V-SHAPED NEGATIVE ELBO IN VAES: RATE ALLOCATION AND AGGREGATE DEPENDENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

We uncover a counterintuitive pattern in variational autoencoders: in a controlled benchmark, reconstruction improves as encoder mutual information increases, yet the negative ELBO first decreases and then increases. Decomposing this objective reveals that aggregate total correlation (TC), a cost arising from dependence among code coordinates, grows enough to outweigh the combined improvement in all remaining terms, including reconstruction. Explaining this reversal requires understanding both the reconstruction benefit of additional information and the accompanying dependence cost. Our geometric analysis of optimal Gaussian coding quantifies how reconstruction gains diminish. To characterize the competing cost, we increase information by reducing Gaussian noise around a fixed representation and derive an exact law governing TC growth. Within this noise family, we establish an interior global minimum of the negative ELBO, bound the minimizing noise precision, and give sufficient conditions for uniqueness. Experiments support the predicted allocation of information and TC growth, while complete objective derivatives locate minima along controlled noise interventions. Two general image benchmarks exhibit similar V-shaped responses along measured KL-rate sweeps. These findings explain how increasing dependence can reverse the objective-level benefit of better reconstruction.

1 INTRODUCTION

Variational autoencoders (VAEs) encode an observation X in a stochastic code Z by maximizing the evidence lower bound (ELBO) on log data likelihood (Kingma and Welling, 2014; Rezende et al., 2014). Retaining more information can improve reconstruction while increasing coding costs. Rate–distortion analyses establish this trade-off (Bae et al., 2023; D'Amato et al., 2025) and show that comparable ELBOs can conceal different representations (Alemi et al., 2018). We ask which cost outweighs continued reconstruction gains, and whether an interior ELBO optimum can arise within a specified model family, with potential implications for principled selection of information levels in VAEs.

In our controlled sweeps, the best ELBO occurs at an intermediate information level, as reflected in the V-shaped response in Fig. 1. Although reconstruction improves along the rising branch, aggregate total correlation (TC), which measures dependence among latent coordinates across the data distribution, grows enough to outweigh the combined improvement in all remaining terms. Understanding this response requires both a law for dependence growth and an account of how that growth competes with reconstruction.

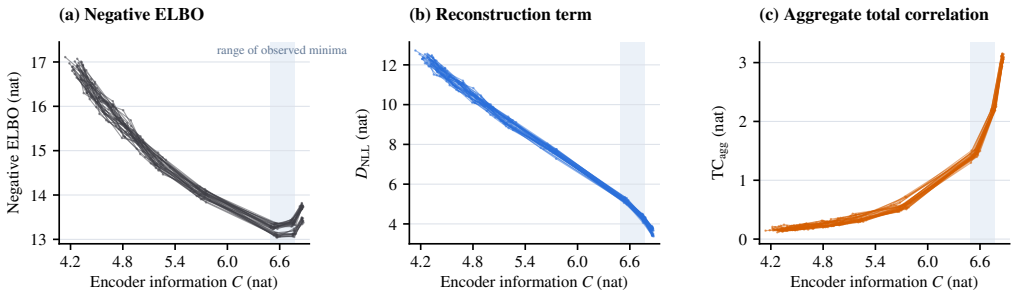
Let $C = I_q(X;Z)$ denote encoder mutual information in nats, and let R_{KL} be the average encoder KL to the prior. Write D_{NLL} for expected reconstruction negative log-likelihood, $\mathrm{TC}_{\mathrm{agg}}$ for the TC of the aggregate posterior $q(z)$, and DWKL for the sum of its marginal KL divergences to the prior. For a factorized prior, the negative ELBO V decomposes exactly as (Chen et al., 2018)

$$V \equiv -\mathrm{ELBO} = D_{\mathrm{NLL}} + R_{\mathrm{KL}} = D_{\mathrm{NLL}} + C + \mathrm{TC}_{\mathrm{agg}} + \mathrm{DWKL}. \tag{1}$$

With this objective, the central comparison is whether increasing TC outweighs improvement in $D_{\mathrm{NLL}} + C + \mathrm{DWKL}$, retaining both information and marginal prior-matching costs.

Prior work supplies the KL decomposition and methods for controlling dependence between and within coordinate groups (Kim and Mnih, 2018; Chen et al., 2018; Esmaeili et al., 2019). The

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.



$$\underbrace{V}_{\text{(a)}} = \underbrace{D_{\mathrm{NLL}}}_{\text{(b)}} + C + \mathrm{DWKL} + \underbrace{\mathrm{TC}_{\mathrm{agg}}}_{\text{(c)}}$$

Figure 1: **Better reconstruction can accompany a worse ELBO.** Eight runs with four prescribed conditions yield 32 trajectories of ten separately optimized models. (a) Negative ELBO exhibits a V-shaped response, with shading marking the range of observed minima. (b) Reconstruction NLL decreases while (c) aggregate TC increases. Panels share measured C, examples, and posterior draws; Gaussian losses omit the common likelihood constant.

decomposition alone does not explain how the balance between reconstruction and coding costs changes with retained information: across trained models, encoder means, covariances, and decoders can vary jointly, complicating the analysis of their distinct contributions. To disentangle these coupled changes, we isolate the reconstruction and dependence responses for separate analysis. Optimal local Gaussian allocation characterizes reconstruction gains, while Gaussian noise reduction around a fixed representation isolates how existing signal dependence becomes expressed in the noisy code. Within the noise family, we then retain the actual decoder response and all coding terms to analyze the complete objective and its interior minima. We compare all objective components on shared data across 320 separately trained models and test the dependence-growth law and minimum localization through noise interventions on 32 fixed models from the same trajectories. Experiments on dSprites and Cars3D assess whether V-shaped responses recur beyond the controlled setting.

Taken together, we discover a reproducible V-shaped response in VAEs and trace its origins through theoretical analysis. Our contributions are fourfold. **First,** covariance-weighted observation geometry characterizes optimal Gaussian allocation and predicts relative retention of generative factors across the evaluated nonlinear models. **Second,** we express TC growth per added nat as the advantage of joint over coordinatewise estimation and characterize strictness by signal dependence. **Third,** within the fixed Gaussian noise family, we establish an interior global minimum of the full negative ELBO even with a finite information ceiling, bound the minimizing precision, and give sufficient conditions for monotone branches and a unique global minimizer (Proposition 2). **Fourth, empirically,** component comparisons establish TC-driven increases in the controlled trained models, while noise interventions test the growth law and the localization of minima from complete objective derivatives.

2 Rate Allocation and Dependence

Section 2.1 defines the terms in Eq. (1). Sections 2.2–2.3 analyze D_{NLL} through optimal local Gaussian allocation and a likelihood-based lower bound. Section 2.4 derives $\mathrm{TC}_{\mathrm{agg}}$ growth with C under noise reduction. Section 2.5 retains C and DWKL in the complete objective V to establish interior optima and compare competing changes.

2.1 Setup and Rate Coordinates

Let $p_{\mathrm{data}}(x)$ be a fixed population or empirical data distribution, $q_\phi(z \mid x)$ the encoder, $p_\theta(x \mid z)$ the decoder likelihood, and $p(z) = \prod_{j=1}^{m} p(z_j)$ a factorized prior on an m-dimensional code. Write $q_C(z \mid x)$ for a given encoder with measured information C, and θ_C for its paired decoder. Its

aggregate posterior is $q_C(z) = \int q_C(z \mid x)p_{\mathrm{data}}(x)\,dx$, and $C = \mathbb{E}_X D_{\mathrm{KL}}(q_C(z \mid X)\|q_C(z))$. For independently trained models, C indexes evaluated operating points compared by finite differences.

The reconstruction term is $D_{\mathrm{NLL}}(C) = \mathbb{E}_{X,Z}[-\log p_{\theta_C}(X \mid Z)]$, with the expectation over the data and encoder. The average encoder KL decomposes as

$$R_{\mathrm{KL}}(C) \equiv \mathbb{E}_X D_{\mathrm{KL}}(q_C(z \mid X)\|p(z)) = C + D_{\mathrm{KL}}(q_C(z)\|p(z)). \tag{2}$$

Define $\mathrm{DWKL}(C) \equiv \sum_j D_{\mathrm{KL}}(q_C(z_j)\|p(z_j))$ and $\mathrm{TC}_{\mathrm{agg}}(C) \equiv D_{\mathrm{KL}}(q_C(z)\|\prod_j q_C(z_j))$ (Watanabe, 1960). In the specified latent coordinates, $R_{\mathrm{KL}} = C + \mathrm{TC}_{\mathrm{agg}} + \mathrm{DWKL}$, giving Eq. (1) (Chen et al., 2018). We focus on reconstruction and TC, as DWKL varies much less with C than either in the controlled right-branch comparisons (Section 3.4).

2.2 Local modes combine data variation and observation geometry

Limited information forces an encoder to prioritize variations by their reconstruction cost. We combine variability with observation sensitivity, measuring squared observation-space error before converting to NLL.

Let Y collect the underlying variables that generate an observation, such as an object's pose, scale, and shape. We call a degree or group of degrees of variation in Y a *generative factor*. This data-generating state differs from the encoder output Z. Model its possible values by a d-dimensional manifold $\mathcal{Y}$, with a smooth immersion $F : \mathcal{Y} \to \mathcal{X} \subset \mathbb{R}^n$ (a map with full-rank local Jacobian) generating $X = F(Y)$. In local coordinates around y, let δY be a centered displacement with covariance $\Sigma_{\mathrm{loc}}(y)$. For a realization δy, write $\delta x = F(y + \delta y) - F(y)$. The Jacobian $J_F(y)$ defines the pullback metric $G(y)$:

$$\delta x \approx J_F(y)\delta y, \qquad G(y) = J_F(y)^\top J_F(y). \tag{3}$$

Thus $\|\delta x\|^2 \approx \delta y^\top G(y)\delta y$ is the squared reconstruction error incurred by omitting this displacement (Arvanitidis et al., 2018; Kato et al., 2020; Nakagawa et al., 2021). Under a local Gaussian approximation, write $\delta Y = \Sigma_{\mathrm{loc}}^{1/2} u$, where $u \sim \mathcal{N}(0, I_d)$ and I_d is the identity matrix. Define the covariance-weighted metric $S(y)$ by

$$S(y) = \Sigma_{\mathrm{loc}}(y)^{1/2} G(y) \Sigma_{\mathrm{loc}}(y)^{1/2} \succeq 0. \tag{4}$$

Let $\lambda_1(y) \geq \cdots \geq \lambda_d(y) \geq 0$ be the eigenvalues of $S(y)$ and $v_j(y)$ their orthonormal eigenvectors. The eigenvalues form its *spectrum*. When $G = \mathrm{diag}(a_j^2)$ and $\Sigma_{\mathrm{loc}} = \mathrm{diag}(\nu_j)$, the mode costs are the products $\nu_j a_j^2$ of variance and squared observation sensitivity. A *local mode* is a variation direction $v_j(y)$ in the standardized displacement u, not a generative factor or a latent coordinate Z_j. The expected uncoded distortion is $\mathbb{E}[u^\top S(y) u] = \sum_j \lambda_j(y)$, so $\lambda_j(y)$ measures the expected squared reconstruction error caused by omitting that mode.

2.3 Diminishing reconstruction benefit under optimal Gaussian allocation

Global information C can encode both neighborhood identity and variation within a neighborhood. We isolate the latter through a local Gaussian coding problem with the neighborhood known to encoder and decoder. Its budget C_{loc} constrains the average conditional coding rate in this surrogate problem. Appendix A relates this construction to the mutual-information chain rule.

In each neighborhood, $w_j = \sqrt{\lambda_j}\, v_j^\top u$ is the scalar variation along mode j, weighted by its reconstruction cost. These variables are independent Gaussians with variances λ_j, and their squared reconstruction errors sum to local observation-space distortion. Following Gaussian rate–distortion theory (Berger, 1971; Cover and Thomas, 2006), a single average rate budget introduces one Lagrange multiplier and hence a common allocation threshold across neighborhoods.

Let $c_j(y) \geq 0$ be the information assigned to mode j at location y. A scalar Gaussian source with variance λ has distortion $D(c) = \lambda e^{-2c}$ at rate c (Berger, 1971; Cover and Thomas, 2006). Let $D^\star(C_{\mathrm{loc}})$ be the minimum expected local squared distortion and $c_j^\star$ its optimal allocation:

$$D^\star(C_{\mathrm{loc}}) = \min_{c_j(y)\geq 0} \mathbb{E}_Y\left[\sum_j \lambda_j(Y) e^{-2c_j(Y)}\right] \quad \text{s.t.} \quad \mathbb{E}_Y\left[\sum_j c_j(Y)\right] \leq C_{\mathrm{loc}}. \tag{5}$$

The Karush–Kuhn–Tucker (KKT) conditions yield *reverse water filling*: every active mode receives positive rate and retains residual error τ. The shared budget determines τ, and $[a]_+ = \max\{a, 0\}$.

Lemma 1 (Gaussian allocation with a shared rate budget) *Assume measurable eigenvalue fields with $0 < \mathbb{E}_Y \sum_j \lambda_j(Y) < \infty$. For every finite $C_{\mathrm{loc}} > 0$, a unique threshold $\tau(C_{\mathrm{loc}})$ gives*

$$c_j^\star(y; C_{\mathrm{loc}}) = \frac{1}{2}\left[\log \frac{\lambda_j(y)}{\tau(C_{\mathrm{loc}})}\right]_+, \qquad D^\star(C_{\mathrm{loc}}) = \mathbb{E}_Y\left[\sum_j \min\{\lambda_j(Y), \tau(C_{\mathrm{loc}})\}\right], \tag{6}$$

with zero rate assigned when $\lambda_j(y) = 0$. For every $C_{\mathrm{loc}} > 0$, $D^\star$ is differentiable and

$$\frac{dD^\star}{dC_{\mathrm{loc}}} = -2\tau(C_{\mathrm{loc}}). \tag{7}$$

The threshold decreases with C_{loc}, and a mode is active exactly when $\lambda_j(y) > \tau(C_{\mathrm{loc}})$.

