Imaginative Generative AI: Crossing the Entropy Wall into Worlds Beyond Imitation

Hossein Goli*, Amin Gohari†, Farzan Farnia‡

Abstract

Generative AI models are primarily designed to imitate the data distribution, an objective that neither corrects diversity lost by a learned generator nor defines how generation should extend beyond the diversity of the data itself. We introduce **Imaginative Generative AI (IGA)**, a framework that makes diversity part of the target-distribution design problem: among distributions close to a reference, IGA selects one whose spectral diversity reaches a prescribed level. Diversity is measured by the **von Neumann entropy** of the generated distribution's kernel covariance operator in a fixed representation space, providing a reference-free representation-guided measure of how broadly probability mass occupies embedding directions. The spectral entropy of the population data distribution defines an **Entropy Wall**. Below the wall, IGA performs **diversity repair**, recovering variation that a learned generator has lost while remaining within the diversity level of the data. Beyond the wall, the data distribution itself becomes infeasible, and IGA deliberately departs from it to produce distributions with greater representation-relative spectral diversity, an operational notion of **imaginative generation**. These regimes form a single regularization path from imitation to imagination and define an i.i.d. target distribution at each prescribed diversity level. We develop the theory of this entropy-constrained projection and show that, under a KL anchor to a pretrained generator, the optimum satisfies a self-consistent exponential-tilt relation. This characterization leads to *IGA Guidance*, a retraining-free inference-time method for score-based and diffusion models, including DDPM and DDIM samplers. Experiments on synthetic and vision benchmarks demonstrate diversity repair below the Entropy Wall and controlled spectral extrapolation beyond it.

1 Introduction

"Imagination is more important than knowledge. Knowledge is limited. Imagination encircles the world."

— Albert Einstein

The typical goal of a generative model is to reproduce the underlying distribution of its training data. For example, a successful image generator is expected to produce realistic images with approximately the same content and variation as the images on which it was trained. This principle underlies the standard paradigms of generative modeling in the literature, including generative adversarial networks (GANs) [1], variational autoencoders (VAEs) [2], and score-based and diffusion models [3–5]. Although these frameworks differ significantly in architecture and training, their population-level goal can be summarized in the following distributional discrepancy minimization:

$$\min_{Q \in \mathcal{P}_{\mathcal{G}}} \mathcal{D}\big(Q; P_{\text{data}}\big), \tag{1}$$

where P_{data} is the underlying data distribution, Q is the distribution produced by the generator over the set of feasible models $\mathcal{P}_{\mathcal{G}}$, and $\mathcal{D}(\cdot;\cdot)$ measures the discrepancy (or divergence) between the two distributions. We refer to this prevailing view of generative modeling as **distributional imitation**.

*Department of Computer Science and Engineering, The Chinese University of Hong Kong, hosseingoli@cse.cuhk.edu.hk
†Department of Information Engineering, The Chinese University of Hong Kong, agohari@ie.cuhk.edu.hk
‡Department of Computer Science and Engineering, The Chinese University of Hong Kong, farnia@cse.cuhk.edu.hk

Imitation is a natural statistical objective, but it also places a ceiling on what the generator is asked to do. Even an ideal solution of (1) is asked to match P_{data}, not to produce a distribution that is systematically more diverse or generate novel and creative content. Moreover, practical generators may not reach even the diversity of their training distribution. The recent study [6] by Farnia, et al. has found that generated samples can exhibit lower spectral diversity than real data when diversity is measured using reference-free diversity measures of the Vendi score [7] and Rényi kernel entropy [8]. This raises a fundamental question:

How should the target of generative modeling be regularized when diversity and novelty, in addition to fidelity, are something we want to control?

To address this question, we propose **Imaginative Generative AI** (**IGA**), a framework that makes diversity part of the target-distribution design problem. Instead of asking only for the distribution closest to a reference distribution, IGA asks for the closest distribution whose diversity score is at least above a given prescribed level. Let P_{ref} denote the reference distribution, which could be the empirical distribution $\hat{P}_n$ of training data or the distribution P_θ of a pretrained generator. Then, IGA solves the following regularized discrepancy minimization problem:

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

where $H(Q)$ measures the spectral entropy (interpreted as diversity) of distribution Q and ρ is the desired diversity level. Under the duality conditions developed in our theoretical analysis, the constrained problem at level ρ can equivalently be stated using a Lagrangian penalty at a corresponding multiplier $\lambda \geq 0$:

$$\min_{Q \in \mathcal{P}_{\mathcal{G}}} \; \mathcal{D}(Q; P_{\text{ref}}) - \lambda\, H(Q) \tag{3}$$

Note that the two terms have complementary roles: The discrepancy term keeps generated samples close to the reference distribution, while the entropy term rewards the spectral diversity in the Vendi score. Setting $\lambda = 0$ recovers standard reference matching; increasing λ gives more weight to the spectral diversity term.

We measure diversity using the **von Neumann entropy (VNE)** of the normalized kernel covariance operator induced by Q in a fixed embedding space. Intuitively, VNE is low when generated samples concentrate along a few embedding directions and high when they spread across many directions; in the empirical setting, it is the logarithm of the Vendi score [7–9]. This measure is reference-free but representation-dependent: it evaluates the diversity of Q without requiring a comparison distribution, while the chosen embedding specifies which variations are meaningful. IGA thus controls spectral diversity relative to a given embedding.

The data distribution itself provides a natural reference level for this diversity. To characterize this wall, we define the underlying distribution's entropy as

$$\rho_\star := H\big(P_{\text{data}}\big). \tag{4}$$

We call $\rho_\star$ the **Entropy Wall**. It is the spectral diversity of the data distribution in the chosen representation. The notion of entropy wall separates two different regimes in the IGA generative modeling approach:

(Regime I) Below the Entropy Wall: Diversity Repair. The below-the-wall regime concerns IGA when we choose entropy lower-bound ρ to satisfy $\rho \leq \rho_\star$.

In this regime, the required diversity level is no greater than the diversity already present in the underlying data distribution. Therefore, in this regime, IGA can then be viewed as **repairing a diversity deficit in a learned generator**: it encourages the generator to recover variation that was present in the data but weakened or lost during training the generative model. Note that, as empirically demonstrated by Farnia et al. in recent work [6], the standard generative models commonly suffer from a diversity bias, and the spectral entropy of their generated data cannot match that of the underlying distribution generating their training samples. In brief, the goal in this regime remains faithful modeling of the data, with an explicit mechanism for counteracting spectral diversity shortfall as shown in [6].

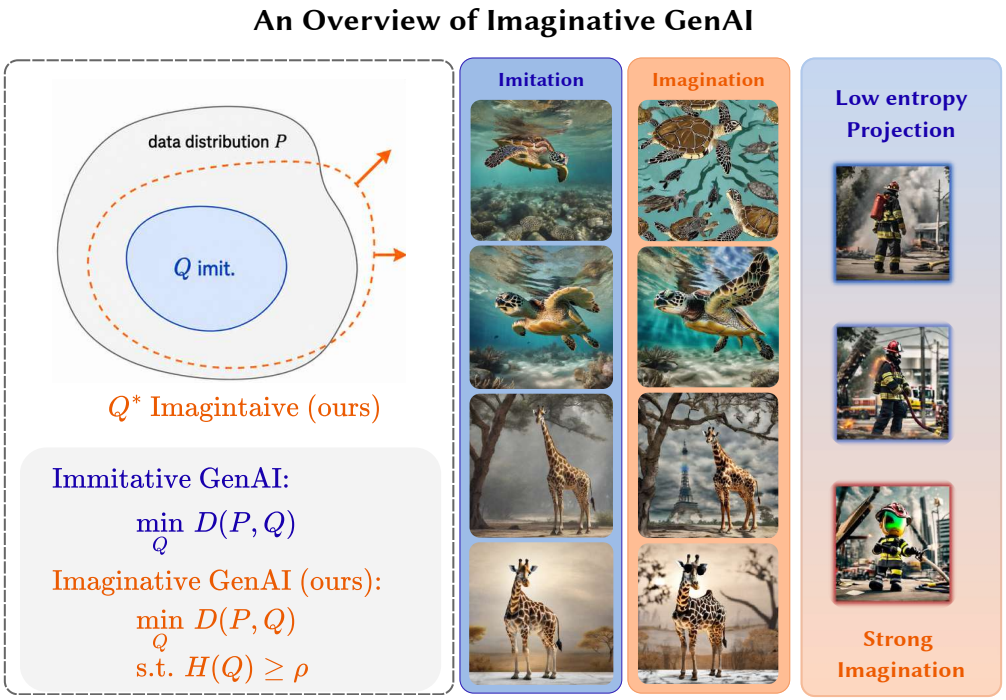


Figure 1: **From Imitation to Imagination.** Imitation aims to return a distribution Q close to P_{data}; IGA returns the closest Q with $H(Q) \geq \rho$, which meets the data distribution below and at the entropy wall $\rho_\star = H(P_{\text{data}})$ and leaves it beyond. Images are Stable Diffusion XL at matched prompts and seeds; the right column sweeps one multiplier from the low-entropy projection to strong extrapolation.

(Regime II) Beyond the Entropy Wall: Imaginative Generation. This regime of applying IGA is when we select the projection lower-bound to satisfy the strict inequality: $\rho > \rho_\star$.

Especially, we highlight that in this regime, the data distribution itself no longer satisfies the diversity constraint, and thus **the IGA solution in** (2) **must intentionally differ from** P_{data}, regardless of whether it is anchored directly to the data or to a pretrained model. The discrepancy term prevents this solution from moving arbitrarily far from the chosen reference P_{ref}, while the entropy constraint pushes it to occupy a broader set of embedding directions. We call this regime *imaginative* because the target has greater spectral diversity than the data distribution that defines the wall.

Here, we use the term **imaginative** to describe generation whose spectral diversity exceeds that of the data in a specified representation space, e.g. CLIP or DINO embedding spaces for images. This representation-relative definition makes imagination operational while allowing the embedding to encode domain-relevant semantics. We evaluate the resulting variation using both the guiding representation and independent measures of sample quality and diversity.

The two regimes form a single regularization path. As the target ρ increases, IGA moves from ordinary imitation to diversity repair and then, after crossing the entropy wall $\rho_\star = H(P_{\text{data}})$ (corresponding to the real data distribution), to deliberate imagination. Therefore, our IGA framework defines diversity enhancement as a property of a single target distribution, enabling i.i.d. generation at a prescribed spectral-diversity level, including levels beyond the entropy of the data. We emphasize that this differs from methods that induce diversity through repulsive or sample-dependent interactions over the course of generating multiple samples, whose outputs form a coupled and generally non-i.i.d. batch [10–12]. IGA therefore provides an i.i.d. distributional alternative to interaction-based diversity promotion, while complementing work that evaluates

Figure 2: **From imitation to imagination with SDXL.** As the IGA multiplier λ increases, SDXL produces progressively stronger structural and compositional variations while preserving the underlying concept.

diversity after generation [7, 8].

Next, we demonstrate that this distribution-level formulation further leads to a practical method for diversity-improved sampling from pretrained score-based and diffusion models. When the reference is a pretrained model P_θ and the discrepancy measure is the KL-divergence $\mathrm{KL}(Q\|P_\theta)$, the analysis in the main body shows that the optimal target takes the form

$$Q^\star(dx) \;\propto\; P_\theta(dx)\cdot \exp\big(\lambda\, G_{Q^\star}(x)\big), \tag{5}$$

where $G_{Q^\star}(x)$ measures how placing probability mass near x changes the VNE spectral entropy term. Because this energy depends on $Q^\star$, the relation is self-consistent. It reweights the base model to increase the spectral diversity of the generated population while remaining close to the original distribution as much as possible.

We specifically show that the application of this formulation to score-based and generative models can be performed by our proposed **IGA Guidance**. IGA Guidance is an inference-time approximation for score-based, DDPM, and DDIM samplers that requires no retraining. Our numerical experiments show promising results of the IGA guidance for large-scale diffusion models. For example, Figure 2 shows the application of IGA-Guidance to the large-scale SD-XL model and how increasing the parameter λ leads to visually more diverse and imaginative image outputs for the input prompt "A Skyscraper". The summary of our contributions are as follows:

- We formulate *IGA*, an entropy-constrained projection framework enabling i.i.d. generation at prescribed VNE diversity levels.
- We introduce the *entropy wall*, separating diversity repair from controlled extrapolation beyond the data's spectral diversity.
- We characterize the IGA regularization path and derive a self-consistent exponential tilt for the KL-anchored optimum.
- We develop *IGA guidance* for inference-time steering of pretrained score-based, DDPM, and DDIM samplers without retraining.

Figure 3: **A multiverse of diversity levels.** Matched-seed PixArt-Σ samples at increasing λ in the IGA framework ($\lambda = 0$ represents the original regularization-free PixArt-Σ), arranged by measured spectral entropy H. *Prompt:* "A skyscraper for a humid coastal city" The entropy wall $\rho_\star = H(P_{\text{data}})$, i.e., the entropy of the real data distribution, separates data-consistent imitated models from higher-diversity imagined distributions.

2 Preliminaries

In this section, we introduce the representation-level and distributional quantities used throughout the paper. We associate every distribution with a normalized kernel covariance matrix whose spectrum defines our notion of diversity and introduce a differentiable smoothed surrogate of the resulting entropy together with its per-sample energy. Extended conventions and an elementary Gibbs-tilt identity are deferred to Appendix C. Throughout this work, P_{data} denotes the population data distribution and $\widehat{P}_N = \frac{1}{N}\sum_{i=1}^{N} \delta_{X_i}$ the empirical measure of N i.i.d. samples from it; the upper-case letter N is reserved for generic empirical sample counts (evaluation batches, generated minibatches), and the lower-case letter n for the size of the training set, whose empirical measure we denote with $\widehat{P}_n$.

2.1 Representation Space and Kernel Covariance Matrix

Consider a measurable representation map $\phi : \mathcal{X} \to \mathbb{R}^d$ satisfying $\|\phi(x)\|_2 = 1$ for every $x \in \mathcal{X}$, which induces the normalized kernel $k(x, x') = \phi(x)^\top \phi(x')$. For a probability distribution Q on $\mathcal{X}$, we associate with the representation its *kernel covariance matrix*

$$\Sigma_Q := \mathbb{E}_{X\sim Q}\big[\phi(X)\phi(X)^\top\big] \in \mathbb{R}^{d\times d}.$$

The unit-norm normalization makes Σ_Q a density matrix, i.e., a positive semidefinite matrix with unit trace, whose spectrum records how the representation of Q distributes its mass across orthogonal feature directions. The covariance spectrum carries the notion of diversity developed next.

The same spectrum is accessible from pairwise similarities. Given samples $x_1, \ldots, x_N$ with Gram matrix $K = [k(x_i, x_j)]_{i,j=1}^N$, the empirical kernel covariance

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

shares its nonzero eigenvalues with $\frac{1}{N}K$. Consequently, every spectral quantity introduced below can be computed from the normalized Gram matrix without explicitly forming the feature vectors.

2.2 Spectral Entropy Measures and Diversity Scores

Following the discussion in [13, 7, 8], we use the following definition for the *von Neumann entropy* of a distribution Q as

$$H_0(Q) := -\operatorname{Tr}\big(\Sigma_Q \log \Sigma_Q\big), \tag{6}$$

which is the Shannon entropy of the covariance spectrum. We note that the exponential of the above quantity $\text{Vendi}(Q) := \exp(H_0(Q))$ is the Vendi score [7]. This can be interpreted as an effective number of occupied feature directions. For example, a spectrum uniform over r orthogonal directions yields $H_0(Q) = \log r$ and $\text{Vendi}(Q) = r$.

This notion of diversity is inherently reference-free, yet it depends on the representation model to embed the data: evaluating $H_0(Q)$ requires no comparison distribution, while the fixed choice of ϕ, or equivalently k, determines which variations count as distinct. Moreover, since $S \mapsto -\operatorname{Tr}(S \log S)$ is concave on density matrices and $Q \mapsto \Sigma_Q$ is affine, H_0 is concave in Q (Lemma 2). At rank-deficient covariance matrices, however, H_0 need not be differentiable.

The guidance analysis of Section 5 requires a well-defined first variation, so we move the covariance uniformly away from the boundary of the density-matrix cone. For $\varepsilon \in (0, 1)$, we define the smoothed covariance and entropy

$$S_Q^\varepsilon := (1-\varepsilon)\Sigma_Q + \varepsilon\,\frac{I_d}{d}, \qquad H_\varepsilon(Q) := -\operatorname{Tr}\big(S_Q^\varepsilon \log S_Q^\varepsilon\big). \tag{7}$$

The spectral floor $S_Q^\varepsilon \succeq \frac{\varepsilon}{d} I_d$ places every eigenvalue in $[\frac{\varepsilon}{d},\, 1-\varepsilon(1-\frac{1}{d})]$ and guarantees that H_ε is differentiable throughout the feasible covariance set.

The derivative of H_ε acts on individual samples through one quantity that recurs at every stage of the paper. We define the *entropy energy*

$$G_Q^\varepsilon(x) := (1-\varepsilon)\,\phi(x)^\top\big(-\log S_Q^\varepsilon\big)\phi(x), \tag{8}$$

which is the first variation of H_ε at Q (Lemma 1, Section 5). Since $-\log S_Q^\varepsilon$ has large eigenvalues precisely where S_Q^ε has small ones, $G_Q^\varepsilon(x)$ is large when $\phi(x)$ aligns with feature directions that Q underrepresents, and the spectral floor gives the uniform bound $0 \le G_Q^\varepsilon(x) \le (1-\varepsilon)\log(d/\varepsilon)$. The energy reappears as the payoff of the spectral adversary at training time (Section 3) and as the exponent of the guidance tilt at sampling time (Section 5).

The two entropy functionals play distinct roles in our analysis: Section 4 uses H_0 to define population spectral diversity, whereas Section 5 uses H_ε to derive the guidance potential. Lemma 3 quantifies their uniform proximity as ε becomes small.

3 The IGA Framework: From Imitative to Imaginative Generative Modeling

Distributional imitation, which is mathematically formulated in (1), requires a generative model to match a reference distribution. In the IGA framework which we formulate in this work, we intentionally augment this objective by requiring the generated distribution to **attain a prescribed level of spectral diversity**, and the resulting formulation separates two questions: (i) what distribution should be targeted, and (ii) how should that target be realized by a generative model?

In what follows, we first present the distribution-level objective function and the regularized form in IGA. We then derive a min-max reformulation of the entropy reward in the IGA optimization, and then develop the framework's application to sampling-time and training-time settings.

3.1 Formulating IGA via Constraining and Penalizing the Spectral Entropy

Consider a reference distribution P_{ref}, which in our applications is either the empirical data distribution $\widehat{P}_n$ for the observed training data or the underlying distribution P_θ of the (already trained) generative model. We also consider the spectral entropy functional H. Given a divergence measure $\mathcal{D}(Q; P_{\text{ref}})$, IGA selects the most faithful distribution whose spectral diversity reaches a target level $\rho \in \mathbb{R}$:

$$\begin{aligned} \underset{Q\in\mathcal{P}}{\text{minimize}} \quad & \mathcal{D}(Q; P_{\text{ref}}) \\ \text{subject to} \quad & H(Q) \ge \rho. \end{aligned} \tag{C_ρ}$$

The divergence objective function anchors the solution to P_{ref}, ensuring the distribution solution remains as close as possible to the reference distribution, while the constraint specifies the desired spectral diversity level. In next section, we review and extend the discussion from [6] to interpret ρ relative to the entropy of the data, distinguishing diversity repair from spectral extrapolation.

We note that due to the convex structure of the above optimization, the application of standard convex duality shows that (C_ρ) is equivalent to the Lagrangian formulation for a corresponding Lagrangian multiplier parameter $\lambda \ge 0$:

$$Q_\lambda = \underset{Q\in\mathcal{P}}{\operatorname{argmin}}\, F_\lambda(Q) := \mathcal{D}(Q; P_{\text{ref}}) - \lambda H(Q) \tag{P_λ}$$

Remark 1. *For every constrained solution satisfying the standard regularity conditions, there is a multiplier $\lambda \geq 0$ for which the same distribution solves (P_λ). Conversely, each Q_λ solves (C_ρ) at its attained diversity level. The precise duality and attainment statements are given in Appendix D.*

A Min-Max Formulation of the IGA Optimization. The spectral entropy term in (P_λ) is a nonlinear function of the covariance spectrum, yet it admits a precise convex dual formulation due to its concavity. For the smoothed entropy, the application of the Gibbs variational principle for the matrix-based entropy converts the smoothed IGA problem into a two-player game.

Proposition 1 (Spectral Min-Max Formulation of IGA Optimization)**.** *Consider smoothed spectral entropy with parameter $\varepsilon \in (0,1)$. Define the spectral adversary class as*

$$\mathbb{T}_\varepsilon := \big\{\Theta \in \mathbb{R}^{d\times d} :\ \Theta = \Theta^\top,\ \mathrm{Tr}(\Theta) = 0,\ \|\Theta\|_{\mathrm{op}} \leq \log(d/\varepsilon)\big\}.$$

Then, the following equivalences hold:

(i) ***Dual representation of the negative entropy.*** *For every distribution Q on $\mathcal{X}$, we have*

$$-H_\varepsilon(Q) = \max_{\Theta\in\mathbb{T}_\varepsilon}\Big\{-(1-\varepsilon)\,\mathbb{E}_{X\sim Q}\big[\phi(X)^\top\Theta\,\phi(X)\big] - \log\mathrm{Tr}\big(e^{-\Theta}\big)\Big\}, \tag{9}$$

and the maximum is attained at the unique traceless matrix

$$\Theta^\star(Q) := -\log S_Q^\varepsilon + \tfrac{1}{d}\,\mathrm{Tr}\big(\log S_Q^\varepsilon\big) I_d \in \mathbb{T}_\varepsilon.$$

(ii) ***Min-Max form of*** (P_λ)***.*** *For every $\lambda \geq 0$, distribution P_{ref}, and every $Q \in \mathcal{P}$, we have*

$$\mathcal{D}(Q;P_{\mathrm{ref}}) - \lambda H_\varepsilon(Q) = \max_{\Theta\in\mathbb{T}_\varepsilon}\ \mathcal{A}_\lambda(Q,\Theta), \tag{10}$$

where $\mathcal{A}_\lambda(Q,\Theta) := \mathcal{D}(Q;P_{\mathrm{ref}}) - \lambda(1-\varepsilon)\,\mathbb{E}_{X\sim Q}\big[\phi(X)^\top\Theta\,\phi(X)\big] - \lambda\log\mathrm{Tr}\big(e^{-\Theta}\big)$.

Therefore, the above results show that (P_λ) can be rewritten as the following two-player min-max problem (game):

$$\min_{Q\in\mathcal{P}}\max_{\Theta\in\mathbb{T}_\varepsilon}\mathcal{A}_\lambda(Q,\Theta)$$

(iii) ***Min-Max and Max-Min Equivalence.*** *If $\mathcal{P}$ is a convex set and $\mathcal{D}(\cdot;P_{\mathrm{ref}})$ is a convex and lower semicontinuous function, the order of minimization and maximization in this game may be interchanged:*

$$\min_{Q\in\mathcal{P}}\max_{\Theta\in\mathbb{T}_\varepsilon}\mathcal{A}_\lambda(Q,\Theta)\ =\ \max_{\Theta\in\mathbb{T}_\varepsilon}\min_{Q\in\mathcal{P}}\mathcal{A}_\lambda(Q,\Theta)$$

Proof. We defer the proof to the Appendix. □

We note that the best response in part (i) is the centered log-spectrum of the smoothed covariance, and its per-sample payoff coincides, up to an additive constant, with the entropy energy: the spectral adversary pays the generator exactly λG_Q^ε of (8), rewarding samples along directions that Q underrepresents (Lemma 7). Therefore, the training-time adversarial payoff and the sampling-time guidance field are induced by the same first-order quantity. The proof of Proposition 1, given in Appendix D.1, relies on Klein's matrix relative-entropy inequality and the Gibbs variational principle for matrix entropy, both of which are proved there in full. We refer to Remark 8 for further structural implications of (10), including the linearization of the entropy reward and the role of smoothing in compactifying the adversary class.

3.2 IGA for Sampling From a Pretrained Model: the Special case of KL-divergence

This subsection presents the application of IGA for sampling from an available (supposedly pretrained) model. At sampling time, the pretrained model is held fixed as the reference distribution in the framework. We specifically consider the reference P_θ as the *underlying* distribution of the pretrained generator, and the Lagrangian IGA optimization problem becomes:

$$Q^\star = \operatorname*{argmin}_{Q\in\mathcal{P}} \Big\{ D(Q, P_\theta) - \lambda H_\varepsilon(Q) \Big\}. \tag{11}$$

Here, our goal is to sample from the optimal distribution $Q^\star$.

We recall that (11) separates the desired target distribution from the algorithm used to sample it: the Bregman divergence determines which departures from the base generator are costly, while the H_ε Lagrangian penalty rewards higher spectral diversity.

In our analysis, we specifically focus on the geometry resulting from choosing the discrepancy measure to be the **KL-divergence**. In this specific case, the optimality conditions yield an explicit density-ratio characterization of the optimal solution. Note that, in the case of KL-divergence, the sampling-based IGA aims to minimize the following objective function:

$$F(Q) := \mathrm{KL}(Q\|P_\theta) - \lambda H_\varepsilon(Q) \tag{12}$$

over $Q \in \mathcal{P}$, where the extended-value convention for the KL term sets $F(Q) = +\infty$ off the set $\{Q \ll P_\theta\}$.

Proposition 2 (Sampling-time target as a self-consistent exponential tilt)**.** *Let $\lambda \geq 0$ and $\varepsilon \in (0, 1)$. If F attains a finite minimum over $\{Q \in \mathcal{P} : Q \ll P_\theta\}$, then the minimizer $Q^\star$ is unique, $Q^\star$ and P_θ are mutually absolutely continuous, and, with the* total reward *defined by*

$$R_Q(x) := \lambda\, G^\varepsilon_Q(x), \tag{13}$$

the density ratio is the exponential tilt

$$\frac{dQ^\star}{dP_\theta}(x) = \frac{\exp\big(R_{Q^\star}(x)\big)}{\mathbb{E}_{X\sim P_\theta}\big[\exp\big(R_{Q^\star}(X)\big)\big]} \tag{14}$$

Proof. We defer the proof to the Appendix. □

Corollary 1 (Score-function relation under the exponential tilt)**.** *Under the assumptions of Proposition 2, suppose that P_θ and $Q^\star$ admit differentiable densities p_θ and $q^\star$, respectively. Then their score functions satisfy*

$$\nabla \log q^\star(x) = \nabla \log p_\theta(x) + \lambda\, \nabla G^\varepsilon_{Q^\star}(x). \tag{15}$$

Proof. The result follows directly by taking the logarithm of (14) and differentiating with respect to x, noting that the log-normalizing constant is independent of x. □

The target is an exponential reweighting of the pretrained law. We highlight that the reweighting is self-consistent rather than externally prescribed, since the total reward $R_{Q^\star}$ depends on the covariance of the unknown target itself. The multiplier λ sets the strength of the reweighting, and the uniform bound on the energy noted after (8) limits how strongly any single sample can be up- or down-weighted, while the KL term confines the redistribution of mass to the support of the base law.

We note that Proposition 2 characterizes the target distribution over clean outputs and does not yet provide a sampler; also, the finite-minimum hypothesis remains to be verified. Section 5 addresses both points: under mild topological conditions the minimizer exists (Theorem 1), the tilt propagates exactly

through the forward noising process to an explicit time-dependent guidance field (Theorem 2), and the field is approximated with denoised predictions at a quantified endpoint error, all without changing the pretrained score network. Unlike guidance by a fixed sample-wise reward, the tilt depends on the target law itself, through its covariance; practical sampling therefore estimates this distribution-level quantity, for example from a pilot batch. We keep the sampling target $Q^\star$ notationally distinct from the training-time optimum $Q_\lambda^{\mathrm{train}}$ of Section 3.3, realized by changing generator parameters.

3.3 IGA for Training Generative Models with Entropy-Regularized Objective

For the training-time application of the imaginative generative modeling in IGA, we change the original divergence minimization in standard generative modeling and include the additional Lagrangian term in the objective function $-\lambda H(Q)$ to promote higher spectral entropy in the trained model.

