G(p, z_i) - x^\star \right) \right\|_2^2, \qquad \text{such that} \quad (1 - M) \odot z_i = (1 - M) \odot z_i^{(0)}, \tag{3}$$

where $\odot$ denotes elementwise multiplication. The constraint keeps the repulsively coupled background noise fixed, so the optimization only improves foreground fidelity.

This procedure is training-free and can be applied to both unconditional diffusion models and text-conditioned models. For latent diffusion models, the image-space mask is downsampled to the latent resolution when constructing M, while the fidelity loss is evaluated in image space.

4.3 Whole-Latent Video Coupling

We consider a latent video diffusion model, where each video is represented and generated through a spatiotemporal latent tensor. Let each initial noise be a spatiotemporal latent tensor $z_i \in \mathbb{R}^{C \times T \times H \times W}$, where T is the temporal dimension and C, H, W are the latent channel and spatial dimensions. A gallery of K videos therefore starts from K such tensors $z_1, \dots, z_k$.

We apply Algorithm 1 across these K video latents, treating each tensor as one sample. Equivalently, we may vectorize each tensor as $z_i \in \mathbb{R}^d$, where $d = CTHW$, and apply the same coupling as in the image setting. Each z_i remains marginally $\mathcal{N}(0, I_d)$, so the single-video input distribution is unchanged. Section 5.4 evaluates this extension on Wan2.1.

5 EXPERIMENTS

We first evaluate repulsive coupling on the Stable Diffusion family (SD1.5, SDXL, and SD3). We then study its composition with complementary diversity methods, including CNO, Particle Guidance, and CADS; test modern image backbones including Qwen-Image and FLUX.1-schnell; and finally extend the evaluation to video generation with Wan2.1. Unless stated otherwise, L2 denotes average pairwise per-pixel squared distance, MSS and Vendi use SSCD features, and galleries have $K = 3$. Additional experiments are paired by prompt; their protocols and uncertainty estimates are in Appendix D.

5.1 DIVERSE TEXT-TO-IMAGE GENERATION

Our setup follows Kim et al. (2026). We evaluate 2,000 COCO prompts using SD1.5, SDXL, and SD3, comparing independent noise, independent noise with Contrastive Noise Optimization (CNO), coupled noise, and coupled noise with CNO. Table 1 reports diversity, CLIP alignment, and PickScore quality; Figures 2 and 3 illustrate generated galleries.

Coupled noise improves diversity across all three models while leaving mean quality and alignment nearly unchanged. For example, on SD3 it increases L2 from 0.1417 to 0.1667, reduces MSS from 0.3007 to 0.2827, and increases Vendi from 2.7256 to 2.7532. Coupling alone also matches or exceeds IID+CNO on several metrics, with essentially the same generation cost as IID. Combining coupling with CNO gives further gains: on SDXL, coupled+CNO achieves the strongest L2, MSS, and Vendi among the four settings. On SD3, it improves CNO's L2 from 0.1612 to 0.1844 and MSS from 0.2804 to 0.2675. The corresponding compute costs and ablations appear in Appendix B.1.

Coupling versus additional IID samples. We compare coupled $K = 3$ galleries with nested IID galleries of sizes $3, 4, 5, 6, 8$ on 100 prompts. For maximum pairwise LPIPS, the IID mean first exceeds the coupled value at $K = 5$, with a clear paired advantage at $K = 6$. For maximum pairwise L2, IID $K = 8$ reaches a similar observed level. At fixed per-image sampling settings, these observed reference points correspond to approximately 1.67–$2.67\times$ as many generated images for IID as for coupled $K = 3$. This budget comparison is specific to the two gallery-diameter objectives and the evaluated configuration; it does not establish a wall-clock speedup or a universal sample-efficiency factor. For mean pairwise distance, increasing the IID sample count leaves its expectation unchanged, whereas coupling changes the pairwise dependence. Appendix D.6 provides the detailed analysis.

Table 1: Diversity and quality on 2,000 COCO prompts per model ($K = 3$). Standard deviations appear in parentheses; Opt. indicates test-time noise optimization. MSS/Vendi use SSCD features.

Model	Method	Opt.	Diversity metrics			Quality metrics	
			L2 ↑	MSS ↓	Vendi ↑	CLIP ↑	PickScore ↑
SD1.5	i.i.d.	✗	0.126 (0.04)	0.150 (0.07)	2.913 (0.07)	0.662	21.407
	i.i.d.+CNO	✓	0.136 (0.04)	0.148 (0.07)	2.916 (0.07)	0.661	21.402
	coupled	✗	0.156 (0.04)	**0.141** (0.07)	**2.921** (0.07)	0.662	21.391
	coupled+CNO	✓	**0.163** (0.05)	0.144 (0.07)	2.920 (0.07)	0.660	21.399
SDXL	i.i.d.	✗	0.087 (0.02)	0.208 (0.08)	2.853 (0.10)	0.670	22.474
	i.i.d.+CNO	✓	0.106 (0.03)	0.187 (0.07)	2.879 (0.09)	0.668	22.419
	coupled	✗	0.105 (0.03)	0.198 (0.08)	2.866 (0.09)	0.669	22.459
	coupled+CNO	✓	**0.122** (0.03)	**0.184** (0.08)	**2.881** (0.09)	0.668	22.412
SD3	i.i.d.	✗	0.142 (0.05)	0.301 (0.10)	2.726 (0.15)	0.664	22.524
	i.i.d.+CNO	✓	0.161 (0.05)	0.280 (0.09)	2.755 (0.15)	0.664	22.505
	coupled	✗	0.167 (0.06)	0.283 (0.09)	2.753 (0.14)	0.664	22.529
	coupled+CNO	✓	**0.184** (0.06)	**0.267** (0.09)	**2.775** (0.14)	0.664	22.515

5.2 COUPLING AS A PLUG-IN FOR DIVERSITY METHODS

Coupling only changes the joint initialization and is therefore compatible with existing diversity-enhancing methods. Since these methods typically start from independent Gaussian noise, coupled initialization can provide complementary diversity gains without changing their algorithms or hyperparameters. We evaluate this combination with CNO, pixel- and DINO-feature Particle Guidance



Figure 2: Independent versus coupled noises with the same prompt "A classic motorcycle in a parking space along the street." The coupled-noise samples exhibit greater visual diversity across pose, viewpoint, object scale, and scene layout while remaining consistent with the prompt.

Table 2: **Coupled initialization improves existing diversity methods.** Relative gains from replacing IID initialization by coupling, on 300 prompts per setting. Reported ranges summarize variation across all tested settings. Quality NI means that mean PickScore drops by at most 1%. Native configurations and per-setting intervals are in Appendix D.2.

Method	Settings	L2 gain ↑	LPIPS gain ↑	Quality NI
CNO	3	20.28–21.15%	3.25–3.61%	Yes
Pixel-PG	3	14.88–15.24%	3.47–3.68%	Yes
DINO-PG	3	14.98–15.12%	3.65–3.76%	Yes
CADS	4	21.01–23.76%	2.87–3.13%	Yes

(PG; Corso et al., 2024), and Condition-Annealed Diffusion Sampling (CADS; Sadat et al., 2024) on 300 held-out COCO prompts. Each comparison changes only the initialization, keeping the method's sampler, steps, guidance, gallery size, and operating point fixed. Across all tested settings, coupled initialization improves both L2 and LPIPS (Table 2). Each setting also satisfies the predeclared mean-PickScore non-inferiority criterion, and coupling adds no measurable generation-time or peak-memory overhead. Experimental settings and detailed results are provided in Appendix D.2.

CADS provides a useful example of how different diversity metrics can be improved by combining complementary methods. In a matched 100-prompt experiment, CADS improves DINO-feature diversity compared with IID initialization but reduces pixel L2 diversity. Coupled initialization restores 21.83% L2 and adds 2.44% LPIPS relative to CADS, while keeping its feature-diversity gain (Table 3). This combination brings together CADS's semantic-feature variation and coupling's pixel/perceptual variation. The same principle applies without an extra optimization or model training.

5.3 Generalization to Modern Image Backbones

We extend the evaluation to Qwen-Image (Wu et al., 2025) and FLUX.1-schnell (Black Forest Labs, 2024). Coupling operates directly on each model's native latent representation, without retraining or changing its denoising settings. On Qwen-Image, it improves L2/LPIPS by $13.91\%/5.68\%$; on FLUX.1-schnell, evaluated on a disjoint 70-prompt split, the gains are $27.20\%/5.99\%$. Coupling also improves feature-space diversity, including SSCD MSS by 11.54% and Vendi by 1.08% on FLUX.1-schnell, supporting transfer beyond the Stable Diffusion family (Table 4). Qwen-Image additionally improves CLIP-image feature diversity. These results suggest that the diversity improvements observed in the Stable Diffusion family generalize to other modern diffusion backbones with different latent representations. The complete feature results and configurations are in Appendix D.3.

5.4 Video Generation

We evaluate whole-latent repulsive coupling on the official pretrained Wan2.1 T2V 1.3B (Team Wan et al., 2025), using 100 fixed COCO prompts and $K = 3$, without retraining or fine-tuning. Each video has 33 frames at 480×832, generated with 30 steps and CFG 5. Cross-video diversity averages all within-gallery pairwise comparisons over nine fixed, time-aligned frames. Coupling increases pixel MSE by 20.69% and LPIPS by 5.96%, respectively (Table 5). Both gains are statistically significant with Benjamini–Hochberg correction across five metrics ($p_{\mathrm{BH}} \approx 2.5 \times 10^{-5}$ for

Table 3: **Complementary diversity dimensions.** SD1.5, 100 prompts, $K = 3$, 100 steps and CFG 9. CADS supplies the stronger DINO-feature diversity; coupling improves its pixel/perceptual diversity.

Initialization / sampler	L2 ↑	LPIPS ↑	DINO MSS ↓	PickScore ↑
IID	0.1553	0.7517	0.6064	21.61
Coupled	0.1895	0.7855	0.5927	21.52
IID + CADS	0.1242	0.7598	0.5394	21.25
Coupled + CADS	0.1513	0.7783	0.5407	21.24

Table 4: **Coupling on modern image backbones.** Relative changes from IID at $K = 3$; n is the number of prompts.

Backbone	n	L2 ↑	LPIPS ↑	MSS ↓	Vendi ↑
Qwen-Image	50	+13.91%	+5.68%	-5.69%	+2.73%
FLUX.1-schnell	70	+27.20%	+5.99%	-11.54%	+1.08%

each). Mean frame-level CLIP does not have significant change. These results support extending initialization coupling to video for increased frame-level visual diversity. Appendix D.10 provides additional results including temporal diagnostics, and qualitative examples.

Table 5: **Video generation results.** Wan2.1 on 100 prompts, with $K = 3$ videos per prompt.

Initialization	Cross-video pixel MSE ↑	Cross-video LPIPS ↑	CLIP ↑
IID	0.1626	0.7497	0.6337
Coupled	0.1963	0.7943	0.6318

5.5 Fixed-Object Background Generation

We use coupling to initialize constrained noise optimization: the background remains repulsively coupled while the foreground is refined. Our LSUN-bedroom evaluation uses 92 valid images, $K = 8$, and SD1.5. Coupling generates diverse backgrounds while maintaining smooth object boundaries; increasing the optimization budget further improves foreground fidelity (Appendix B.3).

We also compare with SD1.5 and SDXL Inpaint on 30 matched cases. Against 20-step SDXL Inpaint, ours with 50 optimization steps achieves higher background LPIPS (0.4238 vs. 0.3611) and lower BRISQUE (22.10 vs. 47.33), but higher foreground RMSE (0.0581 vs. 0.0265), illustrating a diversity–fidelity trade-off. Full settings and results appear in Appendix D.9.

5.6 Additional Experiments and Analysis

We further examine the mechanism and scope of coupled initialization. Additional analyses of noise-to-image diversity transfer, stochastic sampling, gallery-size scaling, alternative coupling schemes, diversity baselines, and gallery diagnostics are provided in Appendix D.

6 Discussion and Limitations

The joint distribution of initial noises is a useful control variable for diffusion generation. Coupling offers a lightweight diversity improvement, complements established sampling methods, and supports structured initialization for fixed-object backgrounds. Results on modern image backbones and a 100-prompt video evaluation suggest a broader role for this design principle.

