RECONSTRUCTION DOES NOT DETERMINE GENERATIVE LEARNABILITY: A THEORY OF RESOURCE-MATCHED VISUAL TOKENIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

A visual tokenizer is typically optimized for reconstruction, while the downstream generator must learn the latent distribution it induces. We show that better reconstruction need not imply easier generation because reconstruction does not identify the generative geometry of the latent space: reconstruction-equivalent tokenizers can induce sharply different finite-resource learning problems. To quantify this gap, we develop Rate–Distortion–Generative Learnability (RDGL), supported by diffusion-scale representation complexity, latent-to-image transport bounds, finite-sample minimax limits, a full-spectrum deterministic equivalent from random matrix theory, and separate capacity and optimization laws. The resulting theory shows that the preferred representation is resource dependent and motivates Capacity-Matched Spectral Tokenization (CMST), which shapes latent geometry to the downstream data, model-capacity, and optimization budgets. Experiments support both the mechanism and the resulting design principle: reconstruction-preserving latent gauges raise fixed-budget gFID from 4.18 to 15.61, while CMST reduces gFID from 4.18 for the matched VA-VAE control to 2.96 under the same downstream protocol. The resulting principle is that a representation's complexity and geometry should be matched to the learner that must model it, rather than optimized for reconstruction alone. The anonymized code is available at `https://anonymous.4open.science/r/ICLR2027-CMST/`.

1 INTRODUCTION

Two-stage image generators first encode images into visual representations and then learn a generative model in the induced latent space (Rombach et al., 2022). Tokenizers are therefore often judged by reconstruction fidelity, suggesting a natural intuition: preserving more visual information should not make downstream generation harder. Recent evidence contradicts this intuition. Richer or wider representations can reconstruct better yet require larger generators, longer optimization, or even yield worse generation under a fixed budget, across both continuous diffusion and discrete autoregressive pipelines (Yao et al., 2025; Ramanujan et al., 2025; Xiong et al., 2025). Figure 1 summarizes this tension. The recurring mismatch raises a more basic question: *what property of a representation is reconstruction failing to control?*

Recent work has improved generative representations through structured or semantic latents, spectral and frequency regularization, equivariance, and explicit learnability diagnostics (Chen et al., 2025a; Skorokhodov et al., 2025; Kouzelis et al., 2025; Chen et al., 2026; Yao et al., 2026; Zhao et al., 2026). These results establish that latent organization matters, but leave a common theoretical question unresolved: *what does reconstruction leave unidentified, and how does that freedom translate into resource-dependent generative difficulty?* Reconstruction constrains only the composite map $D \circ E$, whereas the stage-2 learner must model $Z = E(X)$. Consequently, invertible latent reparameterizations can leave every reconstruction unchanged while altering the covariance, noise geometry, conditioning, and ultimately the difficulty of learning the latent distribution. The reconstruction–generation mismatch is therefore fundamentally an *identifiability problem*, rather than merely insufficient reconstruction fidelity.

We develop a theory of *Rate–Distortion–Generative Learnability (RDGL)* around this distinction. For deterministic continuous representations, diffusion noise scale provides an operational reso-

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

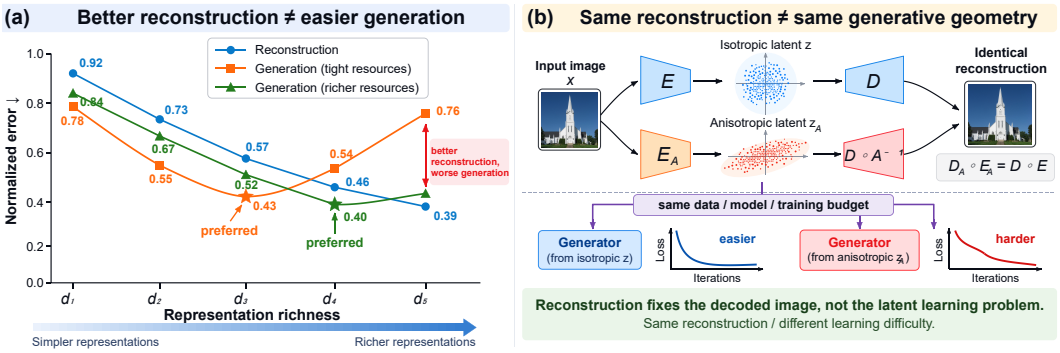


Figure 1: **Reconstruction does not determine generative learnability.** (a) Increasing representation richness can improve reconstruction while making fixed-budget generation harder; with more downstream resources, the preferred operating point shifts toward richer representations. (b) Reconstruction-equivalent latent reparameterizations can induce different generative geometries and therefore different learning difficulty under the same downstream budget.

lution and hence a finite notion of representation complexity. We formalize the reparameterization freedom as a reconstruction gauge and use transport bounds to connect latent modeling error to image-space generation. Under finite data, we derive minimax limits and a full-spectrum deterministic equivalent from random matrix theory, including a low-noise sample-to-representation transition; finite model capacity and finite optimization introduce distinct spectral floors and transients. Together, these laws imply that the preferred representation is resource dependent, and that latent directions should be allocated according to reconstruction value relative to their finite-resource generative burden. This principle leads to *Capacity-Matched Spectral Tokenization (CMST)*, which canonicalizes latent scale and shapes the spectrum using a decoder-calibrated generative tax matched to a specified downstream learner.

We test the theory at three levels. Gauge transformations paired with exact inverse decoders preserve reconstruction while raising fixed-budget gFID from 4.18 to 15.61 as the latent geometry becomes increasingly anisotropic. Independent sweeps over data, model capacity, and optimization budget show that the preferred representation shifts toward richer latents as downstream resources increase. Finally, under a common LightningDiT-XL/1 80-epoch protocol, CMST reduces gFID from 4.18 for the matched VA-VAE control to 2.96, while learning a resource-dependent spectral structure rather than generic isotropy.

Our main contributions are as follows. (1) **A foundation for generative learnability beyond reconstruction.** We introduce diffusion-scale complexity for continuous representations, establish reconstruction-gauge non-identifiability, and connect latent modeling risk to image-space generation through transport bounds. (2) **A finite-resource theory of representation preference.** We derive finite-sample minimax laws, full-spectrum RMT score-learning asymptotics, and separate capacity and optimization mechanisms, yielding resource-dependent RDGL and decoder-aware representation allocation. (3) **A resource-matched tokenizer and controlled validation.** We derive CMST from the finite-resource laws and validate the predicted gauge sensitivity, resource-dependent representation shifts, and fixed-budget generation gains in controlled tokenizer–DiT experiments.

2 Related Work

Generative visual tokenizers and the reconstruction–generation gap. Recent work shows that reconstruction fidelity alone is insufficient for visual tokenization (Lai et al., 2026). VA-VAE identifies an optimization dilemma in which richer continuous latents improve reconstruction but can make downstream diffusion harder to train (Yao et al., 2025). A broader line of work improves generative latents through spectral or frequency structure, equivariance, semantic alignment, and compositionality, including Improving Diffusability, EQ-VAE, AlignTok, VTP, and CompTok (Skorokhodov et al., 2025; Kouzelis et al., 2025; Chen et al., 2026; Yao et al., 2026; Zhao et al., 2026). These results establish that generative usefulness depends on latent organization beyond reconstruction quality; we study this dependence through a finite-resource learning framework spanning data, model capacity, and optimization.

Compression, rate, and resource-dependent tokenization. For discrete autoregressive generation, Ramanujan et al. (2025) connect stage-1 rate–distortion to stage-2 model capacity, while GigaTok studies the reconstruction–generation trade-off induced by tokenizer scaling (Xiong et al., 2025). Recent work further analyzes conditional-entropy collapse and develops unified rate–distortion views of visual quantization (Zheng et al., 2026a; Fang et al., 2026). These studies focus on discrete coding rates, codebooks, and autoregressive predictability. Our setting is instead deterministic continuous tokenization, where rate is defined relative to diffusion noise scale and the central object is finite-resource learning of the induced latent distribution.

Theory of generative learnability and high-dimensional diffusion. MAETok relates diffusion sample complexity to structured Gaussian-mixture latents (Chen et al., 2025a), while Representation Autoencoders identify a capacity obstruction for high-dimensional latent denoising (Zheng et al., 2026b). In parallel, Wang et al. (2026) use random matrix theory to characterize finite-data effects in Gaussian diffusion and validate the resulting spectral predictions on nonlinear diffusion models. We adopt a related tractable Gaussian/RMT viewpoint, but study how representation dimension and spectrum interact jointly with sample size, model capacity, and optimization time. Combined with image-space transport and decoder geometry, these laws lead directly to a resource-matched tokenizer design principle.

3 Problem Setup and Diffusion-Scale Complexity

We study two-stage visual generative models in which a tokenizer $\mathcal{T} = (E, D)$ maps $X \sim P_X$ in $\mathcal{X} \subseteq \mathbb{R}^p$ to a continuous latent $Z = E(X) \in \mathbb{R}^d$ with law $P_Z = E_{\#}P_X$, and a downstream generator learns a latent distribution Q. The resulting image distribution is $P_{\mathrm{gen}} = D_{\#}Q$. We measure tokenizer distortion and downstream modeling error by

$$\delta_{\mathrm{rec}}(\mathcal{T}) = W_2(P_X, D_{\#}P_Z),\ \mathcal{L}_{\mathrm{lat}}(Q; P_Z) = W_2^2(P_Z, Q),\ \mathcal{L}_{\mathrm{img}}(Q; \mathcal{T}) = W_2^2(P_X, D_{\#}Q). \quad (1)$$

Thus δ_{rec} is the image-space discrepancy remaining under exact latent modeling; whenever W_2 is used, distributions have finite second moments (Appendix A summarizes notation).

To make downstream difficulty explicit under finite resources, let $S_n = (Z_1, \ldots, Z_n) \sim P_Z^{\otimes n}$, let $\mathcal{G}_C$ be a generator family with capacity budget C, and let $\mathfrak{A}_{n,C,T}$ denote learning procedures admissible under sample, capacity, and optimization budgets (n, C, T). For a latent distribution class $\mathcal{P} \subseteq \mathcal{P}_2(\mathbb{R}^d)$, define

$$\mathfrak{G}_{n,C,T}(\mathcal{P}) = \inf_{\mathcal{A} \in \mathfrak{A}_{n,C,T}} \sup_{P \in \mathcal{P}} \mathbb{E}_{S_n \sim P^{\otimes n}} \left[W_2^2(P, Q_{\mathcal{A},S_n}) \right]. \quad (2)$$

The three resource axes isolate statistical, capacity, and optimization bottlenecks in the downstream learner.

For deterministic continuous tokenizers, raw mutual information is generally not a finite representation rate: if P_Z is non-atomic, then $I(X; Z) = +\infty$. Diffusion instead supplies an operational resolution through its noise scale. For $\tau > 0$, let $\varepsilon \sim \mathcal{N}(0, I_d)$ be independent of Z, set $Z_\tau = Z + \sqrt{\tau}\,\varepsilon$, and define the *diffusion-scale representation rate* $R_\tau(Z) := I(Z; Z_\tau)$. For $Z \sim \mathcal{N}(\mu, \Sigma)$ with covariance eigenvalues $\lambda_1, \ldots, \lambda_d$,

$$R_\tau(Z) = \frac{1}{2} \sum_{i=1}^{d} \log\left(1 + \frac{\lambda_i}{\tau}\right), \qquad d_{\mathrm{eff}}(\tau) := 2\frac{\partial R_\tau}{\partial \log(1/\tau)} = \sum_{i=1}^{d} \frac{\lambda_i}{\lambda_i + \tau}. \quad (3)$$

Each mode therefore contributes the soft activity $\lambda_i/(\lambda_i + \tau)$: the complexity exposed to a diffusion learner is set by the latent spectrum relative to the current noise scale, rather than by raw dimension alone. Gaussian forward processes reduce to this additive-noise form after rescaling. Appendix B gives the measure-theoretic justification for the continuous-rate construction, its distribution-free I–MMSE extension, and decision-theoretic details of Eq. 2.

4 Reconstruction Does Not Identify Generative Geometry

A reconstruction objective observes only the composite map $D \circ E$, whereas the downstream generator must model the internal coordinates $Z = E(X)$. This creates a latent reparameterization freedom that reconstruction alone cannot resolve. For any measurable bijection $\phi : \mathbb{R}^d \to \mathbb{R}^d$ with measurable inverse, we define $E_\phi = \phi \circ E$ and $D_\phi = D \circ \phi^{-1}$.

Theorem 1 (Reconstruction gauge non-identifiability). *For every such ϕ, $D_\phi(E_\phi(x)) = D(E(x))$ for all $x \in \mathcal{X}$; hence every reconstruction criterion of the form $\mathbb{E}[\ell(X, D(E(X)))]$ is invariant under $(E, D) \mapsto (E_\phi, D_\phi)$.*

For a linear gauge $\phi(z) = Az$, $A \in \mathrm{GL}(d)$, the corresponding isotropic corruption, pulled back to the original coordinates, has noise covariance $\tau A^{-1}A^{-\top}$ and is therefore preserved at the same noise scale only when A is orthogonal. In particular, let $Z \sim \mathcal{N}(0, I_2)$ and $A_a = \mathrm{diag}(a, a^{-1})$, $a \geq 1$. The gauged tokenizers have identical reconstruction and latent dimension, while for every fixed $\tau > 0$,

$$R_\tau(A_a Z) = \frac{1}{2}\log\left[\left(1 + \frac{a^2}{\tau}\right)\left(1 + \frac{a^{-2}}{\tau}\right)\right] \longrightarrow \infty, \qquad \kappa_\tau(a) = \frac{a^2 + \tau}{a^{-2} + \tau} \longrightarrow \infty. \tag{4}$$

Here $\kappa_\tau(a)$ is the condition number of the noised Gaussian score Jacobian. Consequently, neither diffusion-scale rate nor score conditioning is determined by the reconstruction map $D \circ E$ and latent dimension d alone. See Figure 5 in Appendix C.1 for an illustration.

To transfer downstream latent error back to image space, let $\mathcal{S}$ contain the supports of P_Z and Q, and assume that $D|_{\mathcal{S}}$ is bi-Lipschitz: $m_D\|z - z'\|_2 \leq \|D(z) - D(z')\|_2 \leq L_D\|z - z'\|_2$ for all $z, z' \in \mathcal{S}$, with $0 < m_D \leq L_D < \infty$.

Theorem 2 (Transport sandwich). *Every latent distribution Q satisfying the assumptions above satisfies*

$$\Big(m_D W_2(P_Z, Q) - \delta_{\mathrm{rec}}\Big)_+ \leq W_2(P_X, D_\# Q) \leq \delta_{\mathrm{rec}} + L_D W_2(P_Z, Q). \tag{5}$$

Thus, under controlled decoder geometry, latent modeling error necessarily propagates to image space up to the reconstruction discrepancy. The corresponding squared-risk bounds, proofs, and decoder-geometric interpretation are given in Appendix C.

5 Finite-Resource Generative Learnability

5.1 Finite-Sample Limits and Reconstruction–Generation Reversal

We first isolate the statistical bottleneck by removing capacity and optimization restrictions. For fixed $0 < \rho < 1$, let $\mathcal{C}_d(\rho) = \{\Sigma : (1 - \rho)I_d \preceq \Sigma \preceq (1 + \rho)I_d\}$ and $P_\Sigma = \mathcal{N}(0, \Sigma)$. Given $S_n = (Z_1, \ldots, Z_n) \sim P_\Sigma^{\otimes n}$, define

$$\mathfrak{R}_{n,d}(\rho) := \inf_{\widehat{Q}} \sup_{\Sigma \in \mathcal{C}_d(\rho)} \mathbb{E}_\Sigma\left[W_2^2\Big(P_\Sigma, \widehat{Q}(S_n)\Big)\right], \tag{6}$$

where $\widehat{Q}$ may take values in the full space $\mathcal{P}_2(\mathbb{R}^d)$.

Theorem 3 (Finite-sample law and reconstruction–generation reversal). *For every fixed $0 < \rho < 1$, there exist constants $0 < c_\rho \leq C_\rho < \infty$, independent of n and d, such that*

$$c_\rho\, \psi_n(d) \leq \mathfrak{R}_{n,d}(\rho) \leq C_\rho\, \psi_n(d), \qquad \psi_n(d) := \min\left\{d, \frac{d^2}{n}\right\}. \tag{7}$$

Moreover, suppose a d-dimensional tokenizer family induces $P_{Z,\Sigma} = P_\Sigma$ for $\Sigma \in \mathcal{C}_d(\rho)$, and define $\delta_d := \sup_{\Sigma \in \mathcal{C}_d(\rho)} \delta_{\mathrm{rec}}(\Sigma)$. Assume that the corresponding decoders are uniformly bi-Lipschitz on the relevant supports with constants $0 < m_d \leq L_d < \infty$, and let $\mathfrak{R}^{\mathrm{img}}_{n,d}$ denote the corresponding minimax image-space risk. Then

$$\frac{c_\rho m_d^2}{2}\, \psi_n(d) - \delta_d^2 \leq \mathfrak{R}^{\mathrm{img}}_{n,d} \leq 2\delta_d^2 + 2C_\rho L_d^2\, \psi_n(d). \tag{8}$$

Consequently, for two representation families with $d_1 > d_0$ and $\delta_{d_1} < \delta_{d_0}$, if the lower endpoint of Eq. 8 at d_1 exceeds the upper endpoint at d_0, then $\mathfrak{R}^{\mathrm{img}}_{n,d_1} > \mathfrak{R}^{\mathrm{img}}_{n,d_0}$.

The d^2/n regime is the price of estimating second-order latent structure, rather than a "one parameter per coordinate" law. Because Eq. 6 allows arbitrary distribution-valued estimators, the lower bound is not tied to a particular generator architecture or optimizer. Together with the strict reconstruction ordering above, the theorem therefore gives a finite-sample reconstruction–generation reversal. Appendix D gives the Gaussian Bures–Wasserstein geometry, Fano packing lower bound, matching projected-covariance upper bound, and the explicit sufficient reversal condition.

5.2 High-Dimensional Score-Learning Transition

The minimax law above gives a worst-case statistical limit. We next resolve how finite data interacts with the full latent spectrum and diffusion noise scale. Let $Z \sim \mathcal{N}(0, \Sigma_d)$ and $Z_1, \dots, Z_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Sigma_d)$, with $\widehat{\Sigma} = n^{-1} \sum_{i=1}^n Z_i Z_i^\top$. For $Y_\tau = Z + \sqrt{\tau}\varepsilon$, the population and plug-in Gaussian scores are $s_\tau^\star(y) = -(\Sigma_d + \tau I)^{-1} y$ and $\widehat{s}_\tau(y) = -(\widehat{\Sigma} + \tau I)^{-1} y$. Define the normalized excess score risk on an independent test sample by

$$\Phi_d(\tau) := \frac{1}{d} \mathbb{E}\left[\|\widehat{s}_\tau(Y_\tau) - s_\tau^\star(Y_\tau)\|_2^2 \,\middle|\, Z_1, \dots, Z_n \right]. \tag{9}$$

Assume $d/n \to \gamma \in (0, \infty)$, $\sup_d \|\Sigma_d\|_{\text{op}} < \infty$, and that the empirical spectral distribution $H_d = d^{-1} \sum_{i=1}^d \delta_{\lambda_i(\Sigma_d)}$ converges weakly to a compactly supported probability measure H.

Theorem 4 (Spectral deterministic equivalent for score learning)**.** *For every fixed $\tau > 0$, $\Phi_d(\tau)$ converges almost surely to*

$$\Phi(\gamma, \tau, H) = -a'(\tau) - \tau m'(\tau) - 2m(\tau) + h_H(\tau), \qquad h_H(\tau) := \int \frac{dH(\lambda)}{\lambda + \tau}, \tag{10}$$

where primes denote derivatives with respect to τ, and $m(\tau) > 0$ is the unique solution of

$$m(\tau) = \int \frac{dH(\lambda)}{\tau + \chi(\tau)\lambda}, \qquad \chi(\tau) = 1 - \gamma + \gamma\tau m(\tau), \qquad a(\tau) = \int \frac{\lambda\, dH(\lambda)}{\tau + \chi(\tau)\lambda}. \tag{11}$$

Thus the finite-sample score error is a computable functional of the sample-to-representation ratio γ, noise scale τ, and the full limiting spectrum H, rather than of raw latent dimension alone.

Corollary 5 (Low-noise sample-to-representation transition)**.** *For isotropic latents $H = \delta_1$, after the proportional high-dimensional limit, the low-noise score risk satisfies*

$$\Phi_{\text{iso}}(\gamma, \tau) \sim \begin{cases} \dfrac{\gamma(1 + \gamma - \gamma^2)}{(1-\gamma)^3}, & 0 < \gamma < 1, \quad \tau \downarrow 0, \\ \dfrac{1}{2\tau^{3/2}}, & \gamma = 1, \quad \tau \downarrow 0, \\ \dfrac{1 - \gamma^{-1}}{\tau^2}, & \gamma > 1, \quad \tau \downarrow 0. \end{cases} \tag{12}$$

Thus the subcritical score risk diverges as $(1-\gamma)^{-3}$ as $\gamma \uparrow 1$, while for $\gamma > 1$, $\tau^2 \Phi_{\text{iso}}(\gamma, \tau) \to 1 - \gamma^{-1}$, the sample-covariance null-space fraction. Finite noise smooths this transition over $|\gamma - 1| = O(\sqrt{\tau})$. Appendix E gives the complete resolvent calculation, denoising-risk conversion, isotropic closed forms, and the critical scaling law.

5.3 Capacity and Optimization Bottlenecks

We next isolate two non-statistical bottlenecks in the population Gaussian denoising problem. Let $Y_\tau = Z + \sqrt{\tau}\varepsilon$ with $Z \sim \mathcal{N}(0, \Sigma)$, define $\mathcal{L}_\tau(B) = \mathbb{E}\|BY_\tau - Z\|_2^2$, and let $B^\star = \Sigma(\Sigma + \tau I)^{-1}$ be the population-optimal linear denoiser. We measure normalized excess risk by $\mathcal{E}_\tau(B) = d^{-1}\big[\mathcal{L}_\tau(B) - \mathcal{L}_\tau(B^\star)\big]$. Write $\lambda_1 \geq \cdots \geq \lambda_d \geq 0$ for the eigenvalues of Σ. We use matrix rank as an analytically tractable proxy for effective denoising capacity.

Theorem 6 (Spectrum-resolved capacity and optimization laws)**.** *If the denoiser is restricted to $\text{rank}(B) \leq r$, then*

$$\Gamma_{\text{cap}}(\Sigma, r, \tau) := \inf_{\text{rank}(B) \leq r} \mathcal{E}_\tau(B) = \frac{1}{d} \sum_{i > r} \frac{\lambda_i^2}{\lambda_i + \tau}, \tag{13}$$

and an optimum retains the r covariance eigenmodes with largest λ_i.

Starting from $B_0 = 0$, consider $B_{t+1} = B_t - \eta[B_t(\Sigma + \tau I) - \Sigma]$. For $0 < \eta < 2/(\lambda_1 + \tau)$,

$$\Gamma_{\text{opt}}(\Sigma, T, \tau; \eta) := \mathcal{E}_\tau(B_T) = \frac{1}{d} \sum_{i=1}^d \frac{\lambda_i^2}{\lambda_i + \tau} \left[1 - \eta(\lambda_i + \tau)\right]^{2T}. \tag{14}$$

Capacity therefore creates a permanent discarded-mode floor, whereas optimization creates a mode-dependent transient. For isotropic latents, $\Gamma_{\mathrm{cap}}(I_d, r, \tau) = (1-r/d)/(1+\tau)$; for general spectra, the optimization time for a fixed-factor reduction scales as $T = \Theta(\kappa_\tau)$ in the ill-conditioned regime, where $\kappa_\tau = (\lambda_1 + \tau)/(\lambda_d + \tau)$. The two terms add exactly only in the oracle top-r population surrogate; proofs, conditioning analysis, and the joint decomposition are given in Appendix F.

6 Rate–Distortion–Generative Learnability and CMST

6.1 Capacity-Matched Representation

We now turn the finite-resource laws above into a representation-design principle. Consider a family $\{\mathcal{T}_d = (E_d, D_d)\}_d$ indexed by representation size d, viewed through a smooth relaxation when derivatives are used. The diffusion-scale information measure remains R_τ from Eq. 3; here d serves as a tractable coordinate along the representation-complexity curve.

Let $\mathcal{D}(d)$ denote reconstruction distortion in a fixed image-space unit and let $\mathcal{G}(d; n, C, T)$ denote downstream generative difficulty calibrated to the same unit. We define the *rate–distortion–generative-learnability* objective

$$\mathcal{J}_{\mathrm{RDGL}}(d; n, C, T) = \mathcal{D}(d) + \alpha\, \mathcal{G}(d; n, C, T), \qquad \alpha > 0. \tag{15}$$

The finite-sample, capacity, and optimization laws in Section 5 constrain $\mathcal{G}$; their raw surrogate risks are not added before this calibration.

For families satisfying $\mathcal{G}(d; n, C, T) = d\,\phi(\gamma, C, T)$ with $\gamma = d/n$, define $h(\gamma, C, T) := \phi + \gamma\,\partial_\gamma\phi = \partial_d\mathcal{G}$. Theorem 3, for example, gives the benchmark statistical dependence $\phi_{\mathrm{stat}}(\gamma) \asymp \min\{1, \gamma\}$. The smooth analysis below applies piecewise away from nonsmooth crossovers such as $\gamma = 1$; Appendix G also gives the exact discrete rule, which does not require this relaxation.

Theorem 7 (Capacity-matched representation size)**.** *Suppose D and ϕ are C^2 on the relevant interval and $D''(d) + \frac{\alpha}{n}\partial_\gamma h(d/n, C, T) > 0$. Then $\mathcal{J}_{\mathrm{RDGL}}$ is strictly convex there, and any interior minimizer $d^\star$ is unique and satisfies*

$$-\mathcal{D}'(d^\star) = \alpha\, h\left(\frac{d^\star}{n}, C, T\right). \tag{16}$$

Under smooth resource relaxations, if $\partial_\gamma h > 0$, $\partial_C h < 0$, $\partial_T h < 0$, and $h > 0$, then $d^\star$ increases with n, C, and T, and decreases with α.

