IS IMITATION ALL WE NEED FOR GENERATIVE MODELING? SPECTRAL ENTROPY GUIDANCE FOR EXTRAPOLATIVE GENERATION IN DIFFUSION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Distribution matching is a central principle of generative modeling, with models typically designed to reproduce a reference distribution as faithfully as possible. This raises the broader question of whether reference matching should be the sole objective of generative modeling. We study this question through *imaginative generation*, which seeks distributions satisfying prescribed properties while remaining as close as possible to a reference distribution. We instantiate this principle using representation-aware spectral entropy, with complementary formulations based on von Neumann and order-2 Rényi entropies. Both define convex constraint sets and admit a common projection-based interpretation. The spectral entropy of the data distribution defines a natural boundary, which we call the *Entropy Wall*. Below the wall, a spectral entropy requirement can, under suitable conditions, recover variation lost by a learned generator while reducing its discrepancy to the data distribution. Beyond the wall, the data no longer satisfies the prescribed entropy level, and the framework transitions to controlled spectral extrapolation. Under a KL-divergence anchor, we characterize the optimal targets through self-consistent spectral reweighting and derive *spectral entropy guidance*, a retraining-free mechanism for score-based and diffusion models. Our numerical experiments on synthetic and image-generation benchmarks indicate that the proposed guidance method could recover lost variation and enable controlled extrapolative generation beyond the spectral diversity of the data distribution.

1 INTRODUCTION

A central goal of generative modeling is to learn a probability model that faithfully reproduces a reference data distribution. This principle underlies generative adversarial networks (GANs) (Goodfellow et al., 2014), variational autoencoders (VAEs) (Kingma & Welling, 2014), and score-based and diffusion models (Song & Ermon, 2019; Ho et al., 2020; Song et al., 2021). Although these approaches differ in how they are trained and sampled, their population-level goal can broadly be expressed as minimizing a distributional discrepancy:

$$\min_{Q \in \mathcal{P}_{\mathcal{G}}} \mathcal{D}\left(Q; P_{\mathrm{ref}}\right), \tag{1}$$

where P_{ref} denotes a reference distribution, $\mathcal{P}_{\mathcal{G}}$ is the class of distributions realizable by the generator, and $\mathcal{D}$ measures discrepancy between the generated and reference distributions.

Distribution matching provides a natural statistical foundation for generative modeling, yet learned generators often underrepresent portions of the target distribution. Early manifestations include mode collapse and mode dropping in GANs (Goodfellow, 2016), where a generator concentrates on only a subset of the data modes. For example, Arora et al. (2018) used a birthday-paradox test to show that prominent GANs could learn distributions with relatively limited effective support. More recent studies have observed related diversity gaps in modern generative models (Dombrowski et al., 2025; Farnia et al., 2026). Correcting such a deficit remains consistent with faithful distribution matching: the goal is to recover variation already present in the reference distribution.

This phenomenon has motivated a growing class of sampling-time methods for promoting diversity. Some intervene on the conditioning process, as in CADS (Sadat et al., 2024), while others encourage

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

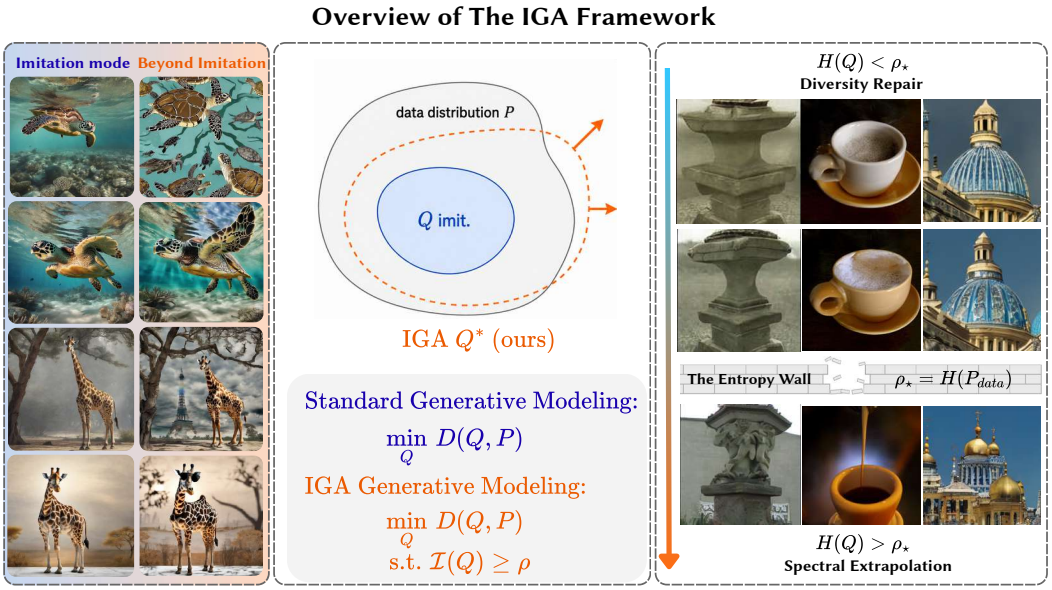


Figure 1: The population spectral entropy defines the Entropy Wall. Requirements at or below this level can support diversity repair; higher requirements prescribe spectral entropy exceeding that of the data in the chosen representation. In both regimes, discrepancy minimization favors closeness to the reference distribution. The right-side images are generated using IGA on DiT-XL/2 trained on ImageNet; the left-side images are text-to-image SDXL samples.

diversity through interactions among jointly or sequentially generated samples, as in Particle Guidance, Vendi Score Guidance, and SPARKE (Corso et al., 2024; Askari Hemmat et al., 2024; Jalali et al., 2025). These methods provide effective mechanisms for modifying the sampling process, and in some cases admit explicit characterizations of the resulting joint sampling law. However, they do not generally formulate diversity control as a population-level operation that maps a reference marginal distribution to the closest target distribution satisfying a prescribed distributional requirement. This leaves open the question of how diversity, novelty, or other desired properties can be incorporated directly into the definition of the generative target itself.

We therefore ask how to specify the desired property while controlling departure from the reference distribution. We propose the *Imaginative Generative AI (IGA)* framework, which expresses this goal as a constrained discrepancy minimization:

$$\begin{aligned} \min_{Q \in \mathcal{P}} \quad & \mathcal{D}\left(Q; P_{\mathrm{ref}}\right) \\ \text{subject to} \quad & \mathcal{I}(Q) \geq \rho, \end{aligned} \tag{2}$$

where $\mathcal{I}(Q)$ measures a desired property of the generated distribution and ρ is the required level. Hence, IGA seeks a distribution satisfying the requirement while remaining as close as possible to the reference distribution.

We instantiate this principle through *spectral diversity*. A fixed unit-normalized representation $\phi : \mathcal{X} \to \mathbb{R}^d$ induces the kernel similarity function $k(x, x') = \phi(x)^\top \phi(x')$ and associates each distribution with the unit-trace kernel covariance matrix $\Sigma_Q = \mathbb{E}_{x \sim Q}[\phi(x)\phi(x)^\top]$ as discussed in (Bach, 2023). Its spectrum describes how probability mass occupies feature directions. Hence, spectral entropy provides a reference-free but representation-dependent measure of variation: it requires no comparison distribution, while the representation determines which variations are distinguished.

We consider two complementary spectral entropies. The first is the von Neumann entropy of the kernel covariance (Bach, 2023):

$$H(Q) = -\operatorname{Tr}\big(\Sigma_Q \log \Sigma_Q\big), \tag{3}$$

whose exponential is the Vendi score (Friedman & Dieng, 2023). The second is the order-2 Rényi spectral entropy $H_{\mathrm{RKE}}(Q) = -\log \operatorname{Tr}\big(\Sigma_Q^2\big)$, whose exponential is the RKE score (Jalali et al., 2023;

Ospanov et al., 2024). These choices induce distinct spectral regularizers but share the geometry required by our projection analysis: H is concave, and H_{RKE} is quasi-concave; hence, both have convex superlevel sets. This enables a unified projection-based treatment of both entropy measures.

The population data distribution defines a natural threshold for spectral regularization:

$$\rho_\star = H(P_{\mathrm{data}}), \tag{4}$$

where H is the spectral entropy function. We call $\rho_\star$ the *Entropy Wall*. When $P_{\mathrm{data}} \in \mathcal{P}$, requirements at or below the wall preserve its feasibility. Under our KL-divergence projection, the resulting target is no farther from the data than the learned reference is. Spectral regularization can therefore repair a diversity deficit while improving agreement with the data. Beyond the wall, the data distribution becomes infeasible. The target must attain greater spectral diversity than the data, while discrepancy minimization controls its departure from the reference. We call this regime *imaginative generation through spectral extrapolation*. The entropy wall separates diversity repair from deliberate extrapolation, relative to the chosen entropy and representation.

To turn this target-distribution formulation into a sampling method, we specialize to KL divergence. The optimal target then takes the form of a self-consistent spectral reweighting of the reference distribution. The von Neumann formulation rewards underrepresented feature directions, while the order-2 Rényi formulation penalizes concentration along dominant directions. We derive the exact correction to the score function and practical approximations, leading to *IGA Guidance*, a retraining-free method for score-based and diffusion models. Once the estimated potential is fixed, sampling trajectories are independent conditional on that estimate. Our experiments examine how these mechanisms affect diversity and agreement with the data, first in controlled synthetic settings and then in pretrained image generators. Our main contributions are as follows. We also note that the extended related work section appears in Appendix A.

- We formulate imaginative generation as a distribution-level extension of reference matching with explicit control of representation-aware spectral diversity.
- We develop a projection theory for spectral-entropy-constrained generation and introduce the entropy wall, separating diversity repair from controlled spectral extrapolation.
- For KL-anchored generation, we characterize the optimal target through self-consistent spectral reweighting and derive spectral entropy guidance for score-based and diffusion models.
- We demonstrate diversity repair and spectral extrapolation on image-generation benchmarks.

2 Preliminaries

We measure diversity through a fixed representation of the samples. This section defines the resulting kernel covariance and the two spectral entropies used in IGA. We write $\mathcal{X}$ for the sample space, P_{data} for the population data distribution, P_{ref} for a reference distribution, and $\mathbb{P}(\mathcal{X})$ for the set of probability measures on $\mathcal{X}$.

Representation Space and Kernel Covariance. We fix a measurable representation $\phi : \mathcal{X} \to \mathbb{R}^d$, normalized such that $\|\phi(x)\|_2 = 1$ for every $x \in \mathcal{X}$. It induces the kernel similarity $k(x, y) = \phi(x)^\top \phi(y)$. For a distribution Q, the population kernel covariance (Bach, 2023) is

$$\Sigma_Q = \mathbb{E}_{x\sim Q}\left[\phi(x)\phi(x)^\top\right]. \tag{5}$$

Feature normalization ensures that $\Sigma_Q \succeq 0$ and $\mathrm{Tr}(\Sigma_Q) = 1$, and therefore its eigenvalues form a probability vector. The spectrum describes how mass is distributed across feature directions: concentration in a few directions gives an uneven spectrum, while broader occupancy gives a more balanced one. This measure requires no comparison distribution, but its interpretation depends on the representation. Given samples $x_1, \ldots, x_N \sim Q$, we estimate the kernel covariance by

$$\widehat{\Sigma}_N = \frac{1}{N}\sum_{i=1}^{N} \phi(x_i)\phi(x_i)^\top.$$

Let K be the Gram matrix with entries $K_{ij} = k(x_i, x_j)$. Then $\widehat{\Sigma}_N$ and $\frac{1}{N}K$ have the same non-zero eigenvalues. Therefore, the spectral quantities we define later can equivalently be calculated from the kernel covariance or the normalized Gram matrix.

Spectral Entropy Measures. Our primary measure of representation-level variation is the von Neumann entropy (VNE) of Σ_Q

$$H(Q) := -\operatorname{Tr}\big(\Sigma_Q \log \Sigma_Q\big) \tag{6}$$

This is the Shannon entropy of the eigenvalues of Σ_Q. Its exponential is the Vendi effective-diversity score (Friedman & Dieng, 2023). Also, we consider the order-2 Rényi kernel entropy (RKE), which is defined as

$$H_{\mathrm{RKE}}(Q) := \log \frac{1}{C_{\mathrm{RKE}}(Q)}, \text{ where} \qquad C_{\mathrm{RKE}}(Q) := \operatorname{Tr}(\Sigma_Q^2), \tag{7}$$

whose exponential is the order-2 RKE effective-diversity score (Jalali et al., 2023; 2025). The inverse score C_{RKE} measures spectral concentration and can also be written in terms of pairwise sample similarities. For independently drawn sample pair $x, x' \sim Q$, we note the following identity:

$$C_{\mathrm{RKE}}(Q) = \mathbb{E}_{x,x' \overset{\text{iid}}{\sim} Q}\big[k(x,x')^2\big]. \tag{8}$$

Smoothed Spectral Entropy Measures. Although (6) is well defined for singular covariances, its derivative involves $\log \Sigma_Q$. For every $\varepsilon \in (0,1)$, we therefore consider the smoothed kernel covariance $S_Q^\varepsilon = (1-\varepsilon)\Sigma_Q + \varepsilon\big(\frac{1}{d} I_d\big)$ and define its entropy as

$$H_\varepsilon(Q) = -\operatorname{Tr}\big(S_Q^\varepsilon \log S_Q^\varepsilon\big) \tag{9}$$

The lower bound $S_Q^\varepsilon \succeq \frac{\varepsilon}{d} I_d$ keeps every eigenvalue away from zero, so the matrix logarithm is well defined and bounded. The smoothing raises entropy, with $H_\varepsilon(Q) \geq H(Q)$, and H_ε converges uniformly to H as $\varepsilon \to 0$ (as shown in Lemma 1).

3 IGA: Spectral Entropy Constraints and Duality

3.1 Spectral-Constrained Reference Matching

IGA seeks the closest distribution to a reference subject to a prescribed level of spectral diversity. Let $\mathcal{P} \subseteq \mathbb{P}(\mathcal{X})$ be a convex set of candidate distributions, and let $\mathcal{J}(Q) = \mathcal{D}(Q; P_{\mathrm{ref}})$ measure discrepancy from the reference. For an entropy requirement ρ, we solve

$$\begin{aligned} Q_\rho \;\in\; \underset{Q \in \mathcal{P}}{\operatorname{argmin}} \quad & \mathcal{J}(Q) \\ \text{s.t.} \quad & H(Q) \geq \rho \end{aligned} \tag{10}$$

Since VNE is concave in Σ_Q and $Q \mapsto \Sigma_Q$ is affine, the feasible region in (10) is a convex set of distributions. Introducing a Lagrange multiplier $\lambda \geq 0$ gives the corresponding penalized objective

$$F_\lambda(Q) = \mathcal{J}(Q) - \lambda H(Q) \tag{11}$$

Appendix B gives the conditions under which the constrained and penalized formulations agree. Varying λ traces a family of targets, placing more weight on diversity as the coefficient increases. Globally optimal solutions satisfy $H(Q_{\lambda_1}) \leq H(Q_{\lambda_2})$ whenever $\lambda_1 < \lambda_2$, and therefore entropy is non-decreasing along this path.

The Entropy Wall. We consider the spectral entropy of the underlying distribution P_{data} as

$$\rho_\star = H(P_{\mathrm{data}}) \tag{12}$$

and we call $\rho_\star$ the *Entropy Wall*. If $\rho \leq \rho_\star$ and $P_{\mathrm{data}} \in \mathcal{P}$, the data distribution remains feasible; therefore, the constraint can restore diversity missing from the reference without requiring greater diversity than the data. On the other hand, when $\rho > \rho_\star$, the data distribution itself becomes infeasible, and every feasible target necessarily has a greater spectral entropy. We refer to these two regimes as *diversity repair* and *spectral extrapolation*, respectively.

With *KL divergence*, this distinction has a precise projection interpretation. Let $\mathcal{J}(Q) = D_{\mathrm{KL}}(Q \| P_{\mathrm{ref}})$ and optimize over all probability laws. Under the finite-KL feasibility conditions of Proposition 3, the unique optimizer Q_ρ satisfies, for every feasible distribution R,

$$D_{\mathrm{KL}}(R \| P_{\mathrm{ref}}) \geq D_{\mathrm{KL}}(R \| Q_\rho) + D_{\mathrm{KL}}(Q_\rho \| P_{\mathrm{ref}}) \tag{13}$$

Below the Entropy Wall $\rho \le \rho_\star$, we can set $R = P_{\mathrm{data}}$ (P_{data} is feasible in this regime), resulting in

$$D_{\mathrm{KL}}(P_{\mathrm{data}} \| Q_\rho) \le D_{\mathrm{KL}}(P_{\mathrm{data}} \| P_{\mathrm{ref}}) - D_{\mathrm{KL}}(Q_\rho \| P_{\mathrm{ref}}). \tag{14}$$

As a result, when the reference under-represents variation already present in the data, increasing spectral diversity can simultaneously move the target closer to P_{data}. Beyond the wall, this projection guarantee no longer applies and IGA deliberately enters the spectral-extrapolation regime.

3.2 Variational Representation of Spectral Entropy

The smoothed entropy H_ε admits a sample-wise variational form. Let $\mathbb{R}^{d\times d}_{\mathrm{sym}}$ denote the symmetric $d \times d$ matrices and define

$$\mathfrak{T}_\varepsilon = \Big\{ \Theta \in \mathbb{R}^{d\times d}_{\mathrm{sym}} : \mathrm{Tr}(\Theta) = 0,\ \|\Theta\|_{\mathrm{op}} \le \log \frac{d}{\varepsilon} \Big\}. \tag{15}$$

For $\Theta \in \mathfrak{T}_\varepsilon$, define

$$\ell_\varepsilon(Q, \Theta) = (1-\varepsilon)\mathbb{E}_Q\big[\phi(X)^\top \Theta \phi(X)\big] + \log \mathrm{Tr}(e^{-\Theta}). \tag{16}$$

Theorem 1. *For every $Q \in \mathbb{P}(\mathcal{X})$, the smoothed entropy admits the variational representation*

$$H_\varepsilon(Q) = \min_{\Theta \in \mathfrak{T}_\varepsilon} \ell_\varepsilon(Q, \Theta), \tag{17}$$

$$\Theta^\star(Q) = -\log S^\varepsilon_Q + \frac{\mathrm{Tr}(\log S^\varepsilon_Q)}{d} I_d, \tag{18}$$

where (18) is the unique minimizer of (17). Furthermore, if $\mathcal{J}$ is convex and lower semicontinuous and the regularity conditions of Appendix B hold, then the following minimax identity holds

$$\begin{aligned} \inf_Q \big\{ \mathcal{J}(Q) - \lambda H_\varepsilon(Q) \big\} &= \inf_Q \max_{\Theta \in \mathfrak{T}_\varepsilon} \big\{ \mathcal{J}(Q) - \lambda \ell_\varepsilon(Q, \Theta) \big\} \\ &= \max_{\Theta \in \mathfrak{T}_\varepsilon} \inf_Q \big\{ \mathcal{J}(Q) - \lambda \ell_\varepsilon(Q, \Theta) \big\}. \end{aligned} \tag{19}$$

Theorem 1 expresses the global entropy through a sample-wise quadratic potential. For fixed Θ, Q enters only through $\mathbb{E}_Q[\phi(X)^\top \Theta \phi(X)]$; for fixed Q, Equation (18) gives the optimal matrix. We use this duality to characterize the KL-anchored target.

4 KL-Divergence Discrepancy Measure and IGA Guidance

4.1 Self-Consistent Spectral Reweighting in the KL-Divergence Setting

We now set $\mathcal{J}(Q) = D_{\mathrm{KL}}(Q\|P)$ with $P = P_{\mathrm{ref}}$[1]. For a distribution Q, define the spectral potential

$$V^\varepsilon_Q(x) = (1-\varepsilon)\phi(x)^\top [-\log S^\varepsilon_Q] \phi(x). \tag{20}$$

Proposition 1. *For every $\lambda \ge 0$, the objective $D_{\mathrm{KL}}(Q\|P) - \lambda H_\varepsilon(Q)$ has a unique minimizer $Q^\star_\varepsilon$, which is characterized by the following fixed-point equation:*

$$\frac{dQ^\star_\varepsilon}{dP}(x) = \frac{\exp\big(\lambda V^\varepsilon_{Q^\star_\varepsilon}(x)\big)}{\mathbb{E}_{x' \sim P}\big[\exp\big(\lambda V^\varepsilon_{Q^\star_\varepsilon}(x')\big)\big]}. \tag{21}$$

Conversely, every distribution Q satisfying (21) is the unique minimizer.

The full proof, including existence and strong-optimality guarantees, is given in Appendix C. Equation (21) shows that IGA upweights samples along weakly represented spectral directions through the coefficient $-\log s$. The reweighting is self-consistent because its potential depends on the covariance of the resulting target. We also prove that $Q^\star_\varepsilon \sim P$, hence the KL IGA redistributes mass within the measure-theoretic support of the reference.

Corollary 1 (IGA target score)**.** *If P has a strictly positive C^1 density p and ϕ is C^1, then*

$$\nabla_x \log q^\star_\varepsilon(x) = \nabla_x \log p(x) + \lambda \nabla_x V^\varepsilon_{Q^\star_\varepsilon}(x), \tag{22}$$

$$\nabla_x V^\varepsilon_Q(x) = 2(1-\varepsilon) D\phi(x)^\top [-\log S^\varepsilon_Q] \phi(x). \tag{23}$$

Hence, IGA modifies the reference score by the gradient of a representation-aware spectral potential.

[1]The divergence $D_{\mathrm{KL}}(Q\|P)$ is infinite unless Q is absolutely continuous with respect to P.