Minimizing $\lambda e^{-2c} + 2\tau c$ over $c \geq 0$ gives residual τ for active modes and zero rate when $\lambda \leq \tau$ (Appendix A). Each active mode therefore offers the same marginal distortion reduction, 2τ per nat, while inactive modes offer no greater reduction.

Reconstruction contribution to the negative ELBO. For a Gaussian likelihood with fixed variance $\sigma_x^2 > 0$, NLL is scaled squared error plus the constant baseline $b_x = \frac{n}{2}\log(2\pi\sigma_x^2)$, where n counts observation coordinates. Optimal Gaussian coding gives

$$D_{\mathrm{NLL}}^\star(C_{\mathrm{loc}}) = \frac{D^\star(C_{\mathrm{loc}})}{2\sigma_x^2} + b_x. \tag{8}$$

The likelihood also gives $D_{\mathrm{NLL}} \geq b_x$ for every decoder, supplying the reconstruction lower bound used in Section 2.5. Using Eq. (7), define the Gaussian marginal reconstruction benefit

$$B_{\mathrm{G}}(C_{\mathrm{loc}}) \equiv -\frac{dD_{\mathrm{NLL}}^\star}{dC_{\mathrm{loc}}} = \frac{\tau(C_{\mathrm{loc}})}{\sigma_x^2}. \tag{9}$$

As C_{loc} grows, $\tau(C_{\mathrm{loc}})$ falls: each additional nat yields a smaller reduction in optimal reconstruction NLL. This local benchmark predicts diminishing marginal reconstruction gains and which local variations receive information first (Section 3.2). The following sections use global information C and the actual decoder response to analyze the complete ELBO.

2.4 Dependence growth under a fixed-representation intervention

Why fix the representation? Rising TC across trained models can reflect changes in mean maps, covariance shapes, and noise levels. To isolate the response to noise reduction, we fix the encoder mean map f, so that $U = f(X)$, and increase retained information by reducing independent Gaussian noise.

Aggregate TC, defined in Section 2.1, measures dependence after averaging over observations. Conditional TC measures dependence within the encoder distribution for a fixed observation:

$$\mathrm{TC}_{\mathrm{cond}}(C) \equiv \mathbb{E}_X D_{\mathrm{KL}}\left(q_C(z \mid X) \,\middle\|\, \prod_j q_C(z_j \mid X)\right). \tag{10}$$

Even independent Gaussian coordinates at each fixed x can have means that co-vary across the dataset, producing aggregate dependence (Chen et al., 2018; Esmaeili et al., 2019).

Lemma 2 (Coordinatewise information accounting) *For any encoder with finite relevant entropies, let $C_{\mathrm{coord}} = \sum_j I(X; Z_j)$ denote the sum of coordinatewise information. Then*

$$C + \mathrm{TC}_{\mathrm{agg}} = C_{\mathrm{coord}} + \mathrm{TC}_{\mathrm{cond}}. \tag{11}$$

A mean-field encoder has conditionally independent coordinates given X, so $\mathrm{TC}_{\mathrm{cond}} = 0$.

The difference $C_{\text{coord}} - C = \text{TC}_{\text{agg}} - \text{TC}_{\text{cond}}$ compares separately counted coordinate information with the joint total. In the independent-noise family below, $\text{TC}_{\text{cond}} = 0$, so comparing coordinatewise and joint information gains gives aggregate TC growth (Appendix B).

Let $U = f(X) \in \mathbb{R}^m$ be a nonconstant deterministic representation with $\mathbb{E}\|U\|^2 < \infty$. Let $t > 0$ be the inverse noise variance and let $\varepsilon \sim \mathcal{N}(0, I_m)$ be independent of X. Consider

$$Z_t = U + t^{-1/2}\varepsilon, \qquad C(t) \equiv I(X; Z_t) = I(U; Z_t). \tag{12}$$

Increasing the *noise precision* t reduces variance per coordinate and increases $C(t)$. The decoder and prior act on Z_t in these coordinates. Let $M_{\text{joint}}(t)$ be the minimum mean-square error in recovering the encoder signal U from the full noisy code. Let $M_{\text{sep}}(t)$ be the summed errors when each U_j uses only $Z_{t,j}$. Conditional means give the optimal estimators:

$$\begin{aligned} M_{\text{joint}}(t) &= \mathbb{E}\|U - \mathbb{E}[U \mid Z_t]\|^2, \\ M_{\text{sep}}(t) &= \sum_j \mathbb{E}\big(U_j - \mathbb{E}[U_j \mid Z_{t,j}]\big)^2. \end{aligned} \tag{13}$$

The gap $M_{\text{sep}} - M_{\text{joint}} \geq 0$ measures the mean-square error reduction from cross-coordinate dependence; normalized by M_{joint}, it gives TC growth per added nat (Eq. (14)).

Proposition 1 (Dependence under Gaussian noise reduction) *For the family in Eq.* (12)*,* $C'(t) > 0$ *and*

$$\frac{d\text{TC}_{\text{agg}}}{dC} = \frac{M_{\text{sep}}(t) - M_{\text{joint}}(t)}{M_{\text{joint}}(t)} \geq 0. \tag{14}$$

For every finite $t > 0$*, the inequality is strict if and only if the coordinates of* U *are not mutually independent.*

The mutual-information–MMSE identity gives $C'(t) = M_{\text{joint}}(t)/2$ and $C'_{\text{coord}}(t) = M_{\text{sep}}(t)/2$ (Guo et al., 2005). Subtracting these derivatives gives the TC response to noise precision; dividing by $C'(t)$ yields Eq. (14), the TC increase per added nat. For mutually independent coordinates of U, independent noise preserves independence, giving $M_{\text{sep}} = M_{\text{joint}}$ and zero slope. Reducing noise therefore strengthens dependence in the code by revealing the signal's shared structure.

Shared factors and coordinate groups. A factor can occupy several coordinates, producing within-group dependence even when groups are independent (Higgins et al., 2018; Pfau et al., 2020; Esmaeili et al., 2019). Such dependence can persist in encoder means despite regularization of sampled codes (Locatello et al., 2019), giving strict TC growth by Proposition 1. Appendices B.2 and C give the growth-law proof and group interpretation, respectively.

2.5 Interior optima and objective reversal

We now retain every term of the negative ELBO along the noise family in Eq. (12). Fix the prior $p(z) = \mathcal{N}(0, I_m)$ and likelihood $p_\theta(x \mid z) = \mathcal{N}(g_\theta(z), \sigma_x^2 I_n)$, with $\sigma_x^2 > 0$. Write $D(t) = D_{\text{NLL}}(t)$. The Gaussian encoder KL gives

$$V(t) = D(t) + \tfrac{1}{2}\mathbb{E}\|U\|^2 + \tfrac{m}{2}\psi(t), \qquad \psi(t) = t^{-1} - 1 + \log t.$$

Thus $m\psi(t)/2$ is the KL cost of the noise variance relative to the prior variance. It vanishes at $t = 1$ and diverges at both endpoints. Since $D(t) \geq b_x$ from Section 2.3, reconstruction cannot offset this divergence.

Proposition 2 (Interior attainment, localization, and shape) *Under the fixed-family assumptions above, if* D *is finite and continuous,* V *attains a global minimum over* $t > 0$*. Every minimizer satisfies* $0 < C(t^\star) < C_{\max} \equiv \lim_{t\to\infty} C(t)$*, even when* $C_{\max}$ *is finite, and lies in the compact interval*

$$\{t > 0 : \psi(t) \leq a_0\}, \qquad a_0 \equiv \frac{2[D(1) - b_x]}{m}.$$

If D *is continuously differentiable, define* $\mathcal{H}(t) = -2t^2 D'(t)$*. Then*

$$V'(t) = \frac{m(t-1) - \mathcal{H}(t)}{2t^2}. \tag{15}$$

If constants $-m < h_- \le h_+$ satisfy $h_- \le \mathcal{H}(t) \le h_+$ for all $t > 0$, every minimizer lies in $[1 + h_-/m,\ 1 + h_+/m]$. The objective strictly decreases below this interval and strictly increases above it. Separately, if $\mathcal{H}$ is differentiable and $\mathcal{H}'(t) < m$ for all $t > 0$, the global minimizer is unique, with strict descent before it and strict ascent afterward. These conclusions transfer through the increasing map $C(t)$.

The location bounds follow from two short comparisons. First, $V(t^\star) \le V(1)$ cancels the fixed mean cost and, using $\psi(1) = 0$, gives

$$\tfrac{m}{2}\psi(t^\star) \le D(1) - D(t^\star) \le D(1) - b_x.$$

Hence a_0 bounds the extra noise KL by the reconstruction improvement available from the unit-variance reference. Second, stationarity in Eq. (15) gives $t^\star = 1 + \mathcal{H}(t^\star)/m$, placing the minimizer in the stated interval whenever $h_- \le \mathcal{H} \le h_+$. The signed response $\mathcal{H}$ is twice the NLL derivative with respect to noise variance. If $\mathcal{H}' < m$, the numerator of V' is strictly increasing and crosses zero exactly once: the objective strictly decreases before the unique minimum and strictly increases afterward. Appendix B.4 provides the full proof and a Lipschitz-decoder threshold for eventual ascent.

Component contributions. Proposition 2 guarantees an interior minimum even for independent signals. To identify TC's contribution using Proposition 1, we decompose the complete derivative:

$$\frac{dV}{dC} = \underbrace{\frac{dD_{\mathrm{NLL}}}{dC}}_{\text{actual frozen-decoder response}} + \underbrace{\frac{M_{\mathrm{sep}}}{M_{\mathrm{joint}}}}_{\text{information + aggregate TC}} + \underbrace{\frac{d\mathrm{DWKL}}{dC}}_{\text{marginal mismatch}}. \tag{16}$$

Every derivative varies t, and the middle term is $1 + d\mathrm{TC}_{\mathrm{agg}}/dC$. At a differentiable minimizer, the actual decoder response balances all coding-cost changes.

Comparisons across independently trained models. We quantify TC's contribution across separately trained models by finite differences, allowing their representations to vary. For models a, b with $C_a < C_b$, write $\Delta Q = Q_b - Q_a$ for any component Q, and $\Delta C = C_b - C_a$. Let $s_{\mathrm{nonTC}} = \Delta(D_{\mathrm{NLL}} + C + \mathrm{DWKL})/\Delta C$ and $s_{\mathrm{TC}} = \Delta\mathrm{TC}_{\mathrm{agg}}/\Delta C$ be the average changes per nat in the non-TC sum and aggregate TC. The exact decomposition gives

$$m_{\mathrm{rev}} = s_{\mathrm{nonTC}} + s_{\mathrm{TC}} = \frac{\Delta V}{\Delta C}. \tag{17}$$

When $s_{\mathrm{nonTC}} < 0$, aggregate TC reverses the improvement in the remaining terms exactly when $s_{\mathrm{TC}} > -s_{\mathrm{nonTC}}$.

3 Experiments and Analysis

Data and evaluation. Known factors Y and a piecewise differentiable generator F supply the local covariances and observation sensitivities needed to test S. Synthetic studies optimize only encoders, using fixed generator-based decoders to compare them under a common reconstruction map. Factor labels and derivatives serve evaluation. The component study uses six scalar factors, interactions in pairs $(1, 2)$ and $(3, 4)$, and six Gaussian code coordinates. Four training masks test sensitivity to posterior covariance restrictions: Block permits covariance within both pairs, Diagonal within neither, and each Split mask disables its named pair. The allocation study has five factor targets and six latent coordinates (Appendix D).

For dSprites (Matthey et al., 2017) and Cars3D (Reed et al., 2015), convolutional mean-field VAEs have ten latent coordinates. Their training objective weights the full encoder KL:

$$\mathcal{L}_{\beta_R} = D_{\mathrm{NLL}} + \beta_R R_{\mathrm{KL}}, \qquad V = D_{\mathrm{NLL}} + R_{\mathrm{KL}}. \tag{18}$$

This is the β-VAE objective (Higgins et al., 2017). We evaluate unweighted V against achieved R_{KL}, using Bernoulli and factorized discretized-logistic likelihoods for dSprites and Cars3D, respectively (Salimans et al., 2017). Losses are compared within each dataset in nat/image.

3.1 THE V-SHAPED RESPONSE IN CONTROLLED MODELS

Eight independent training runs, each with four pre-specified partition conditions and ten control settings, yield 32 trajectories containing 320 separately optimized encoders. Every sweep has a minimum at an interior grid point (Fig. 1(a)). Component evaluation uses the empirical distribution assigning equal weight to 1,024 shared examples and the same posterior draws.

For each condition, reference setting M minimizes mean negative ELBO over the other seven runs, separating selection from evaluation. The endpoint H is the prespecified highest target-KL setting. On the held-out run, $(Q_H - Q_M)/(C_H - C_M)$ gives s_{nonTC}, s_{TC}, and m_{rev} for $Q = D_{\mathrm{NLL}} + C + \mathrm{DWKL}$, $\mathrm{TC}_{\mathrm{agg}}$, and V, respectively, using measured C. Equation (17) tests whether TC growth outweighs the remaining terms' net improvement (Fig. 4(a,b)).