Equation 16 gives the global capacity-matching rule: representation size grows until marginal reconstruction gain matches the generative price imposed on the downstream learner. Appendix G gives the exact comparative statics and discrete single-crossing rule.

The global optimum specifies how much capacity to retain, but not which latent directions deserve it. Let $\mu_Z = \mathbb{E}Z$, $\Sigma v_i = \lambda_i v_i$ with $\|v_i\|_2 = 1$, and $z_i = v_i^\top(Z - \mu_Z)$. For a finite-resource learner, let $e_\tau = \widehat{f}_\tau(Y_\tau) - f^\star_\tau(Y_\tau)$ denote its denoising error, and let ν be a measure over diffusion noise scales. Define the local reconstruction value and image-space generative burden by

$$b_i := \mathbb{E}\left[z_i^2\, \|J_D(Z)v_i\|_2^2\right], \qquad c_i(n, C, T) := \int \mathbb{E}\left[(v_i^\top e_\tau)^2 \|J_D(Z)v_i\|_2^2\right] d\nu(\tau). \tag{17}$$

For a linear decoder, $b_i = \lambda_i w_i$ with $w_i = \mathbb{E}\|J_D(Z)v_i\|_2^2$, illustrating why covariance magnitude alone does not determine image-space value.

Theorem 8 (Geometry-aware capacity matching)**.** *For binary mode retention $u_i \in \{0, 1\}$ and the separable local surrogate $J_{\mathrm{alloc}}(u) = J_0 + \sum_i[(1-u_i)b_i + \alpha u_i c_i]$, each mode is selected independently according to $u_i^\star = \mathbf{1}\{b_i > \alpha c_i\}$.*

More generally, suppose reconstruction value and generative burden are represented by quadratic operators $\mathsf{R} \succeq 0$ and $\mathsf{C} \succ 0$. Among k-dimensional subspaces normalized by $U^\top \mathsf{C} U = I_k$, the maximizer of $\mathrm{tr}(U^\top \mathsf{R} U)$ is spanned by the generalized eigenvectors with the k largest eigenvalues satisfying $Rv = \mu Cv$. When R and C are jointly diagonal in the covariance eigenbasis, $\mu_i = b_i/c_i$, recovering the mode-wise rule above.

Thus capacity should be allocated by reconstruction value relative to finite-resource generative burden, rather than by variance alone. The local derivation, finite-resource burden calibration, fractional hard-budget allocation, and basis-independent geometric construction are developed in Appendix H.

6.2 Capacity-Matched Spectral Tokenization

The preceding theory suggests matching the representation exposed by a tokenizer to the statistical, capacity, and optimization resources of its downstream generator. We instantiate this principle as *Capacity-Matched Spectral Tokenization* (CMST).

Because the global-scale reconstruction gauge can change covariance-based penalties without changing reconstruction, CMST first canonicalizes the latent seen by both decoder and generator:

$$\widetilde{Z} = \frac{Z - \mu_Z}{s_Z}, \qquad s_Z^2 = \frac{1}{d}\operatorname{tr}\Sigma_Z, \qquad \frac{1}{d}\operatorname{tr}\widetilde{\Sigma} = 1. \tag{18}$$

The corresponding decoder is $\widetilde{D}(u) = D(\mu_Z + s_Z u)$, so that $D(Z) = \widetilde{D}(\widetilde{Z})$.

Let $\widetilde{\Sigma} = V \operatorname{diag}(\lambda_1, \dots, \lambda_d) V^\top$ with $\lambda_1 \geq \cdots \geq \lambda_d$, define $w_i = \mathbb{E}\|J_{\widetilde{D}}(\widetilde{Z})v_i\|_2^2$, $\bar{w} = d^{-1}\sum_i w_i$, and let $H_{\widetilde{\Sigma}} = d^{-1}\sum_i \delta_{\lambda_i}$. For a target downstream resource specification $\mathcal{B} = (n, r_C, T, \eta, \nu)$, CMST uses the spectral generative tax

$$\widehat{\Gamma}_{\mathrm{CMST}} = \int \left[\underbrace{\bar{w}\,\tau^2 \Phi\left(\frac{d}{n}, \tau, H_{\widetilde{\Sigma}}\right)}_{\text{finite-data}} + \underbrace{\frac{1}{d}\sum_{i > r_C} w_i \frac{\lambda_i^2}{\lambda_i + \tau}}_{\text{finite-capacity}} + \underbrace{\frac{1}{d}\sum_{i \leq r_C} w_i \frac{\lambda_i^2}{\lambda_i + \tau}\,|1 - \eta(\lambda_i + \tau)|^{2T}}_{\text{finite-optimization}} \right] d\nu(\tau). \tag{19}$$

Here Φ is the high-dimensional score-risk functional from Theorem 4. The statistical term uses the averaged decoder sensitivity $\bar{w}$, while the capacity and optimization terms retain mode-wise weights. Equation 19 is a calibrated design surrogate: outside the oracle population model, finite-sample, capacity, optimization, and nonlinear effects may couple.

CMST is naturally posed as

$$\min_{E,D} \ \mathcal{L}_{\mathrm{rec}} \quad \text{s.t.} \quad \widehat{\Gamma}_{\mathrm{CMST}} \leq B_{\mathrm{gen}}. \tag{20}$$

We optimize Eq. 20 with a single dual multiplier and trace-normalized spectral gates $g_i \in [0, 1]$:

$$s_g^2 = \frac{1}{d}\sum_{i=1}^{d} g_i^2 \lambda_i, \qquad \widetilde{Z}_g = \frac{1}{s_g} V \operatorname{diag}(g_1, \dots, g_d) V^\top \widetilde{Z}, \qquad \lambda_i(g) = \frac{g_i^2 \lambda_i}{s_g^2}, \tag{21}$$

The gating preserves $d^{-1}\sum_i \widetilde{\lambda}_i(g) = 1$. At each spectral refresh, the r_C largest gated eigenvalues define the representable subspace; the spectral basis and decoder sensitivities are then treated as local stop-gradient statistics while the tokenizer and gates are updated by primal–dual optimization. Appendix I gives the scale and image-space calibration, KKT conditions, estimators, and pseudocode, while Appendix J collects limiting regimes and operational instantiations.

7 Experiments

We evaluate the theory along a mechanism–prediction–intervention chain. Reconstruction-preserving gauges isolate the effect of latent geometry; matched sweeps over data, model capacity, and optimization test the resource dependence of the preferred representation; and CMST turns this principle into a resource-matched tokenizer. Neural tokenizer–DiT experiments probe these effects beyond the Gaussian setting, while Appendix L provides controlled Gaussian and decoder-geometric validation.

Experimental setup. Unless otherwise stated, full-scale experiments use ImageNet-1K at 256×256 resolution (Deng et al., 2009). We instantiate CMST on a continuous $f16d32$ tokenizer family and evaluate downstream generative learnability with LightningDiT-XL/1 (Yao et al., 2025), using a fixed 80-epoch training budget and unguided generation as the primary protocol. Controlled comparisons include a reconstruction-only VAE, VA-VAE (Yao et al., 2025), an isotropy-regularized spectral baseline, AlignTok (Chen et al., 2026), and VTP (Yao et al., 2026); broader comparisons to recent tokenizer and end-to-end systems are reported separately under their native protocols. We report rFID and LPIPS for reconstruction and 50K-sample gFID together with FD-DINOv2 for generation (Heusel et al., 2017; Zhang et al., 2018; Stein et al., 2023). Full architectures, resource sweeps, CMST calibration, evaluation details, and compute are provided in Appendix K.

7.1 Reconstruction-Equivalent Latents Can Be Generatively Different

Starting from the same f16d32 tokenizer, we apply volume-preserving anisotropic gauges in the channel-covariance eigenbasis with $a \in \{1, 1.5, 2, 3, 4\}$, pairing every encoder transform with its exact inverse in the decoder; a Haar-random orthogonal gauge serves as the geometry-preserving control. We then train the same LightningDiT-XL/1 from scratch for 80 epochs on every latent space. The exact construction and normalization controls are given in Appendix L.1.

Figure 2 shows a sharp separation despite identical reconstruction maps: the orthogonal control remains at gFID 4.23 versus 4.18 for the identity, whereas the strongest anisotropic gauge reaches 15.61. Across the same intervention, $\kappa_{0.1}$ rises from 4.70 to 273.67, while the Gaussian spectral proxy changes only moderately, linking fixed-budget degradation to the geometry encountered by the diffusion learner. Variance-matched, normalization, and quantitative Gaussian controls are reported in Appendix L.1.

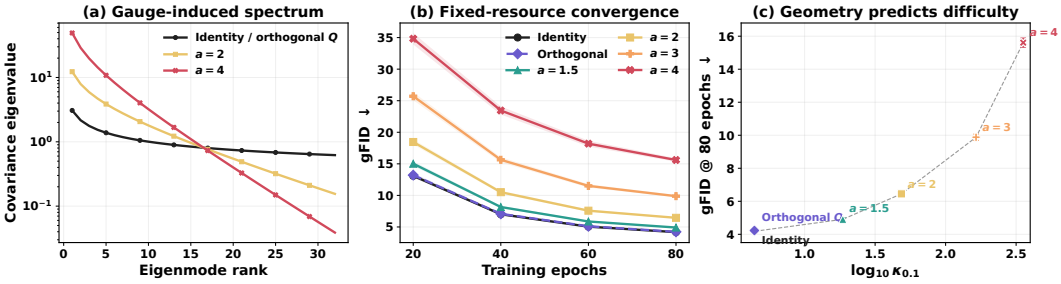


Figure 2: **Reconstruction-equivalent latent spaces can have sharply different generative difficulty.** Left: anisotropic gauges progressively spread the latent spectrum, whereas an orthogonal gauge preserves it. Middle: under identical downstream training, increasing gauge anisotropy systematically slows generation convergence; shaded regions denote ± 1 standard deviation over three runs. Right: fixed-budget gFID increases with the noise-regularized spectral condition number, while the orthogonal control remains near the identity representation. All gauges use the exact inverse decoder and therefore produce the same reconstructions.

7.2 Finite Resources Determine the Preferred Representation

We next test the central RDGL prediction that the preferred representation complexity is resource dependent rather than intrinsic to the tokenizer. We train matched f16 tokenizers with channel widths $c \in \{8, 16, 32, 64\}$, corresponding to $d = 256c$, while keeping the spatial token count fixed. We then vary one downstream resource at a time: the number of distinct training images n, LightningDiT model scale C, or training budget T_{train}. For the data and capacity sweeps, the number of optimizer updates is held fixed. Figure 3 reports the resulting reconstruction–generation trade-offs.

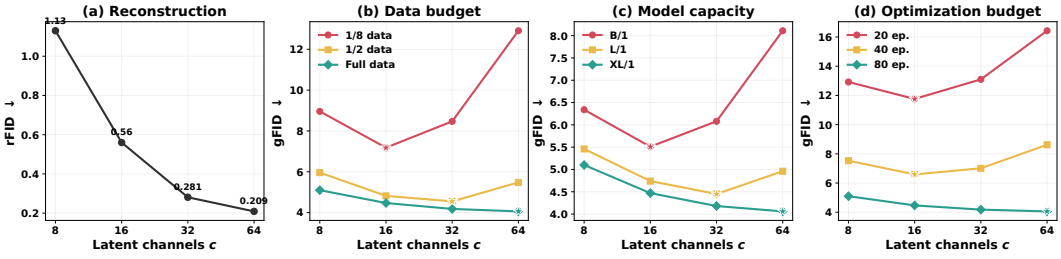


Figure 3: **The preferred representation shifts with downstream resources.** Increasing latent width monotonically improves reconstruction (a), but fixed-budget generation is non-monotonic. With more distinct data (b), larger model capacity (c), or longer optimization (d), the lowest-gFID operating point moves toward richer representations. Stars mark the best width within each resource level; curves show means over three runs.

Better reconstruction therefore need not imply easier generation under a fixed learner. At $1/8$ ImageNet, widening $c = 16 \to 64$ improves rFID from 0.56 to 0.21 but worsens gFID from 7.18 to

12.91, giving a direct neural reconstruction–generation reversal. As data, model capacity, or optimization increase, the lowest-gFID width shifts from 16 toward $32/64$; under the full-data XL/1 budget, $c = 64$ reaches gFID 4.06, consistent with the comparative statics of Theorem 7. Appendix L.2 gives extended and compute-matched sweeps, while Appendix L.3 quantitatively checks the finite-sample and RMT predictions.

7.3 Capacity-Matched Spectral Tokenization

Our default CMST tokenizer uses the same f16d32 architecture and reconstruction objective as the VA-VAE control, but is resource-matched to the full-data LightningDiT-XL/1, 80-epoch budget. Under the common downstream harness, we compare against the reconstruction-only VAE, VA-VAE, isotropy regularization, AlignTok, and VTP; Table 1 reports fixed-budget performance.

Table 1: **Fixed-resource ImageNet-256 evaluation.** All generation results use the common unguided LightningDiT-XL/1 80-epoch protocol. E@FID-4 denotes the first epoch at which gFID falls below 4. VTP† uses additional large-scale tokenizer pretraining data.

Tokenizer	rFID ↓	LPIPS ↓	gFID ↓	FD-DINOv2 ↓	E@FID-4 ↓
Recon-only VAE	**0.254**	**0.048**	8.92 ± 0.14	136.2 ± 1.9	> 80
VA-VAE (Yao et al., 2025)	0.281	0.052	4.18 ± 0.07	91.4 ± 1.2	> 80
Isotropy regularization	0.292	0.055	3.91 ± 0.06	83.5 ± 1.1	75
AlignTok (Chen et al., 2026)	<u>0.267</u>	0.058	3.61 ± 0.06	76.8 ± 1.0	64
VTP† (Yao et al., 2026)	0.360	<u>0.049</u>	<u>3.49 ± 0.05</u>	<u>74.6 ± 0.9</u>	<u>61</u>
CMST (ours)	0.296	0.056	$\mathbf{2.96 \pm 0.04}$	$\mathbf{63.1 \pm 0.8}$	**48**

Relative to matched VA-VAE, CMST reduces gFID from 4.18 to 2.96 (29%) and FD-DINOv2 from 91.4 to 63.1 (31%) with only $0.281 \rightarrow 0.296$ rFID, and reaches gFID 4 by epoch 48. Isotropy improves generation but remains behind CMST, showing that resource matching is not explained by covariance flattening alone. Figure 4 examines the learned spectral mechanism.

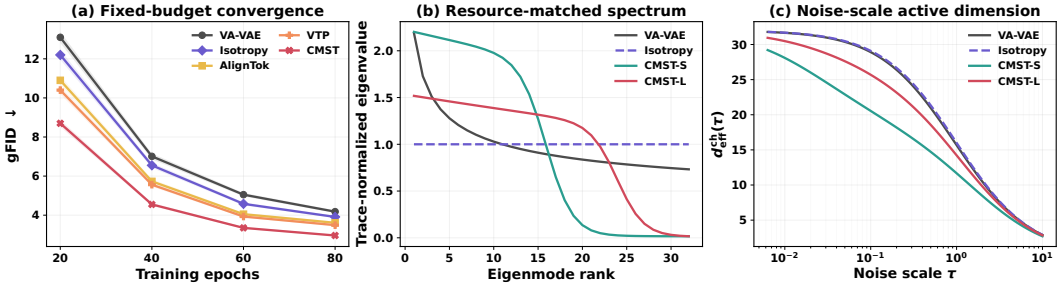


Figure 4: **CMST improves finite-resource learning by shaping, rather than merely flattening, the latent spectrum.** (a) CMST converges faster under the common XL/1 training budget. (b) Isotropy spreads variance uniformly, whereas CMST develops a resource-dependent plateau–tail structure; CMST-S targets a B/1, 40-epoch learner and CMST-L targets the default XL/1, 80-epoch learner. (c) The corresponding channel-level noise-scale active dimension is lower for the more constrained learner and increases when downstream resources permit a richer representation.

CMST-S suppresses more tail modes for the B/1, 40-epoch learner, whereas CMST-L exposes more active directions for XL/1, 80 epochs. Cross-evaluation swaps the preferred CMST variant across these learners, confirming resource-dependent adaptation (Appendix M.2). Decoder-geometry validation is in Appendix L.4; broader robustness, ablations, and overhead are reported in Appendix M.

8 Conclusion

We show that reconstruction quality alone does not determine a visual representation's suitability for generation: reconstruction leaves latent geometry underdetermined, while its learnability depends on downstream data, model capacity, and optimization budget. Our analysis formalizes this mismatch through finite-resource generative learnability, showing that representational detail carries both reconstruction value and a resource-dependent generative cost. This yields the RDGL view that preferred representations depend on the learner that must model them. CMST operationalizes this principle under a specified downstream budget, and experiments validate its mechanism and resource dependence. Overall, visual tokenization should be designed jointly with downstream generative learnability rather than reconstruction in isolation.

AI USE STATEMENT

We used OpenAI's ChatGPT as a research-support tool for early-stage brainstorming and feedback on high-level research framing and methodological presentation, literature discovery and bibliographic checking, and drafting, revising, and polishing limited portions of the manuscript. It was also used for final consistency and minor-error checks. Generative AI was not used to generate source code, run experiments, or produce experimental data or numerical results. All scientific decisions, experimental protocols, analyses, and conclusions were made by the authors. All cited sources were selected by the authors and independently verified against primary publisher or project records, and all AI-assisted text and suggestions were reviewed, verified, and revised by the authors. The authors take full responsibility for the technical content, claims, citations, and final presentation of this work.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we provide an **anonymized code repository** at `https://anonymous.4open.science/r/ICLR2027-CMST/`, and additionally include a snapshot of the anonymized code as part of the **supplementary material**. The repository provides the complete CMST training and evaluation pipeline, including continuous image tokenization, channel-level spectral gating, resource-aware optimization, downstream latent diffusion transformer training and image generation, as well as the Gaussian theory experiments, controlled representation and resource sweeps, decoder-geometry diagnostics, and ablation studies used in the paper.

The manuscript further documents the assumptions, derivations, experimental protocols, and implementation choices needed to reproduce the reported results. Sections 3–6 state the definitions, assumptions, objectives, and finite-resource settings underlying the theoretical development, while Appendices B–J provide detailed derivations and proofs for the diffusion-scale rate, reconstruction-gauge and latent-to-image transport results, finite-sample minimax bounds, random-matrix analysis, capacity and optimization laws, and the RDGL/CMST construction. Section 7 specifies the main evaluation protocol, and Appendix K documents the datasets and evaluation procedures, tokenizer and baseline configurations, downstream generator and resource budgets, practical CMST instantiation, optimization settings, compute environment, random-seed protocol, and statistical reporting. Appendices L–M provide the controlled Gaussian and decoder-geometry validations, complete resource sweeps, extended generation evaluations, cross-resource analyses, ablations, constraint-sensitivity studies, and computational-overhead measurements.

Together, the anonymized code release and the accompanying theoretical and experimental documentation are intended to support independent reproduction of the reported results.

REFERENCES

Zhidong Bai and Jack W. Silverstein. *Spectral Analysis of Large Dimensional Random Matrices*. Springer Series in Statistics. Springer, New York, NY, 2 edition, 2010. ISBN 978-1-4419-0660-1. doi: 10.1007/978-1-4419-0661-8.

Bowei Chen, Sai Bi, Hao Tan, He Zhang, Tianyuan Zhang, Zhengqi Li, Yuanjun Xiong, Jianming Zhang, and Kai Zhang. Aligning visual foundation encoders to tokenizers for diffusion models. In *International Conference on Learning Representations*, 2026. URL `https://openreview.net/forum?id=ajnBafpqmE`.

Hao Chen, Yujin Han, Fangyi Chen, Xiang Li, Yidong Wang, Jindong Wang, Ze Wang, Zicheng Liu, Difan Zou, and Bhiksha Raj. Masked autoencoders are effective tokenizers for diffusion models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 8145–8171. PMLR, 2025a. URL `https://proceedings.mlr.press/v267/chen25v.html`.

Hao Chen, Ze Wang, Xiang Li, Ximeng Sun, Fangyi Chen, Jiang Liu, Jindong Wang, Bhiksha Raj, Zicheng Liu, and Emad Barsoum. Softvq-vae: Efficient 1-dimensional continuous tokenizer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28358–28370, June 2025b.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Xianghong Fang, Wenlong Mou, Yuan Yuan, Dehan Kong, and Tim G. J. Rudner. A unified rate-distortion perspective on vector, product, and scalar quantization, 2026. URL https://arxiv.org/abs/2609.02107.

Dongning Guo, S. Shamai, and S. Verdu. Mutual information and minimum mean-square error in gaussian channels. *IEEE Transactions on Information Theory*, 51(4):1261–1282, 2005. doi: 10.1109/TIT.2005.844072.

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.

Theodoros Kouzelis, Ioannis Kakogeorgiou, Spyros Gidaris, and Nikos Komodakis. EQ-VAE: Equivariance regularized latent space for improved generative image modeling. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 31648–31666. PMLR, 2025. URL https://proceedings.mlr.press/v267/kouzelis25a.html.

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.

Bolin Lai, XuDong Wang, Saketh Rambhatla, James M. Rehg, Zsolt Kira, Rohit Girdhar, and Ishan Misra. Toward diffusible high-dimensional latent spaces: A frequency perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 43450–43460, June 2026.

Xingjian Leng, Jaskirat Singh, Yunzhong Hou, Zhenchang Xing, Saining Xie, and Liang Zheng. Repa-e: Unlocking vae for end-to-end tuning of latent diffusion transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 18262–18272, October 2025.

Jiasen Lu, Liangchen Song, Mingze Xu, Byeongjoo Ahn, Yanjun Wang, Chen Chen, Afshin Dehghan, and Yinfei Yang. Atoken: A unified tokenizer for vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28701–28711, June 2026.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4195–4205, October 2023.

Vivek Ramanujan, Kushal Tirumala, Armen Aghajanyan, Luke Zettlemoyer, and Ali Farhadi. When worse is better: Navigating the compression generation trade-off in visual tokenization. In *Advances in Neural Information Processing Systems*, volume 38, pp. 138949–138976, 2025. doi: 10.52202/085713-4178.

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

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.

J.W. Silverstein and S.I. Choi. Analysis of the limiting spectral distribution of large dimensional random matrices. *Journal of Multivariate Analysis*, 54(2):295–309, 1995. ISSN 0047-259X. doi: 10.1006/jmva.1995.1058. URL https://www.sciencedirect.com/science/article/pii/S0047259X85710585.

Jaskirat Singh, Boyang Zheng, Zongze Wu, Richard Zhang, Eli Shechtman, and Saining Xie. Improved baselines with representation autoencoders, 2026.

Ivan Skorokhodov, Sharath Girish, Benran Hu, Willi Menapace, Yanyu Li, Rameen Abdal, Sergey Tulyakov, and Aliaksandr Siarohin. Improving the diffusability of autoencoders. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 55876–55905. PMLR, 2025. URL https://proceedings.mlr.press/v267/skorokhodov25a.html.

George Stein, Jesse Cresswell, Rasa Hosseinzadeh, Yi Sui, Brendan Ross, Valentin Villecroze, Zhaoyan Liu, Anthony L. Caterini, J. Eric T. Taylor, and Gabriel Loaiza-Ganem. Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Binxu Wang, Jacob A. Zavatone-Veth, and Cengiz Pehlevan. A random matrix theory perspective on the consistency of diffusion models. In *International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2602.02908.

Tianwei Xiong, Jun Hao Liew, Zilong Huang, Jiashi Feng, and Xihui Liu. Gigatok: Scaling visual tokenizers to 3 billion parameters for autoregressive image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 18770–18780, October 2025.

Jingfeng Yao, Bin Yang, and Xinggang Wang. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15703–15712, June 2025.

Jingfeng Yao, Yuda Song, Yucong Zhou, and Xinggang Wang. Towards scalable pre-training of visual tokenizers for generation. In *European Conference on Computer Vision (ECCV)*, 2026.

Hengyu Zeng, Xin Gao, Guanghao Li, Yuxiang Yan, Jiaoyang Ruan, Junpeng Ma, Haoyu Albert Wang, and Jian Pu. Mactok: Robust continuous tokenization for image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 43407–43417, June 2026.

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

Bingchen Zhao, Qiushan Guo, Ye Wang, Yixuan Huang, Zhonghua Zhai, and Yu Tian. Composable visual tokenizers with generator-free diagnostics of learnability, 2026. URL https://arxiv.org/abs/2602.03339.

Bowen Zheng, Weijian Luo, Guang Yang, Colin Zhang, and Tianyang Hu. Taming the entropy cliff: Variable codebook size quantization for autoregressive visual generation, 2026a. URL https://arxiv.org/abs/2605.06207.

Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion transformers with representation autoencoders. In *International Conference on Learning Representations*, 2026b. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/3c4141c12660ad3625eb4ae845e0a6f9-Abstract-Conference.html.

Appendix Contents

A Notation and Conventions **15**

B Continuous Representation Rate and Generative Learnability **16**

B.1 Why Deterministic Mutual Information Is Not a Finite Representation Rate 16

B.2 Noise-Scale Rate and Active Dimension 17

B.3 Two-Stage Risks and Resource-Constrained Learnability 18

C Reconstruction Gauge and Latent-to-Image Transport **18**
C.1 Reconstruction Gauge and Diffusion Geometry 18
C.2 Latent-to-Image Transport 20

D Finite-Sample Minimax Theory **21**
D.1 Gaussian Wasserstein Geometry 21
D.2 Minimax Lower Bound 22
D.3 Matching Upper Bound 22
D.4 Transport to Image Space and Reversal 23

E Random-Matrix Analysis of Gaussian Score Learning **24**
E.1 General Spectral Deterministic Equivalent 24
E.2 Isotropic Limit, Low-Noise Regimes, and Critical Scaling 25

F Population Capacity and Optimization Analysis **26**
F.1 Spectrum-Resolved Capacity 27
F.2 Finite-Time Gradient Dynamics 27
F.3 Joint Capacity–Optimization Law 28

G Capacity-Matched Representation Size **28**
G.1 Continuous Capacity-Matched Optimum 28
G.2 Discrete Representation Sizes 30

H Geometry-Aware Representation Allocation **30**
H.1 Local Decoder Geometry and Generative Burden 30
H.2 Mode-Wise Allocation Laws 31
H.3 Generalized Reconstruction–Generation Geometry 32

I Capacity-Matched Spectral Tokenization **32**
I.1 Canonical Scale and Gauge Degeneracy 32
I.2 Calibrating Generative Taxes to Image-Space Units 33
I.3 Constrained Optimization and Spectral Gates 34
I.4 Estimators and Computational Cost 36
I.5 Detail Offloading as an Architectural Corollary 37

J Concrete Instantiations and Operational Interpretation **37**
J.1 Limiting Regimes of the CMST Objective 37
J.2 Representative Spectral Families 38
J.3 A Worked RDGL Scaling Example 40
J.4 Theory-to-Measurement Dictionary 40

K Experimental Setup and Implementation Details **41**

K.1 Datasets and Evaluation Protocol 41

K.2 Tokenizer Families and Baselines 42

K.3 Downstream Generator and Resource Budgets 42

K.4 Practical CMST Instantiation 43

K.5 Optimization, Compute, and Reproducibility 44

L Theory-Aligned Empirical Validation **44**

L.1 Reconstruction-Gauge Stress Test 44

L.2 Finite-Resource Representation Sweeps 46

L.3 Finite-Sample and RMT Numerical Validation 47

L.4 Decoder Geometry and Modal Allocation 49

M CMST Evaluation and Ablations **50**

M.1 Extended Generation Evaluation 50

M.2 Resource Matching and Spectral Adaptation 51

M.3 Component Ablations and Constraint Sensitivity 52

M.4 Computational Overhead 53

APPENDIX

A NOTATION AND CONVENTIONS

This appendix summarizes the recurring notation used throughout the theory and method. We include the quantities that define the representation, downstream resources, spectral analysis, and CMST objective; variables used only locally inside individual proofs are introduced where needed.