Mathematically, we choose the reference distribution to be the empirical distribution $\widehat{P}_n$ of n training samples $x_1, \ldots, x_n$ (i.e., $\widehat{P}_n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$). Then, the optimization problem for training-time IGA will be computing the optimal solution to the spectral entropy-regularized divergence minimization problem:

$$Q_\lambda^{\mathrm{train}} \;=\; \operatorname*{argmin}_{Q\in\mathcal{P}_{\mathrm{gen}}} \Big\{\mathcal{D}(Q;\widehat{P}_n) - \lambda H(Q)\Big\}. \tag{16}$$

The first term specifies how divergence to the training data distribution is measured, while the second is a distribution-level regularizer: it acts jointly on generated examples and rewards coverage of feature directions that would otherwise be underrepresented. IGA therefore only augments the model's distributional discrepancy objective and can be interpreted as a spectral entropy regularization in the divergence minimization task of training the generative model. Specifically, in the following, we focus on and apply the IGA training framework to the adversarial training of generative adversarial networks (GANs). Further discussion on application of training-time IGA to other generative modeling frameworks is deferred to the Appendix.

Adversarial training and GANs. The min-max format appearing in Proposition 1 composes smoothly with objective functions that are formed in the adversarial-learning formulations of generative modeling. We note that the standard GAN [1, 14, 15] objectives measure the discrepancy through a critic (discriminator) class $\mathcal{D}_{\mathrm{c}}$ and real-valued link functions $u, v : \mathbb{R} \to \mathbb{R}$:

$$\mathcal{D}(Q;\widehat{P}_n) = \max_{D\in\mathcal{D}_{\mathrm{c}}} \Big\{\mathbb{E}_{X\sim\widehat{P}_n}\big[u\big(D(X)\big)\big] - \mathbb{E}_{X\sim Q}\big[v\big(D(X)\big)\big]\Big\}. \tag{17}$$

As notable examples, Wasserstein GANs take a 1-Lipschitz critic class with $u = v = \mathrm{id}$ being the identity map [15]; f-GANs choose $v = f^*$ for the convex conjugate of the convex f function underlying the target f-divergence, recovering the original GAN objective as a special case for JS-divergence [1, 14]. Substituting (17) and the entropy dual (9) into (16) gives an exact reformulation in which the entropy reward joins the discriminator inside a single adversary.

Proposition 3 (IGA-GAN formulation as min-max optimization)**.** *Let $\mathcal{D}(\cdot;\widehat{P}_n)$ admit the representation (17), and let $\lambda \geq 0$, $\varepsilon \in (0,1)$. Then, for every class $\mathcal{Q}$ of distributions, we have the following*

$$\begin{aligned}\min_{Q\in\mathcal{Q}}\big\{\mathcal{D}(Q;\widehat{P}_n) - \lambda H_\varepsilon(Q)\big\} \;=\; \min_{Q\in\mathcal{Q}} \max_{(D,\Theta)\in\mathcal{D}_{\mathrm{c}}\times\mathbb{T}_\varepsilon} \Big\{&\mathbb{E}_{X\sim\widehat{P}_n}\big[u\big(D(X)\big)\big] - \mathbb{E}_{X\sim Q}\big[v\big(D(X)\big)\big] \\ &- \lambda(1-\varepsilon)\,\mathbb{E}_{X\sim Q}\big[\phi(X)^\top \Theta\,\phi(X)\big] - \lambda\log\mathrm{Tr}\big(e^{-\Theta}\big)\Big\},\end{aligned} \tag{18}$$

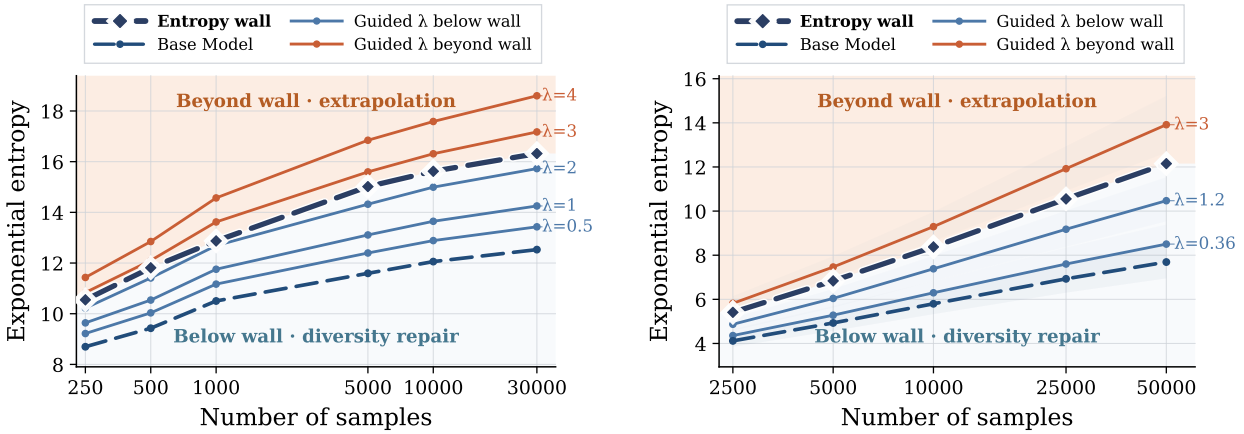
Note that the above has a single maximization over the joint adversary (D,Θ). For every fixed Q the joint maximization decouples across the two components, and the Θ-component is attained at the best response $\Theta^\star(Q)$ of Proposition 1.

Proof. We defer the proof to the Appendix. □

The proof, given in Appendix D.2, relies on a structural observation: the critic and the spectral adversary enter through suprema over independent variables, and such suprema combine additively. Hence, the identity holds pointwise in Q, and neither convexity of $\mathcal{P}_{\text{gen}}$ nor a minimax interchange is used; the interchange remains reserved for the ambient class (Remark 3, Appendix C). Algorithmically, (18) adds one adversary to standard GAN training, and this additional adversary is computationally inexpensive: while the critic D is trained by gradient steps, the spectral player requires no training at all, since its best response is the centered log-spectrum $\Theta^\star(Q)$ and can be computed from an eigendecomposition of the minibatch covariance at $O(d^3)$ cost per refresh.

The two adversaries play complementary roles. The critic enforces the fidelity of individual samples by comparing generated examples against data, whereas Θ acts on the generated distribution as a whole and pays the generator the spectral novelty reward $\lambda G^{\varepsilon}_{Q}(x)$, up to a sample-independent constant, for occupying directions that the current generated law neglects (Lemma 7). We also highlight that freezing Θ at its best response is not a heuristic: the gradients of the generator parameters through the frozen payoff are *exactly* the gradients of $\lambda H_\varepsilon(Q_\vartheta)$ (Proposition 7, Appendix D.2). This envelope-type identity removes the need to differentiate through the eigendecomposition.

4 The Entropy Wall: Spectral Entropy of Real Data as the Boundary between Imitation and Imagination



(a) CelebA-HQ (b) ImageNet

Figure 4: **The Entropy walls on CelebA-HQ and ImageNet.** Real and generated entropy estimates use matched sample sizes. Both base models remain below the data wall; increasing IGA λ closes the deficit and eventually crosses into the imagination regime.

As we discussed in the introduction, the entropy wall is the spectral diversity of the data itself, measured in the chosen representation. We note that the concept of the entropy wall is implied by the discussion in [6], in which the authors reveal the spectral entropy gap between the standard generative models and their target underlying data distributions. In this section, we formalize the concept and propose the term *"Entropy Wall"* to highlight the spectral entropy level of the underlying real data distribution.

Particularly, we highlight that the definition of entropy wall separates two qualitatively different uses of the IGA regularization framework: as long as the required spectral diversity level in IGA stays at or below what the data exhibits, increasing diversity can be read as *repairing* a deficiency of the learned generator; once

the request exceeds it, the data distribution is no longer feasible for (C_ρ), and pushing further is deliberate *extrapolation* beyond the data. Like every diversity statement in this paper, the wall's location depends on the fixed pair (ϕ, k) and is therefore representation-relative.

Definition 1 (Entropy Wall)**.** For spectral entropy function H, the entropy wall is

$$\rho_{\star,H} := H(P_{\text{data}})$$

Based on the above definition, a distribution Q is below, on, or beyond the entropy wall according as $H(Q)$ is $<$, $=$, or $> \rho_{\star,H}$.

Reading the regularization path of Theorem 4 through the wall gives it the statistical interpretation promised in the introduction. Note that we use the notation Q_λ for the optimal solution to the problem with Lagrangian coefficient λ.

Below the wall: diversity repair. When $H(Q_\lambda) \leq H(P_{\text{data}})$, increasing λ moves Q_λ toward higher diversity, and this movement is provably safe for a base-anchored, correctly oriented Bregman objective: every below-wall point on the path is no farther from P_{data}, in the anchoring divergence, than the base model is (Theorem 6, Appendix E). This is the precise sense in which the sub-wall path performs *repair*, counteracting the spectral contraction reported in modern generators [6]. It does not identify Q_λ with P_{data}, nor does it guarantee that every induced semantic change recovers a genuine data mode.

Beyond the wall: spectral extrapolation. When $H(Q_\lambda) > H(P_{\text{data}})$, the same monotone increase in λ means something different: Q_λ is no longer estimating P_{data} but performing representation-relative extrapolation, spreading its mass across directions of the representation more broadly than the data does. The transition is a change of statistical interpretation, not a geometric barrier: the wall can be crossed at arbitrarily small discrepancy whenever a higher-entropy direction exists in $\mathcal{P}$ and the discrepancy is continuous along the mixture path toward it (Proposition 9, Appendix E). Beyond the wall we claim no improved estimation of the data distribution; the regime is evaluated as controlled, representation-relative extrapolation.

5 IGA Guidance for Score-based and Diffusion Models

Here, we focus on diffusion models and apply sampling-time IGA to pre-trained score-based and diffusion models. The developments in this section largely build on the score-function characterization established in Corollary 1. We first establish existence and uniqueness of the tilted target $Q^\star$, then propagate the tilt through the forward noising process, and finally derive retraining-free approximations for score-based, DDPM, and DDIM samplers.

5.1 The target distribution and exact guidance

Assumption 1 (Sampling-time setup)**.** *Given the sample space $\mathcal{X}$, the representation map ϕ is continuous, hence bounded by the unit-norm normalization. P_θ is the original distribution of the pretrained generator. Also, as stated in previous sections, we suppose the hyperparameters satisfy $\lambda \geq 0$, $\varepsilon \in (0, 1)$.*

Under Assumption 1, we minimize the functional F of (12) over $\mathcal{P}$. The KL anchor makes the problem tractable for two reasons: the effective domain $\{Q \ll P_\theta\}$ is convex, and the first variation of $\mathrm{KL}(\cdot\|P_\theta)$ along a density perturbation is simply the log-density ratio. The main difficulty is that H_ε is highly nonlinear in Σ_Q: although Σ_Q is affine in Q, the matrix logarithm couples all eigenvalues. The following lemma resolves this difficulty and justifies the role assigned to the entropy energy at its definition (8); it is the analytical core of Proposition 2 and of the existence theorem below.

Figure 5: **Fashion design with vanilla and IGA SDXL.** Left: vanilla SDXL. Right: IGA SDXL. Corresponding cells use the same initial noise seed. Vanilla samples cluster around beige and gold eveningwear with familiar gown and tailored shapes. IGA adds bright color blocking, asymmetric cuts, mixed materials, and large sculptural or feathered elements.

Figure 6: **Architectural design with vanilla and IGA SDXL.** Left: vanilla SDXL. Right: IGA SDXL. Corresponding cells use the same initial noise seed. Vanilla SDXL mainly produces straight glass towers with similar overall forms. IGA introduces curved shells, open frames, split tops, stacked blocks, and larger changes in color and proportion, while preserving a clear full-building view.

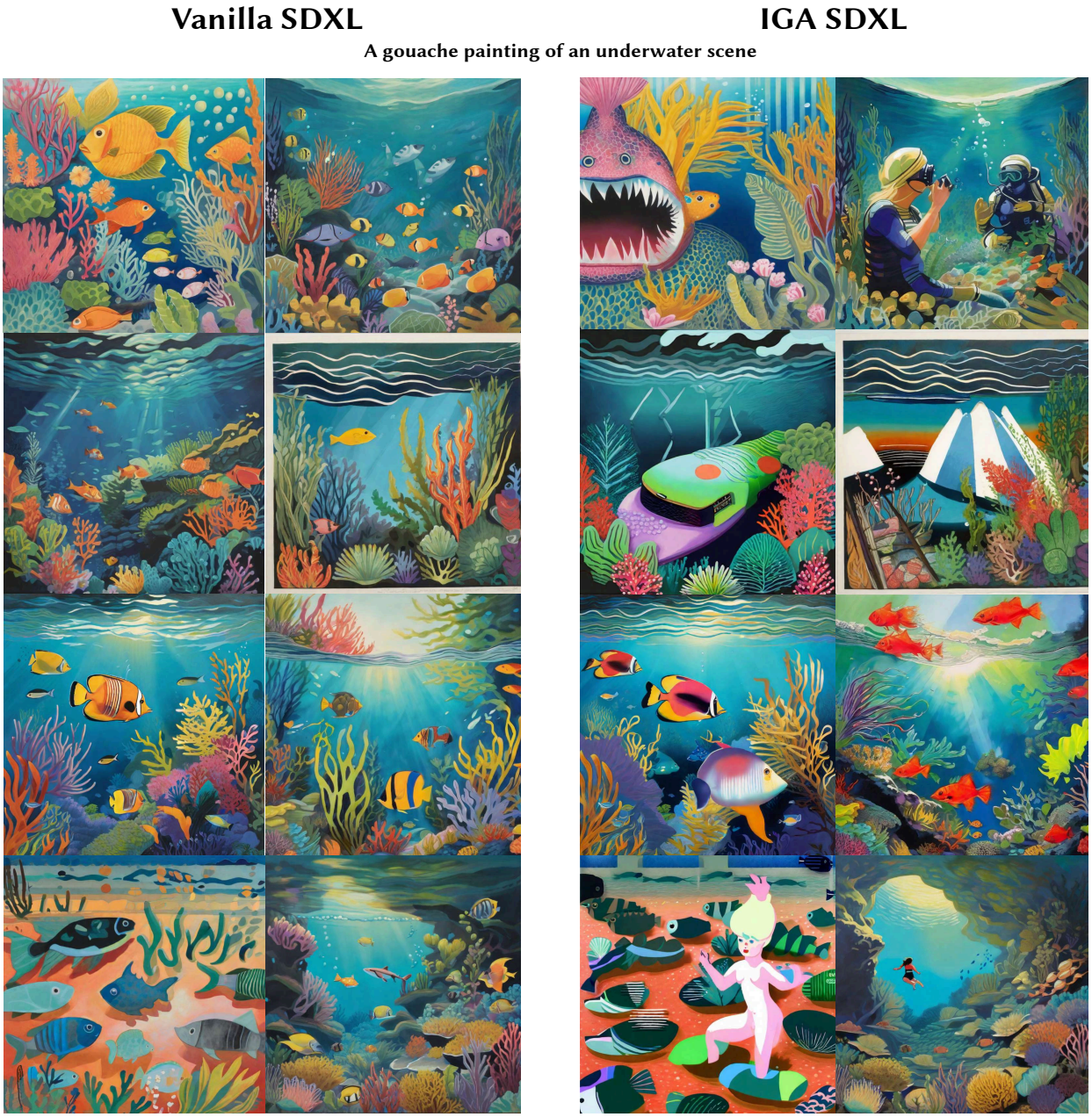


Figure 7: **Stylized underwater scenes with vanilla and IGA SDXL.** Left: vanilla SDXL. Right: IGA SDXL. Corresponding cells use the same initial noise seed. Vanilla SDXL mostly depicts coral reefs and schools of fish. IGA expands the scene content to divers, large creatures, vehicles, built structures, and cave-like spaces, while keeping the gouache rendering style.

Figure 8: **Fantasy throne design with vanilla and IGA SDXL.** Left: vanilla SDXL. Right: IGA SDXL. Corresponding cells use the same initial noise seed. Vanilla SDXL mostly returns ornate high-backed chairs with similar carved frames. IGA introduces spiked metal forms, curved black shells, moss-covered structures, and larger changes in the seat and back, while keeping the throne centered and fully visible.

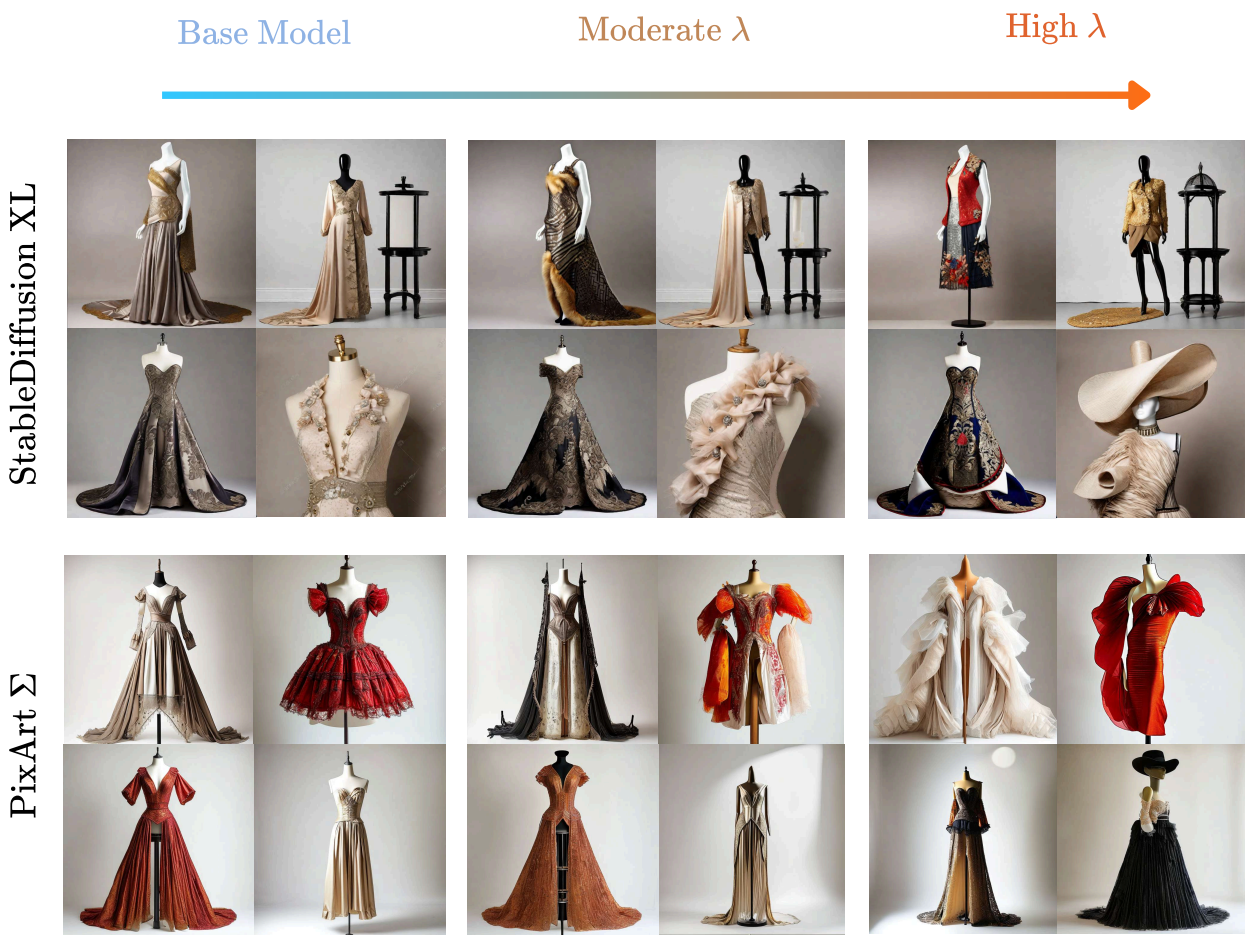


Figure 9: **IGA λ-sweep across SDXL and PixArt-Σ.** Top: SDXL with $\lambda \in \{0, 4, 8\}$. Bottom: PixArt-Σ with $\lambda \in \{0, 10, 20\}$. Within each model, corresponding positions use the same initial noise seed. As λ increases, both models move from familiar dress shapes toward stronger asymmetry, larger volumes and accessories, and wider material and color choices, while staying consistent with the prompt.

Lemma 1 (First variation of the smoothed entropy)**.** *Let ϕ be bounded and measurable with $\|\phi(x)\|_2 = 1$, and let G_Q^ε be the entropy energy* (8)*. Then for every distribution Q and every finite signed measure ν with $\nu(\mathcal{X}) = 0$ such that $Q + t\nu$ is a probability measure for all sufficiently small $t > 0$,*

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

The lemma therefore identifies the total reward R_Q of (13) as the first variation of the reward part $\lambda H_\varepsilon(Q)$ of the objective (proof and further discussion in Appendix F).

Theorem 1. *Under Assumption 1, the following hold for the functional F of* (12)*.*

(i) ***Existence and uniqueness.*** *F has a unique minimizer $Q^\star$, and $Q^\star$ is mutually absolutely continuous with P_θ.*

(ii) ***Self-consistent exponential tilt.*** *$Q^\star$ satisfies the tilt characterization* (14)*.*

(iii) ***Boundedness.*** *$R_{Q^\star}$ is uniformly bounded; consequently $dQ^\star/dP_\theta$ is bounded above and below by positive constants, and $Q^\star$ has the same P_θ-essential support as P_θ.*

Proof. We defer the proof to the Appendix. □

(14) is exactly the tilt announced in (5), with $G_{Q^\star} = G^{\varepsilon}_{Q^\star}$. The theorem verifies the finite-minimum hypothesis of Proposition 2 rather than assuming it: the proof, given in Appendix F, establishes existence by the tightness of KL sublevel sets on a Polish space together with weak lower semicontinuity, and then inherits uniqueness, mutual absolute continuity, and the tilt formula from that proposition. What the theorem adds is that the target is well defined without any attainment hypothesis and that the reward is uniformly bounded, so the tilt redistributes mass within the P_θ-essential support and creates none (Remark 12).

The tilt (14) concerns the clean distribution over x_0, but diffusion samplers generate x_0 as the endpoint of a denoising process that starts from noise at time T, so guidance must be injected at every noise level t. Then, we let $w(x_0) = \exp(R_{Q^\star}(x_0))$ and $Z = \mathbb{E}_{P_\theta} w(X_0)$, so that $Q^\star(dx_0) = \frac{1}{Z} w(x_0) P_\theta(dx_0)$, and let $K_t(dx_t \,|\, x_0)$ denote the forward noising kernel with time-t marginals $p_t, q_t^\star$ under $P_\theta, Q^\star$.

Theorem 2 (Derivation of the exact guidance field)**.** *Let $h_t(x_t) = \mathbb{E}_{P_\theta}[w(X_0) \mid X_t = x_t]$. Then $\frac{dq_t^\star}{dp_t} = \frac{1}{Z} h_t$, and if $p_t, q_t^\star$ admit positive differentiable densities, then we have*

$$\nabla \log q_t^\star(x_t) = \nabla \log p_t(x_t) + u_t(x_t), \qquad u_t(x_t) := \nabla_{x_t} \log h_t(x_t). \tag{19}$$

Note that the function h_t averages the clean-sample reward w over all origins that the base posterior regards as plausible for x_t. We emphasize that h_t is *not* the clean tilt evaluated at a denoised point estimate, and this distinction is what makes the identity exact: applying the tilt before the noising process does not commute with applying it afterward. Adding $u_t = \nabla \log h_t$ to the base score yields a reverse process whose marginals match $q_t^\star$ at *every* noise level (Theorem 7, Appendix F); for continuous-time samplers, the guided score can be used directly in the reverse SDE or the probability-flow ODE.

One caveat accompanies this exactness. The exact reverse process must be initialized at $q_T^\star$, which is not directly samplable, whereas practical samplers initialize from p_T (typically Gaussian noise). The two distributions coincide only when h_T is constant, i.e. when the terminal noise level has erased all reward information. The resulting mismatch enters the end-to-end bound of Theorem 8 as the initialization term $\mathrm{KL}(p_T \| q_T^\star)$, and its magnitude is quantified in Remark 13.

5.2 Practical diffusion guidance

There exist three approximation items that separate the discussed theoretical framework from an implementable sampler. In the following, we make each one explicit and discuss how to address it.

Plug-in guidance fields. The exact field $u_t = \nabla \log h_t$ requires the gradient of a conditional log-moment-generating function under the base posterior $P_\theta(dx_0 \mid x_t)$, which is generally intractable. What is available at every noise level is a denoiser $\widehat{x}_0(x_t, t) \approx \mathbb{E}[X_0 \mid X_t = x_t]$, and the *plug-in* approximation substitutes this point estimate for the posterior average. Two variants differ in how the reward is turned into a vector field: the *chain-rule* variant differentiates $x_t \mapsto R_{Q^\star}(\widehat{x}_0(x_t, t))$ through the denoiser Jacobian $J_{\widehat{x}_0}$, while the cheaper *direct-injection* variant reuses the clean-space gradient as a direction in noisy-sample space:

$$\widetilde{u}_t^{\mathrm{chain}}(x_t) = \omega_t J_{\widehat{x}_0}(x_t, t)^\top \nabla_x R_{Q^\star}\big(\widehat{x}_0(x_t, t)\big), \qquad \widetilde{u}_t^{\mathrm{dir}}(x_t) = \omega_t \nabla_x R_{Q^\star}\big(\widehat{x}_0(x_t, t)\big),$$

with guidance scale $\omega_t \geq 0$. Both are heuristics without a general error bound: $R_{Q^\star}$ is nonlinear, and neither conditional expectation nor differentiation commutes with a point-mass substitution (Definition 2, Remark 14). We always report which variant is used.

Estimating the self-referential reward. The reward $R_{Q^\star}$ depends on the unknown covariance $S^{\varepsilon}_{Q^\star}$, so a practical sampler replaces it by an estimate, and how the estimate is maintained determines the statistical status of the outputs. If the covariance is *frozen*, i.e., computed once from a pilot batch and used to define a single estimated potential $\widehat{R}$ for all subsequent trajectories, the draws are conditionally i.i.d. from the frozen-potential law. If instead the covariance is recomputed on the fly from the batch being generated, each particle's drift depends on the others, and the outputs form an exchangeable but non-i.i.d. interacting particle system (Remark 16). Frozen-potential estimation is therefore the setting in which IGA guidance can be described as sampling from a well-defined target distribution.

Discrete sampler updates. Once a guidance field $\widetilde{u}_t$ is chosen, it is converted to a correction on the noise prediction. In the ε-prediction parameterization, with the standard noise–score convention $\nabla \log p_t(x_t) = -\varepsilon_\theta(x_t,t)/\sqrt{1-\bar{\alpha}_t}$ [5], guiding the score by $+\widetilde{u}_t$ corresponds to

$$\varepsilon^{\mathrm{IGA}}_\theta(x_t,t) = \varepsilon_\theta(x_t,t) - \sqrt{1-\bar{\alpha}_t}\,\widetilde{u}_t(x_t). \tag{20}$$

Substituting (20) into the DDPM posterior mean [4]

$$\mu_\theta = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\varepsilon_\theta\right)$$

and the DDIM update [16] gives

$$\begin{aligned} \mu^{\mathrm{IGA}}_\theta(x_t,t) &= \mu_\theta(x_t,t) + \frac{\beta_t}{\sqrt{\alpha_t}}\,\widetilde{u}_t(x_t), \\ x_{t-1} &= \sqrt{\bar{\alpha}_{t-1}}\,\widehat{x}^{\mathrm{IGA}}_0 + \sqrt{1-\bar{\alpha}_{t-1}-\sigma_t^2}\,\varepsilon^{\mathrm{IGA}}_\theta + \sigma_t z, \end{aligned} \tag{21}$$

with $\widehat{x}^{\mathrm{IGA}}_0 = (x_t - \sqrt{1-\bar{\alpha}_t}\,\varepsilon^{\mathrm{IGA}}_\theta)/\sqrt{\bar{\alpha}_t}$ and $\sigma_t = 0$ for deterministic DDIM; the $\sqrt{1-\bar{\alpha}_t}$ in (20) and the $\beta_t/\sqrt{\alpha_t}$ in (21) cancel algebraically, so the DDPM mean correction is exactly $+(\beta_t/\sqrt{\alpha_t})\widetilde{u}_t$. A model trained with v- or x_0-prediction is first converted to an equivalent ε_θ in the standard way (e.g. $\varepsilon_\theta = \sqrt{\bar{\alpha}_t}\,v_\theta + \sqrt{1-\bar{\alpha}_t}\,x_t$ for v-prediction). These discrete updates are implementations inspired by Theorem 2, and they do not exactly sample $Q^\star$ even when $\widetilde{u}_t = u_t$, because the reverse kernels are discretized (Remark 15). The end-to-end guarantee is given by Theorem 8 in Appendix F. This theorem bounds $\mathrm{KL}(\widehat{Q}\|Q^\star)$ and $\mathrm{TV}(\widehat{Q},Q^\star)$ for the deployed continuous-time process in terms of the initialization mismatch, the score error, and the guidance error; a separate discretization term is required for the implemented sampler.