4.2 IGA Guidance for Diffusion Models

The preceding score formula applies to clean samples. To use it in a diffusion sampler, we must propagate the target through the forward noising process. Let $\Pi_t(dx_t \mid x_0)$ denote the noising kernel, with $P_t = P\Pi_t$ and $Q^\star_{\varepsilon,t} = Q^\star_\varepsilon \Pi_t$, and define $h_t(x_t) := \mathbb{E}_P[\exp\{\lambda V^\varepsilon_{Q^\star_\varepsilon}(X_0)\} \mid X_t = x_t]$. Appendix E shows that the target score at noise level t is

$$\nabla_{x_t} \log q^\star_{\varepsilon,t}(x_t) = \nabla_{x_t} \log p_t(x_t) + \nabla_{x_t} \log h_t(x_t). \tag{24}$$

Therefore, the exact correction averages the clean-sample weight over the possible clean samples underlying x_t, then takes the gradient of its logarithm. Computing this conditional expectation is the main obstacle to applying the exact score formula directly. Our implementation separates potential fitting from guided sampling. We first fit weights $\{\widehat{q}_j\}_{j=1}^N$ on an independent pilot set $\mathcal{B} = \{\widetilde{x}_j\}_{j=1}^N$ using Algorithm 1, and freeze

$$\widehat{\Sigma} = \sum_{j=1}^N \widehat{q}_j \phi(\widetilde{x}_j)\phi(\widetilde{x}_j)^\top, \qquad M_{\mathrm{VNE}} = -(1-\varepsilon)\log\big((1-\varepsilon)\widehat{\Sigma} + \varepsilon(\tfrac{1}{d} I_d)\big). \tag{25}$$

This gives a fixed potential $\widehat{V}(x) = \phi(x)^\top M_{\mathrm{VNE}} \phi(x)$. As in diffusion posterior sampling (Chung et al., 2023) and universal guidance (Bansal et al., 2024), we evaluate the clean-sample objective at the model's denoised prediction $\widehat{x}_0 = (x_t - \sqrt{1-\bar{\alpha}_t}\,\epsilon_\theta(x_t,t))/\sqrt{\bar{\alpha}_t}$. Replacing the conditional log-moment by $\lambda \widehat{V}(\widehat{x}_0)$ gives the practical correction

$$\widetilde{u}_t = \nabla_{x_t}\big[\lambda \widehat{V}(\widehat{x}_0)\big], \qquad \epsilon^{\mathrm{IGA}}_\theta = \epsilon_\theta - \sqrt{1-\bar{\alpha}_t}\,\widetilde{u}_t. \tag{26}$$

The corrected prediction can be used directly with standard DDPM or DDIM samplers. We defer details of the pilot fitting, a direct-gradient approximation, and sampler-specific updates to Appendix F.

4.3 IGA with Order-2 Rényi Kernel Entropy

The same construction extends to order-2 Rényi entropy. The constraint $H_{\mathrm{RKE}}(Q) \geq \rho$ is equivalent to $C_{\mathrm{RKE}}(Q) \leq e^{-\rho}$, and $C_{\mathrm{RKE}}(Q) = \|\Sigma_Q\|_F^2$ is convex in Q. The RKE constraint therefore has the same convex projection geometry as the VNE constraint. The corresponding population potential is

$$V_{\mathrm{RKE},Q}(x) \;=\; -2\,\mathbb{E}_{X'\sim Q}\big[k(x,X')^2\big] \;=\; -2\,\phi(x)^\top \Sigma_Q \phi(x). \tag{27}$$

Equation (27) is the population counterpart of SPARKE's empirical interaction (Jalali et al., 2025). SPARKE couples samples through their generation history; IGA evaluates the interaction under a fixed target Q. With the KL-divergence anchor, the penalized objective becomes $D_{\mathrm{KL}}(Q\|P) + \lambda C_{\mathrm{RKE}}(Q)$. Its unique minimizer satisfies the analogous fixed-point equation (Appendix D):

$$\frac{dQ^\star_{\mathrm{RKE}}}{dP}(x) = \frac{\exp\big(\lambda V_{\mathrm{RKE},Q^\star_{\mathrm{RKE}}}(x)\big)}{\mathbb{E}_{x'\sim P}\big[\exp\big(\lambda V_{\mathrm{RKE},Q^\star_{\mathrm{RKE}}}(x')\big)\big]}. \tag{28}$$

As a result, IGA-RKE defines a single target distribution from which samples can be drawn independently once its potential is fixed. At inference, we fit $\widehat{\Sigma}$ with the RKE pilot objective and use $M_{\mathrm{RKE}} = -2\widehat{\Sigma}$ in (26), in place of the VNE matrix logarithm.

5 Numerical Evaluation

We test whether IGA recovers variation missing from a base generator and increases diversity beyond the data entropy wall. We first track probability mass on controlled synthetic distributions, then evaluate pretrained image generators, including text-conditioned generation, using von Neumann and order-2 Rényi entropy. We report unsmoothed spectral entropies and their exponentials: the Vendi score for IGA-VNE (Friedman & Dieng, 2023) and the RKE score for IGA RKE (Jalali et al., 2023; Ospanov et al., 2024). Unless stated otherwise, we fit the IGA potential on an independent pilot batch and freeze it for evaluation, varying λ to control the distribution. Comparisons match sample counts, including between real and generated samples. Appendix J gives the full experimental protocol.

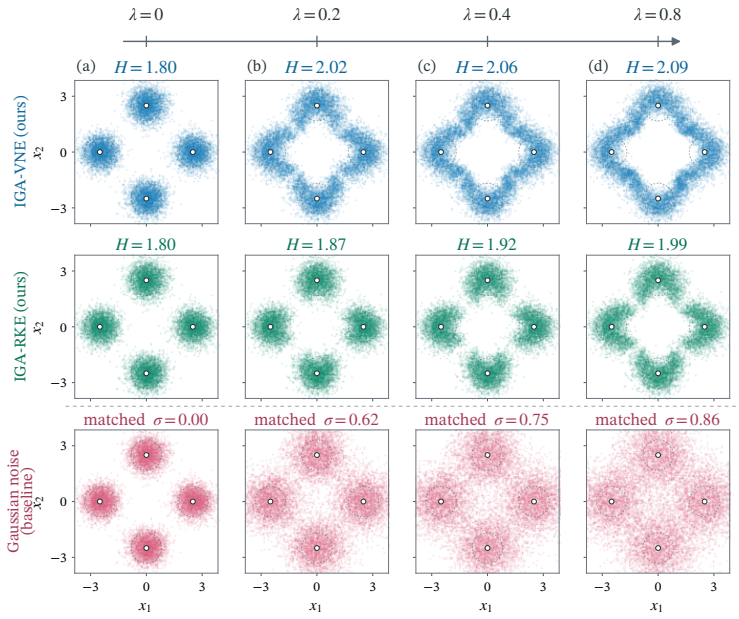


Figure 2: The application of IGA-VNE and IGA-RKE to the score function of a synthetic Gaussian mixture model. The IGA methods found a less noisy extension of the data manifold than the baseline of additive Gaussian noise $\mathcal{N}(\mathbf{0}, \sigma^2 I)$, with σ chosen to match the same spectral entropy level.

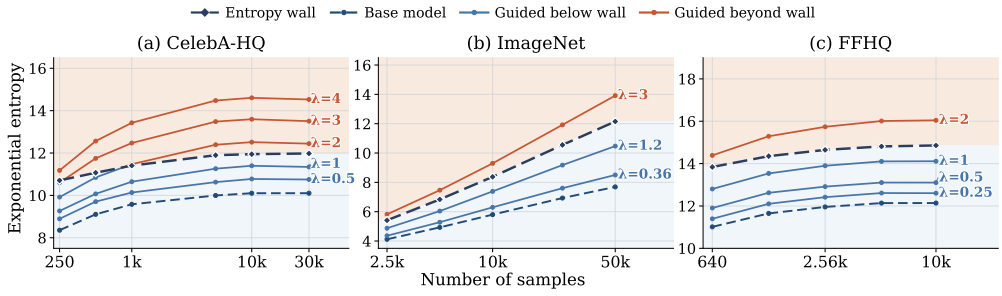


Figure 3: Vendi scores of real and generated images on CelebA-HQ, ImageNet, and FFHQ, evaluated at matched sample counts.

5.1 Synthetic Experiments with Known Ground Truth

Synthetic distributions with known ground truth let us distinguish structured diversity from added noise and repair from extrapolation. For a multimodal DDIM (Song et al., 2021), we compare IGA with Gaussian noise matched to IGA-VNE's representation-space entropy. Figure 2 shows that IGA-VNE connects neighboring modes into a ring, whereas noise broadens them diffusely; IGA-RKE fills the gaps less strongly at the displayed coefficients.

Appendix G.1 compares how the two objectives redistribute mass on a known helix at a common KL budget, and Appendix G.2 shows repair and extrapolation of component weights in an eight-component mixture. On a learned nonlinear manifold (Figure 7), IGA-VNE restores endpoint mass underrepresented by a DDPM (Ho et al., 2020) near the entropy wall, while further extrapolation increases both rare-region coverage and covariance discrepancy.

5.2 Repair and Extrapolation on Image Benchmarks

We next test whether IGA repairs diversity deficits and crosses the entropy wall in pretrained image generators. We evaluate CelebA-HQ (Karras et al., 2018), ImageNet (Russakovsky et al., 2015) using the 100-class subset of Tian et al. (2020), and FFHQ (Karras et al., 2019), comparing real and generated images at matched sample counts. On all three benchmarks, the base models fall below the data entropy; increasing λ with IGA-VNE closes this gap and crosses the wall (Figure 3).



Figure 4: **Base and IGA-guided text-to-image samples.** Samples for the prompts "A Designer fashion dress on a mannequin" and "Architectural visualization of a skyscraper" from SDXL (top) and PixArt-Σ (bottom). Base samples are on the left and guided samples on the right.

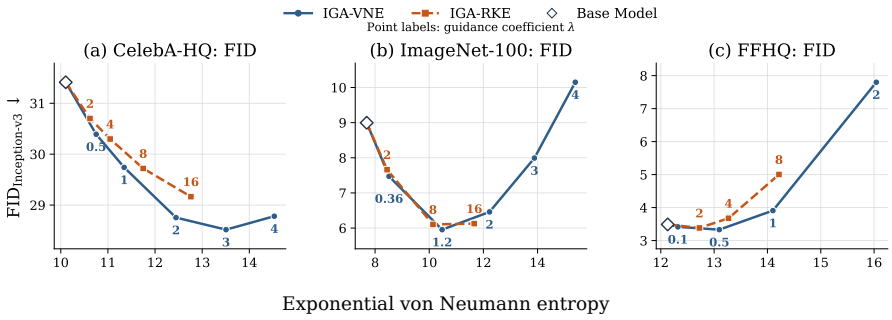


Figure 5: **Inception-space FID along the IGA paths.** FID versus the DINOv2-space Vendi score on CelebA-HQ, ImageNet, and FFHQ. Curves show IGA-VNE and IGA-RKE; diamonds mark the base models and point labels give λ.

We measure the distance to the real data distribution using FID and KID in Inception-v3 space (Heusel et al., 2017; Stein et al., 2023; Bińkowski et al., 2018), independently of the DINOv2 features used for guidance (Oquab et al., 2024). Figure 5 shows that increasing the IGA coefficient λ initially brings the generated distribution closer to the real data while increasing diversity. At larger λ, FID can rise as diversity continues to grow. The corresponding KID paths and comparisons across feature representations and entropy measures appear in Appendix I.2.

Table 1 compares IGA with CADS (Sadat et al., 2024), SPARKE (Jalali et al., 2025), and Particle Guidance (Corso et al., 2024) on ImageNet. At the repair settings, both IGA objectives improve FID, KID, DINOv2 distances, and recall over the base model. Stronger guidance gives IGA-VNE the highest recall, while IGA-RKE achieves the lowest FID and KID among the reported methods. CelebA-HQ and FFHQ comparisons appear in Tables 2 and 3 in Appendix I; guidance settings are listed in Appendix I.1. Further studies find similar guidance paths when fitting on training images instead of generated samples (Appendix I.3) and initial FID/KID improvements across four ImageNet diffusion backbones (Appendix I.4). The fitted weights and appearance changes on the face benchmarks are examined in Appendix H.5; matched ImageNet samples appear in Appendix H.6.

Table 1: **ImageNet comparison.** Estimates are rounded for display; ± uncertainties retain their reported precision. Precision/Recall are DINOv2 percentages. IGA (R), (E), and (H) denote repair, extrapolation, and high diversity; exact λ values are in Appendix I.1.

		Inception-v3		DINOv2		Coverage & Diversity			
	Method	FID ↓	KID $\times 10^3$ ↓	FD ↓	KD $\times 10^2$ ↓	Prec. (%) ↑	Rec. (%) ↑	H ↑	H_{RKE} ↑
ImageNet	Base	9.00±0.06	3.20±0.04	118.3±0.3	4.59±0.002	67.0±0.2	35.9±0.2	2.04±0.002	0.72±0.001
	CADS	6.92±0.05	1.69±0.05	131.1±0.6	**4.52±0.002**	50.1±0.1	54.3±0.3	2.61±0.003	0.96±0.001
	Particle	8.23±0.04	2.32±0.04	120.3±0.5	4.55±0.002	61.1±0.2	40.6±0.2	2.20±0.002	0.78±0.001
	SPARKE	8.37±0.08	2.90±0.04	115.2±0.4	4.58±0.002	**66.3±0.3**	36.3±0.3	2.06±0.002	0.72±0.001
	IGA-VNE (R)	7.01±0.06	1.86±0.02	114.1±0.6	4.56±0.002	61.7±0.2	39.6±0.5	2.18±0.002	0.77±0.001
	IGA-VNE (E)	10.14±0.05	2.40±0.10	196.3±0.8	4.55±0.002	42.6±0.1	**56.0±0.3**	**2.73±0.001**	**1.02±0.001**
	IGA-RKE (R)	6.74±0.06	1.65±0.04	**113.5±0.6**	4.56±0.001	61.1±0.1	40.1±0.5	2.21±0.002	0.79±0.001
	IGA-RKE (H)	**6.13±0.02**	**0.60±0.01**	132.5±0.8	4.54±0.001	52.9±0.2	47.3±0.3	2.46±0.002	0.90±0.001

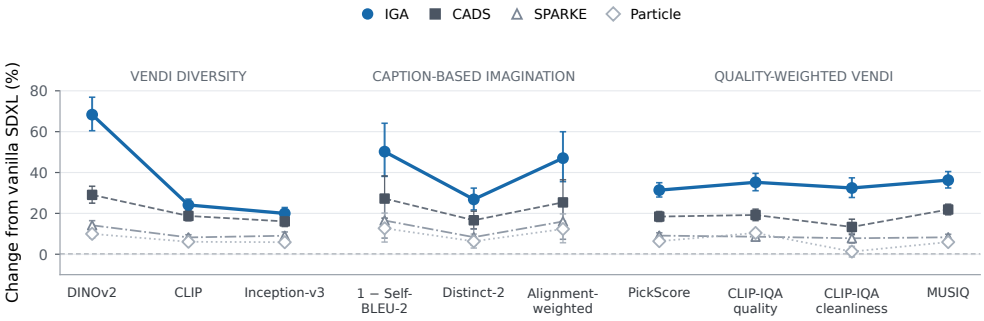


Figure 6: **Prompt-conditional diversity on SDXL.** Percentage change from vanilla SDXL across matched prompts; error bars are 95% paired prompt-bootstrap intervals.

5.3 Prompt-Conditioned Generation

We also test IGA on SDXL (Podell et al., 2024) using a randomly-selected block of 100 MS-COCO (Lin et al., 2014) validation set prompts. Figure 6 compares IGA with CADS, SPARKE, and Particle Guidance. Under the tested settings, IGA increases visual, caption-based, and quality-weighted diversity relative to vanilla SDXL across the reported measures. Figure 4 displays the qualitative imagination-like effect of IGA-VNE spectral entropy guidance on SDXL-generated data for two prompts on fashion and architecture design. Additional SDXL and PixArt-Σ (Chen et al., 2024) visualizations appear in Appendix H, with full evaluation details in Appendix J.3.

6 Conclusion and Discussion

We introduced IGA as a target-distribution framework that complements reference matching with an explicit distributional requirement. Focusing on representation-aware spectral diversity, we showed that von Neumann entropy yields a convex constrained formulation and identified the data entropy as an *Entropy Wall* separating diversity repair from spectral extrapolation. For KL-anchored generation, we derived a self-consistent spectral reweighting of the reference distribution and its corresponding score correction, leading to retraining-free IGA Guidance for diffusion models. We further developed an order-2 RKE formulation that admits an analogous population-level interpretation and a computationally simpler spectral potential. Our experiments highlight both regimes of the framework, from correcting diversity deficits of generators to controlling spectral diversity beyond the data level.

More broadly, the IGA framework allows the distributional requirement to be expressed through a general functional $\mathcal{I}(Q)$ in (2). We focused on spectral entropy because it measures variation in a representation space and gives convex constraint sets for both formulations. These properties make the targets tractable and support the projection guarantees. Other choices of $\mathcal{I}$ could encode different properties of a generated population; identifying choices with similarly useful geometry is a direction for future work. The induced extrapolation depends on the chosen criterion. Within that specification, IGA makes the desired departure from a reference an explicit part of the generative objective.

AI USE STATEMENT

We used generative AI tools during manuscript editing and revision to assist with language polishing, organization, formatting, and refinement of theoretical formulations and proofs. All AI-assisted material was reviewed and verified by the authors, who take full responsibility for the final content of the submission.

ETHICS STATEMENT

This work studies diversity control in generative models through theoretical analysis and experiments with synthetic data, existing datasets, and pretrained models. Qualitative examples illustrate model behavior and variation in visual design, including garments displayed on mannequins. Their inclusion is not intended to promote or demote any particular appearance standard or social stereotype. Generated outputs may reflect biases in the underlying data and models, and increased spectral diversity does not by itself establish fairness or suitability for downstream use.

REPRODUCIBILITY STATEMENT

The theoretical assumptions and proofs are provided in the appendices. Appendix F describes the implementation and approximation details, while Appendix J documents the experimental configurations and evaluation protocols. Additional numerical and qualitative results appear in Appendices G–I.

REFERENCES

Yuang Ai, Qihang Fan, Xuefeng Hu, Zhenheng Yang, Ran He, and Huaibo Huang. DiCo: Revitalizing ConvNets for scalable and efficient diffusion modeling. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=UnslcaZSnb.

Sanjeev Arora, Andrej Risteski, and Yi Zhang. Do GANs learn the distribution? some theory and empirics. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=BJehNfW0-.

Reyhane Askari Hemmat, Melissa Hall, Alicia Yi Sun, Candace Ross, Michal Drozdzal, and Adriana Romero-Soriano. Improving geo-diversity of generated images with contextualized Vendi score guidance. In *European Conference on Computer Vision*. Springer, 2024.

Francis Bach. Information theory with kernel methods. *IEEE Transactions on Information Theory*, 69(2):752–775, 2023.

Arpit Bansal, Hong-Min Chu, Avi Schwarzschild, Roni Sengupta, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Universal guidance for diffusion models. In *International Conference on Learning Representations*, volume 2024, 2024.

Mikołaj Bińkowski, Dougal J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=r1lUOzWCW.

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

Hyungjin Chung, Jeongsol Kim, Michael Thompson Mccann, Marc Louis Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=OnD9zGAGT0k.

Gabriele Corso, Yilun Xu, Valentin De Bortoli, Regina Barzilay, and Tommi Jaakkola. Particle guidance: non-iid diverse sampling with diffusion models. In *International Conference on Learning Representations*, 2024.

Imre Csiszár and Frantisek Matus. Information projections revisited. *IEEE Transactions on Information Theory*, 49(6), 2003.

Mischa Dombrowski, Weitong Zhang, Sarah Cechnicka, Hadrien Reynaud, and Bernhard Kainz. Image generation diversity issues and how to tame them. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2025.

Nelson Dunford and Billy James Pettis. Linear operations on summable functions. *Transactions of the American Mathematical Society*, 47(3), 1940.

Ahmed Elgammal, Bingchen Liu, Mohamed Elhoseiny, and Marian Mazzone. CAN: Creative adversarial networks, generating "art" by learning about styles and deviating from style norms. In *Proceedings of the Eighth International Conference on Computational Creativity*, 2017.

Farzan Farnia, Mohammad Jalali, and Azim Ospanov. Exposing diversity bias in deep generative models: Statistical origins and correction of diversity error. *arXiv preprint arXiv:2602.14682*, 2026.

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

Songwei Ge, Vedanuj Goswami, C. Lawrence Zitnick, and Devi Parikh. Creative sketch generation. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=gwnoVHIES05.

Ian Goodfellow. NIPS 2016 tutorial: Generative adversarial networks. *arXiv preprint arXiv:1701.00160*, 2016.

Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33, 2020.

Mohammad Jalali, Cheuk Ting Li, and Farzan Farnia. An information-theoretic evaluation of generative models in learning multi-modal distributions. *Advances in Neural Information Processing Systems*, 36, 2023.

Mohammad Jalali, Haoyu Lei, Amin Gohari, and Farzan Farnia. SPARKE: Scalable prompt-aware diversity and novelty guidance in diffusion models via RKE score. *Advances in Neural Information Processing Systems*, 38, 2025.

Liwei Jiang, Yuanjun Chai, Margaret Li, Mickel Liu, Raymond Fok, Nouha Dziri, Yulia Tsvetkov, Maarten Sap, and Yejin Choi. Artificial Hivemind: The open-ended homogeneity of language models (and beyond). In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025. URL https://openreview.net/forum?id=saDOrrnNTz.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=Hk99zCeAb.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*. IEEE, 2019.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35, 2022.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. MUSIQ: Multi-scale image quality transformer. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 2021.

Pum Jun Kim, Yoojin Jang, Jisu Kim, and Jaejun Yoo. TopP&R: Robust support estimation approach for evaluating fidelity and diversity in generative models. *Advances in Neural Information Processing Systems*, 36, 2023.

Diederik P. Kingma and Max Welling. Auto-encoding variational Bayes. In *International Conference on Learning Representations*, 2014.

Michael Kirchhof, James Thornton, Louis Béthune, Pierre Ablin, Eugene Ndiaye, and Marco Cuturi. Shielded diffusion: Generating novel and diverse images using sparse repellency. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 30911–30942, 2025. URL https://proceedings.mlr.press/v267/kirchhof25a.html.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-Pic: An open dataset of user preferences for text-to-image generation. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=G5RwHpBUv0.

Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. *Advances in neural information processing systems*, 32, 2019.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 2016. doi: 10.18653/v1/N16-1014.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*. PMLR, 2022.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European conference on computer vision*. Springer, 2014.

Jack Lu, Ryan Teehan, and Mengye Ren. ProCreate, don't reproduce! propulsive energy diffusion for creative generation. In *European Conference on Computer Vision*. Springer, 2024.

Zichen Miao, Jiang Wang, Ze Wang, Zhengyuan Yang, Lijuan Wang, Qiang Qiu, and Zicheng Liu. Training diffusion models towards diverse image generation with reinforcement learning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2024.

Quan Nguyen and Adji Bousso Dieng. Quality-weighted Vendi scores and their application to diverse experimental design. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*. PMLR, 2024. URL https://proceedings.mlr.press/v235/nguyen24d.html.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=a68SUt6zFt. Featured Certification.

Azim Ospanov, Jingwei Zhang, Mohammad Jalali, Xuenan Cao, Andrej Bogdanov, and Farzan Farnia. Towards a scalable reference-free evaluation of generative models. *Advances in Neural Information Processing Systems*, 37, 2024.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin (eds.), *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318. Association for Computational Linguistics, 2002. doi: 10.3115/1073083.1073135. URL https://aclanthology.org/P02-1040/.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 2023.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *International Conference on Learning Representations*, volume 2024, 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*. PmLR, 2021.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. *Advances in neural information processing systems*, 20, 2007.

Elad Richardson, Kfir Goldberg, Yuval Alaluf, and Daniel Cohen-Or. ConceptLab: Creative concept generation using VLM-guided diffusion prior constraints. *ACM Transactions on Graphics*, 43(3), 2024.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*. ieee, 2022.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. ImageNet large scale visual recognition challenge. *International journal of computer vision*, 115(3), 2015.

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*, 2024.

Mehdi SM Sajjadi, Olivier Bachem, Mario Lucic, Olivier Bousquet, and Sylvain Gelly. Assessing generative models via precision and recall. *Advances in neural information processing systems*, 31, 2018.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training GANs. *Advances in neural information processing systems*, 29, 2016.

Vikash Sehwag, Caner Hazirbas, Albert Gordo, Firat Ozgenel, and Cristian Canton. Generating high fidelity data from low-density regions using diffusion models. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, 2022.

Maurice Sion. On general minimax theorems. *Pacific Journal of Mathematics*, 8(1):171–176, 1958. doi: 10.2140/pjm.1958.8.171.

Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Diffusion art or digital forgery? investigating data replication in diffusion models. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2023.

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

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

George Stein, Jesse Cresswell, Rasa Hosseinzadeh, Yi Sui, Brendan Ross, Valentin Villecroze, Zhaoyan Liu, Anthony L Caterini, Eric Taylor, and Gabriel Loaiza-Ganem. Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models. *Advances in Neural Information Processing Systems*, 2023.

Danica J Sutherland and Jeff Schneider. On the error of random fourier features. *arXiv preprint arXiv:1506.02785*, 2015.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *European conference on computer vision*. Springer, 2020.

Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring CLIP for assessing the look and feel of images. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 2555–2563, 2023a.

Zhendong Wang, Yifan Jiang, Huangjie Zheng, Peihao Wang, Pengcheng He, Zhangyang Wang, Weizhu Chen, and Mingyuan Zhou. Patch diffusion: Faster and more data-efficient training of diffusion models. In *Advances in Neural Information Processing Systems*, volume 36, 2023b.

Ankit Yadav, Arpit Garg, Ta Duc Huy, and Lingqiao Liu. STRIDE: Training-free diversity guidance via PCA-directed feature perturbation in single-step diffusion models. *arXiv preprint arXiv:2605.11494*, 2026.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. Texygen: A benchmarking platform for text generation models. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, 2018.

A Related Work

Diversity in generative models. Diversity loss is a persistent problem across generative modeling. In generative adversarial networks (GANs), mode collapse and limited effective support motivated minibatch discrimination and support-size diagnostics (Salimans et al., 2016; Arora et al., 2018). Language models likewise tend toward generic, repetitive, or homogeneous outputs, motivating diversity-aware objectives and decoding strategies (Li et al., 2016; Jiang et al., 2025). Related deficits have been documented in image diffusion models, together with a systematic gap between real and generated diversity for which finite-sample entropy underestimation is one statistical source (Dombrowski et al., 2025; Farnia et al., 2026). These findings motivate diversity regularization and correction methods, including our study, where diversity is analyzed and specified as a property of the target distribution.

Measuring novelty and diversity. The Vendi Score exponentiates von Neumann entropy to measure effective diversity (Friedman & Dieng, 2023), whereas RKE uses the order-two Rényi entropy to provide a tractable score with a mode-count interpretation (Jalali et al., 2023). Precision and recall instead separate sample fidelity from coverage of the data distribution (Sajjadi et al., 2018; Kynkäänniemi et al., 2019; Kim et al., 2023). Our proposed framework moves from post-hoc evaluation to regularizing the generative objective under a separate fidelity term, and uses population-data entropy as a wall separating recovery of lost diversity from deliberate extrapolation beyond the data.

Promoting novelty and diversity. Existing interventions typically specialize either training or generation. At training time, diversity has been promoted through reinforcement learning with explicit diversity rewards (Miao et al., 2024) and through diversity-aware diffusion modules (Dombrowski et al., 2025). At inference, CADS anneals noise in the conditioning signal (Sadat et al., 2024); c-VSG and SPARKE are non-i.i.d. sampling-time methods that promote diversity through contextualized Vendi and conditional RKE interactions over generated samples, respectively (Askari Hemmat et al., 2024; Jalali et al., 2025). STRIDE perturbs intermediate features in distilled one- and few-step generators (Yadav et al., 2026). Particle Guidance evolves an interacting set under a pairwise diversity potential and is explicitly non-i.i.d. (Corso et al., 2024), while SPELL repels trajectories from protected, concurrent, or previously generated images (Kirchhof et al., 2025). Our proposed spectral entropy guidance fits a distribution-level spectral potential on an independent pilot and freezes it during sampling, leading to i.i.d. data generation.

Creative generation. Creative generation has been pursued through deviation from learned styles, novel composition, rare-region sampling, and repulsion from exemplars. Creative Adversarial Networks depart from established artistic styles (Elgammal et al., 2017); DoodlerGAN recombines object parts into unseen sketches (Ge et al., 2021); and ConceptLab searches for new category members (Richardson et al., 2024). For diffusion models, evidence of training-data replication sharpens the distinction between creativity and reproduction (Somepalli et al., 2023). Low-density sampling explores rare regions of the learned distribution (Sehwag et al., 2022), whereas ProCreate pushes generations away from reference images (Lu et al., 2024). Our proposed framework follows a different approach and aims to induce higher spectral diversity in the outputs of generative models, with the goal of learning generative models capable of generating more diverse data than their underlying training distribution.

B Technical Results for Spectral Constraints

Throughout our analysis, we follow the notation in Section 2 and assume that d is finite, $\phi : \mathcal{X} \to \mathbb{R}^d$ is measurable, and $\|\phi(x)\|_2 = 1$. Note that the matrix Σ_Q is the uncentered feature second moment in (5). We note that this matrix is clearly positive semidefinite and has unit trace as $\|\phi(x)\|_2 = 1$ holds everywhere. We consider the standard trace inner product $\langle A, B\rangle_F = \mathrm{Tr}(AB)$ for symmetric matrices and define

$$\mathcal{S}_d = \{S \in \mathbb{R}^{d\times d}_{\mathrm{sym}} : S \succeq 0,\ \mathrm{Tr}(S) = 1\}, \qquad \mathsf{h}(S) = -\mathrm{Tr}(S \log S).$$

Therefore, our earlier definitions can be rewritten as $H(Q) = \mathsf{h}(\Sigma_Q)$ and $H_\varepsilon(Q) = \mathsf{h}(S^\varepsilon_Q)$.

B.1 Matrix Entropy and Spectral Smoothing

We first state the matrix-entropy basic properties used in the proofs. For $S, T \in \mathcal{S}_d$, we define the matrix-based relative entropy as

$$D_{\mathrm{mat}}(S\|T) = \mathrm{Tr}\big[S(\log S - \log T)\big] \quad \text{if } \mathrm{Range}(S) \subseteq \mathrm{Range}(T), \tag{29}$$

Note that the function h is continuous and concave on $\mathcal{S}_d$, with $0 \le \mathsf{h}(S) \le \log d$. Moreover, it is known that $D_{\mathrm{mat}}(S\|T) \ge 0$, with equality if and only if $S = T$.

Lemma 1. *Both H and smoothed entropy H_ε are concave in Q. For $0 < \varepsilon < 1$, define $b_\varepsilon = h_{\mathrm{b}}(\varepsilon) + \varepsilon \log d$, where $h_{\mathrm{b}}(\varepsilon) = -\varepsilon \log \varepsilon - (1-\varepsilon)\log(1-\varepsilon)$. Then, the following holds*

$$0 \le H_\varepsilon(Q) - H(Q) \le b_\varepsilon \qquad \text{for every } Q \in \mathbb{P}(\mathcal{X}). \tag{30}$$

Specifically, $H_\varepsilon \to H$ holds uniformly over probability distributions as ε tends to 0.

Proof. Concavity follows the concavity of the VNE entropy function and the affine dependence of Σ_Q and S^ε_Q on Q. We also have that

$$H_\varepsilon(Q) \ge (1-\varepsilon)H(Q) + \varepsilon \log d \ge H(Q).$$

For the upper bound, let $s = (s_1, \ldots, s_d)$ be the eigenvalue vector of Σ_Q. Since the identity commutes with Σ_Q, the eigenvalues of S^ε_Q are $(1-\varepsilon)s_i + \varepsilon/d$. These are the probabilities of a random index obtained by drawing from s with probability $1-\varepsilon$ and uniformly with probability ε. The entropy of this index is at most its joint entropy with the Bernoulli mixture indicator. Therefore, we obtain

$$H_\varepsilon(Q) \le h_{\mathrm{b}}(\varepsilon) + (1-\varepsilon)H(Q) + \varepsilon \log d.$$

Subtracting $H(Q) \ge 0$ proves (30), and note that the bound tends to zero independently of Q. The proof is hence complete. □

B.2 Constrained Optimization and the Regularization Path

Lemma 2. *Assume $\mathcal{P}$ is a convex set of probability distributions, $\mathcal{J} : \mathcal{P} \to [0, +\infty]$ is convex, and $\Psi : \mathcal{P} \to \mathbb{R}$ is concave. Suppose the following problem has an optimizer and admits $Q^\circ \in \mathrm{dom}\, \mathcal{J}$ with $\Psi(Q^\circ) > \tau$:*

$$p^\star \ = \ \min_{Q\in\mathcal{P}} \mathcal{J}(Q) \quad \text{s.t.} \quad \Psi(Q) \ge \tau \tag{31}$$

Then, there exists a finite multiplier $\lambda^\star \ge 0$ such that every optimizer $Q^\star$ minimizes $\mathcal{J}(Q) - \lambda^\star \Psi(Q)$ over $\mathcal{P}$ and satisfies $\lambda^\star(\Psi(Q^\star) - \tau) = 0$.

Proof. Consider $g(Q) = \tau - \Psi(Q)$ and define the perturbation value $v(s)$ as follows:

$$v(s) = \inf\{\mathcal{J}(Q) : Q \in \mathcal{P},\ g(Q) \le s\}.$$

The convexity of g and $\mathcal{J}$ implies that v is a convex function. This map is also non-increasing and bounded below by zero. Since $g(Q^\circ) < 0$ and $\mathcal{J}(Q^\circ) < \infty$, v is finite on an open interval containing zero. Therefore, it has a finite subgradient ξ at zero. Monotonicity gives $\xi \le 0$, and therefore we set $\lambda^\star = -\xi$.

For every $Q \in \operatorname{dom} \mathcal{J}$, the subgradient inequality gives the following:

$$\mathcal{J}(Q) \ge v(g(Q)) \ge v(0) + \xi g(Q) = p^\star - \lambda^\star g(Q).$$

The same lower bound for $\mathcal{J}(Q) + \lambda^\star g(Q)$ is trivial outside $\operatorname{dom} \mathcal{J}$. For an optimizer $Q^\star$, feasibility and $\mathcal{J}(Q^\star) = p^\star$ force $\lambda^\star g(Q^\star) = 0$. Hence, $Q^\star$ minimizes the Lagrangian, and removing its constant term $\lambda^\star \tau$ proves the result. □

Lemma 3. *Let Ψ be real-valued, and suppose Q_i globally minimizes $\mathcal{J}(Q) - \lambda_i \Psi(Q)$ with a finite objective value, where $0 \le \lambda_1 < \lambda_2$. Then, we have $\Psi(Q_1) \le \Psi(Q_2)$.*

Proof. The two assumed optimality inequalities are as follows:

$$\mathcal{J}(Q_1) - \lambda_1 \Psi(Q_1) \le \mathcal{J}(Q_2) - \lambda_1 \Psi(Q_2),$$
$$\mathcal{J}(Q_2) - \lambda_2 \Psi(Q_2) \le \mathcal{J}(Q_1) - \lambda_2 \Psi(Q_1).$$

Adding the above two inequalities results in $(\lambda_2 - \lambda_1)(\Psi(Q_2) - \Psi(Q_1)) \ge 0$. Also, taking $\Psi = H$ proves the claim in the main text. □

B.3 Proof of the Spectral Variational Representation in Theorem 1

Proof of the variational part of Theorem 1. Fix Q, write $S = S_Q^\varepsilon$, and define the positive-definite Gibbs state $\Gamma_\Theta = e^{-\Theta} / \operatorname{Tr}(e^{-\Theta})$. Direct expansion of (29) gives

$$\operatorname{Tr}(S\Theta) + \log \operatorname{Tr}(e^{-\Theta}) = \mathsf{h}(S) + D_{\mathrm{mat}}(S \| \Gamma_\Theta). \tag{32}$$

For trace-zero Θ, the left-hand side equals $\ell_\varepsilon(Q, \Theta)$ because $\operatorname{Tr}(S\Theta) = (1 - \varepsilon) \operatorname{Tr}(\Sigma_Q \Theta)$. By the non-negativity of the matrix-based KL-divergence, equality with $H_\varepsilon(Q)$ holds exactly when $\Gamma_\Theta = S$. The unique trace-zero solution is

$$\Theta^\star(Q) = -\log S_Q^\varepsilon + \frac{\operatorname{Tr}(\log S_Q^\varepsilon)}{d} I_d.$$

To check feasibility, pick $L_\varepsilon = \log(d/\varepsilon)$. Every eigenvalue s_i of S lies in $[\varepsilon/d, 1]$, and therefore $\log s_i \in [-L_\varepsilon, 0]$. The eigenvalues of $\Theta^\star$ are $\frac{1}{d}\sum_j \log s_j - \log s_i$, each in $[-L_\varepsilon, L_\varepsilon]$. Hence $\Theta^\star \in \mathfrak{T}_\varepsilon$, proving both the representation and uniqueness. □

B.4 Minimax Equality and Saddle-Point Attainment

The following assumptions make the regularity conditions in Theorem 1 explicit. The convex class $\mathcal{P}$ is a subset of a Hausdorff locally convex topological vector space of measures; $Q \mapsto \Sigma_Q$ is continuous on $\mathcal{P}$; and $\mathcal{J} : \mathcal{P} \to [0, +\infty]$ is proper, convex, and lower semicontinuous. Write $\mathcal{Q} = \operatorname{dom} \mathcal{J}$ and, for $\lambda \ge 0$, set

$$\mathscr{L}_{\varepsilon,\lambda}(Q, \Theta) = \mathcal{J}(Q) - \lambda \ell_\varepsilon(Q, \Theta). \tag{33}$$

All distribution optimizations in (19) are over $\mathcal{Q}$. This restriction removes only distributions with infinite objective.

Proposition 2 (Minimax regularity and attainment)**.** *Under the preceding assumptions, (19) holds and its outer maximum is attained. If $\mathcal{J}$ has compact sublevel sets, the primal minimum and each fixed-Θ inner minimum are also attained. Whenever the primal minimum is attained, any primal optimizer $Q^\star$ and any dual optimizer $\Theta^\star$ satisfy*

$$\mathscr{L}_{\varepsilon,\lambda}(Q^\star, \Theta) \le \mathscr{L}_{\varepsilon,\lambda}(Q^\star, \Theta^\star) \le \mathscr{L}_{\varepsilon,\lambda}(Q, \Theta^\star) \tag{34}$$

for all $Q \in \mathcal{Q}$ and $\Theta \in \mathfrak{T}_\varepsilon$.

Proof. The variational representation already proved gives

$$\mathcal{J}(Q) - \lambda H_\varepsilon(Q) = \max_{\Theta \in \mathfrak{T}_\varepsilon} \mathscr{L}_{\varepsilon,\lambda}(Q, \Theta).$$

The matrix set $\mathfrak{T}_\varepsilon$ is compact and convex. For fixed Θ, the saddle objective is convex and lower semicontinuous in Q. For fixed Q, it is continuous and concave in Θ. To verify the latter claim without a matrix-commutativity assumption, (32) implies

$$\log \operatorname{Tr}(e^{-\Theta}) = \max_{S \in \mathcal{S}_d} \{\mathsf{h}(S) - \operatorname{Tr}(S\Theta)\},$$

with equality at $S = \Gamma_\Theta$. The right-hand side is a supremum of affine functions of Θ and is therefore convex. Therefore, Sion's minimax theorem (Sion, 1958) applies, and the application holds for $(\Theta, Q) \mapsto -\mathscr{L}_{\varepsilon,\lambda}(Q, \Theta)$ with the compact variable Θ on the minimizing side. Negating the resulting equality proves (19).

The dual value $\gamma(\Theta) = \inf_{Q \in \mathcal{Q}} \mathscr{L}_{\varepsilon,\lambda}(Q, \Theta)$ is upper semicontinuous as it is an infimum of continuous functions. Also, it is finite, as ℓ_ε is uniformly bounded on $\mathcal{P} \times \mathfrak{T}_\varepsilon$, $\mathcal{J} \geq 0$, and properness supplies a finite upper bound by evaluation at any $Q_0 \in \mathcal{Q}$. Therefore, γ attains its maximum.

If $\mathcal{J}$ has compact sublevel sets, each objective under consideration is lower semicontinuous and differs from $\mathcal{J}$ by a bounded continuous function. A non-empty sublevel set sufficiently close to its infimum is consequently compact, which gives attainment. Finally, let v be the common minimax value and let $Q^\star, \Theta^\star$ be primal and dual optimizers. Primal optimality gives $\mathscr{L}(Q^\star, \Theta) \leq v$ for every Θ, and dual optimality gives $\mathscr{L}(Q, \Theta^\star) \geq v$ for every Q. At their intersection equality holds, proving (34). □

C KL-Anchored Targets and the Entropy Wall

Throughout this section, $P = P_{\mathrm{ref}}$ and unconstrained optimizations are over $\mathbb{P}(\mathcal{X})$. For $Q \ll P$, we write $r = dQ/dP$ and $D_{\mathrm{KL}}(Q\|P) = \int r \log r \, dP$; otherwise, $D_{\mathrm{KL}}(Q\|P) = +\infty$. We retain the fixed normalized finite-dimensional representation from the preliminaries.

C.1 Compactness and an Optimality Inequality

Lemma 4 (KL compactness and covariance continuity)**.** *Define the set P-absolutely-continuous probability distributions as $\mathfrak{D}_P = \{r \in L^1(P) : r \geq 0, \int r \, dP = 1\}$, equipped with $\sigma(L^1(P), L^\infty(P))$. Relative entropy is weakly lower semicontinuous, strictly convex on its finite domain, and has weakly compact sublevel sets. The covariance map and the functionals H, H_ε, and C_{RKE} are weakly continuous on $\mathfrak{D}_P$, with $\varepsilon \in (0, 1)$ fixed.*

Proof. We first note that $\mathfrak{D}_P$ is norm-closed and convex, and therefore is weakly closed. Fatou's lemma, applied along an almost-everywhere convergent subsequence, gives norm lower semicontinuity of relative entropy. Convexity then gives weak lower semicontinuity, while strict convexity follows from that of $u \log u$. Since this function is bounded below and grows superlinearly, every KL sublevel is uniformly integrable. The Dunford-Pettis theorem and weak closedness therefore show the weak compactness (Dunford & Pettis, 1940).

Finally, each covariance entry is the bounded linear functional $r \mapsto \int r\phi_i\phi_j \, dP$, since $\phi_i\phi_j \in L^\infty(P)$. The remaining claims follow from continuity of the spectral functions on the finite-dimensional set of unit-trace positive semidefinite matrices. □

Lemma 5 (KL optimality gap)**.** *Let $\mathcal{C} \subseteq \{Q : Q \ll P\}$ be convex and $\Psi : \mathcal{C} \to \mathbb{R}$ be concave. Suppose $Q^\star$ minimizes $G(Q) = D_{\mathrm{KL}}(Q\|P) - \lambda\Psi(Q)$ over $\mathcal{C}$, where $\lambda \geq 0$, with finite optimal value. Then every $R \in \mathcal{C}$ with finite $D_{\mathrm{KL}}(R\|P)$ satisfies*

$$G(R) - G(Q^\star) \geq D_{\mathrm{KL}}(R\|Q^\star). \tag{35}$$

Proof. We consider $Q_t = (1-t)Q^\star + tR \in \mathcal{C}$ for $0 < t < 1$. The KL mixture identity and concavity of Ψ give

$$G(Q_t) \leq (1-t)G(Q^\star) + tG(R) - (1-t)D_{\mathrm{KL}}(Q^\star\|Q_t) - tD_{\mathrm{KL}}(R\|Q_t).$$

All terms are finite because $Q^\star, R$ have finite KL to P, and Q_t dominates each by a positive multiple. Using $G(Q_t) \geq G(Q^\star)$ and dropping a nonnegative term, we obtain $G(R) - G(Q^\star) \geq D_{\mathrm{KL}}(R\|Q_t)$. As t tends to 0, Q_t converges to $Q^\star$ in the sense of TV distance, and hence the lower semicontinuity of KL in its second argument proves the claim. □

C.2 The Canonical Constrained Projection

Proposition 3 (Constrained VNE projection). *Suppose $\mathcal{C}_\rho = \{Q \in \mathbb{P}(\mathcal{X}) : H(Q) \geq \rho\}$ contains a distribution with finite KL to P. Then minimizing $D_{\mathrm{KL}}(Q\|P)$ over $\mathcal{C}_\rho$ has a unique solution Q_ρ, and every feasible R with finite KL satisfies*

$$D_{\mathrm{KL}}(R\|P) \geq D_{\mathrm{KL}}(R\|Q_\rho) + D_{\mathrm{KL}}(Q_\rho\|P). \tag{36}$$

If some Q° additionally satisfies $D_{\mathrm{KL}}(Q^\circ\|P) < \infty$ and $H(Q^\circ) > \rho$, there exists a finite $\lambda^\star \geq 0$ such that Q_ρ minimizes $D_{\mathrm{KL}}(Q\|P) - \lambda^\star H(Q)$ over all probability distributions and $\lambda^\star(H(Q_\rho) - \rho) = 0$. Moreover, $Q_\rho = P$ when $\rho \leq H(P)$, whereas $H(Q_\rho) = \rho$ when $\rho > H(P)$.