3.2 GAUSSIAN ALLOCATION AND LEARNED FACTOR RETENTION

Gaussian channels with known optima. We numerically optimize rates across six Gaussian modes to test the allocation and diminishing gains predicted by Lemma 1 (Fig. 2(a,b)). Across three initializations, relative distortion error is at most 2.96×10^{-4}, and the full-activation grid discrepancy is 0.1554 nat. Active-mode counts are nondecreasing in all three initializations (Fig. 2(b)).

Which factors does a nonlinear VAE retain? We test predicted retention priorities in nonlinear VAEs, comparing S with G and Σ_{loc} to assess the benefit of combining variation and sensitivity. Each of eight seeds supplies six rate settings, six factor-space neighborhoods, and five evaluation targets, giving 180 model–neighborhood–factor records per seed.

Using factor-indexed strengths, we apply Eq. (6) with the measured encoder KL as the scoring budget and one S-calibrated threshold shared by all three predictors. The resulting *allocation score* predicts information assigned to a factor in a neighborhood. We measure factor retention using a covariance-based Gaussian canonical-correlation (CCA) information proxy from the full noisy code and test-set factor-prediction R^2 from a ridge-linear probe of the full encoder mean. The probe uses 2,048 fitting examples and 1,024 disjoint test examples per neighborhood.

Within each seed, Spearman correlations measure whether larger allocation scores correspond to higher CCA proxies or probe R^2 across the 180 records. The comparison spans rates, neighborhoods, and factors, including their common rate trends. Mean correlations over eight seeds are $0.879/0.689/0.631$ for CCA and $0.493/0.249/0.149$ for probe R^2, in $S/G/\Sigma_{\mathrm{loc}}$ order (Fig. 2(c)). The probe-ranking gain over G (0.244, positive in all eight seeds) supports combining variation and sensitivity to predict relative factor retention (Appendix F).

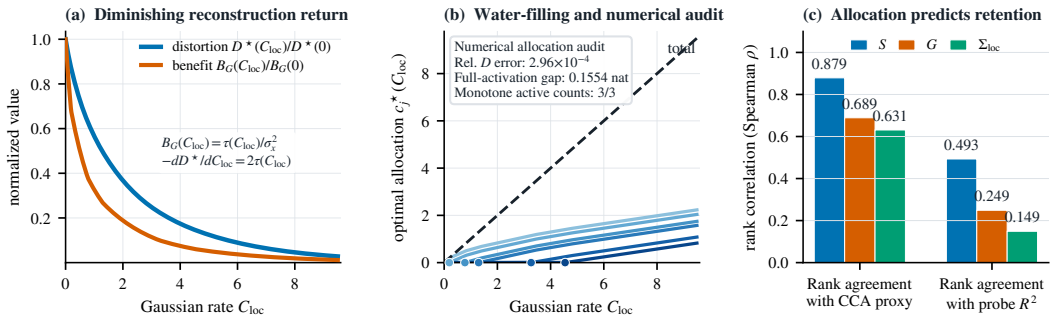


Figure 2: **Gaussian allocation and factor retention (Section 3.2).** (a) Analytic distortion and marginal benefit B_{G}. (b) Analytic allocations and numerical checks for the same spectrum. (c) Mean within-seed Spearman correlations (eight seeds) of allocation scores with the CCA proxy (left) and probe R^2 (right), across rates, neighborhoods, and factors.

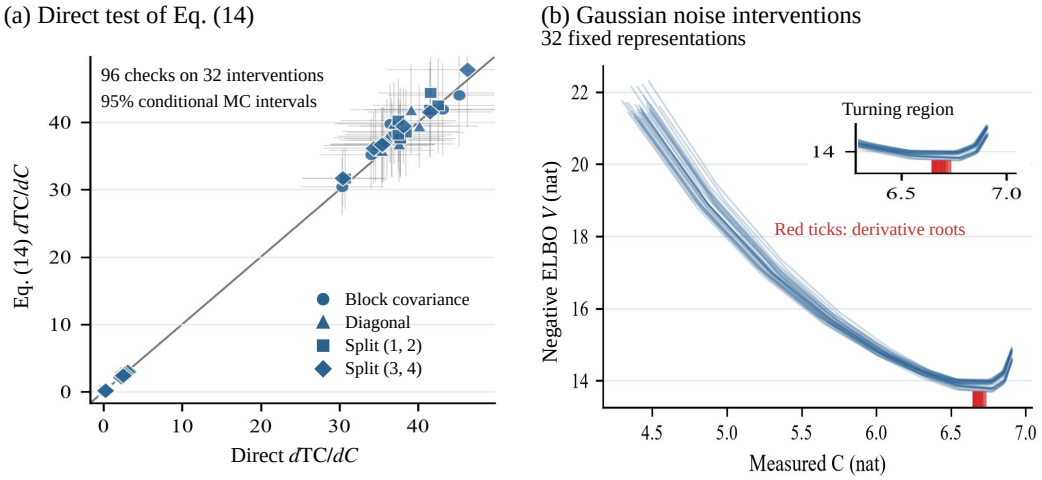


Figure 3: **Dependence growth and minimum localization.** (a) Density slopes versus Eq. (14), with pointwise 95% conditional Monte Carlo intervals (96 checks). Symbols denote training masks (Section 3); all interventions use isotropic noise. (b) 32 fixed models; red ticks mark negative-to-positive crossings of Eq. (16).

3.3 Dependence growth and reversal under Gaussian noise interventions

For each of the 32 controlled sweeps, we freeze the highest-information model's encoder mean map, decoder, prior, and likelihood. We replace its covariance by $t^{-1}I_m$ at ten noise precisions, measuring $C(t)$ and $V(t)$ to obtain 320 pairs. All evaluations share 1,024 inputs and four noise draws per input, without retraining.

Two calculations of dependence growth. For the empirical mixture of 1,024 frozen means, we compute $M_{\text{joint}}(t)$ and $M_{\text{sep}}(t)$ from joint and coordinatewise observations of U, respectively. We compare their Eq. (14) prediction with direct density derivatives.

Direct joint and marginal densities give C and TC at paired perturbations $s \pm 0.001$, where $s = \log t$. Their finite differences separately estimate the derivative ratio (Appendix E).

Across $32 \times 3 = 96$ checks at the first, seventh, and tenth precision levels, mean absolute discrepancy in $d\text{TC}_{\text{agg}}/ds$ is 0.0047 nat per unit s. Errors in $d\text{TC}_{\text{agg}}/dC$ are 0.0033, 0.0198, and 1.2039 at low, middle, and highest precision. Low- and middle-precision estimates agree; small dC/ds near $C_{\max} \leq \log 1024$ amplifies ratio errors. Finite support can also induce dependence absent from the population (Appendix E.1).

Locating minima and resolving the TC contribution. We differentiate the frozen decoder's loss by reparameterization (Kingma and Welling, 2014; Rezende et al., 2014) and evaluate the marginal-KL derivative directly. Together with the information and TC derivatives, these give Eq. (16). The derivative's negative-to-positive crossing provides a full-objective diagnostic, fixed before comparison with the sampled minimum. Across the 32 interventions, mean absolute error relative to the sampled minimum is 0.0967 nat. Every estimate lies between the samples neighboring that minimum (Fig. 3(b)). These ranges describe sampling resolution, not confidence intervals.

The same estimates closely match the minima across independently optimized VAEs in the original controlled sweeps: mean absolute error is 0.1242 nat, with all 32 estimates inside the corresponding neighboring-sample ranges. Substituting the local Gaussian benefit for the actual decoder response gives a larger intervention error of 0.8344 nat, with none of the 32 estimates inside those ranges. This contrast motivates retaining the measured decoder derivative. At the first sample after each intervention minimum, TC growth outweighs a negative combined slope of all remaining terms in all 32 cases (Appendix E).

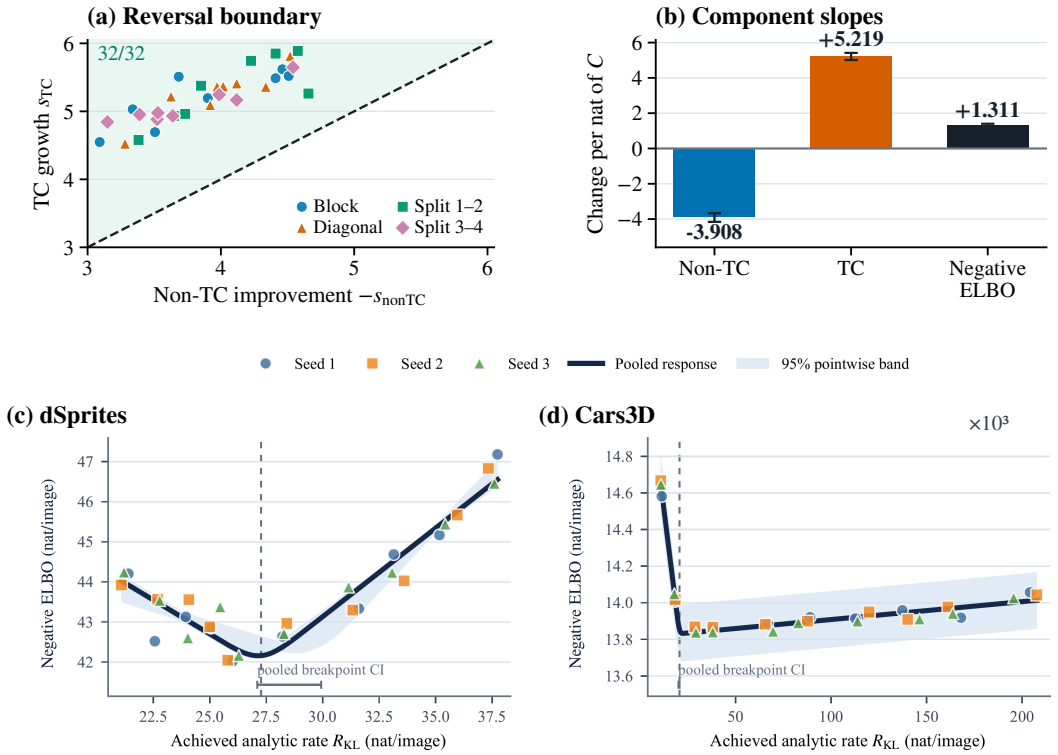


Figure 4: **TC-driven increases and V-shaped responses.** (a) $M \to H$ contrasts; symbols mark training masks. Above the diagonal, TC growth exceeds non-TC improvement. (b) Mean component slopes and run-clustered intervals. (c,d) Three seeds per dataset, pooled curves and pointwise hierarchical-bootstrap 95% bands; nat/image (Cars3D: $\times 10^3$). Cars3D tests use fixed checkpoints.

3.4 Component attribution on the original trained models

To quantify TC's contribution to the rising branch, we examine 32 paired $M \to H$ contrasts. Mean TC and non-TC slopes are $+5.219$ and -3.908 per nat of C, yielding a reversal slope of $+1.311$ (95% CI $[1.210, 1.403]$). All 32 contrasts lie above the reversal boundary (Fig. 4(a)).

DWKL contributes only $+0.016$ per nat of C, or 0.309% of the TC slope (95% CI $[0.202\%, 0.441\%]$; Appendix D), supporting our focus on reconstruction and TC in this paper.

3.5 V-shaped responses on external image benchmarks

At epoch 300, both image benchmarks show V-shaped responses in all three seeds (Fig. 4(c,d)). Pooled dSprites slopes are -0.340 and $+0.442$; the Cars3D intermediate-rate reference improves over low and high rates by 781.56 and 75.69 nat/image. All 95% intervals exclude zero. These recurrences are consistent with classical rate–distortion trade-offs (Alemi et al., 2018); TC attribution is established in the controlled models (Appendix D).

4 Conclusion

We trace the V-shaped negative ELBO in VAEs to aggregate dependence outweighing reconstruction gains. Local Gaussian allocation quantifies diminishing gains and predicts retention priorities, while fixed-representation noise analysis characterizes TC growth and establishes an interior optimum. Controlled experiments support these predictions and show agreement between intervention and trained-model minima. dSprites and Cars3D reproduce V-shaped responses along KL-rate sweeps.

Reproducibility Statement. Appendices A–C provide assumptions and proofs for allocation, dependence growth, interior attainment, and the geometric interpretations. Appendix D specifies the six-factor component benchmark, evaluation design, marginal-mismatch analysis, and external sweeps. Appendix E describes the fixed-model noise intervention, density calculations, decoder derivatives, minimum localization, and uncertainty estimates. Appendix F separates the generator, VAE, and factor readouts, and defines the allocation scores and ranking statistics. The Supplementary archive provides experimental code, configurations, and archived results, including row-level allocation readouts; its README describes the verification workflows and required inputs.

AI Use Statement. Generative AI tools assisted with literature discovery, proof verification, and English-language drafting and editing. The authors are responsible for reviewing all AI-assisted work and for the accuracy, originality, citations, and integrity of the final manuscript and its supporting materials.

REFERENCES

Alemi, A. A., Poole, B., Fischer, I., Dillon, J. V., Saurous, R. A., and Murphy, K. Fixing a broken ELBO. *ICML*, pp. 159–168, 2018.

Arvanitidis, G., Hansen, L. K., and Hauberg, S. Latent space oddity: On the curvature of deep generative models. *ICLR*, 2018.