6 Numerical Evaluation

Throughout our numerical study, we aim to empirically address the following questions:

1. Do widely used pretrained generative models show a diversity deficit and existence of the entropy wall?
2. Does our proposed framework address this deficit, and are there values of λ that reach and go beyond the entropy wall? If so, as predicted by Theorem 6, is there an initial repair region in which the pretrained model moves closer to the data distribution it was meant to imitate?
3. Does crossing the wall produce structured, novel, and imaginative variations across different pretrained generative models, including text-conditional models? We test our theory and hypothesis on CelebA-HQ and ImageNet, and then ask what the resulting variation looks like in a large text-conditioned model.

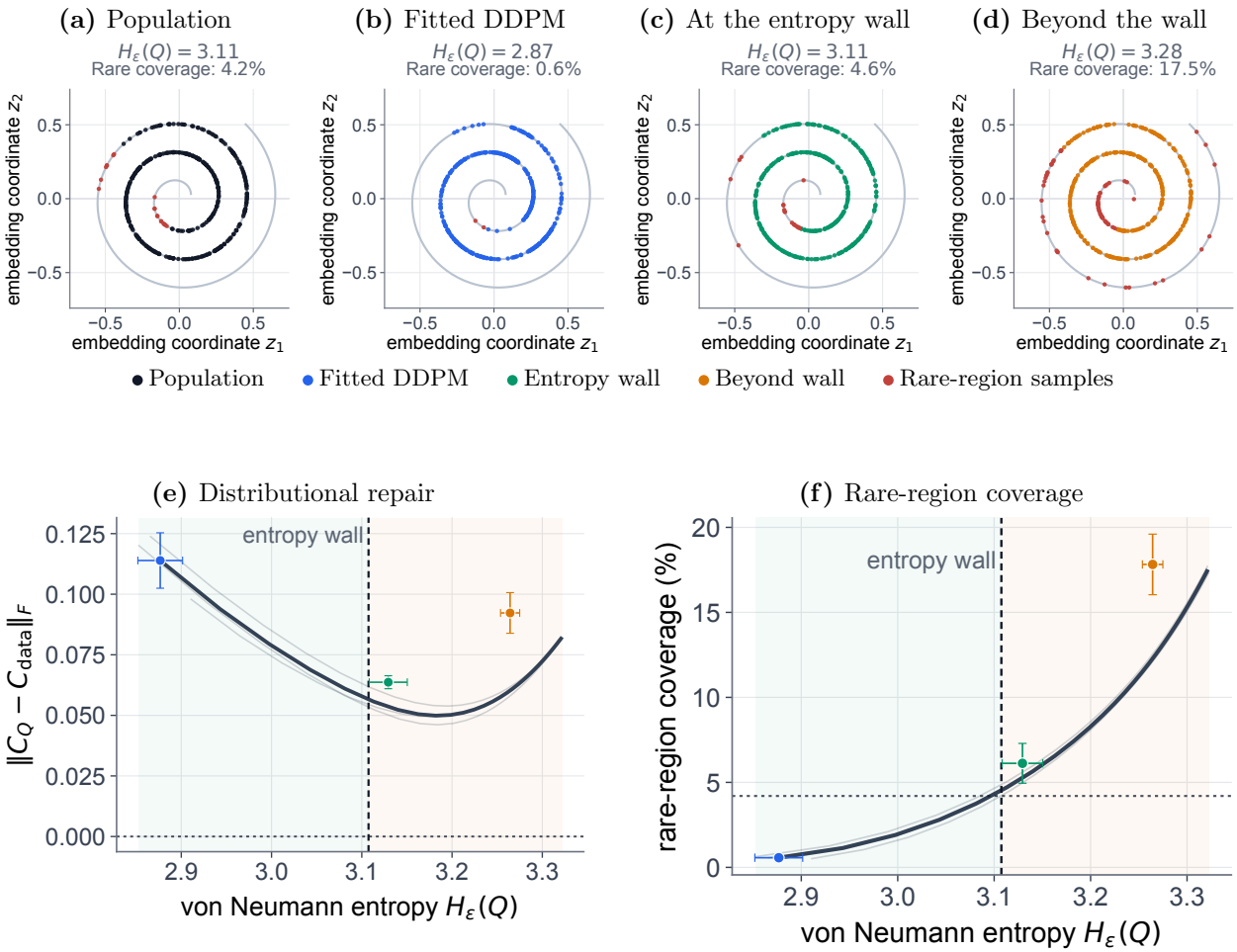


Figure 10: **Repair and extrapolation on a nonlinear manifold.** The fitted DDPM underrepresents the spiral endpoints. IGA restores population-level entropy and rare-region coverage near the wall and increases endpoint exploration beyond it. Panels (e,f) trace the covariance discrepancy and rare-region coverage; thin curves denote individual seeds and thick curves their mean.

6.1 Experimental Protocol

Our experiments cover sampling-time IGA on controlled synthetic distributions and real-world image benchmarks, a training-time study on MNIST, and qualitative text-conditioned generation with SDXL. Because these settings use different models and evaluation criteria, we state only the shared experimental conventions here and introduce the setting-specific configurations in the corresponding subsections.

Entropy representation and evaluation. Across all sampling-time experiments, we set the spectral floor of (7) to $\varepsilon = 10^{-3}$, for which Lemma 3 bounds the gap between H_ε and H_0 by 0.015 nats, several times smaller than the smallest entropy difference we report. IGA guidance uses the smoothed entropy H_ε, while, unless stated otherwise, we report the unsmoothed von Neumann entropy H_0 and its exponential, $\exp(H_0)$, corresponding to the Vendi score [7]. In all sampling-time experiments, we fix the coefficient of the IGA score correction to its theoretically prescribed value of one and vary only λ, so each point along the reported path corresponds to a different target $Q^\star_\lambda$, rather than to a different guidance strength.

For the real-image benchmarks, we use the CLS-token embeddings of DINOv2 ViT-B/14 [17] and approximate

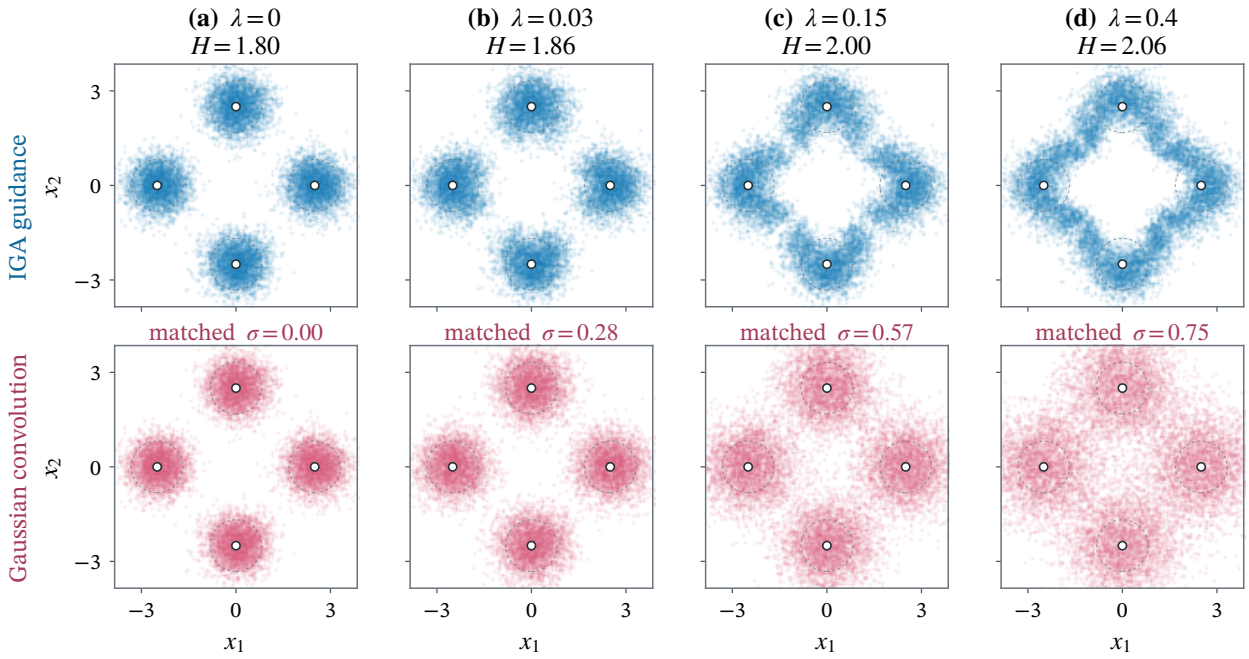


Figure 11: **IGA fills underrepresented inter-mode regions more coherently than entropy-matched noise.** Top: base and IGA-guided DDIM samples as λ increases. Bottom: Gaussian-convolved samples with σ selected to match the entropy of the corresponding IGA distribution. IGA connects the gaps between modes while preserving the original modal structure; Gaussian convolution broadens each mode isotropically.

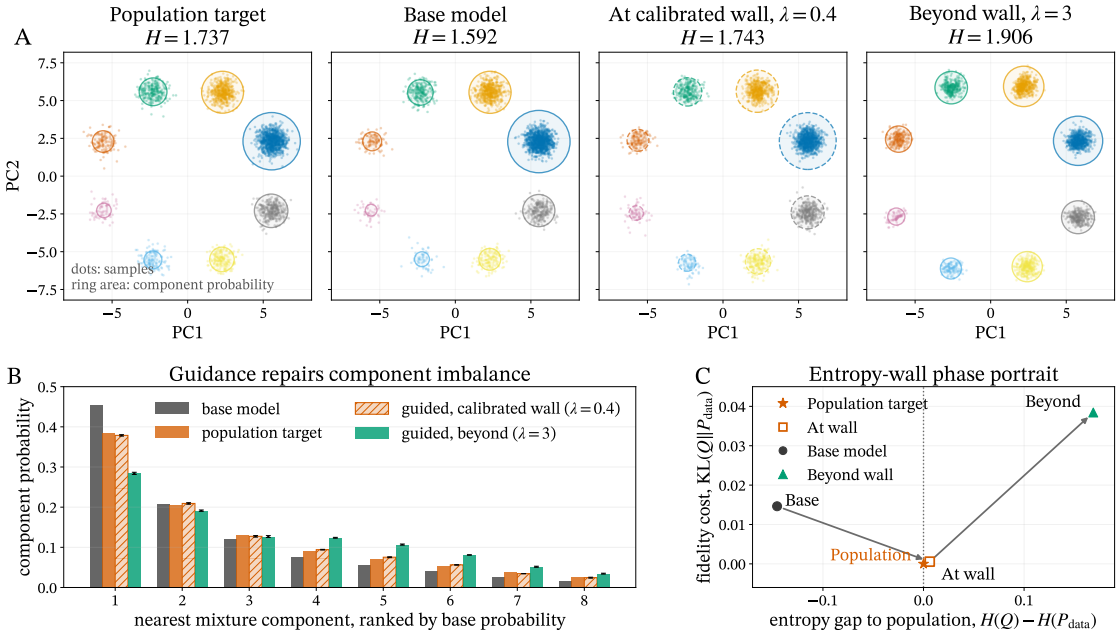


Figure 12: **The entropy wall separates repair from extrapolation in a controlled mixture. (A)** Population, base, calibrated-wall, and beyond-wall distributions. **(B)** IGA repairs the component imbalance at the wall and produces more uniform weights beyond it. **(C)** Population KL first decreases and then increases as the path crosses the entropy wall.

an RBF kernel on these embeddings using 1024 random Fourier features [18], with the RBF bandwidth selected by the median heuristic. Because empirical spectral entropy depends on the number of samples, we compare real and generated distributions using matched sample sizes when locating the empirical entropy wall. In DINOv2 feature space, we report Fréchet distance, kernel distance, and recall. As an evaluation independent of the guidance representation, we additionally report FID and KID in Inception-v3 feature space [19–21]. Synthetic and training-time experiments use the problem-specific metrics introduced in their respective subsections.

Pilot estimation. In all sampling-time experiments, we use the chain-rule plug-in guidance field. The covariance entering the IGA potential is estimated from an independent pilot batch of 2048 samples drawn from the unguided base model and then frozen during subsequent sampling. Thus, conditioned on the frozen pilot estimate, individual sampling trajectories are independent.

6.2 Numerical Application of IGA in Post-hoc Sampling-time Mode

6.2.1 Synthetic Experiments with Known Groundtruth Model

We first study three controlled settings in which the population distribution is known. These experiments allow us to evaluate whether the IGA path approaches the population below the entropy wall and departs from it beyond the wall. The three settings provide complementary evidence: a nonlinear manifold illustrates rare-region repair, an entropy-matched control distinguishes IGA from isotropic noise for increasing entropy, and a finite mixture exposes the redistribution of probability mass across modes.

Rare-region repair on a nonlinear manifold. Figure 10 considers a DDPM [4] trained on a one-dimensional population embedded in $\mathbb{R}^{128}$. The fitted model captures the dominant central portion of the manifold but substantially underrepresents its endpoints, leading to lower entropy and reduced rare-region coverage. Here, rare regions are defined as the portions outside the central 70% of the normalized manifold coordinate.

Increasing λ initially corrects this contraction. Near the entropy wall, IGA recovers both the population entropy and the missing endpoint mass. Panels 10e and 10f show the corresponding transition: the covariance discrepancy decreases as the path approaches the wall, while rare-region coverage increases. Beyond the wall, coverage continues to grow, but the discrepancy to the population turns upward. Thus, the same path first repairs variation lost by the fitted model and then promotes exploration beyond the population level.

Structured coverage versus entropy-matched noise. The spiral experiment shows that IGA directs probability toward underrepresented regions. To determine whether this behavior could be reproduced by simply adding noise, Figure 11 compares IGA with an entropy-matched Gaussian-convolution baseline applied to a multimodal DDIM model [16]. For each IGA setting, the convolution scale σ is selected by bisection so that $Q_\sigma = P_{\mathrm{DDIM}} * \mathcal{N}(0, \sigma^2 I)$ attains the same representation-space von Neumann entropy.

Despite matching entropy, the two methods distribute their additional mass differently. Gaussian convolution broadens every mode approximately isotropically, producing increasingly diffuse clouds around the original modal centers. IGA instead selectively fills the underrepresented regions between neighboring modes. As λ increases, these inter-mode regions form a coherent ring while the original modes remain visible. The entropy increase produced by IGA therefore reflects structure-aware redistribution rather than an undirected increase in noise.

Population-level confirmation in a controlled mixture. Figure 12 provides a complementary view using an eight-component mixture with known population weights. The base distribution overweights its

most frequent components and underrepresents the remaining modes, resulting in lower entropy than the population. Increasing λ initially corrects this imbalance: at the calibrated wall, the guided distribution approximately recovers both the population entropy and its component probabilities. Beyond the wall, the component probabilities become more uniform than those of the population.

Panel C of Figure 12 makes the change in regime explicit. Along the below-wall portion of the path, the population KL decreases as IGA repairs the component imbalance. After the wall is crossed, entropy continues to increase while KL turns upward. The path therefore first approaches the population through diversity repair and subsequently departs from it through deliberate extrapolation.

Together, these controlled experiments show that IGA restores underrepresented population structure below the entropy wall and enters an extrapolative regime beyond it. They further show that the increase in diversity arises from selective redistribution toward underrepresented regions rather than isotropic perturbation. We next examine whether the same progression appears in pretrained diffusion models on real-world image benchmarks.

6.2.2 Real-World Image Distribution Benchmarks

Having established the repair-to-extrapolation transition in controlled settings, we next ask whether the same progression appears in pretrained diffusion models on real-world image distributions. We evaluate sampling-time IGA on unconditional CelebA-HQ and class-conditional ImageNet generation.

Benchmark settings. On CelebA-HQ [22], we guide the pretrained `google/ddpm-ema-celebahq-256` DDPM [4] at 256×256, using deterministic DDIM sampling [16] for 100 steps. On ImageNet, we use the standard ImageNet-100 subset introduced by Tian et al. [23], consisting of their fixed 100-class subset of ILSVRC-2012 [24]. We guide the class-conditional `facebook/DiT-XL-2-256` model [25] for 50 DDIM steps, using classifier-free guidance [26] at scale 2.0.

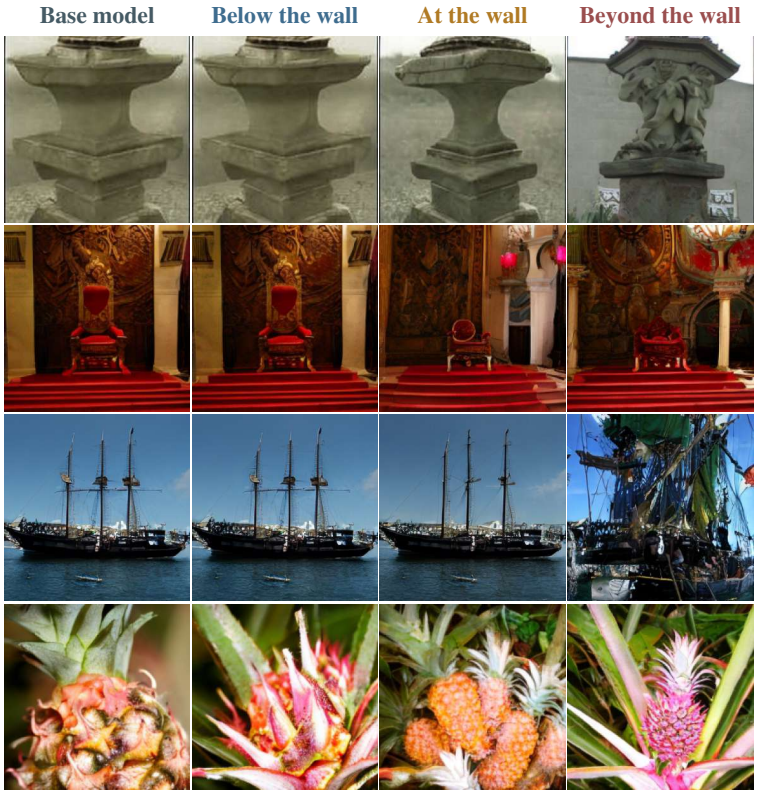
The coefficient on the IGA score correction is fixed to one, as prescribed by the sampling-time construction. We vary only the entropy multiplier λ. Each point along the reported path therefore corresponds to a different entropy-regularized target, rather than to a stronger or weaker application of the same guidance field. The DDIM sampler implements the corresponding unit-scale plug-in correction at each denoising step.

Diversity deficit and wall crossing. Because empirical spectral entropy depends on sample size, Figure 4 compares real and generated distributions using matched numbers of samples. On both datasets, the base model remains below the corresponding empirical data wall throughout the evaluated sample-size range. Increasing λ progressively closes this deficit, reaches the wall at an intermediate point, and crosses it for larger values. The tested path therefore spans three interpretable regimes: a diversity-deficient base model, below-wall repair, and beyond-wall imagination.

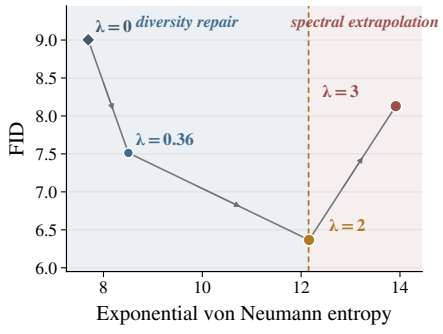
The ImageNet path across the wall. Figure 13 summarizes the progression on ImageNet. The matched samples in Figure 13a show the transition from the base model through below-wall repair and into beyond-wall extrapolation. The FID–entropy phase portrait in Figure 13b shows the corresponding distributional trend: FID initially decreases as the entropy deficit is repaired and turns after the target approaches and crosses the empirical wall. To examine how the additional entropy is obtained, we define the cumulative spectral-occupancy ratio

$$T_\lambda(r) = \log \frac{\sum_{i=r}^{d} v_i^\top S_\lambda v_i}{\sum_{i=r}^{d} v_i^\top S_{\text{data}} v_i},$$

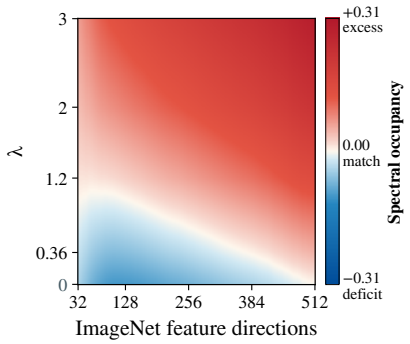
where S_λ and S_{data} denote the generated and data covariance matrices in DINOv2 feature space, respectively, and the data-covariance eigenvectors v_i are ordered from dominant to rare. Negative values indicate an occupancy deficit relative to the data, whereas positive values indicate excess occupancy. Figure 13c shows



(a) Qualitative transition across the entropy wall.



(b) FID–entropy phase portrait.



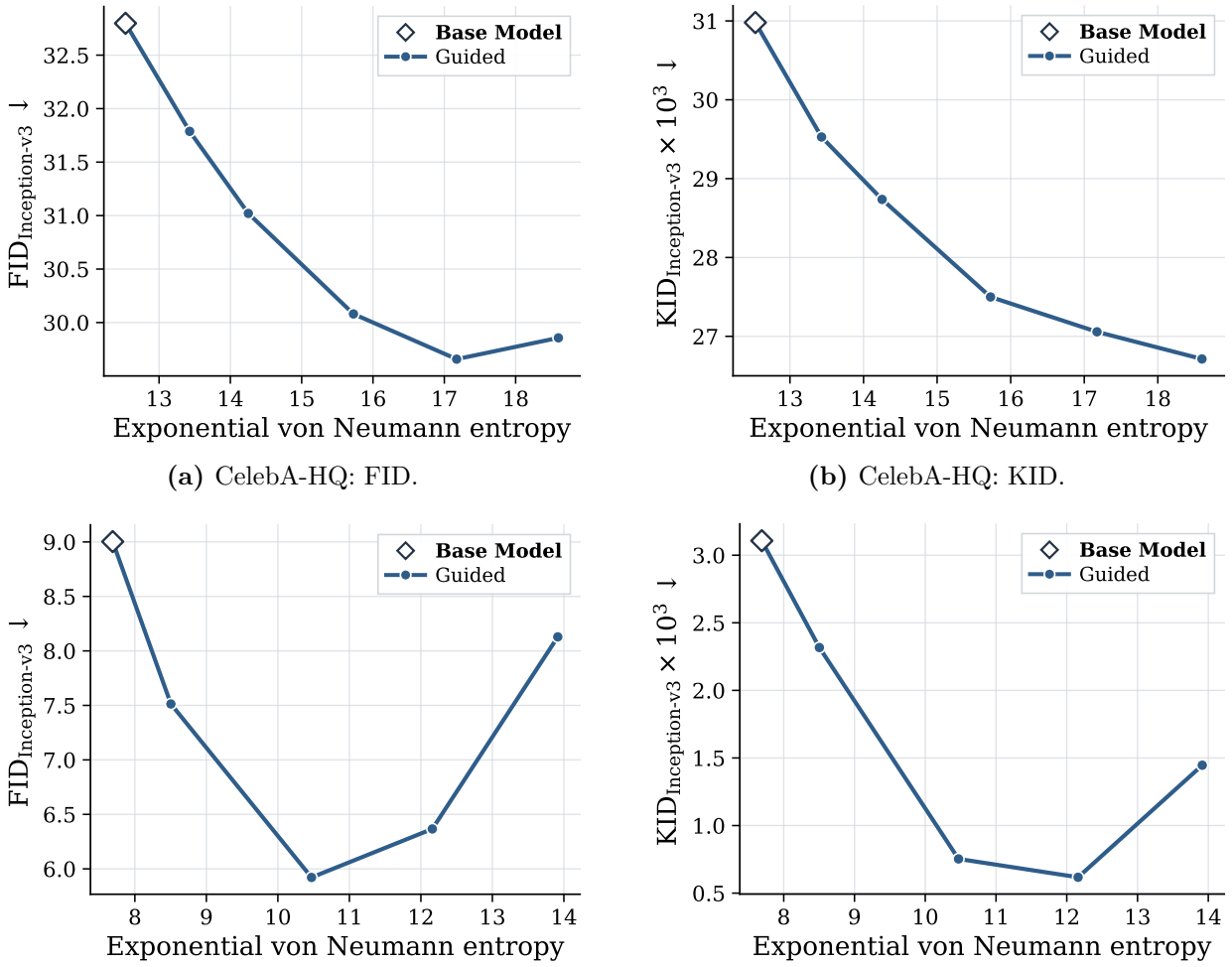
(c) Spectral occupancy.

Figure 13: **IGA across the ImageNet entropy wall. (a)** Matched samples along the IGA target path as λ increases. **(b)** FID initially decreases as entropy approaches the empirical wall and turns beyond it. **(c)** Cumulative spectral occupancy relative to the data, across feature directions ordered from dominant to rare.

that increasing λ progressively closes the deficit across underrepresented directions and produces excess occupancy after the wall is crossed. IGA therefore gains entropy by allocating more probability to directions that the base generator covers insufficiently.

Repair below the wall. Figure 14 shows a consistent initial repair regime on both benchmarks. At the beginning of the below-wall path, spectral diversity and recall increase while the reported feature-space distances decrease relative to the base model. IGA therefore recovers variation missing from the pretrained generator while improving its agreement with the data under both DINOv2 and Inception-v3 representations. This behavior is consistent with the repair result of Theorem 6. As the target approaches the empirical wall, entropy and recall continue to increase, while the different distances attain their minima at nearby but nonidentical values of λ. This is expected: Inception-v3 and DINOv2 encode different properties of image distributions and need not identify the same target as closest to the data.

Imagination beyond the wall. Once λ moves the target beyond the wall, entropy and recall continue to increase, but the feature-space distances no longer decrease uniformly. Their turning points depend on the dataset, metric, and evaluation representation. This should not be interpreted as a direct measurement of declining perceptual image quality. FID, KID, FD, and KD measure distributional departure from the data in particular feature spaces. Beyond the wall, their increase instead indicates that the generated distribution is moving farther from the data reference while occupying additional feature directions.



(a) CelebA-HQ: FID. **(b)** CelebA-HQ: KID.

(c) ImageNet: FID. **(d)** ImageNet: KID.

Figure 14: **Independent Inception-v3 evaluation along the IGA target path.** FID and KID are measured in Inception-v3 feature space, independently of the DINOv2 representation used to define spectral diversity and the empirical entropy wall. On both CelebA-HQ and ImageNet, the initial below-wall portion of the path improves distributional agreement with the data while diversity increases. At larger values of λ, the behavior becomes metric- and dataset-dependent, with distributional distances eventually flattening or turning as the target enters the extrapolative regime. Corresponding DINOv2-space FD and KD curves are reported in Appendix H.

The two benchmarks therefore exhibit the same overall progression: IGA first repairs a measurable diversity deficit and then enters a different statistical regime after crossing the wall. The central result is not a single optimal value of λ, but an interpretable target path whose meaning changes from distributional repair to deliberate spectral extrapolation.

Reference diversity-guidance methods. Table 1 includes CADS [10] and SPARKE [12] as reference points rather than like-for-like baselines. CADS perturbs the conditioning signal and is therefore reported only on class-conditional ImageNet, while the evaluated SPARKE configuration uses joint batch guidance and produces coupled samples. Neither method defines its operating point relative to the data entropy or distinguishes below-wall repair from beyond-wall extrapolation. In contrast, IGA traces a wall-calibrated family of target distributions and, once its potential is estimated and frozen, applies the same guidance

Table 1: **Comparison with diversity-guidance methods.** Distributional distances, coverage, and spectral diversity on CelebA-HQ and ImageNet. Shaded rows trace the IGA target path as λ increases. Bold indicates the lowest distributional distance or highest recall within each dataset block.