Proof. We verify the hypotheses of the information-projection theorem. Concavity of H makes $\mathcal{C}_\rho$ convex. It is also closed in total variation: boundedness of $\phi_i\phi_j$ gives convergence of the covariance entries, and continuity of matrix entropy preserves the constraint. Therefore, finite-KL feasibility and Csiszár & Matus (2003, Theorem 1 and Remarks 1–2) give existence, uniqueness, and (36). Under strict feasibility, Lemma 2 additionally gives the multiplier representation and complementary slackness.

If $\rho \leq H(P)$, the feasible distribution P uniquely minimizes KL with value zero. Otherwise, suppose $H(Q_\rho) > \rho$. For sufficiently small $t > 0$, continuity makes $(1-t)Q_\rho + tP$ feasible, while

$$D_{\mathrm{KL}}\big((1-t)Q_\rho + tP\|P\big) \leq (1-t)D_{\mathrm{KL}}(Q_\rho\|P) < D_{\mathrm{KL}}(Q_\rho\|P).$$

Here $Q_\rho \neq P$, and therefore its finite KL is positive. This contradiction proves $H(Q_\rho) = \rho$. □

If $D_{\mathrm{KL}}(P_{\mathrm{data}}\|P) < \infty$ and $\rho \leq \rho_\star = H(P_{\mathrm{data}})$, setting $R = P_{\mathrm{data}}$ gives (14) without strict feasibility. Nontrivial repair occurs for $H(P) < \rho \leq \rho_\star$, where the constraint is active. Above $\rho_\star$, every feasible target exceeds the data VNE, provided finite-KL feasibility remains possible. These conclusions concern the unsmoothed entropy H.

C.3 First Variation of Smoothed VNE

For positive-definite unit-trace matrices S, T, we write $D_{\mathrm{mat}}(S\|T) = \mathrm{Tr}[S(\log S - \log T)]$.

Lemma 6 (First variation and exact entropy remainder). *Let $0 < \varepsilon < 1$ and let ν be a finite signed measure with $\nu(\mathcal{X}) = 0$ such that $Q + t\nu$ is a probability distribution for sufficiently small $t \geq 0$. Then*

$$\frac{d}{dt} H_\varepsilon(Q + t\nu)\Big|_{t=0+} = \int V_Q^\varepsilon \, d\nu. \tag{37}$$

For any probability distributions Q, R, the supporting inequality has the exact remainder

$$H_\varepsilon(R) + \int V_R^\varepsilon \, d(Q - R) - H_\varepsilon(Q) = D_{\mathrm{mat}}(S_Q^\varepsilon \| S_R^\varepsilon) \geq 0. \tag{38}$$

Proof. We define $\Delta_\nu = \int \phi\phi^\top \, d\nu$. Normalization gives $\mathrm{Tr}(\Delta_\nu) = 0$, and $S_{Q+t\nu}^\varepsilon = S_Q^\varepsilon + t(1-\varepsilon)\Delta_\nu$. The differential of trace entropy therefore yields

$$\frac{d}{dt} H_\varepsilon(Q + t\nu)\Big|_{t=0+} = -(1-\varepsilon)\,\mathrm{Tr}\big[(\log S_Q^\varepsilon + I_d)\Delta_\nu\big] \;=\; \int V_Q^\varepsilon \, d\nu,$$

where the identity term vanishes because Δ_ν has zero trace. For the remainder, we use $S_Q^\varepsilon - S_R^\varepsilon = (1-\varepsilon)(\Sigma_Q - \Sigma_R)$ to obtain

$$\int V_R^\varepsilon \, d(Q - R) = -\,\mathrm{Tr}\big[(S_Q^\varepsilon - S_R^\varepsilon)\log S_R^\varepsilon\big].$$

Substitution into (38) gives the stated identity, and the non-negativity follows from Klein's inequality. The proof is hence complete. □

C.4 The Self-Consistent KL Target

Lemma 7. *For bounded measurable $f : \mathcal{X} \to \mathbb{R}$, define $Z_f = \mathbb{E}_P[e^f]$ and $dQ_f/dP = e^f/Z_f$. Then, for every probability distribution Q,*

$$D_{\mathrm{KL}}(Q\|P) - \mathbb{E}_Q[f] = D_{\mathrm{KL}}(Q\|Q_f) - \log Z_f. \tag{39}$$

Consequently, the left-hand side is uniquely minimized by Q_f.

Proof. Since f is bounded, $Q_f \sim P$ and $\log(dQ_f/dP) = f - \log Z_f$ is bounded. Substituting this identity into $D_{\mathrm{KL}}(Q\|Q_f)$ gives (39) for $Q \ll P$, including infinite-KL cases; otherwise both divergences are infinite. The minimization claim follows from nonnegativity of KL, with equality only at $Q = Q_f$. □

Proof of Proposition 1. We fix $\varepsilon \in (0, 1)$ and $\lambda \geq 0$, and write $\mathcal{F}_{\varepsilon,\lambda}(Q) = D_{\mathrm{KL}}(Q\|P) - \lambda H_\varepsilon(Q)$. Since H_ε is bounded and weakly continuous, the objective is weakly lower semicontinuous and its sublevel sets lie in KL sublevel sets. Lemma 4, together with the finite-objective candidate P, gives attainment. Strict convexity of KL and concavity of H_ε give uniqueness.

We next apply Proposition 2 on $\mathfrak{D}_P$ with its weak $L^1(P)$ topology. For $\lambda > 0$, the resulting saddle point has spectral component $\Theta^\star(Q^\star_\varepsilon)$. Fixing this matrix, the distribution-dependent terms are

$$D_{\mathrm{KL}}(Q\|P) - \lambda(1-\varepsilon)\mathbb{E}_Q\big[\phi^\top \Theta^\star(Q^\star_\varepsilon)\phi\big].$$

The scalar identity component contributes only a constant because $\|\phi(x)\|_2 = 1$. Hence Lemma 7, with $f = \lambda V^\varepsilon_{Q^\star_\varepsilon}$, gives (21). For $\lambda = 0$, the optimizer is P and the formula holds directly.

For the converse, let $Q^\star_\varepsilon$ be any distribution satisfying (21). Its bounded potential gives a bounded log-density ratio and finite KL. Combining the logarithm of the tilt with (38), we obtain, for every Q with finite KL to P,

$$\begin{aligned}\mathcal{F}_{\varepsilon,\lambda}(Q) - \mathcal{F}_{\varepsilon,\lambda}(Q^\star_\varepsilon) = D_{\mathrm{KL}}(Q\|Q^\star_\varepsilon)\\ + \lambda D_{\mathrm{mat}}(S^\varepsilon_Q \| S^\varepsilon_{Q^\star_\varepsilon}).\end{aligned} \tag{40}$$

Both terms are nonnegative, and the first vanishes only when $Q = Q^\star_\varepsilon$. distributions with infinite KL have infinite objective, proving the converse and the stronger optimality claim.

Finally, the eigenvalues of S^ε_Q lie in $[\varepsilon/d, 1]$, and hence $0 \leq V^\varepsilon_Q(x) \leq (1-\varepsilon)\log(d/\varepsilon)$. With $w_\varepsilon = \lambda(1-\varepsilon)\log(d/\varepsilon)$, normalizing the exponential weight gives

$$e^{-w_\varepsilon} \leq \frac{dQ^\star_\varepsilon}{dP} \leq e^{w_\varepsilon} \qquad P\text{-a.s.} \tag{41}$$

In particular, $Q^\star_\varepsilon \sim P$. □

C.5 Relation to the Unsmoothed Target

Proposition 4. *For fixed $\lambda \geq 0$, the objective $\mathcal{F}_{0,\lambda}(Q) = D_{\mathrm{KL}}(Q\|P) - \lambda H(Q)$ has a unique optimizer $Q^\star$. For $0 < \varepsilon < 1$, its smoothed counterpart satisfies*

$$D_{\mathrm{KL}}(Q^\star_\varepsilon\|Q^\star) \leq \mathcal{F}_{0,\lambda}(Q^\star_\varepsilon) - \mathcal{F}_{0,\lambda}(Q^\star) \leq \lambda b_\varepsilon, \tag{42}$$

where $b_\varepsilon = h_{\mathrm{b}}(\varepsilon) + \varepsilon \log d$ is the bound from Lemma 1 and $h_{\mathrm{b}}(u) = -u\log u - (1-u)\log(1-u)$.

Proof. Existence and uniqueness follow as in the smoothed case, using boundedness, concavity, and weak continuity of H. Lemma 5, applied with $\Psi = H$ and $R = Q^\star_\varepsilon$, gives the first inequality. For the second, we define $e_\varepsilon(Q) = H_\varepsilon(Q) - H(Q)$. By Lemma 1, $0 \leq e_\varepsilon(Q) \leq b_\varepsilon$, and therefore the optimality of $Q^\star_\varepsilon$ yields

$$\mathcal{F}_{0,\lambda}(Q^\star_\varepsilon) - \mathcal{F}_{0,\lambda}(Q^\star) \;\leq\; \lambda\big[e_\varepsilon(Q^\star_\varepsilon) - e_\varepsilon(Q^\star)\big] \;\leq\; \lambda b_\varepsilon.$$

Since $b_\varepsilon \to 0$, this proves convergence in the stated KL orientation for fixed λ, without asserting convergence of spatial scores. □

C.6 Proof of Corollary 1

Proof of Corollary 1. We work on an open set $U \subseteq \mathbb{R}^D$ contained in $\mathcal{X}$, where P has a strictly positive C^1 Lebesgue density p and ϕ is C^1. The tilt defines the positive C^1 density version $q_\varepsilon^\star(x) = Z^{-1}p(x)\exp\{\lambda V_{Q_\varepsilon^\star}^\varepsilon(x)\}$ on U. Taking its logarithmic gradient gives (22), since the covariance and normalizing constant do not depend on x.

For fixed symmetric M, we have $\nabla_x(\phi^\top M\phi) = 2D\phi^\top M\phi$, with all terms evaluated at x. Substituting $M = -(1-\varepsilon)\log S_Q^\varepsilon$ gives (23). Finally, differentiating $\phi^\top\phi = 1$ yields $D\phi^\top\phi = 0$; hence, scalar identity shifts of M do not affect the score. □

D Order-2 RKE: Projection and Target Distribution

This section proves the extension in Section 4.3, using the preliminaries' definitions of H_{RKE} and C_{RKE}. Also, we write $\mathcal{R}_{\mathrm{RKE}}(Q) = -C_{\mathrm{RKE}}(Q)$ which is the concave reward function.

Proposition 5 (RKE geometry and variational representation)**.** *The functional C_{RKE} is convex, and H_{RKE} has convex superlevel sets. With $\mathfrak{A} = \{A \in \mathbb{R}^{d\times d}_{\mathrm{sym}} : \|A\|_F \le 1\}$, we have*

$$-C_{\mathrm{RKE}}(Q) = \min_{A\in\mathfrak{A}} \Big\{ \|A\|_F^2 - 2\mathbb{E}_{x\sim Q}[\phi(x)^\top A\phi(x)] \Big\}, \tag{43}$$

and the minimizer is uniquely $A^\star(Q) = \Sigma_Q$. For every admissible mass-preserving signed perturbation ν,

$$\frac{d}{dt}\mathcal{R}_{\mathrm{RKE}}(Q+t\nu)\Big|_{t=0+} = \int V_{\mathrm{RKE},Q}\, d\nu, \tag{44}$$

where $V_{\mathrm{RKE},Q}$ is the potential in (27).

Proof. The affine map $Q \mapsto \Sigma_Q$ followed by the squared Frobenius norm is convex. Since $H_{\mathrm{RKE}}(Q) \ge \rho$ if and only if $C_{\mathrm{RKE}}(Q) \le e^{-\rho}$, its superlevel sets are convex. Completing the square gives

$$\|A\|_F^2 - 2\,\mathrm{Tr}(A\Sigma_Q) = \|A-\Sigma_Q\|_F^2 - \|\Sigma_Q\|_F^2 .$$

The covariance belongs to $\mathfrak{A}$ because it is positive semidefinite with unit trace. This proves (43) and uniqueness. For $\Delta_\nu = \int \phi\phi^\top\, d\nu$, direct differentiation provides the following

$$\frac{d}{dt}[-\,\mathrm{Tr}((\Sigma_Q + t\Delta_\nu)^2)]\Big|_{t=0+} = -2\,\mathrm{Tr}(\Sigma_Q\Delta_\nu) = \int V_{\mathrm{RKE},Q}\, d\nu.$$

The equality $\phi(x)^\top\Sigma_Q\phi(x) = \mathbb{E}_{X'\sim Q}[k(x,X')^2]$ identifies its pairwise form. For later use, completion of the square also gives the exact supporting remainder

$$\mathcal{R}_{\mathrm{RKE}}(R) + \int V_{\mathrm{RKE},R}\, d(Q-R) - \mathcal{R}_{\mathrm{RKE}}(Q) = \|\Sigma_Q - \Sigma_R\|_F^2 . \tag{45}$$

The above finishes the proof. □

Specifically, the first variation of H_{RKE} itself is represented by $V_{\mathrm{RKE},Q}/C_{\mathrm{RKE}}(Q)$. The potential in the main text is instead the variation of $-C_{\mathrm{RKE}}$, because the convex constraint is expressed in terms of purity. The associated multiplier therefore belongs to the purity constraint; it is not a fixed linear penalty on H_{RKE}.

Proposition 6 (KL-anchored RKE target)**.** *For every $P \in \mathbb{P}(\mathcal{X})$ and $\lambda \ge 0$, the objective $\mathcal{F}_{\mathrm{RKE},\lambda}(Q) = D_{\mathrm{KL}}(Q\|P) + \lambda C_{\mathrm{RKE}}(Q)$ has a unique minimizer $Q^\star_{\mathrm{RKE}}$. It satisfies (28), and any distribution satisfying that equation is the minimizer. Moreover, the following holds P-almost sure*

$$e^{-2\lambda} \le \frac{dQ^\star_{\mathrm{RKE}}}{dP} \le e^{2\lambda}, \tag{46}$$

and, for every Q with finite KL, we have the following

$$\mathcal{F}_{\mathrm{RKE},\lambda}(Q) - \mathcal{F}_{\mathrm{RKE},\lambda}(Q^\star_{\mathrm{RKE}}) = D_{\mathrm{KL}}(Q\|Q^\star_{\mathrm{RKE}}) + \lambda\left\|\Sigma_Q - \Sigma_{Q^\star_{\mathrm{RKE}}}\right\|_F^2 . \tag{47}$$

Proof. Since $\frac{1}{d} \le C_{\mathrm{RKE}} \le 1$, a minimizing sequence has bounded KL. Lemma 4 and continuity of C_{RKE} give existence. Strict convexity of KL and convexity of C_{RKE} give uniqueness.

Put $\ell_{\mathrm{RKE}}(Q, A) = \|A\|_F^2 - 2\,\mathrm{Tr}(A\Sigma_Q)$. By (43),

$$\mathcal{F}_{\mathrm{RKE},\lambda}(Q) = \max_{A \in \mathfrak{A}} \{D_{\mathrm{KL}}(Q\|P) - \lambda \ell_{\mathrm{RKE}}(Q, A)\}.$$

The expression in braces is weakly lower semicontinuous and convex in the density of Q, and continuous and concave in A. The set $\mathfrak{A}$ is compact and convex, and $|\ell_{\mathrm{RKE}}| \le 3$. Sion's theorem and compact KL sublevels therefore give an attained primal–dual saddle, by the same argument as Proposition 2. For $\lambda > 0$, its unique spectral best response is $A = \Sigma_{Q^\star_{\mathrm{RKE}}}$. At this matrix, minimization in Q reduces, up to a constant, to

$$D_{\mathrm{KL}}(Q\|P) + 2\lambda \mathbb{E}_Q[\phi^\top \Sigma_{Q^\star_{\mathrm{RKE}}} \phi].$$

Lemma 7 gives exactly (28). The case $\lambda = 0$ again gives $Q^\star_{\mathrm{RKE}} = P$.

Conversely, the logarithm of any self-consistent tilt and (45) give (47) by the same three-point expansion used in (40). Explicitly, the linear term from the log-density ratio is $-2\lambda\,\mathrm{Tr}[\Sigma_{Q^\star_{\mathrm{RKE}}}(\Sigma_Q - \Sigma_{Q^\star_{\mathrm{RKE}}})]$; adding the purity difference leaves exactly the squared norm in (47). This proves global optimality and the converse. Finally, $-2 \le V_{\mathrm{RKE},Q} \le 0$, and as a result normalization of its exponential gives (46). □

For the constrained RKE problem, finite-KL feasibility is sufficient for existence, uniqueness, and the Pythagorean inequality: use weak continuity of C_{RKE}, its convex sublevel sets, and Lemma 5. Under strict finite-KL feasibility, Lemma 2 with $\Psi = -C_{\mathrm{RKE}}$ and $\tau = -e^{-\rho}$ supplies a finite multiplier. The Lagrangian is $D_{\mathrm{KL}}(Q\|P) + \lambda(C_{\mathrm{RKE}}(Q) - e^{-\rho})$, whose last term is constant in Q.

The data-relative comparison uses $H_{\mathrm{RKE}}(P_{\mathrm{data}})$ on the RKE scale. It does not, by itself, establish that the canonical VNE Entropy Wall has been crossed, since the two entropies need not order different covariance spectra identically.

E Exact Propagation of the Target under Forward Noising

The noisy-score relation requires no differentiation through the target covariance. It follows from propagating a fixed likelihood ratio through the forward kernel. We state the result for any bounded log-weight, and as a result, it applies to both spectral potentials and to a frozen pilot potential.

Lemma 8 (Propagation of a bounded exponential tilt)**.** *Let P be a probability distribution on a standard Borel space, let $\Pi_t(dx_t \mid x_0)$ be a Markov kernel to another standard Borel space, and let f be bounded and measurable. Define $dQ_f/dP = e^f/Z_f$, where $Z_f = \mathbb{E}_P[e^f]$, and set $P_t = P\Pi_t$, $Q_{f,t} = Q_f\Pi_t$. Under the reference joint distribution, let $h_t(x_t) = \mathbb{E}_P[e^{f(X_0)} \mid X_t = x_t]$. Then*

$$\frac{dQ_{f,t}}{dP_t}(x_t) = \frac{h_t(x_t)}{Z_f} \qquad P_t\text{-a.s.} \tag{48}$$

If P_t and $Q_{f,t}$ have positive C^1 densities on an open set $U \subseteq \mathbb{R}^D$, the version $h_t = Z_f q_{f,t}/p_t$ is positive and C^1 there and satisfies

$$\nabla_{x_t} \log q_{f,t}(x_t) = \nabla_{x_t} \log p_t(x_t) + \nabla_{x_t} \log h_t(x_t), \qquad x_t \in U. \tag{49}$$

Proof. For any bounded measurable test function g, the tower property gives

$$\begin{aligned}\int g(x_t)\, Q_{f,t}(dx_t) &= \frac{1}{Z_f}\mathbb{E}_P[g(X_t)e^{f(X_0)}] \\ &= \frac{1}{Z_f}\mathbb{E}_P[g(X_t)h_t(X_t)] \\ &= \int g(x_t)\frac{h_t(x_t)}{Z_f}\, P_t(dx_t).\end{aligned}$$

This proves (48). Boundedness of f gives positive finite lower and upper bounds for h_t, up to P_t-null sets. On U, choosing $h_t = Z_f q_{f,t}/p_t$ and differentiating its logarithm proves (49). This step does not interchange differentiation and conditional expectation. □

Taking $f = \lambda V^{\varepsilon}_{Q^\star_\varepsilon}$ gives $Q_{f,t} = Q^\star_{\varepsilon,t}$ and proves (24). Taking $f = \lambda V_{\mathrm{RKE},Q^\star_{\mathrm{RKE}}}$ gives the RKE counterpart. For Gaussian forward noising with positive variance, the marginals are positive smooth Gaussian mixtures; hence, the density assumptions hold. At zero noise, the clean-score regularity conditions must be checked separately.

F Implementation and Approximation Details for IGA Guidance

F.1 Finite-Sample Pilot Optimization

For VNE, we fit a potential on a fixed, independent pilot rather than on the trajectories being evaluated. Writing $z_i = \phi(x_i)$ and $C_q = \sum_i q_i z_i z_i^\top$, the finite-support problem minimizes

$$\sum_{i=1}^{N} q_i \log(N q_i) + \lambda \operatorname{Tr}\big[\mathcal{S}_\varepsilon(C_q) \log \mathcal{S}_\varepsilon(C_q)\big] \quad \text{over } q \in \Delta_N, \qquad \Delta_N = \{q \in \mathbb{R}^N_+ : \textstyle\sum_i q_i = 1\}.$$

Algorithm 1 iterates its self-consistency equations, where $\mathcal{S}_\varepsilon$ is the smoothing map defined in the algorithm.