Bae, J., Zhang, M. R., Ruan, M., Wang, E., Hasegawa, S., Ba, J., and Grosse, R. B. Multi-Rate VAE: Train once, get the full rate-distortion curve. *ICLR*, 2023.

Berger, T. *Rate Distortion Theory: A Mathematical Basis for Data Compression*. Prentice-Hall, 1971.

Bozkurt, A., Esmaeili, B., Tristan, J.-B., Brooks, D. H., Dy, J., and van de Meent, J.-W. Rate-regularization and generalization in VAEs. *AISTATS*, pp. 3880–3888, 2021.

Chen, R. T. Q., Li, X., Grosse, R., and Duvenaud, D. Isolating sources of disentanglement in variational autoencoders. *NeurIPS*, 2018.

Cover, T. M. and Thomas, J. A. *Elements of Information Theory*, 2nd ed. Wiley, 2006.

D'Amato, L., Lancia, G. L., and Pezzulo, G. The geometry of efficient codes: How rate-distortion trade-offs distort the latent representations of generative models. *PLOS Computational Biology*, 21(5):e1012952, 2025. doi: 10.1371/journal.pcbi.1012952.

Esmaeili, B., Wu, H., Jain, S., Bozkurt, A., Siddharth, N., Paige, B., Brooks, D. H., Dy, J., and van de Meent, J.-W. Structured disentangled representations. *AISTATS*, pp. 2525–2534, 2019.

Fumero, M., Cosmo, L., Melzi, S., and Rodolà, E. Learning disentangled representations via product manifold projection. *ICML*, pp. 3530–3540, 2021.

Guo, D., Shamai, S., and Verdú, S. 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.

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., and Lerchner, A. β-VAE: Learning basic visual concepts with a constrained variational framework. *ICLR*, 2017.

Higgins, I., Amos, D., Pfau, D., Racanière, S., Matthey, L., Rezende, D., and Lerchner, A. Towards a definition of disentangled representations. *arXiv preprint arXiv:1812.02230*, 2018.

Kato, K., Zhou, J., Sasaki, T., and Nakagawa, A. Rate-distortion optimization guided autoencoder for isometric embedding in Euclidean latent space. *ICML*, pp. 5166–5176, 2020.

Kim, H. and Mnih, A. Disentangling by factorising. *ICML*, 2018.

Kingma, D. P. and Welling, M. Auto-encoding variational Bayes. *ICLR*, 2014.

Kumar, A. and Poole, B. On implicit regularization in β-VAEs. *ICML*, pp. 5480–5490, 2020.

Locatello, F., Bauer, S., Lucic, M., Rätsch, G., Gelly, S., Schölkopf, B., and Bachem, O. Challenging common assumptions in the unsupervised learning of disentangled representations. *ICML*, pp. 4114–4124, 2019.

Matthey, L., Higgins, I., Hassabis, D., and Lerchner, A. dSprites: Disentanglement testing sprites dataset. 2017. URL `https://github.com/google-deepmind/dsprites-dataset`.

Nakagawa, A., Kato, K., and Suzuki, T. Quantitative understanding of VAE as a non-linearly scaled isometric embedding. *ICML*, pp. 7916–7926, 2021.

Pfau, D., Higgins, I., Botev, A., and Racanière, S. Disentangling by subspace diffusion. *NeurIPS*, 2020.

Reed, S. E., Zhang, Y., Zhang, Y., and Lee, H. Deep visual analogy-making. *Advances in Neural Information Processing Systems*, 28, 2015.

Rezende, D. J., Mohamed, S., and Wierstra, D. Stochastic backpropagation and approximate inference in deep generative models. *ICML*, pp. 1278–1286, 2014.

Rolinek, M., Zietlow, D., and Martius, G. Variational autoencoders pursue PCA directions (by accident). *CVPR*, 2019.

Salimans, T., Karpathy, A., Chen, X., and Kingma, D. P. PixelCNN++: Improving the PixelCNN with discretized logistic mixture likelihood and other modifications. *ICLR*, 2017.

Shu, R. and Ermon, S. Bit prioritization in variational autoencoders via progressive coding. *ICML*, pp. 20141–20155, 2022.

Träuble, F., Creager, E., Kilbertus, N., Locatello, F., Dittadi, A., Goyal, A., Schölkopf, B., and Bauer, S. On disentangled representations learned from correlated data. *ICML*, pp. 10401–10412, 2021.

Watanabe, S. Information theoretical analysis of multivariate correlation. *IBM Journal of Research and Development*, 4(1):66–82, 1960.

Xiao, T. Z. and Bamler, R. Trading information between latents in hierarchical variational autoencoders. *ICLR*, 2023.

A Proof of the Gaussian Allocation Lemma

Local-channel interpretation. For a finite neighborhood label $J = J(X)$ determined by the observation, the mutual-information chain rule gives, for the same encoder,

$$C = I(X;Z) = I(J;Z) + I(X;Z \mid J).$$

The first term measures information about neighborhood identity, and the second measures information retained within neighborhoods. Equation (5) models the latter task through a separate Gaussian surrogate: the neighborhood is supplied to both encoder and decoder, and its conditional modes are independent Gaussians. The budget C_{loc} bounds the expected coding rate in that surrogate. It is not automatically equal to $I(X;Z \mid J)$ for a trained nonlinear encoder. All neighborhoods share one expected budget and hence one threshold. The Gaussian experiment numerically optimizes rates in the independent-channel setting. The nonlinear experiment calibrates the surrogate's allocation scores using measured encoder KL and tests their retention rankings.

The covariance-weighted metric follows directly from the local quadratic distortion. Under the linearized observation map and centered local displacements,

$$\mathbb{E}[\|\delta X\|_2^2 \mid Y = y] \approx \operatorname{tr}(G(y)\Sigma_{\mathrm{loc}}(y)) = \operatorname{tr} S(y) = \sum_j \lambda_j(y). \tag{19}$$

This is the distortion incurred when the local displacement is omitted entirely. To express partial reconstruction, let $\widehat{u}$ be a reconstruction of the standardized displacement u in the same neighborhood.

Its intrinsic-coordinate residual is $\Sigma_{\mathrm{loc}}^{1/2}(u-\widehat{u})$. Define $w_j = \sqrt{\lambda_j}\, v_j^\top u$ and $\widehat{w}_j = \sqrt{\lambda_j}\, v_j^\top \widehat{u}$. Under the local linearization, the squared observation-space reconstruction error is

$$(u-\widehat{u})^\top S(y)(u-\widehat{u}) = \sum_j \lambda_j \big(v_j^\top (u-\widehat{u})\big)^2 = \sum_j (w_j - \widehat{w}_j)^2.$$

The independent weighted modes have variances λ_j, which justifies the scalar Gaussian distortions in Eq. (5). For an eigenvector $v_j(y)$ with $\lambda_j(y) > 0$, define $e_j(y) = \Sigma_{\mathrm{loc}}(y)^{1/2} v_j(y)/\sqrt{\lambda_j(y)}$. Then $e_i(y)^\top G(y) e_j(y)$ equals one when $i = j$ and zero otherwise. These are mutually orthonormal tangent directions under the observation metric.

Existence and uniqueness of the threshold. For $\tau > 0$, define the rate assigned by the proposed allocation as

$$\mathcal{R}(\tau) \equiv \mathbb{E}_Y \sum_j \frac{1}{2} \left[\log \frac{\lambda_j(Y)}{\tau}\right]_+, \tag{20}$$

with a zero contribution when $\lambda_j(Y) = 0$. The inequality $[\log a]_+ \le a$ for $a > 0$ gives $\mathcal{R}(\tau) \le \mathbb{E}_Y \sum_j \lambda_j(Y)/(2\tau) < \infty$. On any compact interval of positive thresholds, the same integrable bound permits dominated convergence, so $\mathcal{R}$ is continuous. This bound also gives $\mathcal{R}(\tau) \to 0$ as $\tau \to \infty$. Since the eigenvalues have positive total expectation, at least one is positive on a set of positive probability. Monotone convergence then gives $\mathcal{R}(\tau) \to \infty$ as $\tau \downarrow 0$.

The function is strictly decreasing wherever it is positive. In fact, if $0 < \tau_1 < \tau_2$ and $\mathcal{R}(\tau_2) > 0$, the set of modes with $\lambda_j(Y) > \tau_2$ has positive expected count. Each such mode contributes $\frac{1}{2}\log(\tau_2/\tau_1) > 0$ to $\mathcal{R}(\tau_1) - \mathcal{R}(\tau_2)$. Continuity and these limits establish a unique solution of $\mathcal{R}(\tau) = C_{\mathrm{loc}}$ for every finite $C_{\mathrm{loc}} > 0$. The threshold, expressed as the inverse function $\tau(C_{\mathrm{loc}})$, decreases with C_{loc}.

Optimality of the allocation. Fix this threshold. At each location and mode, minimize the convex scalar function $\lambda_j(y)e^{-2c} + 2\tau c$ over $c \ge 0$. For an active mode, stationarity gives $\lambda_j(y)e^{-2c} = \tau$. At the boundary, $c = 0$ is optimal when $\lambda_j(y) \le \tau$. These are exactly the two branches of Eq. (6). To verify global optimality, write $D(c)$ and $R(c)$ for the expected distortion and rate of any feasible allocation. Pointwise minimization and integration yield

$$D(c^\star) + 2\tau C_{\mathrm{loc}} \le D(c) + 2\tau R(c) \le D(c) + 2\tau C_{\mathrm{loc}}. \tag{21}$$

Thus $D(c^\star) \le D(c)$. Substituting the allocation gives the clipped-spectrum expression for $D^\star(C_{\mathrm{loc}})$. The multiplier 2τ is the marginal reduction in distortion available from relaxing the rate constraint.

Marginal distortion and the Gaussian likelihood. Let $N(\tau) \equiv \mathbb{E}_Y \sum_j \mathbf{1}\{\lambda_j(Y) > \tau\}$ be the expected number of active modes. At thresholds for which $\lambda_j(Y) = \tau$ has zero probability for every j, differentiation under the expectation gives

$$\frac{d\mathcal{R}}{d\tau} = -\frac{N(\tau)}{2\tau}, \qquad \frac{d}{d\tau}\mathbb{E}_Y \sum_j \min\{\lambda_j(Y), \tau\} = N(\tau). \tag{22}$$

Their ratio proves Eq. (7) away from activation thresholds. At a threshold, scalar optimality gives $D^\star(C_2) \ge D^\star(C_1) - 2\tau(C_1)(C_2 - C_1)$ for positive local budgets C_1, C_2. Applying this inequality in both directions and using continuity of τ proves the same derivative there. For an n-dimensional Gaussian observation model with fixed variance σ_x^2, the reconstruction NLL is

$$D_{\mathrm{NLL}}(C) = \frac{\mathbb{E}\|X - \widehat{x}_{\theta_C}(Z)\|^2}{2\sigma_x^2} + \frac{n}{2}\log(2\pi\sigma_x^2). \tag{23}$$

For the local Gaussian problem, optimal encoder–decoder pairs attain $D^\star(C_{\mathrm{loc}})$. Substitution into Eq. (23) gives the reconstruction derivative $-\tau(C_{\mathrm{loc}})/\sigma_x^2$, proving Eq. (9).

Within the same Gaussian problem, let $D_{\mathrm{rec}}(C_{\mathrm{loc}})$ be the squared distortion of a feasible family and define $\epsilon_D(C_{\mathrm{loc}}) = D_{\mathrm{rec}}(C_{\mathrm{loc}}) - D^\star(C_{\mathrm{loc}}) \ge 0$. Wherever derivatives exist,

$$-\frac{dD_{\mathrm{NLL}}}{dC_{\mathrm{loc}}} = \frac{\tau(C_{\mathrm{loc}})}{\sigma_x^2} - \frac{\epsilon_D'(C_{\mathrm{loc}})}{2\sigma_x^2}.$$

The derivative ϵ_D' gives the correction to the optimal marginal benefit and can have either sign. Accordingly, Eq. (16) uses the measured decoder response along the fixed-model intervention.

Relation to earlier VAE allocation analyses. Rolinek et al. (2019) relate diagonal posterior noise to local decoder orthogonality and PCA-like directions. For hierarchical VAEs, Shu and Ermon (2022) prioritize information useful for progressive lossy compression, while Xiao and Bamler (2023) allocate rates across latent layers and study their downstream consequences. Here the allocation assigns information to each mode of the covariance-weighted metric in each intrinsic-state neighborhood. The nonlinear experiment tests this choice against observation sensitivity and factor variation separately. The numerical Gaussian allocation check tests its distortion and full-activation predictions.