			Inception-v3		DINOv2		Coverage & diversity	
	Method	Vendi	KID $\times 10^3 \downarrow$	FID $\downarrow$	KD $\times 10^2 \downarrow$	FD $\downarrow$	Recall $\uparrow$	H_0
CelebA-HQ	Base	12.5	30.98	32.80	12.331	158.7	0.611	2.528
	SPARKE [12]	23.0	50.314	54.77	30.294	792.5	0.479	3.134
	IGA, $\lambda = 0.5$	13.4	29.53	31.79	12.018	152.5	0.627	2.597
	IGA, $\lambda = 1$	14.3	28.74	31.02	**11.876**	**151.0**	0.649	2.657
	IGA, $\lambda = 2$	15.7	27.50	30.08	12.049	156.4	0.685	2.755
	IGA, $\lambda = 3$	17.2	27.06	**29.66**	12.616	169.8	0.713	2.843
	IGA, $\lambda = 4$	18.6	**26.71**	29.86	13.272	192.1	**0.732**	2.923
ImageNet	Base	7.7	3.107	9.00	4.595	118.3	0.598	2.040
	CADS [10]	13.1	1.260	6.05	4.518	123.3	0.645	2.571
	SPARKE [12]	7.9	2.912	7.89	4.593	115.2	0.470	2.061
	IGA, $\lambda = 0.36$	8.5	2.317	7.51	4.572	**114.3**	0.636	2.141
	IGA, $\lambda = 1.2$	10.5	0.753	**5.92**	4.550	123.0	0.696	2.349
	IGA, $\lambda = 2$	12.2	**0.617**	6.37	**4.546**	143.5	0.727	2.498
	IGA, $\lambda = 3$	13.9	1.446	8.13	4.548	170.7	**0.760**	2.633

independently to each sampling trajectory. The table therefore provides numerical context under common evaluation metrics rather than a comparison of identical objectives or guarantees.

Across both benchmarks, the empirical picture is consistent. The pretrained model begins below the entropy wall; below-wall values of λ repair part of this deficit while increasing coverage and reducing distributional distances; and larger values cross the wall, where additional coverage is accompanied by a representation-dependent departure from the data distribution. IGA therefore exposes an interpretable target path from diversity repair to controlled imagination.

6.3 Numerical Application of Training-Time IGA

The main empirical focus of this paper is sampling-time IGA, which can be applied to a pretrained generator without retraining. Nevertheless, the same distribution-level regularization principle also extends naturally to model training. Section 3.3 formulates training-time IGA as (16), where the model is trained to balance fidelity to the empirical data distribution with the spectral entropy of its generated distribution. The corresponding MNIST results are reported in Appendix H.2.

We evaluate this training-time realization using a GAN on MNIST. Following Proposition 3, the entropy reward is added to the adversarial objective through the joint-adversary formulation in (18). For each generated minibatch, we compute the spectral adversary at its closed-form best response and hold it fixed during the generator update. By Proposition 7, this frozen payoff gives the exact gradient of the minibatch entropy at the refresh point.

Implementation details. We use a convolutional GAN with a 64-dimensional latent and batch size 128, trained for 20 epochs with Adam using learning rate 2×10^{-4} and $(\beta_1, \beta_2) = (0.5, 0.999)$ for both generator and discriminator. We use the non-saturating logistic generator objective and one discriminator update per generator update. The IGA representation is the unit-normalized 64-dimensional embedding of a frozen MNIST classifier, while an architecturally distinct frozen classifier with a 96-dimensional embedding is used for independent evaluation. We set $\varepsilon = 0.05$ and report means and standard errors over five random seeds.

The baseline GAN exhibits a noticeable imbalance in generated digit frequencies despite being trained on the nearly balanced MNIST distribution. Figure 20 in Appendix H.2 shows that moderate IGA regularization redistributes probability mass toward digit classes underrepresented by the baseline generator. At the best intermediate settings, the total variation distance between the generated and empirical class distributions decreases by 37.1%, while the Fréchet distance measured in the feature space of a separate evaluator network decreases by 40.1%. The improvement in both metrics indicates that the effect is not limited to the class-frequency statistic. Their nonmonotone dependence on the IGA multiplier also illustrates the tradeoff between the GAN fidelity objective and the distribution-level entropy reward.

6.4 IGA Application to Prompt-Conditioned Generative Models

We finally test sampling-time IGA on two text-to-image models: Stable Diffusion XL (SDXL) [27] and PixArt-Σ [28]. We use `stable-diffusion-xl-base-1.0` at 768×768 and `PixArt-Sigma-XL-2-1024-MS` at 1024×1024, with deterministic DDIM sampling for 50 steps. The IGA score correction is fixed at unit scale, and only λ is varied. Figure 9 shows the resulting path, using $\lambda \in \{0, 4, 8\}$ for SDXL and $\lambda \in \{0, 10, 20\}$ for PixArt-Σ. Within each model, the initial noise seeds are matched across the sweep. Increasing λ leads to larger changes in garment shape, volume, material, and color while remaining consistent with the prompt.

Figures 5–8 compare vanilla and IGA SDXL across fashion, architecture, underwater painting, and throne design. The base samples tend to stay near familiar forms, whereas IGA produces sculptural garments, curved and stacked towers, underwater scenes with divers and vehicles, and more varied throne structures. Across these examples, the main changes are in shape, structure, and scene composition rather than only color or texture. Additional qualitative results for PixArt-Σ are provided in Appendix B.

7 Conclusion and Discussion

Generative modeling is typically formulated as distributional imitation, i.e., the ultimate goal is to generate fresh samples from the underlying distribution of real training samples. However, as shown in [6], such an approach can empirically lead to a model that generates high-quality samples while remaining systematically less diverse than the target real distribution. Our work introduces *Imaginative Generative AI (IGA)*, a distribution-level framework that incorporates spectral entropy as an explicit and controllable diversity component of the generative modeling objective. The real data distribution's spectral entropy establishes an *Entropy Wall* in the application of IGA: below this wall, IGA entropy regularization repairs diversity lost during training while remaining compatible with the diversity of the data; beyond the wall, the generated distribution intentionally attains greater representation-relative spectral diversity than the real data.

We note that the IGA regularization principle can be applied to both the training of a generative model and post-hoc sampling from a pretrained model. Therefore, IGA provides a general framework for diversity regularization and imaginative generation. Beginning with improving the imitation regime, the approach first counteracts spectral-diversity deficits and encourages the recovery of variation underrepresented by the learned generator. Upon reaching the Entropy Wall, additional regularization transitions into a controlled extrapolative regime, balancing increased representation-relative diversity with closeness to the reference distribution. During sampling, once the IGA guidance potential is fixed, the resulting target enables independent and identically distributed generation, eliminating the need for an interacting batch.

Our numerical results support the application of IGA for both diversity repair and imaginative data generation. Pretrained diffusion models demonstrate a measurable entropy deficit compared with matched real-data samples; moderate IGA guidance addresses this deficit, enhancing diversity and, in several cases, distributional fidelity. Stronger guidance crosses the Entropy Wall and generates structured variation beyond the data reference level, including qualitatively novel and imaginative changes in large text-to-image models. These findings indicate that diversity enhancement does not need to be treated as an architecture-specific heuristic or as an uncontrolled deviation from quality. Instead, IGA offers a general regularization

framework for systematically transitioning from imitation, through diversity repair, to controlled imaginative extrapolation.

The notion of imagination in IGA is intentionally representation-relative: exceeding the Entropy Wall means exceeding the spectral diversity of the data in a specified embedding space, rather than satisfying a representation-independent notion of creativity or novelty. Consequently, the choice of representation, the reference distribution, and the fidelity discrepancy remain important modeling decisions. Subject to these choices, the Entropy Wall provides an explicit and measurable boundary between improving imitation and deliberately moving beyond it, making the transition from imitation to imagination mathematically well-defined and controllable.

References

[1] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 27, 2014.

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

[3] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems*, volume 32, pages 11895–11907, 2019.

[4] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851, 2020.

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

[6] 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.

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

[8] Mohammad Jalali, Cheuk Ting Li, and Farzan Farnia. An information-theoretic evaluation of generative models in learning multi-modal distributions. In *Advances in Neural Information Processing Systems*, volume 36, pages 9931–9943, 2023.

[9] Azim Ospanov, Jingwei Zhang, Mohammad Jalali, Xuenan Cao, Andrej Bogdanov, and Farzan Farnia. Towards a scalable reference-free evaluation of generative models. In *Advances in Neural Information Processing Systems*, volume 37, pages 120892–120927, 2024.

[10] Seyedmorteza Sadat, Jakob Buhmann, Derek Bradley, Otmar Hilliges, and Romann M. Weber. CADS: Unleashing the diversity of diffusion models through condition-annealed sampling. In *International Conference on Learning Representations*, 2024.

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

[12] Mohammad Jalali, Haoyu Lei, Amin Gohari, and Farzan Farnia. SPARKE: Scalable prompt-aware diversity and novelty guidance in diffusion models via RKE score. In *Advances in Neural Information Processing Systems*, volume 38, pages 119943–119980, 2025.

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

[14] Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-GAN: Training generative neural samplers using variational divergence minimization. In *Advances in Neural Information Processing Systems*, volume 29, pages 271–279, 2016.

[15] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 214–223. PMLR, 2017.

[16] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021.

[17] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy 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. *arXiv preprint arXiv:2304.07193*, 2023.

[18] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, volume 20, pages 1177–1184, 2007.

[19] 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.

[20] Mikołaj Bińkowski, Danica J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. In *International Conference on Learning Representations*, 2018.

[21] Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[22] 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.

[23] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *European Conference on Computer Vision*, 2020.

[24] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[25] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023.

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

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

[28] 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 (ECCV)*, pages 74–91, 2024.

[29] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training GANs. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

[30] Sanjeev Arora, Andrej Risteski, and Yi Zhang. Do GANs learn the distribution? some theory and empirics. In *International Conference on Learning Representations*, 2018.

[31] 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*, pages 110–119. Association for Computational Linguistics, 2016.

[32] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *International Conference on Learning Representations*, 2020.

[33] 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 *Advances in Neural Information Processing Systems*, volume 38, 2025. Datasets and Benchmarks Track.

[34] Mischa Dombrowski, Weitong Zhang, Sarah Cechnicka, Hadrien Reynaud, and Bernhard Kainz. Image generation diversity issues and how to tame them. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3029–3039, 2025.

[35] Jingwei Zhang, Cheuk Ting Li, and Farzan Farnia. An interpretable evaluation of entropy-based novelty of generative models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 59148–59172. PMLR, 2024.

[36] Mohammad Jalali, Azim Ospanov, Amin Gohari, and Farzan Farnia. Conditional Vendi score: Prompt-aware diversity evaluation for text-guided generative AI models. In *The 29th International Conference on Artificial Intelligence and Statistics*, 2026.

[37] Azim Ospanov, Mohammad Jalali, and Farzan Farnia. Scendi score: Prompt-aware diversity evaluation via schur complement of CLIP embeddings. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16927–16937, 2025.

[38] Azim Ospanov and Farzan Farnia. Do Vendi scores converge with finite samples? truncated Vendi score for finite-sample convergence guarantees. In *Proceedings of the Forty-first Conference on Uncertainty in Artificial Intelligence*, volume 286 of *Proceedings of Machine Learning Research*, pages 3272–3299. PMLR, 2025.

[39] 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*, pages 37667–37682. PMLR, 2024.

[40] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10844–10853, 2024.

[41] Reyhane Askari Hemmat, Melissa Hall, Alicia Sun, Candace Ross, Michal Drozdzal, and Adriana Romero-Soriano. Improving geo-diversity of generated images with contextualized Vendi score guidance. In *Computer Vision – ECCV 2024*, volume 15145 of *Lecture Notes in Computer Science*, pages 213–229. Springer, 2025.

[42] 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.

[43] 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*, pages 30911–30942. PMLR, 2025.

[44] 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*, pages 96–103, 2017.

[45] Songwei Ge, Vedanuj Goswami, C. Lawrence Zitnick, and Devi Parikh. Creative sketch generation. In *International Conference on Learning Representations*, 2021.

[46] 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):1–14, 2024.

[47] Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Diffusion art or digital forgery? investigating data replication in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6048–6058, 2023.

[48] Vikash Sehwag, Caner Hazirbas, Albert Gordo, Firat Ozgenel, and Cristian Canton Ferrer. Generating high fidelity data from low-density regions using diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11492–11501, 2022.

[49] Jack Lu, Ryan Teehan, and Mengye Ren. ProCreate, don't reproduce! propulsive energy diffusion for creative generation. In *Computer Vision – ECCV 2024*, volume 15118 of *Lecture Notes in Computer Science*, pages 397–414. Springer, 2025.

[50] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 1278–1286. PMLR, 2014.

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 [29, 30]. Language models likewise tend toward generic, repetitive, or homogeneous outputs, motivating diversity-aware objectives and decoding strategies [31–33]. More recently, Rényi Kernel Entropy (RKE) evaluations have shown that modern generators can produce high-quality samples while still missing modes [8]. 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 [34, 6]. These findings motivate IGA's premise: diversity should be specified as a property of the target distribution rather than left as a by-product of distribution fitting.

Measuring novelty and diversity. Reference-free measures assess variation within a distribution through similarities among its samples. The Vendi Score uses the von Neumann entropy of a normalized kernel matrix [7], while RKE provides a tractable order-two counterpart with mode-count interpretations [8]. Kernel-based Entropic Novelty compares which modes are more strongly expressed than in a reference distribution [35]. Conditional Vendi and Scendi extend diversity evaluation to prompt-conditioned generators [36, 37], while scalable and truncated variants address computational cost and finite-sample estimation [9, 38]. A complementary line folds sample quality directly into the diversity score itself, yielding quality-weighted Vendi scores [39]. IGA moves this spectral perspective from post-hoc evaluation into the generative objective itself. Population-data entropy then defines an entropy wall that separates 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 [40] and through diversity-aware diffusion modules [34]. At inference, CADS anneals noise in the conditioning signal [10]; c-VSG and SPARKE guide generation using contextualized Vendi and conditional RKE, respectively [41, 12]; and STRIDE perturbs intermediate features in distilled one- and few-step generators [42]. Particle Guidance instead evolves an interacting set under a pairwise diversity potential and is explicitly non-i.i.d. [11], while SPELL repels trajectories from protected, concurrent, or previously generated images [43].

IGA instead provides a single distribution-level regularizer with both training- and sampling-time realizations. It does not define diversity through the active batch or generation history: during sampling, its potential is estimated beforehand, held fixed, and applied independently to each initialized trajectory, yielding i.i.d. samples from the approximated IGA target. The entropy wall additionally identifies whether regularization repairs diversity lost during learning or intentionally moves beyond the population data. IGA thus unifies diversity control across end-to-end learning and post-hoc sampling while providing a principled transition from imitation to extrapolation.

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 [44]; DoodlerGAN recombines object parts into unseen sketches [45]; and ConceptLab searches for new category members [46]. For diffusion models, evidence of training-data replication sharpens the distinction between creativity and reproduction [47]. Low-density sampling explores rare regions of the learned distribution [48], whereas ProCreate pushes generations away from reference images [49]. IGA instead gives creativity a distributional interpretation: crossing the entropy wall produces a target whose representation-relative spectral diversity exceeds that of the population data, while the fidelity term controls departure from the reference. The same definition governs both training and sampling.

B Additional Qualitative Results

We provide additional qualitative comparisons on PixArt-Σ. Unless stated otherwise, IGA uses $\lambda = 20$. Across prompts, the vanilla model often concentrates on a narrow set of familiar forms or compositions, while IGA produces broader structural and semantic variation. This is particularly visible in the underwater example, where vanilla PixArt repeatedly generates very similar coral-reef scenes, whereas IGA explores substantially different subjects and layouts while retaining the requested rendering style.

Figure 15: **Fantasy throne design with PixArt-Σ.** Vanilla PixArt (left) and IGA with $\lambda = 20$ (right). IGA produces broader variation in silhouette and structure.

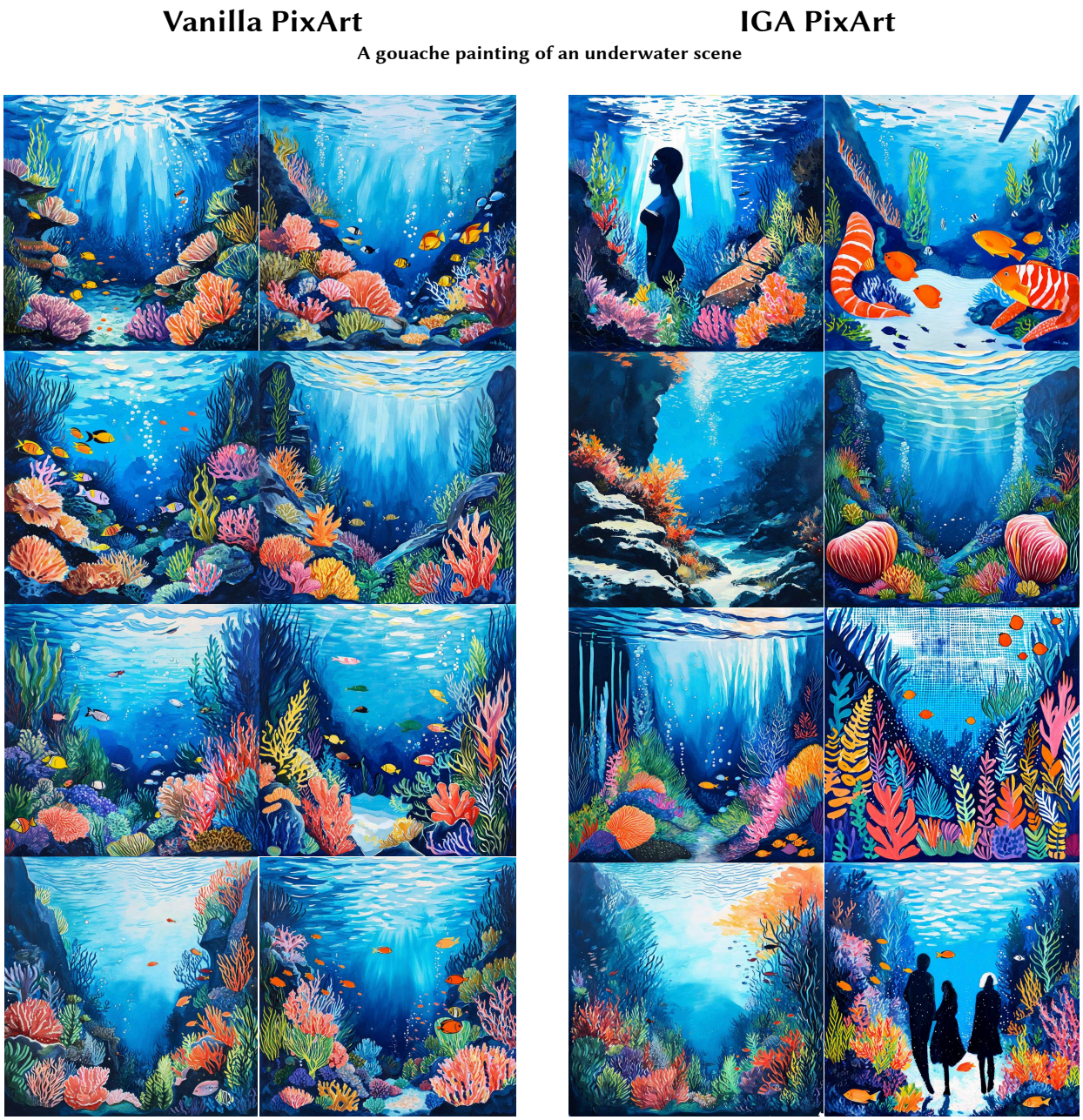


Figure 16: **Underwater scenes with PixArt-Σ.** Vanilla PixArt (left) largely collapses to the same coral-reef scene across seeds. IGA with $\lambda = 40$ (right) produces substantially broader variation in subjects and composition while preserving the gouache style.

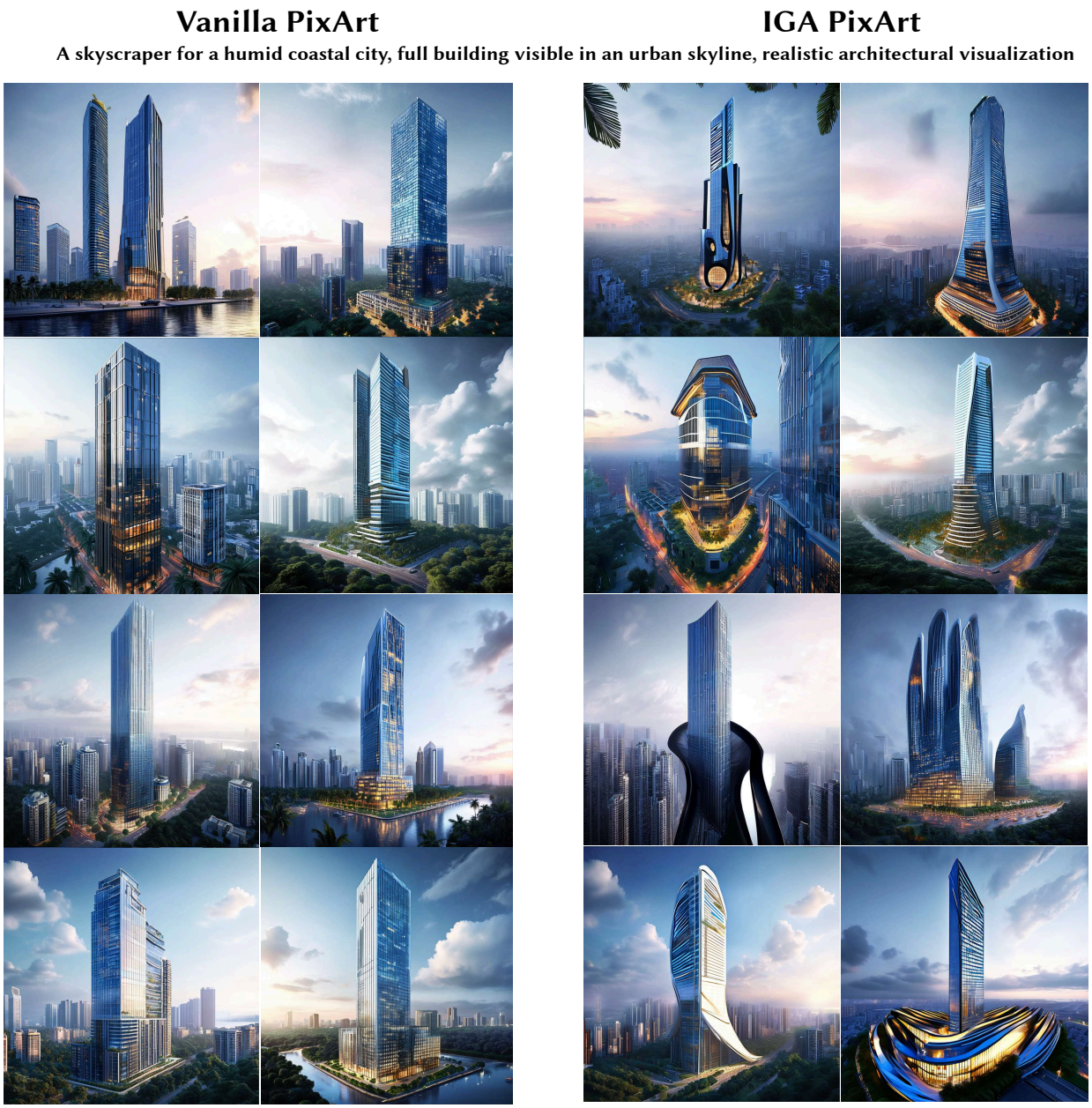


Figure 17: **Architectural design with PixArt-Σ.** Vanilla PixArt (left) and IGA with $\lambda = 20$ (right). IGA introduces stronger geometric and structural variation while retaining the skyscraper concept.

Figure 18: **Fashion design with PixArt-Σ.** Vanilla PixArt (left) and IGA with $\lambda = 20$ (right). IGA produces broader variation in garment silhouette, volume, material, and color, while remaining consistent with the wearable haute-couture prompt.

C Extended Preliminaries and Conventions

We begin by stating two conventions that are used throughout the paper. We would like to further clarify that some of the theoretical lemmas and basic statements are also discussed in [6].

Remark 2. *For any discrepancy $\mathcal{D}$ we write $\mathcal{D}(Q; P_{\mathrm{ref}})$ with the optimized distribution Q first and the reference P_{ref} second. For symmetric discrepancies this is cosmetic. However, for KL and general Bregman divergences the ordering cannot be generally swapped.*

In our notation, the KL anchor is always $\mathrm{KL}(Q\|P_{\mathrm{ref}})$, and a Bregman anchor is always $D_\Phi(Q, P_{\mathrm{ref}})$. A data-first object such as $\mathrm{KL}(\widehat{P}_n\|Q)$ or $D_\Phi(\widehat{P}_n, Q)$ is a different problem and does not inherit the guarantees below (see also Remark 10 and, for the maximum-likelihood setting where the data-first orientation is forced, Proposition 8).

Remark 3. *$\mathcal{P}$ is assumed to be the convex ambient class of probability measures on $\mathcal{X}$, used for the convex-analytic theory (concavity of entropy, convex duality, Bregman projection). $\mathcal{P}_{\mathrm{gen}}$ is the possibly nonconvex class realizable by a fixed architecture, used for training. Convexity, strong duality, and Pythagorean statements are proved on $\mathcal{P}$ and never silently transferred to $\mathcal{P}_{\mathrm{gen}}$; a trained or guided model reaches the ambient optimum only in an approximation-theoretic sense.*

Assumption 2 (Finite-dimensional embedding)**.** *Unless stated otherwise, $\phi : \mathcal{X} \to \mathbb{R}^d$ with $\|\phi(x)\|_2 = 1$ is measurable, and $\Sigma_Q \in \mathbb{R}^{d\times d}$ with $\Sigma_Q \succeq 0$ and $\mathrm{Tr}(\Sigma_Q) = 1$. When weak continuity of $Q \mapsto \Sigma_Q$ is invoked, we further assume $\mathcal{X}$ is Polish and ϕ is bounded and continuous.*

Remark 4 (Use of entropy symbol H)**.** *When a result holds for either entropy functional we let H denote a fixed but arbitrary choice from $\{H_0, H_\varepsilon\}$ (as in (C_ρ), (P_λ), Theorem 4, and Definition 1); H is never used to mix the two within a single statement. When differentiability of the entropy is needed we specialize to the smoothed H_ε (Section 5), and when defining the population wall in its main statistical interpretation we use the unsmoothed H_0 (Section 4). Lemma 3 controls the gap between the two.*

Lemma 2 (Concavity of H_0 and H_ε in Q)**.** *If $Q \mapsto \Sigma_Q$ is affine, then $Q \mapsto H_0(Q)$ and $Q \mapsto H_\varepsilon(Q)$ are concave.*

Proof. Note that $S \mapsto -\mathrm{Tr}(S \log S)$ is a concave functional of density matrices (unit-trace PSD matrices), and a concave function composed with an affine map is concave. This proves concavity of H_0.

For H_ε, the map $Q \mapsto S_Q^\varepsilon = (1-\varepsilon)\Sigma_Q + \varepsilon\frac{1}{d}I_d$ is also affine in Q, and therefore the same argument applies. □

Lemma 3 (Bounding Smoothing Gap of Spectral Entropy)**.** *Let $d \geq 2$, let Σ be a $d \times d$ density matrix, and set $S = (1-\varepsilon)\Sigma + \varepsilon\frac{1}{d}I_d$ for $\varepsilon \in [0,1]$. Writing $\tau = \varepsilon(1-\frac{1}{d})$ and $h_2(t) = -t\log t - (1-t)\log(1-t)$,*

$$\big|\mathsf{H}(S) - \mathsf{H}(\Sigma)\big| \leq \tau\log(d-1) + h_2(\tau).$$

Since $H_\varepsilon(Q) = \mathsf{H}(S_Q^\varepsilon)$ and $H_0(Q) = \mathsf{H}(\Sigma_Q)$, this implies the following for every Q:

$$\big|H_\varepsilon(Q) - H_0(Q)\big| \leq \tau\log(d-1) + h_2(\tau).$$

In particular, $|\rho_{\star,H_\varepsilon} - \rho_\star|$ admits the same bound. The bound is achieved when Σ is rank one: then $\frac{1}{2}\|S-\Sigma\|_1 = \tau$ exactly, S has eigenvalues $1-\tau$, $\frac{\tau}{d-1}, \ldots, \frac{\tau}{d-1}$, and $\big|\mathsf{H}(S) - \mathsf{H}(\Sigma)\big| = \tau\log(d-1) + h_2(\tau)$.