Conventions. The symbol $\#$ denotes pushforward of a probability measure, and $\preceq$ denotes the Loewner order. We use $\|\cdot\|_2$, $\|\cdot\|_F$, and $\|\cdot\|_{\mathrm{op}}$ for Euclidean, Frobenius, and operator norms, respectively. Hats denote empirical, plug-in, or finite-sample quantities, while $\star$ denotes an optimal or population-oracle quantity. Tildes mark transformed or calibrated quantities and are defined explicitly where they appear. Subscripted λ_i always denotes a covariance eigenvalue; un-subscripted Lagrange multipliers such as $\lambda^\star$ or λ_{dual} are distinguished by context. We likewise distinguish the scalar capacity budget C from the generative-cost operator C, and the decoder D from the reconstruction curve $\mathcal{D}$. The symbols and terms used throughout this paper are summarized in Table 2.

Table 2: Summary of notation and conventions.

Symbol	Meaning
Representation and generative risks	
$X \sim P_X,\ \mathcal{X}$	Image random variable, its distribution, and image space $\mathcal{X} \subseteq \mathbb{R}^p$.
$\mathcal{T} = (E, D)$	Deterministic tokenizer with encoder $E : \mathcal{X} \to \mathbb{R}^d$ and decoder $D : \mathbb{R}^d \to \mathcal{X}$.
$Z,\ P_Z$	Latent representation $Z = E(X)$ and induced latent law $P_Z = E_\# P_X$.
$Q,\ P_{\mathrm{gen}}$	Learned latent distribution and generated image distribution $P_{\mathrm{gen}} = D_\# Q$.
$d,\ n,\ C,\ T$	Representation dimension/size, downstream sample budget, model-capacity budget, and optimization budget.
$S_n,\ \mathcal{G}_C,\ \mathfrak{A}_{n,C,T}$	Latent training sample, generator family under capacity C, and admissible learning procedures under resources (n, C, T).
$W_2,\ \mathcal{P}_2(\mathbb{R}^d)$	2-Wasserstein distance and probability measures on $\mathbb{R}^d$ with finite second moment.
$r_{\mathrm{rec}},\ \delta_{\mathrm{rec}}$	Pointwise RMS reconstruction error and distributional reconstruction discrepancy $W_2(P_X, D_\# P_Z)$.
$\mathcal{L}_{\mathrm{lat}},\ \mathcal{L}_{\mathrm{img}}$	Latent and image-space generative risks, $W_2^2(P_Z, Q)$ and $W_2^2(P_X, D_\# Q)$.
$\mathfrak{G}_{n,C,T}$	Resource-constrained minimax generative learnability over a latent distribution class.
$m_D,\ L_D,\ \kappa_D$	Lower/upper bi-Lipschitz constants of the decoder and $\kappa_D = L_D/m_D$.
Diffusion-scale rate and spectral geometry	
$Z_\tau,\ \tau$	Gaussian-corrupted latent $Z_\tau = Z + \sqrt{\tau}\varepsilon$; τ is the noise variance.
$R_\tau(Z)$	Diffusion-scale representation rate $I(Z; Z_\tau)$.
$d_{\mathrm{eff}}(\tau)$	Noise-scale active dimension $2\,\partial R_\tau / \partial \log(1/\tau)$.
$\mathrm{mmse}_Z(\tau),\ K_\tau$	Scalar MMSE of estimating Z from Z_τ and the normalized MMSE error-covariance operator; $\mathrm{tr}\, K_\tau = d_{\mathrm{eff}}(\tau)$.
$\Sigma,\ \lambda_i,\ v_i$	Latent covariance and its eigenpairs, $\Sigma v_i = \lambda_i v_i$.
$G_D(z)$	Decoder pullback metric $G_D(z) = J_D(z)^\top J_D(z)$.
$\gamma = d/n$	Sample-to-representation aspect ratio used in the high-dimensional analysis.
$H_d,\ H$	Empirical spectral distribution of Σ_d and its limiting spectral law.
Finite-resource statistical, capacity, and optimization laws	
$\mathcal{C}_d(\rho)$	Uniformly conditioned Gaussian covariance class $\{\, \Sigma : (1-\rho) I_d \preceq \Sigma \preceq (1+\rho) I_d \,\}$.
$\mathfrak{R}_{n,d}(\rho),\ \psi_n(d)$	Gaussian minimax latent risk and its characteristic scaling $\psi_n(d) = \min\{d, d^2/n\}$.
$\Phi_d(\tau),\ \Phi(\gamma, \tau, H)$	Finite-sample normalized score-estimation risk and its RMT deterministic equivalent.
$\Psi_d(\tau)$	Normalized excess denoising risk, $\Psi_d(\tau) = \tau^2 \Phi_d(\tau)$.
$\omega_\tau(\lambda)$	Modal denoising value $\omega_\tau(\lambda) = \lambda^2/(\lambda + \tau)$.

Symbol	Meaning
$\Gamma_{\mathrm{cap}},\ \Gamma_{\mathrm{opt}}$	Population excess risks induced by finite rank and finite optimization time.
$r,\ r_C$	Rank constraint in the Gaussian surrogate and effective denoising-rank budget used by CMST.
η, κ_τ	Effective optimization step size and noise-regularized spectral condition number $\kappa_\tau = (\lambda_1 + \tau)/(\lambda_d + \tau)$.
RDGL and geometry-aware allocation	
$\mathcal{D}(d)$	Reconstruction distortion along a representation-size family.
$\mathcal{G}(d; n, C, T)$	Calibrated downstream generative tax for representation size d and resources (n, C, T).
$\mathcal{J}_{\mathrm{RDGL}}$	Rate–Distortion–Generative Learnability objective $\mathcal{D}(d) + \alpha\mathcal{G}(d; n, C, T)$.
α	Trade-off weight between reconstruction distortion and downstream generative difficulty.
$h(\gamma, C, T)$	Marginal generative tax $\phi + \gamma\,\partial_\gamma\phi$ when $\mathcal{G} = d\,\phi(\gamma, C, T)$.
$\Delta_{\mathrm{rec}},\ \Delta_{\mathrm{gen}}$	Discrete marginal reconstruction gain and generative cost of increasing representation size.
$b_i,\ w_i$	Local reconstruction value of mode i and decoder sensitivity $w_i = \mathbb{E}\|J_D(Z)v_i\|_2^2$.
$\widetilde{c}_i,\ c_i$	Latent modal denoising excess and its corresponding local image-space generative burden.
u_i	Binary/fractional mode-retention variable.
$\mathsf{R},\ \mathsf{C}$	Quadratic reconstruction-value and generative-cost operators in the generalized spectral allocation problem.
Capacity-Matched Spectral Tokenization	
$\widetilde{Z},\ \widetilde{D}$	Centered, trace-normalized latent representation and the corresponding decoder in canonicalized coordinates.
$\widetilde{\Sigma},\ H_{\widetilde{\Sigma}}$	Covariance of $\widetilde{Z}$ and its empirical spectral measure $d^{-1}\sum_i \delta_{\lambda_i}$.
$\bar{w}$	Mean decoder sensitivity, $\bar{w} = d^{-1}\sum_i w_i$.
ν	Measure over diffusion noise scales used to aggregate the CMST spectral tax.
$\mathcal{B} = (n, r_C, T, \eta, \nu)$	Target downstream resource specification used by CMST.
$\widehat{\Gamma}_{\mathrm{CMST}}$	Theory-calibrated CMST spectral tax combining finite-data, capacity, and optimization burdens.
B_{gen}	Allowed generative-difficulty budget in the constrained CMST formulation.
$g_i,\ s_g$	Spectral gate for mode i and the normalization factor that preserves unit average variance.
$\lambda_i(g)$	Trace-normalized covariance eigenvalue after spectral gating, $\lambda_i(g) = g_i^2\lambda_i/s_g^2$.
$\mathcal{S}_C(g)$	Indices of the r_C largest gated eigenvalues, i.e. the modes retained by the rank-capacity surrogate.
λ_{dual}	Dual variable used in the primal–dual implementation of the CMST constraint.

B Continuous Representation Rate and Generative Learnability

This appendix provides the technical details underlying the rate and finite-resource learnability definitions in Section 3.

B.1 Why Deterministic Mutual Information Is Not a Finite Representation Rate

For discrete representations, entropy or coding rate provides a natural finite measure of representation complexity. Deterministic continuous encoders behave differently.

Proposition 9 (Mutual information of a deterministic continuous representation)**.** *Let $Z = E(X)$ for a measurable deterministic encoder E, and suppose that P_Z is non-atomic. Then*

$$I(X; Z) = D_{\mathrm{KL}}(P_{X,Z} \,\|\, P_X \otimes P_Z) = +\infty.$$

Proof. Let $\mathcal{G}_E = \{(x, z) : z = E(x)\}$ be the graph of the encoder. Since $Z = E(X)$ almost surely, $P_{X,Z}(\mathcal{G}_E) = 1$. In contrast,

$$(P_X \otimes P_Z)(\mathcal{G}_E) = \int P_Z(\{E(x)\})\, dP_X(x) = 0,$$

because P_Z is non-atomic. Hence $P_{X,Z}$ is singular with respect to $P_X \otimes P_Z$, so the corresponding Kullback–Leibler divergence is infinite. □

Proposition 9 rules out $I(X; Z)$ as a generic finite rate for deterministic continuous tokenizers. A conventional alternative is to introduce an explicit resolution. For a quantizer Q_Δ with cell width Δ, define $R_\Delta(Z) = H(Q_\Delta(Z))$; its small-scale growth yields the information dimension

$$d_I(Z) = \lim_{\Delta \downarrow 0} \frac{H(Q_\Delta(Z))}{\log(1/\Delta)},$$

whenever the limit exists. In latent diffusion, the forward noise schedule itself supplies a family of operational resolutions, motivating the diffusion-scale rate used in the main text.

B.2 Noise-Scale Rate and Active Dimension

Let $Z_\tau = Z + \sqrt{\tau}\varepsilon$ with $\varepsilon \sim \mathcal{N}(0, I_d)$ independent of Z. For $Z \sim \mathcal{N}(\mu, \Sigma)$, $Z_\tau \sim \mathcal{N}(\mu, \Sigma + \tau I)$, and

$$\begin{aligned} R_\tau(Z) &= h(Z_\tau) - h(Z_\tau \mid Z) \\ &= \frac{1}{2}\log\det(\Sigma + \tau I) - \frac{1}{2}\log\det(\tau I) = \frac{1}{2}\log\det\left(I + \frac{\Sigma}{\tau}\right). \end{aligned} \tag{22}$$

Thus, if $\lambda_1, \ldots, \lambda_d$ are the eigenvalues of Σ,

$$R_\tau(Z) = \frac{1}{2}\sum_{i=1}^{d} \log\left(1 + \frac{\lambda_i}{\tau}\right).$$

Differentiating with respect to logarithmic inverse-noise scale gives

$$\begin{aligned} d_{\text{eff}}(\tau) &= 2\frac{\partial R_\tau}{\partial \log(1/\tau)} = -2\tau\frac{\partial R_\tau}{\partial \tau} \\ &= \sum_{i=1}^{d} \frac{\lambda_i}{\lambda_i + \tau} = \operatorname{tr} K_\tau, \qquad K_\tau := \Sigma(\Sigma + \tau I)^{-1}. \end{aligned} \tag{23}$$

Consequently, $0 \le d_{\text{eff}}(\tau) \le \operatorname{rank}(\Sigma) \le d$, $d_{\text{eff}}(\tau)$ is nonincreasing in τ, and

$$\lim_{\tau \downarrow 0} d_{\text{eff}}(\tau) = \operatorname{rank}(\Sigma), \qquad \lim_{\tau \to \infty} d_{\text{eff}}(\tau) = 0.$$

Each eigenmode contributes the soft activity weight $\lambda_i/(\lambda_i + \tau)$.

The additive-noise parameterization covers the usual Gaussian forward processes. If $Y_t = \alpha_t Z + \sigma_t \varepsilon$ with $\alpha_t > 0$, then

$$\frac{Y_t}{\alpha_t} = Z + \sqrt{\tau_t}\varepsilon, \qquad \tau_t = \frac{\sigma_t^2}{\alpha_t^2}.$$

The same rate admits a distribution-free denoising interpretation. For any Z with finite second moment, define

$$\operatorname{mmse}_Z(\tau) := \mathbb{E}\left[\|Z - \mathbb{E}[Z \mid Z_\tau]\|_2^2\right].$$

The I–MMSE identity (Guo et al., 2005) yields

$$d_{\text{eff}}(\tau) = \frac{\operatorname{mmse}_Z(\tau)}{\tau}, \qquad R_\tau(Z) = \frac{1}{2}\int_\tau^\infty \frac{\operatorname{mmse}_Z(u)}{u^2}\, du. \tag{24}$$

Hence R_τ accumulates the recoverable content of the representation over noise scales.

A corresponding matrix-valued quantity is

$$K_\tau := \frac{1}{\tau}\mathbb{E}\left[\big(Z - \mathbb{E}[Z \mid Z_\tau]\big)\big(Z - \mathbb{E}[Z \mid Z_\tau]\big)^\top\right],$$

for which $d_{\mathrm{eff}}(\tau) = \operatorname{tr} K_\tau$. For Gaussian Z this reduces to $K_\tau = \Sigma(\Sigma + \tau I)^{-1}$, recovering Eq. 23. Equivalently, if

$$\mathcal{J}(Z_\tau) = \mathbb{E}\left[\|\nabla \log p_{Z_\tau}(Z_\tau)\|_2^2\right]$$

denotes scalar Fisher information, Tweedie's identity gives

$$d_{\mathrm{eff}}(\tau) = d - \tau \mathcal{J}(Z_\tau).$$

For nonzero K_τ, one may further summarize how its active mass is distributed through the participation-ratio effective rank

$$r_{\mathrm{eff}}(\tau) := \frac{(\operatorname{tr} K_\tau)^2}{\operatorname{tr}(K_\tau^2)}. \tag{25}$$

When active directions contribute comparably, $r_{\mathrm{eff}}(\tau)$ approaches their number; when the activity is strongly concentrated, it can be substantially smaller. These quantities are spectral summaries; the high-dimensional analysis in Section 5.2 instead retains the full limiting spectral distribution.

B.3 Two-Stage Risks and Resource-Constrained Learnability

The main text uses the distributional reconstruction discrepancy δ_{rec} because it interfaces directly with Wasserstein generative risk. For completeness, the pointwise RMS reconstruction error is

$$r_{\mathrm{rec}}^2(\mathcal{T}) = \mathbb{E}_{X\sim P_X}\left[\|X - D(E(X))\|_2^2\right], \qquad \delta_{\mathrm{rec}}(\mathcal{T}) = W_2(P_X, D_\# P_Z). \tag{26}$$

The coupling $(X, D(E(X)))$ gives $\delta_{\mathrm{rec}}(\mathcal{T}) \le r_{\mathrm{rec}}(\mathcal{T})$. Thus r_{rec} is a natural tokenizer-training distortion, whereas δ_{rec} is the image-space discrepancy remaining when the latent law is modeled exactly.

For a fixed learning procedure $\mathcal{A}$ and latent law P, define its pointwise generative risk as

$$\mathcal{R}_{\mathrm{gen}}(\mathcal{A}; P) = \mathbb{E}_{S_n\sim P^{\otimes n}}\left[W_2^2(P, Q_{\mathcal{A},S_n})\right].$$

The distinction between this pointwise risk and the minimax quantity $\mathfrak{G}_{n,C,T}(\mathcal{P})$ in Eq. 2 is essential. An expression such as

$$\inf_{\mathcal{A}} \mathbb{E}_{S_n\sim P^{\otimes n}} W_2^2(P, Q_{\mathcal{A},S_n})$$

is degenerate if the procedure itself may depend on the unknown P: an oracle could simply encode and return P. Statistical learnability therefore refers either to a fixed procedure evaluated at P, or to a single procedure that must operate uniformly over a class $\mathcal{P}$.

The resource constraints play distinct roles. The sample budget n limits statistical estimation, the capacity budget C restricts the generator family, and the optimization budget T limits how closely training can approach the best model available within that family.

C Reconstruction Gauge and Latent-to-Image Transport

C.1 Reconstruction Gauge and Diffusion Geometry

For an invertible latent reparameterization $\phi : \mathbb{R}^d \to \mathbb{R}^d$, recall $E_\phi = \phi \circ E$ and $D_\phi = D \circ \phi^{-1}$. For the linear subgroup $\phi(z) = Az$, $A \in \mathrm{GL}(d)$, we write $E_A = AE$ and $D_A = D \circ A^{-1}$ and refer to $\{(E_A, D_A) : A \in \mathrm{GL}(d)\}$ as the *reconstruction gauge orbit*.

Proof of Theorem 1. By construction,

$$D_\phi(E_\phi(x)) = D\big(\phi^{-1}(\phi(E(x)))\big) = D(E(x))$$

for every $x \in \mathcal{X}$. Hence, for any measurable reconstruction loss ℓ for which the expectation exists,

$$\mathbb{E}\left[\ell(X, D_\phi(E_\phi(X)))\right] = \mathbb{E}\left[\ell(X, D(E(X)))\right].$$

Thus the entire reconstruction map, rather than only a particular scalar loss, is invariant under invertible latent reparameterization.

For a linear gauge $A \in \mathrm{GL}(d)$, let $Z_A = AZ$. Applying the standard isotropic corruption in the transformed coordinates gives

$$Z_{A,\tau} = AZ + \sqrt{\tau}\,\varepsilon.$$

Pulling this process back through A^{-1} yields

$$A^{-1}Z_{A,\tau} = Z + \sqrt{\tau}\,A^{-1}\varepsilon, \qquad \mathrm{Cov}(A^{-1}\varepsilon) = A^{-1}A^{-\top}.$$

The pulled-back corruption has covariance τI at the same noise scale if and only if $A^{-1}A^{-\top} = I$, equivalently $AA^{\top} = I$. Thus only orthogonal linear gauges preserve standard isotropic diffusion exactly.

Now let $Z \sim \mathcal{N}(0, I_2)$ and $A_a = \mathrm{diag}(a, a^{-1})$. Then

$$A_a Z \sim \mathcal{N}\left(0, \Sigma_a\right), \qquad \Sigma_a = \begin{bmatrix} a^2 & 0 \\ 0 & a^{-2} \end{bmatrix}.$$

Since $\det A_a = 1$, this is a volume-preserving gauge, while reconstruction remains exactly unchanged by the first part of the theorem.

By Eq. 3,

$$R_\tau(A_a Z) = \frac{1}{2}\left[\log\left(1 + \frac{a^2}{\tau}\right) + \log\left(1 + \frac{a^{-2}}{\tau}\right)\right].$$

For fixed $\tau > 0$, the first term grows as $\log a + O(1)$ and the second converges to zero, so $R_\tau(A_a Z) \to \infty$.

The noised latent distribution is $\mathcal{N}(0, \Sigma_a + \tau I)$, whose score is

$$s_{a,\tau}(y) = -(\Sigma_a + \tau I)^{-1} y.$$

Its Jacobian is $-(\Sigma_a + \tau I)^{-1}$, and therefore

$$\mathrm{cond}(\nabla s_{a,\tau}) = \mathrm{cond}\big((\Sigma_a + \tau I)^{-1}\big) = \mathrm{cond}(\Sigma_a + \tau I) = \frac{a^2 + \tau}{a^{-2} + \tau} \longrightarrow \infty.$$

This proves Eq. 4. Since all members of the family have identical $D \circ E$ and identical latent dimension $d = 2$, no function of $(D \circ E, d)$ can determine either R_τ or the conditioning of the noised Gaussian score over all continuous tokenizers. □

Figure 5 visualizes the volume-preserving Gaussian gauge used in the proof, contrasting its unchanged reconstruction map with its altered diffusion geometry.

The same mismatch admits a decoder-geometric interpretation. If ϕ is a C^1 diffeomorphism and

$$G_D(z) := J_D(z)^{\top} J_D(z)$$

denotes the decoder pullback metric, then for $z' = \phi(z)$,

$$J_{D_\phi}(z') = J_D(z) J_{\phi^{-1}}(z'), \qquad G_{D_\phi}(z') = J_{\phi^{-1}}(z')^{\top} G_D(z) J_{\phi^{-1}}(z').$$

In particular, for $\phi(z) = Az$,

$$G_{D_A}(Az) = A^{-\top} G_D(z) A^{-1}.$$

Moreover,

$$\|D(z + \delta z) - D(z)\|_2^2 = \delta z^{\top} G_D(z) \delta z + o(\|\delta z\|_2^2).$$

Thus reconstruction geometry co-transforms with the decoder, whereas standard latent diffusion continues to measure corruption in the fixed Euclidean metric. This mismatch underlies the geometry-aware allocation developed later in Section 6.1.

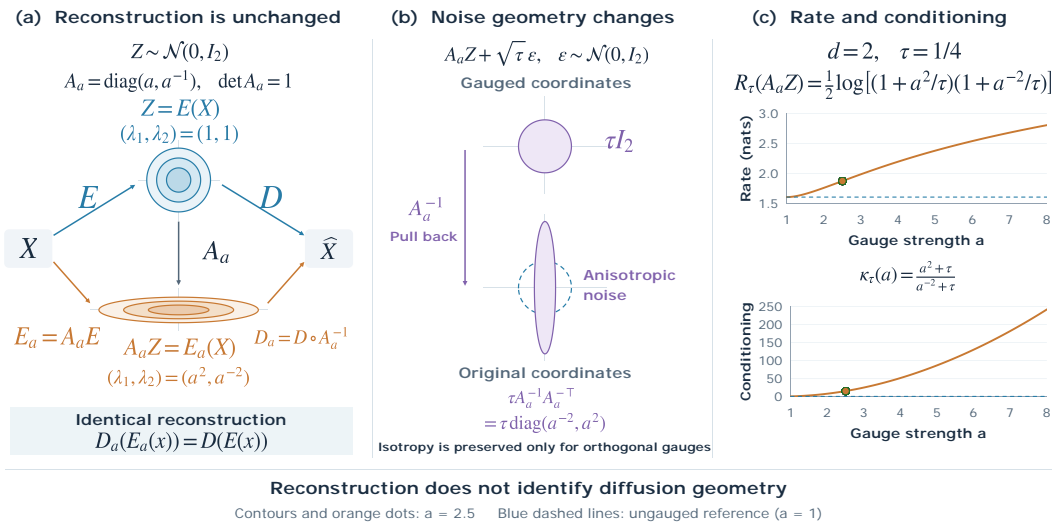


Figure 5: **Same reconstruction, different diffusion geometry.** Let $Z \sim \mathcal{N}(0, I_2)$ and $A_a = \mathrm{diag}(a, a^{-1})$, with $a \geq 1$. **(a)** The paired transformations $E_a = A_a E$ and $D_a = D \circ A_a^{-1}$ preserve the reconstruction map exactly, although the latent contours become anisotropic. **(b)** Isotropic corruption in the gauged coordinates pulls back to noise with covariance $\tau A_a^{-1} A_a^{-\top} = \tau\,\mathrm{diag}(a^{-2}, a^2)$ in the original coordinates. The dashed circle shows the original isotropic noise covariance τI_2 for comparison. **(c)** Exact analytical curves at fixed $d = 2$ and $\tau = 1/4$ show the diffusion-scale rate and the condition number of the noised Gaussian score Jacobian as functions of a. Dashed horizontal lines mark the ungauged values at $a = 1$; dots mark $a = 2.5$, used for the contours in panels (a)–(b). Reconstruction invariance therefore does not imply invariance of diffusion geometry.

C.2 Latent-to-Image Transport

Let P and Q be distributions supported on $\mathcal{S}$. If D is L_D-Lipschitz on $\mathcal{S}$, pushing any coupling $\pi \in \Pi(P, Q)$ through (D, D) gives

$$\int \|D(z) - D(z')\|_2^2 \, d\pi \leq L_D^2 \int \|z - z'\|_2^2 \, d\pi.$$

Taking the infimum over couplings yields

$$W_2(D_\# P, D_\# Q) \leq L_D W_2(P, Q). \tag{27}$$

Conversely, if $\|D(z) - D(z')\|_2 \geq m_D \|z - z'\|_2$ on $\mathcal{S}$, then $D|_{\mathcal{S}}$ is injective and its inverse on $D(\mathcal{S})$ is $1/m_D$-Lipschitz. Pulling couplings back through this inverse gives

$$W_2(D_\# P, D_\# Q) \geq m_D W_2(P, Q). \tag{28}$$

Proof of Theorem 2. The triangle inequality and Eq. 27 give

$$\begin{aligned} W_2(P_X, D_\# Q) &\leq W_2(P_X, D_\# P_Z) + W_2(D_\# P_Z, D_\# Q) \\ &\leq \delta_{\mathrm{rec}} + L_D W_2(P_Z, Q). \end{aligned}$$

Similarly, the reverse triangle inequality and Eq. 28 give

$$\begin{aligned} W_2(P_X, D_\# Q) &\geq W_2(D_\# P_Z, D_\# Q) - W_2(P_X, D_\# P_Z) \\ &\geq m_D W_2(P_Z, Q) - \delta_{\mathrm{rec}}. \end{aligned}$$

Combining this with nonnegativity of W_2 proves Eq. 5. □

Squaring the two sides of Eq. 5 and using $(a - b)_+^2 \geq a^2/2 - b^2$ and $(a + b)^2 \leq 2a^2 + 2b^2$ for $a, b \geq 0$ yields, in the notation of Eq. 1,

$$\frac{m_D^2}{2} \mathcal{L}_{\mathrm{lat}}(Q; P_Z) - \delta_{\mathrm{rec}}^2 \leq \mathcal{L}_{\mathrm{img}}(Q; \mathcal{T}) \leq 2\delta_{\mathrm{rec}}^2 + 2L_D^2 \mathcal{L}_{\mathrm{lat}}(Q; P_Z). \tag{29}$$

The ratio $\kappa_D = L_D / m_D$ is the corresponding decoder condition number.