Our evidence is strongest for compact galleries and pixel/perceptual diversity; feature-space gains vary by backbone and representation, and benefits diminish with K. Future directions include semantic couplings, user-centered selection, and larger video studies, extending control over gallery diversity and quality while preserving individual marginals.

AI USE STATEMENT

Generative AI tools assisted implementation of standard methods, code inspection, organization and interpretation of experimental records, checking of theoretical arguments and proofs, manuscript polishing and editing. A separate exploratory study used an LLM to compare generated galleries; its setup and reported outcomes are described in Appendix D.11. Image and video results were produced by the specified diffusion models. The authors are responsible for the final manuscript and its claims.

REPRODUCIBILITY STATEMENT

Appendix A gives mathematical details and proofs. Appendices B.1, B.3, and D describe model settings, prompt splits, evaluation metrics, and supplementary results. The implementation is available at `https://anonymous.4open.science/r/diffusion-coupling-CE03/`.

REFERENCES

Yves F Atchadé and Pierre E Jacob. Unbiased Markov Chain Monte Carlo: what, why, and how. *arXiv preprint arXiv:2406.06851*, 2024.

Yuanhao Ban, Ruochen Wang, Tianyi Zhou, Boqing Gong, Cho-Jui Hsieh, and Minhao Cheng. The crystal ball hypothesis in diffusion models: Anticipating object positions from initial noise. In *The Thirteenth International Conference on Learning Representations*, 2025. URL `https://openreview.net/forum?id=GpdO9r73xT`.

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*, pp. 3462–3483. PMLR, 2024.

Niloy Biswas, Pierre E Jacob, and Paul Vanetti. Estimating convergence of markov chains with l-lag couplings. *Advances in neural information processing systems*, 32, 2019.

Black Forest Labs. FLUX.1 [schnell] model card. `https://huggingface.co/black-forest-labs/FLUX.1-schnell`, 2024.

Hamadi Chihaoui, Abdelhak Lemkhenter, and Paolo Favaro. Blind image restoration via fast diffusion inversion. *Advances in Neural Information Processing Systems*, 37:34513–34532, 2024.

Gabriele Corso, Yilun Xu, Valentin De Bortoli, Regina Barzilay, and Tommi S. Jaakkola. Particle guidance: non-i.i.d. diverse sampling with diffusion models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL `https://openreview.net/forum?id=KqbCvIFBY7`.

George Deligiannidis, Pierre E Jacob, El Mahdi Khribch, and Guanyang Wang. On importance sampling and independent metropolis-hastings with an unbounded weight function. *arXiv preprint arXiv:2411.09514*, 2024.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Luca Eyring, Shyamgopal Karthik, Karsten Roth, Alexey Dosovitskiy, and Zeynep Akata. ReNO: Enhancing one-step text-to-image models through reward-based noise optimization. *Advances in Neural Information Processing Systems*, 37:125487–125519, 2024.

Peter W Glynn and Chang-han Rhee. Exact estimation for markov chain equilibrium expectations. *Journal of Applied Probability*, 51(A):377–389, 2014.

Xiefan Guo, Jinlin Liu, Miaomiao Cui, Jiankai Li, Hongyu Yang, and Di Huang. Initno: Boosting text-to-image diffusion models via initial noise optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9380–9389, 2024.

T Harris. On chains of infinite order. *Pacific J. Math.*, 5:707–724, 1955.

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

Pierre E Jacob, John O'Leary, and Yves F Atchadé. Unbiased markov chain monte carlo methods with couplings. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82(3): 543–600, 2020.

Jing Jia, Sifan Liu, Bowen Song, Wei Yuan, Liyue Shen, and Guanyang Wang. Antithetic noise in diffusion models. In *The Fourteenth International Conference on Learning Representations*, 2026a. URL https://openreview.net/forum?id=9yFORC1tu3.

Jing Jia, Wei Yuan, Sifan Liu, Liyue Shen, and Guanyang Wang. Weak diffusion priors can still achieve strong inverse-problem performance. *arXiv preprint arXiv:2601.22443*, 2026b.

Valen E Johnson. A coupling-regeneration scheme for diagnosing convergence in markov chain monte carlo algorithms. *Journal of the American Statistical Association*, 93(441):238–248, 1998.

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

David A Levin and Yuval Peres. *Markov chains and mixing times*, volume 107. American Mathematical Soc., 2017.

Torgny Lindvall. *Lectures on the coupling method*. Courier Corporation, 2002.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Nanye Ma, Shangyuan Tong, Haolin Jia, Hexiang Hu, Yu-Chuan Su, Mingda Zhang, Xuan Yang, Yandong Li, Tommi Jaakkola, Xuhui Jia, et al. Scaling inference time compute for diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 2523–2534, 2025.

Radford M Neal. Circularly-coupled markov chain sampling. *arXiv preprint arXiv:1711.04399*, 2017.

Ed Pizzi, Sreya Dutta Roy, Sugosh Nagavara Ravindra, Priya Goyal, and Matthijs Douze. A self-supervised descriptor for image copy detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14532–14542, 2022.

James Gary Propp and David Bruce Wilson. Exact sampling with coupled markov chains and applications to statistical mechanics. *Random Structures & Algorithms*, 9(1-2):223–252, 1996.

Zipeng Qi, Lichen Bai, Haoyi Xiong, and Zeke Xie. Not all noises are created equally: Diffusion noise selection and optimization. *arXiv preprint arXiv:2407.14041*, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PmLR, 2021.

Seyedmorteza Sadat, Jakob Buhmann, Derek Bradley, Otmar Hilliges, and Romann Weber. Cads: Unleashing the diversity of diffusion models through condition-annealed sampling. In *International Conference on Learning Representations*, volume 2024, pp. 23723–23755, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/6769bfdeb7313ded45714aabf759adfe-Paper-Conference.pdf.

Bowen Song, Zecheng Zhang, Zhaoxu Luo, Jason Hu, Wei Yuan, Jing Jia, Zhengxu Tang, Guanyang Wang, and Liyue Shen. CCS: Controllable and constrained sampling with diffusion models via initial noise perturbation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=64lk7JQAU1.

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

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

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

Team Wan, Ang Wang, Baole Ai, Bin Wen, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. URL https://arxiv.org/abs/2503.20314.

Yuxiang Wan, Ryan Devera, Wenjie Zhang, and Ju Sun. Fmplug: Plug-in foundation flow-matching priors for inverse problems. *arXiv preprint arXiv:2508.00721*, 2025.

Hengkang Wang, Xu Zhang, Taihui Li, Yuxiang Wan, Tiancong Chen, and Ju Sun. Dmplug: A plug-in method for solving inverse problems with diffusion models. *Advances in Neural Information Processing Systems*, 37:117881–117916, 2024.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, et al. Qwen-Image technical report. *arXiv preprint arXiv:2508.02324*, 2025. URL https://arxiv.org/abs/2508.02324.

Katherine Xu, Lingzhi Zhang, and Jianbo Shi. Good seed makes a good crop: Discovering secret seeds in text-to-image diffusion models. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 3024–3034. IEEE, 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 conference on computer vision and pattern recognition*, pp. 586–595, 2018.

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

A Mathematical Details and Proofs

Throughout this appendix, K denotes the gallery size and d denotes the dimension of each initial-noise vector. We write I_d for the $d \times d$ identity matrix and $\mathbf{1}_K$ for the all-ones vector in $\mathbb{R}^K$. All Gaussian random vectors are allowed to be degenerate when the covariance matrix is positive semidefinite but not positive definite.

A.1 Repulsive Gaussian coupling

We first verify the Gaussian coupling used in Algorithm 1.

Lemma A.1 (Repulsive Gaussian coupling). *Let* $U_1, \dots, U_K \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_d)$*, define*

$$\bar{U} := \frac{1}{K}\sum_{j=1}^{K} U_j, \qquad Z_i := \sqrt{\frac{K}{K-1}}\,(U_i - \bar{U}), \qquad i = 1, \dots, K.$$

Then each $Z_i \sim \mathcal{N}(0, I_d)$*. Moreover, for* $i \neq j$*,*

$$\mathrm{Cov}(Z_i, Z_j) = -\frac{1}{K-1} I_d.$$

In addition,

$$\sum_{i=1}^{K} Z_i = 0 \qquad \textit{almost surely}.$$

Proof. Since Z_i is a linear transformation of jointly Gaussian random vectors, $(Z_1, \dots, Z_K)$ is jointly Gaussian. Also,

$$\mathbb{E}[Z_i] = 0.$$

For the marginal covariance, note that

$$\mathrm{Cov}(U_i, \bar{U}) = \frac{1}{K} I_d, \qquad \mathrm{Cov}(\bar{U}, \bar{U}) = \frac{1}{K} I_d.$$

Therefore,

$$\begin{aligned}\mathrm{Cov}(U_i - \bar{U}) &= \mathrm{Cov}(U_i) - 2\,\mathrm{Cov}(U_i, \bar{U}) + \mathrm{Cov}(\bar{U}) \\ &= I_d - \frac{2}{K} I_d + \frac{1}{K} I_d = \frac{K-1}{K} I_d.\end{aligned}$$

Multiplying by the factor $K/(K-1)$ gives

$$\mathrm{Cov}(Z_i) = I_d.$$

Thus $Z_i \sim \mathcal{N}(0, I_d)$.

For $i \neq j$,

$$\mathrm{Cov}(U_i, U_j) = 0, \qquad \mathrm{Cov}(U_i, \bar{U}) = \mathrm{Cov}(U_j, \bar{U}) = \frac{1}{K} I_d.$$

Hence

$$\begin{aligned}\mathrm{Cov}(U_i - \bar{U}, U_j - \bar{U}) &= \mathrm{Cov}(U_i, U_j) - \mathrm{Cov}(U_i, \bar{U}) - \mathrm{Cov}(\bar{U}, U_j) + \mathrm{Cov}(\bar{U}, \bar{U}) \\ &= 0 - \frac{1}{K} I_d - \frac{1}{K} I_d + \frac{1}{K} I_d = -\frac{1}{K} I_d.\end{aligned}$$

After multiplying by $K/(K-1)$, we obtain

$$\mathrm{Cov}(Z_i, Z_j) = -\frac{1}{K-1} I_d.$$

Finally,

$$\sum_{i=1}^{K} Z_i = \sqrt{\frac{K}{K-1}} \sum_{i=1}^{K} (U_i - \bar{U}) = \sqrt{\frac{K}{K-1}} \left(\sum_{i=1}^{K} U_i - K\bar{U} \right) = 0.$$

This proves the result. □

A.2 Proof of Proposition 3.1

We prove Proposition 3.1 from the main paper.

Proof. First consider the existence statement. If the stated equicorrelated Gaussian coupling exists, then its block covariance matrix has the form

$$\Sigma_c = R_c \otimes I_d,$$

where

$$R_c = \begin{pmatrix} 1 & c & \cdots & c \\ c & 1 & \cdots & c \\ \vdots & \vdots & \ddots & \vdots \\ c & c & \cdots & 1 \end{pmatrix} = (1-c)I_K + c\,\mathbf{1}_K\mathbf{1}_K^\top.$$

A centered Gaussian vector with covariance Σ_c exists if and only if Σ_c is positive semidefinite. Since $\Sigma_c = R_c \otimes I_d$, this is equivalent to R_c being positive semidefinite.

The eigenvalues of R_c are explicit. For any $v \in \mathbb{R}^K$ satisfying $\mathbf{1}_K^\top v = 0$,

$$R_c v = (1-c)v.$$

Hence $1-c$ is an eigenvalue with multiplicity $K-1$. Also,

$$R_c\mathbf{1}_K = \big(1 + (K-1)c\big)\mathbf{1}_K.$$

Thus the remaining eigenvalue is $1+(K-1)c$. Therefore R_c is positive semidefinite if and only if

$$1 - c \geq 0 \qquad \text{and} \qquad 1 + (K-1)c \geq 0.$$

Equivalently,

$$-\frac{1}{K-1} \leq c \leq 1.$$

This proves the existence part.