B Information Accounting, Dependence Growth, and Interior Optima

B.1 Coordinatewise identity and extension to groups

The information coordinate is the expected KL from a conditional encoder to its aggregate posterior:

$$C = I_{q_C}(X;Z) = \mathbb{E}_X D_{\mathrm{KL}}(q_C(z \mid X) \| q_C(z)). \tag{24}$$

We prove the more general identity for coordinate groups and recover Lemma 2 as its singleton case. Fix a partition $\pi = \{B_1, \dots, B_K\}$ of the m latent coordinates, and let Z_{B_k} denote the subvector indexed by B_k. Define

$$\begin{aligned} \mathrm{TC}_{\mathrm{agg},\pi} &= D_{\mathrm{KL}}\left(q_C(z) \,\middle\|\, \prod_k q_C(z_{B_k}) \right), \\ \mathrm{TC}_{\mathrm{cond},\pi} &= \mathbb{E}_X D_{\mathrm{KL}}\left(q_C(z \mid X) \,\middle\|\, \prod_k q_C(z_{B_k} \mid X) \right). \end{aligned} \tag{25}$$

Let h denote differential entropy, with $h(Z_{B_k} \mid X = x)$ evaluated under the conditional law at a fixed observation. When the relevant entropies are finite,

$$\begin{aligned} \sum_k I(X; Z_{B_k}) - I(X;Z) = &\left[\sum_k h(Z_{B_k}) - h(Z) \right] \\ &- \mathbb{E}_X \left[\sum_k h(Z_{B_k} \mid X = x) - h(Z \mid X = x) \right]. \end{aligned} \tag{26}$$

The first bracket measures aggregate block TC, and the expected second bracket measures conditional block TC. Hence

$$\sum_k I(X; Z_{B_k}) - C = \mathrm{TC}_{\mathrm{agg},\pi} - \mathrm{TC}_{\mathrm{cond},\pi}. \tag{27}$$

Taking singleton blocks gives $C_{\mathrm{coord}} - C = \mathrm{TC}_{\mathrm{agg}} - \mathrm{TC}_{\mathrm{cond}}$, which is Lemma 2. A mean-field encoder factorizes over coordinates at every x, and hence over any partition, making the conditional term zero. The identity compares terms at a fixed encoder. To analyze growth as information increases, we now specify a family of encoders.

B.2 Proof of Proposition 1

Let $U = f(X) \in \mathbb{R}^m$ be a measurable representation with $\mathbb{E}\|U\|^2 < \infty$ that is not almost surely constant under the specified data law; equivalently, $\mathrm{tr}\,\mathrm{Cov}(U) > 0$. This is the meaning of nonconstant throughout the fixed-noise-family analysis. For $t > 0$, let $Z_t = U + t^{-1/2}\varepsilon$, where $\varepsilon \sim \mathcal{N}(0, I_m)$ is independent of X. We again prove the groupwise extension for a fixed partition $\pi = \{B_1, \dots, B_K\}$. Write $\mathrm{TC}_\pi(Z_t)$ for the aggregate block TC from Eq. (25) evaluated on this family. The additive Gaussian density and finite second moment make the relevant output entropies and information quantities finite, including for non-Gaussian, singular, or discrete U.

For each block, define the separate conditional-mean estimator $\widehat{U}^{\,\mathrm{sep}}_{B_k} = \mathbb{E}[U_{B_k} \mid Z_{t,B_k}]$. Define the joint estimator by $\widehat{U}^{\,\mathrm{joint}} = \mathbb{E}[U \mid Z_t]$ and let $\widehat{U}^{\,\mathrm{joint}}_{B_k}$ be its block restriction. Conditional means

minimize squared error. Therefore

$$M_{\mathrm{sep},\pi}(t) = \sum_k \mathbb{E}\|U_{B_k} - \widehat{U}^{\,\mathrm{sep}}_{B_k}\|^2,$$
$$M_{\mathrm{joint}}(t) = \mathbb{E}\|U - \widehat{U}^{\,\mathrm{joint}}\|^2 = \sum_k \mathbb{E}\|U_{B_k} - \widehat{U}^{\,\mathrm{joint}}_{B_k}\|^2. \tag{28}$$

The Markov chains $X \to U \to Z_t$ and $X \to U_{B_k} \to Z_{t,B_k}$ have deterministic first steps. Consequently, $I(X;Z_t) = I(U;Z_t)$ and $I(X;Z_{t,B_k}) = I(U_{B_k};Z_{t,B_k})$. The noise blocks are independent, so conditional TC is zero. Equation (27) gives

$$\mathrm{TC}_\pi(Z_t) = \sum_k I(U_{B_k}; \sqrt{t}\,U_{B_k} + \varepsilon_{B_k}) - I(U; \sqrt{t}\,U + \varepsilon). \tag{29}$$

Rescaling each block by $\sqrt{t}$ preserves its information and TC. The vector I–MMSE theorem therefore applies to each block and the full vector for every $t > 0$ (Guo et al., 2005, Theorem 2). Using natural logarithms,

$$C'(t) = \frac{1}{2} M_{\mathrm{joint}}(t), \qquad \frac{d}{dt}\mathrm{TC}_\pi(Z_t) = \frac{1}{2}\left[M_{\mathrm{sep},\pi}(t) - M_{\mathrm{joint}}(t)\right]. \tag{30}$$

The signal distribution and representation map remain fixed throughout these derivatives.

To establish the sign and its exact strictness condition, apply the orthogonality property of conditional expectation. The joint estimation error is orthogonal to every square-integrable function of Z_t. Expanding the difference between separate and joint squared errors thus gives

$$M_{\mathrm{sep},\pi}(t) - M_{\mathrm{joint}}(t) = \sum_k \mathbb{E}\left\|\widehat{U}^{\,\mathrm{joint}}_{B_k} - \widehat{U}^{\,\mathrm{sep}}_{B_k}\right\|^2 \geq 0. \tag{31}$$

Equality holds exactly when every block's conditional mean is unchanged by observing the other blocks, almost surely. Strict growth therefore means that observing other blocks reduces the minimum mean-square error for at least one block.

Equality holds exactly for independent signal blocks. Fix a finite $t > 0$. Let p_t be the density of Z_t and let $p_{t,k}$ be its marginal density on block B_k. Gaussian convolution makes these densities strictly positive and smooth. Differentiating their convolution formulas gives

$$\nabla_{z_{B_k}} \log p_t(z) = t\big(\widehat{U}^{\,\mathrm{joint}}_{B_k}(z) - z_{B_k}\big),$$
$$\nabla_{z_{B_k}} \log p_{t,k}(z_{B_k}) = t\big(\widehat{U}^{\,\mathrm{sep}}_{B_k}(z_{B_k}) - z_{B_k}\big).$$

If equality holds in Eq. (31), the two conditional means coincide almost surely for every block. Hence the gradient of $\log[p_t(z)/\prod_k p_{t,k}(z_{B_k})]$ vanishes almost everywhere. Positivity and smoothness extend this equality to every $z \in \mathbb{R}^m$, so the density ratio is constant. Normalization gives $p_t(z) = \prod_k p_{t,k}(z_{B_k})$.

To recover independence of the signal blocks, let $\varphi_U(\omega) = \mathbb{E}e^{\mathrm{i}\omega^\top U}$ be the characteristic function of U, with $\mathrm{i}^2 = -1$, and define $\varphi_{U_{B_k}}$ analogously. The output factorization implies

$$\varphi_U(\omega) e^{-\|\omega\|^2/(2t)} = \prod_k \left[\varphi_{U_{B_k}}(\omega_{B_k}) e^{-\|\omega_{B_k}\|^2/(2t)}\right].$$

Cancelling the nonzero Gaussian factor yields $\varphi_U(\omega) = \prod_k \varphi_{U_{B_k}}(\omega_{B_k})$, which characterizes mutual independence of the signal blocks. Conversely, independent signal blocks and independent noise give a factorized posterior, so the joint and separate estimators coincide. Thus equality at any finite $t > 0$ is equivalent to mutual independence of the blocks of U. Otherwise, the estimation advantage is strictly positive at every finite $t > 0$. The argument requires neither a density nor Gaussianity of U.

For a nonconstant U, $M_{\mathrm{joint}}(t) > 0$ at finite t. Indeed, the nonsingular Gaussian likelihood is strictly positive on the support of U, so conditioning on a finite noisy observation cannot turn a nonconstant prior law into a point mass. Thus $C'(t) > 0$ and t can locally be reparameterized by C. Dividing the two derivatives in Eq. (30) gives the groupwise result

$$\frac{d\mathrm{TC}_\pi(Z_t)}{dC} = \frac{M_{\mathrm{sep},\pi}(t) - M_{\mathrm{joint}}(t)}{M_{\mathrm{joint}}(t)} \geq 0.$$

For singleton blocks, $M_{\mathrm{sep},\pi} = M_{\mathrm{sep}}$ and $\mathrm{TC}_\pi = \mathrm{TC}_{\mathrm{agg}}$, proving Eq. (14). The equality characterization above establishes the stated necessary and sufficient condition for strict growth.

An equivalent data-processing argument. For $0 < t_1 < t_2$, construct an independent standard Gaussian vector η and set $\widetilde{Z}_{t_1} = Z_{t_2} + \sqrt{t_1^{-1} - t_2^{-1}}\,\eta$. The joint law of $(U, \widetilde{Z}_{t_1})$ equals that of (U, Z_{t_1}). This is a product channel across blocks. Applying KL contraction to the joint output law and the product of its block marginals yields $\mathrm{TC}_\pi(Z_{t_1}) \le \mathrm{TC}_\pi(Z_{t_2})$ (Cover and Thomas, 2006). Ordinary data processing also yields $C(t_1) \le C(t_2)$. The I–MMSE argument adds the explicit slope and strictness condition.

B.3 Scope, counterexamples, and the objective

Proposition 1 applies to any fixed nonconstant signal law with finite second moment under independent Gaussian noise with a common changing scale. A fixed positive diagonal noise covariance can be reduced to this setting by coordinatewise rescaling, which preserves every block TC. Input-dependent noise, changes in relative noise scales across coordinates, and changes in encoder maps require separate analysis.

If the blocks of U are independent, the blocks of Z_t remain independent for all t. Then $M_{\mathrm{sep},\pi} = M_{\mathrm{joint}}$ and TC remains zero although C increases. This rules out strict monotonicity without a dependence condition. Even with dependence, strict growth alone does not imply that the slope increases with C.

Changes in the signal map can even reverse the sign of the TC response to information. Let $Y_0 \sim \mathcal{N}(0,1)$. At $t = 1$, the map $U = (Y_0, Y_0)$ gives $C = \frac{1}{2}\log 3$ and $\mathrm{TC} = \log 2 - \frac{1}{2}\log 3 > 0$. At $t = 2$, the map $U = (\sqrt{2}Y_0, 0)$ gives $C = \frac{1}{2}\log 5$ and $\mathrm{TC} = 0$. Both conditional encoders have independent isotropic Gaussian noise of variance $1/t$. Thus higher information can coexist with lower TC when the encoder map changes. The second map is a rotation of the first signal representation, illustrating the dependence of coordinatewise TC on the chosen coordinates.

For any encoder with finite relevant terms, combining the objective decomposition with Eq. (11) gives

$$\begin{aligned} -\mathrm{ELBO} &= D_{\mathrm{NLL}} + C + \mathrm{DWKL} + \mathrm{TC}_{\mathrm{agg}} \\ &= D_{\mathrm{NLL}} + \mathrm{DWKL} + C_{\mathrm{coord}} + \mathrm{TC}_{\mathrm{cond}}. \end{aligned} \tag{32}$$

The second line also applies to full-covariance encoders. Its conditional-TC term vanishes for the independent-noise family.

For singleton blocks, differentiating Eq. (32) along Eq. (12) and using I–MMSE gives $dC_{\mathrm{coord}}/dC = M_{\mathrm{sep}}/M_{\mathrm{joint}}$. The independent noise makes $\mathrm{TC}_{\mathrm{cond}} = 0$, proving Eq. (16). Locating a zero of this objective derivative requires the decoder reconstruction derivative and marginal-KL derivative as well as the dependence slope. The intervention in Appendix E evaluates all three contributions in the same, unscaled latent coordinates. Information and TC are invariant under the blockwise rescaling used in the proof, but the decoder likelihood and prior mismatch generally are not.

For a fixed finite dataset, encoder mutual information under its empirical distribution is bounded by the entropy of the sample identity. The general monotonicity result still applies, but the unbounded high-rate asymptotics in Appendix C.5 describe a continuous Gaussian population model. They are not asymptotic statements about the finite external image sweeps.

Observed responses and optimal budget values. The trained response compares separately optimized models ordered by measured C. It is distinct from the globally optimal value $V_{\le}(c) = \inf_{\phi,\theta:\, I_{q_\phi}(X;Z)\le c}[-\mathrm{ELBO}(\phi,\theta)]$ under an information upper bound. Increasing c enlarges this feasible set, so $V_{\le}$ is non-increasing. The observed rising branch therefore concerns the evaluated models and interventions, not this nested-constraint optimal value.

B.4 Interior attainment, localization, and shape under additional assumptions

We prove Proposition 2 for the same fixed signal and unscaled latent coordinates as Eq. (12). All expectations use one fixed data law, which may be a population distribution or a finite empirical distribution. The decoder mean g_θ, its variance σ_x^2, and the standard normal prior remain fixed. No Gaussian assumption on U is needed.

Reconstruction lower bound and exact encoder KL. The Gaussian likelihood gives, for every $t > 0$,

$$D_{\mathrm{NLL}}(t) = b_x + \frac{\mathbb{E}\|X - g_\theta(U + t^{-1/2}\varepsilon)\|^2}{2\sigma_x^2} \geq b_x, \qquad b_x = \frac{n}{2}\log(2\pi\sigma_x^2). \tag{33}$$

This lower bound follows directly from the fixed Gaussian likelihood. The constant b_x may be negative because a continuous density can exceed one. Subtracting it translates the objective without changing any minimizer.