D FINITE-SAMPLE MINIMAX THEORY

This appendix proves Theorem 3. Throughout, $\rho \in (0,1)$ is fixed; constants depending only on ρ are denoted by c_ρ, C_ρ and may change from line to line.

D.1 GAUSSIAN WASSERSTEIN GEOMETRY

For centered Gaussians,

$$W_2^2(\mathcal{N}(0,\Sigma_1),\mathcal{N}(0,\Sigma_2)) = \operatorname{tr}\Sigma_1 + \operatorname{tr}\Sigma_2 - 2\operatorname{tr}\left(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2}\right)^{1/2}.$$

On the uniformly conditioned family $\mathcal{C}_d(\rho)$, this metric is uniformly equivalent, up to ρ-dependent constants, to squared Frobenius distance.

Lemma 10 (Bures–Frobenius equivalence)**.** *There exist constants $0 < c_\rho \le C_\rho < \infty$ such that, for all $\Sigma_1, \Sigma_2 \in \mathcal{C}_d(\rho)$,*

$$c_\rho\|\Sigma_1 - \Sigma_2\|_F^2 \le W_2^2(\mathcal{N}(0,\Sigma_1),\mathcal{N}(0,\Sigma_2)) \le C_\rho\|\Sigma_1 - \Sigma_2\|_F^2. \tag{30}$$

Proof. The Gaussian W_2 distance admits the Procrustes representation

$$W_2^2 = \min_{U^\top U = I}\left\|\Sigma_1^{1/2} - \Sigma_2^{1/2}U\right\|_F^2.$$

Let $A = \Sigma_1^{1/2}$, $B = \Sigma_2^{1/2}$, and let $U_\star$ attain the minimum. Since $\|A\|_{\mathrm{op}}, \|B\|_{\mathrm{op}} \le \sqrt{1+\rho}$,

$$\|\Sigma_1 - \Sigma_2\|_F = \|AA^\top - (BU_\star)(BU_\star)^\top\|_F \le 2\sqrt{1+\rho}\,\|A - BU_\star\|_F,$$

which gives the lower inequality in Eq. 30.

For the reverse direction, the spectra are bounded below by $1-\rho > 0$. The matrix square-root map is Lipschitz on this spectral interval, so

$$\|\Sigma_1^{1/2} - \Sigma_2^{1/2}\|_F \le C_\rho\|\Sigma_1 - \Sigma_2\|_F.$$

Choosing $U = I$ in the Procrustes representation proves the upper bound. □

We also require the corresponding local control of Gaussian relative entropy.

Lemma 11 (Gaussian KL control)**.** *For all $\Sigma_1, \Sigma_2 \in \mathcal{C}_d(\rho)$,*

$$D_{\mathrm{KL}}(\mathcal{N}(0,\Sigma_1)\,\|\,\mathcal{N}(0,\Sigma_2)) \le C_\rho\|\Sigma_1 - \Sigma_2\|_F^2.$$

Consequently, for n independent samples,

$$D_{\mathrm{KL}}\big(P_{\Sigma_1}^{\otimes n}\,\big\|\,P_{\Sigma_2}^{\otimes n}\big) \le C_\rho n\|\Sigma_1 - \Sigma_2\|_F^2.$$

Proof. Let $M = \Sigma_2^{-1/2}\Sigma_1\Sigma_2^{-1/2}$. Then

$$2D_{\mathrm{KL}} = \operatorname{tr} M - d - \log\det M.$$

The eigenvalues of M lie in a compact interval bounded away from zero and infinity, with bounds depending only on ρ. On this interval, $x - 1 - \log x \le C_\rho(x-1)^2$, and therefore

$$D_{\mathrm{KL}} \le C_\rho\left\|\Sigma_2^{-1/2}(\Sigma_1 - \Sigma_2)\Sigma_2^{-1/2}\right\|_F^2 \le C_\rho\|\Sigma_1 - \Sigma_2\|_F^2.$$

Tensorization gives the n-sample bound. □

D.2 Minimax Lower Bound

We use a packing of the $d(d+1)/2$-dimensional symmetric matrix space.

Lemma 12 (Bounded spectral packing)**.** *There exist universal constants $c_0, c_1 > 0$ and symmetric matrices $A_1, \dots, A_M$ such that*

$$\log M \geq c_0 d^2, \qquad \|A_j\|_{\mathrm{op}} \leq 1, \qquad \|A_i - A_j\|_F^2 \geq c_1 d$$

for every $i \neq j$.

A standard Varshamov–Gilbert construction gives such a family. One may start from symmetric Rademacher matrices, normalize them by $\sqrt{d}$, retain a constant-probability subset with uniformly bounded operator norm, and extract a Hamming packing of cardinality $\exp(\Theta(d^2))$. The normalization converts a $\Theta(d^2)$ Hamming separation into a $\Theta(d)$ squared Frobenius separation.

Lower bound in Theorem 3. Let $A_1, \dots, A_M$ be given by Lemma 12 and set

$$\Sigma_j = I + \varepsilon A_j, \qquad \varepsilon = c'_\rho \min\left\{1, \sqrt{\frac{d}{n}}\right\},$$

with $c'_\rho > 0$ sufficiently small. Since $\|A_j\|_{\mathrm{op}} \leq 1$, all Σ_j belong to $\mathcal{C}_d(\rho)$.

By Lemma 10,

$$W_2^2(P_{\Sigma_i}, P_{\Sigma_j}) \geq c_\rho \varepsilon^2 \|A_i - A_j\|_F^2 \geq c_\rho \varepsilon^2 d,$$

so the squared packing radius satisfies

$$\varepsilon^2 d \asymp_\rho \min\left\{d, \frac{d^2}{n}\right\}.$$

Lemma 11, together with $\|A_i - A_j\|_F^2 \leq 4d$, gives

$$D_{\mathrm{KL}}\left(P_{\Sigma_i}^{\otimes n} \,\middle\|\, P_{\Sigma_j}^{\otimes n}\right) \leq C_\rho n \varepsilon^2 d.$$

If $n \geq d$, this is $O(d^2)$; if $n < d$, it is $O(nd) \leq O(d^2)$. Choosing c'_ρ sufficiently small makes the pairwise divergences a fixed small fraction of $\log M$.

Fano's inequality therefore yields a constant probability of confusing the packed targets. The argument applies to estimators taking values in the entire Wasserstein space $\mathcal{P}_2(\mathbb{R}^d)$: assigning any estimate $\widehat{Q}$ to its nearest packed Gaussian target converts a sufficiently accurate distribution estimator into an estimator of the packing index. Hence

$$\inf_{\widehat{Q}} \sup_{\Sigma \in \mathcal{C}_d(\rho)} \mathbb{E}_\Sigma W_2^2(P_\Sigma, \widehat{Q}) \geq c_\rho \min\left\{d, \frac{d^2}{n}\right\}.$$

□

D.3 Matching Upper Bound

Let

$$\widehat{\Sigma} = \frac{1}{n} \sum_{i=1}^n Z_i Z_i^\top$$

be the sample covariance, and let $\widehat{\Sigma}_{\mathcal{C}}$ denote its Frobenius projection onto the closed convex set $\mathcal{C}_d(\rho)$. Since the true Σ belongs to this set,

$$\|\widehat{\Sigma}_{\mathcal{C}} - \Sigma\|_F \leq \|\widehat{\Sigma} - \Sigma\|_F.$$

For Gaussian samples,

$$\mathbb{E}_\Sigma \|\widehat{\Sigma} - \Sigma\|_F^2 = \frac{1}{n}\left[(\operatorname{tr}\Sigma)^2 + \operatorname{tr}(\Sigma^2)\right] \leq C_\rho \frac{d^2}{n}. \tag{31}$$

Using $\widehat{Q} = \mathcal{N}(0, \widehat{\Sigma}_{\mathcal{C}})$ and Lemma 10,

$$\sup_{\Sigma \in \mathcal{C}_d(\rho)} \mathbb{E}_\Sigma W_2^2(P_\Sigma, \widehat{Q}) \le C_\rho \frac{d^2}{n}.$$

For the data-poor regime, the constant estimator $\widehat{Q}_0 = \mathcal{N}(0, I_d)$ satisfies

$$W_2^2(P_\Sigma, \widehat{Q}_0) \le C_\rho \|\Sigma - I_d\|_F^2 \le C_\rho d.$$

Taking the better of these estimators gives

$$\mathfrak{R}_{n,d}(\rho) \le C_\rho \min\left\{ d, \frac{d^2}{n} \right\},$$

which completes the proof of Eq. 7.

If the Gaussian mean is also unknown, estimating it contributes

$$\mathbb{E}\|\widehat{\mu} - \mu\|_2^2 = \frac{\operatorname{tr} \Sigma}{n} = O_\rho\left(\frac{d}{n}\right),$$

which does not change the covariance-dominated minimax order.

D.4 Transport to Image Space and Reversal

For the image-space statement, consider a representation family indexed by $\Sigma \in \mathcal{C}_d(\rho)$ with image distribution $P_{X,\Sigma}$, decoder D_Σ, and induced latent law P_Σ. Define

$$\delta_{\mathrm{rec}}(\Sigma) = W_2(P_{X,\Sigma}, D_{\Sigma\#} P_\Sigma), \qquad \delta_d = \sup_{\Sigma \in \mathcal{C}_d(\rho)} \delta_{\mathrm{rec}}(\Sigma),$$

and

$$\mathfrak{R}_{n,d}^{\mathrm{img}} := \inf_{\widehat{Q}} \sup_{\Sigma \in \mathcal{C}_d(\rho)} \mathbb{E}_\Sigma \left[W_2^2\Big(P_{X,\Sigma}, D_{\Sigma\#} \widehat{Q}(S_n) \Big) \right].$$

Assume that the decoders are uniformly bi-Lipschitz on the relevant supports with constants $0 < m_d \le L_d < \infty$.

For every $\Sigma \in \mathcal{C}_d(\rho)$ and every latent estimator $\widehat{Q}$, Eq. 29 gives

$$\mathcal{L}_{\mathrm{img}} \ge \frac{m_d^2}{2} \mathcal{L}_{\mathrm{lat}} - \delta_{\mathrm{rec}}(\Sigma)^2 \ge \frac{m_d^2}{2} \mathcal{L}_{\mathrm{lat}} - \delta_d^2.$$

Taking expectation, supremum over Σ, and then infimum over latent estimators yields

$$\mathfrak{R}_{n,d}^{\mathrm{img}} \ge \frac{m_d^2}{2} \mathfrak{R}_{n,d}(\rho) - \delta_d^2 \ge \frac{c_\rho m_d^2}{2} \psi_n(d) - \delta_d^2.$$

For the upper bound, use the estimator constructed in Appendix D.3. Again by Eq. 29,

$$\mathcal{L}_{\mathrm{img}} \le 2\delta_d^2 + 2L_d^2 \mathcal{L}_{\mathrm{lat}},$$

and therefore

$$\mathfrak{R}_{n,d}^{\mathrm{img}} \le 2\delta_d^2 + 2C_\rho L_d^2 \psi_n(d).$$

Together these inequalities prove Eq. 8.

For two representation families with $d_1 > d_0$ and $\delta_{d_1} < \delta_{d_0}$, a sufficient condition for strict reversal is

$$\frac{c_\rho m_{d_1}^2}{2} \psi_n(d_1) - \delta_{d_1}^2 > 2\delta_{d_0}^2 + 2C_\rho L_{d_0}^2 \psi_n(d_0). \tag{32}$$

Indeed, the left-hand side is a lower bound on $\mathfrak{R}_{n,d_1}^{\mathrm{img}}$, whereas the right-hand side is an achievable upper bound on $\mathfrak{R}_{n,d_0}^{\mathrm{img}}$. Hence

$$\mathfrak{R}_{n,d_1}^{\mathrm{img}} > \mathfrak{R}_{n,d_0}^{\mathrm{img}},$$

despite $\delta_{d_1} < \delta_{d_0}$. This establishes the reconstruction–generation reversal.

Figure 6 visualizes the sufficient condition in Eq. 32.

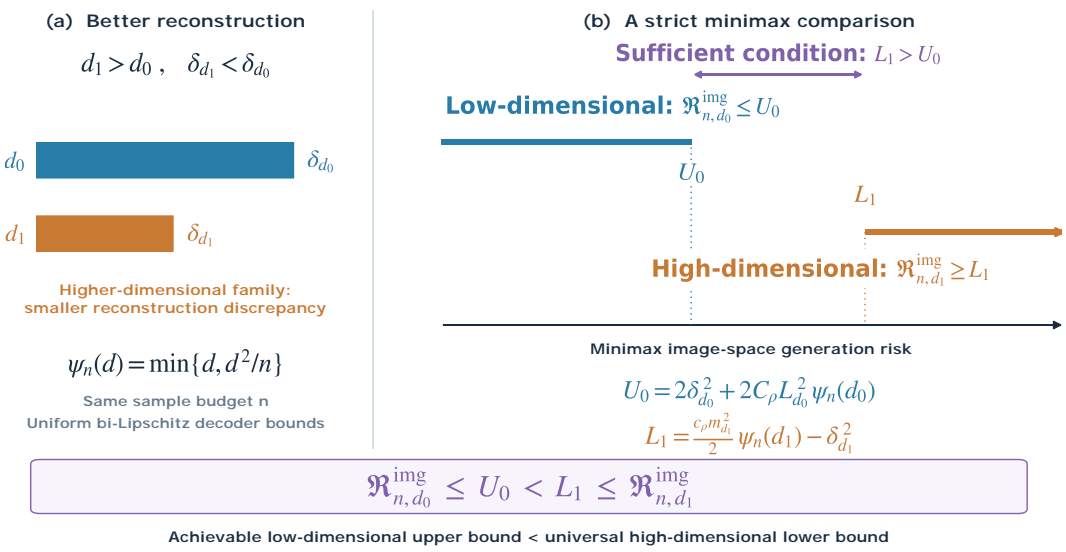


Figure 6: **Why reconstruction–generation reversal is a strict comparison.** The diagram is schematic and not to scale. **(a)** Two representation families satisfy $d_1 > d_0$ and $\delta_{d_1} < \delta_{d_0}$ at the same sample budget n. **(b)** Under the theorem's uniform decoder assumptions, $U_0 = 2\delta^2_{d_0} + 2C_\rho L^2_{d_0} \psi_n(d_0)$ is an achievable upper bound on the low-dimensional minimax image-space risk, while $L_1 = (c_\rho m^2_{d_1}/2)\psi_n(d_1) - \delta^2_{d_1}$ is a lower bound on the high-dimensional minimax risk, with $\psi_n(d) = \min\{d, d^2/n\}$. The sufficient condition $L_1 > U_0$ yields $\mathfrak{R}^{\text{img}}_{n,d_0} \le U_0 < L_1 \le \mathfrak{R}^{\text{img}}_{n,d_1}$ despite the opposite reconstruction ordering. The lower bound permits arbitrary distribution-valued latent estimators in $\mathcal{P}_2(\mathbb{R}^{d_1})$; it is not restricted to Gaussian-output generators.

E Random-Matrix Analysis of Gaussian Score Learning

This appendix derives the deterministic equivalent and low-noise transition in Section 5.2. We use standard deterministic-equivalent results for sample-covariance resolvents and their Stieltjes transforms (Silverstein & Choi, 1995; Bai & Silverstein, 2010); all subsequent expressions are derived explicitly.

E.1 General Spectral Deterministic Equivalent

Let

$$\widehat{A}_\tau = (\widehat{\Sigma} + \tau I)^{-1}, \qquad A_\tau = (\Sigma_d + \tau I)^{-1}.$$

Conditioned on the training samples, $Y_\tau \sim \mathcal{N}(0, \Sigma_d + \tau I)$, so

$$\begin{aligned}\Phi_d(\tau) &= \frac{1}{d} \operatorname{tr}\left[(\widehat{A}_\tau - A_\tau)(\Sigma_d + \tau I)(\widehat{A}_\tau - A_\tau)\right] \\ &= \frac{1}{d} \operatorname{tr}(\widehat{A}_\tau \Sigma_d \widehat{A}_\tau) + \frac{\tau}{d} \operatorname{tr} \widehat{A}^2_\tau - \frac{2}{d} \operatorname{tr} \widehat{A}_\tau + \frac{1}{d} \operatorname{tr} A_\tau. \end{aligned} \tag{33}$$

For $z \in \mathbb{C} \setminus \mathbb{R}_+$, let $Q_d(z) = (\widehat{\Sigma} - zI)^{-1}$. Under the assumptions of Theorem 4, the standard sample-covariance resolvent deterministic equivalent gives, at $z = -\tau$,

$$\begin{aligned}\frac{1}{d} \operatorname{tr} \widehat{A}_\tau &\longrightarrow m(\tau), \\ \frac{1}{d} \operatorname{tr}(\Sigma_d \widehat{A}_\tau) &\longrightarrow a(\tau), \end{aligned} \tag{34}$$

almost surely, where

$$m(\tau) = \int \frac{dH(\lambda)}{\tau + \chi(\tau)\lambda}, \qquad a(\tau) = \int \frac{\lambda\, dH(\lambda)}{\tau + \chi(\tau)\lambda}, \qquad \chi(\tau) = 1 - \gamma + \gamma\tau m(\tau).$$

These quantities satisfy

$$\chi(\tau)a(\tau) + \tau m(\tau) = 1, \tag{35}$$

obtained by integrating $[\chi(\tau)\lambda + \tau]/[\chi(\tau)\lambda + \tau] = 1$.

Because the resolvents are analytic away from the nonnegative real axis and the deterministic equivalents converge locally uniformly there, their derivatives also converge. Hence

$$\frac{1}{d}\operatorname{tr}\widehat{A}_\tau^2 \longrightarrow -m'(\tau), \qquad \frac{1}{d}\operatorname{tr}(\Sigma_d\widehat{A}_\tau^2) \longrightarrow -a'(\tau).$$

Using cyclicity of the trace, $\operatorname{tr}(\widehat{A}_\tau\Sigma_d\widehat{A}_\tau) = \operatorname{tr}(\Sigma_d\widehat{A}_\tau^2)$, while

$$\frac{1}{d}\operatorname{tr}A_\tau \longrightarrow h_H(\tau) = \int \frac{dH(\lambda)}{\lambda+\tau}.$$

Substitution into Eq. 33 proves Eq. 10.

For completeness, define

$$J_k(\tau) = \int \frac{\lambda^k\, dH(\lambda)}{[\tau + \chi(\tau)\lambda]^2}, \qquad k = 0, 1, 2.$$

Differentiating the fixed-point equations gives

$$m' = -\frac{J_0 + \gamma m J_1}{1 + \gamma\tau J_1}, \qquad \chi' = \gamma(m + \tau m'), \qquad a' = -J_1 - \chi' J_2. \tag{36}$$

Thus $\Phi(\gamma, \tau, H)$ is determined by the scalar fixed point in Eq. 11 and one-dimensional spectral integrals.

E.2 Isotropic Limit, Low-Noise Regimes, and Critical Scaling

For $H = \delta_1$, we have $a(\tau) = m(\tau) = q(\gamma, \tau)$, and Eq. 11 reduces to

$$q = \frac{1}{\tau + 1 - \gamma + \gamma\tau q}.$$

Equivalently, $\gamma\tau q^2 + (\tau + 1 - \gamma)q - 1 = 0$, whose positive root is

$$q(\gamma, \tau) = \frac{\sqrt{(\tau + 1 - \gamma)^2 + 4\gamma\tau} - (\tau + 1 - \gamma)}{2\gamma\tau}. \tag{37}$$

Let $u(\gamma, \tau) = -\partial_\tau q(\gamma, \tau)$. Implicit differentiation gives

$$u = \frac{\gamma q^2 + q}{2\gamma\tau q + \tau + 1 - \gamma}, \qquad \Phi_{\text{iso}}(\gamma, \tau) = (1 + \tau)u - 2q + \frac{1}{1+\tau}. \tag{38}$$

We now derive the three low-noise regimes.

For $0 < \gamma < 1$,

$$q(\gamma, \tau) \longrightarrow \frac{1}{1-\gamma}, \qquad u(\gamma, \tau) \longrightarrow \frac{1}{(1-\gamma)^3},$$

and therefore

$$\begin{aligned}\Phi_{\text{iso}}(\gamma, 0^+) &= \frac{1}{(1-\gamma)^3} - \frac{2}{1-\gamma} + 1 \\ &= \frac{\gamma(1 + \gamma - \gamma^2)}{(1-\gamma)^3}.\end{aligned} \tag{39}$$

At the critical ratio $\gamma = 1$, Eq. 37 gives

$$q(1, \tau) = \frac{\sqrt{\tau^2 + 4\tau} - \tau}{2\tau} = \tau^{-1/2} - \frac{1}{2} + O(\tau^{1/2}),$$

and consequently

$$u(1,\tau) = \frac{1}{2\tau^{3/2}} + O(\tau^{-1/2}).$$

Hence

$$\Phi_{\mathrm{iso}}(1,\tau) = \frac{1}{2\tau^{3/2}} + O(\tau^{-1}). \tag{40}$$

For $\gamma > 1$, the Marchenko–Pastur law has a zero-eigenvalue mass $1 - \gamma^{-1}$. Equivalently, Eq. 37 yields

$$q(\gamma,\tau) = \frac{1-\gamma^{-1}}{\tau} + O(1), \qquad u(\gamma,\tau) = \frac{1-\gamma^{-1}}{\tau^2} + O(\tau^{-1}),$$

and therefore

$$\Phi_{\mathrm{iso}}(\gamma,\tau) = \frac{1-\gamma^{-1}}{\tau^2} + O(\tau^{-1}). \tag{41}$$

Equations 39–41 give Eq. 12.

The corresponding denoising excess follows from Tweedie's identity. Conditioned on the training sample,

$$f^\star_\tau(y) = \mathbb{E}[Z \mid Y_\tau = y] = y + \tau s^\star_\tau(y), \qquad \widehat{f}_\tau(y) = y + \tau \widehat{s}_\tau(y),$$

and hence

$$\frac{1}{d}\mathbb{E}\left[\left\|\widehat{f}_\tau(Y_\tau) - f^\star_\tau(Y_\tau)\right\|_2^2 \,\middle|\, Z_1,\dots,Z_n\right] = \tau^2 \Phi_d(\tau).$$

Therefore, after the proportional high-dimensional limit,

$$\lim_{\tau\downarrow 0} \Psi_{\mathrm{iso}}(\gamma,\tau) = \begin{cases} 0, & \gamma \le 1, \\ 1-\gamma^{-1}, & \gamma > 1. \end{cases}$$

For $\gamma > 1$, the surviving floor is precisely the zero-eigenvalue mass of the sample covariance.

Finally, consider the critical window $\gamma = 1 + s\sqrt{\tau}$ with fixed $s \in \mathbb{R}$. Write $t = \sqrt{\tau}$ and $\Delta_s = \sqrt{s^2+4}$. From Eq. 37,

$$q(1+st, t^2) = \frac{s + \Delta_s}{2t} + O(1).$$

Setting $c_s = (s + \Delta_s)/2$ and using Eq. 38,

$$\gamma q^2 + q = \frac{c_s^2}{t^2} + O(t^{-1}), \qquad 2\gamma\tau q + \tau + 1 - \gamma = t\Delta_s + O(t^2),$$

so

$$u(1+st, t^2) = \frac{c_s^2}{\Delta_s t^3} + O(t^{-2}).$$

The remaining terms in $\Phi_{\mathrm{iso}} = (1+\tau)u - 2q + (1+\tau)^{-1}$ are lower order, and therefore

$$\lim_{\tau\downarrow 0} \tau^{3/2}\Phi_{\mathrm{iso}}(1+s\sqrt{\tau},\tau) = \mathcal{F}(s) := \frac{\left(s+\sqrt{s^2+4}\right)^2}{4\sqrt{s^2+4}}. \tag{42}$$

Moreover, $\mathcal{F}(s) \sim |s|^{-3}$ as $s \to -\infty$ and $\mathcal{F}(s) \sim s$ as $s \to +\infty$. Thus the critical law matches the subcritical cubic blow-up and supercritical null-space regime on the two sides of the $|\gamma - 1| = O(\sqrt{\tau})$ window.

The denoising floor and critical scaling are visualized in Figure 7(a,b).

F Population Capacity and Optimization Analysis

This appendix proves Theorem 6 and derives the exact joint capacity–optimization decomposition used to motivate the finite-resource surrogate in the main text. Throughout, let $C_\tau = \Sigma + \tau I$ and write $\Sigma = U\,\mathrm{diag}(\lambda_1,\dots,\lambda_d)U^\top$ with $\lambda_1 \ge \cdots \ge \lambda_d \ge 0$.