Algorithm 1 Fixed-point pilot estimation for IGA Guidance

Require: Base sampler P_θ, unit-normalized feature map $\phi : \mathcal{X} \to \mathbb{R}^d$, $\lambda \geq 0$, $0 < \varepsilon < 1$
Require: Pilot size $N \geq 1$, damping $0 < \gamma \leq 1$, tolerance $\delta > 0$, iteration limit $K \geq 1$

Draw $x_{1:N} \overset{\text{i.i.d.}}{\sim} P_\theta$ once and set $z_i \leftarrow \phi(x_i)$
Define $\Pi_0(B) = B - \operatorname{Tr}(B) I_d/d$ and $\mathcal{S}_\varepsilon(C) = (1-\varepsilon)C + \varepsilon I_d/d$
$C \leftarrow N^{-1} \sum_i z_i z_i^\top$; $\Theta \leftarrow \Pi_0(-\log \mathcal{S}_\varepsilon(C))$
for $k = 0, \ldots, K-1$ **do**
  $a_i \leftarrow \lambda(1-\varepsilon) z_i^\top \Theta z_i$ for all i
  $q_i \leftarrow \exp(a_i) / \sum_j \exp(a_j)$ for all i
  $C \leftarrow \sum_i q_i z_i z_i^\top$
  $G \leftarrow e^{-\Theta} / \operatorname{Tr}(e^{-\Theta})$
  $r \leftarrow \big\| \Pi_0\big(\mathcal{S}_\varepsilon(C) - G\big) \big\|_{\mathrm{F}}$
  if $r \leq \delta$ **then**
    break
  end if
  $\widetilde{\Theta} \leftarrow \Pi_0(-\log \mathcal{S}_\varepsilon(C))$
  $\Theta \leftarrow \Pi_0\big((1-\gamma)\Theta + \gamma \widetilde{\Theta}\big)$
end for
$a_i \leftarrow \lambda(1-\varepsilon) z_i^\top \Theta z_i$ for all i; $q_i \leftarrow \exp(a_i) / \sum_j \exp(a_j)$ for all i
return (q, Θ)

The resulting VNE matrix is $M = (1-\varepsilon)\Theta$; trace centering changes the per-sample potential only by a constant because $\|\phi(x)\|_2 = 1$. A small residual indicates approximate finite-support stationarity, but the damped iteration has no general convergence guarantee; direct finite-dimensional optimization remains an alternative if it fails to converge.

F.2 Frozen Potentials and Conditional Independence

Let $\mathcal{B} = \{\widetilde{x}_j\}_{j=1}^N$ be a nonempty pilot, constructed separately from the output trajectories, and let $\widehat{\Sigma}$ be its fitted weighted covariance. The two frozen matrices are

$$M_{\mathrm{VNE}} = -(1-\varepsilon) \log\big((1-\varepsilon)\widehat{\Sigma} + \varepsilon I_d/d\big), \quad M_{\mathrm{RKE}} = -2\widehat{\Sigma}. \tag{50}$$

For either choice, write $\widehat{V}(x) = \phi(x)^\top M \phi(x)$. Conditional on the pilot, this specifies the exact frozen-potential distribution

$$\frac{d\widehat{Q}}{dP}(x) = \frac{e^{\lambda \widehat{V}(x)}}{\mathbb{E}_P[e^{\lambda \widehat{V}(X)}]}. \tag{51}$$

It does not follow that the practical diffusion sampler draws exactly from $\widehat{Q}$.

For comparison, a one-step shortcut sets $\widehat{\Sigma}$ to the uniform pilot covariance. Replacing the covariance in either population potential by a candidate $\Sigma \in \mathcal{S}_d$, exponentially tilting P, and taking the resulting covariance defines a map $\mathcal{T}(\Sigma)$. Propositions 1 and 6 identify the exact target covariance as its fixed point. The uniform pilot estimates Σ_P and yields only the first iterate $\mathcal{T}(\Sigma_P)$ in the population limit. The finite-support fit instead reweights the pilot; no convergence of its damped fixed-point iteration is assumed without checking the residual.

For a fixed condition, fixed pilot, and identical sampler settings, independent initializations and independent sampler randomness produce conditionally i.i.d. outputs from the implemented sampler's distribution. If the shared pilot is random and then marginalized out, the joint distribution is a mixture of such product distributions and need not be independent. This dependence issue is separate from approximation of $Q^\star_\varepsilon$ or $\widehat{Q}$.

F.3 Denoised-Point and Jacobian Approximations

The exact noisy correction for (51) is $\nabla_{x_t} \log \widehat{h}_t(x_t)$, where $\widehat{h}_t(x_t) = \mathbb{E}_P[e^{\lambda \widehat{V}(X_0)} \mid X_t = x_t]$. IGA evaluates the log-weight at the denoised prediction, replacing $\log \widehat{h}_t(x_t)$ by $\lambda \widehat{V}(\widehat{x}_0(x_t, t))$. Therefore, covariance freezing and posterior approximation are distinct steps.

Jensen's inequality gives $\log \widehat{h}_t \geq \lambda \mathbb{E}_P[\widehat{V}(X_0) \mid X_t]$, but supplies no general comparison with $\lambda \widehat{V}(\mathbb{E}_P[X_0 \mid X_t])$ when that posterior mean exists. Nor does it give a bound for evaluating an approximate denoiser. Even convergence of scalar log-weights would not alone establish convergence of their gradients. We therefore make no universal score-error or small-noise exactness claim for the plug-in rule.

Assume the feature map and denoiser are differentiable at the evaluated points. Holding M fixed, the exact chain rule for this plug-in objective is

$$\begin{aligned} g_t &= 2\lambda D\phi(\widehat{x}_0)^\top M \phi(\widehat{x}_0), \\ \widetilde{u}_t^{\mathrm{chain}} &= D\widehat{x}_0(x_t, t)^\top g_t. \end{aligned} \tag{52}$$

The required products can be evaluated by reverse-mode differentiation without forming dense Jacobians. The lower-cost direct variant sets $\widetilde{u}_t^{\mathrm{direct}} = g_t$, replacing the entire denoised-prediction Jacobian by the identity. In contrast,

$$D\widehat{x}_0(x_t, t) = \frac{I_D - \sqrt{1 - \bar{\alpha}_t}\, D_x \epsilon_\theta(x_t, t)}{\sqrt{\bar{\alpha}_t}}. \tag{53}$$

Stopping gradients only through ϵ_θ would give $I_D/\sqrt{\bar{\alpha}_t}$, not I_D. Neither Jacobian approximation is asserted to recover the exact posterior score correction. Any clipping, gradient detachment, or rescaling used in an implementation changes the stated rule and must be specified with the experiment.

F.4 Noise Parameterization and Sampling Algorithm

For a variance-preserving forward process, write $X_t = \sqrt{\bar{\alpha}_t} X_0 + \sqrt{1 - \bar{\alpha}_t}\, Z$, with $Z \sim \mathcal{N}(0, I_D)$ independent of X_0. Gaussian differentiation gives the exact score identity

$$\nabla \log p_t(x_t) = -\frac{\epsilon^\star_P(x_t, t)}{\sqrt{1 - \bar{\alpha}_t}}, \qquad \epsilon^\star_P(x_t, t) = \mathbb{E}_P[Z \mid X_t = x_t]. \tag{54}$$

Indeed, differentiate the Gaussian likelihood inside $p_t(x_t) = \int \Pi_t(x_t \mid x_0) P(dx_0)$ and divide by $p_t(x_t)$. For positive noise variance the Gaussian density and its first derivatives are bounded functions of the residual, justifying differentiation under the integral locally in x_t. Substitution of the learned approximation $\epsilon_\theta \approx \epsilon^\star_P$ gives the sign and scale of the correction in (26).

Algorithm 2 uses this correction without changing pretrained parameters. The pilot and all randomness for the output trajectory are independent; the matrix remains fixed throughout the trajectory. The main-text update is the chain-rule variant, while the direct variant is explicitly identified as an additional approximation.

Algorithm 2 IGA Guidance with a Frozen Spectral Potential

Require: Noise predictor ϵ_θ; normalized feature map ϕ; nonempty independent pilot $\mathcal{B}$; $\lambda \geq 0$; spectral choice VNE or RKE, with $\varepsilon \in (0,1)$ for VNE.
Require: Diffusion schedule; DDPM or DDIM sampler; variant $v \in \{\text{chain}, \text{direct}\}$.
Fit $\widehat{\Sigma}$ on $\mathcal{B}$ for the chosen entropy objective (Algorithm 1 for VNE).
Form M by (50) and hold it fixed.
Draw $x_T \sim \mathcal{N}(0, I_D)$ independently of $\mathcal{B}$.
for $t = T, \dots, 1$ **do**
  $e_t \leftarrow \epsilon_\theta(x_t, t)$.
  $\widehat{x}_0 \leftarrow (x_t - \sqrt{1-\bar{\alpha}_t}\, e_t)/\sqrt{\bar{\alpha}_t}$.
  $g_t \leftarrow 2\lambda D\phi(\widehat{x}_0)^\top M \phi(\widehat{x}_0)$.
  if $v = \text{chain}$ **then**
    $\widetilde{u}_t \leftarrow D\widehat{x}_0(x_t, t)^\top g_t$.
  else
    $\widetilde{u}_t \leftarrow g_t$.
  end if
  $e_t^{\mathrm{IGA}} \leftarrow e_t - \sqrt{1-\bar{\alpha}_t}\, \widetilde{u}_t$.
  $x_{t-1} \leftarrow \mathrm{SamplerStep}_t(x_t, e_t^{\mathrm{IGA}})$, using fresh sampler randomness.
end for
return x_0 (decoded for latent diffusion).

The exact target at the terminal noise level is $Q^\star_{\varepsilon,T}$, or its RKE counterpart, not necessarily $\mathcal{N}(0, I_D)$. They agree with the Gaussian initialization when forward noising removes all dependence on X_0; at a finite nonzero signal level, terminal initialization is another approximation. Learned-score error, posterior approximation, and reverse-time discretization also affect the sampling distribution. Even an exact noisy-score identity alone does not establish exact sampling by a finite-step DDPM or DDIM implementation.

F.5 DDPM and DDIM Updates

We give the unmodified noise-parameterized sampler formulas, with no clipping of clean predictions (Ho et al., 2020; Song et al., 2021). Let $0 < \beta_t < 1$, $\alpha_t = 1 - \beta_t$, $\bar{\alpha}_t = \prod_{j=1}^t \alpha_j$, and $\bar{\alpha}_0 = 1$.

DDPM. The standard reverse mean and its guided counterpart satisfy

$$\begin{aligned} \mu_\theta(x_t, t) &= \frac{1}{\sqrt{\alpha_t}}\Big(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x_t, t)\Big), \\ \mu_\theta^{\mathrm{IGA}}(x_t, t) &= \mu_\theta(x_t, t) + \frac{\beta_t}{\sqrt{\alpha_t}}\widetilde{u}_t(x_t). \end{aligned} \tag{55}$$

The second line follows by direct substitution of (26) into the first. The update is $x_{t-1} = \mu_\theta^{\mathrm{IGA}}(x_t, t) + \upsilon_t z_t$, where $z_t \sim \mathcal{N}(0, I_D)$ is fresh noise for $t > 1$ and $z_1 = 0$. The chosen reverse variance υ_t^2 is unchanged; a standard choice is $\upsilon_t^2 = \beta_t(1-\bar{\alpha}_{t-1})/(1-\bar{\alpha}_t)$. The mean shift is an exact algebraic identity for this parameterization, not a claim that the target's finite-step reverse conditional is Gaussian with that variance.

DDIM. For $\eta \in [0,1]$, set

$$\begin{aligned} \sigma_t &= \eta\sqrt{\frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}}\sqrt{1-\frac{\bar{\alpha}_t}{\bar{\alpha}_{t-1}}}, \\ \widehat{x}_0^{\mathrm{IGA}} &= \frac{x_t - \sqrt{1-\bar{\alpha}_t}\,\epsilon_\theta^{\mathrm{IGA}}(x_t, t)}{\sqrt{\bar{\alpha}_t}}. \end{aligned} \tag{56}$$

Recompute this clean prediction from the corrected noise prediction, then apply

$$x_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\,\widehat{x}_0^{\mathrm{IGA}} + \sqrt{1-\bar{\alpha}_{t-1}-\sigma_t^2}\,\epsilon_\theta^{\mathrm{IGA}}(x_t, t) + \sigma_t z_t. \tag{57}$$

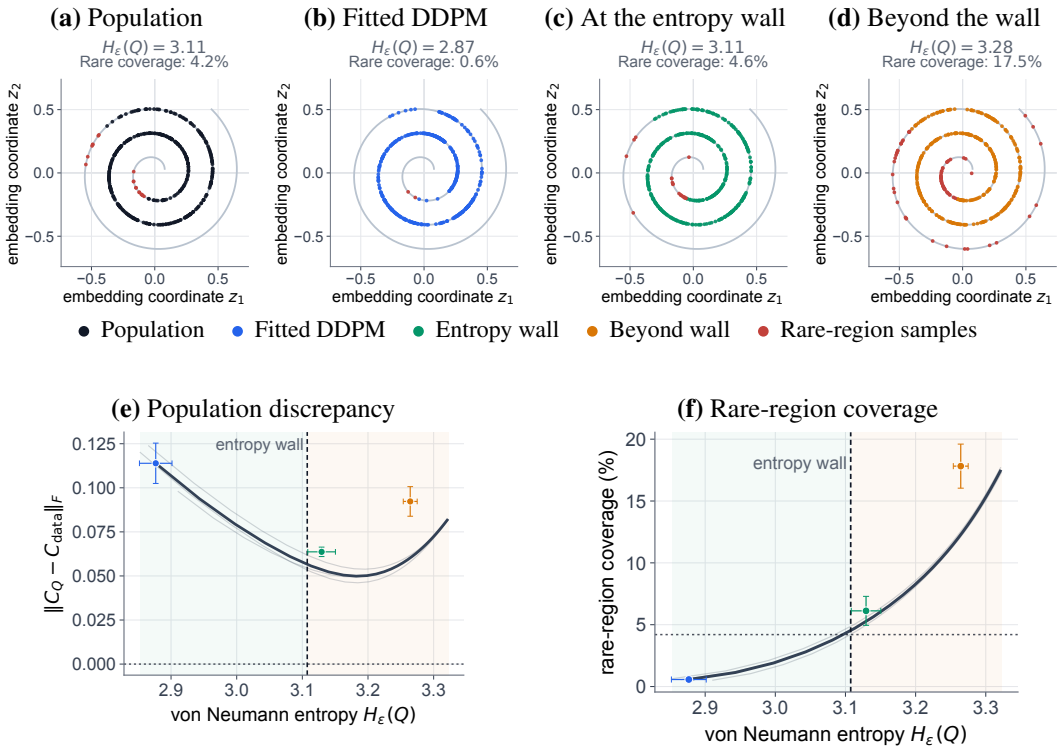


Figure 7: **Repair and extrapolation on a learned nonlinear manifold.** Population entropy, population discrepancy, and rare-region coverage as functions of λ. Rare-region coverage is the fraction of samples outside the central 70% of the normalized manifold coordinate.

The radicand is nonnegative for the stated range of η. Deterministic DDIM corresponds to $\eta = 0$; otherwise $z_t \sim \mathcal{N}(0, I_D)$ is independent across updates. For a subsampled DDIM schedule, replace $t-1$ by the preceding retained noise level in both (56) and (57).

F.6 Normalized Features and Conditional Generation

If $\phi(x) = \psi(x)/\|\psi(x)\|_2$, the differentiability statements require ψ to be differentiable and nonzero at the evaluated points. Its Jacobian is

$$D\phi(x) = \frac{(I_d - \phi(x)\phi(x)^\top) D\psi(x)}{\|\psi(x)\|_2}. \tag{58}$$

This identity explains the tangent relation $D\phi(x)^\top \phi(x) = 0$ and must be respected when differentiating through feature normalization.

For latent diffusion, x_t and $\widehat{x}_0$ are latent variables, and ψ may include decoding and feature extraction. In conditional generation, all target definitions are applied at a fixed condition c, with reference $P(\cdot \mid c)$ and a pilot generated under that same condition. The conditional independence claim concerns repeated generation at fixed c.

G Additional Synthetic Experiments

G.1 Spectral Objectives on a Known Helix

We use an analytically specified helix to isolate how VNE and RKE redistribute probability mass. Unlike the learned manifold experiment in Figure 7, this setting allows us to optimize both objectives exactly, without fitting a diffusion model. The clean support is $x(t) = (\cos t, \sin t, 0.105t)$, for $t \in [-4.2\pi, 4.2\pi]$, and the optimization uses a 2,708-dimensional one-hot spectral representation.

The base distribution concentrates on a few portions of the curve, with a 10^{-4} uniform component giving positive probability throughout the support.

To compare the objectives at the same departure from the base, we choose λ separately for each so that $D_{\mathrm{KL}}(Q\|P_{\mathrm{base}}) = 2$: $\lambda = 25.0913$ for Rényi-2 and $\lambda = 1.0370$ for VNE. Figure 8 shows that RKE redistributes mass mainly among the more populated parts of the helix, whereas VNE gives appreciable mass to portions that the base rarely visits. This agrees with the spectral potentials: RKE penalizes large eigenvalues, while VNE's logarithmic weights place greater emphasis on weak directions. The comparison concerns these different responses at a common KL budget; no separate data distribution is specified from which to define an entropy wall.

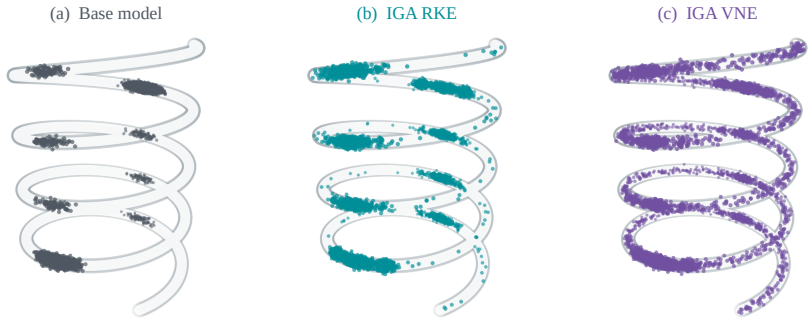


Figure 8: IGA-VNE and IGA-RKE on a known helix at equal KL divergence from the base distribution. The gray curve shows the helix

G.2 Repair and Extrapolation in an Eight-Component Mixture

We use an eight-component mixture with known population weights to test whether IGA repairs the base model's imbalance across modes. At the calibrated wall ($\lambda = 0.4$), guidance nearly recovers those weights (Figure 9). At $\lambda = 3$, the weights become more uniform than the population's: KL divergence to the population falls toward the wall and then rises beyond it, even as entropy continues to increase.

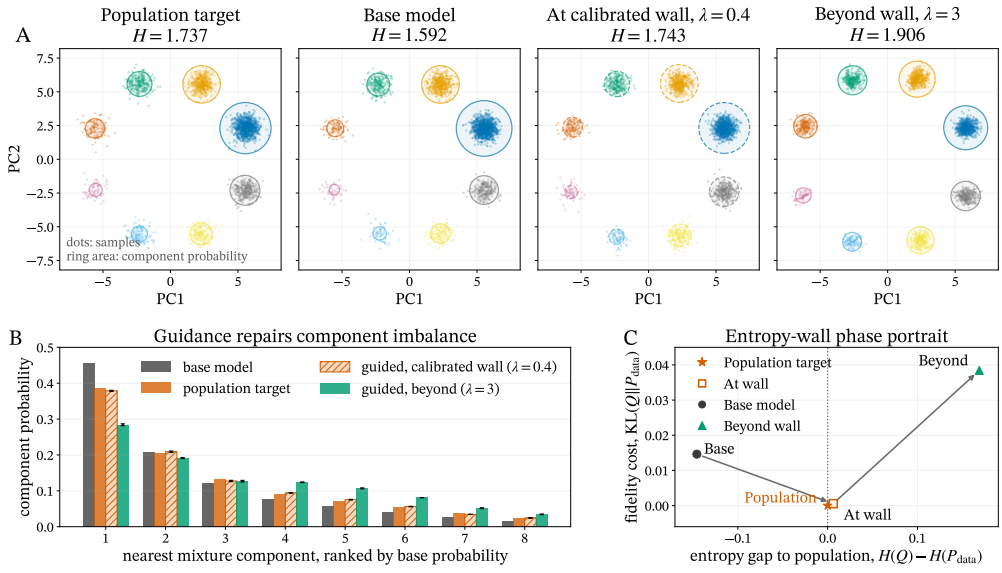


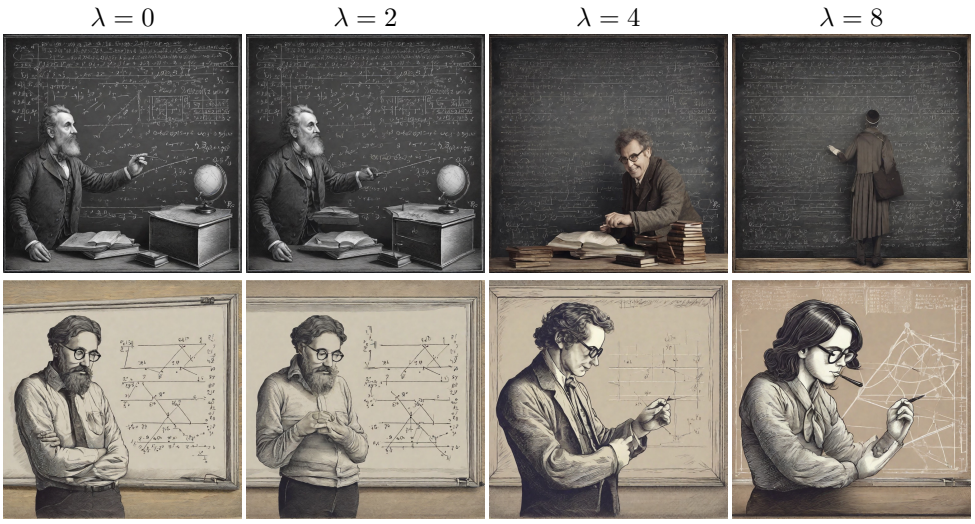
Figure 9: IGA on an eight-component mixture: (A) samples, with ring areas indicating component probabilities; (B) component weights; and (C) KL divergence to the population.