For $q_t(z \mid x) = \mathcal{N}(f(x), t^{-1}I_m)$, the Gaussian KL formula gives (Kingma and Welling, 2014)

$$\begin{aligned} R_{\mathrm{KL}}(t) &= \tfrac{1}{2}\mathbb{E}\|U\|^2 + \tfrac{m}{2}\psi(t), \qquad \psi(t) = t^{-1} - 1 + \log t, \\ V(t) &= D_{\mathrm{NLL}}(t) + R_{\mathrm{KL}}(t) \geq b_x + \tfrac{1}{2}\mathbb{E}\|U\|^2 + \tfrac{m}{2}\psi(t). \end{aligned} \tag{34}$$

Since $\psi(t) \to \infty$ as $t \downarrow 0$ and as $t \to \infty$,

$$\lim_{t\downarrow 0} V(t) = \lim_{t\to\infty} V(t) = +\infty. \tag{35}$$

At the first endpoint, the encoder variance diverges. At the second, its conditional covariance collapses relative to the nonsingular prior. Either incurs an unbounded KL cost, while reconstruction has only a finite amount to gain relative to any finite reference loss.

Attainment inside the precision domain. The assumed finiteness of reconstruction makes $V(1)$ finite. Equation (35) supplies $0 < a < 1 < b < \infty$ such that $V(t) > V(1)$ outside $[a, b]$. Continuity and the extreme value theorem give a minimizer on $[a, b]$. Since 1 belongs to this interval, that minimizer is also global on $(0, \infty)$. Every global minimizer lies in this compact interval, so the minimum is attained at a finite, positive precision.

The continuity assumption is mild for the intended decoder class. For example, suppose $\mathbb{E}\|X\|^2 < \infty$ and g_θ is continuous with at most linear growth, $\|g_\theta(z)\| \leq A + B\|z\|$ for finite constants A, B. On any compact positive interval of t, the squared residual in Eq. (33) is dominated by an integrable multiple of $1 + \|X\|^2 + \|U\|^2 + \|\varepsilon\|^2$. Dominated convergence then establishes finite, continuous reconstruction. Fixed finite ReLU networks satisfy this growth condition, as do bounded continuous decoders.

Why the information at an objective minimum is strictly below its ceiling. Let $K_U = \mathrm{Cov}(U)$. Gaussian maximum entropy and finite second moment give

$$0 \leq C(t) \leq \tfrac{1}{2}\log\det(I_m + tK_U) \leq \tfrac{t}{2}\operatorname{tr} K_U, \tag{36}$$

so $C(t) \to 0$ as $t \downarrow 0$. Proposition 1 gives $C'(t) = M_{\mathrm{joint}}(t)/2 > 0$ at every finite $t > 0$. Hence C is continuous and strictly increasing, and the limit $C_{\max} = \lim_{t\to\infty} C(t) \in (0, \infty]$ exists. For any finite minimizing $t^\star$, choose a finite $t_2 > t^\star$. Then

$$0 < C(t^\star) < C(t_2) \leq C_{\max}. \tag{37}$$

This proves strict interiority without assuming $C_{\max} = \infty$. No finite precision attains the information ceiling. Equivalently, the inverse map $t(C)$ defines $V(t(C))$ on $(0, C_{\max})$, and its global minima correspond exactly to those in the precision coordinate.

For a nonconstant signal with finite support, $C_{\max} = H(U)$, the discrete entropy of its distinct values. As noise vanishes, nearest-support decoding has error probability tending to zero, and the finite-alphabet conditional entropy tends to zero. If all N equally weighted reference means are distinct, this gives $C_{\max} = \log N$. Repeated means instead reduce the ceiling to $H(U) \leq \log N$. In either case, Gaussian overlap prevents attainment at finite t. A largest *sampled* information value is different: the proposition does not ensure that an arbitrarily truncated grid contains a global minimizer.

A quantitative bound. Reference precision $t = 1$ matches the conditional noise variance to the prior variance. At every minimizer, $V(t^\star) \leq V(1)$. Combining this with Eq. (34) cancels $\mathbb{E}\|U\|^2/2$ and yields

$$\psi(t^\star) \leq a_0, \qquad a_0 \equiv \frac{2[D_{\mathrm{NLL}}(1) - b_x]}{m} = \frac{\mathbb{E}\|X - g_\theta(U + \varepsilon)\|^2}{m\sigma_x^2}. \tag{38}$$

The derivative $\psi'(t) = (t-1)/t^2$ is negative below one and positive above one. For $a_0 > 0$, the two solutions of $\psi(t) = a_0$ therefore give finite endpoints $0 < t_- < 1 < t_+ < \infty$, and every minimizer lies in $[t_-, t_+]$. If $a_0 = 0$, that interval reduces to $\{1\}$. Monotonicity also confines the minimizing information levels to $[C(t_-), C(t_+)] \subset (0, C_{\max})$. The available reduction in reference reconstruction loss thus bounds the additional noise-mismatch cost worth paying.

Actual reconstruction response and decoder geometry. For $D(t) = D_{\mathrm{NLL}}(t) \in C^1(0,\infty)$, differentiating the complete KL in Eq. (34) gives Eq. (15). To interpret its reconstruction term, write $v = t^{-1}$ and $\widetilde{D}(v) = D(1/v)$. Then

$$\mathcal{H}(t) = -2t^2 D'(t) = 2\widetilde{D}'(1/t). \tag{39}$$

Thus $\mathcal{H}$ is twice the signed NLL response to noise variance for the fixed decoder.

For a C^2 decoder with bounded first and second derivatives and $\mathbb{E}\|X\|^2 < \infty$, Gaussian covariance differentiation (Rezende et al., 2014) also gives

$$\mathcal{H}(t) = \frac{1}{\sigma_x^2}\mathbb{E}\left[\|J_{g_\theta}(Z_t)\|_F^2 - (X - g_\theta(Z_t))^\top \Delta g_\theta(Z_t)\right]. \tag{40}$$

Here J_{g_θ} is the decoder Jacobian, $\|\cdot\|_F$ denotes the Frobenius norm, and Δg_θ is the vector of Laplacians of the output coordinates. Indeed, for $\ell_x(z) = \|x - g_\theta(z)\|^2/(2\sigma_x^2)$, direct differentiation yields $\Delta\ell_x = (\|J_{g_\theta}\|_F^2 - (x - g_\theta)^\top \Delta g_\theta)/\sigma_x^2$. The Gaussian identity $d\mathbb{E}[\ell_X(U + \sqrt{v}\,\varepsilon)]/dv = \mathbb{E}[\Delta\ell_X]/2$ proves Eq. (40). The stated bounds and finite second moments justify differentiation on compact positive variance intervals. The first term quantifies noise amplification. The second captures the interaction between reconstruction residuals and decoder curvature. For nonsmooth decoders, Eq. (39) remains the definition: pointwise zero second derivatives inside ReLU regions do not remove contributions from crossing their boundaries.

These quantities can provide response bounds. For example, on any interval I of precisions, suppose $j_- \le \mathbb{E}\|J_{g_\theta}(Z_t)\|_F^2 \le j_+$, $\mathbb{E}\|X - g_\theta(Z_t)\|^2 \le r_I^2$, and $\mathbb{E}\|\Delta g_\theta(Z_t)\|^2 \le k_I^2$. Cauchy–Schwarz gives

$$\frac{j_- - r_I k_I}{\sigma_x^2} \le \mathcal{H}(t) \le \frac{j_+ + r_I k_I}{\sigma_x^2}, \qquad t \in I. \tag{41}$$

Bounds established only on I yield slope signs only there. If I contains every global minimizer, as does the interval from Eq. (38), stationary-point bounds can tighten that interval by intersection. Global branch statements require the bounds on the full positive axis.

Proof of localization and conditional uniqueness. At every minimizer, differentiability and interior attainment imply $V'(t^\star) = 0$. Equation (15) therefore gives

$$t^\star = 1 + \frac{\mathcal{H}(t^\star)}{m}. \tag{42}$$

The response bounds in Proposition 2 place every minimizer in the stated interval. Below its lower endpoint, $m(t-1) - \mathcal{H}(t) \le m(t-1) - h_- < 0$; above its upper endpoint, $m(t-1) - \mathcal{H}(t) \ge m(t-1) - h_+ > 0$. Since $2t^2 > 0$, these are precisely the claimed derivative signs. The condition $h_- > -m$ makes the lower endpoint strictly positive. If it is unavailable, an upper response bound still gives ascent above $1 + h_+/m$ whenever that threshold is positive.

For uniqueness, response bounds are unnecessary. Define $F(t) = m(t-1) - \mathcal{H}(t)$. If $\mathcal{H}'(t) < m$ everywhere, $F'(t) > 0$, so F is strictly increasing. Interior attainment supplies a zero at a global minimizer. Strict monotonicity makes this zero unique. Hence F, and therefore V', is negative before it and positive afterward. Because $C'(t) > 0$, the signs transfer to the information coordinate, including when $C_{\max}$ is finite. This change of coordinate preserves minima and monotonicity, but need not preserve convexity.

When $D \in C^2(0,\infty)$, the uniqueness condition has an equivalent curvature interpretation:

$$\frac{d^2}{dv^2}V(1/v) = \widetilde{D}''(v) + \frac{m}{2v^2} > 0 \quad \Longleftrightarrow \quad \mathcal{H}'(t) < m, \qquad v = 1/t. \tag{43}$$

The Gaussian KL curvature can dominate a nonconvex reconstruction response. Convexity of $\widetilde{D}$ is sufficient for Eq. (43). The stated inequality gives a more general condition for the same shape guarantee.

An explicit upper bound and guaranteed eventual ascent. A complementary result avoids global response envelopes and second derivatives. Suppose g_θ is globally L-Lipschitz and $\mathbb{E}\|X\|^2 < \infty$, in addition to the fixed-family assumptions. Set

$$\begin{aligned} r_0 &= \big(\mathbb{E}\|X - g_\theta(U)\|^2\big)^{1/2}, \quad a_L = \frac{L r_0}{\sigma_x^2\sqrt{m}}, \\ b_L &= 1 + \frac{L^2}{\sigma_x^2}, \qquad\qquad T_L = \left(\frac{a_L + \sqrt{a_L^2 + 4b_L}}{2}\right)^2. \end{aligned} \tag{44}$$

Then $V'(t) > 0$ for every $t > T_L$. All stationary points and all global minimizers consequently satisfy $t \le T_L$, and their information levels are at most $C(T_L) < C_{\max}$. This certifies an increasing right tail even when a global constant upper bound on $\mathcal{H}$ is unavailable.

To prove the claim, Lipschitz continuity gives an almost-everywhere Jacobian with operator norm at most L. Reparameterization and Cauchy–Schwarz imply

$$|D'(t)| \le \frac{L\sqrt{m}}{2\sigma_x^2 t^{3/2}} \left(r_0 + L\sqrt{m/t}\right). \tag{45}$$

Specifically, differentiate the squared residual at $Z_t = U + t^{-1/2}\varepsilon$ to obtain $D'(t) = -(2\sigma_x^2 t^{3/2})^{-1}\mathbb{E}[(g_\theta(Z_t) - X)^\top J_{g_\theta}(Z_t)\varepsilon]$. Its absolute value is bounded using $\mathbb{E}\|J_{g_\theta}(Z_t)\varepsilon\|^2 \le L^2 m$ and the L^2 triangle inequality $(\mathbb{E}\|g_\theta(Z_t) - X\|^2)^{1/2} \le r_0 + L\sqrt{m/t}$. Gaussian noise assigns probability zero to the decoder's nondifferentiability set, and the same bounds justify differentiation under the expectation on compact positive intervals. Combining Eq. (45) with the exact KL derivative yields

$$2t^2 V'(t) \ge m\big(t - a_L\sqrt{t} - b_L\big). \tag{46}$$

The quadratic in $\sqrt{t}$ is positive precisely above its positive root, which gives T_L. Finite feedforward ReLU networks with finite weights are globally Lipschitz. A product of layer operator norms supplies a valid, potentially conservative L.

Affine illustration. For $g_\theta(z) = Az + b$, independent zero-mean noise gives $D(t) = D_0 + \|A\|_F^2/(2\sigma_x^2 t)$, where $D_0 = b_x + \mathbb{E}\|X - AU - b\|^2/(2\sigma_x^2)$. Thus $\mathcal{H}(t) = \|A\|_F^2/\sigma_x^2$ is constant and

$$t^\star = 1 + \frac{\|A\|_F^2}{m\sigma_x^2}, \qquad V(t) - V(t^\star) = \frac{m}{2}\left(\frac{t^\star}{t} - 1 + \log\frac{t}{t^\star}\right). \tag{47}$$

The optimum is unique for any signal law satisfying the proposition. This exact special case illustrates the response balance and is consistent with earlier analyses linking variational covariance to decoder geometry (Kumar and Poole, 2020; Nakagawa et al., 2021).

Stationarity, shape, and attribution. Write $C^\star = C(t^\star)$. Where D_{NLL} is differentiable, each interior minimizer satisfies

$$\begin{aligned} -\left.\frac{dD_{\mathrm{NLL}}}{dC}\right|_{C^\star} &= 1 + \left.\frac{d\mathrm{TC}_{\mathrm{agg}}}{dC}\right|_{C^\star} + \left.\frac{d\mathrm{DWKL}}{dC}\right|_{C^\star} \\ &= \left.\frac{M_{\mathrm{sep}}}{M_{\mathrm{joint}}}\right|_{t^\star} + \left.\frac{d\mathrm{DWKL}}{dC}\right|_{C^\star}. \end{aligned} \tag{48}$$

This is the stationarity condition for Eq. (16). The basic assumptions ensure attainment, and $\mathcal{H}'(t) < m$ ensures a unique minimum with strict descent and ascent. These results cover arbitrary signal dependence. TC-driven reversal on an evaluated interval is established separately by Eq. (17), which compares TC growth with the combined change in the remaining components.