F.1 Spectrum-Resolved Capacity

Expanding the population loss gives

$$\mathcal{L}_\tau(B) = \operatorname{tr}(BC_\tau B^\top) - 2\operatorname{tr}(B\Sigma) + \operatorname{tr}\Sigma.$$

Its unique minimizer for $\tau > 0$ is $B^\star = \Sigma C_\tau^{-1}$. Completing the square yields

$$\mathcal{E}_\tau(B) = \frac{1}{d}\left\|(B - B^\star)C_\tau^{1/2}\right\|_F^2. \tag{43}$$

Proof of the capacity part of Theorem 6. Set $M = BC_\tau^{1/2}$. Since $C_\tau^{1/2}$ is invertible, $\operatorname{rank}(M) = \operatorname{rank}(B)$, so Eq. 43 reduces the rank-constrained problem to

$$\inf_{\operatorname{rank}(M)\le r} \|M - M^\star\|_F^2, \qquad M^\star = B^\star C_\tau^{1/2} = \Sigma C_\tau^{-1/2}.$$

In the eigenbasis of Σ,

$$M^\star = U \operatorname{diag}\left(\frac{\lambda_1}{\sqrt{\lambda_1+\tau}}, \dots, \frac{\lambda_d}{\sqrt{\lambda_d+\tau}}\right) U^\top.$$

Its squared singular values are $\omega_\tau(\lambda_i) = \lambda_i^2/(\lambda_i+\tau)$. Since $\omega_\tau(\lambda)$ is increasing for $\lambda \ge 0$, the Eckart–Young theorem retains the r largest covariance eigenmodes and gives

$$\inf_{\operatorname{rank}(B)\le r} [\mathcal{L}_\tau(B) - \mathcal{L}_\tau(B^\star)] = \sum_{i>r} \frac{\lambda_i^2}{\lambda_i+\tau}.$$

Dividing by d proves Eq. 13. □

The modal value satisfies $\omega_\tau(\lambda) \to \lambda$ as $\tau \downarrow 0$ and $\omega_\tau(\lambda) \sim \lambda^2/\tau$ as $\tau \to \infty$. For $\Sigma = I_d$,

$$\Gamma_{\text{cap}}(I_d, r, \tau) = \frac{d-r}{d(1+\tau)} \longrightarrow 1 - \frac{r}{d} \qquad (\tau \downarrow 0).$$

This resembles the statistical low-noise floor $1 - \gamma^{-1} = 1 - n/d$ for $\gamma > 1$ derived in Section 5.2, but the mechanisms differ: the former is imposed by architectural rank, whereas the latter comes from the rank deficiency of the empirical covariance.

F.2 Finite-Time Gradient Dynamics

The Euclidean gradient of the unnormalized population loss is

$$\nabla_B \mathcal{L}_\tau(B) = 2\big[BC_\tau - \Sigma\big].$$

Absorbing the factor 2 into the effective step size gives

$$B_{t+1} = B_t - \eta[B_t C_\tau - \Sigma].$$

Proof of the optimization part of Theorem 6. Since $B_0 = 0$ and both C_τ and Σ are diagonal in the eigenbasis U, every iterate has the form $B_t = U \operatorname{diag}(b_{1,t}, \dots, b_{d,t})U^\top$. Each mode satisfies

$$b_{i,t+1} = b_{i,t} - \eta\left[(\lambda_i+\tau)b_{i,t} - \lambda_i\right].$$

Its fixed point is $b_i^\star = \lambda_i/(\lambda_i+\tau)$, and therefore

$$b_{i,t+1} - b_i^\star = [1 - \eta(\lambda_i+\tau)]\,(b_{i,t} - b_i^\star).$$

With $b_{i,0} = 0$,

$$b_{i,T} - b_i^\star = -b_i^\star\left[1 - \eta(\lambda_i+\tau)\right]^T.$$

Substitution into Eq. 43 gives Eq. 14. Convergence of every mode is guaranteed by $0 < \eta < 2/(\lambda_1+\tau)$. □

For a fixed step size, the worst modal contraction is

$$\rho(\eta) = \max_i |1 - \eta(\lambda_i + \tau)| .$$

Since the effective quadratic spectrum lies in $[\lambda_d + \tau, \lambda_1 + \tau]$, the minimax constant step size equalizes the endpoint contractions. Hence

$$\eta^\star = \frac{2}{\lambda_1 + \lambda_d + 2\tau}, \qquad \rho^\star = \frac{\kappa_\tau - 1}{\kappa_\tau + 1}, \qquad \kappa_\tau = \frac{\lambda_1 + \tau}{\lambda_d + \tau}. \tag{44}$$

Moreover,

$$\Gamma_{\mathrm{opt}}(T) \le (\rho^\star)^{2T} \Gamma_{\mathrm{opt}}(0).$$

Thus achieving $\Gamma_{\mathrm{opt}}(T) \le \epsilon \Gamma_{\mathrm{opt}}(0)$ is guaranteed whenever

$$T \ge \frac{\log(1/\epsilon)}{2 \log\left((\kappa_\tau + 1)/(\kappa_\tau - 1)\right)} = O\left(\kappa_\tau \log \frac{1}{\epsilon}\right)$$

for large κ_τ. In particular, the worst-case spectral contraction requires $T = \Theta(\kappa_\tau)$ iterations for any fixed-factor reduction in the ill-conditioned regime.

F.3 Joint Capacity–Optimization Law

Let $P_r = U_r U_r^\top$ project onto the top-r covariance eigenvectors and restrict the gradient dynamics to this oracle subspace:

$$B_{t+1} = B_t - \eta P_r \left[B_t C_\tau - \Sigma\right] P_r, \qquad B_0 = 0.$$

For $i \le r$, the modal dynamics are those derived above, whereas for $i > r$ the coefficient remains zero. Consequently,

$$\mathcal{E}_\tau(B_T) = \frac{1}{d} \sum_{i > r} \omega_\tau(\lambda_i) + \frac{1}{d} \sum_{i \le r} \omega_\tau(\lambda_i) \left[1 - \eta(\lambda_i + \tau)\right]^{2T} . \tag{45}$$

The first term is the permanent approximation floor induced by finite rank, while the second is the optimization transient on representable modes. This exact additivity relies on the oracle population linear model. When the retained subspace must itself be estimated from finite data, or when capacity is implemented through a nonlinear parameterization, statistical, capacity, and optimization effects generally couple rather than decompose exactly.

Figure 7 contrasts the finite-data behavior of the Gaussian plug-in learner with the capacity and optimization mechanisms of the oracle population model.

G Capacity-Matched Representation Size

This appendix proves Theorem 7 and derives its exact comparative statics and discrete counterpart.

G.1 Continuous Capacity-Matched Optimum

Assume

$$\mathcal{G}(d; n, C, T) = d\, \phi(d/n, C, T).$$

Writing $\gamma = d/n$ and $h(\gamma, C, T) = \phi + \gamma \phi_\gamma$, the RDGL objective becomes

$$\mathcal{J}_{\mathrm{RDGL}}(d) = \mathcal{D}(d) + \alpha d\, \phi(d/n, C, T).$$

Differentiating with respect to d gives

$$\partial_d \mathcal{J}_{\mathrm{RDGL}} = \mathcal{D}'(d) + \alpha h(\gamma, C, T).$$

Hence every interior stationary point satisfies Eq. 16. A second derivative gives

$$\partial_d^2 \mathcal{J}_{\mathrm{RDGL}} = \mathcal{D}''(d) + \frac{\alpha}{n} h_\gamma(\gamma, C, T).$$

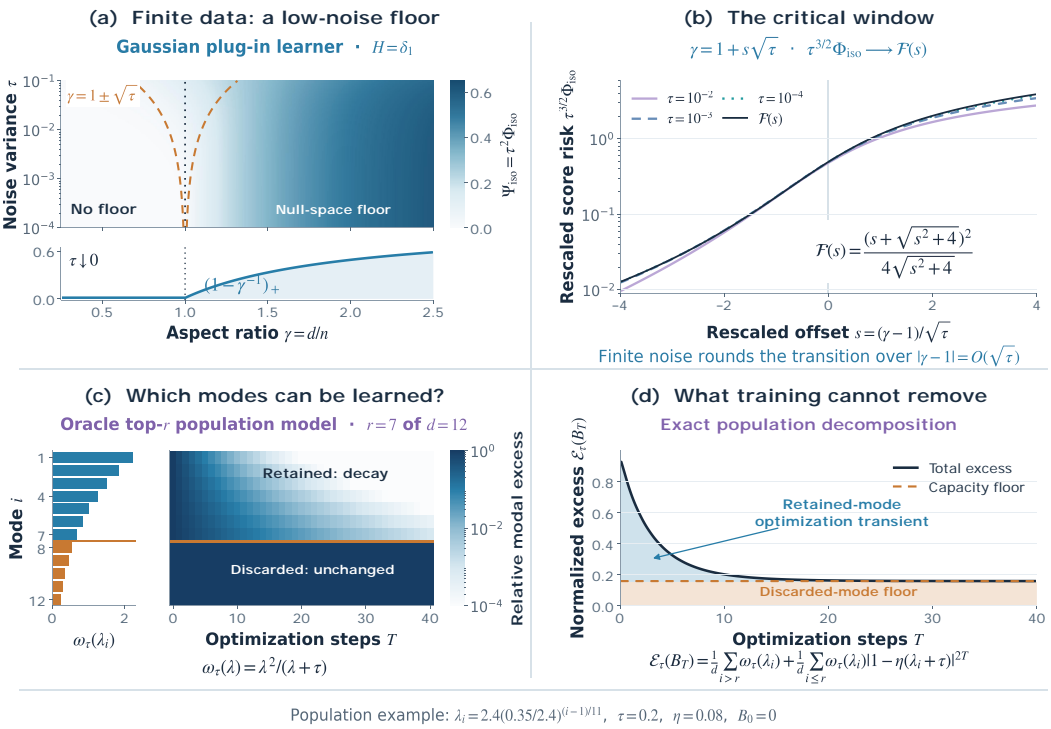


Figure 7: **Spectral signatures of finite resources.** All panels are analytical illustrations, not empirical training results. **(a)** For isotropic Gaussian latents, the heatmap shows the high-dimensional denoising excess $\Psi_{\text{iso}} = \tau^2\Phi_{\text{iso}}$. The dotted line marks $\gamma = 1$; the dashed curves $\gamma = 1 \pm \sqrt{\tau}$ indicate the critical-window scale, rather than exact phase boundaries at finite noise. The lower strip shows the subsequent low-noise limit $(1 - \gamma^{-1})_+$. **(b)** Rescaled analytical score-risk curves approach $\mathcal{F}(s)$ as $\tau \downarrow 0$, with $s = (\gamma - 1)/\sqrt{\tau}$; the vertical axis is logarithmic. **(c)** In an oracle top-r population model, bars show $\omega_\tau(\lambda_i) = \lambda_i^2/(\lambda_i + \tau)$. The heatmap shows each mode's excess divided by its initial value: $|1 - \eta(\lambda_i + \tau)|^{2T}$ for $i \le r$ and 1 otherwise. Colors use a logarithmic scale, with values below 10^{-4} clipped to the lightest color. The horizontal line separates retained and discarded modes. **(d)** The same population model yields an exact sum of a discarded-mode capacity floor and a decaying retained-mode transient. Panels (c,d) use $d = 12$, $r = 7$, $B_0 = 0$, $\lambda_i = 2.4(0.35/2.4)^{(i-1)/11}$, $\tau = 0.2$, and $\eta = 0.08$. The finite-data and population analyses are shown separately; no general additive decomposition across all three resources is assumed.

The positivity condition in Theorem 7 therefore makes the objective strictly convex on the interval, so an interior stationary point is unique and globally optimal there.

For the comparative statics, define

$$F(d, n, C, T, \alpha) = \mathcal{D}'(d) + \alpha h(d/n, C, T).$$

At the optimum, $F(d^\star, n, C, T, \alpha) = 0$ and

$$F_d = \mathcal{D}''(d^\star) + \frac{\alpha}{n} h_\gamma = \Xi^\star > 0.$$

Holding d fixed when differentiating the remaining arguments gives

$$F_n = -\alpha h_\gamma \frac{d}{n^2}, \qquad F_C = \alpha h_C, \qquad F_T = \alpha h_T, \qquad F_\alpha = h.$$

The implicit-function theorem therefore yields

$$\frac{\partial d^\star}{\partial n} = \frac{\alpha h_\gamma d^\star/n^2}{\Xi^\star}, \qquad \frac{\partial d^\star}{\partial C} = -\frac{\alpha h_C}{\Xi^\star}, \qquad \frac{\partial d^\star}{\partial T} = -\frac{\alpha h_T}{\Xi^\star}, \qquad \frac{\partial d^\star}{\partial \alpha} = -\frac{h}{\Xi^\star}, \tag{46}$$

with the derivatives of h evaluated at $(d^\star/n, C, T)$.

Thus $h_\gamma > 0$, $h_C < 0$, $h_T < 0$, and $h > 0$ imply that the optimal representation grows with sample size, downstream model capacity, and optimization budget, while a larger generative penalty favors stronger compression. These assumptions concern the calibrated tax $\mathcal{G}$ and do not require the raw statistical, capacity, and optimization surrogates in Section 5 to add exactly.

G.2 Discrete Representation Sizes

For integer representation size, define

$$\Delta_{\mathrm{rec}}(d) = \mathcal{D}(d) - \mathcal{D}(d+1), \qquad \Delta_{\mathrm{gen}}(d) = \mathcal{G}(d+1; n, C, T) - \mathcal{G}(d; n, C, T).$$

Since

$$\mathcal{J}_{\mathrm{RDGL}}(d+1) - \mathcal{J}_{\mathrm{RDGL}}(d) = -\Delta_{\mathrm{rec}}(d) + \alpha\Delta_{\mathrm{gen}}(d),$$

one has the exact marginal rule

$$\mathcal{J}_{\mathrm{RDGL}}(d+1) < \mathcal{J}_{\mathrm{RDGL}}(d) \quad \Longleftrightarrow \quad \Delta_{\mathrm{rec}}(d) > \alpha\Delta_{\mathrm{gen}}(d). \tag{47}$$

If $\Delta_{\mathrm{rec}}(d)$ is nonincreasing and $\Delta_{\mathrm{gen}}(d)$ is nondecreasing, then

$$s_d := \Delta_{\mathrm{rec}}(d) - \alpha\Delta_{\mathrm{gen}}(d)$$

is nonincreasing, while $\mathcal{J}_{\mathrm{RDGL}}(d+1) - \mathcal{J}_{\mathrm{RDGL}}(d) = -s_d$ is nondecreasing. The objective therefore decreases until the marginal inequality in Eq. 47 reverses.

For admissible sizes $d \in \{d_{\min}, \dots, d_{\max}\}$, if the crossing is interior and no equality occurs, the unique minimizer is

$$d^\star = 1 + \max\left\{d \in \{d_{\min}, \dots, d_{\max} - 1\} : \Delta_{\mathrm{rec}}(d) > \alpha\Delta_{\mathrm{gen}}(d)\right\}.$$

If no increment is beneficial then $d^\star = d_{\min}$; if every increment is beneficial then $d^\star = d_{\max}$. This is the discrete counterpart of Eq. 16.

H Geometry-Aware Representation Allocation

This appendix derives the local reconstruction value and generative burden in Eq. 17, proves Theorem 8, and develops the corresponding hard-budget and basis-independent formulations.

H.1 Local Decoder Geometry and Generative Burden

Let $\mu_Z = \mathbb{E}Z$, let v_i be a covariance eigenvector, and define $z_i = v_i^\top(Z - \mu_Z)$. Attenuating the corresponding centered mode gives

$$Z_{\epsilon,i} = Z - \epsilon z_i v_i.$$

Assume that D is continuously differentiable in a neighborhood of the support of P_Z and that the Taylor remainder is dominated by an integrable random variable. Then

$$D(Z_{\epsilon,i}) = D(Z) - \epsilon z_i J_D(Z) v_i + o(\epsilon|z_i|),$$

and hence

$$\|D(Z) - D(Z_{\epsilon,i})\|_2^2 = \epsilon^2 z_i^2 \|J_D(Z)v_i\|_2^2 + o(\epsilon^2 z_i^2).$$

Taking expectations gives

$$\mathbb{E}\|D(Z) - D(Z_{\epsilon,i})\|_2^2 = \epsilon^2 b_i + o(\epsilon^2),$$

with b_i as defined in Eq. 17.

Writing

$$w_i := \mathbb{E}\|J_D(Z)v_i\|_2^2,$$

one obtains

$$b_i = \lambda_i w_i$$

whenever z_i^2 decouples from the local decoder sensitivity; this is exact for a linear decoder. For nonlinear decoders, b_i retains the correlation between modal amplitude and decoder sensitivity and is therefore the more intrinsic quantity.

Equivalently, with the decoder pullback metric

$$G_D(Z) = J_D(Z)^\top J_D(Z),$$

the reconstruction value can be written as

$$b_i = \mathbb{E}\left[z_i^2\, v_i^\top G_D(Z) v_i\right].$$

Thus equal covariance eigenvalues need not imply equal image-space reconstruction value.

We next make the finite-resource burden explicit. For $e_\tau = \widehat{f}_\tau(Y_\tau) - f_\tau^\star(Y_\tau)$, define

$$c_i(\tau; n, C, T) = \mathbb{E}\left[(v_i^\top e_\tau)^2 \|J_D(Z) v_i\|_2^2\right], \qquad c_i(n, C, T) = \int c_i(\tau; n, C, T)\, d\nu(\tau).$$

Under a local decorrelation approximation,

$$c_i(\tau; n, C, T) \approx w_i\, \widetilde{c}_i(\tau; n, C, T), \qquad \widetilde{c}_i(\tau; n, C, T) := \mathbb{E}\left[(v_i^\top e_\tau)^2\right].$$

For the oracle rank-r, T-step Gaussian surrogate of Section 5.3,

$$\widetilde{c}_i(\tau; r, T) = \omega_\tau(\lambda_i) \begin{cases} |1 - \eta(\lambda_i + \tau)|^{2T}, & i \le r, \\ 1, & i > r. \end{cases}$$

Thus discarded directions incur their full modal denoising cost, whereas retained modes carry only the finite-time optimization transient.

For the empirical-covariance plug-in learner of Section 5.2,

$$\frac{1}{d}\sum_{i=1}^{d} \widetilde{c}_i(\tau) = \mathbb{E}_{S_n}\left[\Psi_d(\tau)\right], \qquad \Psi_d(\tau) = \tau^2 \Phi_d(\tau).$$

The RMT law therefore supplies the aggregate finite-sample burden, while the population surrogate above resolves capacity and optimization mode by mode.

H.2 Mode-Wise Allocation Laws

For binary retention variables $u_i \in \{0, 1\}$, the local RDGL surrogate is

$$\mathcal{J}_{\mathrm{alloc}}(u) = \mathcal{J}_0 + \sum_{i=1}^{d} \left[(1 - u_i) b_i + \alpha u_i c_i\right]. \tag{48}$$

Proof of the mode-wise statement in Theorem 8. The contribution of mode i to Eq. 48 is b_i when $u_i = 0$ and αc_i when $u_i = 1$. Since the objective is separable, retaining the mode is optimal exactly when $\alpha c_i < b_i$, which proves the mode-wise statement of Theorem 8. □

A hard generative budget leads to a related fractional relaxation:

$$\max_{0 \le u_i \le 1} \sum_i b_i u_i \qquad \text{s.t.} \qquad \sum_i c_i u_i \le B. \tag{49}$$

Introducing a multiplier $\lambda \ge 0$ gives

$$\mathcal{L}(u, \lambda) = \sum_i (b_i - \lambda c_i) u_i + \lambda B.$$

For fixed λ, each coordinate is optimized independently. At an active budget, complementary slackness therefore yields

$$u_i^\star = \begin{cases} 1, & b_i / c_i > \lambda^\star, \\ 0, & b_i / c_i < \lambda^\star, \end{cases} \tag{50}$$

with fractional retention possible only on the boundary $b_i / c_i = \lambda^\star$. If the ratios are distinct, at most one mode is fractional.

For strictly binary variables and a hard budget, Eq. 49 becomes a 0–1 knapsack problem, so a universal ordering by b_i / c_i is not guaranteed. The mode-wise threshold in Theorem 8 remains exact for the unconstrained Lagrangian surrogate Eq. 48 and for the fractional relaxation above.

H.3 Generalized Reconstruction–Generation Geometry

The covariance eigenbasis is convenient for the Gaussian spectral laws, but the decoder-induced reconstruction geometry need not commute with the corresponding generative-cost geometry. Let $\mathsf{R} \succeq 0$ denote a quadratic reconstruction-value operator and $\mathsf{C} \succ 0$ a quadratic generative-cost operator on the same latent coordinate system. For a direction v,

$$\mathcal{U}(v) = \frac{v^\top \mathsf{R} v}{v^\top \mathsf{C} v}$$

is the generalized reconstruction value per unit generative burden.

For a k-dimensional retained subspace, consider

$$\max_{U^\top \mathsf{C} U = I_k} \operatorname{tr}(U^\top \mathsf{R} U).$$

Proof of the generalized statement in Theorem 8. Let $V = \mathsf{C}^{1/2} U$. Then $U^\top \mathsf{C} U = I_k$ is equivalent to $V^\top V = I_k$, while

$$\operatorname{tr}(U^\top \mathsf{R} U) = \operatorname{tr}\left[V^\top \mathsf{C}^{-1/2} \mathsf{R} \mathsf{C}^{-1/2} V\right].$$

By the Ky Fan maximum principle, the optimum is attained by the top k eigenvectors of $\mathsf{C}^{-1/2}\mathsf{R}\mathsf{C}^{-1/2}$. Mapping them back through $\mathsf{C}^{-1/2}$ gives generalized eigenvectors satisfying $Rv = \mu C v$, as stated in Theorem 8. □

When $\mathsf{R} = \operatorname{diag}(b_1, \dots, b_d)$ and $\mathsf{C} = \operatorname{diag}(c_1, \dots, c_d)$ in the same basis, the generalized eigenvalues are $\mu_i = b_i / c_i$. Thus both formulations are governed by the same reconstruction-to-generation utility ratios: the mode-wise Lagrangian rule in Theorem 8 retains modes above the threshold α, whereas the fixed-k generalized spectral problem retains the k largest ratios.

A natural mean-metric approximation is

$$\overline{G}_D = \mathbb{E}[G_D(Z)].$$

If $\Sigma \succ 0$, whitening the centered latent as $\xi = \Sigma^{-1/2}(Z - \mu_Z)$ gives the reconstruction operator

$$\mathsf{R} = \Sigma^{1/2} \overline{G}_D \Sigma^{1/2}.$$

The generative operator C can analogously be constructed from the schedule-aggregated finite-resource denoising burden. If an estimated C is only positive semidefinite, one may use $\mathsf{C}_\varepsilon = \mathsf{C} + \varepsilon I$ and interpret the construction as $\varepsilon \downarrow 0$.

Figure 8 connects the global size principle in Appendix G to the mode-wise and generalized allocation rules developed here.

Section 6.2 turns these capacity-matching principles into a computable spectral tokenizer objective.

I Capacity-Matched Spectral Tokenization

This appendix provides the scale calibration, image-space interpretation, constrained optimization, implementation details, and architectural extension underlying CMST in Section 6.2.

I.1 Canonical Scale and Gauge Degeneracy

The normalization in Eq. 18 is necessary because reconstruction alone does not fix latent scale. Consider the global gauge $E_a = aE$ and $D_a(z) = D(z/a)$. Reconstruction is unchanged, whereas a covariance eigenvalue transforms as $\lambda_i \mapsto a^2 \lambda_i$. The population modal cost becomes

$$\omega_\tau(a^2 \lambda_i) = \frac{a^4 \lambda_i^2}{a^2 \lambda_i + \tau} \longrightarrow 0 \qquad (a \downarrow 0).$$

A covariance-based spectral penalty can therefore be reduced by global rescaling without altering the reconstructed image.

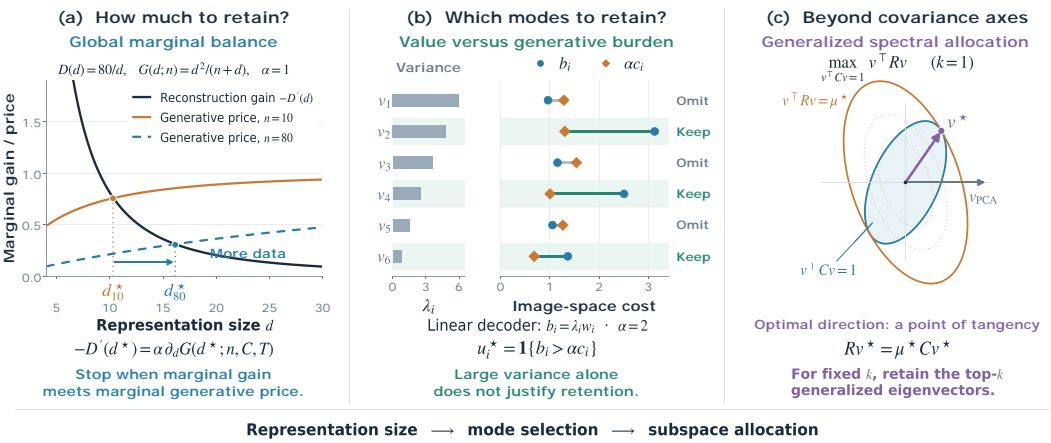


Figure 8: **How much representation to retain, and which directions.** All panels are analytical illustrations. **(a)** For the smooth proxy in Appendix J.3, $D(d) = 80/d$, $G(d;n) = d^2/(n+d)$, and $\alpha = 1$, intersections of marginal reconstruction gain and generative price determine the unique optimal sizes. Increasing n from 10 to 80 moves $d^\star$ from approximately 10.28 to 16.13. **(b)** An illustrative linear-decoder example orders six modes by decreasing variance. Blue circles show reconstruction values $b_i = \lambda_i w_i$; orange diamonds show weighted generative burdens αc_i, with $\alpha = 2$. The separable penalized surrogate retains modes $2, 4, 6$, demonstrating that retention need not follow variance order. This is the mode-wise Lagrangian rule, rather than a solution to a binary hard-budget knapsack problem. **(c)** In a two-dimensional example, the blue ellipse is $v^\top Cv = 1$ and orange ellipses are reconstruction-value contours. The highest attainable contour touches the cost ellipse at $\pm v^\star$, where $Rv^\star = \mu^\star Cv^\star$. The optimal direction differs from the leading covariance direction v_{PCA}. For fixed k, the corresponding subspace is spanned by the top-k generalized eigenvectors under $U^\top CU = I_k$.

For $\mu_Z = \mathbb{E}Z$ and $s_Z^2 = d^{-1}\operatorname{tr}\Sigma_Z$, define

$$\widetilde{Z} = \frac{Z - \mu_Z}{s_Z}, \qquad \widetilde{D}(u) = D(\mu_Z + s_Z u).$$

Then $D(Z) = \widetilde{D}(\widetilde{Z})$ and $d^{-1}\operatorname{tr}\operatorname{Cov}(\widetilde{Z}) = 1$.