H Additional Qualitative Results

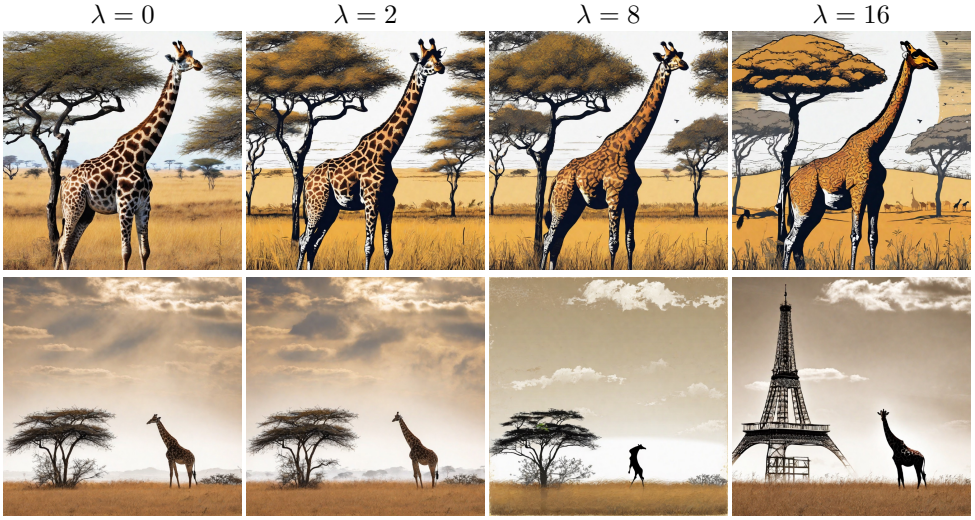
The numerical metrics describe how the generated distribution changes as a whole. Here we examine what those changes look like, first in text-conditioned SDXL and PixArt-Σ (Chen et al., 2024) generation and then on the face and ImageNet benchmarks.

H.1 Prompt-conditioned guidance paths

We first hold the prompt and initial noise fixed so that each sequence shows the effect of changing λ. Figure 10 illustrates this with two prompts. The mathematician scenes retain the blackboard setting while changing the subject's appearance and pose. For the giraffe prompt, one sequence gradually shifts toward illustration, while another introduces a tower beside the animal. These changes arise from the spectral guidance without an instruction to make a particular edit. The examples illustrate the range of possible changes; the aggregate prompt-level evaluation is reported in Appendix J.3.



(a) "A brilliant mathematician at a blackboard"



(b) "A giraffe in the savanna"

Figure 10: **SDXL paths for two fixed prompts.** Within each row, the prompt and initial noise stay fixed; only λ changes.

H.2 GUIDANCE PATHS ACROSS MODELS

To examine whether these effects extend across generators, we apply IGA to SDXL and PixArt-Σ using the same fashion-design prompt. Figure 11 fixes the initial noise within each model and varies λ. Both models introduce larger changes in garment shape, materials, and accessories as the coefficient increases.

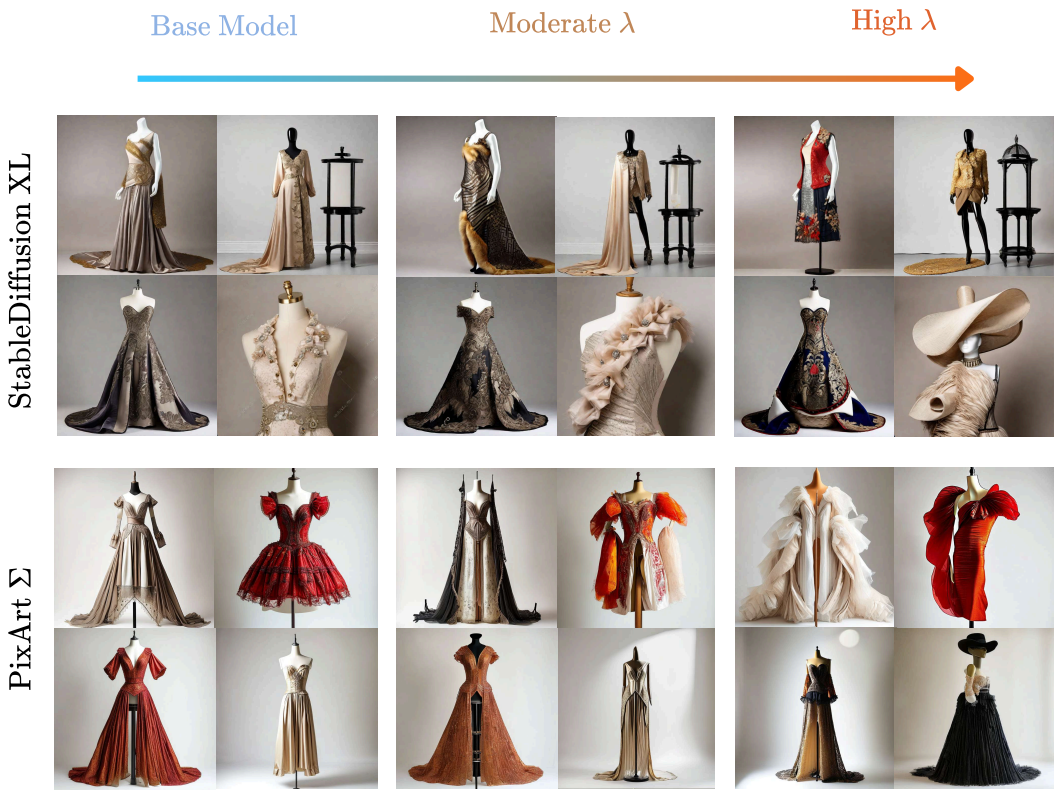


Figure 11: Generation with SDXL (top) and PixArt-Σ (bottom) as λ increases. Corresponding samples within each model use the same prompt and initial noise.

H.3 SDXL COMPARISONS

We compare vanilla and guided SDXL across fashion, architecture, underwater illustration, and fantasy design in Figures 12–15. Matched initial noise makes it possible to inspect the effect of guidance on each example. Across these prompts, the changes include new object shapes, a wider range of materials, and different scene compositions.



Figure 12: Fashion designs from vanilla SDXL (left) and IGA-guided SDXL (right), using the same prompt and matched initial noise.



Figure 13: Architectural designs from vanilla SDXL (left) and IGA-guided SDXL (right), using the same prompt and matched initial noise.

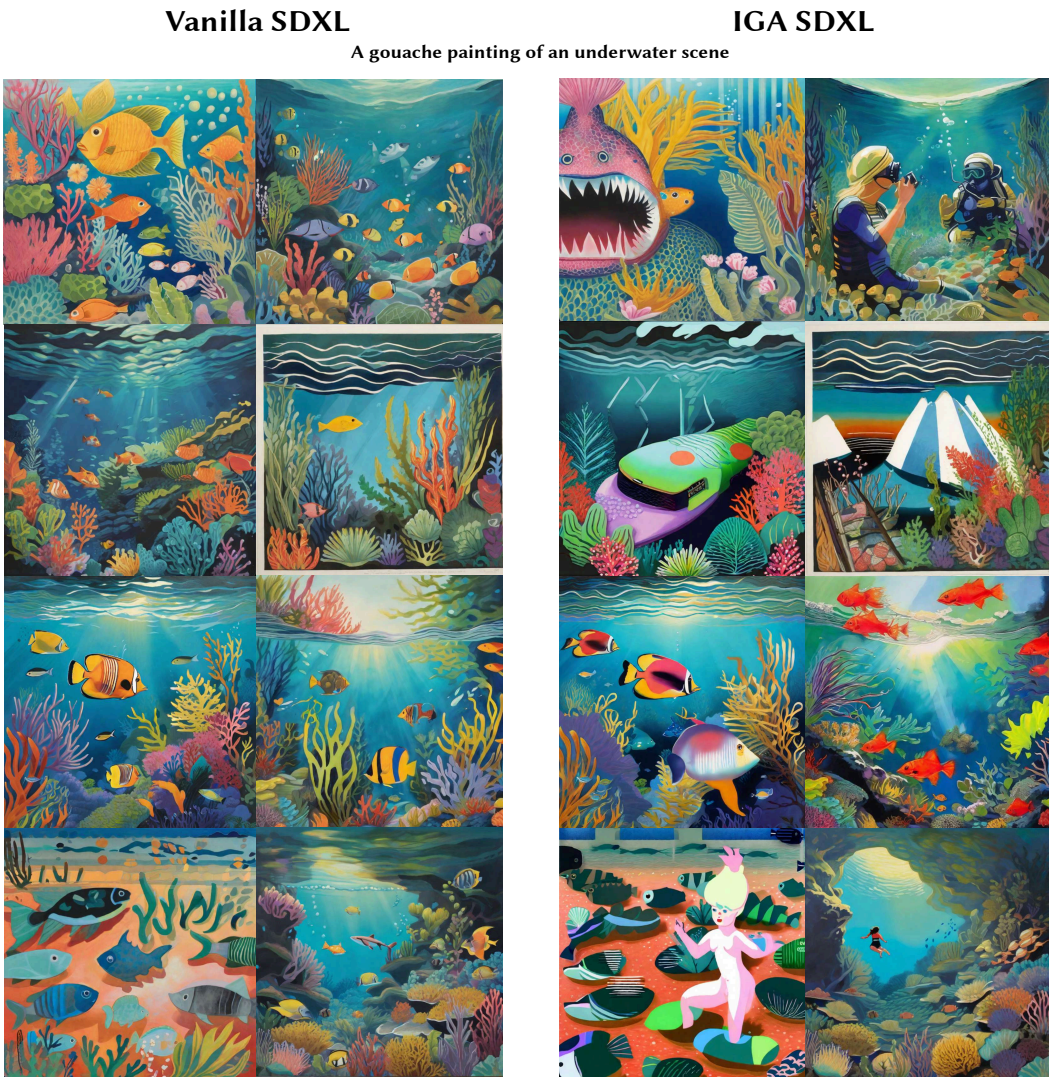


Figure 14: Underwater scenes from vanilla SDXL (left) and IGA-guided SDXL (right), using the same prompt and matched initial noise.



Figure 15: Fantasy throne designs from vanilla SDXL (left) and IGA-guided SDXL (right), using the same prompt and matched initial noise.

H.4 PIXART-Σ COMPARISONS

We repeat the four prompt comparisons with PixArt-Σ in Figures 16–19. The guided samples show similar changes in form and composition across this second model. We use $\lambda = 20$, except for the underwater comparison, which uses $\lambda = 40$.



Figure 16: Fashion designs from vanilla PixArt-Σ (left) and IGA-guided PixArt-Σ with $\lambda = 20$ (right).

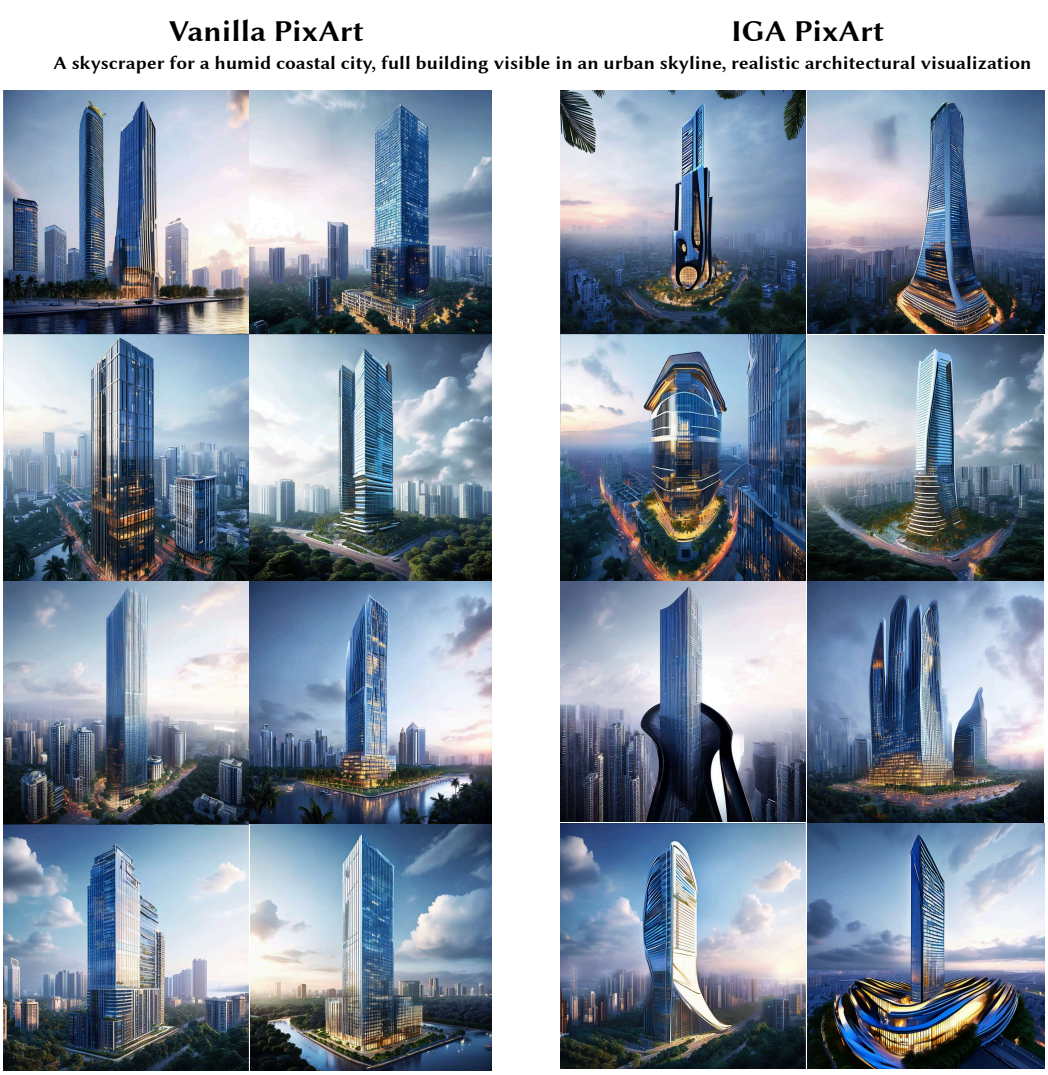


Figure 17: Architectural designs from vanilla PixArt-Σ (left) and IGA-guided PixArt-Σ with $\lambda = 20$ (right).

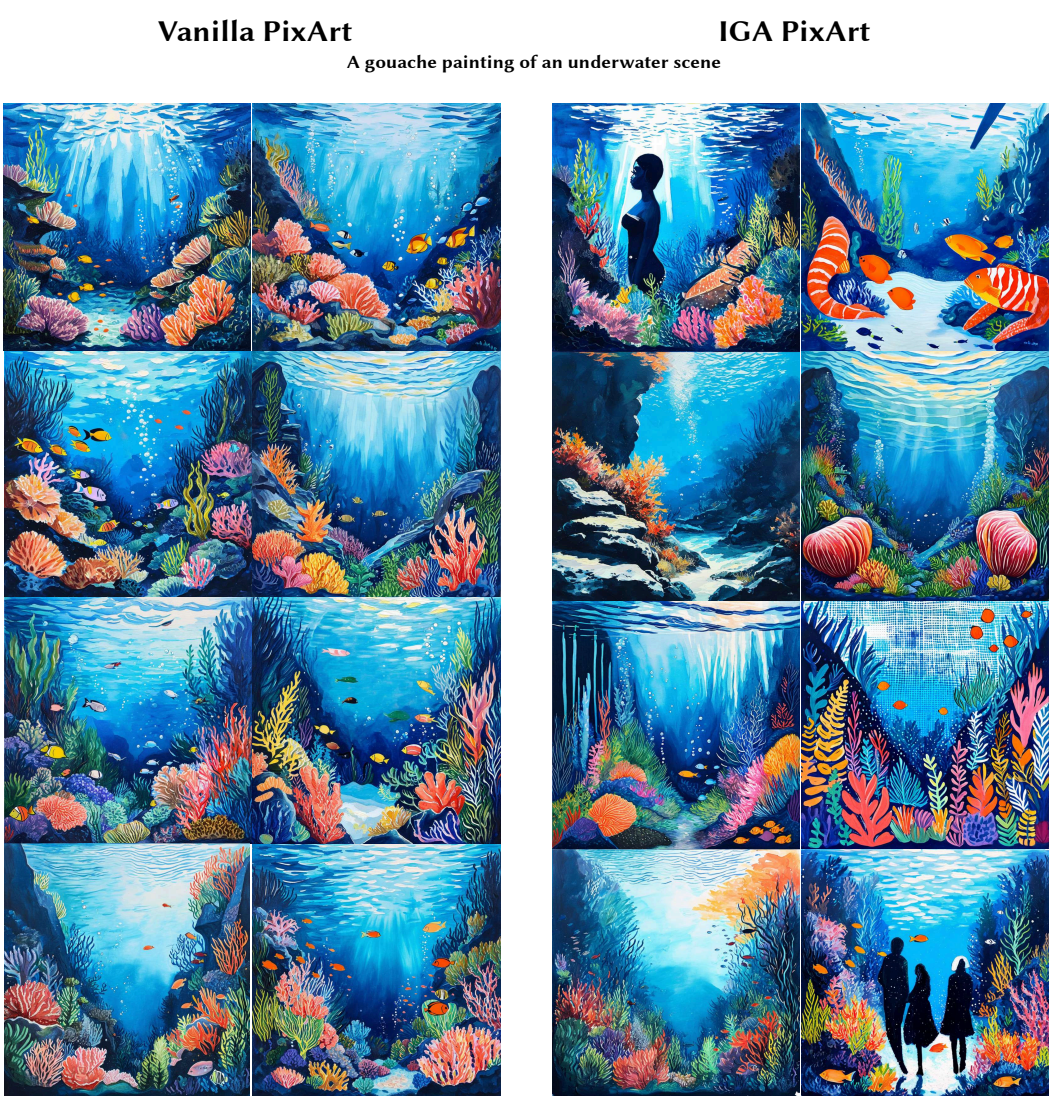


Figure 18: Underwater scenes from vanilla PixArt-Σ (left) and IGA-guided PixArt-Σ with $\lambda = 40$ (right).



Figure 19: Fantasy throne designs from vanilla PixArt-Σ (left) and IGA-guided PixArt-Σ with $\lambda = 20$ (right).

H.5 SPECTRAL REWEIGHTING AND APPEARANCE SHIFTS IN GENERATED FACES

To understand how spectral guidance changes the face distributions, we first examine which images the fitted target favors. The multiplier $w_\lambda(x) = \widehat{q}_\lambda(x)/p(x)$ measures the weight assigned to an image relative to the base distribution. Figures 20 and 21 plot these weights on base-generated FFHQ and CelebA-HQ faces, using one shared DINOv2 t-SNE map for each dataset. High-weight faces tend to lie in less crowded regions, while many low-weight faces lie in dense clusters. The maps give a qualitative view of how the fitted potential favors less represented appearances.

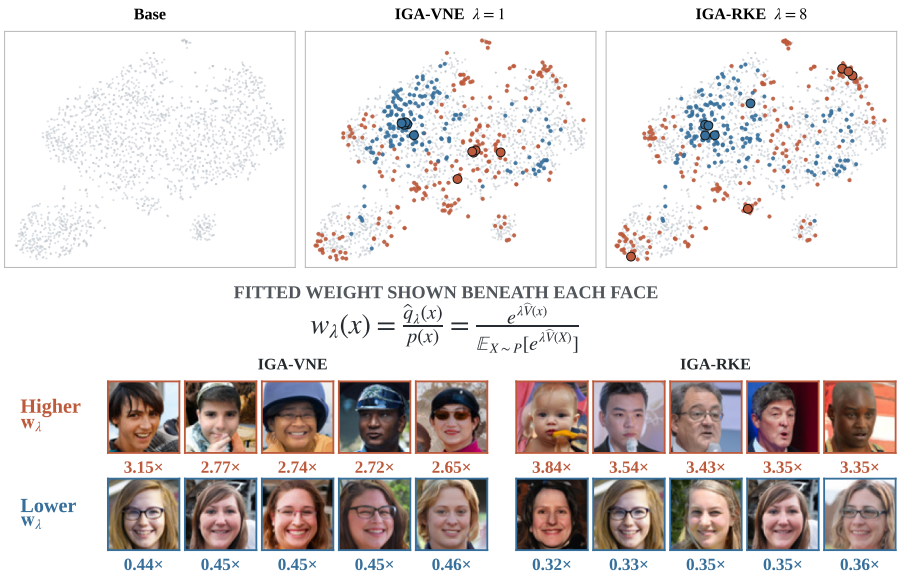


Figure 20: **Fitted IGA weights on FFHQ base outputs.** Orange and blue mark the highest and lowest weight deciles; outlined portraits show examples from each tail with their multipliers.

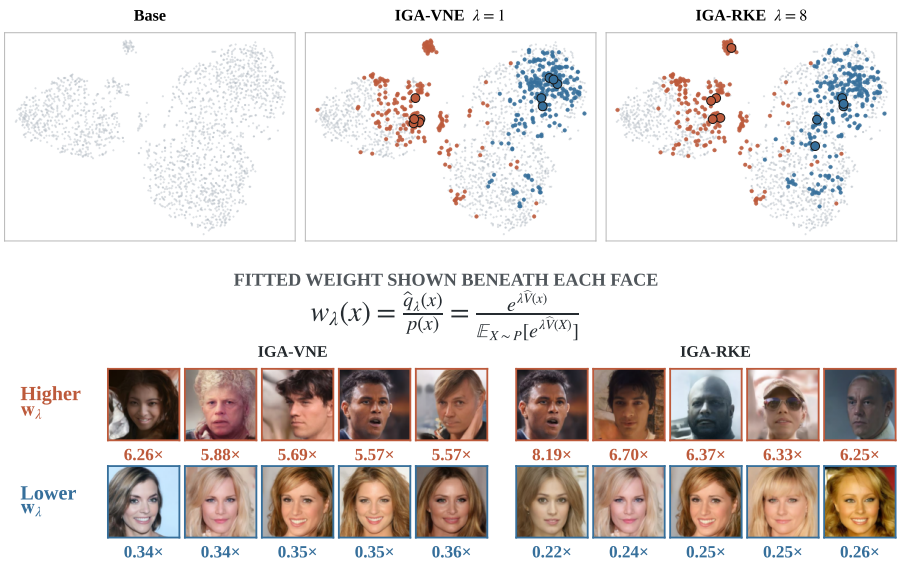


Figure 21: **Fitted IGA weights on CelebA-HQ base outputs.** Orange and blue mark the highest and lowest weight deciles; outlined portraits show examples from each tail with their multipliers.