We prove the minimax property. Let $\gamma \in \Gamma_K$ be any coupling of K standard Gaussian vectors $Z_1, \ldots, Z_K \in \mathbb{R}^d$. Since each marginal is $\mathcal{N}(0, I_d)$,

$$\mathbb{E}\|Z_i\|_2^2 = d, \qquad i = 1, \ldots, K.$$

By nonnegativity of squared norm,

$$0 \leq \mathbb{E}\left\|\sum_{i=1}^K Z_i\right\|_2^2.$$

Expanding the right-hand side gives

$$\begin{aligned} 0 &\leq \sum_{i=1}^K \mathbb{E}\|Z_i\|_2^2 + 2\sum_{1\leq i<j\leq K} \mathbb{E}\langle Z_i, Z_j\rangle \\ &= Kd + 2\sum_{1\leq i<j\leq K} \mathbb{E}\langle Z_i, Z_j\rangle. \end{aligned}$$

Hence

$$\frac{2}{K(K-1)} \sum_{1\leq i<j\leq K} \frac{1}{d}\mathbb{E}\langle Z_i, Z_j\rangle \geq -\frac{1}{K-1}.$$

The maximum of a finite set is at least its average, so

$$\max_{1\leq i<j\leq K} \frac{1}{d}\mathbb{E}\langle Z_i, Z_j\rangle \geq -\frac{1}{K-1}.$$

Since this lower bound holds for every coupling $\gamma \in \Gamma_K$,

$$\inf_{\gamma\in\Gamma_K} \max_{1\leq i<j\leq K} \frac{1}{d}\mathbb{E}_\gamma\langle Z_i, Z_j\rangle \geq -\frac{1}{K-1}.$$

It remains to show that the lower bound is attainable. By the first part of the proposition, the equicorrelated Gaussian coupling with $c = -1/(K-1)$ exists. For this coupling,

$$\frac{1}{d}\mathbb{E}\langle Z_i, Z_j\rangle = -\frac{1}{K-1}, \qquad i \neq j.$$

Therefore,

$$\max_{1\le i<j\le K} \frac{1}{d}\mathbb{E}\langle Z_i, Z_j\rangle = -\frac{1}{K-1}.$$

This proves the minimax claim and completes the proof. □

A.3 Optimality of repulsive coupling

Proposition A.2 (Optimality of repulsive coupling for linear-feature diversity)**.** *Let $k \ge 2$, $d \ge 1$, $m \ge 1$, and let $J \in \mathbb{R}^{m\times d}$ be fixed. Let Γ_k denote the set of all couplings of k standard Gaussian vectors in $\mathbb{R}^d$. For $\gamma \in \Gamma_k$, write*

$$(z_1, \dots, z_k) \sim \gamma, \qquad y_i = Jz_i \in \mathbb{R}^m.$$

Define the average squared feature separation

$$\mathcal{D}_J(\gamma) := \frac{2}{k(k-1)} \sum_{1\le i<j\le k} \mathbb{E}_\gamma \|y_i - y_j\|_2^2.$$

Then

$$\mathcal{D}_J(\gamma) \le \frac{2k}{k-1}\|J\|_F^2 \qquad \text{for every } \gamma \in \Gamma_k.$$

Moreover, equality holds if and only if

$$\sum_{i=1}^k Jz_i = 0 \qquad \gamma\text{-a.s.}$$

In particular, the repulsive Gaussian coupling

$$z_i = \sqrt{\frac{k}{k-1}}\,(u_i - \bar{u}), \qquad \bar{u} = \frac{1}{k}\sum_{j=1}^k u_j, \qquad u_1, \dots, u_k \overset{\text{i.i.d.}}{\sim} N(0, I_d),$$

attains this upper bound.

Now let $\Gamma_k^{\mathrm{G}} \subset \Gamma_k$ denote the set of possibly degenerate jointly Gaussian couplings with standard Gaussian marginals. For $\tau > 0$, define the average RBF feature similarity

$$\mathcal{S}_{J,\tau}(\gamma) := \frac{2}{k(k-1)} \sum_{1\le i<j\le k} \mathbb{E}_\gamma \exp\left(-\frac{\|y_i - y_j\|_2^2}{2\tau^2}\right).$$

Then, for every $\gamma \in \Gamma_k^{\mathrm{G}}$,

$$\mathcal{S}_{J,\tau}(\gamma) \ge \det\left(I_m + \frac{2k}{(k-1)\tau^2} JJ^\top\right)^{-1/2}.$$

The repulsive Gaussian coupling attains equality. Consequently, among all jointly Gaussian couplings with standard Gaussian marginals, the repulsive Gaussian coupling minimizes the expected average RBF similarity in any fixed linear feature space $z \mapsto Jz$. The same minimizer also applies to any nonnegative weighted sum, or nonnegative weighted integral, of such RBF similarities over bandwidths τ.

Proof. We first prove the statement for squared feature separation. For deterministic vectors $y_1, \dots, y_k \in \mathbb{R}^m$, the following identity holds:

$$\sum_{1\le i<j\le k} \|y_i - y_j\|_2^2 = k\sum_{i=1}^k \|y_i\|_2^2 - \left\|\sum_{i=1}^k y_i\right\|_2^2.$$

Applying this identity with $y_i = Jz_i$ and taking expectation under γ, we get

$$\sum_{1\le i<j\le k} \mathbb{E}_\gamma \|Jz_i - Jz_j\|_2^2 = k\sum_{i=1}^k \mathbb{E}_\gamma \|Jz_i\|_2^2 - \mathbb{E}_\gamma \left\|\sum_{i=1}^k Jz_i\right\|_2^2.$$

Since each marginal satisfies $z_i \sim N(0, I_d)$,

$$\mathbb{E}_\gamma \|Jz_i\|_2^2 = \operatorname{tr}(JJ^\top) = \|J\|_F^2.$$

Therefore

$$\mathcal{D}_J(\gamma) = \frac{2k}{k-1}\|J\|_F^2 - \frac{2}{k(k-1)}\mathbb{E}_\gamma \left\|\sum_{i=1}^k Jz_i\right\|_2^2.$$

The second term is nonnegative, so

$$\mathcal{D}_J(\gamma) \le \frac{2k}{k-1}\|J\|_F^2.$$

Equality holds if and only if

$$\mathbb{E}_\gamma \left\|\sum_{i=1}^k Jz_i\right\|_2^2 = 0,$$

which is equivalent to

$$\sum_{i=1}^k Jz_i = 0 \qquad \gamma\text{-a.s.}$$

For the repulsive Gaussian coupling,

$$\sum_{i=1}^k z_i = \sqrt{\frac{k}{k-1}}\sum_{i=1}^k (u_i - \bar{u}) = 0 \qquad \text{a.s.}$$

Hence $\sum_i Jz_i = 0$ almost surely, and the repulsive Gaussian coupling attains the upper bound.

We now prove the RBF similarity claim. Fix $\gamma \in \Gamma_k^{\mathrm{G}}$. For each pair $i < j$, define

$$V_{ij} := \operatorname{Cov}(y_i - y_j) = \operatorname{Cov}(Jz_i - Jz_j).$$

Since γ is jointly Gaussian, $y_i - y_j$ is a centered Gaussian vector in $\mathbb{R}^m$ with covariance V_{ij}. Hence

$$\mathbb{E}_\gamma \exp\left(-\frac{\|y_i - y_j\|_2^2}{2\tau^2}\right) = \det\left(I_m + \frac{1}{\tau^2}V_{ij}\right)^{-1/2}.$$

Define

$$g_\tau(V) := \det\left(I_m + \frac{1}{\tau^2}V\right)^{-1/2}$$

for positive semidefinite $V \in \mathbb{R}^{m\times m}$. The function g_τ is convex on the positive semidefinite cone. Indeed,

$$V \mapsto -\frac{1}{2}\log\det\left(I_m + \frac{1}{\tau^2}V\right)$$

is convex, and the exponential function is convex and nondecreasing. Also, g_τ is decreasing in the positive semidefinite order: if $0 \preceq A \preceq B$, then

$$\det\left(I_m + \frac{1}{\tau^2}A\right) \le \det\left(I_m + \frac{1}{\tau^2}B\right),$$

and therefore $g_\tau(A) \ge g_\tau(B)$.

Let

$$\bar{V} := \frac{2}{k(k-1)}\sum_{1\le i<j\le k} V_{ij}.$$

Using the matrix form of the pairwise-difference identity,

$$\sum_{1\le i<j\le k}(y_i-y_j)(y_i-y_j)^\top = k\sum_{i=1}^{k} y_i y_i^\top - \left(\sum_{i=1}^{k} y_i\right)\left(\sum_{i=1}^{k} y_i\right)^\top.$$

Taking expectations gives

$$\sum_{1\le i<j\le k} V_{ij} = k\sum_{i=1}^{k}\mathrm{Cov}(y_i) - \mathrm{Cov}\left(\sum_{i=1}^{k} y_i\right).$$

Since $\mathrm{Cov}(y_i) = JJ^\top$ for every i,

$$\bar{V} = \frac{2k}{k-1}JJ^\top - \frac{2}{k(k-1)}\,\mathrm{Cov}\left(\sum_{i=1}^{k} y_i\right).$$

Thus

$$\bar{V} \preceq V_\star, \qquad V_\star := \frac{2k}{k-1}JJ^\top.$$

By Jensen's inequality and the convexity of g_τ,

$$\mathcal{S}_{J,\tau}(\gamma) = \frac{2}{k(k-1)}\sum_{1\le i<j\le k} g_\tau(V_{ij}) \ge g_\tau(\bar{V}).$$

Since g_τ is decreasing in the positive semidefinite order and $\bar{V} \preceq V_\star$,

$$g_\tau(\bar{V}) \ge g_\tau(V_\star).$$

Combining the two inequalities yields

$$\mathcal{S}_{J,\tau}(\gamma) \ge g_\tau(V_\star) = \det\left(I_m + \frac{2k}{(k-1)\tau^2}JJ^\top\right)^{-1/2}.$$

It remains to check that the repulsive Gaussian coupling attains equality. Under this coupling,

$$\mathrm{Cov}(z_i, z_j) = -\frac{1}{k-1}I_d \qquad \text{for } i \ne j.$$

Therefore, for every pair $i < j$,

$$\mathrm{Cov}(z_i - z_j) = I_d + I_d - \mathrm{Cov}(z_i, z_j) - \mathrm{Cov}(z_j, z_i) = \frac{2k}{k-1}I_d.$$

Hence

$$V_{ij} = \mathrm{Cov}(Jz_i - Jz_j) = \frac{2k}{k-1}JJ^\top = V_\star$$

for every $i < j$. Thus

$$\mathcal{S}_{J,\tau}(\gamma) = g_\tau(V_\star) = \det\left(I_m + \frac{2k}{(k-1)\tau^2}JJ^\top\right)^{-1/2}.$$

So the lower bound is tight, and it is attained by the repulsive Gaussian coupling.

Finally, since the same repulsive Gaussian coupling attains the lower bound for every bandwidth $\tau > 0$, any nonnegative weighted sum or nonnegative weighted integral over such RBF similarities is minimized by the same coupling. This completes the proof. □

A.4 Proof of Proposition 3.2

Proof. Let $Z \sim \mathcal{N}(0, I_d)$ be an independent standard Gaussian noise. Since $Z_i \sim \mathcal{N}(0, I_d)$ for each i, we have

$$Z_i \overset{d}{=} Z.$$

Because G is deterministic, the map $z \mapsto G(p, z)$ is fixed once the prompt p is fixed. Hence applying this same map to two random variables with the same distribution gives random variables with the same distribution:

$$X_i = G(p, Z_i) \overset{d}{=} G(p, Z).$$

Thus each X_i has the same marginal distribution as the output generated from an independent standard Gaussian initial noise.