The normalization removes translation and global-scale degeneracies but intentionally retains anisotropy, since anisotropic geometry changes the difficulty of isotropic diffusion. Decoder sensitivity also prevents aligned anisotropic rescaling from being treated as free. Under a diagonal rescaling aligned with a latent mode, λ_i and $w_i = \mathbb{E}\|J_D(Z)v_i\|_2^2$ transform inversely, so the factorized low-noise quantity $\lambda_i w_i$ is invariant. The exact local reconstruction value b_i in Eq. 17 has the corresponding invariance.

I.2 Calibrating Generative Taxes to Image-Space Units

Let

$$G_{\widetilde{D}}(z) = J_{\widetilde{D}}(z)^\top J_{\widetilde{D}}(z)$$

be the decoder pullback metric. For a small latent perturbation e,

$$\|\widetilde{D}(z+e) - \widetilde{D}(z)\|_2^2 = e^\top G_{\widetilde{D}}(z)e + o(\|e\|_2^2).$$

Thus latent denoising error must be measured through decoder geometry to obtain a local image-space distortion.

For a covariance eigenvector v_i, define

$$w_i = \mathbb{E}\left[v_i^\top G_{\widetilde{D}}(\widetilde{Z})v_i\right] = \mathbb{E}\|J_{\widetilde{D}}(\widetilde{Z})v_i\|_2^2.$$

When modal denoising error is approximately decorrelated from local decoder sensitivity, a latent modal MSE $\widetilde{c}_i$ contributes approximately $w_i\widetilde{c}_i$ image-space MSE. This is the local calibration used in Eq. 17 and Appendix H.1.

The finite-capacity and optimization laws are mode resolved. They therefore give

$$\Gamma^{\mathrm{img}}_{\mathrm{cap}}(\tau) = \frac{1}{d} \sum_{i>r_C} w_i \omega_\tau(\lambda_i)$$

and

$$\Gamma^{\mathrm{img}}_{\mathrm{opt}}(\tau) = \frac{1}{d} \sum_{i\le r_C} w_i \omega_\tau(\lambda_i) \left|1-\eta(\lambda_i+\tau)\right|^{2T}.$$

These are the decoder-weighted versions of the two terms in Eq. 45.

The RMT result in Section 5.2 instead controls aggregate normalized denoising excess,

$$\Psi(\gamma,\tau,H) = \tau^2 \Phi(\gamma,\tau,H).$$

In CMST, the asymptotic functional $\Phi(\gamma,\tau,H)$ is evaluated at $\gamma = d/n$ and the current empirical spectral measure $H_{\widetilde{\Sigma}}$. This is a plug-in use of the deterministic equivalent as a finite-dimensional design surrogate, rather than an identity for the exact finite-d score risk. We therefore use the rotationally averaged decoder gain

$$\bar{w} = \frac{1}{d}\sum_i w_i = \frac{1}{d}\mathbb{E}\|J_{\widetilde{D}}(\widetilde{Z})\|_F^2$$

to obtain

$$\Gamma^{\mathrm{img}}_{\mathrm{stat}}(\tau) = \bar{w}\,\tau^2 \Phi(\gamma,\tau,H).$$

This calibration is exact when the denoising-error covariance is isotropic and independent of the local metric orientation; otherwise it is a rotationally averaged approximation.

More generally, if

$$m_D^2 I \preceq G_{\widetilde{D}}(z) \preceq L_D^2 I,$$

then every latent error e satisfies

$$m_D^2 \mathbb{E}\|e\|_2^2 \le \mathbb{E}\left[e^\top G_{\widetilde{D}}(\widetilde{Z}) e\right] \le L_D^2 \mathbb{E}\|e\|_2^2.$$

Hence the latent finite-resource laws induce corresponding local image-space bounds even without rotational averaging.

Integrating the calibrated statistical, capacity, and optimization terms over the diffusion schedule gives Eq. 19. The capacity and optimization terms inherit the exact oracle population decomposition from Eq. 45; adding the finite-sample RMT term produces the theory-guided CMST surrogate once statistical and representational effects interact.

I.3 Constrained Optimization and Spectral Gates

The constrained objective in Eq. 20 has Lagrangian

$$\mathscr{L}(\theta,\psi,g,\lambda) = \mathcal{L}_{\mathrm{rec}}(\theta,\psi,g) + \lambda\left[\widehat{\Gamma}_{\mathrm{CMST}}(\theta,\psi,g) - B_{\mathrm{gen}}\right], \qquad \lambda \ge 0,$$

where (θ,ψ) parameterize the encoder and decoder and $g \in [0,1]^d$ denotes the spectral gates. We restrict attention to nondegenerate gate configurations with $s_g > 0$, so that the trace-normalized gated representation in Eq. 21 is well defined.

For the gated representation in Eq. 21, let

$$\mathcal{S}_C(g) = \mathrm{Top}_{r_C}\left\{\lambda_i(g)\right\}_{i=1}^d$$

denote the indices of the r_C largest gated eigenvalues. The capacity term in Eq. 19 is then evaluated over $i \notin \mathcal{S}_C(g)$ and the optimization term over $i \in \mathcal{S}_C(g)$.

This choice follows the covariance-rank capacity surrogate of Theorem 6, whose oracle rank-r_C solution retains the leading covariance modes. It should be distinguished from the more general geometry-aware allocation problem in Theorem 8: the latter optimizes an abstract retained subspace using reconstruction-to-generation utility, whereas CMST instantiates the specific spectral rank surrogate used to derive its capacity and optimization taxes.

At a spectral refresh, the decoder sensitivities are reevaluated as

$$w_i(g) = \mathbb{E}\left[\|J_{\widetilde{D}}(\widetilde{Z}_g)v_i\|_2^2\right].$$

Between refreshes, the basis V and sensitivities w_i are treated as fixed local statistics. Away from spectral ties, $\mathcal{S}_C(g)$ is locally constant, so the gated tax is differentiable with respect to the tokenizer outputs and gates within each such region.

At a regular constrained optimum, the KKT conditions include

$$\lambda^\star\left[\widehat{\Gamma}_{\mathrm{CMST}} - B_{\mathrm{gen}}\right] = 0, \qquad \widehat{\Gamma}_{\mathrm{CMST}} \le B_{\mathrm{gen}}, \qquad \lambda^\star \ge 0.$$

For every interior gate $0 < g_i^\star < 1$,

$$-\frac{\partial \mathcal{L}_{\mathrm{rec}}}{\partial g_i} = \lambda^\star \frac{\partial \widehat{\Gamma}_{\mathrm{CMST}}}{\partial g_i}. \tag{51}$$

Boundary gates satisfy the corresponding one-sided inequalities. Equation 51 is the continuous spectral analogue of the mode-wise rule in Theorem 8: a direction is expanded until its marginal reconstruction benefit balances the marginal generative price imposed by the target downstream learner.

A practical implementation need not differentiate through every spectral refresh. The estimated basis and decoder sensitivities can be treated as stop-gradient quantities between refreshes, while the projected variances

$$\lambda_i = v_i^\top \widehat{\Sigma} v_i$$

and gated eigenvalues in Eq. 21 remain differentiable through the tokenizer outputs and gates.

Algorithm 1 summarizes the resulting primal–dual procedure. The gate constraint can be enforced by projected updates or by a bounded parameterization; throughout we maintain $s_g > 0$ so that the trace normalization is well defined.

Algorithm 1 Capacity-Matched Spectral Tokenization

Require: $(E_\theta, \widetilde{D}_\psi)$, gates $\mathbf{g}$, dual λ_{dual}, budget $\mathcal{B} = (n, r_C, T, \eta, \nu)$, constraint B_{gen}, batch size M, refresh interval K_{spec}, and learning rates $\eta_{\mathrm{tok}}, \eta_{\mathrm{dual}}$

for $t = 1, 2, \ldots$ **do**
  $Z \leftarrow E_\theta(X), \quad \widetilde{Z} \leftarrow \textsc{Canon}(Z)$
  if $t = 1$ or $t \bmod K_{\mathrm{spec}} = 0$ **then**
    $\widehat{\Sigma} \leftarrow M^{-1}\widetilde{Z}^\top \widetilde{Z}$
    $V \leftarrow \textsc{StopGrad}\Big(\textsc{EigVecs}(\widehat{\Sigma})\Big)$
  end if
  $\lambda_i \leftarrow M^{-1}\|\widetilde{Z}v_i\|_2^2, \quad i = 1, \ldots, d$
  $s_g^2 \leftarrow d^{-1}\sum_i g_i^2 \lambda_i$
  $\widetilde{Z}_g \leftarrow s_g^{-1}\widetilde{Z}V\,\mathrm{diag}(\mathbf{g})V^\top$
  $\lambda_i^{(g)} \leftarrow g_i^2\lambda_i / s_g^2$
  $\mathcal{S}_C \leftarrow \textsc{Top}_{r_C}(\boldsymbol{\lambda}^{(g)})$
  if $t = 1$ or $t \bmod K_{\mathrm{spec}} = 0$ **then**
    $w_i \leftarrow \textsc{StopGrad}\Big(M^{-1}\sum_{m=1}^M \|J_{\widetilde{D}_\psi}(\widetilde{z}_g^{(m)})v_i\|_2^2\Big)$ ▷ JVP
  end if
  $\widehat{\Gamma} \leftarrow \textsc{CMSTTax}(\{\lambda_i^{(g)}, w_i\}_{i=1}^d, \mathcal{S}_C; \mathcal{B})$
  $\mathcal{L} \leftarrow \mathcal{L}_{\mathrm{rec}}(X, \widetilde{D}_\psi(\widetilde{Z}_g)) + \lambda_{\mathrm{dual}}(\widehat{\Gamma} - B_{\mathrm{gen}})$
  $(\theta, \psi) \leftarrow (\theta, \psi) - \eta_{\mathrm{tok}}\nabla_{\theta,\psi}\mathcal{L}$
  $\mathbf{g} \leftarrow \Pi_{[0,1]^d}\left(\mathbf{g} - \eta_{\mathrm{tok}}\nabla_{\mathbf{g}}\mathcal{L}\right)$
  $\lambda_{\mathrm{dual}} \leftarrow \Big[\lambda_{\mathrm{dual}} + \eta_{\mathrm{dual}}(\widehat{\Gamma} - B_{\mathrm{gen}})\Big]_+$
end for

Figure 9 summarizes the computational dependencies and optimization loop of Algorithm 1.

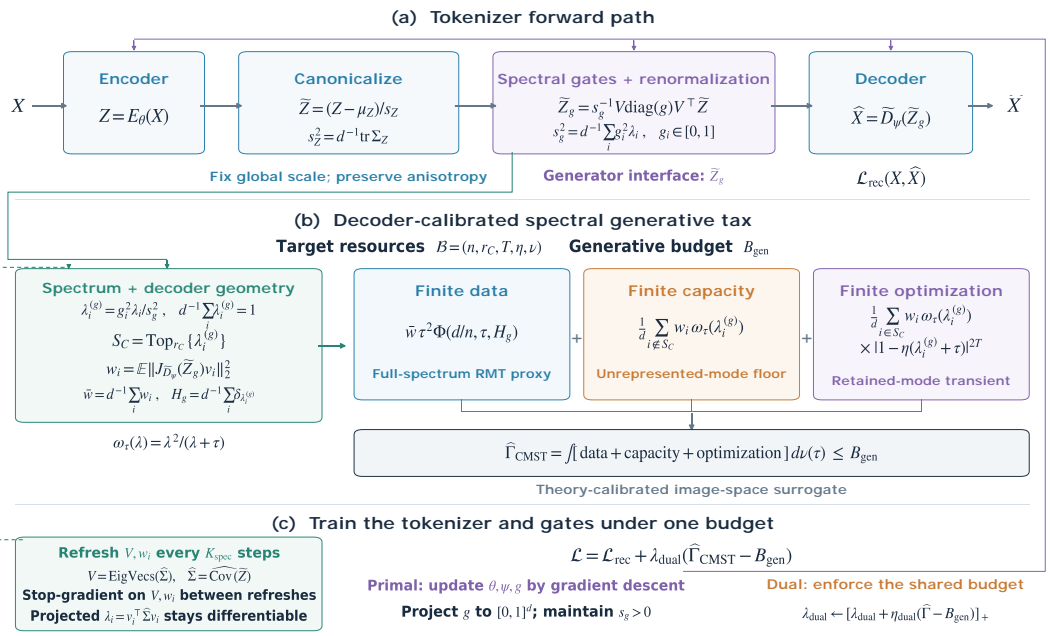


Figure 9: **CMST: from resource specifications to a trainable spectral constraint.** **(a)** The encoder output is centered and globally scale-normalized, then spectrally gated and renormalized. The decoder and downstream generator use the same gated latent interface $\widetilde{Z}_g$. **(b)** Given $\mathcal{B} = (n, r_C, T, \eta, \nu)$, CMST evaluates three spectral taxes using the gated eigenvalues $\lambda_i^{(g)}$. Here $H_g = d^{-1}\sum_i \delta_{\lambda_i^{(g)}}$ and S_C contains the indices of the r_C largest gated eigenvalues. The statistical term uses the averaged decoder sensitivity $\bar{w}$; capacity and optimization retain the modal weights w_i. Integration over noise scales gives one image-space design surrogate $\widehat{\Gamma}_{\mathrm{CMST}}$, constrained by B_{gen}. The three terms need not decompose the exact risk additively. **(c)** Primal updates train the encoder, decoder, and gates; one nonnegative dual multiplier enforces the shared budget. The basis V and decoder sensitivities w_i are periodically refreshed and treated as stop-gradient statistics between refreshes, while projected variances and gated eigenvalues remain differentiable. Decoder sensitivities can be estimated through JVPs. Purple feedback denotes primal updates; the dashed green route denotes periodic spectral refresh. In the dual update, $\widehat{\Gamma}$ abbreviates $\widehat{\Gamma}_{\mathrm{CMST}}$.

I.4 Estimators and Computational Cost

In implementation, the centering and trace-normalization statistics in Eq. 18 can be estimated from the spectral-refresh batch or maintained as running statistics.

Suppose a spectral refresh uses M centered latent vectors in $\mathbb{R}^d$, arranged as $Z \in \mathbb{R}^{M \times d}$. Explicit covariance construction

$$\widehat{\Sigma} = \frac{1}{M} Z^\top Z$$

requires $O(Md^2)$ arithmetic and $O(d^2)$ covariance storage, followed by $O(d^3)$ for a full eigendecomposition. Once the spectrum is available, evaluating Eq. 19 at K quadrature or sampled noise scales costs $O(Kd)$.

Decoder sensitivities do not require a dense Jacobian. Each

$$w_i = \mathbb{E}\|J_{\widetilde{D}}(\widetilde{Z}) v_i\|_2^2$$

can be estimated with a Jacobian–vector product. Tracking k directions therefore requires approximately k directional decoder evaluations per spectral refresh; these can be batched and refreshed substantially less frequently than ordinary tokenizer gradients.

When d is too large to materialize $\widehat{\Sigma}$, covariance–vector products are sufficient:

$$\widehat{\Sigma}v = \frac{1}{M} Z^\top (Zv),$$

with $O(Md)$ cost per product. With k probes and m Lanczos iterations, stochastic Lanczos quadrature can estimate quantities of the form $\operatorname{tr} f(\widehat{\Sigma})$ in $O(kmMd)$ time without storing a $d \times d$ covariance matrix. Randomized or block Lanczos can similarly track only the leading spectral subspace required by the capacity term. CMST can therefore interpolate between an exact channel-level implementation and a matrix-free approximation for very large latent spaces.

I.5 Detail Offloading as an Architectural Corollary

The allocation law distinguishes generative burden from reconstruction value. A direction can be simultaneously expensive for the main generator and important for reconstruction; such information need not be discarded.

Partition the representation as $Z = (Z_{\mathrm{gen}}, Z_{\mathrm{detail}})$, let a main generator model Z_{gen}, and let $K(dz_{\mathrm{detail}} \mid z_{\mathrm{gen}})$ denote a conditional refinement kernel.

Corollary 13 (Exact detail offloading)**.** *If* $Q_{\mathrm{gen}} = P_{Z_{\mathrm{gen}}}$ *and*

$$K(\cdot \mid z_{\mathrm{gen}}) = P_{Z_{\mathrm{detail}} | Z_{\mathrm{gen}} = z_{\mathrm{gen}}}$$

almost surely, then

$$Q_{\mathrm{gen}}(dz_{\mathrm{gen}}) K(dz_{\mathrm{detail}} \mid z_{\mathrm{gen}}) = P_Z(dz_{\mathrm{gen}}, dz_{\mathrm{detail}}).$$

Proof. The statement follows directly from the disintegration

$$P_Z(dz_{\mathrm{gen}}, dz_{\mathrm{detail}}) = P_{Z_{\mathrm{gen}}}(dz_{\mathrm{gen}}) P_{Z_{\mathrm{detail}} | Z_{\mathrm{gen}}}(dz_{\mathrm{detail}} \mid z_{\mathrm{gen}}).$$

□

Hence high-value, high-cost information may be moved from the globally modeled latent into a conditional refinement path. Its cost is transferred rather than removed: the main generator becomes simpler while the refinement mechanism assumes the conditional modeling burden.

J Concrete Instantiations and Operational Interpretation

This appendix instantiates the abstract quantities in the main text in several representative regimes. The purpose is interpretive rather than to introduce new theory: we show how the CMST tax simplifies under limiting resources, evaluate the spectral quantities for common covariance profiles, work out an explicit capacity-matched representation-size example, and summarize how the theoretical objects can be estimated in practice.

J.1 Limiting Regimes of the CMST Objective

We first record several reductions of Eq. 19. These limits isolate the role of each resource axis and provide useful sanity checks for the full objective.

Infinite-data limit. Let $\gamma = d/n \downarrow 0$ while the limiting spectral law H and $\tau > 0$ remain fixed. In the fixed point of Eq. 11,

$$\chi(\tau) \longrightarrow 1, \qquad m(\tau) \longrightarrow h_H(\tau) = \int \frac{dH(\lambda)}{\lambda + \tau}.$$

Moreover,

$$a(\tau) = \int \frac{\lambda \, dH(\lambda)}{\lambda + \tau} = 1 - \tau m(\tau),$$

so $a' = -m - \tau m'$. Substitution into Eq. 10 gives

$$\Phi(0, \tau, H) = -a' - \tau m' - 2m + h_H = 0.$$

Thus the finite-data component of CMST vanishes as $n/d \to \infty$, leaving only capacity and optimization costs.

Full-capacity and fully optimized limits. If $r_C = d$, the finite-capacity term in Eq. 19 is identically zero. If, in addition,

$$|1 - \eta(\lambda_i + \tau)| < 1$$

for every represented mode and for ν-almost every τ, then the finite-optimization term vanishes as $T \to \infty$. Consequently,

$$r_C = d, \quad T \to \infty \qquad \Longrightarrow \qquad \widehat{\Gamma}_{\mathrm{CMST}} = \int \bar{w}\, \tau^2 \Phi\left(\frac{d}{n}, \tau, H_{\widetilde{\Sigma}}\right) d\nu(\tau).$$

Combining this with the infinite-data limit sends the complete CMST tax to zero, as expected for an unrestricted population learner.

Uniform decoder geometry. If the decoder has equal average sensitivity along all covariance modes, $w_i \equiv w$, then $\bar{w} = w$ and the image-space calibration factors out:

$$\widehat{\Gamma}_{\mathrm{CMST}} = w \int \left[\tau^2 \Phi\left(\frac{d}{n}, \tau, H_{\widetilde{\Sigma}}\right) + \frac{1}{d}\sum_{i > r_C} \omega_\tau(\lambda_i) + \frac{1}{d}\sum_{i \le r_C} \omega_\tau(\lambda_i)\, |1 - \eta(\lambda_i + \tau)|^{2T}\right] d\nu(\tau).$$

Hence decoder geometry acts only as a global scale in this special case, and CMST reduces to a purely latent-spectral criterion.

Isotropic latent. For a trace-normalized isotropic representation, $\widetilde{\Sigma} = I_d$, so $H_{\widetilde{\Sigma}} = \delta_1$ and $\omega_\tau(1) = (1+\tau)^{-1}$. If also $w_i \equiv w$ and $q_C := r_C/d$, then

$$\widehat{\Gamma}_{\mathrm{CMST}} = w \int \left[\tau^2 \Phi_{\mathrm{iso}}(\gamma, \tau) + \frac{1 - q_C}{1 + \tau} + \frac{q_C}{1 + \tau}\, |1 - \eta(1 + \tau)|^{2T}\right] d\nu(\tau). \tag{52}$$

Equation 52 makes the three bottlenecks particularly transparent: the data term depends only on the aspect ratio and noise scale, the capacity floor is proportional to the discarded rank fraction, and the optimization transient is the retained rank fraction multiplied by the common modal contraction.

J.2 Representative Spectral Families

We next evaluate the main spectral quantities for three elementary covariance profiles. These families are analytical templates rather than assumptions on any particular tokenizer.

Isotropic spectrum. For $\lambda_i = \lambda$ for all i, Eq. 3 gives

$$R_\tau = \frac{d}{2} \log\left(1 + \frac{\lambda}{\tau}\right), \qquad d_{\mathrm{eff}}(\tau) = d \frac{\lambda}{\lambda + \tau}.$$

The population rank-r capacity tax is

$$\Gamma_{\mathrm{cap}} = \left(1 - \frac{r}{d}\right) \frac{\lambda^2}{\lambda + \tau},$$

and, within the oracle rank-r subspace, Eq. 45 reduces to

$$\Gamma_{r,T} = \frac{\lambda^2}{\lambda + \tau} \left[1 - \frac{r}{d} + \frac{r}{d}\, |1 - \eta(\lambda + \tau)|^{2T}\right].$$

The canonical CMST normalization corresponds to $\lambda = 1$.

Spiked spectrum. Consider a two-level spectrum with k spike modes and $d - k$ bulk modes,

$$\lambda_i = \begin{cases} \lambda_{\mathrm{s}}, & i \le k, \\ \lambda_{\mathrm{b}}, & i > k, \end{cases} \qquad \lambda_{\mathrm{s}} > \lambda_{\mathrm{b}} \ge 0. \tag{53}$$

For a canonicalized representation one additionally has $k\lambda_{\mathrm{s}} + (d - k)\lambda_{\mathrm{b}} = d$. The diffusion-scale rate and active dimension become

$$R_\tau = \frac{1}{2}\left[k \log\left(1 + \frac{\lambda_{\mathrm{s}}}{\tau}\right) + (d - k) \log\left(1 + \frac{\lambda_{\mathrm{b}}}{\tau}\right)\right],$$

and

$$d_{\text{eff}}(\tau) = k\frac{\lambda_{\text{s}}}{\lambda_{\text{s}}+\tau} + (d-k)\frac{\lambda_{\text{b}}}{\lambda_{\text{b}}+\tau}.$$

Thus, when $\lambda_{\text{b}} \ll \tau \ll \lambda_{\text{s}}$, the spike modes remain almost fully active while each bulk mode contributes only approximately λ_{b}/τ.

For the oracle rank-r surrogate, define

$$r_{\text{s}} := \min\{r, k\}, \qquad r_{\text{b}} := (r-k)_+,$$

and

$$\omega_{\text{s}} := \frac{\lambda_{\text{s}}^2}{\lambda_{\text{s}}+\tau}, \qquad \omega_{\text{b}} := \frac{\lambda_{\text{b}}^2}{\lambda_{\text{b}}+\tau},$$

with $\rho_{\text{s}} = 1-\eta(\lambda_{\text{s}}+\tau)$ and $\rho_{\text{b}} = 1-\eta(\lambda_{\text{b}}+\tau)$. Then the exact joint population cost is

$$\Gamma_{r,T} = \frac{1}{d}\left[(k-r_{\text{s}})\omega_{\text{s}} + (d-k-r_{\text{b}})\omega_{\text{b}} + r_{\text{s}}\omega_{\text{s}}\rho_{\text{s}}^{2T} + r_{\text{b}}\omega_{\text{b}}\rho_{\text{b}}^{2T}\right]. \tag{54}$$

This expression separates a permanent cost from discarded spikes or bulk directions and a transient cost from the retained modes.

Power-law spectrum. Consider

$$\lambda_i = c\, i^{-\beta}, \qquad \beta > 1,$$

where $c > 0$ can be rescaled to satisfy the CMST trace normalization if desired. Define the noise-dependent spectral index

$$i_\tau := \left(\frac{c}{\tau}\right)^{1/\beta}.$$

In the interior regime $1 \ll i_\tau \ll d$,

$$d_{\text{eff}}(\tau) = \sum_{i=1}^{d} \frac{1}{1+(i/i_\tau)^\beta}.$$

Approximating the sum by an integral gives

$$d_{\text{eff}}(\tau) \sim C_\beta\, i_\tau, \qquad C_\beta := \int_0^\infty \frac{dx}{1+x^\beta} = \frac{\pi}{\beta}\csc\left(\frac{\pi}{\beta}\right). \tag{55}$$

Similarly,

$$R_\tau \sim \frac{\beta C_\beta}{2}\, i_\tau,$$

since

$$\int_0^\infty \log(1+x^{-\beta})\, dx = \beta C_\beta.$$

Thus both diffusion-scale rate and active dimension are controlled, up to constants, by the number of eigenvalues above the current noise scale.

The rank-r capacity tail exhibits a corresponding crossover. Since $\omega_\tau(\lambda) \approx \lambda$ for $\lambda \gg \tau$ and $\omega_\tau(\lambda) \approx \lambda^2/\tau$ for $\lambda \ll \tau$, integral approximation yields

$$\Gamma_{\text{cap}}(r,\tau) \asymp \begin{cases} \dfrac{c}{d(\beta-1)} r^{1-\beta}, & 1 \ll r \ll i_\tau, \\ \dfrac{c^2}{d\tau(2\beta-1)} r^{1-2\beta}, & i_\tau \ll r \ll d. \end{cases} \tag{56}$$

Hence the effective noise-scale rank $i_\tau \asymp d_{\text{eff}}(\tau)$ marks where the decay law of the unresolved capacity burden changes. The borderline case $\beta = 1$ introduces logarithmic factors and is omitted here for brevity.