In particular, for a signal with finite support, $C(t)$ and $\mathrm{TC}_{\mathrm{agg}}(t)$ remain bounded, whereas $R_{\mathrm{KL}}(t) \to \infty$ as $t \to \infty$. Consequently $\mathrm{DWKL}(t) \to \infty$ at that endpoint. Small measured DWKL changes on the evaluated interval are compatible with this limit, but cannot be extrapolated to the entire path. Appendix E.1 describes the corresponding finite-support limits of information and TC.

C Geometric Partitions and Gaussian Dependence

The main argument uses the dependence of the code's actual coordinates. This appendix explains how that dependence changes under regrouping, relates Gaussian dependence to local geometric coordinates, and gives explicit noise-growth examples. These results supply interpretations of Proposition 1, while its direct empirical test is in Section 3.3.

C.1 Partition refinement and the prior

A geometric factor can require a vector representation (Higgins et al., 2018; Pfau et al., 2020; Fumero et al., 2021). Coordinates within that vector can be dependent even when distinct factors are independent. Suppose π refines a coarse coordinate partition ρ. For each $B \in \rho$, let $\pi|_B = \{A \in \pi : A \subseteq B\}$. With finite relevant entropies,

$$\begin{aligned} \mathrm{TC}_\pi(Z) - \mathrm{TC}_\rho(Z) &= \sum_{B\in\rho} \left[\sum_{A\in\pi|_B} h(Z_A) - h(Z_B) \right] \\ &= \sum_{B\in\rho} \mathrm{TC}_{\pi|_B}(Z_B). \end{aligned} \tag{49}$$

Consequently,

$$\mathrm{TC}_\pi(Z) = \mathrm{TC}_\rho(Z) + \sum_{B\in\rho} \mathrm{TC}_{\pi|_B}(Z_B). \tag{50}$$

This compares the same probability law under two partitions. It does not by itself imply a derivative with respect to information.

For a factorized prior, define $p(z_B) = \prod_{j\in B} p(z_j)$. The same decomposition gives

$$D_{\mathrm{KL}}(q(z)\|p(z)) = \mathrm{TC}_\rho(Z) + \sum_{B\in\rho} D_{\mathrm{KL}}(q(z_B)\|p(z_B)). \tag{51}$$

Within-block dependence remains inside the block KL. Merely regrouping coordinates leaves the total prior mismatch and ELBO unchanged. A prior that models within-block dependence, or a separately weighted hierarchical objective, changes the optimization problem (Esmaeili et al., 2019).

C.2 What geometric factorization does and does not imply

For a Riemannian product $\mathcal{Y} = \mathcal{Y}_1 \times \cdots \times \mathcal{Y}_K$ equipped with its product metric, the tangent spaces to the factors form orthogonal subspaces preserved by parallel transport. The de Rham decomposition relates this product structure to invariant subspaces of the holonomy group. The global converse requires a simply connected, geodesically complete manifold (Pfau et al., 2020). Here holonomy means the linear action obtained by transporting tangent vectors around closed loops.

If the holonomy action on a factor's tangent space is irreducible and its dimension exceeds one, that space has no proper nonzero subspace invariant under every such loop. A collection of fixed scalar axes therefore cannot represent the same transport-invariant decomposition. This supplies a conditional geometric reason to preserve an entire block. It does not imply that any particular covariance is anisotropic, that every transport mixes the chosen axes, or that transport angles grow with encoder information.

Geometric and statistical factorization are also distinct. A product manifold need not carry a product probability measure, and semantic factors can be statistically correlated in observed data (Träuble et al., 2021). Symmetry-based definitions allow multi-dimensional factor representations (Higgins et al., 2018). Unsupervised observations alone do not identify a unique semantic factorization without suitable inductive biases (Locatello et al., 2019). The appropriate factorization therefore depends on the data distribution and the modeling assumptions.

C.3 A Gaussian coding interpretation of geometric coordinates

Comparing local geometric directions. For each positive mode in Section 2.2, the standardized eigenvector $v_j(y)$ maps to the unit tangent direction $e_j(y) = \Sigma_{\mathrm{loc}}(y)^{1/2} v_j(y)/\sqrt{\lambda_j(y)}$, as derived

in Appendix A. Fixing transport paths, parallel transport under G compares these tangent directions in a common reference space while preserving lengths and angles. On a Riemannian product, it preserves each factor's tangent subspace, although axes may rotate within a factor (Pfau et al., 2020).

In a tangent block of dimension s, let $\Lambda = \mathrm{diag}(\nu_1, \ldots, \nu_s) \succ 0$ be the covariance in its local eigenframe. Let Q be the orthogonal matrix expressing that frame relative to a transported orthonormal reference, and set $\Gamma = Q\Lambda Q^\top$. The Gaussian law in the reference frame is $W \sim \mathcal{N}(0, \Gamma)$. Restrict π to this block, with $\Gamma[B, B]$ its principal submatrix for $B \in \pi$. Define its block TC by

$$b_\pi(\Gamma) \equiv \mathrm{TC}_\pi(W) = \frac{1}{2} \log \frac{\prod_{B\in\pi} \det \Gamma[B, B]}{\det \Gamma}. \tag{52}$$

Theorem 1 (Gaussian block coding cost) *At a common per-coordinate squared-error target below every eigenvalue of Γ, $b_\pi(\Gamma)$ equals the excess Gaussian rate for separately coding the prescribed blocks over coding them jointly. It is non-negative, independent of that target, and zero exactly when Γ is block diagonal under π.*

Here separate coding means independently coding each marginal block B under total squared-error distortion $|B|\delta$, without using cross-block information, where δ is the common per-coordinate target. The determinant ratio measures the resulting excess rate for the specified Gaussian law. Appendix C.4 distinguishes this local Gaussian quantity from the dependence of an aggregate mixture.

Proof of Theorem 1. Work in an s-dimensional tangent block with orthonormal coordinates for the metric G. Let $\Lambda = \mathrm{diag}(\nu_1, \ldots, \nu_s) \succ 0$ be its covariance and let the orthogonal Q express its eigenframe relative to the transported reference. The represented Gaussian law has covariance $\Gamma = Q\Lambda Q^\top$. Fix a partition π of these s coordinates and a per-coordinate squared-error target $\delta > 0$ smaller than every eigenvalue of Γ. This local block dimension need not equal the learned encoder's ambient dimension m.

Let $R_{\mathrm{joint}}(\delta)$ be the Gaussian rate–distortion value for jointly coding these coordinates with total distortion $s\delta$. Let $R_{\mathrm{sep},\pi}(\delta)$ be the sum of marginal block rates, allocating total distortion $|B|\delta$ to each block B. Every joint and block eigenmode is active at this common water level. The Gaussian rate–distortion formula gives (Berger, 1971; Cover and Thomas, 2006)

$$R_{\mathrm{joint}}(\delta) = \frac{1}{2} \log \frac{\det \Gamma}{\delta^s}, \qquad R_{\mathrm{sep},\pi}(\delta) = \frac{1}{2} \sum_{B\in\pi} \log \frac{\det \Gamma[B, B]}{\delta^{|B|}}. \tag{53}$$

Because $\sum_B |B| = s$, subtracting cancels every power of δ:

$$R_{\mathrm{sep},\pi}(\delta) - R_{\mathrm{joint}}(\delta) = \frac{1}{2} \log \frac{\prod_{B\in\pi} \det \Gamma[B, B]}{\det \Gamma} = b_\pi(\Gamma). \tag{54}$$

The Gaussian entropy formula gives the same expression for $\mathrm{TC}_\pi(W)$, where $W \sim \mathcal{N}(0, \Gamma)$. Fischer's determinant inequality makes the expression non-negative, with equality exactly when Γ is block diagonal under π. This proves the theorem.

For two singleton blocks, take Q to be a rotation by θ and $\Lambda = \mathrm{diag}(\nu_1, \nu_2)$. Multiplication gives

$$b_\pi(\Gamma) = \frac{1}{2} \log \left[1 + \frac{(\nu_1 - \nu_2)^2}{\nu_1 \nu_2} \sin^2 \theta \cos^2 \theta \right]. \tag{55}$$

Isotropy or alignment makes this zero. A fixed Γ gives a fixed rate excess even when the distortion target decreases. Dependence growth therefore requires a changing represented law, as specified next.

C.4 A FIXED FRAME UNDER DECREASING ENCODER NOISE

In a Gaussian channel of dimension m, let the fixed signal have the centered Gaussian law $U \sim \mathcal{N}(0, K)$, where $K = Q\Lambda Q^\top$, allowing $\Lambda \succeq 0$. Under Eq. (12), the aggregate covariance is $K + t^{-1} I_m$. The rescaled code $\sqrt{t}\, Z_t$ has covariance $I_m + tK \succ 0$ even if the signal is rank deficient. Its eigenframe remains fixed while its eigenvalues relative to the unit noise change with t. The

rescaling preserves information and TC, so Gaussian entropy gives

$$\begin{aligned} C(t) &= \frac{1}{2}\log\det(I_m + tK), \\ \mathrm{TC}_\pi(Z_t) &= \frac{1}{2}\left[\sum_{B\in\pi}\log\det(I_{|B|} + tK[B,B]) - \log\det(I_m + tK)\right]. \end{aligned} \tag{56}$$

In two dimensions, let Q be a rotation through angle θ and $\Lambda = \mathrm{diag}(\nu_1, \nu_2)$ with $\nu_1, \nu_2 \geq 0$. For singleton blocks, substituting the eigenvalues $1 + t\nu_1$ and $1 + t\nu_2$ into Eq. (55) gives

$$\mathrm{TC}(Z_t) = \frac{1}{2}\log\left[1 + \frac{t^2(\nu_1 - \nu_2)^2}{(1 + t\nu_1)(1 + t\nu_2)}\sin^2\theta\cos^2\theta\right]. \tag{57}$$

For fixed nonnegative ν_1, ν_2, differentiation of its variable factor gives

$$\frac{d}{dt}\frac{t^2}{(1 + t\nu_1)(1 + t\nu_2)} = \frac{2t + t^2(\nu_1 + \nu_2)}{(1 + t\nu_1)^2(1 + t\nu_2)^2} > 0. \tag{58}$$

Thus TC grows strictly when $\nu_1 \neq \nu_2$ and $\sin\theta\cos\theta \neq 0$. For a nonzero signal, $C'(t) > 0$, so the sign is preserved when parameterizing by C. If both signal eigenvalues are positive, Eq. (57) converges to Eq. (55). The high-rate dependence need not diverge.

One factor in two coordinates. For $U = (Y_0, Y_0)$ with $Y_0 \sim \mathcal{N}(0, 1)$, the signal covariance has eigenvalues 2 and 0 and fixed angle $\pi/4$. Equation (56) yields

$$C(t) = \frac{1}{2}\log(1 + 2t), \qquad \mathrm{TC}(Z_t) = \log(1 + t) - \frac{1}{2}\log(1 + 2t). \tag{59}$$

Solving for t gives $t = (e^{2C} - 1)/2$, hence

$$\mathrm{TC}(C) = \log\frac{e^C + e^{-C}}{2} = \log\cosh C, \qquad \frac{d\mathrm{TC}}{dC} = \tanh C, \tag{60}$$

and $\mathrm{TC}(C) = C - \log 2 + o(1)$ as $C \to \infty$. As noise decreases, the code concentrates near the fixed one-dimensional signal subspace. The example demonstrates information shared across coordinates. It does not assert that duplicating a scalar factor is necessary. A rotation can place the signal in one coordinate, as the counterexample in Appendix B.3 illustrates.

Two independent factor groups in Figure 5. Let A_0 and B_0 be independent standard Gaussian scalar factors. Set $U = (A_0, A_0, A_0, B_0, B_0, B_0)$ in Eq. (12). The coarse partition is $\rho = \{\{1, 2, 3\}, \{4, 5, 6\}\}$ and π is the singleton partition. The two factor groups remain independent for every t, so $\mathrm{TC}_\rho = 0$. Each 3×3 signal block has one nonzero eigenvalue, equal to three. The Gaussian entropy formula gives

$$C(t) = \log(1 + 3t), \qquad \mathrm{TC}_\pi(t) = 3\log(1 + t) - \log(1 + 3t). \tag{61}$$

Within each coarse group, the off-diagonal correlation is $t/(1 + t)$. All between-group correlations are zero. Figure 5 uses $t = 0.3$ and $t = 20$, equivalent to independent encoder-noise variances $1/t$ around the fixed signal. The heatmaps share the scale $[0, 1]$. Correlations remove changes in marginal scale, making the increasing within-group dependence visible. These matrices completely characterize dependence for this Gaussian example, but not for arbitrary non-Gaussian aggregates.

Local Gaussian laws and the aggregate posterior. The log-determinant identities above concern the stated Gaussian distribution. A mixture over neighborhood centers generally is not Gaussian, and the average of local Gaussian TCs need not equal the TC of that mixture. Proposition 1 avoids this issue by acting on the full aggregate law induced by $U = f(X)$. Its proof does not assume that law is Gaussian. In the trained nonlinear benchmark, aggregate density evaluation supplies the objective components directly.