Now let q be any integrable single-image score. Since $X_i \overset{d}{=} G(p, Z)$, we have

$$\mathbb{E}\big[q(p, X_i)\big] = \mathbb{E}_{Z\sim\mathcal{N}(0,I_d)}\big[q(p, G(p, Z))\big], \qquad i = 1, \dots, K.$$

By linearity of expectation,

$$\begin{aligned}
\mathbb{E}\left[K^{-1}\sum_{i=1}^{K} q(p, X_i)\right] &= K^{-1}\sum_{i=1}^{K}\mathbb{E}\big[q(p, X_i)\big] \\
&= K^{-1}\sum_{i=1}^{K}\mathbb{E}_{Z\sim\mathcal{N}(0,I_d)}\big[q(p, G(p, Z))\big] \\
&= \mathbb{E}_{Z\sim\mathcal{N}(0,I_d)}\big[q(p, G(p, Z))\big].
\end{aligned}$$

This proves the claim. □

A.5 Proof of Theorem 3.3

We prove Theorem 3.3 from the main paper.

Proof. We first prove the forward direction. Let

$$u_1, \dots, u_r \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_d),$$

and let $A \in \mathbb{R}^{M\times r}$ have unit-norm rows:

$$\sum_{\ell=1}^{r} A_{m\ell}^2 = 1, \qquad m = 1, \dots, M.$$

Define

$$z_m = \sum_{\ell=1}^{r} A_{m\ell} u_\ell, \qquad m = 1, \dots, M.$$

Since each z_m is a linear combination of independent Gaussian random vectors, $(z_1, \dots, z_M)$ is jointly Gaussian. Also,

$$\mathbb{E}[z_m] = 0.$$

For the marginal covariance,

$$\begin{aligned}
\mathrm{Cov}(z_m) &= \mathrm{Cov}\left(\sum_{\ell=1}^{r} A_{m\ell} u_\ell\right) \\
&= \sum_{\ell=1}^{r} A_{m\ell}^2 \,\mathrm{Cov}(u_\ell) \\
&= \left(\sum_{\ell=1}^{r} A_{m\ell}^2\right) I_d = I_d.
\end{aligned}$$

Hence

$$z_m \sim \mathcal{N}(0, I_d), \qquad m = 1, \dots, M.$$

Thus, the construction gives a valid coupling of M standard Gaussian noise vectors.

For the cross-covariance, using the independence of $u_1, \dots, u_r$,

$$\begin{aligned}\mathrm{Cov}(z_i, z_j) &= \mathrm{Cov}\left(\sum_{\ell=1}^{r} A_{i\ell} u_\ell, \sum_{s=1}^{r} A_{js} u_s\right) \\ &= \sum_{\ell=1}^{r} A_{i\ell} A_{j\ell} I_d \\ &= \left(\sum_{\ell=1}^{r} A_{i\ell} A_{j\ell}\right) I_d.\end{aligned}$$

Equivalently,

$$\mathrm{Cov}(z_i, z_j) = (AA^\top)_{ij} I_d.$$

We now prove the converse. Suppose $(z_1, \dots, z_M)$ is a centered jointly Gaussian coupling such that each $z_m \sim \mathcal{N}(0, I_d)$ and

$$\mathrm{Cov}(z_i, z_j) = \rho_{ij} I_d.$$

Let $R = (\rho_{ij})_{i,j=1}^{M}$. Since R is the sample-level covariance matrix of a jointly Gaussian vector, R is positive semidefinite. Also, because each marginal covariance is I_d,

$$R_{ii} = \rho_{ii} = 1, \qquad i = 1, \dots, M.$$

Assume $\mathrm{rank}(R) \le r$. Let $s = \mathrm{rank}(R)$. Since R is positive semidefinite, there exists $B \in \mathbb{R}^{M \times s}$ such that

$$R = BB^\top.$$

If $s < r$, pad B with $r - s$ zero columns and define

$$A = [B \;\; 0] \in \mathbb{R}^{M \times r}.$$

If $s = r$, set $A = B$. Then

$$AA^\top = R.$$

Moreover, for every row,

$$\|A_{i,:}\|_2^2 = (AA^\top)_{ii} = R_{ii} = 1.$$

Thus A has unit-norm rows.

Now, construct new Gaussian noise vectors by drawing

$$u_1, \dots, u_r \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_d)$$

and setting

$$\tilde{z}_m = \sum_{\ell=1}^{r} A_{m\ell} u_\ell.$$

By the forward direction,

$$\mathrm{Cov}(\tilde{z}_i, \tilde{z}_j) = (AA^\top)_{ij} I_d = R_{ij} I_d = \rho_{ij} I_d.$$

The constructed vector $(\tilde{z}_1, \dots, \tilde{z}_M)$ is centered Gaussian and has the same block covariance matrix as $(z_1, \dots, z_M)$. Since centered Gaussian laws are determined by their covariance matrices,

$$(\tilde{z}_1, \dots, \tilde{z}_M) \overset{d}{=} (z_1, \dots, z_M).$$

This proves the converse and completes the proof. □

A.6 First-order effect of a coupling

Theorem A.3 (First-order effect of a Gaussian coupling)**.** *Fix a prompt p, a gallery size K, and a smooth gallery-level objective*

$$H_p : \mathbb{R}^{Kd} \to \mathbb{R}, \qquad H_p(z_1, \dots, z_K).$$

Let $R \in \mathbb{R}^{K\times K}$ be a positive semidefinite correlation matrix with $R_{ii} = 1$, and let

$$B := R - I_K.$$

For $t \in [0,1]$, define

$$R_t := I_K + tB = (1-t)I_K + tR,$$

and let

$$(Z_1^{(t)}, \dots, Z_K^{(t)}) \sim \mathcal{N}(0, R_t \otimes I_d).$$

Define

$$\Psi(t) := \mathbb{E}\left[H_p(Z_1^{(t)}, \dots, Z_K^{(t)})\right].$$

For $1 \le i < j \le K$, write

$$\mathcal{D}_{ij} H_p(z_1, \dots, z_K) := \operatorname{tr}\big(\nabla_{ij}^2 H_p(z_1, \dots, z_K)\big) = \sum_{\ell=1}^{d} \frac{\partial^2 H_p}{\partial z_{i,\ell} \partial z_{j,\ell}}.$$

Assume the derivatives below are integrable. Then

$$\Psi(1) - \Psi(0) = \int_0^1 \sum_{1\le i<j\le K} B_{ij}\, \mathbb{E}_t\Big[\mathcal{D}_{ij} H_p(Z_1^{(t)}, \dots, Z_K^{(t)})\Big]\, dt,$$

where $\mathbb{E}_t$ denotes expectation under the Gaussian law with covariance $R_t \otimes I_d$. Consequently, the first-order expansion around the independent coupling is

$$\begin{aligned}\mathbb{E}_R[H_p(Z_1, \dots, Z_K)] = \mathbb{E}_{\text{iid}}[H_p(Z_1, \dots, Z_K)] \\ + \sum_{1\le i<j\le K} B_{ij}\, \mathbb{E}_{\text{iid}}[\mathcal{D}_{ij} H_p(Z_1, \dots, Z_K)] + \operatorname{Rem}(B),\end{aligned}$$

where

$$\begin{aligned}\operatorname{Rem}(B) = \int_0^1 (1-t) \sum_{1\le i<j\le K} \sum_{1\le k<\ell\le K} B_{ij} B_{k\ell} \\ \cdot\, \mathbb{E}_t\Big[\mathcal{D}_{k\ell}\mathcal{D}_{ij} H_p(Z_1^{(t)}, \dots, Z_K^{(t)})\Big]\, dt.\end{aligned}$$

In particular, if

$$\Big|\mathbb{E}_t\Big[\mathcal{D}_{k\ell}\mathcal{D}_{ij} H_p(Z_1^{(t)}, \dots, Z_K^{(t)})\Big]\Big| \le M$$

for all $t \in [0,1]$ and all index pairs, then

$$|\operatorname{Rem}(B)| \le \frac{M}{2}\left(\sum_{1\le i<j\le K} |B_{ij}|\right)^2.$$

Proof. Let

$$\Sigma_t := R_t \otimes I_d.$$

Since $R_t = (1-t)I_K + tR$, the matrix R_t is positive semidefinite for all $t \in [0,1]$. Hence Σ_t is a valid covariance matrix.

We use the standard Gaussian differentiation identity: if $X_t \sim \mathcal{N}(0, \Sigma_t)$ and f is smooth enough, then

$$\frac{d}{dt}\mathbb{E}[f(X_t)] = \frac{1}{2}\sum_{a,b} \frac{d(\Sigma_t)_{ab}}{dt}\mathbb{E}\left[\frac{\partial^2 f(X_t)}{\partial x_a \partial x_b}\right].$$

Apply this identity to

$$f(z_1, \dots, z_K) = H_p(z_1, \dots, z_K).$$

For coordinates $z_{i,r}$ and $z_{j,s}$,

$$\frac{d}{dt} \operatorname{Cov}(Z_{i,r}^{(t)}, Z_{j,s}^{(t)}) = B_{ij} \mathbf{1}\{r = s\}.$$

The diagonal terms vanish because $B_{ii} = 0$. For every pair $i < j$, the off-diagonal covariance appears twice in the full covariance matrix, once as (i, j) and once as (j, i). These two terms cancel the factor $1/2$. Therefore

$$\Psi'(t) = \sum_{1 \le i < j \le K} B_{ij} \mathbb{E}_t \left[\sum_{r=1}^{d} \frac{\partial^2 H_p}{\partial z_{i,r} \partial z_{j,r}} (Z_1^{(t)}, \dots, Z_K^{(t)}) \right].$$

That is,

$$\Psi'(t) = \sum_{1 \le i < j \le K} B_{ij} \mathbb{E}_t \Big[\mathcal{D}_{ij} H_p(Z_1^{(t)}, \dots, Z_K^{(t)}) \Big].$$

Integrating from $t = 0$ to $t = 1$ gives the exact interpolation identity.

For the Taylor expansion, write

$$\Psi(1) = \Psi(0) + \Psi'(0) + \int_0^1 (1 - t) \Psi''(t)\, dt.$$

The expression for $\Psi'(0)$ gives the first-order term under independent noise. Applying the same Gaussian differentiation identity once more to $\mathcal{D}_{ij} H_p$ gives

$$\Psi''(t) = \sum_{1 \le i < j \le K} \sum_{1 \le k < \ell \le K} B_{ij} B_{k\ell} \mathbb{E}_t \Big[\mathcal{D}_{k\ell} \mathcal{D}_{ij} H_p(Z_1^{(t)}, \dots, Z_K^{(t)}) \Big].$$

Substituting this into Taylor's formula gives the stated remainder. The bound follows by taking absolute values and using the assumed uniform bound M. □

A.7 Exact Nonlinear Feature-Distance Characterization

The local-linear interpretation can be complemented by an exact characterization for nonlinear generators. Fix a prompt and let $F : \mathbb{R}^d \to \mathbb{R}^m$ denote the generator composed with a chosen feature map, with $\mathbb{E}\|F(Z)\|_2^2 < \infty$ for $Z \sim \mathcal{N}(0, I_d)$. This covers image pixels and normalized image features without requiring the generator to be locally affine.

Proposition A.4 (Squared feature distance under Gaussian coupling)**.** *Let (X_c, Y_c) be jointly Gaussian with standard-Gaussian marginals and $\operatorname{Cov}(X_c, Y_c) = cI_d$, where $c \in [-1, 1]$. Expand F in the orthonormal multivariate Hermite basis as*

$$F(z) = \sum_{\alpha \in \mathbb{N}_0^d} f_\alpha h_\alpha(z), \qquad A_r = \sum_{|\alpha| = r} \|f_\alpha\|_2^2.$$

Then the expected squared feature distance satisfies

$$D_F(c) := \mathbb{E}\|F(X_c) - F(Y_c)\|_2^2 = 2 \sum_{r \ge 1} (1 - c^r) A_r. \tag{4}$$

In particular, $D_F'(0) = -2A_1$. For $a \in (0, 1]$,

$$D_F(-a) - D_F(0) = 2 \sum_{q \ge 1} a^{2q-1} A_{2q-1} - 2 \sum_{q \ge 1} a^{2q} A_{2q}. \tag{5}$$

Thus finite negative correlation improves expected squared feature distance exactly when the weighted odd-order energy exceeds the weighted even-order energy.