We next ask whether newly generated guided faces move toward the real reference in recognizable aspects of appearance. Using OpenCLIP, we compare 128 matched generated seeds per setting with 1,024 real images per dataset. The probe assigns each image a normalized similarity across a group of appearance prompts; we refer to the resulting relative weights as prompt mass. This measures image–text association rather than verified attributes or demographic frequencies. Full probe definitions and the image-bootstrap intervals are described in Appendix J.2.

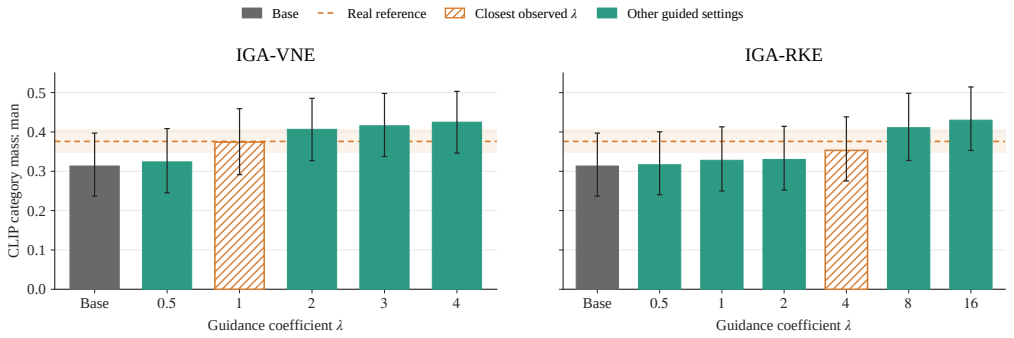


Figure 22: Mean normalized OpenCLIP similarity to the prompt *man*, relative to *woman*, on CelebA-HQ. The line and band show the real-data mean and 95% interval; hatching marks each method's closest mean.

In Figure 22, the base falls below the real-reference mean. Moderate guidance brings the mean closer to the reference, while stronger guidance moves it beyond the reference level.

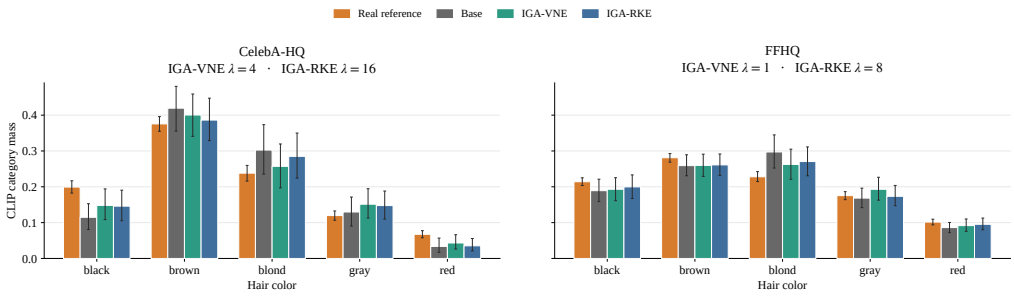


Figure 23: **Hair-color prompt distributions under IGA guidance.** Real, base, and guided mass across five prompts on CelebA-HQ and FFHQ; error bars show image-bootstrap 95% intervals.

Figure 23 provides a second view of the guided distributions. On CelebA-HQ, the mean black-, brown-, and blond-hair prompt masses move toward the real reference. On FFHQ, the excess blond-hair prompt mass also decreases. Together with the reweighting maps, these probes connect the spectral changes to differences in the appearance of newly generated samples.

Finally, we check whether the guided sampler reaches images favored by the fitted potential. Figures 24 and 25 evaluate w_λ on newly generated IGA outputs and average it over the full bank at each coefficient: 10,240 images for FFHQ and 30,000 for CelebA-HQ. Its average over the base is one by normalization. For both objectives, the mean fitted probability ratio of IGA samples relative to the base rises with λ, reaching plotted maxima of $2.37\times$ (VNE) and $2.53\times$ (RKE) on FFHQ, and $13.88\times$ (VNE) and $1.66\times$ (RKE) on CelebA-HQ. The curves summarize the complete banks, while the portraits illustrate individual high-weight guided outputs.

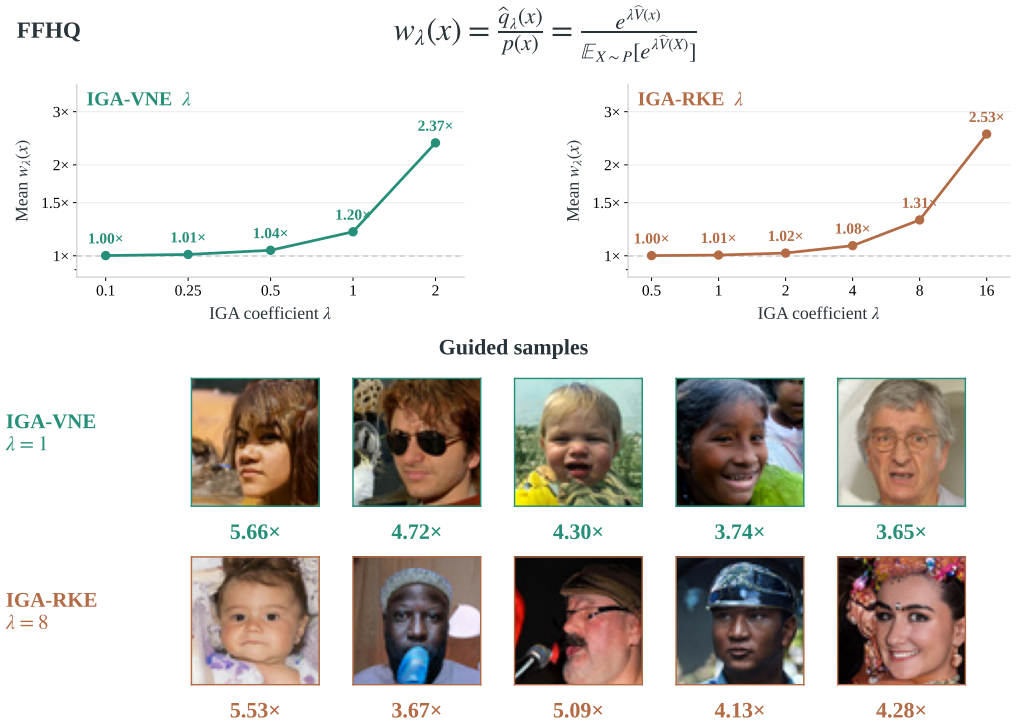


Figure 24: IGA weights on generated FFHQ images. Top: mean w_λ over 10,240 images per setting; the dashed line marks the base mean of one. Bottom: selected high-weight samples from IGA-VNE ($\lambda = 1$) and IGA-RKE ($\lambda = 8$), labeled with their weights.

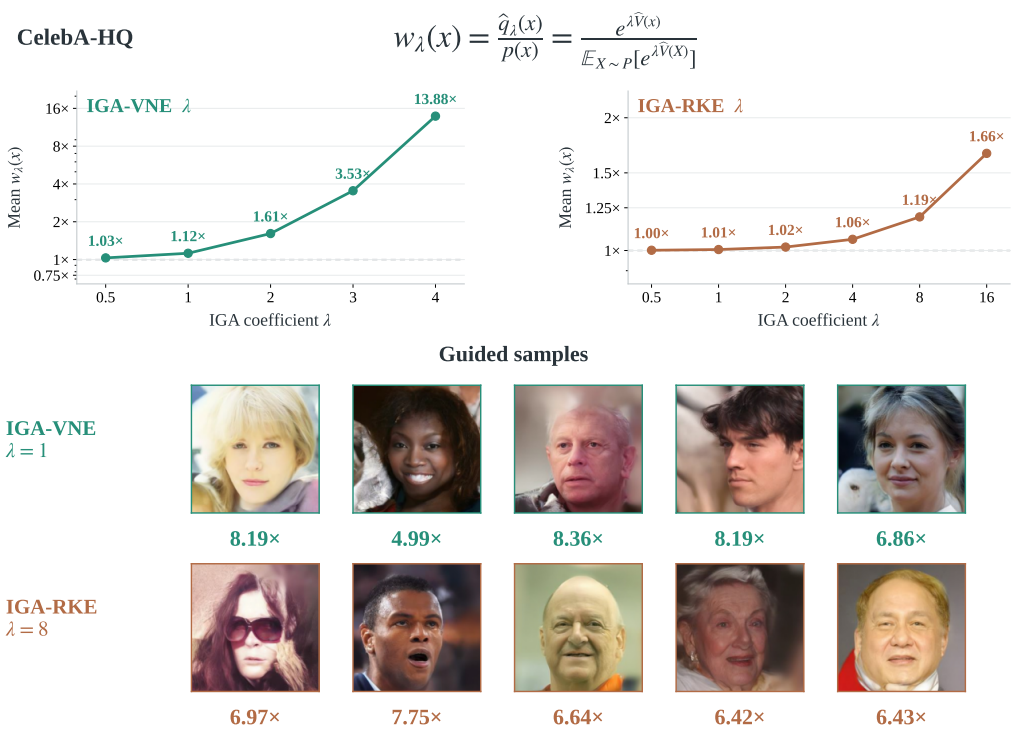


Figure 25: IGA weights on generated CelebA-HQ images. Top: mean w_λ over 30,000 images per setting; the dashed line marks the base mean of one. Bottom: selected high-weight samples from IGA-VNE ($\lambda = 1$) and IGA-RKE ($\lambda = 8$), labeled with their weights.

H.6 IMAGENET SAMPLES ACROSS THE ENTROPY WALL

Figure 26 shows matched-seed ImageNet samples along the IGA path. Changes are modest below the empirical entropy wall and become more pronounced at and beyond it.

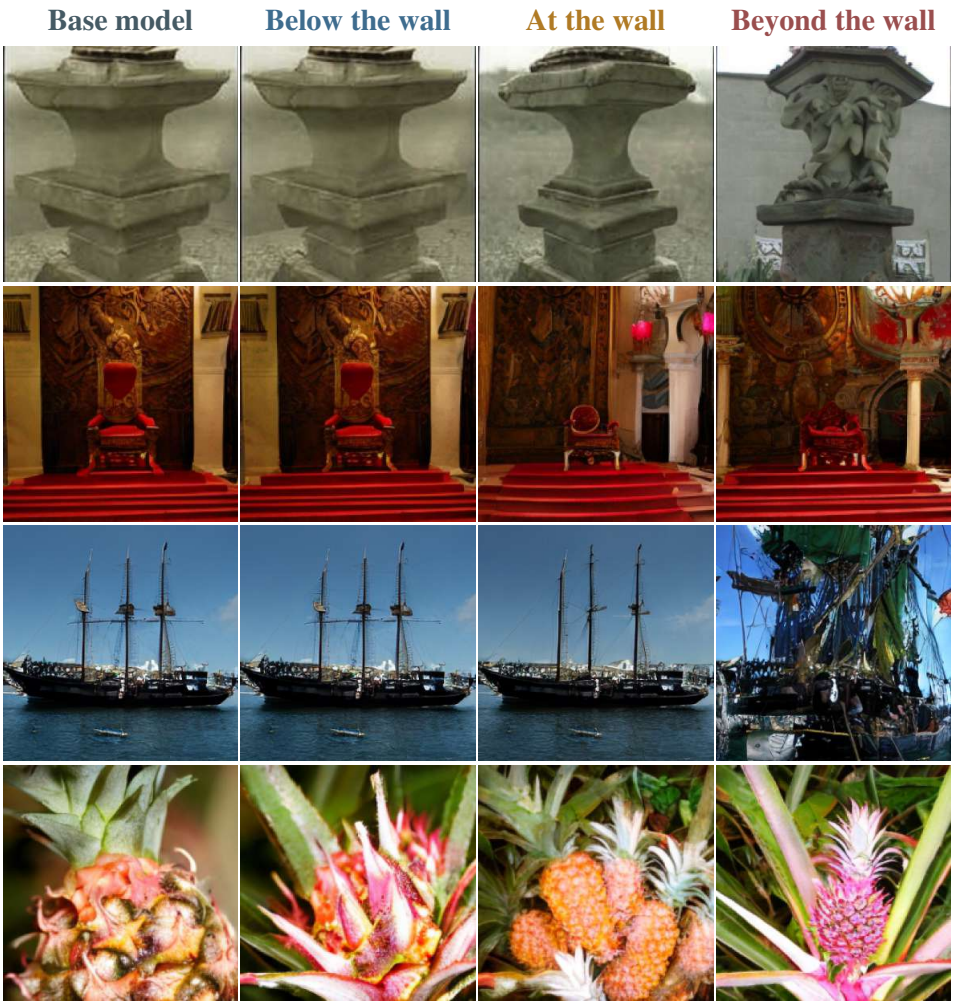


Figure 26: Class-conditional ImageNet samples from the base model and at increasing IGA coefficients. Each row follows one class and seed across the four labeled entropy regimes.

I ADDITIONAL NUMERICAL RESULTS

We extend the image-benchmark analysis with CelebA-HQ and FFHQ method comparisons, distance curves in both evaluation spaces, and studies of the fitting data and generator architecture. The entropy-wall curves for all three benchmarks appear in Figure 3.

CelebA-HQ. Table 2 compares IGA with the base model, Particle Guidance, and SPARKE. Both IGA repair settings improve FID, KID, DINOv2 distances, and recall over the base. At the extrapolation settings, Inception distances decrease further while DINOv2 distances increase, illustrating the dependence of agreement on the evaluation representation. IGA-VNE attains the highest recall among the reported methods.

Table 2: **CelebA-HQ benchmark comparison.** Each setting uses 30,000 generated images. Metrics and notation follow Table 1; ± denotes one five-group jackknife standard error. IGA (R) and (E) denote repair and extrapolation; exact λ values are in Appendix I.1.

		Inception-v3		DINOv2		Coverage & Diversity			
	Method	FID ↓	KID $\times 10^3$ ↓	FD ↓	KD $\times 10^2$ ↓	Prec. (%) ↑	Rec. (%) ↑	H ↑	H_{RKE} ↑
CelebA-HQ	Base	31.41±0.15	29.62±0.21	156.5±1.0	12.55±0.052	56.9±0.5	30.9±0.1	2.31±0.003	0.86±0.001
	Particle	30.93±0.11	29.23±0.17	154.7±1.0	12.48±0.052	56.0±0.4	31.3±0.3	2.34±0.003	0.87±0.001
	SPARKE	31.81±0.06	28.08±0.09	301.7±2.0	17.78±0.070	44.4±0.6	36.5±0.4	2.66±0.003	**1.02±0.001**
	IGA-VNE (R)	29.74±0.08	27.66±0.12	152.2±0.9	12.27±0.034	55.0±0.5	34.2±0.6	2.43±0.004	0.91±0.002
	IGA-VNE (E)	**28.78±0.09**	**25.66±0.10**	205.3±1.7	14.05±0.074	47.4±0.4	**40.7±0.7**	**2.68±0.004**	**1.02±0.001**
	IGA-RKE (R)	30.30±0.12	28.13±0.13	**152.1±0.8**	**12.26±0.036**	**56.2±0.5**	32.9±0.3	2.40±0.003	0.91±0.001
	IGA-RKE (E)	29.17±0.09	26.51±0.11	165.2±1.2	12.79±0.053	52.8±0.4	36.3±0.8	2.55±0.003	0.97±0.001

FFHQ. Table 3 gives the method comparison corresponding to Figure 3(c). Both IGA repair settings improve the reported distances and recall over the base; the extrapolation settings increase entropy and recall further while increasing the distances to the real reference.

Table 3: **FFHQ benchmark comparison.** Each setting uses 10,240 generated images. Metrics and notation follow Table 1; ± denotes one five-group jackknife standard error.

		Inception-v3		DINOv2		Coverage & Diversity			
	Method	FID ↓	KID $\times 10^3$ ↓	FD ↓	KD $\times 10^2$ ↓	Prec. (%) ↑	Rec. (%) ↑	H_0 ↑	H_{RKE} ↑
FFHQ	Base	3.49±0.11	1.07±0.16	210.3±1.6	7.95±0.132	83.9±0.3	56.5±2.3	2.50±0.003	0.87±0.001
	Particle	4.47±0.05	2.17±0.15	223.4±1.4	8.98±0.109	81.6±0.4	56.5±2.0	2.50±0.003	0.87±0.001
	SPARKE	4.12±0.11	1.81±0.10	220.4±1.4	8.31±0.102	82.4±0.4	58.4±1.1	2.52±0.003	0.88±0.001
	IGA-VNE (R)	**3.32±0.11**	0.87±0.14	207.5±1.4	7.59±0.137	82.3±0.3	60.1±1.3	2.53±0.003	0.88±0.001
	IGA-VNE (E)	7.80±0.15	3.60±0.10	325.9±1.6	9.96±0.067	71.3±0.4	**67.5±1.1**	**2.78±0.002**	0.98±0.001
	IGA-RKE (R)	3.38±0.14	**0.77±0.11**	**204.2±1.0**	**7.29±0.100**	**82.7±0.6**	59.5±1.5	2.54±0.003	0.89±0.001
	IGA-RKE (E)	9.03±0.16	4.20±0.12	305.2±3.2	9.66±0.104	76.5±0.4	61.6±1.5	2.76±0.002	**1.00±0.001**

I.1 Image-Benchmark Operating Points

The IGA λ settings for the (VNE-R, VNE-E, RKE-R, RKE-H/E) rows of Tables 1, 2, and 3 are $(0.5, 4, 4, 16)$ on ImageNet, $(1, 4, 4, 16)$ on CelebA-HQ, and $(0.25, 2, 2, 16)$ on FFHQ. The final RKE row denotes high diversity on ImageNet and extrapolation on the other two datasets.

I.2 Distributional Paths Across Feature Spaces

We use two evaluation spaces to examine which conclusions persist beyond a single representation. Figure 27 shows the Inception-v3 KID paths corresponding to the main-text FID paths. Figure 28 complements the Inception-v3 results in the main text with DINOv2 Fréchet and kernel distances. Moderate guidance reduces these distances on all three benchmarks. At higher diversity, the Fréchet distances rise, while the kernel-distance response varies by dataset. For example, the CelebA-HQ DINOv2 kernel distance turns upward even though its Inception KID continues to decrease. These differences explain why we evaluate the paths in both spaces.

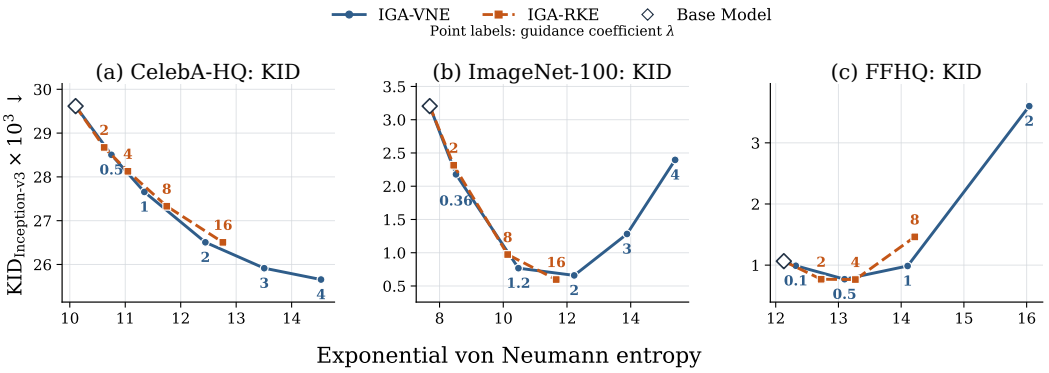


Figure 27: Inception-v3 KID versus DINOv2-space Vendi score on CelebA-HQ, ImageNet, and FFHQ. Curves show IGA-VNE and IGA-RKE; diamonds mark the base models and point labels give λ.

Figures 29 and 30 show the same distance measurements against the DINOv2 RKE score $\exp(H_{\mathrm{RKE}})$. This view lets us examine the paths using the order-2 diversity measure as well as Vendi.

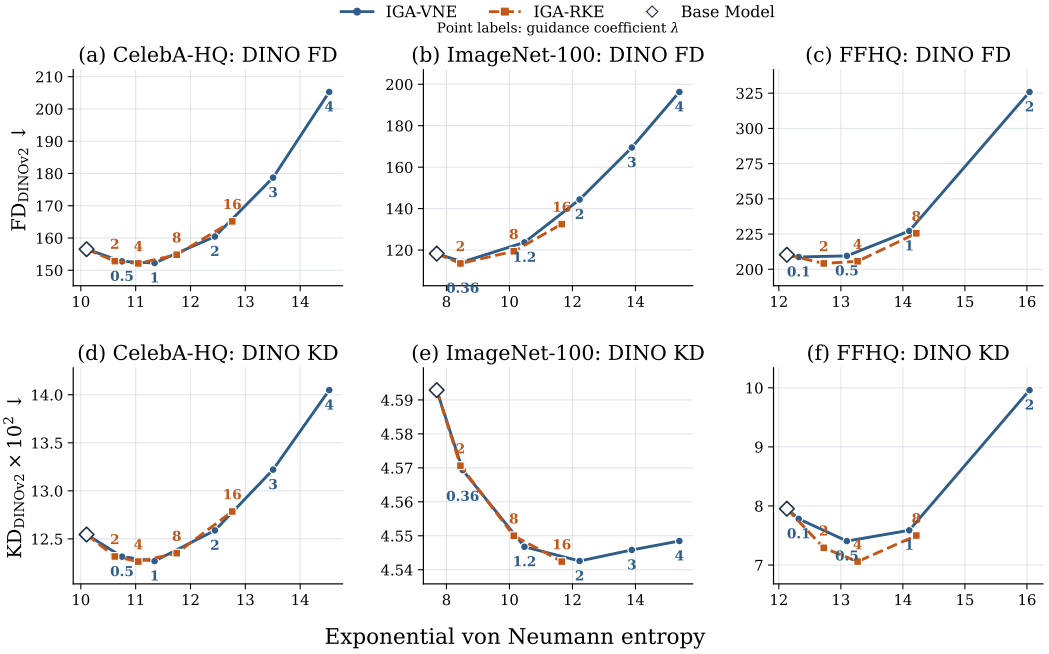


Figure 28: **DINOv2-space distances versus Vendi score.** Fréchet distance (FD, top) and kernel distance (KD, bottom) on CelebA-HQ, ImageNet, and FFHQ. Curves show IGA-VNE and IGA-RKE; diamonds mark the base models and point labels give λ.