Proof. First, we bound the trace distance between S and Σ. Since $S - \Sigma = \varepsilon(\frac{1}{d}I_d - \Sigma)$, we have $\frac{1}{2}\|S-\Sigma\|_1 = \varepsilon\cdot\frac{1}{2}\big\|\frac{1}{d}I_d - \Sigma\big\|_1$. If Σ has eigenvalues p_i, then $\frac{1}{2}\big\|\frac{1}{d}I_d - \Sigma\big\|_1 = \frac{1}{2}\sum_i |p_i - \frac{1}{d}|$, which over the probability simplex is maximized at a vertex $p = e_j$, giving $1 - \frac{1}{d}$. Hence

$$\tfrac{1}{2}\|S-\Sigma\|_1 \leq \varepsilon(1-\frac{1}{d}) = \tau.$$

Next, we apply the Fannes–Audenaert inequality, showing that for $d \times d$ density matrices A, B with $\frac{1}{2}\|A-B\|_1 \le t \le 1-\frac{1}{d}$,

$$|\mathsf{H}(A) - \mathsf{H}(B)| \le t\log(d-1) + h_2(t).$$

The right-hand side is non-decreasing in t on $[0, 1-\frac{1}{d}]$: its derivative $\log(d-1) + \log\frac{1-t}{t}$ is non-negative there, vanishing only at $t = 1-\frac{1}{d}$. Applying the inequality at $t = \frac{1}{2}\|S-\Sigma\|_1 \le \tau \le 1-\frac{1}{d}$ (here $d \ge 2$ ensures $\log(d-1) \ge 0$) gives the claim with $A = S$ and $B = \Sigma$. Finally, the consequence for H_ε, H_0 follows by the stated identities, and the wall bound follows by taking $Q = P_{\mathrm{data}}$. For rank-one $\Sigma = vv^\top$, the eigenvalues of $\frac{1}{d}I_d - \Sigma$ are $\frac{1}{d}-1$ (once) and $\frac{1}{d}$ (with multiplicity $d-1$), so $\frac{1}{2}\|S-\Sigma\|_1 = \tau$ exactly; the spectrum of S is then $\big(1-\tau,\ \frac{\tau}{d-1}, \dots, \frac{\tau}{d-1}\big)$, whence $\mathsf{H}(\Sigma) = 0$ and $\mathsf{H}(S) = \tau\log(d-1) + h_2(\tau)$, so equality holds. □

The last ingredient is an elementary identity for exponential tilts. For a *fixed* bounded reward G, it identifies the minimizer of the KL-anchored linear objective in closed form; the sampling-time tilt of Theorem 1 is its self-consistent analogue, in which G is the entropy energy evaluated at the optimum itself.

Lemma 4 (Elementary Gibbs identity)**.** *Let P be a probability law and G measurable with $Z_G = \mathbb{E}_P[e^G] < \infty$; define $P^G(dx) = Z_G^{-1} e^{G(x)} P(dx)$. For every $Q \ll P$,*

$$\mathrm{KL}(Q\|P) - \mathbb{E}_Q[G] = \mathrm{KL}(Q\|P^G) - \log Z_G, \tag{22}$$

so P^G is the unique minimizer over $\{Q \ll P\}$ of the left-hand side whenever it is finite.

Proof. On $\{Q \ll P\}$ we have $\log(dQ/dP^G) = \log(dQ/dP) - G + \log Z_G$. Integrating against Q gives

$$\mathrm{KL}(Q\|P^G) = \mathrm{KL}(Q\|P) - \mathbb{E}_Q[G] + \log Z_G,$$

which is (22). The left-hand side equals $\mathrm{KL}(Q\|P^G) - \log Z_G$, minimized (over $Q \ll P$, equivalently $Q \ll P^G$ since the two are equivalent) uniquely at $Q = P^G$, where $\mathrm{KL} = 0$. □

D The Constrained–Penalized Correspondence, the Spectral Game, and Training

This appendix proves the results of Section 3. We first state and prove the constrained–penalized correspondence invoked in Section 3.1, together with the monotone regularization path; Appendix D.1 then proves the spectral min–max representation, and Appendix D.2 the training-time instantiations of Section 3.3.

Theorem 3 (Constrained–penalized correspondence)**.** *Suppose that:*

(i) *$\mathcal{P}$ is a nonempty compact convex subset of a locally convex Hausdorff space of finite signed measures;*

(ii) *$Q \mapsto \mathcal{D}(Q; P_{\mathrm{ref}})$ is proper, convex, and lower semicontinuous on $\mathcal{P}$;*

(iii) *$Q \mapsto \Sigma_Q$ is affine and continuous, so that $H \in \{H_0, H_\varepsilon\}$ is concave (Lemma 2) and upper semicontinuous;*

(iv) (Slater condition) *there exists $\bar{Q} \in \mathcal{P}$ with $\mathcal{D}(\bar{Q}; P_{\mathrm{ref}}) < \infty$ and $H(\bar{Q}) > \rho$.*

Then the following hold.

(a) Attainment. *The feasible set $\mathcal{P} \cap \{H \ge \rho\}$ is nonempty and compact, and both the constrained minimum in (C_ρ) and the inner minimum defining F_λ are attained.*

(b) Strong duality.

$$\min_{\substack{Q\in\mathcal{P}\\ H(Q)\ge\rho}} \mathcal{D}(Q; P_{\mathrm{ref}}) \;=\; \max_{\lambda\ge 0} \Big\{ \min_{Q\in\mathcal{P}} \big\{\mathcal{D}(Q; P_{\mathrm{ref}}) - \lambda H(Q)\big\} + \lambda\rho \Big\}. \tag{23}$$

(c) Optimal multiplier. *There exists* $\lambda^\star \geq 0$ *such that every solution* $Q^\star$ *of* (C_ρ) *minimizes* $F_{\lambda^\star}$ *and satisfies the complementary-slackness identity* $\lambda^\star\big(\rho - H(Q^\star)\big) = 0$.

Theorem 3 justifies replacing (C_ρ) by (P_λ) at the specific multiplier $\lambda^\star$ dual to ρ; it does *not* claim that every $\lambda \geq 0$ corresponds to a user-chosen target level. The proof proceeds through the value function of the constrained problem, after recording an unconditional min–max identity (Proposition 4).

Proposition 4 (Exact primal min–max identity)**.** *For any feasible set* $\mathcal{Q}$ *and arbitrary functionals* J, H*,*

$$\inf_{\substack{Q\in\mathcal{Q}\\ H(Q)\geq\rho}} J(Q) = \inf_{Q\in\mathcal{Q}} \sup_{\lambda\geq 0}\big\{J(Q) + \lambda(\rho - H(Q))\big\}.$$

Proof. For fixed Q, $\sup_{\lambda\geq 0}\{J(Q)+\lambda(\rho-H(Q))\}$ equals $J(Q)$ if $H(Q) \geq \rho$ (the coefficient of λ is nonpositive, so the supremum is at $\lambda = 0$) and $+\infty$ if $H(Q) < \rho$ (the coefficient is positive, so the expression diverges as $\lambda \to \infty$). Taking the infimum over $Q \in \mathcal{Q}$ retains only feasible Q and reproduces the constrained value. □

Proof of Theorem 3. By hypotheses (ii)–(iii), $J(Q) = \mathcal{D}(Q; P_{\mathrm{ref}})$ is proper, convex, l.s.c. on the convex set $\mathcal{P}$, and $H \in \{H_0, H_\varepsilon\}$ is concave (Lemma 2) and u.s.c., so the feasible set $\mathcal{P} \cap \{H \geq \rho\}$ is convex and closed.

First, we verify feasibility and attainment. By Slater (iv) the feasible set contains $\bar{Q}$, hence is nonempty; it is a closed subset of the compact $\mathcal{P}$ (i), hence compact. A l.s.c. function attains its minimum on a nonempty compact set, so the constrained minimum in (C_ρ) is attained; likewise, for each $\lambda \geq 0$, $Q \mapsto J(Q) - \lambda H(Q)$ is l.s.c. on the compact $\mathcal{P}$ and attains its minimum, so F_λ has a minimizer and the displayed min's in (23) are justified.

The remainder of the proof runs through the value function of the constrained problem,

$$v(r) = \inf_{Q\in\mathcal{P}}\big\{J(Q) : H(Q) \geq r\big\},$$

with $v(r) = +\infty$ if no feasible Q exists; by the previous paragraph, $v(\rho)$ is finite and attained. Note also that J, being l.s.c. on the compact $\mathcal{P}$, is bounded below on $\mathcal{P}$, so $v(r) \geq \inf_{\mathcal{P}} J > -\infty$ for every r.

Next, we establish the two structural properties of v. The value function is nondecreasing: if $r_1 \leq r_2$ then $\{H \geq r_2\} \subseteq \{H \geq r_1\}$, so the infimum over the smaller set is at least as large, i.e. $v(r_1) \leq v(r_2)$. The value function is also convex. To see this, fix $r_1, r_2 \in \mathbb{R}$, $\theta \in [0,1]$, and $\eta > 0$, and choose feasible Q_i (that is, $H(Q_i) \geq r_i$) with $J(Q_i) \leq v(r_i) + \eta$. The mixture $Q_\theta = \theta Q_1 + (1-\theta)Q_2 \in \mathcal{P}$ then satisfies $H(Q_\theta) \geq \theta r_1 + (1-\theta) r_2$ by concavity of H, and, by convexity of J,

$$J(Q_\theta) \leq \theta J(Q_1) + (1-\theta) J(Q_2) \leq \theta v(r_1) + (1-\theta) v(r_2) + \eta.$$

Hence $v(\theta r_1 + (1-\theta) r_2) \leq \theta v(r_1) + (1-\theta) v(r_2) + \eta$, and letting $\eta \downarrow 0$ gives convexity.

Then, we show that v is subdifferentiable at the target level ρ. By the Slater condition there is $\bar{Q} \in \mathcal{P}$ with $J(\bar{Q}) < \infty$ and $H(\bar{Q}) > \rho$; hence $v(r) \leq J(\bar{Q}) < \infty$ for all $r \leq H(\bar{Q})$. Combined with the lower bound above, v is finite on $(-\infty, H(\bar{Q})]$, an interval whose interior contains ρ. A finite convex function on an open interval is subdifferentiable at every interior point; pick $\lambda^\star \in \partial v(\rho)$. Since v is nondecreasing, $\lambda^\star \geq 0$.

With the multiplier $\lambda^\star$ in hand, we can prove strong duality. The subgradient inequality gives, for every $Q \in \mathcal{P}$,

$$J(Q) \geq v(H(Q)) \geq v(\rho) + \lambda^\star\big(H(Q) - \rho\big),$$

hence $J(Q) - \lambda^\star H(Q) \geq v(\rho) - \lambda^\star \rho$; taking the infimum over $Q \in \mathcal{P}$ yields $\inf_{Q\in\mathcal{P}}\{J - \lambda^\star H\} + \lambda^\star \rho \geq v(\rho)$. Conversely, weak duality holds: for any $\lambda \geq 0$ and any feasible Q (that is, $H(Q) \geq \rho$),

$$J(Q) \geq J(Q) - \lambda\big(H(Q) - \rho\big) \geq \inf_{Q'\in\mathcal{P}}\{J - \lambda H\} + \lambda\rho,$$

and taking the infimum over feasible Q gives $v(\rho) \geq \sup_{\lambda \geq 0}\{\inf_{Q'}\{J - \lambda H\} + \lambda\rho\}$. The two inequalities together yield (23), with the outer supremum attained at $\lambda^\star$.

Finally, we establish complementary slackness. Let $Q^\star$ solve (C_ρ). Feasibility gives $H(Q^\star) \geq \rho$, and by strong duality

$$J(Q^\star) = v(\rho) = \inf_Q\{J - \lambda^\star H\} + \lambda^\star\rho.$$

On the one hand, $J(Q^\star) - \lambda^\star H(Q^\star) \geq \inf_Q\{J - \lambda^\star H\} = v(\rho) - \lambda^\star\rho$. On the other hand, feasibility and $\lambda^\star \geq 0$ give $J(Q^\star) - \lambda^\star H(Q^\star) \leq J(Q^\star) - \lambda^\star\rho = v(\rho) - \lambda^\star\rho$. The two bounds match, forcing $\lambda^\star\big(H(Q^\star) - \rho\big) = 0$ and $J(Q^\star) - \lambda^\star H(Q^\star) = \inf_Q\{J - \lambda^\star H\}$, i.e. $Q^\star$ minimizes $F_{\lambda^\star}$. □

Remark 5 (Which hypotheses do what)**.** *The hypotheses of Theorem 3 play three separable roles.* Duality. *The value-function argument shows that strong duality and the existence of an optimal multiplier $\lambda^\star \in \partial v(\rho)$ require only convexity of $\mathcal{P}$, convex l.s.c. J, concave H, the Slater condition (iv), and the value function v being proper and finite near ρ, i.e. $v(\rho) > -\infty$ (equivalently, J bounded below on the feasible set), in addition to $v(\rho) < \infty$ from Slater. Convexity and Slater alone do not guarantee $v(\rho) > -\infty$: if J is unbounded below on $\mathcal{P}$ then $v \equiv -\infty$, no finite subgradient exists, and the duality statement is vacuous. In the compact setting of Theorem 3 this cannot happen, because an l.s.c. J on the compact $\mathcal{P}$ is bounded below; the properness caveat matters only in the noncompact variant.* Attainment. *The compactness in (i), with l.s.c. J and u.s.c. H, is otherwise used only to guarantee attainment of the constrained minimum, of the inner minima defining F_λ, and hence of the displayed* min / max *in* (23). Noncompact classes. *In settings where $\{Q \ll P_\theta\}$ is convex but not compact (Section 5), existence and attainment are instead obtained by the direct method of Theorem 1, where $J = \mathrm{KL}(\cdot\|P_\theta) \geq 0$ is automatically bounded below.*

Proposition 5 (Saddle representation)**.** *Under the hypotheses of Theorem 3, with $Q^\star$ a solution of* (C_ρ) *and $\lambda^\star$ the optimal multiplier, $(Q^\star, \lambda^\star)$ is a saddle point of $\mathcal{L}(Q, \lambda) = J(Q) + \lambda(\rho - H(Q))$ on $\mathcal{P} \times [0, \infty)$:*

$$\mathcal{L}(Q^\star, \lambda) \leq \mathcal{L}(Q^\star, \lambda^\star) \leq \mathcal{L}(Q, \lambda^\star) \qquad \forall Q \in \mathcal{P},\ \lambda \geq 0,$$

and consequently $\inf_{Q\in\mathcal{P}} \sup_{\lambda\geq 0} \mathcal{L}(Q, \lambda) = \sup_{\lambda\geq 0} \inf_{Q\in\mathcal{P}} \mathcal{L}(Q, \lambda)$.

Proof. Since $Q^\star$ minimizes $F_{\lambda^\star} = J - \lambda^\star H$ over $\mathcal{P}$ (Theorem 3) and $\mathcal{L}(\cdot, \lambda^\star) = F_{\lambda^\star}(\cdot) + \lambda^\star\rho$, the right inequality $\mathcal{L}(Q^\star, \lambda^\star) \leq \mathcal{L}(Q, \lambda^\star)$ holds for all $Q \in \mathcal{P}$. For the left inequality, $\mathcal{L}(Q^\star, \lambda) = J(Q^\star) + \lambda(\rho - H(Q^\star))$ is nonincreasing in $\lambda \geq 0$ because $\rho - H(Q^\star) \leq 0$; together with $\lambda^\star(\rho - H(Q^\star)) = 0$ this gives $\mathcal{L}(Q^\star, \lambda) \leq \mathcal{L}(Q^\star, \lambda^\star)$ for all $\lambda \geq 0$. The equality of the two mixed extrema is the standard consequence of a saddle point. □

Even without convexity, the multiplier acts as a monotone control on global minimizers; the following statement, summarized in Section 3.1, applies both to the ambient problem and to a nonconvex generator family.

Theorem 4 (Monotone regularization path)**.** *Let $\mathcal{Q} \in \{\mathcal{P}, \mathcal{P}_{\mathrm{gen}}\}$, write $J(Q) = \mathcal{D}(Q; P_{\mathrm{ref}})$, and suppose that a minimizer $Q_\lambda \in \arg\min_{Q\in\mathcal{Q}} F_\lambda(Q)$ exists for every $\lambda \geq 0$. Then, for any $0 \leq \lambda_1 < \lambda_2$ and any choices of minimizers Q_{λ_1} and Q_{λ_2},*

$$H(Q_{\lambda_2}) \geq H(Q_{\lambda_1}), \qquad J(Q_{\lambda_2}) \geq J(Q_{\lambda_1}).$$

Proof of Theorem 4. Let $0 \leq \lambda_1 < \lambda_2$ and let $Q_{\lambda_1}, Q_{\lambda_2}$ be any minimizers of $F_{\lambda_1}, F_{\lambda_2}$ over $\mathcal{Q}$. Abbreviate $H_i = H(Q_{\lambda_i})$, $J_i = J(Q_{\lambda_i})$. Optimality of Q_{λ_1} at λ_1 and of Q_{λ_2} at λ_2 gives

$$J_1 - \lambda_1 H_1 \leq J_2 - \lambda_1 H_2, \qquad J_2 - \lambda_2 H_2 \leq J_1 - \lambda_2 H_1.$$

Adding these two inequalities cancels J_1, J_2 and yields $(\lambda_2 - \lambda_1)(H_2 - H_1) \geq 0$, so $H_2 \geq H_1$. Substituting into the first inequality, rearranged as $J_1 - J_2 \leq \lambda_1(H_1 - H_2) \leq 0$, gives $J_2 \geq J_1$. No convexity, differentiability, uniqueness, or path continuity is used; the argument is valid for any selection of minimizers, so the ordering holds even when minimizers are nonunique. □

Proposition 6 (Every penalized optimizer is a constrained optimizer at its attained level)**.** *Under the hypotheses of Theorem 4, set* $\rho_\lambda := H(Q_\lambda)$*. Then* $Q_\lambda \in \operatorname*{argmin}_{Q\in\mathcal{Q}} \{J(Q) : H(Q) \geq \rho_\lambda\}$.

Proof. Take any $Q \in \mathcal{Q}$ with $H(Q) \geq \rho_\lambda$. Penalized optimality gives $J(Q_\lambda) - \lambda H(Q_\lambda) \leq J(Q) - \lambda H(Q)$, so

$$J(Q_\lambda) \leq J(Q) + \lambda(H(Q_\lambda) - H(Q)) = J(Q) - \lambda(H(Q) - \rho_\lambda) \leq J(Q),$$

using $\lambda \geq 0$ and $H(Q) \geq \rho_\lambda$. □

Remark 6 (Constrained and penalized problems are not interchangeable in general)**.** *Outside the convex setting the two problems need not share solutions for a prescribed* λ *or* ρ*. Proposition 6 matches each* Q_λ *to its* own *attained level* ρ_λ*, while Theorem 3 recovers a* prescribed *level* ρ *only under its convexity and Slater hypotheses. On a nonconvex* $\mathcal{P}_{\text{gen}}$ *the saddle representation (Proposition 5) can fail with a positive duality gap; we therefore do not claim per-*λ *constrained–penalized equivalence for* $\mathcal{P}_{\text{gen}}$*.*

D.1 The spectral min–max representation: proofs

This appendix proves Proposition 1 and the best-response identity quoted in Sections 3.1 and 3.3. The main tool is the Gibbs variational principle for matrix entropy, which we derive from Klein's inequality; both results are proved in full.

Lemma 5 (Klein's inequality for matrix relative entropy)**.** *Let* S, T *be* $d \times d$ *density matrices with* $T \succ 0$*. Then*

$$\operatorname{Tr}\big(S\log S - S\log T\big) \geq \operatorname{Tr}(S) - \operatorname{Tr}(T) = 0,$$

with equality if and only if $S = T$*.*

Proof. Write spectral decompositions $S = \sum_i \alpha_i u_i u_i^\top$ and $T = \sum_j \beta_j v_j v_j^\top$ with orthonormal bases $(u_i), (v_j)$, eigenvalues $\alpha_i \geq 0$, $\beta_j > 0$, and set $c_{ij} := (u_i^\top v_j)^2$. The matrix (c_{ij}) is doubly stochastic: $\sum_j c_{ij} = \|u_i\|^2 = 1$ and $\sum_i c_{ij} = \|v_j\|^2 = 1$, since each basis is orthonormal. Expanding the traces in these bases,

$$\operatorname{Tr}(S\log S) = \sum_i \alpha_i \log \alpha_i = \sum_{i,j} c_{ij}\, \alpha_i \log \alpha_i, \qquad \operatorname{Tr}(S\log T) = \sum_{i,j} c_{ij}\, \alpha_i \log \beta_j,$$

using $\sum_j c_{ij} = 1$ for the first identity and $u_i^\top (\log T) u_i = \sum_j c_{ij} \log \beta_j$ for the second. The scalar inequality $x \log x - x \log y \geq x - y$, valid for $x \geq 0$, $y > 0$ (with $0 \log 0 = 0$; it is the tangent-line inequality for the convex function $x \mapsto x \log x$ at y), holds with equality if and only if $x = y$: for $x > 0$ this is strict convexity, and at $x = 0$ the inequality reads $0 \geq -y$, strict since $y > 0$. Applying it termwise,

$$\operatorname{Tr}\big(S\log S - S\log T\big) = \sum_{i,j} c_{ij}\big(\alpha_i \log \alpha_i - \alpha_i \log \beta_j\big) \geq \sum_{i,j} c_{ij}(\alpha_i - \beta_j) = \sum_i \alpha_i - \sum_j \beta_j = 0,$$

where the last step again uses double stochasticity. If equality holds, then every pair (i, j) with $c_{ij} > 0$ satisfies $\alpha_i = \beta_j$. Fix j and expand $v_j = \sum_i (u_i^\top v_j) u_i$; then

$$S v_j = \sum_i \alpha_i (u_i^\top v_j)\, u_i = \sum_i \beta_j (u_i^\top v_j)\, u_i = \beta_j v_j,$$

since every index i contributing a nonzero coefficient has $c_{ij} > 0$, hence $\alpha_i = \beta_j$. Thus S acts as β_j on each v_j, so $S = \sum_j \beta_j v_j v_j^\top = T$. Conversely $S = T$ gives equality trivially. □

Lemma 6 (Gibbs variational principle for matrix entropy)**.** *For every $d \times d$ density matrix S,*

$$\sup_{\Theta=\Theta^\top} \big\{ -\operatorname{Tr}(S\Theta) - \log \operatorname{Tr}(e^{-\Theta}) \big\} = -\mathsf{H}(S),$$

where the supremum runs over all symmetric $d \times d$ matrices. The objective is invariant under $\Theta \mapsto \Theta + cI_d$ for $c \in \mathbb{R}$. If $S \succ 0$, the supremum is attained exactly at the family $\Theta = -\log S + cI_d$, $c \in \mathbb{R}$; if S is singular, the supremum is not attained, but is approached along $\Theta_\delta = -\log(S + \delta I_d)$ as $\delta \downarrow 0$, whose objective value is $\sum_i \alpha_i \log(\alpha_i + \delta) - \log(1 + \delta d)$ in terms of the eigenvalues α_i of S.

Proof. First, the invariance: replacing Θ by $\Theta + cI_d$ changes $-\operatorname{Tr}(S\Theta)$ by $-c \operatorname{Tr}(S) = -c$ and changes $-\log \operatorname{Tr}(e^{-\Theta - cI}) = -\log(e^{-c} \operatorname{Tr}(e^{-\Theta}))$ by $+c$, so the objective is unchanged.

Next, the upper bound. For symmetric Θ, let $R_\Theta := e^{-\Theta} / \operatorname{Tr}(e^{-\Theta})$, a positive-definite density matrix with $\log R_\Theta = -\Theta - \log \operatorname{Tr}(e^{-\Theta})\, I_d$. Klein's inequality (Lemma 5) with $T = R_\Theta$ gives $\operatorname{Tr}(S \log S) - \operatorname{Tr}(S \log R_\Theta) \geq 0$, which expands to

$$0 \leq \operatorname{Tr}(S \log S) + \operatorname{Tr}(S\Theta) + \log \operatorname{Tr}(e^{-\Theta}),$$

i.e. $-\operatorname{Tr}(S\Theta) - \log \operatorname{Tr}(e^{-\Theta}) \leq -\mathsf{H}(S)$, with equality if and only if $R_\Theta = S$.

Then, attainment. If $S \succ 0$, the equation $R_\Theta = S$ has the solutions $\Theta = -\log S + cI_d$, $c \in \mathbb{R}$, and no others: $R_\Theta = S$ forces $-\Theta = \log S + \log \operatorname{Tr}(e^{-\Theta}) I_d$. If S is singular, no symmetric Θ satisfies $R_\Theta = S$, since $R_\Theta \succ 0$ always; hence the supremum is not attained. Finally, evaluating the objective at $\Theta_\delta = -\log(S + \delta I_d)$ gives

$$\begin{aligned} -\operatorname{Tr}(S\Theta_\delta) - \log \operatorname{Tr}(e^{-\Theta_\delta}) &= \operatorname{Tr}\big(S \log(S + \delta I_d)\big) - \log \operatorname{Tr}(S + \delta I_d) \\ &= \sum_i \alpha_i \log(\alpha_i + \delta) - \log(1 + \delta d), \end{aligned}$$

which converges to $\sum_i \alpha_i \log \alpha_i = -\mathsf{H}(S)$ as $\delta \downarrow 0$ (the terms with $\alpha_i = 0$ contribute $0 \cdot \log \delta = 0$). Hence the supremum equals $-\mathsf{H}(S)$ in all cases. □

Remark 7 (Smoothed versus unsmoothed dual, and concavity as a byproduct)**.** *Lemma 6 explains why the min–max form is stated for the* smoothed *entropy. For the unsmoothed H_0, the covariance Σ_Q can be singular, in which case the supremum is not attained, and the near-maximizers $\Theta_\delta = -\log(\Sigma_Q + \delta I_d)$ have operator norm growing like* $\log(1/\delta)$*: no compact adversary class captures the supremum uniformly over all Q. The spectral floor $S_Q^\varepsilon \succeq (\varepsilon/d) I_d$ removes both obstructions: it confines the best response to the compact class $\mathbb{T}_\varepsilon$ of Proposition 1 and guarantees attainment. The lemma also yields an independent proof of fact (a) of the organization paragraph: it displays $-\mathsf{H}(S)$ as a supremum of affine functions of S, hence convex, so H is concave, which also yields Lemma 2.*

Remark 8 (Further readings of the min–max form)**.** *Two structural readings of* (10) *complement the closed-form best response discussed in Section 3.1.* Linearization: *for fixed Θ, the inner objective depends on Q only through the expectation of the per-sample payoff $\phi(x)^\top \Theta\, \phi(x)$, so the distribution-level reward $-\lambda H_\varepsilon(Q)$ becomes an ordinary expected loss at the cost of one $d \times d$ symmetric adversarial variable; Section 3.3 exploits this directly, where the spectral player joins the discriminator as a second adversary that admits a closed-form best response.* Why smoothing: *for the unsmoothed entropy the supremum runs over an unbounded matrix class and is not attained at rank-deficient covariances (Remark 7), whereas the spectral floor $\varepsilon \frac{1}{d} I_d$ confines the adversary to the compact class $\mathbb{T}_\varepsilon$ and guarantees attainment; the interchange in part (iii) then follows from Sion's minimax theorem, whose compactness requirement is satisfied by the Θ-side alone.*

Proof of Proposition 1. First, part (i). Since $S_Q^\varepsilon \succeq (\varepsilon/d) I_d \succ 0$, Lemma 6 gives

$$-H_\varepsilon(Q) = -\mathsf{H}(S_Q^\varepsilon) = \sup_{\Theta=\Theta^\top} \big\{ -\operatorname{Tr}(S_Q^\varepsilon \Theta) - \log \operatorname{Tr}(e^{-\Theta}) \big\},$$

attained exactly at the family $-\log S_Q^\varepsilon + cI_d$. Imposing $\mathrm{Tr}(\Theta) = 0$ pins the constant at $c = \frac{1}{d}\mathrm{Tr}(\log S_Q^\varepsilon)$, which is the matrix $\Theta^\star(Q)$ of the statement; it is the unique traceless maximizer, since attainment forces membership in the family. For the operator-norm bound, the eigenvalues of S_Q^ε lie in $[\varepsilon/d,\, 1-\varepsilon(1-\frac{1}{d})] \subseteq [\varepsilon/d, 1]$, so the eigenvalues $\ell_1, \dots, \ell_d$ of $-\log S_Q^\varepsilon$ lie in $[0, \log(d/\varepsilon)]$; centering replaces ℓ_i by $\ell_i - \bar{\ell}$ with $\bar{\ell} = \frac{1}{d}\sum_j \ell_j \in [0, \log(d/\varepsilon)]$, so each centered eigenvalue satisfies $|\ell_i - \bar{\ell}| \le \max_j \ell_j - \min_j \ell_j \le \log(d/\varepsilon)$. Hence $\Theta^\star(Q) \in \mathbb{T}_\varepsilon$, and restricting the supremum to $\mathbb{T}_\varepsilon$ preserves both the value and the attainment, upgrading sup to max. It remains to pass to the per-sample form: for traceless Θ,

$$\mathrm{Tr}(S_Q^\varepsilon \Theta) = (1-\varepsilon)\,\mathrm{Tr}(\Sigma_Q \Theta) + \frac{\varepsilon}{d}\,\mathrm{Tr}(\Theta) = (1-\varepsilon)\,\mathbb{E}_{X\sim Q}\big[\phi(X)^\top \Theta\, \phi(X)\big],$$

using $\mathrm{Tr}(\Sigma_Q\Theta) = \mathbb{E}_Q[\mathrm{Tr}(\phi\phi^\top\Theta)] = \mathbb{E}_Q[\phi^\top\Theta\phi]$. This is (9).