Proof. The Gaussian Hermite cross-moment identity is

$$\mathbb{E}[h_\alpha(X_c)h_\beta(Y_c)] = \mathbf{1}\{\alpha = \beta\}c^{|\alpha|}.$$

To see this, expand both sides of

$$\mathbb{E}\Big[e^{t^\top X_c - \|t\|^2/2} e^{s^\top Y_c - \|s\|^2/2}\Big] = e^{c\, t^\top s}$$

in the generating functions of the Hermite polynomials and compare coefficients. Applying the cross-moment identity to finite Hermite expansions and summing over output components gives

$$\mathbb{E}\langle F(X_c), F(Y_c)\rangle = \sum_{r\geq 0} c^r A_r.$$

The same formula holds for square-integrable F by L^2 approximation and Cauchy–Schwarz, including the degenerate endpoints $c = \pm 1$. Parseval's identity gives $\mathbb{E}\|F(Z)\|^2 = \sum_{r\geq 0} A_r < \infty$. Expanding the squared difference proves (4); its constant term cancels. The terms of degree at least two satisfy $|\sum_{r\geq 2} c^r A_r| \leq c^2 \sum_{r\geq 2} A_r$, so differentiation at zero gives $D_F'(0) = -2A_1$. Separating the odd and even powers at $c = -a$ proves (5). □

Connection to the generator. First-order Hermites satisfy $h_{e_j}(z) = z_j$, so $A_1 = \sum_j \|\mathbb{E}[F(Z)Z_j]\|^2$. If F has weak derivatives satisfying Gaussian integration by parts, then $A_1 = \|\mathbb{E}[J_F(Z)]\|_F^2$. Whenever $A_1 > 0$, a sufficiently small move from independence toward negative correlation therefore increases expected squared feature distance. For a linear map $F(z) = b + Jz$, all higher-order energies vanish and (4) reduces exactly to $2(1-c)\|J\|_F^2$.

Finite-strength behavior captures a broader range of responses. For $F(z) = z$ in one dimension, $D_F(c) = 2(1-c)$; for $F(z) = z^2$, $D_F(c) = 4(1-c^2)$. Negative correlation increases the former and decreases the latter relative to independence. The odd/even condition consequently explains why the measured response can depend on the generator, prompt, and representation. For unit-normalized features, mean cosine similarity equals $1 - D_F(c)/2$, so the same characterization applies to that pairwise similarity. Vendi and best-of-K objectives require separate analysis.

Gallery size. Every pair in the strength-s equicorrelated gallery has $c = -s/(K-1)$, so (4) also gives its expected average pairwise squared feature distance. At maximal repulsion,

$$\left|D_F\left(-\frac{1}{K-1}\right) - D_F(0)\right| \leq \frac{2}{K-1}\sum_{r\geq 1} A_r,$$

and the change is $2A_1/(K-1) + O((K-1)^{-2})$ for a fixed F as K grows. This gives a quantitative reason to expect smaller average pairwise effects in larger galleries, consistent with the experiments in Appendix D.5.

B Details and additional experiments

The image experiments in this appendix used eight NVIDIA L40S GPUs. Additional experiments in Appendix D used NVIDIA GH200 GPUs; timings are compared within each hardware and sampling configuration. Section B.1.1 provides the experimental configurations. Section B.1.5 describes how amortized coupling optimization is applied in this setting.

B.1 Diversity generation

B.1.1 Experimental configurations

We evaluate diverse text-to-image generation on Stable Diffusion 1.5 (SD1.5), Stable Diffusion XL (SDXL), and Stable Diffusion 3 (SD3). For each prompt, we generate a gallery of $K = 3$ images and compare four methods: i.i.d. Gaussian initial noise, i.i.d. noise followed by Contrastive Noise Optimization (Kim et al., 2026) (CNO), repulsively coupled Gaussian noise, and repulsively coupled Gaussian noise followed by CNO. Before any CNO updates, coupled initialization preserves

Table 6: CNO and sampling hyperparameters for diverse generation.

Model	CFG scale	CNO steps	Learning rate	Pooling window	Sampling steps
SD1.5	6.0	3	0.01	16	50
SDXL	6.0	3	0.01	16	50
SD3	7.0	3	0.001	32	28

Table 7: Compute cost for diverse generation. Time is measured per prompt for one $K = 3$ gallery. Relative time is normalized by i.i.d. sampling for each model.

Model	Method	Time / prompt	Relative time
SD1.5	i.i.d.	8.42s	$1.00\times$
	i.i.d. + CNO	8.92s	$1.06\times$
	coupled	8.42s	$1.00\times$
	coupled + CNO	8.92s	$1.06\times$
SDXL	i.i.d.	18.27s	$1.00\times$
	i.i.d. + CNO	19.06s	$1.04\times$
	coupled	18.27s	$1.00\times$
	coupled + CNO	19.06s	$1.04\times$
SD3	i.i.d.	64.73s	$1.00\times$
	i.i.d. + CNO	71.88s	$1.11\times$
	coupled	64.73s	$1.00\times$
	coupled + CNO	71.88s	$1.11\times$

each noise's standard-Gaussian marginal and changes only the dependence within the gallery; the subsequent direct noise optimization need not preserve this marginal.

For CNO, we follow the model-specific hyperparameters of the original CNO paper, summarized in Table 6. Unlike the original CNO evaluation, which uses a batch size 5 for SD1.5 and SD3, and batch size 3 for SDXL, we use a fixed gallery size $K = 3$ for all models to keep the comparison with coupled-noise initialization uniform.

We evaluate diversity using average pairwise L2 distance, MSS, and Vendi score. L2 is computed directly in pixel space as the mean pairwise per-pixel squared distance. MSS and Vendi are computed from SSCD image features Pizzi et al. (2022): MSS is the mean off-diagonal cosine similarity within each gallery, while Vendi is computed from the eigenvalue entropy of the SSCD feature similarity matrix.

B.1.2 Compute Cost Comparison

We compare the generation cost of the four sampling settings in Table 7. The reported time is the average wall-clock time to generate one $K = 3$ gallery for a single prompt. Since repulsive coupling only applies a linear transformation to the initial Gaussian noises, its cost is effectively the same as i.i.d. sampling. CNO introduces an optimization overhead because it optimizes the noises before the final diffusion sampling pass; in our setting, this overhead is modest since only 3 optimization steps are used.

B.1.3 Ablation study on the Gallery Size

We additionally study how the effect of repulsive coupling changes with the gallery size K. We run an SD1.5 ablation on 500 COCO prompts and compare i.i.d. Gaussian noise with repulsively coupled Gaussian noise for $K \in \{2, 3, 4, 5\}$. Results are shown in Table 8. Repulsive coupling improves all three diversity metrics for every value of K: it increases pixel-space L2, decreases MSS, and increases Vendi. The L2 gain is largest for $K = 2$ and decreases as K grows, from a 49.7% relative improvement at $K = 2$ to a 13.0% improvement at $K = 5$. Even for larger galleries,

Table 8: Ablation on gallery size for SD1.5 diverse generation over 500 COCO prompts. We compare i.i.d. Gaussian noise and repulsively coupled Gaussian noise without CNO.

K	Method	L2 ↑	MSS ↓	Vendi ↑
2	i.i.d.	0.1276	0.1550	1.9679
	coupled	**0.1911**	**0.1414**	**1.9721**
3	i.i.d.	0.1267	0.1504	2.9136
	coupled	**0.1567**	**0.1370**	**2.9255**
4	i.i.d.	0.1260	0.1502	3.8372
	coupled	**0.1458**	**0.1426**	**3.8519**
5	i.i.d.	0.1257	0.1497	4.7419
	coupled	**0.1421**	**0.1439**	**4.7592**

Table 9: Ablation on equicorrelation strength for SD1.5 diverse generation over 100 COCO prompts with gallery size $K = 3$. We vary the off-diagonal correlation c in the Gaussian coupling.

Correlation c	L2 ↑	MSS ↓	Vendi ↑
0	0.1283	0.1542	2.9066
-0.125	0.1353	0.1429	2.9176
-0.25	0.1428	0.1447	2.9204
-0.375	0.1497	**0.1332**	**2.9300**
-0.5	**0.1574**	0.1346	2.9281

the coupled initialization continues to improve SSCD-based diversity, as reflected by lower MSS and higher Vendi across all settings.

B.1.4 Ablation study on the correlation strength

We additionally study how the strength of the equicorrelated coupling affects diversity for a fixed gallery size. We run an SD1.5 ablation on 100 COCO prompts with gallery size $K = 3$, and vary the off-diagonal correlation c in the equicorrelated Gaussian coupling from independent sampling, $c = 0$, to the maximally repulsive valid value, $c = -1/(K-1) = -1/2$. We evaluate

$$c \in \{0, -0.125, -0.25, -0.375, -0.5\}.$$

Results are shown in Table 9. Increasing the amount of negative correlation consistently improves pixel-space L2, from 0.1283 at $c = 0$ to 0.1574 at $c = -0.5$. SSCD-based diversity also improves: MSS decreases and Vendi increases for all negative-correlation settings compared with independent sampling. The best MSS and Vendi occur at $c = -0.375$, while the $c = -0.5$ result is extremely close. The results show that stronger repulsion generally improves diversity, with the theoretical boundary $c = -1/(K-1)$ performing best or near-best across the reported metrics.

B.1.5 Learning the optimal coupling

We also experimented with learned couplings optimized directly for image-space diversity. For this experiment, we parameterize a batch coupling matrix $A \in \mathbb{R}^{K\times K}$ following 3.3, sample independent basis noises $\epsilon_1, \ldots, \epsilon_K$, and form coupled initial noises

$$z_i = \sum_{j=1}^{K} A_{ij}\epsilon_j.$$

Each row of A is constrained to have unit norm, so every z_i remains marginally standard Gaussian. We initialize A either as the identity coupling, and update it by projected gradient steps on the row-normalized sphere.

The optimization uses Stable Diffusion 1.5 and treats the final generated images as the forward signal. Since differentiating through the full denoising trajectory is expensive, we use a skip-gradient

Table 10: Brightness-clustered generation.

Model	Method	Gap ↑	Contrast ↑	Margin ↑
SD1.5	i.i.d.	0.0448	0.0673	0.0336
	equicorr.	0.0494	0.0746	0.0384
	learned	**0.0702**	**0.0873**	**0.0462**
SDXL	i.i.d.	0.0355	0.0524	0.0256
	equicorr.	0.0364	0.0543	0.0271
	learned	**0.0461**	**0.0604**	**0.0286**
SD3	i.i.d.	0.0478	0.0715	0.0353
	equicorr.	0.0508	0.0761	0.0380
	learned	**0.0652**	**0.0845**	**0.0412**

estimator: we first run the deterministic sampler from the coupled initial noises to obtain final latents z_i^{final}, then use

$$\tilde{z}_i = \text{stopgrad}(z_i^{\text{final}}) + z_i - \text{stopgrad}(z_i)$$

before VAE decoding. We tested gallery sizes $K = 4$ and $K = 5$ using pairwise SSIM and LPIPS as the optimization objective. In all cases, the learned coupling quickly approaches the analytic repulsive structure. For example, for $K = 4$, the best learned sample-level correlation matrix using SSIM is

$$C = AA^\top = \begin{pmatrix} 1.000 & -0.339 & -0.332 & -0.333 \\ -0.339 & 1.000 & -0.335 & -0.335 \\ -0.332 & -0.335 & 1.000 & -0.326 \\ -0.333 & -0.335 & -0.326 & 1.000 \end{pmatrix}.$$

This is very close to the repulsive coupling, whose off-diagonal entries are $-1/(K-1) = -1/3$. This provides an empirical check that, when the learning objective directly rewards image-space diversity, amortized optimization recovers essentially the same repulsive dependence structure as the analytic construction.

B.2 Brightness-Clustered Generation

Noise coupling can also be learned to induce structured, user-specified relationships among generated samples. We consider a brightness-clustered generation task in which a four-image gallery is encouraged to split into two brighter and two darker samples. We follow the amortized optimization framework (2) and learn a coupling matrix A that maps independent Gaussian basis noises $\epsilon_1, \ldots, \epsilon_4$ into coupled initial noises $z_i = \sum_{j=1}^4 A_{ij}\epsilon_j$. Given generated images $x_i = G(p, z_i)$, let b_i denote the average brightness of image i, computed from the value channel. We optimize A using the objective

$$\left| \frac{b_1 + b_2}{2} - \frac{b_3 + b_4}{2} \right| + \lambda \frac{|b_1 - b_2| + |b_3 - b_4|}{2}, \tag{6}$$

where the first term encourages a bright–dark split between the two image pairs, and the second term weakly encourages non-collapse within each pair. After learning A, generation is performed without any per-prompt optimization: we sample independent basis noises, apply the learned coupling, and run the pretrained diffusion sampler normally.