J.3 A WORKED RDGL SCALING EXAMPLE

The marginal law in Eq. 16 can be made explicit for a simple representation family. Suppose reconstruction distortion follows a diminishing-return curve

$$\mathcal{D}(d) = Ad^{-p}, \qquad A > 0, \quad p > 0.$$

The statistical minimax law in Eq. 7 scales as $\min\{d, d^2/n\}$. To avoid introducing a nondifferentiable kink at $d = n$, consider the smooth proxy

$$\mathcal{G}^{\mathrm{sm}}_{\mathrm{stat}}(d; n) = \zeta \frac{d^2}{n + d}, \qquad \zeta > 0. \tag{57}$$

This proxy preserves the minimax scaling up to universal constants:

$$\frac{\zeta}{2} \min\left\{d, \frac{d^2}{n}\right\} \le \mathcal{G}^{\mathrm{sm}}_{\mathrm{stat}}(d; n) \le \zeta \min\left\{d, \frac{d^2}{n}\right\}.$$

The corresponding RDGL objective is

$$\mathcal{J}(d) = Ad^{-p} + \alpha\zeta \frac{d^2}{n + d}.$$

Its derivatives are

$$\mathcal{J}'(d) = -pAd^{-p-1} + \alpha\zeta \frac{d(2n + d)}{(n + d)^2},$$

and

$$\mathcal{J}''(d) = p(p + 1)Ad^{-p-2} + 2\alpha\zeta \frac{n^2}{(n + d)^3} > 0.$$

Thus the smooth example has a unique interior optimum whenever the stationary point lies in the admissible range, determined by

$$pAd^{-p-1} = \alpha\zeta \frac{d(2n + d)}{(n + d)^2}. \tag{58}$$

Two limiting regimes give transparent scaling laws. If $d^\star \ll n$, then $d(2n + d)/(n + d)^2 \sim 2d/n$, so

$$d^\star \sim \left(\frac{pA\, n}{2\alpha\zeta}\right)^{1/(p+2)}. \tag{59}$$

Hence the preferred representation size grows sublinearly with the available sample budget.

If instead $d^\star \gg n$, then $d(2n + d)/(n + d)^2 \to 1$, yielding

$$d^\star \sim \left(\frac{pA}{\alpha\zeta}\right)^{1/(p+1)}. \tag{60}$$

In this regime the statistical burden has already entered its linear-in-d branch, so the leading-order optimum no longer depends on n.

These reductions make the comparative statics in Theorem 7 explicit: larger n supports a richer representation in the data-rich regime, while increasing the generative penalty α reduces $d^\star$ in both regimes. More generally, if ζ is interpreted as an effective resource-dependent calibration that decreases when downstream capacity or optimization improves, the same expressions predict an increase in the capacity-matched representation size.

J.4 THEORY-TO-MEASUREMENT DICTIONARY

The quantities above are theoretical objects, but most admit direct empirical or plug-in counterparts. Table 3 summarizes the mapping used by theory-aligned diagnostics and by CMST itself. The table is not meant to prescribe a unique estimator; it records the simplest operational counterpart of each quantity.

Table 3: Theory-to-measurement dictionary for the main quantities used in RDGL and CMST.

Theoretical quantity	**Operational estimate**	**Role**
$\widetilde{\Sigma}$, $H_{\widetilde{\Sigma}}$	Center and trace-normalize the latent batch, then estimate its covariance and eigenspectrum; for large d, leading modes or spectral functionals can be estimated by Lanczos or stochastic Lanczos quadrature.	Latent geometry entering R_τ, d_{eff}, RMT, and CMST.
R_τ, $d_{\text{eff}}(\tau)$	Plug the estimated eigenvalues into $\frac{1}{2}\sum_i \log(1+\lambda_i/\tau)$ and $\sum_i \lambda_i/(\lambda_i+\tau)$.	Noise-scale representation complexity.
$\Phi(\gamma, \tau, H)$	Set $\gamma = d/n$, replace H by the current empirical spectral measure, and solve the scalar fixed point in Eq. 11.	Finite-data component of the CMST tax.
w_i	Average directional decoder sensitivity $\|J_{\widetilde{D}}(\widetilde{Z})v_i\|_2^2$ using decoder JVPs over a latent batch.	Converts modal latent error into local image-space units.
b_i	Monte Carlo estimate of $z_i^2\|J_D(Z)v_i\|_2^2$.	Reconstruction value of latent mode i.
c_i	If $f_\tau^\star$ is available, directly average $(v_i^\top e_\tau)^2\|J_D(Z)v_i\|_2^2$ over samples and noise scales, or use $c_i \approx w_i\widetilde{c}_i$. For real tokenizer latents where $f_\tau^\star$ is unavailable, use a frozen high-resource denoiser as reference and denote the resulting proxy by c_i^{emp}.	Finite-resource generative burden of mode i and its empirical reference-based counterpart.
Γ_{cap}, Γ_{opt}	Insert the estimated spectrum into Eqs. 13 and 14, using r_C for the capacity surrogate and (T, η) for the optimization surrogate.	Population capacity and optimization surrogates.
$\widehat{\Gamma}_{\text{CMST}}$	Evaluate Eq. 19 on the empirical spectrum and decoder sensitivities, with numerical quadrature or sampled noise scales from ν.	Resource-matched tokenizer constraint or penalty.
$\mathcal{D}(d)$, $\mathcal{G}(d; n, C, T)$	Estimate along a tokenizer-size sweep; finite differences provide empirical counterparts of the marginal reconstruction gain and generative cost.	Tests the RDGL balance and capacity-matched size prediction.

K Experimental Setup and Implementation Details

This appendix specifies the common experimental protocol used throughout Section 7. We separate controlled comparisons, where tokenizer or downstream resources are changed while the remaining pipeline is held fixed, from broader system-level comparisons that inherit substantially different latent topologies, pretraining data, or joint-training procedures. This distinction is important because our primary goal is to isolate representation-dependent generative difficulty rather than to construct a heterogeneous generation leaderboard.

K.1 Datasets and Evaluation Protocol

ImageNet. Our primary benchmark is ImageNet-1K (Deng et al., 2009) at 256×256 resolution. Tokenizers are trained on the ImageNet training split and evaluated on the validation split. Unless an experiment explicitly varies the statistical budget, the downstream generator uses the full training set. For finite-data experiments, we construct class-stratified subsets containing $\{1/8, 1/4, 1/2, 1\}$ of the training images. Throughout these experiments, n denotes the number of *independent images*, consistent with the statistical formulation in the main text; spatial latent locations are never counted as independent samples when defining $\gamma = d/n$. A fixed held-out subset of the ImageNet training split is reserved for hyperparameter and CMST-resource calibration; the validation split is used only for final evaluation.

Controlled Gaussian experiments. Experiments that directly test the finite-sample and random-matrix predictions use synthetic centered Gaussian distributions with explicitly specified covariance spectra. The finite-sample experiment uses uniformly conditioned covariance families, while the RMT experiments consider isotropic, spiked, and tokenizer-matched spectral laws as specified in Appendix L.3. These experiments report the finite-dimensional quantities appearing in the corresponding theory.

Reconstruction metrics. We report reconstruction FID (rFID) and LPIPS (Zhang et al., 2018) as the primary distributional and perceptual reconstruction measures. All reconstruction metrics are computed on the same 50K validation-image reconstructions whenever applicable.

Generation metrics. Our primary downstream metric is unguided gFID-50K (Heusel et al., 2017), evaluated from 50K generated images against the standard ImageNet reference statistics. Because Inception features can favor particular model families, we additionally report Fréchet distance in the DINOv2-ViT-L/14 representation space (FD-DINOv2) following Stein et al. (2023). Inception Score, precision, recall, and guided FID are included in the extended results. Theory-facing comparisons use unguided generation unless explicitly stated, so that representation learnability is not confounded by method-specific classifier-free-guidance tuning.

K.2 Tokenizer Families and Baselines

Controlled tokenizer family. The main controlled experiments use the convolutional continuous-autoencoder family underlying VA-VAE (Yao et al., 2025). At the default $f16d32$ operating point, a 256×256 image is encoded into a $16 \times 16 \times 32$ latent tensor. For representation-size experiments, we keep the spatial downsampling ratio fixed at $f = 16$ and vary only the channel dimension $c \in \{8, 16, 32, 64\}$. The corresponding flattened representation sizes are $d = 256c \in \{2048, 4096, 8192, 16384\}$. Holding the spatial token count fixed avoids conflating representation richness with the quadratic sequence-length cost of the downstream Transformer.

We use the same encoder–decoder architecture, reconstruction loss, data preprocessing, and training duration for the following causal controls: a reconstruction-only VAE, VA-VAE, VA-VAE with an isotropy penalty, and CMST. The isotropy control uses $\mathcal{L}_{\text{iso}} = c^{-1}\|\widetilde{\Sigma}_{\text{ch}} - I\|_F^2$ and tests whether generic covariance flattening can explain the gains of resource-dependent spectral matching.

External tokenizer baselines. We additionally evaluate AlignTok (Chen et al., 2026) and VTP (Yao et al., 2026) with their released tokenizers while using the same downstream evaluation harness whenever their native representation permits direct integration. Both methods explicitly target more generatively useful latent representations, making them substantially more informative baselines for our setting than reconstruction-only autoencoders.

Broader system-level references. For broader system-level context, we also report published or officially released results from SoftVQ-VAE (Chen et al., 2025b), REPA-E (Leng et al., 2025), MacTok (Zeng et al., 2026), AToken (Lu et al., 2026), and RAEv2 (Singh et al., 2026). These methods differ materially in token topology, pretrained encoders, generator architecture, pretraining scale, or end-to-end optimization, and are therefore not treated as controlled comparisons. Table 4 summarizes their role in our evaluation, while Table 11 reports a representative native-protocol subset with directly available ImageNet-256 generation results.

K.3 Downstream Generator and Resource Budgets

Default generator. Unless otherwise specified, downstream generative learnability is evaluated with LightningDiT-XL/1 (Yao et al., 2025). For $f16d32$ latents, we follow the released LightningDiT recipe: global batch size 1024, AdamW with learning rate 2×10^{-4}, linear transport, velocity prediction, log-normal time sampling, and cosine-weighted denoising loss. The default finite optimization budget is 80 epochs, corresponding to approximately 10^5 parameter updates on ImageNet-1K.

We use unguided sampling for the primary comparisons. Final generation metrics use 250-step Euler integration and 50K generated samples. Guided results use a common CFG-scale grid $\{4, 6, 8, 10\}$

Table 4: Baseline taxonomy used in the experiments. "Controlled" indicates that the comparison is evaluated under our common downstream protocol; "native" denotes a system-level reference evaluated under the protocol of the corresponding work.

Method	Year	Representation	Role
Recon-only VAE	–	2D continuous	Controlled
VA-VAE (Yao et al., 2025)	2025	2D continuous	Controlled
Isotropy regularization	–	2D continuous	Controlled
AlignTok (Chen et al., 2026)	2026	continuous	Controlled/external
VTP (Yao et al., 2026)	2026	continuous	Controlled/external
CMST (ours)	–	2D continuous	Controlled
SoftVQ-VAE (Chen et al., 2025b)	2025	1D continuous	Native
REPA-E (Leng et al., 2025)	2025	VAE + end-to-end	Native
MacTok (Zeng et al., 2026)	2026	1D continuous	Native
AToken (Lu et al., 2026)	2026	unified tokenizer	Native
RAEv2 (Singh et al., 2026)	2026	representation AE	Native

for all controlled methods; the scale is selected on the held-out calibration split and then frozen for the final 50K-sample evaluation.

Finite data, capacity, and optimization. The three resource axes are varied one at a time. For the finite-data sweep, we use class-stratified subsets containing $\{1/8, 1/4, 1/2, 1\}$ of ImageNet while holding the optimization budget fixed at 10^5 parameter updates and the global batch size at 1024; smaller subsets are therefore sampled with replacement as needed. This isolates the number of distinct training examples n from optimization time. Model capacity is varied on the full dataset using LightningDiT-B/1, L/1, and XL/1, again for 10^5 updates. Optimization is varied with the full dataset and fixed XL/1 architecture using checkpoints at approximately 20, 40, and 80 full-ImageNet epochs; selected experiments extend training to 160 epochs. Neural-network scale is treated as an empirical realization of the abstract capacity budget C, while optimizer updates are denoted T_{train} and are not identified with the linear-gradient iteration count in the population surrogate.

K.4 Practical CMST Instantiation

Structured spectral space. For 2D tokenizer latents, the full flattened covariance can mix spatial positions and would introduce a new architectural degree of freedom unrelated to our main question. We therefore instantiate CMST in a translation-shared channel space. Given latent tensors in $\mathbb{R}^{h\times w\times c}$, we estimate a channel covariance $\widetilde{\Sigma}_{\mathrm{ch}} \in \mathbb{R}^{c\times c}$ by aggregating centered activations over images and spatial locations. The same channel spectral basis and gates are applied at every spatial location. We refer to this implementation as *channel-CMST*. For the statistical component, we retain the flattened representation size $d = hwc$ in $\gamma = d/n$; the c-dimensional channel spectrum is interpreted as a translation-shared spectral law repeated across the hw spatial locations. Thus spatial positions are not treated as additional independent samples. Correspondingly, r_C in channel-CMST denotes a per-location effective channel-rank budget, while $q_C = r_C/c$ is its normalized capacity fraction.

Canonicalization and spectral updates. Latents are centered and trace-normalized before evaluating the spectral tax, as prescribed by Eq. 18. Spectral statistics are refreshed every 500 tokenizer updates using 2,048 images. Since the main experiments use at most $c = 64$ channels, we use exact eigendecomposition rather than an approximate Lanczos estimator. Decoder sensitivities w_i are estimated by batched Jacobian–vector products and are treated as stop-gradient quantities between spectral refreshes.

Noise-scale integration. The measure ν in Eq. 19 is chosen as the pushforward of the downstream LightningDiT training-time distribution under the corresponding noise-scale map $\tau(t)$. We approximate the integral with eight equal-mass quadrature nodes. Diagnostic plots that explicitly vary τ use a separate logarithmically spaced grid over $[10^{-3}, 10^{1}]$.

Capacity calibration. The rank parameter r_C in the Gaussian surrogate is not interpreted as the literal matrix rank of a nonlinear DiT. Instead, we write $q_C = r_C/c$ and treat it as an effective spectral-capacity calibration variable. For each downstream generator scale, q_C is selected once from $\{0.25, 0.5, 0.75, 1.0\}$ using a held-out calibration split and is then frozen across tokenizer comparisons. The same principle is used for the effective optimization parameters entering the CMST surrogate: they are calibrated once for each downstream resource regime and are not re-tuned per tokenizer.

Constraint and gates. All gates are initialized at $g_i = 1$. We parameterize g_i by a sigmoid-transformed unconstrained variable and optimize the tokenizer and gates jointly with the reconstruction objective. The dual variable is updated by projected ascent. For trade-off experiments, we express the constraint relative to the initial unconstrained spectral tax, $B_{\mathrm{gen}} = \rho_{\mathrm{gen}} \widehat{\Gamma}^{(0)}_{\mathrm{CMST}}$, and use $\rho_{\mathrm{gen}} \in \{0.6, 0.7, 0.8, 0.9, 1.0\}$, with 0.8 as the default operating point.

K.5 Optimization, Compute, and Reproducibility

Tokenizer optimization. Controlled $f16d32$ tokenizer variants are trained for 50 epochs on ImageNet using the VA-VAE architecture and objective as the common base. The reconstruction-only, isotropy, and CMST variants use identical initialization, data ordering, augmentations, optimizer, and reconstruction losses; only the representation regularizer differs. Unless otherwise noted, tokenizer optimization uses a base learning rate of 10^{-4}, an effective global batch size of 256 (8 images per GPU with two-step gradient accumulation on 16 GPUs), and mixed-BF16 training.

Compute environment. Experiments are run on two nodes with eight NVIDIA H100 SXM 80GB GPUs per node. Full LightningDiT runs use one 8-GPU node, while tokenizer training and large resource sweeps use up to both nodes. Within a node, GPUs communicate through NVLink/NVSwitch; distributed training uses NCCL. Our common environment uses Ubuntu 24.04, Python 3.11, PyTorch 2.14, and CUDA 13.0. Third-party tokenizers are loaded using their officially released environments when necessary, after which latent features are exported to a common sharded format and all downstream controlled evaluations use our unified LightningDiT environment.

The complete study uses approximately 5.0×10^3 H100 GPU-hours, including tokenizer training, downstream generator sweeps, controlled replications, ablations, and final 50K-sample evaluations. Individual large-scale generator runs use at most eight GPUs.

Statistical reporting. Primary controlled comparisons and all configurations reported in the main resource sweeps are repeated with three random seeds and reported as mean $\pm$ standard deviation. Exploratory configurations not used for a reported comparison may be screened with a common seed before replication. Gaussian simulations use 100 Monte Carlo repetitions; Table 9 reports the resulting finite-dimensional discrepancies. Hyperparameters are selected using the held-out calibration split and frozen before final evaluation.

L Theory-Aligned Empirical Validation

This appendix provides controlled numerical and neural-network experiments connecting the analytical results in Sections 4–6.1 to finite-dimensional Gaussian models and realistic tokenizer latents. We reserve direct quantitative comparisons for settings matching the corresponding analytical assumptions, while full neural generators are used to test whether the predicted mechanisms persist beyond those assumptions.

L.1 Reconstruction-Gauge Stress Test

Gauge construction. Let $z \in \mathbb{R}^c$ denote the channel vector at one spatial position of the pretrained $f16d32$ tokenizer, with $c = 32$. We first compute a single channel mean μ_Z and global scale $s_Z^2 = c^{-1} \operatorname{tr} \Sigma_{\mathrm{ch}}$ from the original training latents and define $\widetilde{z} = (z - \mu_Z)/s_Z$. These statistics are fixed for every intervention. Writing $\widetilde{\Sigma}_{\mathrm{ch}} := \mathrm{Cov}(\widetilde{z}) = V \operatorname{diag}(\lambda_1, \dots, \lambda_c) V^\top$ with $\lambda_1 \geq \cdots \geq \lambda_c$, we set $s_i = 1 - 2(i-1)/(c-1)$ and $A_a = V \operatorname{diag}(a^{s_1}, \dots, a^{s_c}) V^\top$. Since $\sum_i s_i = 0$, $\det A_a = 1$.

The transformed encoder and decoder are $E_a(x) = A_a\widetilde{E}(x)$ and $D_a(u) = D(\mu_Z + s_Z A_a^{-1}u)$, where the same transformation is applied independently at every spatial position. Consequently, $D_a(E_a(x)) = D(E(x))$ exactly in real arithmetic.

We evaluate $a \in \{1, 1.5, 2, 3, 4\}$. As a geometry-preserving control, we draw a fixed Haar-random orthogonal matrix $Q \in \mathbb{R}^{c\times c}$ and analogously use $E_Q(x) = Q\widetilde{E}(x)$ and $D_Q(u) = D(\mu_Z + s_Z Q^\top u)$. All downstream generators use the same isotropic corruption schedule and the same global latent scale. In particular, we do not recompute per-gauge channel-wise standardization in the primary intervention, since doing so would partially undo the coordinate transformation being tested.

Spectral diagnostics and exact reconstruction. For descriptive purposes, we evaluate the Gaussian covariance proxy $R_\tau^{\mathrm{G}} = \frac{1}{2}\sum_i \log(1 + \lambda_i/\tau)$ and the noise-regularized condition number $\kappa_\tau = (\lambda_{\max} + \tau)/(\lambda_{\min} + \tau)$ at $\tau = 0.1$. We use the superscript G because the empirical tokenizer latents need not be Gaussian; R_τ^{G} is therefore a spectral diagnostic rather than an exact estimate of $I(Z; Z_\tau)$. Table 5 confirms that the inverse-decoder construction leaves reconstruction unchanged to numerical precision while substantially altering the spectral geometry and downstream generation.

Table 5: **Exact reconstruction invariance under latent gauges.** $R_{0.1}^{\mathrm{G}}$ and $\kappa_{0.1}$ are computed from the empirical channel spectrum. Generation uses LightningDiT-XL/1 for 80 epochs without CFG; gFID reports mean $\pm$ standard deviation over three runs.

Gauge	$R_{0.1}^{\mathrm{G}}$	$\kappa_{0.1}$	Max. pixel diff. ↓	rFID ↓	gFID ↓
Identity	36.67	4.70	0	0.281	4.18 ± 0.07
Orthogonal Q	36.67	4.70	1.8×10^{-7}	0.281	4.23 ± 0.08
$A_{1.5}$	37.25	18.31	2.0×10^{-7}	0.281	4.91 ± 0.09
A_2	37.93	44.26	2.1×10^{-7}	0.281	6.45 ± 0.12
A_3	39.28	134.75	2.5×10^{-7}	0.281	9.88 ± 0.19
A_4	40.51	273.67	2.9×10^{-7}	0.281	15.61 ± 0.28

The orthogonal transformation changes the channel coordinates but preserves both covariance eigenvalues and isotropic diffusion geometry, and correspondingly produces no meaningful change in fixed-budget generation. In contrast, the non-orthogonal gauges preserve the same decoded image while progressively stretching the latent spectrum. Their effect appears early during training and remains substantial at 80 epochs, rather than merely producing a small offset near convergence.

Noise-resolved denoising difficulty. The final gFID gap could in principle arise only at sampling time or from a small difference accumulated late in training. We therefore inspect the learning problem directly using the same trained checkpoints. For each gauge and noise scale τ, we corrupt held-out latents using the downstream forward process, convert the model prediction back to a clean-latent estimate $\widehat{Z}_\theta$, and measure the energy-normalized denoising error $\mathcal{E}_{\mathrm{den}}(\tau) = \mathbb{E}\|\widehat{Z}_\theta - Z\|_2^2/\mathbb{E}\|Z - \mathbb{E}Z\|_2^2$. This normalization removes differences in total latent energy and makes the quantity comparable across gauges. Figure 10 reports the resulting noise-resolved errors for the identity, orthogonal, and two representative anisotropic gauges.

The denoising diagnostic mirrors the final generation behavior but localizes the discrepancy to the diffusion learning problem itself. Identity and orthogonal coordinates remain nearly indistinguishable throughout the schedule, while the $a = 2$ and $a = 4$ gauges exhibit progressively larger errors, with the strongest separation around the low-to-intermediate noise regime. At very high noise, the curves partially reconverge because little recoverable information remains under any parameterization. We emphasize that $\mathcal{E}_{\mathrm{den}}$ is an empirical neural-network diagnostic rather than the Gaussian excess risk Ψ from Section 5.2; its role is to test whether the same geometry-dependent ordering persists in the nonlinear downstream learner.

Removing global-scale and preprocessing explanations. The volume-preserving family above intentionally follows the construction in Theorem 1, but its total second moment can increase with

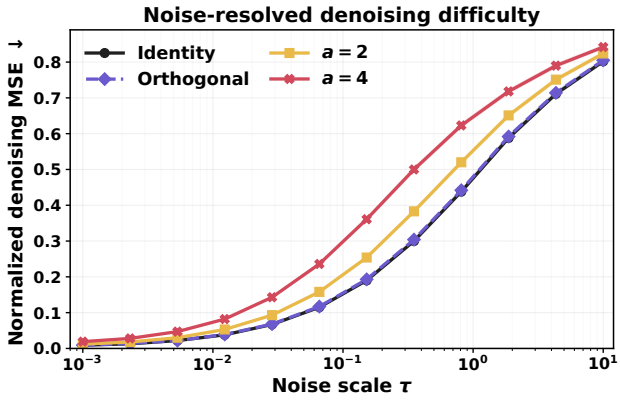


Figure 10: **Gauge anisotropy increases held-out denoising difficulty across noise scales.** Curves show energy-normalized clean-latent denoising MSE at the 80-epoch checkpoints; shaded regions denote ± 1 standard deviation over three independently trained generators. Identity and orthogonal gauges remain nearly indistinguishable, whereas anisotropic gauges incur their largest relative penalty at low-to-intermediate noise levels.

anisotropy. We therefore repeat the experiment with $\bar{A}_a = A_a \Big/ \sqrt{c^{-1}\,\mathrm{tr}\Big(A_a \widetilde{\Sigma}_{\mathrm{ch}} A_a^\top\Big)}$, which enforces unit average latent variance and is again paired with its exact inverse in the decoder. We also repeat the experiment after recomputing conventional per-channel standardization statistics for every transformed latent space. Table 6 shows that both interventions reduce the magnitude of the degradation but leave the ordering intact. Hence the phenomenon cannot be explained solely by a change in global latent energy or by the absence of channel standardization.

Table 6: **Gauge robustness controls.** Unguided gFID-50K after 80 epochs. Trace matching removes changes in total latent variance; per-channel standardization further compensates marginal channel scales but does not undo the full non-orthogonal geometry.

Gauge severity	Volume-preserving	Trace-matched	Channel-standardized
$a = 1$	4.18 ± 0.07	4.18 ± 0.07	4.18 ± 0.07
$a = 1.5$	4.91 ± 0.09	4.62 ± 0.08	4.47 ± 0.09
$a = 2$	6.45 ± 0.12	5.41 ± 0.11	5.06 ± 0.10
$a = 3$	9.88 ± 0.19	7.49 ± 0.16	6.81 ± 0.15
$a = 4$	15.61 ± 0.28	10.27 ± 0.22	8.94 ± 0.20

Together, these controls isolate the mechanism in Theorem 1: the reconstruction map remains fixed while the geometry seen by an isotropic diffusion learner changes substantially, producing large fixed-budget generation gaps.

L.2 Finite-Resource Representation Sweeps

The experiments in this section isolate the three resource axes entering the RDGL formulation. All tokenizer variants use the same $f16$ spatial resolution and differ only in channel width $c \in \{8, 16, 32, 64\}$. Thus the Transformer sequence length and attention cost remain fixed, while the flattened representation size is $d = 256c$. For the data sweep we hold LightningDiT-XL/1 and 10^5 optimizer updates fixed and vary only the number of distinct training images; for the capacity sweep we use the full dataset and 10^5 updates while varying B/1, L/1, and XL/1; for the optimization sweep we fix the full dataset and XL/1 architecture and vary training duration.

Increasing c produces a monotonic reconstruction improvement while changing the downstream generator only through its small input/output adapters. Table 7 reports the corresponding reconstruction and compute statistics.