Next, part (ii). Multiplying (9) by $\lambda \ge 0$ preserves the maximum (for $\lambda = 0$ both sides of the resulting identity vanish identically on $\mathbb{T}_\varepsilon$, since $-\lambda H_\varepsilon(Q) = 0$ and the Θ-dependent terms carry the factor λ), and adding $\mathcal{D}(Q; P_{\mathrm{ref}})$, which does not depend on Θ, gives the pointwise identity (10). Since the two sides agree as functions of Q, the problem (P_λ) of minimizing the left side over $\mathcal{P}$ is the two-player game of minimizing the right side, as claimed.

Finally, part (iii). $\mathbb{T}_\varepsilon$ is convex and compact. For fixed Θ, $Q \mapsto \mathcal{A}_\lambda(Q, \Theta)$ is convex and l.s.c.: the expectation term is affine in Q (and weakly continuous when the continuity clause of Assumption 2 is in force, ϕ being bounded and continuous), and $\mathcal{D}(\cdot; P_{\mathrm{ref}})$ is convex l.s.c. by hypothesis. For fixed Q, $\Theta \mapsto \mathcal{A}_\lambda(Q, \Theta)$ is concave and continuous: the expectation term is linear in Θ, and $-\lambda \log \mathrm{Tr}(e^{-\Theta})$ is concave, since Lemma 6 exhibits $\Theta \mapsto -\log \mathrm{Tr}(e^{-\Theta})$ as an infimum over S of affine functions of Θ (namely $-\log \mathrm{Tr}(e^{-\Theta}) = \inf_S\{\mathrm{Tr}(S\Theta) - \mathsf{H}(S)\}$, the dual reading of the same variational identity). Sion's minimax theorem requires compactness of only one side, here the Θ-side $\mathbb{T}_\varepsilon$, so no compactness of $\mathcal{P}$ is needed, and the interchange holds as stated. □

Lemma 7 (Best response and the entropy energy)**.** *For every distribution Q and every $x \in \mathcal{X}$,*

$$\lambda(1-\varepsilon)\,\phi(x)^\top \Theta^\star(Q)\,\phi(x) = \lambda\, G_Q^\varepsilon(x) - c_Q,$$

where $c_Q := \frac{\lambda(1-\varepsilon)}{d}\,\mathrm{Tr}\big(-\log S_Q^\varepsilon\big) \in \big[0,\ \lambda(1-\varepsilon)\log(d/\varepsilon)\big]$ is a constant independent of x. In particular, the per-sample payoff of the best-responding spectral adversary equals the entropy energy (8)*, up to an additive constant independent of x.*

Proof. By definition $\Theta^\star(Q) = -\log S_Q^\varepsilon + \frac{1}{d}\mathrm{Tr}(\log S_Q^\varepsilon) I_d$, so

$$\begin{aligned}\lambda(1-\varepsilon)\,\phi(x)^\top \Theta^\star(Q)\,\phi(x) &= \lambda(1-\varepsilon)\,\phi(x)^\top\big(-\log S_Q^\varepsilon\big)\phi(x) \\ &\quad + \frac{\lambda(1-\varepsilon)}{d}\,\mathrm{Tr}\big(\log S_Q^\varepsilon\big)\,\|\phi(x)\|_2^2 .\end{aligned}$$

The first term is $\lambda\, G_Q^\varepsilon(x)$ by (8), and since $\|\phi(x)\|_2 = 1$ the second term is the constant $-c_Q$. The range of c_Q follows because the eigenvalues of $-\log S_Q^\varepsilon$ lie in $[0, \log(d/\varepsilon)]$, so their average lies in the same interval. □

D.2 Training-time instantiations: adversarial and maximum-likelihood models

This appendix proves the results invoked in Section 3.3: the joint-adversary identity for adversarially trained generators (Proposition 3), the exactness of gradients computed through the frozen spectral adversary (Proposition 7), and the likelihood–KL–ELBO relations underlying the maximum-likelihood instantiation (Proposition 8).

Proof of Proposition 3. Fix $Q \in \mathcal{Q}$. By the critic representation (17) and the entropy dual (9),

$$\mathcal{D}(Q; \widehat{P}_n) - \lambda H_\varepsilon(Q) = \sup_{D \in \mathcal{D}_c} A_Q(D) + \max_{\Theta \in \mathbb{T}_\varepsilon} B_Q(\Theta),$$

where $A_Q(D) = \mathbb{E}_{\widehat{P}_n}[u(D)] - \mathbb{E}_Q[v(D)]$ and $B_Q(\Theta) = -\lambda(1-\varepsilon)\mathbb{E}_Q[\phi^\top \Theta \phi] - \lambda \log \mathrm{Tr}(e^{-\Theta})$; here the passage from (9) to $-\lambda H_\varepsilon(Q) = \max_\Theta B_Q(\Theta)$ is part (ii) of the proof of Proposition 1 (multiplication by $\lambda \geq 0$, with the degenerate case $\lambda = 0$ giving $B_Q \equiv 0 = -\lambda H_\varepsilon(Q)$). Since D and Θ range over independent sets and the two objectives share no variable, the suprema add:

$$\sup_{D \in \mathcal{D}_c} A_Q(D) + \max_{\Theta \in \mathbb{T}_\varepsilon} B_Q(\Theta) = \sup_{D \in \mathcal{D}_c} \max_{\Theta \in \mathbb{T}_\varepsilon} \big\{ A_Q(D) + B_Q(\Theta) \big\},$$

and the right-hand side is the inner expression of (18). The identity therefore holds *pointwise* in Q, and taking the infimum over an arbitrary class $\mathcal{Q}$, convex or not, preserves it. Attainment of the Θ-maximum at $\Theta^\star(Q)$ is part (i) of Proposition 1. No convexity of $\mathcal{Q}$ was used and no minimax interchange was performed. □

Proposition 7 (Frozen spectral adversary yields exact entropy gradients)**.** *Let $Z \sim P_Z$ on a latent space $\mathcal{Z}$, let $g_\vartheta : \mathcal{Z} \to \mathcal{X}$ be measurable for each $\vartheta \in \mathbb{R}^p$ and differentiable in ϑ at P_Z-a.e. z, let ϕ be differentiable on an open set containing the relevant ranges with $\|\phi\|_2 \equiv 1$, and let Q_ϑ denote the law of $g_\vartheta(Z)$. Suppose there are a neighborhood U of ϑ_0 and $L \in L^1(P_Z)$ such that*

$$\big\| \nabla_\vartheta \big[\phi(g_\vartheta(z)) \phi(g_\vartheta(z))^\top \big] \big\| \leq L(z) \qquad \text{for all } \vartheta \in U \text{ and } P_Z\text{-a.e. } z.$$

Then $\vartheta \mapsto H_\varepsilon(Q_\vartheta)$ is differentiable at ϑ_0 and

$$\nabla_\vartheta \, H_\varepsilon(Q_\vartheta)\Big|_{\vartheta_0} = (1-\varepsilon)\, \mathbb{E}_Z \Big[\nabla_\vartheta \, \phi\big(g_\vartheta(Z)\big)^\top \, \Theta^\star(Q_{\vartheta_0}) \, \phi\big(g_\vartheta(Z)\big) \Big|_{\vartheta_0} \Big],$$

i.e. the exact gradient of the entropy coincides with the gradient of the expected per-sample payoff in which the spectral adversary is frozen *at its best response $\Theta^\star(Q_{\vartheta_0})$.*

Proof. Write $S(\vartheta) := S^\varepsilon_{Q_\vartheta} = (1-\varepsilon)\, \mathbb{E}_Z\big[\phi(g_\vartheta(Z))\phi(g_\vartheta(Z))^\top\big] + \varepsilon \frac{1}{d} I_d$. First, the domination hypothesis justifies differentiation under the expectation: $\vartheta \mapsto S(\vartheta)$ is differentiable at ϑ_0 with $\partial_{\vartheta_k} S(\vartheta_0) = (1-\varepsilon)\, \mathbb{E}_Z\big[\partial_{\vartheta_k}\big(\phi(g_\vartheta(Z))\phi(g_\vartheta(Z))^\top\big)\big|_{\vartheta_0}\big]$. Next, as in the proof of Lemma 1, the matrix entropy H is Fréchet differentiable at every positive-definite matrix with $D\mathsf{H}(S)[B] = -\,\mathrm{Tr}\big(B(\log S + I_d)\big)$ for symmetric B, and this applies at $S(\vartheta_0) \succeq (\varepsilon/d) I_d \succ 0$. By the chain rule,

$$\partial_{\vartheta_k} H_\varepsilon(Q_\vartheta)\Big|_{\vartheta_0} = -\,\mathrm{Tr}\Big(\partial_{\vartheta_k} S(\vartheta_0)\, \big(\log S(\vartheta_0) + I_d\big)\Big).$$

Then, the trace term drops out: since $\|\phi\|_2^2 \equiv 1$, we have $\mathrm{Tr}\big(\partial_{\vartheta_k} S(\vartheta_0)\big) = (1-\varepsilon)\, \partial_{\vartheta_k} \mathbb{E}_Z\big[\|\phi(g_\vartheta(Z))\|_2^2\big] = \partial_{\vartheta_k}(1-\varepsilon) = 0$, so the I_d contribution vanishes and

$$\partial_{\vartheta_k} H_\varepsilon(Q_\vartheta)\Big|_{\vartheta_0} = \mathrm{Tr}\Big(\partial_{\vartheta_k} S(\vartheta_0)\, \big(-\log S(\vartheta_0)\big)\Big) = (1-\varepsilon)\, \mathbb{E}_Z\Big[\partial_{\vartheta_k}\, \phi^\top \big(-\log S(\vartheta_0)\big)\phi\Big|_{\vartheta_0}\Big],$$

where the matrix $-\log S(\vartheta_0)$ is held fixed under the derivative. Finally, replacing $-\log S(\vartheta_0)$ by $\Theta^\star(Q_{\vartheta_0}) = -\log S(\vartheta_0) + \frac{1}{d}\,\mathrm{Tr}(\log S(\vartheta_0)) I_d$ changes the per-sample payoff by a multiple of $\|\phi\|_2^2 \equiv 1$, whose ϑ-gradient is zero; hence the displayed identity. In the language of the min–max game (10), this is an envelope (Danskin-type) statement: the inner maximum is attained at the unique $\Theta^\star(Q_{\vartheta_0})$, and differentiating the value equals differentiating at the frozen maximizer. The direct computation above proves the identity without invoking any general envelope theorem. □

Maximum-likelihood training and VAEs. Deep maximum-likelihood models fit an explicit density q_ϑ by minimizing the empirical negative log-likelihood, which is the per-sample empirical proxy for the *data-first* divergence: up to an additive constant independent of ϑ, $\mathbb{E}_{P_{\mathrm{data}}}[-\log q_\vartheta(X)] = \mathrm{KL}(P_{\mathrm{data}} \| Q_\vartheta) + \mathrm{const}$ (Proposition 8 below). This orientation is forced at training time: the model-first quantity $\mathrm{KL}(Q_\vartheta \| \widehat{P}_n)$ is

typically infinite for a continuously supported model against an atomic empirical reference, whereas the likelihood is finite and estimable sample by sample. The IGA-regularized maximum-likelihood objective is

$$\min_{\vartheta} \ \mathbb{E}_{X\sim\widehat{P}_n}\big[-\log q_\vartheta(X)\big] - \lambda\, H_\varepsilon(Q_\vartheta), \tag{24}$$

and when the likelihood is intractable, as in variational autoencoders, the negative evidence lower bound takes its place [2, 50]:

$$\min_{\vartheta,\eta} \ \mathbb{E}_{X\sim\widehat{P}_n}\big[-\mathrm{ELBO}(X;\vartheta,\eta)\big] - \lambda\, H_\varepsilon(Q_\vartheta). \tag{25}$$

Proposition 8 records the exact relation between the two: the negative ELBO exceeds the negative log-likelihood by the encoder-posterior gap $\mathrm{KL}(r_\eta(\cdot \mid x) \,\|\, p_\vartheta(\cdot \mid x)) \geq 0$, a λ-independent quantity, so (25) is (24) plus a nonnegative gap that only the encoder parameters η tighten; the same reading applies to diffusion models trained through variational bounds [4].

We highlight two features of this combination. First, the entropy regularizer is *likelihood-free*: evaluating $H_\varepsilon(Q_\vartheta)$ requires only samples from the decoder, never density values, so it applies to any latent-variable model whose sampler is differentiable, alongside a fidelity term that does require likelihoods. Second, the orientation caveat of Remark 10 applies: the repair guarantee of Theorem 6 is proved for base-anchored objectives and does not transfer to this data-first geometry, while the monotone path of Theorem 4, which is orientation- and convexity-agnostic, continues to describe the global minimizers of (24) and (25) as λ grows. When the reference is instead a smooth law, such as a pretrained teacher in fine-tuning rather than $\widehat{P}_n$, the model-first KL anchor becomes directly usable.

Proposition 8 (Likelihood, data-first KL, and the ELBO)**.** *Let ν be a σ-finite measure on $\mathcal{X}$ and let each model law Q_ϑ have ν-density q_ϑ.*

(i) Likelihood is data-first KL. *Suppose $P_{\mathrm{data}} \ll \nu$ with density p_0 and $\mathbb{E}_{P_{\mathrm{data}}}|\log p_0(X)| < \infty$. Then, for every ϑ,*

$$\mathbb{E}_{P_{\mathrm{data}}}\big[-\log q_\vartheta(X)\big] = \mathrm{KL}(P_{\mathrm{data}}\|Q_\vartheta) + h_\nu(P_{\mathrm{data}}),$$

where $h_\nu(P_{\mathrm{data}}) := -\mathbb{E}_{P_{\mathrm{data}}}\big[\log p_0(X)\big]$ is finite and independent of ϑ, and the two sides are finite or $+\infty$ together.

(ii) Structure of the data-first divergence. *For fixed P, the map $Q \mapsto \mathrm{KL}(P\|Q)$ is convex on $\mathcal{P}$, and if $\mathcal{X}$ is Polish it is weakly lower semicontinuous. Consequently Theorem 4 applies to the objectives* (24) *and* (25) *whenever global minimizers exist, while Theorem 6, proved for the base-anchored orientation, does not transfer (Remark 10).*

(iii) ELBO gap. *Let $q_\vartheta(x) = \int p_\vartheta(x \mid z)\, p_Z(dz)$ be a latent-variable model and $r_\eta(\cdot \mid x)$ an encoder with $r_\eta(\cdot \mid x) \ll p_\vartheta(\cdot \mid x)$, where $p_\vartheta(\cdot \mid x)$ is the model posterior. Then, for every x with $q_\vartheta(x) \in (0,\infty)$,*

$$-\mathrm{ELBO}(x;\vartheta,\eta) = -\log q_\vartheta(x) + \mathrm{KL}\big(r_\eta(\cdot \mid x) \,\big\|\, p_\vartheta(\cdot \mid x)\big) \ \geq \ -\log q_\vartheta(x),$$

with equality if and only if the encoder matches the model posterior at x.

Proof. First, part (i). Decompose $-\log q_\vartheta = \log(p_0/q_\vartheta) - \log p_0$ on the set $\{p_0 > 0\}$, which carries full P_{data}-mass. The second term integrates to $h_\nu(P_{\mathrm{data}})$, finite by hypothesis. For the first term, set $r := q_\vartheta/p_0$ on $\{p_0 > 0\}$; the positive part of $\log r$ is P_{data}-integrable, since $\log r \leq r - 1$ gives $\mathbb{E}_{P_{\mathrm{data}}}[(\log r)_+] \leq \mathbb{E}_{P_{\mathrm{data}}}[r] = \int_{\{p_0>0\}} q_\vartheta \, d\nu \leq 1$, so $\mathbb{E}_{P_{\mathrm{data}}}[-\log r] = \mathbb{E}_{P_{\mathrm{data}}}[\log(p_0/q_\vartheta)]$ is well defined in $(-\infty,+\infty]$. If $P_{\mathrm{data}} \ll Q_\vartheta$, then $dP_{\mathrm{data}}/dQ_\vartheta = p_0/q_\vartheta$ holds P_{data}-a.s. and $\mathbb{E}_{P_{\mathrm{data}}}[\log(p_0/q_\vartheta)] = \mathrm{KL}(P_{\mathrm{data}}\|Q_\vartheta)$ by definition. If $P_{\mathrm{data}} \not\ll Q_\vartheta$, pick A with $Q_\vartheta(A) = 0 < P_{\mathrm{data}}(A)$; then $q_\vartheta = 0$ ν-a.e. on A, so $\log(p_0/q_\vartheta) = +\infty$ on a set of positive P_{data}-measure and, the negative part being integrable, $\mathbb{E}_{P_{\mathrm{data}}}[\log(p_0/q_\vartheta)] = +\infty = \mathrm{KL}(P_{\mathrm{data}}\|Q_\vartheta)$ under the extended-value convention. In both cases the displayed identity holds, with both sides finite or $+\infty$ together since $h_\nu(P_{\mathrm{data}})$ is finite.

Next, part (ii). For convexity, fix Q_0, Q_1 and $\theta \in (0,1)$, and let ν' be a σ-finite measure dominating P, Q_0, and Q_1 (for instance $P + Q_0 + Q_1$), with densities p, q_0, q_1; the mixture $Q_\theta = (1-\theta)Q_0 + \theta Q_1$ has density $q_\theta = (1-\theta)q_0 + \theta q_1$. For fixed x with $p(x) > 0$, the map $q \mapsto p(x)\log(p(x)/q)$ is convex in $q > 0$ (as $-\log$ is convex), and extends convexly to $q \geq 0$ with value $+\infty$ at $q = 0$; composing with the affine $\theta \mapsto q_\theta(x)$ and integrating $d\nu'$ preserves convexity, giving $\mathrm{KL}(P\|Q_\theta) \leq (1-\theta)\mathrm{KL}(P\|Q_0) + \theta\mathrm{KL}(P\|Q_1)$. For lower semicontinuity, we invoke the Donsker–Varadhan variational formula, a standard fact: for probability measures on a Polish space,

$$\mathrm{KL}(P\|Q) = \sup_{f \in C_b(\mathcal{X})} \big\{\mathbb{E}_P[f] - \log \mathbb{E}_Q[e^f]\big\}.$$

For each fixed $f \in C_b(\mathcal{X})$, the map $Q \mapsto \mathbb{E}_P[f] - \log \mathbb{E}_Q[e^f]$ is weakly continuous: e^f is bounded continuous, so $Q \mapsto \mathbb{E}_Q[e^f]$ is weakly continuous with values in the compact interval $[e^{-\|f\|_\infty}, e^{\|f\|_\infty}] \subset (0,\infty)$, on which log is continuous. A supremum of weakly continuous functions is weakly lower semicontinuous, which proves the claim. The consequences for Theorems 4 and 6 are as stated: the former uses only the existence of global minimizers and is agnostic to orientation and convexity, while the latter's three-point argument differentiates the Bregman divergence in its *first* argument and is unavailable in the data-first orientation.

Finally, part (iii). Write the ELBO with encoder r_η:

$$\mathrm{ELBO}(x; \vartheta, \eta) = \mathbb{E}_{Z \sim r_\eta(\cdot|x)}\big[\log p_\vartheta(x \mid Z) + \log p_Z(Z) - \log r_\eta(Z \mid x)\big],$$

with $\log p_Z$ understood as the density of the prior with respect to the latent reference measure. By Bayes' rule, $p_\vartheta(z \mid x) = p_\vartheta(x \mid z)\, p_Z(z)/q_\vartheta(x)$ for $q_\vartheta(x) \in (0,\infty)$, so $\log p_\vartheta(x \mid z) + \log p_Z(z) = \log p_\vartheta(z \mid x) + \log q_\vartheta(x)$, and substituting,

$$\begin{aligned}\mathrm{ELBO}(x; \vartheta, \eta) &= \log q_\vartheta(x) - \mathbb{E}_{Z \sim r_\eta(\cdot|x)}\Big[\log \frac{r_\eta(Z \mid x)}{p_\vartheta(Z \mid x)}\Big] \\ &= \log q_\vartheta(x) - \mathrm{KL}\big(r_\eta(\cdot \mid x) \,\big\|\, p_\vartheta(\cdot \mid x)\big).\end{aligned}$$

Negating gives the display; nonnegativity of KL gives the inequality, with equality if and only if $r_\eta(\cdot \mid x) = p_\vartheta(\cdot \mid x)$. □

E Proofs for Section 4

This appendix proves the results of Section 4: the bias and consistency of the empirical entropy wall in Theorem 5, the sub-wall repair guarantee, and the wall-crossing statement (Proposition 9).

Theorem 5 (Bias and consistency of the empirical wall)**.** *Let $X_1, X_2, \ldots$ be i.i.d. from P_{data}, let $\phi : \mathcal{X} \to \mathbb{R}^d$ satisfy $\|\phi(x)\|_2 = 1$, and let $\widehat{P}_N = \frac{1}{N}\sum_{i=1}^N \delta_{X_i}$. Then, for either $H \in \{H_0, H_\varepsilon\}$ and every $N \geq 1$:*

(i) downward bias: $\mathbb{E}\, H(\widehat{P}_N) \leq H(P_{\mathrm{data}})$;

(ii) monotonicity in the sample size: $\mathbb{E}\, H(\widehat{P}_{N+1}) \geq \mathbb{E}\, H(\widehat{P}_N)$;

(iii) consistency: $\mathbb{E}\, H(\widehat{P}_N) \longrightarrow H(P_{\mathrm{data}})$ *as* $N \to \infty$.

Proof of Theorem 5. We prove the three parts in turn.

First, for part (i), recall from Lemma 2 that $H \in \{H_0, H_\varepsilon\}$ is concave in its distribution argument through the affine map $Q \mapsto \Sigma_Q$. The empirical covariance is unbiased:

$$\mathbb{E}\, \Sigma_{\widehat{P}_N} = \frac{1}{N}\sum_{i=1}^N \mathbb{E}\big[\phi(X_i)\phi(X_i)^\top\big] = \Sigma_{P_{\mathrm{data}}}.$$

Jensen's inequality for the concave map $\Sigma \mapsto H$ then gives

$$\mathbb{E}\, H(\widehat{P}_N) \le H\big(\mathbb{E}\, \Sigma_{\widehat{P}_N}\big) = H(P_{\mathrm{data}}).$$

Next, for part (ii), we use a leave-one-out averaging identity. Fix $N+1$ samples and, for $j = 1, \dots, N+1$, let $\widehat{P}_N^{(-j)} = \frac{1}{N} \sum_{i \neq j} \delta_{X_i}$. Each index appears in exactly N of the $N+1$ leave-one-out measures, so

$$\widehat{P}_{N+1} = \frac{1}{N+1} \sum_{j=1}^{N+1} \widehat{P}_N^{(-j)}, \qquad \text{equivalently} \qquad \Sigma_{\widehat{P}_{N+1}} = \frac{1}{N+1} \sum_{j=1}^{N+1} \Sigma_{\widehat{P}_N^{(-j)}}.$$

Concavity of H gives, pathwise,

$$H(\widehat{P}_{N+1}) \ge \frac{1}{N+1} \sum_{j=1}^{N+1} H\big(\widehat{P}_N^{(-j)}\big).$$

Each $\widehat{P}_N^{(-j)}$ has the same distribution as $\widehat{P}_N$, so taking expectations yields $\mathbb{E}H(\widehat{P}_{N+1}) \ge \mathbb{E}H(\widehat{P}_N)$.

Finally, for part (iii), note that since $\|\phi(x)\|_2 = 1$, the summands $\phi(X_i)\phi(X_i)^\top$ are i.i.d. bounded random matrices with mean $\Sigma_{P_{\mathrm{data}}}$, so by the (matrix) strong law of large numbers $\Sigma_{\widehat{P}_N} \to \Sigma_{P_{\mathrm{data}}}$ almost surely. Both H_0 and H_ε are continuous functions of the covariance matrix on the compact set of density matrices and bounded in $[0, \log d]$, so $H(\widehat{P}_N) \to H(P_{\mathrm{data}})$ a.s.; dominated convergence then gives $\mathbb{E}H(\widehat{P}_N) \to H(P_{\mathrm{data}})$. □

Remark 9 (Interpretation and experimental consequence)**.** *Theorem 5 upgrades the informal assumption of downward diversity bias to a theorem for the empirical law; it does not by itself establish that a trained generator Q_0 satisfies $H_0(Q_0) < \rho_\star$, which remains a separate empirical claim, consistent with reported spectral deficits in modern generators. Because $\widehat{\rho}_\star = H_0(\widehat{P}_N)$ is downward biased, and a finite generated batch inherits the same downward bias when estimating $H_0(Q_\lambda)$, wall-crossing plots should use matched sample sizes, repeated subsampling, or a bias-aware estimator.*

Theorem 6 (Population discrepancy improves up to the wall)**.** *Let Φ be Fréchet differentiable and strictly convex on an open convex set containing $\mathcal{P}$, with Bregman divergence $D_\Phi(P, Q) = \Phi(P) - \Phi(Q) - \langle \nabla\Phi(Q), P - Q\rangle$, and let $Q_\lambda \in \operatorname*{argmin}_{Q \in \mathcal{P}} \{D_\Phi(Q, Q_0) - \lambda H(Q)\}$ with H concave and $\mathcal{P}$ convex. If $P_{\mathrm{data}} \in \mathcal{P}$ and $H(Q_\lambda) \le H(P_{\mathrm{data}})$, then*

$$D_\Phi(P_{\mathrm{data}}, Q_\lambda) + D_\Phi(Q_\lambda, Q_0) \le D_\Phi(P_{\mathrm{data}}, Q_0),$$

and hence in particular $D_\Phi(P_{\mathrm{data}}, Q_\lambda) \le D_\Phi(P_{\mathrm{data}}, Q_0)$.

Proof. By Proposition 6 applied to $J(Q) = D_\Phi(Q, Q_0)$ over the convex $\mathcal{P}$, $Q_\lambda \in \arg\min\{D_\Phi(Q, Q_0) : Q \in \mathcal{P},\ H(Q) \ge \rho_\lambda\}$ with $\rho_\lambda = H(Q_\lambda)$. The feasible set $\mathcal{C} = \mathcal{P} \cap \{H \ge \rho_\lambda\}$ is convex (H concave). The below-wall hypothesis $H(P_{\mathrm{data}}) \ge \rho_\lambda$ and $P_{\mathrm{data}} \in \mathcal{P}$ give $P_{\mathrm{data}} \in \mathcal{C}$.