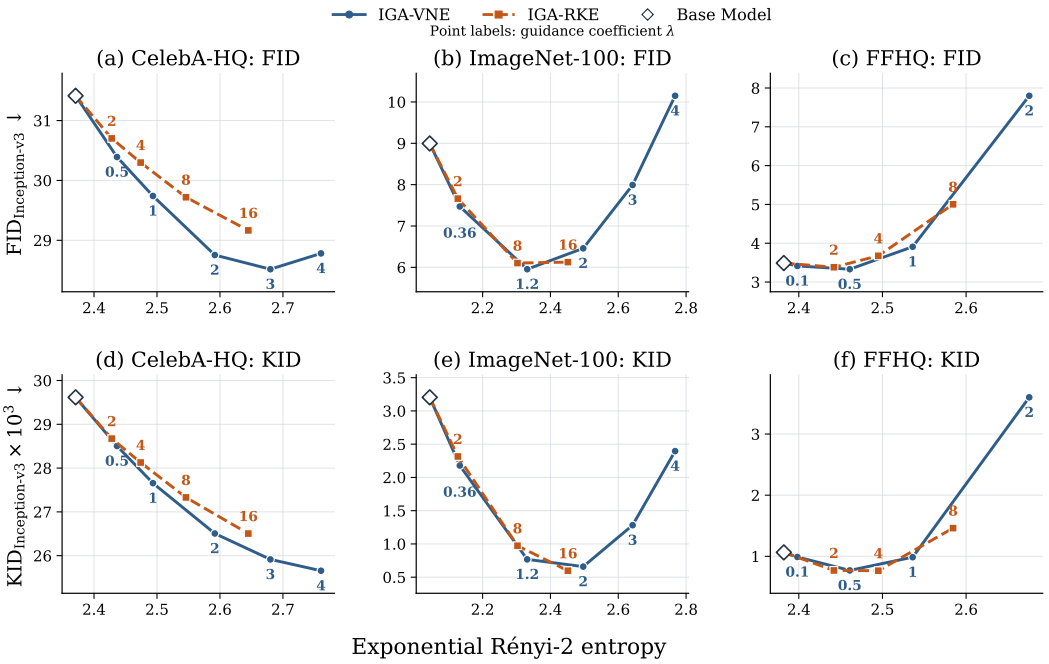


Figure 29: **Inception-v3 distances versus RKE score.** FID (top) and KID (bottom) on CelebA-HQ, ImageNet, and FFHQ; point labels give λ.

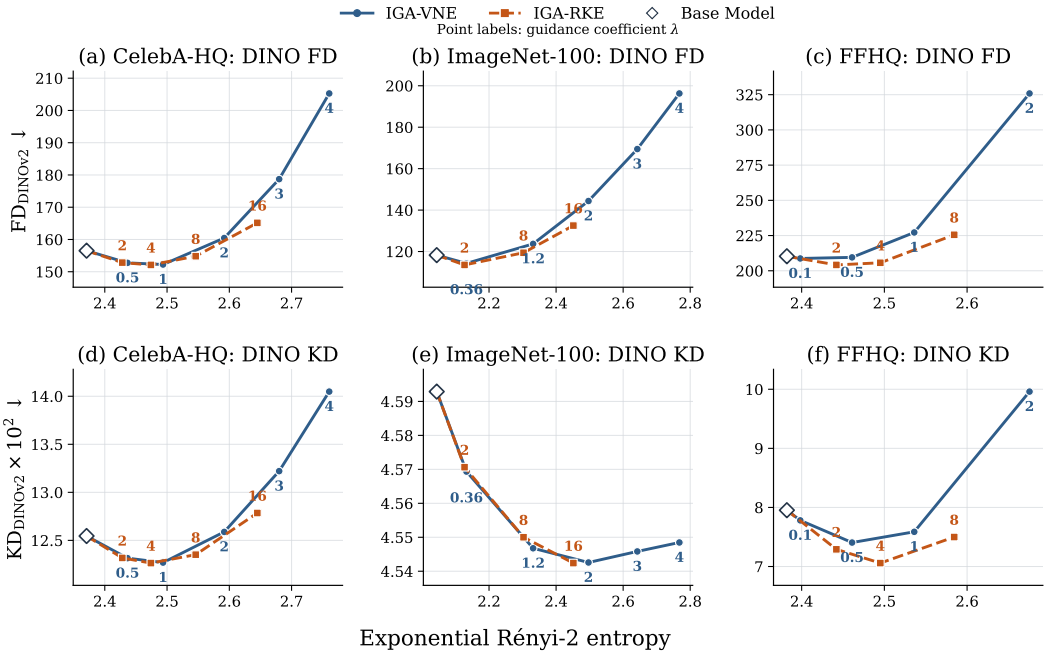


Figure 30: **DINOv2-space distances versus RKE score.** Fréchet distance (FD, top) and kernel distance (KD, bottom) on CelebA-HQ, ImageNet, and FFHQ; point labels give λ.

I.3 Training-Image Versus Generated-Pilot Fitting

We study whether the source of the optimization samples changes the resulting guidance path. We replace the base-generated pilot with training images when solving for the IGA potential, then use

the fitted potential to guide the same pretrained sampler. Figures 31 and 32 show broadly similar relationships between diversity and Inception distances for both fitting sources. The generated-pilot construction uses the base-model reference in the projection result of Proposition 3; fitting on training images instead uses the empirical data distribution as the reference. Bank sizes and sampling details are given in Appendix J.

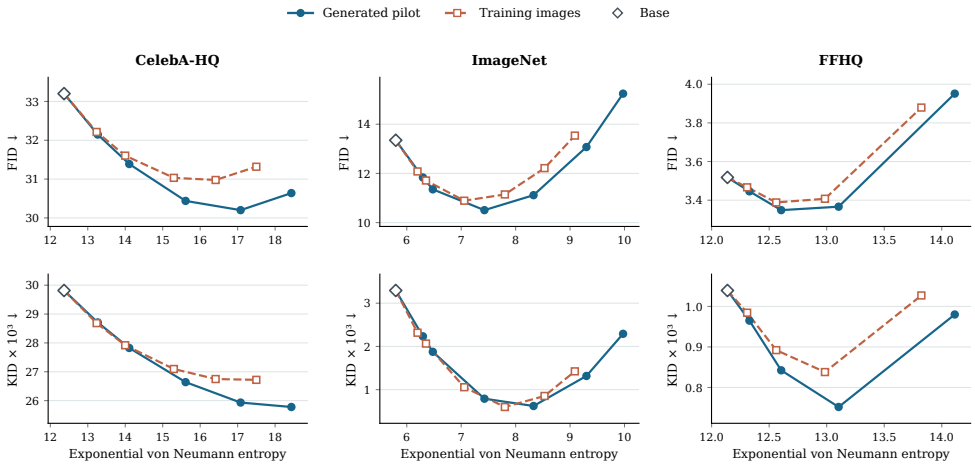


Figure 31: IGA-VNE fitted using generated samples (blue circles) or training images (orange squares). FID (top) and KID (bottom) are plotted against the Vendi score; diamonds mark the base models.

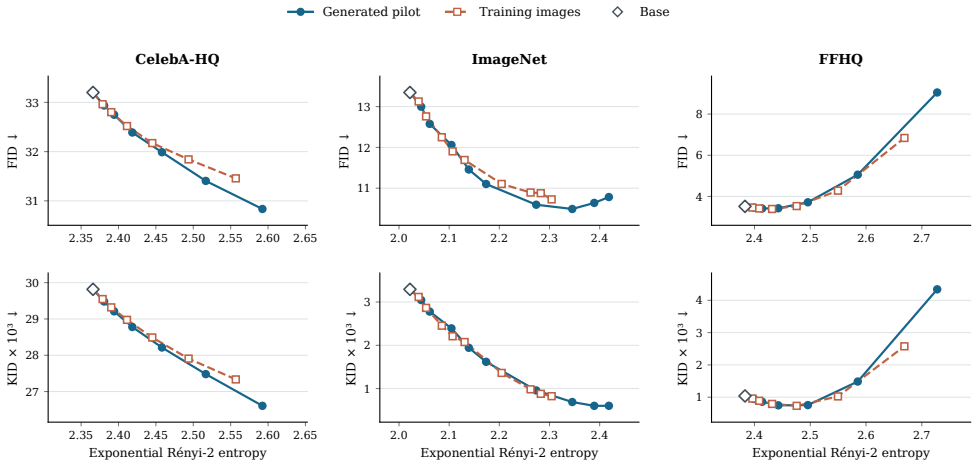


Figure 32: IGA-RKE fitted using generated samples (blue circles) or training images (orange squares). FID (top) and KID (bottom) are plotted against the RKE score; diamonds mark the base models.

I.4 Transfer Across Diffusion Backbones

To test whether the ImageNet behavior depends on DiT, we evaluate four released class-conditional 256×256 backbones: DiT-XL/2 (transformer) (Peebles & Xie, 2023), CompVis LDM (latent U-Net) (Rombach et al., 2022), DiCo-XL (ConvNet) (Ai et al., 2025), and Patch Diffusion (patch-aware ADM) (Wang et al., 2023b). These cover distinct denoiser designs under a common benchmark. For every backbone, each metric evaluation uses 10,000 generated images (100 per class) against a matched 10,000-image real reference. Figure 33 shows a baseline entropy deficit and an initial FID/KID decrease for every backbone as diversity rises. All four reach the real-data entropy level, although the distance minima differ; Patch Diffusion's FID and KID continue to improve over the measured range. The repair-to-extrapolation path is therefore not specific to DiT.

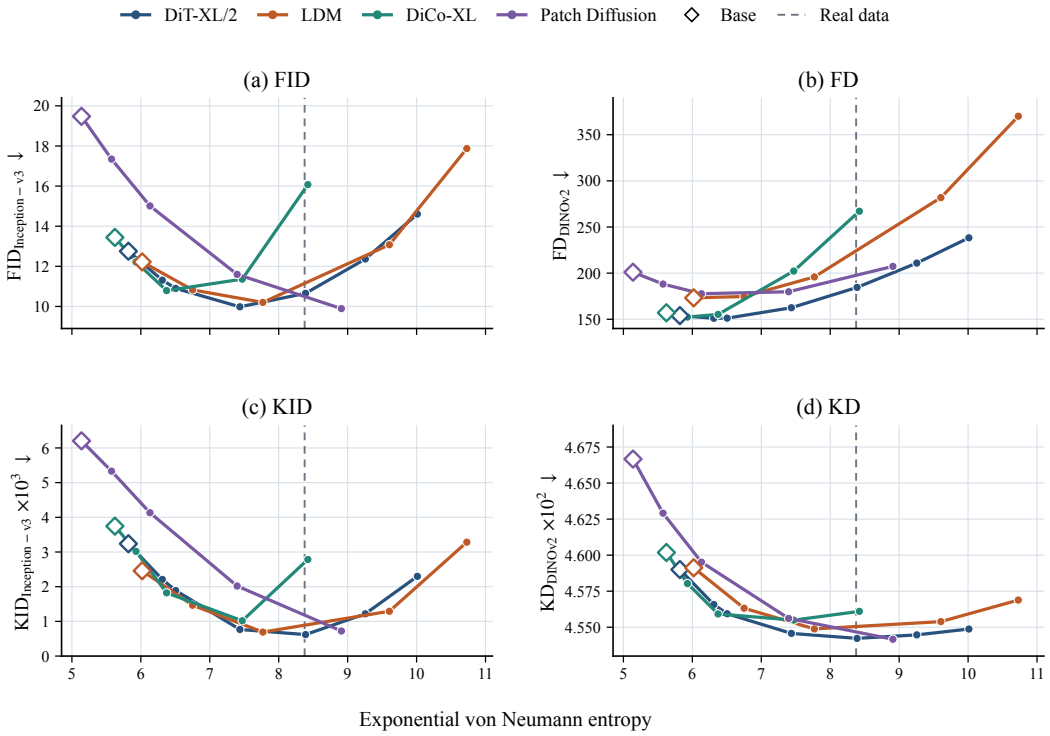


Figure 33: IGA across four ImageNet diffusion models. Inception-v3 distances (FID/KID) and DINOv2 distances (FD/KD) are plotted against the Vendi score. Hollow diamonds mark the base models; the dashed line marks the real-data score.

J Experimental Details and Evaluation Protocol

We give the metric definitions and the image-benchmark and prompt-conditioned settings used in Section 5. Additional controlled synthetic results appear in Appendix G.

J.1 Evaluation metrics

Spectral entropy and effective rank (Friedman & Dieng, 2023; Jalali et al., 2023; Ospanov et al., 2024). We report the unsmoothed von Neumann entropy H_0 and its exponential, the Vendi score, together with the order-2 Rényi entropy $H_{\mathrm{RKE}} = -\log \mathrm{Tr}(\Sigma_Q^2)$ and its exponential, the RKE score. Higher values indicate greater feature-space diversity. For SDXL, we compute Vendi in DINOv2 (Oquab et al., 2024), CLIP (Radford et al., 2021), and Inception-v3 spaces to check that the conclusion is not representation-specific.

Fréchet and kernel distances (Heusel et al., 2017; Bińkowski et al., 2018). Fréchet distance compares feature means and covariances, whereas kernel distance is a nonparametric two-sample statistic; lower is better for both. We report them in Inception-v3 and DINOv2 spaces to test whether diversity gains remain close to the data distribution across representations.

Precision and recall (Kynkäänniemi et al., 2019). We form full-bank, pooled manifolds of ℓ_2-normalized DINOv2 features with $k = 3$, excluding self-neighbors. On ImageNet, classes are concatenated before computation, rather than averaged class by class. Recall estimates coverage of the real-data manifold; precision estimates how much generated mass remains supported by that manifold.

Image-benchmark uncertainty. The $\pm$ values in Tables 1, 2, and 3 are one five-group delete-one jackknife standard error. We keep the real reference and fitted IGA potential fixed; these are not independent-seed intervals.

Caption-based imagination (Papineni et al., 2002; Zhu et al., 2018; Li et al., 2016). Each image is independently recaptioned with BLIP (Li et al., 2022). Within a 32-image prompt panel, we report $1 - \mathrm{SelfBLEU}_2$ (less repetition) and Distinct-2, the fraction of unique caption bigrams. This generate–reception–measure pipeline asks whether visual diversity changes what an independent captioner says, rather than only pixels or style. We also report the alignment-weighted score $D^{\mathrm{align}} = (1 - \mathrm{SelfBLEU}_2)\max(A, 0)$, where A is the mean CLIP text-embedding cosine similarity between the recaptions and the source prompt, to reward semantic variation that remains prompt-aligned.

Prompt fidelity and perceptual quality (Kirstain et al., 2023; Wang et al., 2023a; Ke et al., 2021). PickScore measures learned prompt–image preference; CLIP-IQA and MUSIQ assess image quality without a reference. Higher is better, and together they test whether diversity is purchased through prompt drift or visible degradation.

Quality-weighted Vendi (Nguyen & Dieng, 2024). Quality-weighted Vendi combines each panel's Vendi score with its mean perceptual quality, penalizing diverse but low-quality outputs.

J.2 Image-benchmark configurations

Shared IGA settings. We fit each potential on an independent generated pilot using DINOv2 ViT-B/14 CLS features (Oquab et al., 2024) and a fixed 1024-dimensional random Fourier feature map (Rahimi & Recht, 2007) with both sine and cosine features to ensure a universal unit-norm as discussed in (Sutherland & Schneider, 2015). The VNE fit uses $\varepsilon = 10^{-3}$. ImageNet fits one potential per class; the two unconditional benchmarks fit one each. The potential is frozen before evaluation. We use a guidance-score coefficient of one over the 20–80% portion of reverse sampling steps. For FFHQ, we use the EDM VP FFHQ-64 model with 40 Heun steps (Karras et al., 2022).

Table 4: **Image-benchmark setup.** Counts are pilot / real reference / generated images per setting. The λ grids include the unguided base point.

Dataset	Generator and sampler	Pilot / real / generated	IGA λ grids
ImageNet-100	DiT-XL/2-256; 50-step ancestral DDPM; CFG 2	1,024/class / 50,000 / 50,000	VNE $\{0, 0.36, 0.5, 1.2, 2, 3, 4\}$ RKE $\{0, 2, 4, 8, 16\}$
CelebA-HQ	EMA DDPM-256; 100 stochastic DDPM steps	4,096 / 30,000 / 30,000	VNE $\{0, 0.5, 1, 2, 3, 4\}$ RKE $\{0, 0.5, 1, 2, 4, 8, 16\}$
FFHQ	EDM VP FFHQ-64; 40 Heun steps	2,048 / 1,024† / 10,240	VNE $\{0, 0.1, 0.25, 0.5, 1, 2\}$ RKE $\{0, 0.5, 1, 2, 4, 8, 16\}$

†FFHQ FID uses official full-dataset reference moments; the 1,024-image bank is used for KID and DINOv2 metrics.

Comparator settings. Table 5 summarizes the settings for CADS (Sadat et al., 2024), SPARKE (Jalali et al., 2025), and Particle Guidance (Corso et al., 2024). SPARKE uses normalized gradients every ten steps; Particle Guidance evolves groups of four coupled samples. All use the corresponding benchmark sampler, and CADS applies only to class-conditional ImageNet.

Table 5: **Baseline settings used in Tables 1, 2, and 3.**

Method	ImageNet-100	CelebA-HQ	FFHQ
CADS	$s = 0.10$, $(\tau_1, \tau_2) = (0.55, 0.90)$, $\psi = 1$	–	–
SPARKE	$c = 0.03$; batch 32; same-class guidance	$c = 0.003$; batch 64; history 1,000	$c = 0.003$; five histories
Particle Guidance	$c = 30$	$c = 10$	$c = 30$

Evaluation and fitting-bank comparison. FID and KID use Inception-v3 features; FD, KD, precision, and recall use DINOv2. KID and DINO KD average 1,000-image subsamples (50 draws on ImageNet and FFHQ, 100 on CelebA-HQ). The matched-count comparison in Figure 3 applies to the spectral-entropy curves; for FFHQ in Table 3, KID and DINOv2 KD use matched 1,000-image subsamples, FID uses official real-data moments, and DINOv2 FD, precision, and recall use the full 1,024-real/10,240-generated banks. ImageNet spectral entropy is averaged across classes before exponentiation; the CelebA-HQ and FFHQ table estimates use the fixed random Fourier feature covariance. For the comparison in Figures 31 and 32, the training-image fit uses 10,000 CelebA-HQ images, all 126,689 ImageNet-100 training images, or 10,000 FFHQ images. Each path point evaluates 10,000 matched generated images.

Face maps and appearance probes. The face maps plot frozen pilot weights $w_\lambda = \widehat{q}_\lambda / p$. Their 2,048-point overlays are unscreened subsets of full base banks (10,240 FFHQ; 30,000 CelebA-HQ); upper portraits are legible selections from the 40 largest weights, and lower portraits are the five smallest. The prompt probes use OpenCLIP ViT-L/14-336 (OpenAI weights), 128 matched generated seeds, and 1,024 real images per dataset. Each prompt embedding averages its bare phrase with "a portrait photo of …" and "a close-up photograph of …" before normalization. Masses average the within-group softmax of 100 times image–text cosine. The hair-color probe uses a separate fixed seed cohort. Intervals use image bootstrap; divergence changes use 2,000 paired bootstrap replicates.

J.3 Prompt-conditioned SDXL evaluation

We give the protocol and metric definitions for the SDXL comparison in Figure 6.

Evaluation set and common protocol. We evaluate Stable Diffusion XL base 1.0 (Podell et al., 2024) on a randomly-selected block of 100 prompts from the MS-COCO 2017 validation split (Lin et al., 2014). For every method and prompt, we generate 32 images at 768×768 resolution using 50 deterministic DDIM steps. Initial latent noise is matched across methods wherever the sampling procedure permits. Each method therefore contributes 3,200 evaluation images, for 16,000 images in the five-way primary comparison.

Table 6: **Key hyperparameters for the prompt-conditioned SDXL comparison.** Settings shared by all methods are given in the text.

Method	Important hyperparameters	Sampling structure
IGA	$\lambda = 4$; DINOv2 CLS with a fixed 512-D cosine projection; active for $t/999 \in [0.35, 0.75]$	32-image fixed support; fitted potential frozen; independent trajectories
CADS (Sadat et al., 2024)	$s = 0.10$, $\tau_1 = 0.6$, $\tau_2 = 0.9$, $\psi = 1$	Annealed noise on SDXL text conditioning; independent
SPARKE (Jalali et al., 2025)	$c = 0.03$, $\sigma_{\text{image}} = 0.8$; Gaussian latent kernel; guidance every 10 steps	Sequential memory bank
Particle Guidance (Corso et al., 2024)	$c = 30$; latent-space RBF repulsion	32 jointly coupled particles

Metric computation and uncertainty. We apply the metrics in Section J.1 within each 32-image prompt panel. The external DINOv2, CLIP ViT-B/32, and Inception-v3 representations used for evaluation remain distinct from the projected DINOv2 representation used by IGA Guidance. All prompt-conditional endpoints are aggregated within each panel and reported relative to vanilla SDXL using matched prompts. The 95% confidence intervals are computed from 10,000 paired prompt-bootstrap replicates, resampling entire prompt panels. This retains method–baseline pairing and accounts for the within-panel dependence induced by SPARKE and Particle Guidance.