In our experiment, we choose $\lambda = 0.35$. The coupling matrix A is learned on 600 COCO prompts using SD1.5 and then applied amortized, without per-prompt optimization, to held-out prompts across SD1.5, SDXL, and SD3. We compare independent noise, repulsive equicorrelated noise with correlation $-1/3$, and the learned coupling.

We report three metrics: the pair gap $|(b_1 + b_2)/2 - (b_3 + b_4)/2|$ the contrast objective from Eq. (6), and sorted margin $b_{(3)} - b_{(2)}$, where $b_{(1)} \le b_{(2)} \le b_{(3)} \le b_{(4)}$ are the sorted brightness values. This metric measures the gap between the darker and brighter halves of the gallery.

The learned brightness coupling produces a clear block structure in the noise correlations. The two samples intended to form the bright cluster are positively correlated, and the two samples intended

to form the dark cluster are also positively correlated, while samples across the two clusters are strongly negatively correlated. In one learned solution, the induced correlation matrix $C = AA^\top$ is

$$C = \begin{pmatrix} 1.00 & 0.47 & -0.76 & -0.74 \\ 0.47 & 1.00 & -0.75 & -0.77 \\ -0.76 & -0.75 & 1.00 & 0.55 \\ -0.74 & -0.77 & 0.55 & 1.00 \end{pmatrix}.$$

This structure matches the target relationship: samples (1,2) share noise components and therefore tend to move together in brightness, samples (3,4) do the same, and the two groups are coupled in opposite directions to encourage a bright–dark split across the gallery.

Table 10 shows a cleaner brightness split than independent and repulsive noise for the reported learned coupling, including transfer from SD1.5 to SDXL and SD3. This illustrates how coupling optimization can express a specified gallery relationship and reuse it without additional test-time optimization. Figure 4 provides examples.

B.3 Background generation with fixed object

B.3.1 Experiment Details

We evaluate fixed-object background generation on 100 LSUN bedroom images, using YOLOv8 segmentation masks for the bed foreground. Failed detections or generations are excluded, leaving 92 valid images for the Stable Diffusion 1.5 experiment. All metrics are computed with the same foreground mask used for generation. Boundary-TV is computed in a 5-pixel dilation-minus-erosion band around the mask boundary.

We use `runwayml/stable-diffusion-v1-5` with prompt "a bedroom", DDIM inversion/sampling with 20 steps, and classifier-free guidance scale 1. We generate $K = 8$ samples per image. Our method uses DDIM-inverted foreground noise, repulsively coupled background noise, Adam optimization with learning rate 0.01, and checkpointed optimization steps at 10, 20, 30, 50, and 100 iterations. The optimization uses the same masked foreground MSE objective and resets the background latent after each Adam step. SDEdit is implemented with a DDIM scheduler, 20 denoising steps, guidance scale 1, strengths $\{0.2, 0.4, 0.6, 0.8\}$, and foreground pasting before evaluation. In our experiments, strengths of 0.2, 0.4, and 0.6 produced very low diversity, so the table reports results at strength 0.8. Stable Diffusion inpainting uses 20 steps, guidance scale 1, strength 1, and the inverted foreground mask as the editable region.

B.3.2 Effect of optimization steps

Table 11 reports how the number of optimization steps affects each evaluation axis for SD1.5 fixed-object background generation. The results show a controllable trade-off between background diversity and foreground fidelity, governed by a single hyperparameter.

As the number of optimization steps increases from 10 to 100, foreground preservation improves monotonically: RMSE-F decreases from 0.086 to 0.048, indicating that additional optimization progressively pulls the masked region closer to the input object. At the same time, background diversity decreases: LPIPS-B drops from 0.453 to 0.373 and L2-B from 0.045 to 0.029. This is consistent with the structure of the objective in (3), which only constrains the foreground; longer optimization tightens this constraint and indirectly reduces variation in the surrounding pixels through the coupling between foreground and background in the denoising trajectory.

Boundary-TV remains essentially flat across all settings 0.012–0.013, indicating smooth object–background transitions across the tested optimization budgets. Global naturalness, measured by BRISQUE, degrades mildly with more steps (from 17.52 to 23.02).

Across all optimization budgets, our method achieves substantially lower BRISQUE and Boundary-TV than both SDEdit and SD inpaint, indicating that the coupled-noise approach consistently produces more natural images that blend smoothly with the preserved object. With sufficient optimization steps, foreground fidelity also surpasses that of SD inpaint. SDEdit retains the strongest foreground preservation by construction, but it is dominated on every other axis we measure (background diversity, boundary smoothness, and global naturalness).

Table 11: Effect of optimization steps for SD1.5 fixed-object background generation. We report mean (standard deviation). LPIPS-B and L2-B measure background diversity; RMSE-F measures foreground preservation; Boundary-TV measures local roughness near the object boundary; BRISQUE measures global image naturalness.

Steps	LPIPS-B ↑	L2-B ↑	RMSE-F ↓	Boundary-TV ↓	BRISQUE ↓
10	0.453 (0.160)	0.045 (0.013)	0.086 (0.025)	0.012 (0.004)	17.52 (9.57)
20	0.433 (0.166)	0.040 (0.014)	0.072 (0.021)	0.012 (0.004)	18.12 (10.84)
30	0.418 (0.171)	0.037 (0.014)	0.065 (0.019)	0.012 (0.004)	19.11 (11.05)
50	0.399 (0.174)	0.033 (0.015)	0.057 (0.016)	0.012 (0.004)	20.24 (10.63)
100	0.373 (0.177)	0.029 (0.016)	0.048 (0.014)	0.013 (0.004)	23.02 (11.15)

Table 12: Unconditional fixed-object background generation. We report mean (standard deviation). LPIPS-B and L2- B measure background diversity; RMSE-F measures foreground preservation; Boundary-TV measures local roughness near the object boundary; BRISQUE measures global image naturalness.

Method	LPIPS-B ↑	L2-B ↑	RMSE-F ↓	Boundary-TV ↓	BRISQUE ↓
DDIM inv.	–	–	0.20 (0.06)	**0.015** (0.006)	32.98 (7.82)
Object-inv. + random BG	0.39 (0.16)	0.22 (0.09)	0.32 (0.08)	0.017 (0.011)	38.37 (11.00)
SDEdit	0.39 (0.12)	0.14 (0.04)	0.03 (0.01)	0.041 (0.010)	**24.45** (8.35)
Ours	**0.40** (0.15)	**0.25** (0.10)	0.05 (0.02)	0.039 (0.014)	31.46 (11.91)

B.3.3 Unconditional diffusion models

We evaluate the unconditional fixed-object task on 100 LSUN bedroom images with YOLOv8 bed masks as foreground objects. We use the pretrained pixel-space diffusion model google/ddpm-bedroom-256 and generate a gallery of $K = 4$ images for each input. Our method uses DDIM inversion to initialize the foreground noise, repulsively coupled background noise for diversity, and 100 optimization steps to improve foreground fidelity while keeping the background noise fixed. We compare against DDIM inversion, Object-inv. + random BG, and SDEdit. DDIM inversion reconstructs from the inverted noise and therefore has no meaningful background diversity. Object-inv. + random BG uses the inverted foreground noise with random background noise.

Object-inv. + random BG produces diverse backgrounds, but has poor foreground preservation, with RMSE-F 0.32. This shows that fixing only the object noise is insufficient after denoising. SDEdit-80 achieves the best foreground preservation and BRISQUE, with RMSE-F 0.03 and BRISQUE 24.45, but has lower background diversity than ours, especially in L2-B, 0.14 versus 0.25. Our method achieves the strongest background diversity among these baselines, with LPIPS-B 0.40 and L2-B 0.25, while keeping foreground RMSE much lower than Object-inv. + random BG, 0.05 versus 0.32.

C Visualizations

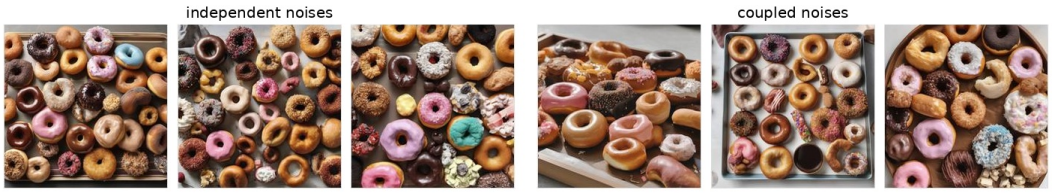

Figure 3: Independent versus coupled noises with the same prompt “Assortment of doughnuts and other snack items on a serving tray.” generated from SDXL.



Figure 4: Examples of brightness-clustered generation with Stable Diffusion 1.5. The displayed four-image galleries exhibit a two-group bright/dark structure under the learned coupling.

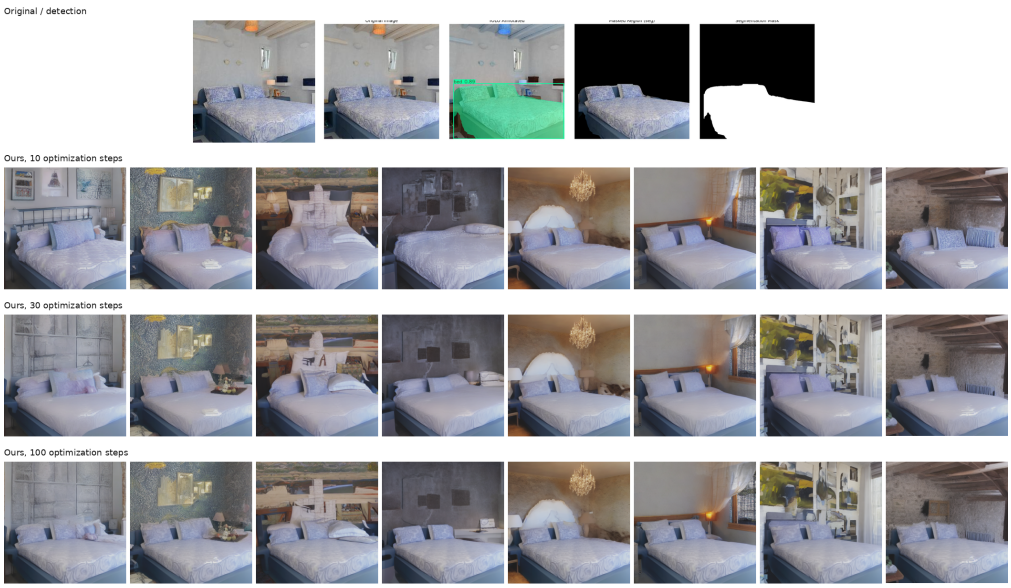

Figure 5: Fixed-object background generation on an LSUN bedroom image. The first row shows the input image, YOLOv8 bed detection, masked foreground, and segmentation mask. The remaining rows show our method after 10, 30, and 100 optimization steps. Across optimization budgets, the bed is preserved while the surrounding room layout, wall texture, lighting, and decor vary substantially. Increasing the number of optimization steps improves foreground fidelity while maintaining diverse and natural backgrounds.

Table 13: Fixed-object background generation with SD 1.5. We report mean (standard deviation). LPIPS-B and L2-B measure background diversity; RMSE-F measures foreground preservation; Boundary-TV measures local roughness near the boundary; BRISQUE measures global naturalness.