Table 7: **Representation richness and reconstruction.** The spatial latent resolution is fixed at 16×16; only the channel width changes. Generator parameter count and step time change negligibly compared with the Transformer backbone.

c	Flattened d	rFID ↓	LPIPS ↓	XL/1 params.	Step time
8	2,048	1.13	0.104	674.7M	1.00×
16	4,096	0.56	0.073	674.8M	1.00×
32	8,192	0.281	0.052	674.9M	1.01×
64	16,384	**0.209**	**0.046**	675.1M	1.02×

The complete resource sweeps are given in Table 8. In each row, bold and underlined entries denote the best and second-best gFID, respectively. The preferred width is smallest when any one resource is restrictive and moves toward larger representations as that bottleneck is relaxed. Importantly, the finite-data sweep uses the same number of optimizer updates at every data fraction, so its crossover cannot be attributed to shorter training on the smaller subsets.

Table 8: **Full finite-resource representation sweep.** Entries are unguided gFID-50K (mean ± standard deviation over three runs). Only the resource named in the first two columns is changed within each block.

Axis	Resource level	$c = 8$	$c = 16$	$c = 32$	$c = 64$	Best c
Data	$1/8$	8.96 ± 0.15	$\mathbf{7.18 \pm 0.11}$	$\underline{8.47 \pm 0.14}$	12.91 ± 0.23	16
	$1/4$	7.31 ± 0.13	$\mathbf{5.73 \pm 0.09}$	$\underline{6.15 \pm 0.10}$	8.79 ± 0.17	16
	$1/2$	5.96 ± 0.10	$\underline{4.82 \pm 0.08}$	$\mathbf{4.55 \pm 0.07}$	5.48 ± 0.09	32
	Full	5.10 ± 0.09	4.47 ± 0.07	$\underline{4.18 \pm 0.07}$	$\mathbf{4.06 \pm 0.06}$	64
Capacity	B/1	6.34 ± 0.10	$\mathbf{5.51 \pm 0.08}$	$\underline{6.08 \pm 0.10}$	8.11 ± 0.14	16
	L/1	5.46 ± 0.09	$\underline{4.74 \pm 0.07}$	$\mathbf{4.45 \pm 0.07}$	4.96 ± 0.08	32
	XL/1	5.10 ± 0.09	4.47 ± 0.07	$\underline{4.18 \pm 0.07}$	$\mathbf{4.06 \pm 0.06}$	64
Optimization	20 ep.	$\underline{12.92 \pm 0.22}$	$\mathbf{11.76 \pm 0.18}$	13.10 ± 0.28	16.43 ± 0.31	16
	40 ep.	7.53 ± 0.13	$\mathbf{6.59 \pm 0.11}$	$\underline{7.01 \pm 0.15}$	8.62 ± 0.18	16
	80 ep.	5.10 ± 0.09	4.47 ± 0.07	$\underline{4.18 \pm 0.07}$	$\mathbf{4.06 \pm 0.06}$	64
	160 ep.	4.74 ± 0.08	4.13 ± 0.06	$\underline{3.82 \pm 0.05}$	$\mathbf{3.60 \pm 0.05}$	64
Capacity–time	B/1, 160 ep.	5.92 ± 0.10	$\mathbf{5.24 \pm 0.08}$	$\underline{5.81 \pm 0.09}$	7.20 ± 0.13	16

The distinction between permanent capacity limitations and optimization transients becomes clearer when training is extended. Figure 11 shows both the long-horizon convergence and the reconstruction–generation plane. With XL/1, wider representations progressively overtake narrower ones as optimization proceeds, and $c = 64$ is best by 80–160 epochs. In contrast, B/1 still prefers $c = 16$ after 160 epochs (Table 8), suggesting that additional optimization alone does not remove a sufficiently tight capacity bottleneck.

Together, these sweeps recover both central finite-resource predictions in a realistic tokenizer–DiT pipeline: reconstruction–generation reversal under tight resources and a systematic shift toward richer preferred representations as data, model capacity, or optimization budget increases.

L.3 Finite-Sample and RMT Numerical Validation

The neural experiments above test whether the finite-resource mechanisms persist beyond the tractable models used in the theory. We now return to the Gaussian setting, where the relevant quantities can be measured directly and compared quantitatively with the analytical predictions. These experiments provide finite-dimensional checks of the matching estimators and random-matrix asymptotics.

For the finite-sample experiment, we fix $\rho = 0.5$, draw covariance eigenvalues independently from $\mathrm{Unif}[1 - \rho, 1 + \rho]$, and apply a Haar-random orthogonal eigenbasis. We evaluate the matching estimator from Appendix D.3: the projected sample covariance is used for $n \geq d$ and the identity estimator for $n < d$, matching the two branches of the upper-bound construction. We vary $d \in \{64, 128, 256, 512\}$ and the sample ratio n/d from $1/4$ to 8, measuring the exact Gaussian W_2^2 risk through the Bures formula. For the score-learning experiment, we additionally consider isotropic, spiked, and tokenizer-matched spectral laws. The spiked family places 10% of its mass at $\lambda_s = 4$

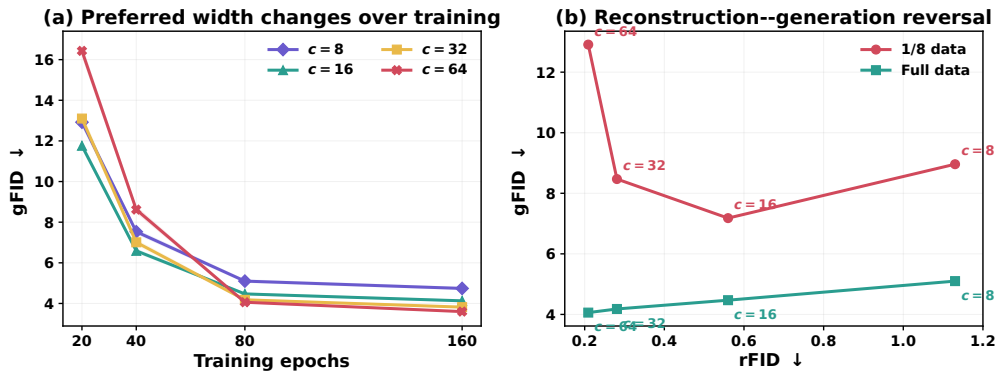


Figure 11: **Long-horizon resource matching and reconstruction–generation reversal.** Left: with full ImageNet and XL/1, narrow representations are easier early in training, but richer representations eventually overtake them as the optimization budget increases. Right: each curve traces increasing channel width $c = 8 \to 16 \to 32 \to 64$ in the reconstruction–generation plane. Under $1/8$ data, reconstruction continues to improve while generation reverses beyond $c = 16$; with full data, the same richer representations become usable.

and rescales the bulk to unit mean variance; the tokenizer-matched law is obtained by resampling the empirical channel spectrum of the default VA-VAE tokenizer. Figure 12 summarizes the results. Figure 12(b) uses $d = 512$ with n chosen to match the displayed aspect ratio $\gamma = d/n$; Monte Carlo markers average 100 trials per setting.

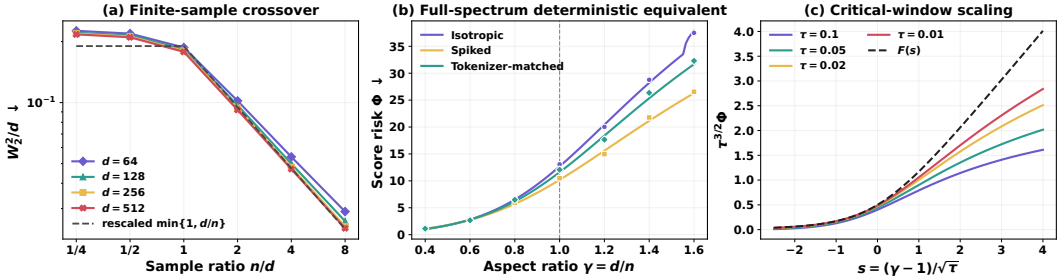


Figure 12: **Finite-dimensional numerical validation of the statistical and RMT predictions.** (a) The matching estimator exhibits the predicted crossover from an $O(d)$ data-poor plateau to the $O(d^2/n)$ data-rich regime; the dashed curve is a single rescaled reference $\min\{1, d/n\}$ rather than an exact constant prediction. (b) Finite-d Monte Carlo score risks closely track the full-spectrum deterministic equivalent at $\tau = 0.1$ for isotropic, spiked, and tokenizer-matched spectra. (c) Under $\gamma = 1 + s\sqrt{\tau}$, the rescaled isotropic curves $\tau^{3/2}\Phi$ collapse toward the critical law $F(s)$ as $\tau \downarrow 0$.

Panel (a) recovers both branches of the finite-sample scaling: when $n \lesssim d$, normalized risk remains approximately constant, whereas for $n \gg d$ it decays proportionally to d/n. The curves for different dimensions nearly collapse after normalization by d, consistent with the characteristic scaling $\psi_n(d) = \min\{d, d^2/n\}$ rather than a one-parameter-per-coordinate law. Panels (b)–(c) provide a stronger quantitative check of Section 5.2: the deterministic equivalent remains accurate for non-isotropic spectra, and the finite-noise transition sharpens into the predicted $|\gamma - 1| = O(\sqrt{\tau})$ critical window.

To quantify finite-dimensional convergence beyond the plotted examples, Table 9 reports the median relative discrepancy between Monte Carlo Φ_d and the deterministic equivalent over $\gamma \in \{0.5, 0.75, 1.0, 1.25, 1.5\}$ and $\tau \in \{0.05, 0.1, 0.2\}$.

The approximation improves steadily with dimension and remains comparably accurate for the tokenizer-matched spectrum, suggesting that the dependence on the full spectral law H is operational rather than restricted to isotropic examples. The larger error for the tokenizer-matched spectrum is expected because its mass is less regular than either analytical template; nevertheless, the

Table 9: **Finite-d error of the RMT deterministic equivalent.** Entries are median relative errors (%) over the stated (γ, τ) grid and 100 Monte Carlo trials per setting.

Spectrum	$d = 128$	$d = 256$	$d = 512$	$d = 1024$
Isotropic	7.2	4.8	3.1	2.0
Spiked	8.4	5.6	3.7	2.4
Tokenizer-matched	9.0	6.2	4.0	2.6

deterministic equivalent becomes quantitatively accurate at moderate dimensions and provides the spectral functional used by CMST as a design surrogate rather than an exact finite-d identity.

L.4 Decoder Geometry and Modal Allocation

The spectral experiments above operate primarily through covariance geometry. The allocation principle in Section 6.1, however, predicts that variance alone is insufficient once latent directions have different image-space value. We test this prediction using the frozen decoder of the default $f16d32$ VA-VAE tokenizer. All quantities below are estimated on 10K images from the held-out calibration split; the finite-resource burden uses the B/1, 40-epoch downstream learner and the same eight noise-scale quadrature nodes used by CMST.

For each covariance eigenvector v_i, we attenuate the centered latent by $Z_{\epsilon,i} = Z - \epsilon z_i v_i$ and measure the resulting image-space squared error. The local theory predicts $\mathbb{E}\|D(Z) - D(Z_{\epsilon,i})\|_2^2 = \epsilon^2 b_i + o(\epsilon^2)$. We estimate $w_i = \mathbb{E}\|J_D(Z)v_i\|_2^2$ using decoder JVPs.

For real tokenizer latents, the population-optimal denoiser $f_\tau^\star$ is unavailable. We therefore use the full-data XL/1 160-epoch checkpoint as a frozen high-resource reference f_τ^{ref} and define $e_\tau^{\mathrm{emp}} = \widehat{f}_\tau^{\mathrm{B}/1,40}(Y_\tau) - f_\tau^{\mathrm{ref}}(Y_\tau)$. The corresponding reference-based burden c_i^{emp} and latent modal excess $\widetilde{c}_i^{\mathrm{emp}}$ are computed using the same schedule quadrature as CMST. These quantities are empirical proxies for the theoretical c_i and $\widetilde{c}_i$ in Eq. 17, rather than oracle measurements of $f_\tau^\star$. Finally, we vary the retained subspace size $k \in \{8, 12, 16, 20, 24\}$ and compare variance ranking, reconstruction-value ranking, b_i/c_i^{emp}, and the basis-independent generalized allocation of Theorem 8. Figure 13 summarizes all three tests.

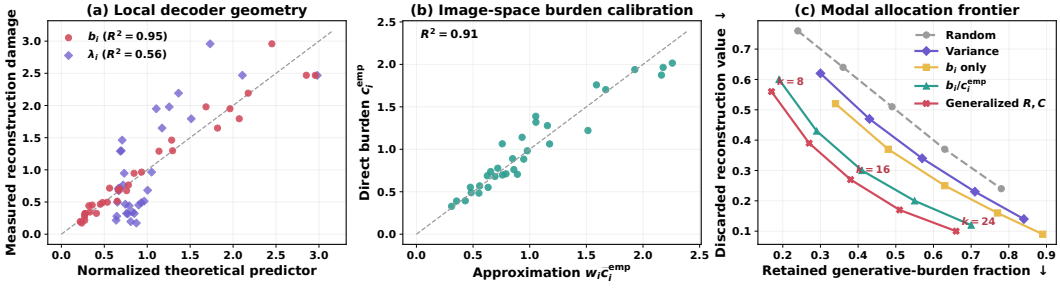


Figure 13: **Decoder geometry predicts image-space value and improves modal allocation.** (a) The local reconstruction value b_i closely predicts the measured effect of attenuating individual latent modes, whereas covariance magnitude λ_i alone is substantially less predictive. (b) The decoder-weighted approximation $w_i\widetilde{c}_i^{\mathrm{emp}}$ tracks the reference-based finite-resource burden c_i^{emp}. (c) Varying the retained dimension traces a reconstruction–generation allocation frontier: geometry-aware b_i/c_i^{emp} selection and the generalized R, C construction dominate variance-only and reconstruction-only rankings over most operating points. Lower-left is better.

At $\epsilon = 0.1$, normalized b_i explains 95% of the across-mode variation in measured reconstruction damage, compared with 56% for covariance magnitude alone; the same ordering persists for $\epsilon \in \{0.05, 0.1, 0.2\}$. The local factorization used by CMST is also accurate but not exact: $w_i\widetilde{c}_i^{\mathrm{emp}}$ explains 91% of the variation in directly estimated c_i^{emp}, consistent with decoder sensitivity being an important image-space calibration rather than a purely latent statistic.

The allocation frontier provides the corresponding decision-level check. At $k = 16$, top-variance retention discards 34% of the estimated reconstruction value while exposing 57% of the total finite-resource burden. Ranking by b_i/c_i^{emp} improves these quantities to 30% and 41%, respectively, while the generalized R, C subspace reaches 27% discarded value at 38% burden. Reconstruction-only selection attains slightly lower distortion but pays substantially more generative burden, illustrating why neither variance nor reconstruction value alone defines the preferred subspace. These measurements support the empirical counterparts of the local quantities in Eq. 17 and the modal and generalized allocation rules of Theorem 8.

M CMST Evaluation and Ablations

This appendix complements Section 7.3 with extended generation metrics, native-protocol context, generator-family robustness, resource-matched cross-evaluation, mechanism ablations, constraint sensitivity, and computational measurements. Controlled comparisons use the common $f16d32$ tokenizer interface and, unless explicitly testing generator-family robustness, the common LightningDiT evaluation pipeline; system-level results using materially different latent spaces or downstream architectures are reported separately.

M.1 Extended Generation Evaluation

The main-text conclusion is not specific to Inception FID. Table 10 reports Inception Score (Salimans et al., 2016), improved precision and recall (Kynkäänniemi et al., 2019), and guided FID under the same downstream evaluation harness and a common CFG search grid.

Table 10: **Extended ImageNet-256 generation metrics.** All methods use the common downstream evaluation harness. Precision and recall use the metric of Kynkäänniemi et al. (2019); all controlled methods share the same evaluation implementation.

Tokenizer	gFID ↓	$\mathrm{FID}_{\mathrm{cfg}}$ ↓	IS ↑	Precision ↑	Recall ↑
Recon-only VAE	8.92	3.36	195.4	0.72	0.58
VA-VAE (Yao et al., 2025)	4.18	2.06	251.6	0.78	0.62
Isotropy regularization	3.91	1.94	260.8	0.79	0.62
AlignTok (Chen et al., 2026)	3.61	1.88	268.5	0.80	0.63
VTP (Yao et al., 2026)	<u>3.49</u>	<u>1.82</u>	<u>271.3</u>	0.79	<u>0.64</u>
CMST (ours)	**2.96**	**1.67**	**286.1**	**0.81**	**0.65**

CMST improves both fidelity-oriented and coverage-oriented metrics, rather than obtaining the lower gFID through a precision–recall collapse. Its advantage also remains under CFG, although the relative gain is smaller because guidance partially compensates for representation-dependent modeling error at inference time.

For context, Table 11 lists representative recent systems under the protocols reported by their respective papers. These numbers are *not* used as controlled evidence for CMST because tokenizer topology, generator architecture, training recipe, and in some cases end-to-end optimization differ substantially.

Table 11: **Native-protocol context on ImageNet-256.** Numbers in the first four rows follow the evaluation protocol of the cited work and should not be read as an apples-to-apples ranking against CMST.

System	Generator / protocol	Training budget	Unguided gFID ↓
VA-VAE (Yao et al., 2025)	LightningDiT-XL/1	64 ep.	5.14
AlignTok (Chen et al., 2026)	LightningDiT (∼673M)	80K steps	4.05
REPA-E (Leng et al., 2025)	E2E-VAE + REPA	80 ep.	3.46
RAEv2 (Singh et al., 2026)	$\mathrm{DiT}^{\mathrm{DH}}$-XL	80 ep.	**1.06**
CMST (ours)	LightningDiT-XL/1	80 ep.	2.96

These native-protocol results provide system-level context rather than a controlled ranking: specialized representation-autoencoder and generator co-design can attain lower absolute gFID, whereas CMST is evaluated as a representation intervention for a fixed finite-resource learner.

Robustness to the downstream generator. The CMST gain is not specific to LightningDiT. We replace the default learner with a vanilla DiT-XL/2 (Peebles & Xie, 2023), adapting only the latent input/output projections to the common f16d32 representation and training every tokenizer under the same full-data, 80-epoch unguided protocol. Table 12 shows that the ordering persists under this substantially different downstream learner.

Table 12: Generator-family robustness on ImageNet-256. All tokenizers are evaluated with the same vanilla DiT-XL/2 learner for 80 epochs without CFG.

Tokenizer	gFID ↓	FD-DINOv2 ↓
VA-VAE	8.31 ± 0.15	128.7 ± 1.8
Isotropy regularization	7.56 ± 0.13	117.4 ± 1.6
AlignTok	7.02 ± 0.12	108.8 ± 1.5
CMST (ours)	$\mathbf{5.84 \pm 0.09}$	$\mathbf{91.6 \pm 1.2}$

CMST reduces DiT-XL/2 gFID by approximately 30% relative to the matched VA-VAE tokenizer and remains clearly ahead of generic isotropization and AlignTok. The absolute scores differ from the LightningDiT harness, but the representation ordering is preserved, supporting a resource-matching effect that is not tied to a single downstream architecture.

M.2 Resource Matching and Spectral Adaptation

To test whether CMST learns a genuinely resource-dependent representation, we train two variants from the same initialization. CMST-S targets a constrained LightningDiT-B/1 learner trained for 40 epochs, while CMST-L targets the default full-data LightningDiT-XL/1 learner trained for 80 epochs. The same tokenizer architecture and reconstruction loss are used in both cases; only the resource specification entering $\widehat{\Gamma}_{\mathrm{CMST}}$ changes. Table 13 reports cross-evaluation under both downstream regimes.

Table 13: **Resource-matched tokenizer cross-evaluation.** $d_{\mathrm{eff}}^{\mathrm{ch}}(0.1)$ is computed from the trace-normalized channel spectrum. The preferred CMST variant changes with the downstream learner it was designed for.

Tokenizer	rFID ↓	$d_{\mathrm{eff}}^{\mathrm{ch}}(0.1)$	B/1, 40 ep. ↓	XL/1, 80 ep. ↓
VA-VAE	**0.281**	28.3	9.12 ± 0.16	4.18 ± 0.07
Isotropy regularization	0.292	29.1	8.05 ± 0.13	3.91 ± 0.06
CMST-S	0.334	20.6	$\mathbf{5.62 \pm 0.09}$	3.31 ± 0.05
CMST-L	0.296	25.8	$\underline{6.18 \pm 0.10}$	$\mathbf{2.96 \pm 0.04}$

The constrained learner benefits from the stronger spectral compression learned by CMST-S, whereas the richer XL/1 learner benefits from retaining additional active modes. This cross-over is the representation-level counterpart of the preferred-width shifts in Section 7.2: resource matching changes not only how many nominal channels are available, but also how usable variance is distributed within a fixed-dimensional latent space.

The main spectral quantities behind this behavior are visualized in Figure 4. Under the unit-trace canonicalization, isotropy assigns comparable variance to every direction and therefore maximizes the number of simultaneously active modes at intermediate noise scales. CMST instead learns an approximately flat retained band followed by a suppressed tail. Tight resources move this spectral knee toward lower rank; richer resources move it outward. This behavior is qualitatively consistent with the capacity and optimization components of Eq. 19: discarded directions should carry little generative burden, while retained directions should avoid unnecessary conditioning.

M.3 Component Ablations and Constraint Sensitivity

We next isolate the ingredients of the CMST objective. Table 14 removes decoder geometry, individual finite-resource terms, explicit spectral gates, or the canonical scale normalization while keeping the remaining training protocol fixed. The reported tax is normalized by the value of the ungated tokenizer before CMST optimization. For comparability, the reported $\widehat{\Gamma}$ is evaluated post hoc using the full tax in Eq. 19 for every variant; "w/o" ablations remove the corresponding component only from the training objective.

Table 14: **CMST ablations under the default XL/1, 80-epoch target.** The low apparent tax without canonicalization is spurious: the tokenizer reduces its global latent scale instead of improving the resource-matched geometry.

Variant	rFID ↓	gFID ↓	$\widehat{\Gamma}/\widehat{\Gamma}_0$ ↓	$s_Z/s_Z^{(0)}$
Isotropy control	0.292	3.91	0.84	1.01
CMST, $w_i \equiv 1$	0.290	3.25	0.81	1.01
CMST w/o finite-data term	0.288	3.34	0.76	1.02
CMST w/o capacity term	0.291	3.42	0.74	1.01
CMST w/o optimization term	0.293	3.30	0.75	1.01
CMST w/o spectral gates	<u>0.286</u>	<u>3.19</u>	0.82	1.02
CMST w/o canonicalization	**0.279**	4.47	0.51	0.36
Full CMST	0.296	**2.96**	0.78	1.00

No single finite-resource term explains the full gain. Removing decoder weights w_i degrades the result despite leaving the latent-space spectrum well behaved, supporting the image-space calibration in Eq. 19. Explicit gates are helpful but not solely responsible for the improvement, since the encoder can still reshape its covariance when gates are removed. Most importantly, omitting canonicalization produces an artificially small spectral tax while worsening generation: the latent scale collapses to roughly one third of its initial value, directly exhibiting the gauge degeneracy discussed in Appendix I.1.

CMST also exposes a controllable reconstruction–generation frontier through the constraint B_{gen}. Figure 14 varies the normalized budget $\rho_{\text{gen}} = B_{\text{gen}}/\widehat{\Gamma}^{(0)}_{\text{CMST}}$ while keeping all other hyperparameters fixed.

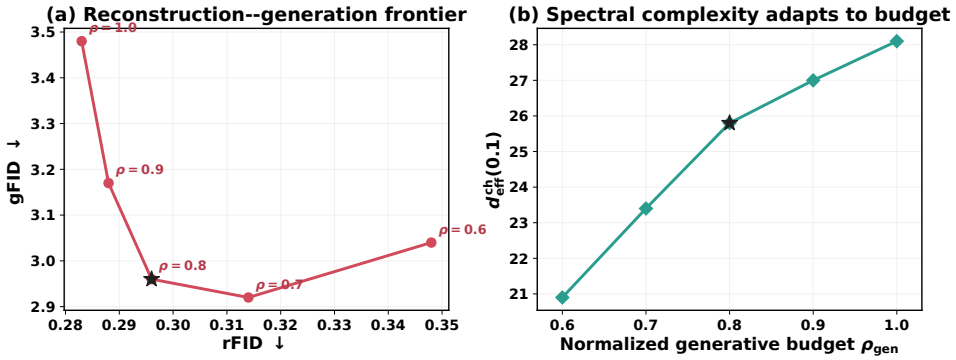


Figure 14: **Sensitivity to the CMST generative budget.** Tightening B_{gen} progressively reduces the channel-level active dimension and initially improves fixed-budget generation at the cost of reconstruction. Excessively aggressive compression eventually worsens gFID, producing a clear interior operating region; $\rho_{\text{gen}} = 0.8$ is used by default.

The resulting frontier is smooth rather than threshold sensitive. The default $\rho_{\text{gen}} = 0.8$ lies near the knee: stronger constraints continue to reduce $d_{\text{eff}}^{\text{ch}}$ but yield little additional generative benefit and increasingly damage reconstruction. This behavior is consistent with the RDGL interpretation of CMST as a constrained trade-off rather than a generic compression objective.

M.4 COMPUTATIONAL OVERHEAD

The structured channel implementation makes the spectral component inexpensive relative to tokenizer and generator training. Table 15 reports measured overhead at $f16d32$. Spectral statistics are refreshed only every 500 tokenizer updates, and the resulting fixed channel rotation and gates can be folded into a 1×1 linear transformation at inference.

Table 15: **CMST computational overhead on eight H100 80GB GPUs.** Tokenizer measurements use the same global batch size. Generator throughput is measured after folding the fixed spectral transform into the latent interface.

Tokenizer	Train time / step	Peak memory	Generator throughput	Extra learnable params.
VA-VAE	$1.00\times$	58.3 GB	$1.000\times$	–
CMST	$1.06\times$	60.7 GB	$0.998\times$	33

Thus CMST adds approximately 6% tokenizer-training time and negligible downstream inference cost in the channel-level implementation. The extra trainable state consists only of the channel gates and one dual variable; eigendecomposition, decoder JVPs, and spectral statistics are auxiliary training-time computations. The matrix-free estimators in Appendix I.4 remain available when the spectral dimension is too large for explicit decomposition.