Because $Q \mapsto D_\Phi(Q, Q_0)$ is convex and differentiable in its first argument (with $\nabla_Q D_\Phi(Q, Q_0) = \nabla\Phi(Q) - \nabla\Phi(Q_0)$) and $\mathcal{C}$ is convex, first-order optimality of the minimizer Q_λ over $\mathcal{C}$ gives, for every feasible P,

$$\langle \nabla\Phi(Q_\lambda) - \nabla\Phi(Q_0), P - Q_\lambda \rangle \ge 0.$$

Setting $P = P_{\mathrm{data}} \in \mathcal{C}$ and using the three-point Bregman identity

$$\begin{aligned} D_\Phi(P_{\mathrm{data}}, Q_0) = D_\Phi(P_{\mathrm{data}}, Q_\lambda) + D_\Phi(Q_\lambda, Q_0) \\ + \langle \nabla\Phi(Q_\lambda) - \nabla\Phi(Q_0), P_{\mathrm{data}} - Q_\lambda \rangle, \end{aligned}$$

together with nonnegativity of the inner-product term, yields $D_\Phi(P_{\mathrm{data}}, Q_0) \ge D_\Phi(P_{\mathrm{data}}, Q_\lambda) + D_\Phi(Q_\lambda, Q_0)$. Nonnegativity of $D_\Phi(Q_\lambda, Q_0)$ gives the second inequality. □

Remark 10 (What the repair theorem does not cover)**.** *Theorem 6 concerns the base-anchored orientation $D_\Phi(Q, Q_0)$ (optimized law first) and does not transfer to a data-first objective $D_\Phi(\widehat{P}_n, Q)$: the gradient of a Bregman divergence in its* second *argument involves the Hessian of Φ and is not $\nabla\Phi(Q) - \nabla\Phi(\widehat{P}_n)$, so the three-point identity no longer collapses the cross term. In particular, the maximum-likelihood instantiation of Section 3.3, whose fidelity is the data-first KL by Proposition 8(i), inherits the monotone path of Theorem 4 but not the repair guarantee. Two clean options remain: (i) the base-anchored objective $D_\Phi(Q, Q_0) - \lambda H(Q)$, especially with $Q_0 = P_\theta$. (ii) a symmetric Hilbertian discrepancy such as squared MMD, for which the orientation is immaterial.*

Proposition 9 (Local wall crossing)**.** *Let $\mathcal{P}$ be convex, $P_{\mathrm{data}} \in \mathcal{P}$, and suppose there exists $R \in \mathcal{P}$ with $H_0(R) > H_0(P_{\mathrm{data}})$. Put $Q_t = (1-t)P_{\mathrm{data}} + tR$ for $t \in (0, 1]$. If $\mathcal{D}(Q_t; P_{\mathrm{data}}) \to 0$ as $t \downarrow 0$, then for every $\delta > 0$ there exists $Q \in \mathcal{P}$ with $\mathcal{D}(Q; P_{\mathrm{data}}) \le \delta$ and $H_0(Q) > H_0(P_{\mathrm{data}})$.*

Proof. By convexity of $\mathcal{P}$, $Q_t \in \mathcal{P}$. By concavity of H_0 (Lemma 2), $H_0(Q_t) \ge (1-t)H_0(P_{\mathrm{data}}) + tH_0(R)$; since $H_0(R) > H_0(P_{\mathrm{data}})$, the right-hand side equals $H_0(P_{\mathrm{data}}) + t(H_0(R) - H_0(P_{\mathrm{data}})) > H_0(P_{\mathrm{data}})$ for all $t \in (0, 1]$, so $H_0(Q_t) > H_0(P_{\mathrm{data}})$ (plain concavity suffices; no strictness is used). By hypothesis choose $t_\delta > 0$ with $\mathcal{D}(Q_{t_\delta}; P_{\mathrm{data}}) \le \delta$; then $Q = Q_{t_\delta}$ satisfies both requirements. □

F Proofs for Section 3.2 and Section 5

This appendix proves the sampling-time results in the following order. We first record a tightness lemma for KL sublevel sets, which drives the existence argument in Theorem 1, and compute the first variation of the smoothed entropy (Lemma 1). We then record the formal stationarity condition behind the density-ratio representations (Remark 11), prove the tilt characterization (Proposition 2), and derive Theorem 1 by combining the existence argument with that proposition. The remainder of the appendix treats the propagation of the tilt through the noising process and the endpoint guarantees.

Lemma 8 (Relative-entropy sublevel tightness)**.** *Let P be a probability measure on a Polish space and $C \ge 0$. The sublevel set $\{Q : \mathrm{KL}(Q\|P) \le C\}$ is tight. Concretely, for any measurable A with $0 < P(A) < 1$ and any Q with $\mathrm{KL}(Q\|P) \le C$,*

$$Q(A) \;\le\; \frac{C+1}{\log(1/P(A))}. \tag{26}$$

Proof. By the data-processing inequality applied to the binary partition $\{A, A^c\}$,

$$\mathrm{KL}(Q\|P) \ge d_2\big(Q(A) \,\|\, P(A)\big), \qquad d_2(q\|p) = q\log\tfrac{q}{p} + (1-q)\log\tfrac{1-q}{1-p},$$

where d_2 is the binary KL, with the usual conventions and $d_2(q\|p) = +\infty$ if $p \in \{0, 1\}$ while $q \notin \{0, 1\}$. Write $p = P(A) \in (0, 1)$ and $q = Q(A)$. Since $1 - p \le 1$ gives $\log\frac{1-q}{1-p} \ge \log(1-q)$, and $t\log t \ge -e^{-1}$ on $[0, 1]$, the terms $q\log q$ and $(1-q)\log\frac{1-q}{1-p}$ together contribute at least $-2/e \ge -1$; hence

$$d_2(q\|p) = q\log(1/p) + q\log q + (1-q)\log\tfrac{1-q}{1-p} \ge q\log(1/p) - 1;$$

hence $Q(A)\log(1/P(A)) \le \mathrm{KL}(Q\|P) + 1 \le C + 1$, which is (26).

For tightness, fix $\eta > 0$. Since every probability measure on a Polish space is tight, there is a compact K with $P(K^c)$ as small as desired; taking $P(K^c)$ small enough that $(C+1)/\log(1/P(K^c)) \le \eta$ yields $\sup_{\mathrm{KL}(Q\|P)\le C} Q(K^c) \le \eta$. □

Proof of Lemma 1. Write $A_\nu := \int \phi(x)\phi(x)^\top d\nu(x)$, a symmetric matrix (finite since ϕ is bounded and ν is finite). Since $Q \mapsto \Sigma_Q$ is affine, $\Sigma_{Q+t\nu} = \Sigma_Q + tA_\nu$ and hence $S^\varepsilon_{Q+t\nu} = S^\varepsilon_Q + t(1-\varepsilon)A_\nu$ for all t for which

$Q + t\nu$ is a probability measure. The matrix entropy $\mathsf{H}(S) = -\operatorname{Tr}(S \log S)$ is Fréchet differentiable at every positive definite S, with derivative $D\mathsf{H}(S)[B] = -\operatorname{Tr}\big(B(\log S + I_d)\big)$ for symmetric B; this applies at $S = S_Q^\varepsilon$ because $S_Q^\varepsilon \succeq (\varepsilon/d) I_d \succ 0$, and the perturbed matrices $S_Q^\varepsilon + t(1-\varepsilon)A_\nu$ remain in a compact neighborhood of positive definite matrices for small t. By the chain rule along the affine path,

$$\begin{aligned}\frac{d}{dt}\, H_\varepsilon(Q + t\nu)\Big|_{t=0^+} &= -(1-\varepsilon)\ \operatorname{Tr}\big(A_\nu(\log S_Q^\varepsilon + I_d)\big) \\ &= (1-\varepsilon)\int \phi(x)^\top\big(-\log S_Q^\varepsilon\big)\phi(x)\, d\nu(x) \\ &\qquad - (1-\varepsilon)\int \|\phi(x)\|_2^2\ d\nu(x).\end{aligned}$$

Since $\|\phi(x)\|_2^2 = 1$, the last integral equals $\nu(\mathcal{X}) = 0$, and the first term is $\int G_Q^\varepsilon\, d\nu$ by definition (8). □

We note that the normalization $\|\phi\|_2 = 1$ makes the trace term of the derivative drop out, so G_Q^ε represents the first variation of the entropy functional: the first variation of H_ε at Q, along any admissible mass-preserving perturbation, integrates G_Q^ε against the perturbation.

Remark 11 (Bregman stationarity)**.** *For a general Bregman anchor, a formal first-order condition explains how the geometry of Φ converts the entropy first variation into a displacement of the law: an interior optimizer $Q^\star$ of* (11) *satisfies*

$$\nabla\Phi(Q^\star) - \nabla\Phi(P_\theta) = \lambda\, g_{Q^\star} + c,$$

where $g_Q = G_Q^\varepsilon$ is the first variation of H_ε (Lemma 1) and c is the scalar multiplier of the unit-mass constraint. We do not rely on this identity: in the KL geometry, the derivation below obtains the density ratio directly from the first-variation computation, with no interiority hypothesis.

We now prove the tilt characterization of Section 3.2.

Proof of Proposition 2. Throughout, write $\mathcal{P}_\theta = \{Q \in \mathcal{P} : Q \ll P_\theta\}$, a convex set, let F be extended by $+\infty$ off $\mathcal{P}_\theta$, and recall from the statement that F attains a finite minimum on $\mathcal{P}_\theta$; let $Q^\star$ be any minimizer and $q^\star = dQ^\star/dP_\theta$. We first prove uniqueness, then mutual absolute continuity, then derive the tilt from the first-variation computation, and finally bound the density ratio.

Uniqueness. On its finite domain the KL term is strictly convex in Q, and $-\lambda H_\varepsilon$ is convex by Lemma 2. Hence F is strictly convex where finite, and its minimizer is unique.

Mutual absolute continuity. We have $Q^\star \ll P_\theta$ with $\mathrm{KL}(Q^\star\|P_\theta) < \infty$ by finiteness of the minimum; it remains to prove $q^\star > 0$ P_θ-a.s. Suppose instead that $q^\star = 0$ on a measurable set A with $P_\theta(A) > 0$. Let $R = P_\theta(\cdot \mid A)$ and $Q_t = (1-t)Q^\star + tR \in \mathcal{P}_\theta$ for $t \in (0,1)$. Exactly as in the corresponding computation for the KL term (splitting the integral over A, where the density of Q_t is $t\,\mathbf{1}_A/P_\theta(A)$, and A^c, where it is scaled by $1-t$),

$$\mathrm{KL}(Q_t\|P_\theta) - \mathrm{KL}(Q^\star\|P_\theta) = t\log t + O(t).$$

The remaining entropy term of F changes by only $O(t)$. Indeed, $S_{Q_t}^\varepsilon - S_{Q^\star}^\varepsilon = t(1-\varepsilon)(\Sigma_R - \Sigma_{Q^\star})$ has norm $O(t)$ (with constants depending only on $\|\phi\|_2 = 1$), and $S \mapsto -\operatorname{Tr}(S\log S)$ is Lipschitz on the compact spectral range $[\varepsilon/d, 1]$, its derivative $-(\log S + I)$ being bounded in operator norm by $\log(d/\varepsilon) + 1$ there. Therefore

$$F(Q_t) - F(Q^\star) = t\log t + O(t) < 0 \qquad \text{for small } t > 0,$$

since $t\log t \to 0^-$ dominates $O(t)$; this contradicts optimality. Thus $q^\star > 0$ P_θ-a.s. and $Q^\star \sim P_\theta$.

First variation and the tilt. For bounded measurable h with $\mathbb{E}_{Q^\star} h = 0$, set $dQ_t = (1+th)\, dQ^\star$, a valid probability law for $|t| \le 1/(1+\|h\|_\infty)$, and let $\nu = h\, dQ^\star$, a finite signed measure with $\nu(\mathcal{X}) = 0$. By Lemma 1

and the identity $g_Q = G^\varepsilon_Q$, the entropy term has derivative $\frac{d}{dt}H_\varepsilon(Q_t)|_{t=0} = \int g_{Q^\star}\, h\, dQ^\star$. The KL term has derivative

$$\frac{d}{dt}\mathrm{KL}(Q_t\|P_\theta)\Big|_{t=0} = \int (\log q^\star + 1)\, h\, dQ^\star = \int \log q^\star\, h\, dQ^\star,$$

where the $+1$ term vanishes since $\mathbb{E}_{Q^\star}h = 0$; differentiation under the integral is justified by dominated convergence, as h is bounded and $\int q^\star|\log q^\star|\, dP_\theta < \infty$ from $\mathrm{KL}(Q^\star\|P_\theta) < \infty$ together with the uniform bound $t\log t \geq -e^{-1}$. First-order optimality $\frac{d}{dt}F(Q_t)|_{t=0} = 0$ for all such h therefore gives

$$\int \Big(\log q^\star - \lambda g_{Q^\star}\Big)\, h\, dQ^\star = 0$$

for all bounded h with $\mathbb{E}_{Q^\star}h = 0$, so the integrand in parentheses is $Q^\star$-a.s. (and, by mutual absolute continuity, P_θ-a.s.) equal to a constant. Exponentiating and normalizing, with the constant absorbed into the normalizer, yields (14).

Boundedness of the ratio. The eigenvalues of S^ε_Q lie in $[\varepsilon/d, 1]$, so $\big\|-\log S^\varepsilon_Q\big\|_{\mathrm{op}} \leq \log(d/\varepsilon)$ and $0 \leq \lambda g_Q \leq \lambda(1-\varepsilon)\log(d/\varepsilon)$ uniformly over Q and x. The exponent in (14) is therefore uniformly bounded, so the normalizer lies in $(0,\infty)$ and the density ratio is bounded above and below by positive constants. □

Proof of Theorem 1. We first establish existence, then obtain the remaining claims from Proposition 2.

Existence. Work on $\mathcal{P}_\theta = \{Q : Q \ll P_\theta\}$, a convex set, and extend F by $+\infty$ off it. The infimum is finite: $F(P_\theta) = -\lambda H_\varepsilon(P_\theta) \in [-\lambda\log d, 0]$, while $F \geq -\lambda\log d > -\infty$ termwise. Let (Q_n) be a minimizing sequence. Since $0 \leq H_\varepsilon(Q) \leq \log d$, boundedness of $F(Q_n)$ implies $\sup_n \mathrm{KL}(Q_n\|P_\theta) =: C < \infty$. By Lemma 8, the KL sublevel set $\{Q : \mathrm{KL}(Q\|P_\theta) \leq C\}$ is tight, so (Q_n) is tight. By Prokhorov's theorem a subsequence converges weakly to some $Q^\star$. Since ϕ is bounded and continuous, $Q \mapsto \Sigma_Q, S^\varepsilon_Q$ are weakly continuous; because $S^\varepsilon_Q \succeq (\varepsilon/d)I_d \succ 0$ uniformly, $S \mapsto -\operatorname{Tr}(S\log S)$ is continuous on the relevant compact spectral range, so $Q \mapsto H_\varepsilon(Q)$ is weakly continuous. The map $Q \mapsto \mathrm{KL}(Q\|P_\theta)$ is weakly l.s.c.: by the Donsker–Varadhan formula, $\mathrm{KL}(Q\|P_\theta) = \sup_{f\in C_b(\mathcal{X})}\{\mathbb{E}_Q[f] - \log\mathbb{E}_{P_\theta}[e^f]\}$ is a supremum of weakly continuous functions of Q, exactly as in the proof of Proposition 8(ii) with the roles of the two arguments exchanged. Therefore $F(Q^\star) \leq \liminf_n F(Q_n)$, so $Q^\star$ attains the infimum; in particular $\mathrm{KL}(Q^\star\|P_\theta) < \infty$, so $Q^\star \ll P_\theta$.

Specialization. By the existence step, F attains a finite minimum on $\mathcal{P}_\theta$, so Proposition 2 applies: the minimizer $Q^\star$ is unique and mutually absolutely continuous with P_θ, which proves part (i), and the tilt (14) holds with the total reward $R_{Q^\star}$ of (13), proving part (ii).

Boundedness. The spectral floor gives $0 \leq \lambda\, G^\varepsilon_{Q^\star} \leq \lambda(1-\varepsilon)\log(d/\varepsilon)$. Thus $R_{Q^\star}$ is uniformly bounded, so $Z = \mathbb{E}_{P_\theta}\exp(R_{Q^\star}) \in (0,\infty)$ and $dQ^\star/dP_\theta$ is bounded above and below by positive constants; hence $Q^\star$ has the same P_θ-essential support as P_θ, proving part (iii). □

Remark 12 (Scope of the tilt characterization)**.** *The characterization* (14) *is a fixed point: $R_{Q^\star}$ depends on $Q^\star$ through $S^\varepsilon_{Q^\star}$ (contrast Lemma 4, where the reward is fixed and the tilt is explicit). The KL anchor only reweights within* $\mathrm{supp}(P_\theta)$ *and creates no mass where $P_\theta = 0$. Strict convexity proves uniqueness of the target law but does not imply that any particular fixed-point iteration is contractive; convergence of a numerical solver requires a separate argument.*

Proof of Theorem 2. For measurable A,

$$q^\star_t(A) = \int K_t(A \mid x_0)\, Q^\star(dx_0) = Z^{-1}\int K_t(A \mid x_0)\, w(x_0)\, P_\theta(dx_0).$$

Disintegrate the base joint law of (X_0, X_t) as $K_t(dx_t \mid x_0)P_\theta(dx_0) = P_\theta(dx_0 \mid x_t)\, p_t(dx_t)$. By Fubini's theorem (applicable since w is bounded, by Theorem 1),

$$q^\star_t(A) = Z^{-1}\int_A \Big(\int w(x_0)P_\theta(dx_0 \mid x_t)\Big)p_t(dx_t) = Z^{-1}\int_A h_t(x_t)\, p_t(dx_t),$$

with $h_t(x_t) = \mathbb{E}_{P_\theta}[w(X_0) \mid X_t = x_t]$. Since A was arbitrary, $dq_t^\star/dp_t = h_t/Z$. If both marginals have positive differentiable densities, then $\log q_t^\star = \log p_t + \log h_t - \log Z$; as Z is constant in x_t, $\nabla \log q_t^\star = \nabla \log p_t + \nabla \log h_t$. □

Theorem 7 (Exact reverse process)**.** *Suppose the forward SDE $dX_t = f(X_t, t)\,dt + g(t)\,dW_t$ (for t increasing from 0 to T) admits strictly positive differentiable marginal densities and satisfies the standard regularity conditions for time reversal and the probability-flow construction. We use the standard reverse-time convention in which the displayed equations are integrated with* decreasing *t from T to 0. Then the reverse-time SDE, initialized at $q_T^\star$,*

$$dX_t = \big[f(X_t, t) - g(t)^2\big(\nabla \log p_t(X_t) + u_t(X_t)\big)\big]\,dt + g(t)\,d\overline{W}_t,$$

has time-zero law exactly $Q^\star$, where $\overline{W}$ is a reverse-time Brownian motion. The probability-flow ODE

$$\dot{X}_t = f(X_t, t) - \tfrac{1}{2}\,g(t)^2\big(\nabla \log p_t(X_t) + u_t(X_t)\big),$$

likewise integrated with decreasing t, has the same one-time marginals. Equivalently, under the forward reparameterization $\tau = T - t$ and $Y_\tau = X_{T-\tau}$, both dynamics run with increasing τ and their drifts are the negatives *of the displayed drifts evaluated at $t = T - \tau$ (the diffusion term is unchanged).*

Proof. Anderson's time-reversal theorem, under the stated regularity, gives the reverse-time SDE (integrated with decreasing t) for the process with marginals $q_t^\star$ as having drift $f - g^2\nabla \log q_t^\star$ and initial law $q_T^\star$. Substituting $\nabla \log q_t^\star = \nabla \log p_t + u_t$ from Theorem 2 yields the stated drift, and the marginals are $q_t^\star$ for all t, in particular $Q^\star$ at $t = 0$. The probability-flow ODE $\dot{X}_t = f - \frac{1}{2}g^2\nabla \log q_t^\star$ is the deterministic process with identical one-time marginals under the $q_T^\star$ initialization. The $\tau = T - t$ statement follows from the chain rule $\frac{d}{d\tau}Y_\tau = -\frac{d}{dt}X_t|_{t=T-\tau}$, which flips the sign of every drift while preserving the (sign-indifferent) diffusion coefficient. If instead one initializes at $p_T \neq q_T^\star$, the time-zero law is not $Q^\star$; the discrepancy is quantified in Theorem 8. □

Remark 13 (Initialization)**.** *Practical samplers initialize from p_T, not $q_T^\star$. These coincide only when h_T is constant. Because $dq_T^\star/dp_T = h_T/Z$, the mismatch is* $\mathrm{KL}(p_T \| q_T^\star) = \mathbb{E}_{p_T} \log \frac{dp_T}{dq_T^\star} = \log Z - \mathbb{E}_{p_T} \log h_T(X_T)$*, which enters any rigorous comparison between a deployed sampler and $Q^\star$ (Theorem 8). Theorem 7 is exact only with the $q_T^\star$ initialization.*

Definition 2 (Plug-in fields)**.** Assume $\mathcal{X} \subseteq \mathbb{R}^D$ and that ϕ and the denoiser are differentiable. With guidance scale $\omega_t \geq 0$,

$$\widetilde{u}_t^{\mathrm{chain}}(x_t) = \frac{\omega_t}{a} J_{\widehat{x}_0}(x_t, t)^\top \nabla_x R_{Q^\star}(\widehat{x}_0(x_t, t)),$$

where $J_{\widehat{x}_0}$ is the Jacobian of the denoiser. The *direct-injection* variant, applicable when clean and noisy states share dimension, is

$$\widetilde{u}_t^{\mathrm{dir}}(x_t) = \omega_t \nabla_x R_{Q^\star}(\widehat{x}_0(x_t, t)),$$

which uses a chosen state-space direction rather than the derivative of the composite map $x_t \mapsto R_{Q^\star}(\widehat{x}_0(x_t, t))$.

Remark 14 (Plug-in is uncontrolled)**.** *No general equality or one-sided bound relates u_t and $\widetilde{u}_t$: $R_{Q^\star}$ is nonlinear, and neither conditional expectation nor differentiation commutes with a point-mass substitution. For the reward $R_{Q^\star} = \lambda G^\varepsilon_{Q^\star}$ of* (13)*,*

$$\nabla_x R_{Q^\star}(x) = J_\phi(x)^\top \big[2\lambda(1-\varepsilon)(-\log S^\varepsilon_{Q^\star})\phi(x)\big],$$

where J_ϕ is the Jacobian of ϕ (the factor 2 comes from differentiating the quadratic form $\phi^\top M \phi$ with symmetric $M = -\log S^\varepsilon_{Q^\star}$, which is the x-gradient of $\lambda G^\varepsilon_{Q^\star}$).

Remark 15 (Discrete updates are approximations)**.** *The DDPM/DDIM updates* (20)–(21) *are algebraically consistent with the corrected noise prediction under the noise–score convention $\nabla \log p_t = -\varepsilon_\theta/\sqrt{1-\bar{\alpha}_t}$, and are discrete implementations inspired by Theorem 2. They do not exactly sample $Q^\star$ even if $\widetilde{u}_t = u_t$; DDIM adds a further ODE discretization and path-selection approximation. A discretization term must therefore be added to the continuous-time bound of Theorem 8.*

Theorem 8 (Endpoint KL and TV bounds)**.** *Let the exact reverse process have initial law $q_T^\star$, base score $s_t = \nabla \log p_t$, and exact guidance $u_t = \nabla \log h_t$; let the deployed process have initial law π_T, learned score $\widehat{s}_t$, and approximate guidance $\widetilde{u}_t$. Put*

$$e_t = \widehat{s}_t - s_t, \qquad \delta_t = \widetilde{u}_t - u_t.$$

Assume that:

(i) *both continuous-time processes share the diffusion coefficient $g(t)I$ with $g(t) > 0$ on $(0, T)$ (nondegeneracy on the open interval); any endpoint degeneracy $g(0) = 0$ or $g(T) = 0$ is handled by truncating to $[\eta, T - \eta]$, applying the bound there, and letting $\eta \downarrow 0$, assuming the resulting integral converges;*

(ii) *the absolute-continuity and Novikov conditions for Girsanov's theorem hold on each such subinterval.*

Then, for the orientation $\mathrm{KL}(\widehat{Q}\|Q^\star)$ between the clean endpoint laws $Q^\star, \widehat{Q}$ (deployed law first),

$$\mathrm{KL}(\widehat{Q}\|Q^\star) \le \mathrm{KL}(\pi_T\|q_T^\star) + \frac{1}{2}\int_0^T g(t)^2\, \mathbb{E}_{\widehat{\mathbb{P}}}\big[\|e_t(X_t) + \delta_t(X_t)\|_2^2\big]\, dt,$$

and, by Pinsker's inequality (again for the orientation $\mathrm{KL}(\widehat{Q}\|Q^\star)$),

$$\mathrm{TV}(\widehat{Q}, Q^\star) \le \Big[\tfrac{1}{2}\mathrm{KL}(\pi_T\|q_T^\star) + \tfrac{1}{4}\int_0^T g(t)^2\, \mathbb{E}_{\widehat{\mathbb{P}}}\big[\|e_t(X_t) + \delta_t(X_t)\|_2^2\big]\, dt\Big]^{1/2}.$$

Proof. Let $\mathbb{P}^\star$ be the path law of the exact reverse process (initial law $q_T^\star$, drift $b_t^\star = f - g^2(s_t + u_t)$) and $\widehat{\mathbb{P}}$ the path law of the deployed process (initial law π_T, drift $\widehat{b}_t = f - g^2(\widehat{s}_t + \widetilde{u}_t)$), both with diffusion coefficient $g(t)I$. The drift difference is $\widehat{b}_t - b_t^\star = -g(t)^2(e_t + \delta_t)$. First, we decompose the path-space relative entropy in the direction $\mathrm{KL}(\widehat{\mathbb{P}}\|\mathbb{P}^\star)$ (deployed first) by the chain rule over the initial time T,

$$\mathrm{KL}(\widehat{\mathbb{P}}\|\mathbb{P}^\star) = \mathrm{KL}(\pi_T\|q_T^\star) + \mathbb{E}_{\pi_T}\mathrm{KL}\big(\widehat{\mathbb{P}}(\cdot \mid X_T)\,\|\,\mathbb{P}^\star(\cdot \mid X_T)\big).$$

Next, we evaluate the conditional term. Conditionally on X_T, the two processes share the diffusion coefficient $g(t)I$, which is nondegenerate on $(0, T)$ (or on each $[\eta, T - \eta]$, with $\eta \downarrow 0$ afterwards), and differ only in drift, so on that interval the change of measure is absolutely continuous and Girsanov's theorem applies, giving

$$\begin{aligned}\mathrm{KL}\big(\widehat{\mathbb{P}}(\cdot \mid X_T)\,\|\,\mathbb{P}^\star(\cdot \mid X_T)\big) &= \tfrac{1}{2}\,\mathbb{E}_{\widehat{\mathbb{P}}}\Big[\int_0^T \big\|g(t)^{-1}(\widehat{b}_t - b_t^\star)\big\|^2\, dt \;\Big|\; X_T\Big] \\ &= \tfrac{1}{2}\,\mathbb{E}_{\widehat{\mathbb{P}}}\Big[\int_0^T g(t)^2\,\|e_t + \delta_t\|^2\, dt \;\Big|\; X_T\Big],\end{aligned}$$

using $g(t)^{-1}(\widehat{b}_t - b_t^\star) = -g(t)(e_t + \delta_t)$. Averaging over $X_T \sim \pi_T$, we arrive at

$$\mathrm{KL}(\widehat{\mathbb{P}}\|\mathbb{P}^\star) = \mathrm{KL}(\pi_T\|q_T^\star) + \frac{1}{2}\int_0^T g(t)^2\, \mathbb{E}_{\widehat{\mathbb{P}}}\,\|e_t + \delta_t\|^2\, dt.$$

Finally, the clean endpoint laws $\widehat{Q}, Q^\star$ are measurable images (the time-0 coordinate) of the path laws, so the data-processing inequality gives $\mathrm{KL}(\widehat{Q}\|Q^\star) \le \mathrm{KL}(\widehat{\mathbb{P}}\|\mathbb{P}^\star)$, which is the stated KL bound. Pinsker's inequality $\mathrm{TV}(\mu, \nu) \le \sqrt{\tfrac{1}{2}\mathrm{KL}(\mu\|\nu)}$ applied to $\widehat{Q}, Q^\star$ gives the TV bound. □

The exact-guidance bound is recovered only when $\pi_T = q_T^\star$ and $e_t \equiv 0$; a separate discretization term is still needed for the implemented DDPM/DDIM sampler (Remark 15).