Method	LPIPS-B ↑	L2-B ↑	RMSE-F ↓	Boundary-TV ↓	BRISQUE ↓
DDIM inv.	0 (0)	0 (0)	0.09 (0.03)	0.015 (0.006)	45.40 (10.95)
Object-inv. + random BG	**0.47** (0.16)	**0.05** (0.01)	0.14 (0.03)	0.023 (0.011)	<u>18.16</u> (8.80)
SDEdit	0.31 (0.12)	0.02 (0.01)	**0.01** (0.00)	0.025 (0.007)	35.99 (9.67)
SD inpaint	0.42 (0.14)	**0.05** (0.01)	<u>0.05</u> (0.03)	<u>0.022</u> (0.007)	34.07 (7.27)
Ours	<u>0.45</u> (0.16)	**0.05** (0.01)	0.09 (0.03)	**0.012** (0.005)	**17.52** (9.57)

D Additional Experiments and Sampling Extensions

D.1 Protocols, Metrics, and Statistical Units

The additional evaluations use fixed COCO-caption subsets and matched prompts across conditions. The strength, SDXL, and composition evaluations use 300 prompts each. The compact CADS and CDF-stratification studies, alternative baselines, and SDE sampling use 100 prompts each; Qwen-Image uses 50 prompts. The initial FLUX evaluation uses 30 prompts, with an additional 70-prompt holdout. The video and large-K evaluations use 100 prompts each.

L2 is the average pairwise mean squared RGB difference on $[0, 1]$ images. LPIPS uses the AlexNet backbone (Zhang et al., 2018). Unless an encoder is specified, MSS and Vendi use SSCD features (Pizzi et al., 2022). We additionally evaluate DINOv2-base and CLIP ViT-L/14 image-feature MSS/Vendi. Lower MSS and higher Vendi indicate greater feature diversity. CLIP denotes text–image alignment, and PickScore and MUSIQ provide quality diagnostics. Diversity metrics describe different aspects of a gallery and need not rank methods identically.

Prompts are the units for paired comparisons. For example, if each method generates three images per prompt, we average the three pairwise distances within each gallery and compare the two methods' averages for that prompt. All reported 95% confidence intervals use paired t intervals. Ranges across experimental settings are labeled separately in the tables. We apply Benjamini–Hochberg correction for multiple testing within each specified comparison family. We consider mean PickScore non-inferior if the lower bound of the 95% confidence interval for its relative change is above -1%. This criterion concerns average quality and does not guarantee the quality of the best or worst image in each gallery.

D.2 Coupling with Existing Diversity Methods

Table 14 expands the main-text ranges. CNO uses SD1.5, DDIM-50, CFG 6, $K = 3$, and one, three, or five optimization steps. PG uses Euler-30, CFG 9, and $K = 4$; pixel coefficients are $10, 30, 60$ and DINO-B/8 feature coefficients are $4, 8, 16$. CADS uses 100 denoising steps, CFG 9, $K = 3$, $\tau_1 = 0.6$, $\tau_2 = 0.9$, full rescaling, and conditioning-noise scales $0.025, 0.10, 0.15, 0.25$. The 0.25 endpoint is also used in the 100-prompt complementary-dimensions example in Table 3.

These comparisons use each method's native protocol. Their purpose is to measure the effect of changing initialization; raw metric values from different step counts, CFG settings, or gallery sizes are not a common-budget ranking. The PG implementation is a current-library reproduction.

Table 14: Per-setting composition gains on 300 prompts. Setting denotes CNO optimization steps, PG coefficient, or CADS conditioning-noise scale. Intervals are relative 95% paired intervals.

Method	Setting	L2 gain	95% CI (%)	LPIPS gain	95% CI (%)
CNO	1	+20.28%	[18.01, 22.55]	+3.36%	[2.69, 4.04]
CNO	3	+21.15%	[18.88, 23.42]	+3.25%	[2.54, 3.96]
CNO	5	+20.48%	[18.20, 22.77]	+3.61%	[2.90, 4.32]
Pixel-PG	10	+15.24%	[13.50, 16.97]	+3.68%	[2.99, 4.36]
Pixel-PG	30	+15.12%	[13.40, 16.85]	+3.59%	[2.90, 4.27]
Pixel-PG	60	+14.88%	[13.18, 16.59]	+3.47%	[2.79, 4.16]
DINO-PG	4	+15.12%	[13.39, 16.85]	+3.68%	[3.00, 4.37]
DINO-PG	8	+15.02%	[13.28, 16.76]	+3.65%	[2.96, 4.34]
DINO-PG	16	+14.98%	[13.23, 16.72]	+3.76%	[3.06, 4.45]
CADS	0.025	+21.01%	[18.62, 23.40]	+2.94%	[2.15, 3.74]
CADS	0.10	+22.72%	[20.33, 25.12]	+2.87%	[2.05, 3.69]
CADS	0.15	+23.27%	[20.79, 25.74]	+2.94%	[2.13, 3.75]
CADS	0.25	+23.76%	[21.14, 26.39]	+3.13%	[2.31, 3.95]

The 100-prompt CADS experiment makes the complementary behavior explicit. Relative to IID, CADS reduces DINO MSS from 0.6064 to 0.5394, while L2 decreases from 0.1553 to 0.1242. Coupled+CADS raises L2 to 0.1513 and LPIPS to 0.7783, while DINO MSS stays near the CADS value at 0.5407. Mean PickScore is 21.25 for CADS and 21.24 for coupled+CADS. Coupling

therefore improves the pixel/perceptual side of the combined method while retaining its feature-space diversity. It need not maximize every diversity axis in isolation.

D.3 Modern Backbones: Settings and Feature Metrics

Qwen-Image uses its native pipeline at 1024×1024, 50 steps, and true CFG 4; FLUX.1-schnell uses 1024×1024, four steps, and guidance scale 0. Both use $K = 3$ and identical settings within each comparison. The FLUX weights were loaded from the public `frankjoshua/FLUX.1-schnell` mirror. We evaluate Qwen-Image on 50 prompts and FLUX.1-schnell on 70 prompts. Table 15 reports all evaluated diversity representations, including pixel-space, perceptual, and feature-space metrics.

Coupled initialization improves diversity metrics across modern backbones, including SSCD, CLIP-image, and DINOv2 features, while the magnitude of improvement varies across representations. These results show that the effect of coupling extends beyond pixel-space diversity and depends on the feature space used for evaluation.

Table 15: Modern-backbone diversity results with relative 95% paired intervals. Qwen-Image is evaluated on 50 prompts and FLUX.1-schnell on a disjoint set of 70 prompts. MSS is better when lower, and all distance/Vendi metrics are better when higher.

Backbone	Metric	IID	Coupled	Δ(%)	95% CI (%)
Qwen-Image	L2	0.0876	0.0998	+13.91%	[7.43, 20.39]
Qwen-Image	LPIPS	0.6172	0.6523	+5.68%	[3.57, 7.78]
Qwen-Image	MSS	0.5014	0.4729	-5.69%	[-9.80, -1.58]
Qwen-Image	Vendi	2.3422	2.4062	+2.73%	[0.93, 4.54]
Qwen-Image	DINO MSS	0.8422	0.8403	-0.22%	[-2.52, 2.07]
Qwen-Image	DINO Vendi	1.4763	1.4813	+0.34%	[-2.95, 3.63]
Qwen-Image	CLIP MSS	0.9446	0.9331	-1.22%	[-1.95, -0.50]
Qwen-Image	CLIP Vendi	1.1931	1.2273	+2.87%	[1.21, 4.53]
FLUX.1-schnell	L2	0.1507	0.1916	+27.20%	[21.92, 32.49]
FLUX.1-schnell	LPIPS	0.7375	0.7817	+5.99%	[4.68, 7.30]
FLUX.1-schnell	MSS	0.2366	0.2093	-11.54%	[-17.17, -5.91]
FLUX.1-schnell	Vendi	2.8194	2.8500	+1.08%	[0.48, 1.69]
FLUX.1-schnell	DINO MSS	0.7148	0.6813	-4.68%	[-7.44, -1.92]
FLUX.1-schnell	DINO Vendi	1.8185	1.8887	+3.86%	[1.35, 6.36]
FLUX.1-schnell	CLIP MSS	0.8551	0.8433	-1.38%	[-2.98, 0.21]
FLUX.1-schnell	CLIP Vendi	1.4531	1.4870	+2.34%	[-0.22, 4.90]

D.4 Coupling Stochastic Sampling Chains

An SDE sampler (Song et al., 2021b) with a fixed discretization can be viewed as a deterministic function of all its random inputs:

$$X_i = F_p(\xi_{i,0}, \xi_{i,1}, \ldots, \xi_{i,T}),$$

where $\xi_{i,0}$ is the initialization and the remaining terms are independent stepwise Gaussian innovations. At each time t, draw a K-tuple $(\xi_{1,t}, \ldots, \xi_{K,t})$ from a marginal-preserving coupling, independently of the tuples at other times. For a fixed chain i, its entire sequence still has the original product-Gaussian law. Applying the same F_p therefore gives exactly the same single-chain output distribution as the original discretized sampler. Only the joint distribution across chains changes. For $K = 2$, one example draws fresh ε_t independently at each time and uses $(\varepsilon_t, -\varepsilon_t)$ across chains.

We evaluate initialization-only, injection-only, and both forms of coupling with SD1.5 Euler-ancestral sampling, 30 steps, CFG 6, $K = 3$, and 100 prompts. Table 16 shows benefits from either location and the largest L2/LPIPS gain when coupling both. This experiment demonstrates

that the construction can remain effective when fresh randomness enters throughout denoising. The marginal-law argument applies independently of whether a particular nonlinear model exhibits a diversity gain.

Table 16: Stochastic sampling on 100 prompts. Gains are relative to the IID Euler-ancestral sampler under otherwise identical settings.

Coupling location	L2 ↑	LPIPS ↑	L2 gain	LPIPS gain
IID	0.1307	0.7616	–	–
Initialization	0.1422	0.7812	+8.85%	+2.58%
Stepwise injections	0.1446	0.7808	+10.68%	+2.53%
Both	0.1566	0.7910	+19.83%	+3.85%

D.5 Noise–Image Relationship and Larger Galleries

The 300-prompt SD1.5 strength experiment uses DDIM-50, CFG 6, $K = 3$, and strengths $s \in \{0, 0.25, 0.5, 0.75, 1\}$. Shared base noises allow matched comparisons across strengths, with

$$z_i(s) = \sqrt{1 + \frac{s}{K-1}}\,(u_i - \bar{u}) + \sqrt{1-s}\,\bar{u}, \qquad \mathrm{Cov}(z_i, z_j) = -\frac{s}{K-1} I_d \quad (i \neq j).$$

The endpoint improves L2 by 23.77% (95% interval $[21.45, 26.09]\%$) and LPIPS by 3.85% ($[3.15, 4.55]\%$). A matching 300-prompt SDXL evaluation yields 20.64% ($[18.79, 22.49]\%$) and 3.58% ($[2.91, 4.24]\%$), respectively.

For each of 900 prompt–pair trajectories, we correlate noise distance and output distance across the five strengths, giving 4,500 matched observations. Correlations are averaged over pairs within each prompt before aggregation. Mean Pearson correlation is 0.596 for pixel L2 (95% interval $[0.547, 0.644]$) and 0.319 for LPIPS ($[0.262, 0.375]$); 84.4% and 69.8% of pair trajectories have positive correlation. Feature-space relationships are weaker. This supports the transfer of noise separation into pixel/perceptual separation for many prompts without requiring every prompt or representation to respond monotonically.

The correlation magnitude of maximal repulsion is $1/(K-1)$, suggesting that gains should become smaller for large galleries. We extend the original $K = 2, 3, 4, 5$ analysis to $K = 16, 32$ on 100 further prompts with DDIM-50 and CFG 6. Table 17 confirms the expected decay. At $K = 16$, L2, LPIPS, and SSCD Vendi still show measurable gains; at $K = 32$, the clearest remaining effect is L2. Compact candidate galleries are therefore the strongest current use case.

Table 17: Larger-gallery evaluation on 100 prompts. Intervals are relative 95% paired intervals. MSS uses the lower-is-better convention.

K	Metric	IID	Coupled	Change	95% CI (%)
16	L2	0.1262	0.1302	+3.18%	[2.21, 4.14]
16	LPIPS	0.7528	0.7579	+0.68%	[0.37, 0.98]
16	MSS	0.1526	0.1508	-1.18%	[-2.80, 0.44]
16	Vendi	13.7080	13.7625	+0.40%	[0.05, 0.74]
32	L2	0.1269	0.1287	+1.42%	[0.83, 2.01]
32	LPIPS	0.7549	0.7558	+0.12%	[-0.07, 0.31]
32	MSS	0.1526	0.1532	+0.44%	[-0.67, 1.55]
32	Vendi	24.9494	24.9235	-0.10%	[-0.44, 0.23]

D.6 Diversity and IID Sample Count

We compare coupled $K = 3$ galleries with nested IID prefixes of sizes $3, 4, 5, 6, 8$ on 100 prompts. The appropriate comparison depends on the endpoint. For maximum pairwise LPIPS, the IID mean first exceeds the coupled value at $K = 5$, with a clear paired advantage at $K = 6$; for maximum pairwise L2, IID $K = 8$ reaches a similar observed level. These correspond to roughly 1.67–$2.67\times$

the image-generation count for these particular gallery-diameter objectives, rather than a universal sampling-efficiency factor.

For average pairwise distance, IID sample count has a different interpretation:

$$\mathbb{E}\left[\frac{2}{K(K-1)}\sum_{i<j} d(X_i, X_j)\right] = \mathbb{E}[d(X_1, X_2)].$$

Its expectation does not grow with K. Coupled $K = 3$ achieves mean pairwise LPIPS 0.774348, compared with 0.754344, 0.753006, 0.750637, 0.750182, and 0.749757 for IID $K = 3, 4, 5, 6, 8$, respectively. Coupling changes the average pairwise relation itself; generating more IID images instead increases the opportunities to find extreme pairs or favorable individual candidates. Vendi is not used to infer a sample multiplier because its upper bound changes with gallery size.

D.7 Alternative Designs and Diversity Baselines

The coupling framework also includes non-Gaussian joint distributions with Gaussian marginals. For each latent coordinate, let $(r_1, \ldots, r_K)$ be a uniformly random permutation of $0, \ldots, K-1$, and draw independent $v_i \sim \mathrm{Unif}[0, 1]$. Set

$$z_i = \Phi^{-1}\left(\frac{r_i + v_i}{K}\right).$$

Repeating this independently across latent coordinates gives a CDF-stratified gallery in which each vector is standard Gaussian, while samples occupy distinct quantile strata coordinatewise. On 100 prompts with SD1.5 DDIM-50, CFG 6, and $K = 3$, repulsive/CDF-stratified coupling improves L2 by $22.64\%/26.35\%$, LPIPS by $3.41\%/2.99\%$, and SSCD MSS by $12.82\%/11.88\%$ in the diversity direction. Table 18 supplies an effective alternative within the broader framework.

Table 18: Gaussian repulsion and non-Gaussian CDF stratification on 100 prompts. Each design preserves standard-Gaussian single-sample marginals.

Initialization	L2 ↑	LPIPS ↑	MSS ↓	Vendi ↑
IID	0.1284	0.7554	0.1542	2.9067
Repulsive	0.1574	0.7811	0.1344	2.9283
CDF-stratified	0.1622	0.7779	0.1359	2.9253

We also compare coupled initialization with simple (non-coupling) alternatives, including temperature scaling, latent SLERP, and linear or cosine CFG schedules, using a separate set of 100 prompts shared across all methods. All use SD1.5 DDIM-50 and $K = 3$. Temperature multiplies the initial noise by $1.05, 1.10$, or 1.20; SLERP uses two independent endpoints and their spherical midpoint, with a linearly interpolated radius. Linear and cosine schedules increase CFG from 1 to 6 over the 50 denoising steps: $w(t) = 1 + 5t$ and $w(t) = 1 + \frac{5}{2}(1 - \cos(\pi t))$, respectively, with $t = j/49$ for $j = 0, \ldots, 49$. Other conditions use fixed CFG 6. Results are provided in Table 19. Among the tested methods and settings, coupled initialization provides the most balanced improvements: it is the only approach that improves all three diversity metrics while keeping mean PickScore essentially unchanged (-0.03%). Temperature scaling achieves larger L2 gains at the expense of LPIPS and PickScore, while linear and cosine CFG schedules improve CLIP-feature diversity but reduce L2, LPIPS, and PickScore.

D.8 Gallery Quality and Diversity Factors

Marginal preservation does not determine the distribution of the best or worst image in a gallery. We therefore evaluate the worst-of-K score, within-gallery score range, and probability that any member falls below the IID 10th-percentile threshold, using CLIP, PickScore, and MUSIQ on 300 prompts for each of SD1.5 and SDXL. These paired audits do not reveal a reliable lower-tail degradation under the declared correction, while finite-sample uncertainty remains relevant to individual gallery outcomes.

Table 19: Alternative diversity strategies on 100 prompts. Entries are relative changes from the matched IID baseline; all displayed diversity columns and PickScore are higher-is-better.

Method	L2	LPIPS	CLIP Vendi	PickScore
Coupled	+21.16%	+2.65%	+1.82%	-0.03%
Temperature 1.05	+38.44%	-8.31%	-0.17%	-0.77%
Temperature 1.10	+83.37%	-13.39%	+0.78%	-3.84%
Temperature 1.20	+113.29%	-24.58%	-13.47%	-21.01%
Latent SLERP	-24.32%	-9.29%	-1.73%	-0.07%
Linear CFG	-25.70%	-2.33%	+10.11%	-2.53%
Cosine CFG	-26.97%	-2.31%	+10.81%	-3.12%

To distinguish forms of diversity, we use frozen YOLOv8x segmentation to measure class-aware spatial layout and object-count differences, and a frozen CLIP ViT-L/14 classifier over 12 fixed style labels. On 300 prompts, coupling increases layout diversity by 3.63% (95% interval $[1.61, 5.65]\%$). Object-count and style changes are less clear. This suggests spatial arrangement as one contributor to the measured perceptual variation, while richer semantic control remains an opportunity for future coupling designs.

D.9 Modern Inpainting Comparison

The additional comparison uses the first 30 valid bed detections from a fixed ordered LSUN-bedroom subset, with identical source images, masks, prompts, and $K = 8$ across methods. This is a separate set from the original 92-case evaluation. Our SD1.5 method uses 10 or 50 foreground-optimization iterations. We compare SD1.5 Inpaint and SDXL Inpaint at their native resolutions (512 and 1024), evaluating resized outputs at 512×512. The SDXL comparison includes 20 and 60 denoising steps.

Table 20: Modern inpainting comparison on 30 paired cases. Ours uses SD1.5 and the indicated optimization budget; numbers in parentheses for SDXL indicate denoising steps. Values are absolute means, avoiding ambiguity in percentage denominators.

Method	LPIPS-B ↑	L2-B ↑	RMSE-F ↓	TV ↓	BRISQUE ↓
Ours (10 opt.)	0.4783	0.0459	0.0864	0.0203	18.78
Ours (50 opt.)	0.4238	0.0333	0.0581	0.0169	22.10
SD1.5 Inpaint	0.4471	0.0523	0.0453	0.0240	36.92
SDXL Inpaint (20)	0.3611	0.0325	0.0265	0.0239	47.33
SDXL Inpaint (60)	0.3818	0.0392	0.0274	0.0245	47.11

At 50 optimization iterations, our LPIPS-B is 17.36% higher than 20-step SDXL Inpaint, while boundary TV and BRISQUE are 29.46% and 53.30% lower, respectively, using SDXL as the denominator. SDXL's foreground RMSE is 54.47% lower when measured relative to ours. These comparisons preserve the central trade-off: greater background variation and favorable boundary/naturalness proxies versus closer foreground reconstruction.

D.10 Video Generation: Protocol and Results

We use `Wan-AI/Wan2.1-T2V-1.3B-Diffusers`, with a BF16 transformer and FP32 VAE, 30 sampling steps, guidance 5, and 33 frames at 480×832. Prompts are 100 fixed COCO captions. Each condition generates three videos per prompt, yielding 600 videos in total.

Coupling acts on the gallery dimension of complete spatiotemporal noise tensors, with IID and coupled conditions sharing the same underlying Gaussian draws for each prompt. Cross-video L2 denotes mean squared RGB difference on $[0, 1]$ pixels. L2 and AlexNet LPIPS compare corresponding frames at indices $0, 4, 8, 12, 16, 20, 24, 28, 32$, averaged over all three video pairs and nine frames within each prompt. Frame-level text CLIP uses ViT-L/14 with score $2.5 \max(\text{cosine}, 0)$, averaged

over all 33 frames and three videos. Adjacent-frame LPIPS and CLIP cosine are temporal diagnostics: changes can reflect motion as well as flicker, so they are not standalone temporal-quality scores.

Table 21 reports means and sample standard deviations. Paired differences are computed as coupled minus IID, with 95% t confidence intervals. We compute two-sided p-values using Monte Carlo sign-flip tests with $100,000$ random sign-flip replicates, followed by Benjamini–Hochberg correction across the five video metrics.

Table 21: Means and sample standard deviations across 100 prompts for Wan2.1 T2V 1.3B with $K = 3$. Temporal measures are diagnostics rather than quality objectives.

Metric	IID mean (SD)	Coupled mean (SD)
Cross-video L2 (pixel MSE)	0.1626 (0.0471)	0.1963 (0.0520)
Cross-video LPIPS	0.7497 (0.0605)	0.7943 (0.0513)
Frame-level text CLIP	0.6337 (0.0758)	0.6318 (0.0776)
Adjacent-frame LPIPS	0.0542 (0.0582)	0.0516 (0.0510)
Adjacent-frame CLIP cosine	0.9873 (0.0145)	0.9880 (0.0137)

Coupling increases mean cross-video L2 by 20.69% and LPIPS by 5.96%, with positive differences on 87 and 88 of 100 prompts, respectively. The mean paired differences are 0.03365 (95% CI $[0.02749, 0.03981]$) for L2 and 0.04466 ($[0.03592, 0.05340]$) for LPIPS. Both remain significant after correction ($p_{\mathrm{BH}} \approx 2.50 \times 10^{-5}$).

Mean frame-level text CLIP changes from 0.6337 to 0.6318, with a paired difference of -0.00192 (95% CI $[-0.01024, 0.00641]$; $p_{\mathrm{BH}} = 0.6513$). Adjacent-frame LPIPS and CLIP cosine also have intervals spanning zero, with adjusted $p = 0.4773$ for both. These results support increased cross-video pixel/perceptual diversity on this prompt set.

The average generation time per three-video gallery on GH200 GPUs is 51.14 seconds for IID and 51.06 seconds for coupled initialization; both reach 18.59 GiB peak allocated GPU memory in this configuration. Figure 6 shows illustrative frames for one prompt.

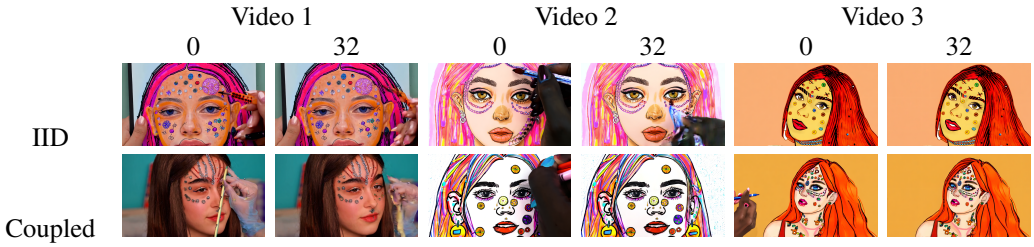

Figure 6: First and last frames of all gallery members for the prompt: “A drawing of a young woman with many facial piercings.”

D.11 Exploratory Automatic Gallery Judgments

An exploratory study compared randomly ordered pairs of $K = 3$ SD3 galleries for 200 COCO prompts using an LLM judge, reported in the study record as ChatGPT 5.6 Luna. The diversity instruction asked the judge to consider subjects, scenes, viewpoints, colors, composition, and other visual attributes and answer A, B, or T for a tie. Coupling versus IID received 86 wins, 75 losses, and 39 ties; coupled+CNO versus IID received 91 wins, 71 losses, and 38 ties. A separate favorite-image question favored the coupled gallery in 115 of 200 cases. These promising preliminary results motivate larger-scale evaluations with human judges and a broader range of gallery-selection tasks.