Remark 16 (i.i.d. sampling holds only for fixed guidance)**.** *At the population level $Q^\star$ is a single law, so independent exact samplers with a fixed potential $R_{Q^\star}$ produce i.i.d. draws from $Q^\star$; this distinguishes IGA from methods that define diversity only through a coupled batch objective. However, $R_{Q^\star}$ depends on the unknown $Q^\star$ through $S^\varepsilon_{Q^\star}$. If a practical algorithm recomputes covariance or entropy gradients from the same batch being generated, each particle's drift depends on the others: the outputs are exchangeable but not independent. An i.i.d. guarantee requires one of the following:* frozen-potential sampling*, in which the potential is estimated in a separate stage, frozen, and used to run independent trajectories;* independent-pilot estimation*, in which the potential is estimated on an independent pilot sample; or a* mean-field analysis*, invoking a propagation-of-chaos argument when the potential is updated from the active batch. Absent these, finite-batch IGA guidance should be described as an interacting particle system.*

G Training-Time IGA for Diffusion Models: Proofs and Discussion

Diffusion models are trained through variational bounds, which places them in the maximum-likelihood family of Appendix D.2. Attaching $-\lambda H_\varepsilon(Q_\vartheta)$ directly to the denoising objective requires samples from Q_ϑ, and hence full reverse rollouts inside the training loop. The framework offers a rollout-free alternative: perform the IGA correction on the *data* before fitting the denoiser. To this end, we apply Proposition 2 with the reference distribution $\widehat{P}_n$; note that the proposition depends on P_θ only through its role as the reference measure, and the finite-minimum hypothesis holds automatically since the feasible set is the simplex over the training atoms. This application yields unique weights

$$q_i^\star = \frac{w_i}{\sum_{j=1}^n w_j}, \qquad w_i = \exp\Big(\lambda\, G^\varepsilon_{Q^\star_\lambda}(x_i)\Big), \tag{27}$$

which form a self-consistent softmax over the training set and define $Q^\star_\lambda = \sum_{i=1}^n q_i^\star \delta_{x_i}$. The weights can be computed as a finite-dimensional convex–concave saddle problem through the spectral dual of Proposition 1, with the spectral adversary and the reweighting playing the detection and response roles described after that proposition. The outcome can be viewed as a distributionally robust reweighting of the dataset, although not a worst-case-loss one (Remark 17, Appendix G). The following proposition shows that training on the reweighted data is justified exactly rather than heuristically:

Proposition 10 (IGA training as divergence minimization toward reweighted data)**.** *Let P_{ref} be a probability measure, let $\lambda \geq 0$, $\varepsilon \in (0,1)$, and let F be the objective* (12) *with P_{ref} in place of P_θ. Suppose F attains a finite minimum over $\{Q \in \mathcal{P} : Q \ll P_{\mathrm{ref}}\}$, at $Q^\star_\lambda$. Then for every Q with $F(Q) < \infty$,*

$$F(Q) - F(Q^\star_\lambda) = \mathrm{KL}(Q \,\|\, Q^\star_\lambda) + \lambda\, B_{-H_\varepsilon}(Q, Q^\star_\lambda), \tag{28}$$

where $B_{-H_\varepsilon}(Q,Q') := H_\varepsilon(Q') - H_\varepsilon(Q) + \int G^\varepsilon_{Q'}\, d(Q - Q') \geq 0$ is the Bregman divergence of the convex functional $-H_\varepsilon$.

Every term on the right-hand side of (28) is a divergence between Q and the IGA-reweighted reference, and both terms vanish exactly at $Q = Q^\star_\lambda$. Therefore, over any generator class, minimizing the IGA objective is equivalent to matching the reweighted law. For a diffusion model, this equivalence justifies *weighted denoising score matching*, i.e., the standard training loss with clean samples drawn according to the weights $q^\star$ in place of uniform weights, which coincides with diffusion training under the data law $Q^\star_\lambda$. When the variational bound is tight and the generator class is expressive, the minimizers of the weighted bound attain the IGA optimum. In general, the weighted bound controls the data-first divergence $\mathrm{KL}(Q^\star_\lambda \| Q_\vartheta)$, whereas the IGA excess (28) is the model-first sum; this is the standard mass-covering versus mode-seeking asymmetry, stated here in an exact form (Corollary 3, Appendix G, which also records the empirical-versus-population role of the reference).

The following proves the results of the diffusion-training paragraph of Section 3.3: the finite-sample IGA reweighting of the data (Corollary 2), the exact decomposition of Proposition 10, its consequence for weighted denoising training (Corollary 3), and the relation to distributionally robust optimization (Remark 17).

Corollary 2 (Finite-sample IGA reweighting)**.** *Let $x_1, \dots, x_n$ be the training samples and $\widehat{P}_n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$. For every $\lambda \geq 0$ and $\varepsilon \in (0,1)$, the objective of* (12) *with reference $\widehat{P}_n$ attains a finite minimum over $\{Q \in \mathcal{P} : Q \ll \widehat{P}_n\}$, and its unique minimizer $Q^\star_\lambda = \sum_{i=1}^n q_i^\star \delta_{x_i}$ has strictly positive weights given by the self-consistent softmax* (27)*.*

Proof. The map $q \mapsto Q_q = \sum_i q_i \delta_{x_i}$ identifies $\{Q \in \mathcal{P} : Q \ll \widehat{P}_n\}$ with the simplex $\Delta_n = \{q \in \mathbb{R}^n : q \geq 0,\ \sum_i q_i = 1\}$; if some training points coincide, the identification merges the corresponding atoms and the argument below is unchanged. On Δ_n each term of the objective is finite and continuous: $\mathrm{KL}(Q_q \| \widehat{P}_n) = \sum_i q_i \log(n q_i)$, continuous with the convention $0 \log 0 = 0$; and $H_\varepsilon(Q_q)$, the composition of the matrix entropy H, continuous on density matrices, with the affine map $q \mapsto (1-\varepsilon)\sum_i q_i \phi(x_i)\phi(x_i)^\top + \frac{\varepsilon}{d} I_d$. A continuous function on the compact set Δ_n attains its minimum, which is finite, so the hypothesis of Proposition 2 holds with $\widehat{P}_n$ in the role of the reference. That proposition gives uniqueness, mutual absolute continuity (equivalently, $q_i^\star > 0$ for every i), and the tilt (14), which on atoms reads $n q_i^\star = w_i / Z$ with $Z = \frac{1}{n}\sum_j w_j$ and with w_i as in (27); cancelling the factors of n gives (27). □

Two roles of the reference. The corollary solves the IGA problem anchored at the *empirical* training distribution exactly; the resulting $Q^\star_\lambda$ is supported on the training set and is the implicit data law of the weighted training scheme below. When the generated law Q_ϑ of a continuously supported model is compared against a reference, $\mathrm{KL}(Q_\vartheta \| \widehat{P}_n) = +\infty$, so the decomposition of Proposition 10 is applied with a population reference, such as P_{data} or a smooth teacher law in fine-tuning, for which $F(Q_\vartheta)$ is finite for absolutely continuous models. The empirical weights (27) are then the plug-in counterpart of the population tilt: the exponent is a fixed continuous function of x once the covariance $S^\varepsilon_{Q^\star_\lambda}$ is given, and the empirical fixed point estimates exactly this covariance. We do not pursue a finite-sample analysis of this plug-in step here; Theorem 5 describes the behavior of the underlying moment estimates.

Proof of Proposition 10. Write $Q^\star = Q^\star_\lambda$ and let

$$T(x) := \lambda\, G^\varepsilon_{Q^\star}(x)$$

be the total reward (13), with P_{ref} in the role of the reference, so that $dQ^\star/dP_{\mathrm{ref}} = e^T/Z$ with $Z = \mathbb{E}_{P_{\mathrm{ref}}}[e^{T(X)}]$ by (14). Since the eigenvalues of S^ε_Q lie in $[\varepsilon/d, 1]$, the exponent is uniformly bounded: $0 \leq T \leq \lambda(1-\varepsilon)\log(d/\varepsilon)$. Fix Q with $F(Q) < \infty$. Because $0 \leq H_\varepsilon \leq \log d$, finiteness of $F(Q)$ is equivalent to $\mathrm{KL}(Q\|P_{\mathrm{ref}}) < \infty$, and in particular $Q \ll P_{\mathrm{ref}}$.

Step 1: chain rule for the KL term. The ratio $dQ^\star/dP_{\mathrm{ref}} = e^T/Z$ is bounded above and below by positive constants, so $Q \ll P_{\mathrm{ref}}$ if and only if $Q \ll Q^\star$, and P_{ref}-almost surely

$$\log \frac{dQ}{dP_{\mathrm{ref}}} = \log \frac{dQ}{dQ^\star} + T - \log Z.$$

The negative part of $\log(dQ/dP_{\mathrm{ref}})$ is Q-integrable (as always for a log-density ratio: $t(\log t)_- \leq e^{-1}$ for $t \geq 0$), and $T - \log Z$ is bounded, hence Q-integrable; therefore all three expectations below are well defined in $(-\infty, +\infty]$ and additivity holds:

$$\mathrm{KL}(Q\|P_{\mathrm{ref}}) = \mathrm{KL}(Q\|Q^\star) + \mathbb{E}_Q[T] - \log Z, \tag{29}$$

with $\mathrm{KL}(Q\|Q^\star) < \infty$ exactly when $\mathrm{KL}(Q\|P_{\mathrm{ref}}) < \infty$.

Step 2: the entropy Bregman term is nonnegative. Let $\nu = Q - Q^\star$, a finite signed measure with $\nu(\mathcal{X}) = 0$. For $t \in [0,1]$, $Q^\star + t\nu = (1-t)Q^\star + tQ \in \mathcal{P}$, and $h(t) := H_\varepsilon(Q^\star + t\nu)$ is concave on $[0,1]$ (Lemma 2, through the affine map $Q \mapsto \Sigma_Q$) with right derivative $h'(0^+) = \int G^\varepsilon_{Q^\star}\, d\nu$ (Lemma 1). Concavity places $h(1)$ below the tangent at 0:

$$H_\varepsilon(Q) \leq H_\varepsilon(Q^\star) + \int G^\varepsilon_{Q^\star}\, d(Q - Q^\star),$$

which is exactly $B_{-H_\varepsilon}(Q, Q^\star) \geq 0$, and by the definition of B_{-H_ε},

$$-\lambda H_\varepsilon(Q) = -\lambda H_\varepsilon(Q^\star) - \lambda \int G^\varepsilon_{Q^\star}\, d(Q - Q^\star) + \lambda\, B_{-H_\varepsilon}(Q, Q^\star). \tag{30}$$

Step 3: assembly. Summing (29) and (30), and substituting $\mathbb{E}_Q[T] = \lambda \mathbb{E}_Q[G^\varepsilon_{Q^\star}]$,

$$F(Q) = \mathrm{KL}(Q\|Q^\star) + \lambda\, B_{-H_\varepsilon}(Q, Q^\star) + C,$$

where the $\mathbb{E}_Q[G^\varepsilon_{Q^\star}]$ terms cancel against the integral in (30), leaving $\lambda \mathbb{E}_{Q^\star}[G^\varepsilon_{Q^\star}]$, and

$$C = \lambda \mathbb{E}_{Q^\star}[G^\varepsilon_{Q^\star}] - \log Z - \lambda H_\varepsilon(Q^\star).$$

Finally, applying (29) at $Q = Q^\star$ gives

$$\mathrm{KL}(Q^\star\|P_{\mathrm{ref}}) = \lambda \mathbb{E}_{Q^\star}[G^\varepsilon_{Q^\star}] - \log Z,$$

so $C = \mathrm{KL}(Q^\star\|P_{\mathrm{ref}}) - \lambda H_\varepsilon(Q^\star) = F(Q^\star)$, which is (28). □

Corollary 3 (Weighted diffusion training)**.** *Let $q^\star$ be the weights of Corollary 2 and let $\ell(x;\vartheta)$ be any per-example diffusion training loss (a denoising-score-matching loss or a negative variational bound). Then:*

(i) *The weighted objective $\sum_{i=1}^n q_i^\star \ell(x_i;\vartheta) = \mathbb{E}_{X\sim Q_\lambda^\star}[\ell(X;\vartheta)]$ coincides with the corresponding standard training objective with data law $Q_\lambda^\star$; no other component of the training pipeline changes.*

(ii) *Let P_{ref} and $Q_\lambda^\star$ be as in Proposition 10. A law Q with $F(Q) < \infty$ attains $\min_{Q' \ll P_{\mathrm{ref}}} F$ if and only if $Q = Q_\lambda^\star$; and for any sequence (ϑ_k) with $\mathrm{KL}(Q_{\vartheta_k}\|Q_\lambda^\star) \to 0$, $F(Q_{\vartheta_k}) \to \min_{Q' \ll P_{\mathrm{ref}}} F(Q')$.*

Proof. Part (i) is the definition of expectation under a finitely supported law. For part (ii), the "only if" direction: if $F(Q) = F(Q_\lambda^\star) < \infty$, then by (28) the two nonnegative terms vanish, in particular $\mathrm{KL}(Q\|Q_\lambda^\star) = 0$, so $Q = Q_\lambda^\star$; the converse is trivial. For the convergence claim, write $Q_k = Q_{\vartheta_k}$ and $\delta_k = \sup_A |Q_k(A) - Q_\lambda^\star(A)|$, so that $\left|\int f\, d(Q_k - Q_\lambda^\star)\right| \leq 2\,\|f\|_\infty\, \delta_k$ for bounded measurable f, and $\delta_k \leq \sqrt{\mathrm{KL}(Q_k\|Q_\lambda^\star)/2} \to 0$ by Pinsker's inequality. By (28) it suffices that each right-hand term vanishes along the sequence. The KL term does by hypothesis. For the Bregman term, each entry of $\Sigma_{Q_k} - \Sigma_{Q_\lambda^\star}$ is $\int \phi_a \phi_b\, d(Q_k - Q_\lambda^\star)$ with $|\phi_a \phi_b| \leq 1$, so $\Sigma_{Q_k} \to \Sigma_{Q_\lambda^\star}$, hence $S^\varepsilon_{Q_k} \to S^\varepsilon_{Q_\lambda^\star}$ within the compact set of density matrices with spectrum in $[\varepsilon/d, 1]$, on which H is continuous, giving $H_\varepsilon(Q_k) \to H_\varepsilon(Q_\lambda^\star)$; and $\left|\int G^\varepsilon_{Q_\lambda^\star}\, d(Q_k - Q_\lambda^\star)\right| \leq 2(1-\varepsilon)\log(d/\varepsilon)\,\delta_k \to 0$. Hence $B_{-H_\varepsilon}(Q_k, Q_\lambda^\star) \to 0$, completing the proof. □

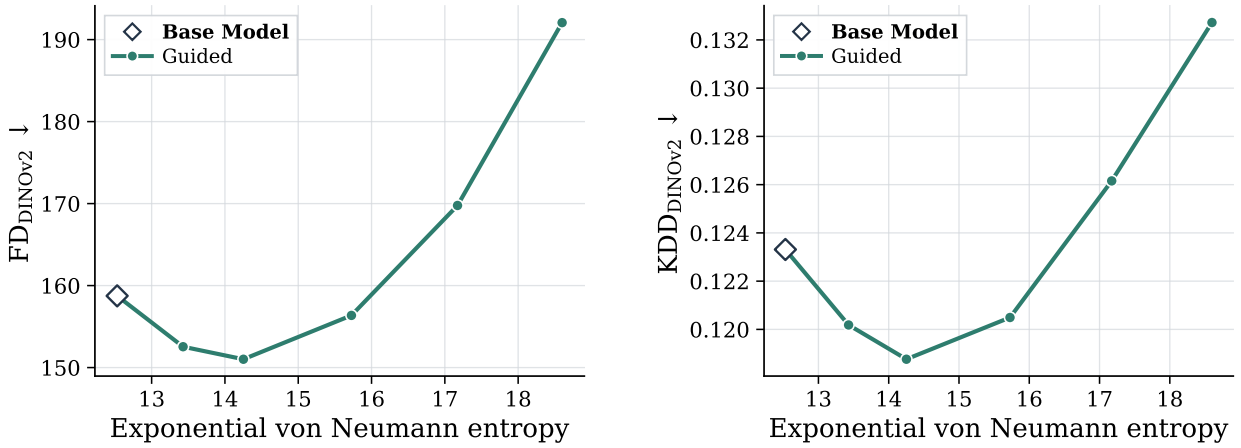
Remark 17 (Saddle computation of the weights, and the relation to DRO)**.** (a) Computation. *By Proposition 1, on Δ_n the weights of Corollary 2 solve the finite-dimensional saddle problem*

$$\min_{q\in\Delta_n} \max_{\Theta\in\mathbb{T}_\varepsilon} \Big\{ \sum_i q_i \log(n q_i) - \lambda(1-\varepsilon) \sum_i q_i\, \phi(x_i)^\top \Theta\, \phi(x_i) - \lambda \log \mathrm{Tr}\big(e^{-\Theta}\big) \Big\},$$

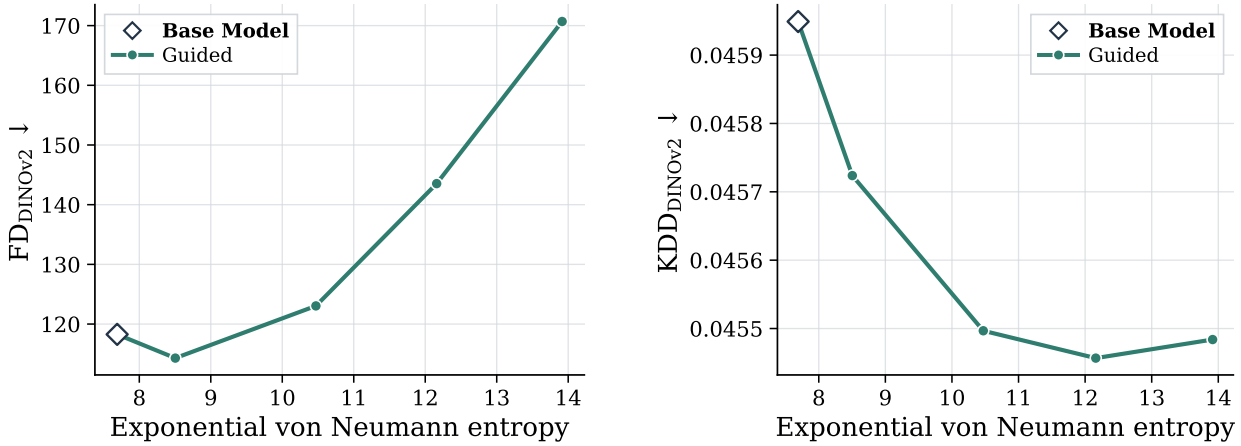
whose objective is convex and continuous in q on the compact Δ_n and concave and continuous in Θ on the compact $\mathbb{T}_\varepsilon$; by Sion's minimax theorem the order of optimization may be interchanged, and both optima are attained. Alternating best responses are natural: at fixed q the inner maximum is the closed form $\Theta^\star(Q_q)$ of Proposition 1, while at fixed Θ the outer minimization is, by Lemma 4 with reward $G(x) = \lambda(1-\varepsilon)\phi(x)^\top \Theta\, \phi(x)$, the explicit softmax $q_i \propto \exp\big(\lambda(1-\varepsilon)\phi(x_i)^\top \Theta\, \phi(x_i)\big)$.

(b) Not worst-case-loss DRO. *It is instructive to contrast* (27) *with KL-penalized distributionally robust training of the denoiser,*

$$\min_{\vartheta} \max_{Q \ll \widehat{P}_n} \big\{ \mathbb{E}_Q[\ell(X;\vartheta)] - \eta\, \mathrm{KL}(Q\|\widehat{P}_n) \big\}, \qquad \eta > 0.$$



(a) CelebA-HQ: Fréchet distance (FD). **(b)** CelebA-HQ: kernel distance (KD).



(c) ImageNet: Fréchet distance (FD). **(d)** ImageNet: kernel distance (KD).

Figure 19: **DINOv2-space distributional distances along the IGA target path.** Fréchet distance (FD) and kernel distance (KD) are evaluated in the DINOv2 representation used for the spectral analysis. The initial portion of the IGA path reduces the diversity deficit while improving or maintaining distributional agreement with the data. As λ increases further, the distance minima occur at metric- and dataset-dependent operating points, consistent with the transition from below-wall diversity repair to beyond-wall spectral extrapolation.

There, by Lemma 4, the inner maximizer reweights the data by the loss, $q_i \propto \exp(\ell(x_i;\vartheta)/\eta)$*: the adversarial reweighting tracks loss hardness and changes with ϑ at every step. The IGA reweighting* (27) *is ϑ-independent and tilts by the* entropy energy*: it up-weights points along spectral directions the data underpopulates, whether or not the current model finds them hard. Comparing the two Gibbs exponents, the schemes produce the same weights only when $\ell(\cdot;\vartheta)$ is, on the training set, an affine function of the IGA exponent, which is a nongeneric coincidence. Replacing the denoiser's training distribution by a worst-case-loss adversary therefore optimizes robustness, not spectral diversity, and is not equivalent to IGA training; the rigorous route to training-time IGA for diffusion models is the reweighting-and-refit composition of Corollaries 2 and 3.*

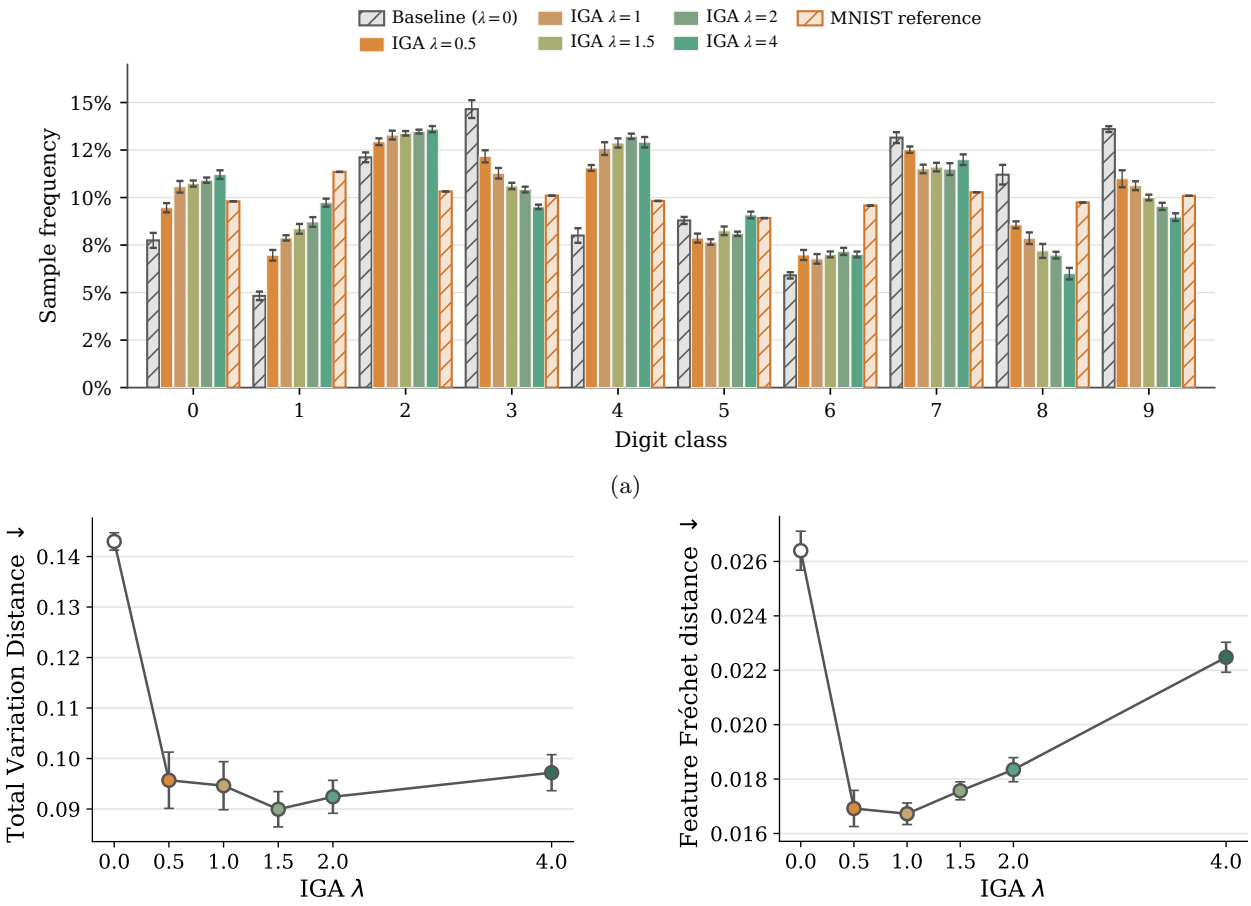


Figure 20: **Training-time IGA reduces class imbalance in an MNIST GAN. (a)** Generated digit frequencies move toward the empirical MNIST distribution as the IGA multiplier increases. **(b,c)** Moderate regularization reduces both class-distribution total variation and independent evaluator-feature Fréchet distance. Results are averaged over five seeds; error bars denote standard error.

H Additional Numerical Results

This appendix collects supplementary numerical results for the sampling-time and training-time experiments. Section H.1 reports Dinov2-space distributional distances for the CelebA-HQ and ImageNet experiments, complementing the independent Inception-v3 evaluation in the main text. Section H.2 reports the MNIST GAN results for the training-time realization of IGA discussed in Section 6.3.

H.1 DINOv2-Space Distributional Distances

The main text reports FID and KID in Inception-v3 feature space as an evaluation independent of the DINOv2 representation used to define the spectral entropy and entropy wall. Here we provide the corresponding DINOv2-space distributional distances. These results complement the independent Inception-v3 evaluation and make explicit the representation dependence of the precise fidelity optimum along the IGA regularization path.

H.2 Training-Time IGA on MNIST

The training-time experiment of Section 6.3 evaluates whether the same spectral-entropy regularizer used for sampling-time IGA can be incorporated directly into adversarial training. Figure 20 reports the resulting class-frequency and distributional-fidelity measurements across the IGA path.

Experimental details. We train a convolutional GAN on MNIST for 20 epochs. The generator maps a 64-dimensional standard-normal latent through a fully connected projection and two transposed-convolution stages to a 28×28 image, while the discriminator uses two strided convolutional layers followed by a linear output. We use batch size 128 and Adam with learning rate 2×10^{-4} and $(\beta_1, \beta_2) = (0.5, 0.999)$ for both networks, with the non-saturating logistic generator objective and one discriminator update per generator update. We evaluate $\lambda \in \{0, 0.5, 1, 1.5, 2, 4\}$ over five random seeds.

The fixed IGA representation is the unit-normalized 64-dimensional embedding of a separately trained MNIST classifier. At each generator update, the spectral adversary is recomputed from the current generated minibatch at its closed-form best response and then held fixed during the generator update, as in Proposition 7. We set $\varepsilon = 0.05$. For independent evaluation, we use a second frozen classifier with a different architecture and a 96-dimensional embedding. For each run, 10,000 generated samples are used to compute class frequencies, total variation distance to the empirical MNIST test-set class distribution, and feature Fréchet distance between generated and real test samples in the independent evaluator space.

As shown in Figure 20(a), increasing λ initially redistributes generated mass away from overrepresented classes and toward classes that are underrepresented by the baseline GAN. This redistribution is reflected quantitatively in panels (b) and (c): both class-distribution total variation and independent evaluator-feature Fréchet distance improve substantially at intermediate values of λ. Their nonmonotone behavior at larger λ illustrates the tradeoff between the GAN fidelity objective and the distribution-level entropy reward.