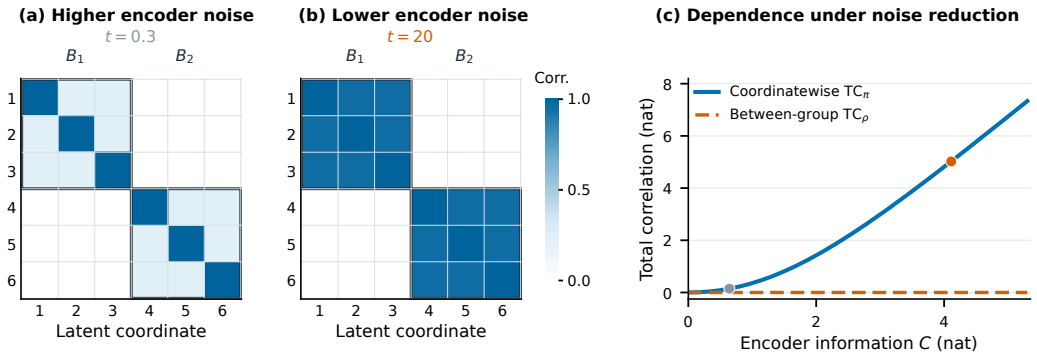


Figure 5: **Dependence can grow within independent factor groups.** An analytic Gaussian example with two independent factors, each represented in three coordinates. (a,b) Aggregate correlation matrices at two encoder-noise levels share one color scale. Thick boxes mark the coarse groups B_1, B_2 of ρ, and grid cells mark the singleton partition π. (c) Coordinatewise TC_π increases with information, while between-group TC_ρ stays zero. Markers match the two matrices. These are analytic distributions, not fitted experimental data.

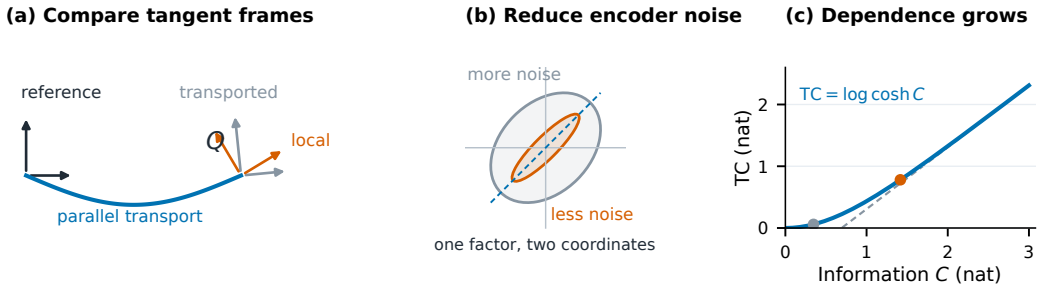


Figure 6: **Geometric interpretation of a shared scalar factor.** (a) Parallel transport provides a common tangent frame. (b) Two coordinates represent one Gaussian factor. Reducing independent Gaussian noise concentrates the code around its fixed signal direction. The ellipses are equal-Mahalanobis-distance contours. (c) The resulting $\mathrm{TC}(C) = \log\cosh C$ approaches $C - \log 2$ at high information rate.

C.5 High-rate slopes from block projection ranks

Let $Y_0 \sim \mathcal{N}(0, I_r)$ be an r-dimensional continuous Gaussian source, and let $A \in \mathbb{R}^{m\times r}$ be fixed with full column rank $r \geq 1$. Consider $U = AY_0$ and independent Gaussian noise of covariance $t^{-1}I_m$ as in Eq. (12). For a block B of a fixed partition π, let A_B denote the rows of A indexed by B, and write $r_B = \mathrm{rank}(A_B)$. The value r_B counts the signal dimensions visible in that block, while r counts the dimensions visible jointly.

Corollary 1 (Gaussian high-rate block dependence) *As $t \to \infty$,*

$$C(t) = \frac{r}{2}\log t + O(1), \qquad \mathrm{TC}_\pi(Z_t) = \frac{\sum_{B\in\pi} r_B - r}{2}\log t + O(1). \tag{62}$$

Equivalently, as $C \to \infty$ along this family,

$$\mathrm{TC}_\pi(C) = \left(\frac{\sum_{B\in\pi} r_B}{r} - 1\right) C + O(1), \qquad \frac{d\mathrm{TC}_\pi}{dC} \longrightarrow \frac{\sum_{B\in\pi} r_B}{r} - 1. \tag{63}$$

Proof. The signal covariance is $K = AA^\top$ and has rank r. Each block covariance $K[B,B] = A_B A_B^\top$ has rank r_B. For a positive semidefinite matrix L with positive eigenvalues $\ell_1, \ldots, \ell_s$, its log

determinant satisfies

$$\log\det(I+tL)=\sum_{i=1}^{s}\log(1+t\ell_i)=s\log t+\sum_{i=1}^{s}\log\ell_i+o(1). \tag{64}$$

The sum is zero when $s=0$. Applying this expansion to Eq. (56) proves both asymptotic formulas. Eliminating $\log t$ gives the first part of Eq. (63). Differentiating the finite eigenvalue sums shows that the derivatives with respect to t have leading terms $r/(2t)$ and $(\sum_B r_B - r)/(2t)$, proving the derivative limit.

Interpretation and limits. Since stacking the block matrices recovers A, $\sum_B r_B \geq r$. The coefficient is positive exactly when counting each block's signal dimensions separately overcounts the joint dimension. Refining a block cannot decrease this sum of ranks. If every coordinate has nonzero signal variance and π is the singleton partition, then $\sum_B r_B = m$ and the coefficient is $(m-r)/r$.

A multi-dimensional semantic factor does not by itself imply this positive coefficient. For example, when A has full row rank and π is a coordinate partition, the high-rate coefficient is zero even if the limiting Gaussian TC is positive. The formula refers to signal projection ranks of the specified continuous model, not a count of semantic labels. It supplies a sufficient asymptotic mechanism rather than a quantitative fit to the controlled full-covariance sweeps.

D Experimental Protocol and Evidence

Synthetic data and the role of generative-factor coordinates. The component benchmark uses a fixed six-factor generator. Its independent scalar generative factors satisfy $Y_j \sim \mathcal{N}(0,\nu_j)$, with $(\nu_1,\ldots,\nu_6)=(4.0,2.4,1.4,0.85,0.50,0.30)$. Let F_0 be the unscaled noiseless observation map. The two interaction coefficients are $\kappa_{12}=1.0$ and $\kappa_{34}=0.7$, and

$$F_0(Y)=\left(Y_1,\ldots,Y_6,\frac{\kappa_{12}Y_1Y_2}{\sqrt{\nu_1\nu_2}},\frac{\kappa_{34}Y_3Y_4}{\sqrt{\nu_3\nu_4}}\right)^{\top}. \tag{65}$$

The first six coordinates preserve the factor values, and the last two introduce nonlinear interactions within the designated pairs. With independent observation noise $\eta\sim\mathcal{N}(0,0.10I_8)$, the raw observation is $X_{\mathrm{raw}}=F_0(Y)+\eta$. The training protocol forms $X=X_{\mathrm{raw}}/a$ using one scalar $a>0$. Its square is the average of the eight coordinate variances in the training data. The Gaussian decoder variance is fixed at $\sigma_x^2=0.10/a^2$. Every rate and partition condition uses the same scaling. The decoder mean is fixed to $F_0(\mathrm{diag}(\sqrt{\nu_1},\ldots,\sqrt{\nu_6})z)/a$. Encoder training uses observations alone; factor labels and derivatives of F_0/a are used for evaluation. The data protocol specifies 60,000 training, 10,000 validation, and 20,000 test examples, with 1,024 fixed test examples used for the component audit.

The separate allocation-ranking study evaluates retention of five scalar factor targets across six local regions and six rate checkpoints per seed. Its generator, six-coordinate latent code, allocation predictors, and readouts are defined in Appendix F.

Availability of experiment records. The accompanying Supplementary archive contains source code, configurations, controlled and intervention results, row-level allocation readouts, and external fit and bootstrap records. Its README maps these materials to the paper and describes direct checks of the archived results. Neural checkpoints and raw image datasets are not included; some original entry points require paths to be configured before execution.

Experimental units. The independent run identifiers are 5201–5208. Within each run, four prescribed conditions each contribute ten separately optimized encoders paired with the same fixed decoder. Each trajectory therefore comprises ten separately trained models. The full sweep contains $8\times4\times10=320$ models. Selecting the tenth model from each trajectory gives $8\times4=32$ distinct intervention anchors. A condition is retained with its parent run in every cluster bootstrap. These counts distinguish the 32 trajectories and anchors from the eight independent resampling units. The encoder minimizes $\widetilde{D}_{\mathrm{NLL}}+R_{\mathrm{KL}}+100|R_{\mathrm{KL}}-r_{\mathrm{target}}|$, where $\widetilde{D}_{\mathrm{NLL}}=\mathbb{E}\|X-g(Z)\|^2/(2\sigma_x^2)$. The

Table 1: Experimental settings and the quantities each study tests.

Study	Model and comparison	Evaluation
Gaussian allocation	Numerical rate allocation over six Gaussian modes, three initializations	Analytic distortion and thresholded full activation
Nonlinear allocation ranking	Six-coordinate latent code, five scalar factor targets, eight seeds	Three allocation predictors compared with two factor-retention readouts over 180 units per seed
Nonlinear component audit	Six-factor observations, six-coordinate Gaussian code with Cholesky covariance parameterization, eight runs, four partition conditions, ten rate settings	Same 1,024 test examples and four posterior draws, full ELBO decomposition, cross-fitted reference, run-clustered inference
Gaussian noise intervention	32 learned mean maps with fixed decoders, isotropic Gaussian noise varied over ten levels	Separate density and MMSE calculations, decoder response, and intervention minima on the same 1,024 inputs
External image sweeps	Ten-coordinate convolutional mean-field VAEs, global KL-weighted training, three seeds and ten weights per dataset	Standard negative ELBO against achieved R_{KL} at epoch 300, paired hierarchical bootstrap

ten targets are $4.4, 4.6, 4.8, 5.0, 5.2, 5.5, 6.2, 8.0, 9.0, 10.0$ nats. Checkpoints minimize validation negative ELBO without the capacity penalty. Controlled Gaussian losses omit the common constant in Eq. (23); differences, derivatives, and minima are unchanged. Block masks on the encoder's Cholesky factor implement the four covariance conditions.

Cross-fitted high-rate attribution. The controlled study contains eight independent training runs, four pre-specified partitions, and ten checkpoints per trajectory. The four partition conditions keep both designated pairs $(1, 2)$ and $(3, 4)$ together, split both, or split only one of the two pairs. All 320 checkpoints are evaluated on the same ordered set of 1,024 test examples, with four fixed posterior-noise draws shared across objective components and partition-density evaluations. For each held-out run and partition, the reference setting M is the one with the lowest negative ELBO averaged over the other seven runs, among the ten rate-control settings. Every selector chooses setting 8, and the endpoint H is fixed at setting 10. The evaluated run is therefore excluded from its own selection and contributes once to the paired $M \to H$ contrast. Confidence intervals use 10,000 bootstrap draws with the selected contrasts fixed, resampling training runs with their four partitions kept together.

Independent marginal-mismatch audit. For each fixed contrast, we calculate $s_A = \Delta A/\Delta C$ directly from the endpoint values for $A = D_{\mathrm{NLL}} + C$, $A = \mathrm{DWKL}$, and $A = \mathrm{TC}_{\mathrm{agg}}$. In particular, DWKL is not obtained as an objective remainder. We first average the four partitions within each run and then average the eight run means. A 10,000-draw percentile bootstrap of run means (seed 20260825) gives

$$\begin{aligned} s_{D_{\mathrm{NLL}}+C} &= -3.924 \quad [-4.172, -3.682], \\ s_{\mathrm{DWKL}} &= +0.0161 \quad [+0.0104, +0.0233], \\ s_{\mathrm{TC}} &= +5.219 \quad [+5.014, +5.417], \\ s_{\mathrm{nonTC}} &= -3.908 \quad [-4.156, -3.669], \\ m_{\mathrm{rev}} &= +1.311 \quad [+1.210, +1.403]. \end{aligned}$$

The pre-specified ratio of aggregate means is $|s_{\mathrm{DWKL}}|/|s_{\mathrm{TC}}| = 0.309\%\ [0.202\%, 0.441\%]$, with the ratio recomputed from the aggregate slopes in each bootstrap draw. Ratios are not averaged over the 32 trajectories. All eight run-mean s_{DWKL} values are positive, whereas the trajectory signs are 25 positive and 7 negative. To assess sensitivity to any single run, we keep the selected contrasts fixed, remove one run's four conditions at a time, and recompute summaries over the remaining seven runs. The eight mean reversal slopes range from 1.281 to 1.351 per nat, compared with 1.311 using all runs. This measures the stability of the average effect magnitude. The resulting DWKL-to-TC ratios range from 0.251% to 0.336%, while the maximum original run-level and trajectory-level ratios are

0.707% and 1.252%. These results support the interval-specific statement that marginal mismatch is small relative to aggregate TC in this controlled comparison.

Gaussian allocation check. We optimize six rates $c_j = C_{\mathrm{loc}}\,\mathrm{softmax}(v)_j$ to minimize the analytic distortion $\sum_j \lambda_j e^{-2c_j}$, using Adam for 1,200 steps at learning rate 0.05 and three initializations (seeds 7401–7403). The spectrum is $(3.4181625, 2.3243505, 1.2920654, 0.9160675, 0.3418162, 0.2050897)$; full precision is retained in the configuration. The sampled budgets are $0.5, 1, 2, 3.5, 4.4, 4.5, 4.6, 4.7, 4.8, 5.1, 5.5, 6.5, 7.5, 8, 8.5, 9.5$ nats. The maximum relative discrepancy from $D^\star(C_{\mathrm{loc}})$ is 2.96×10^{-4}. A mode is counted as numerically active when $c_j > 0.02$ nat. All six first meet this criterion at $C_{\mathrm{loc}} = 4.7$, a gap of 0.1554 nat from the analytic full-activation boundary 4.5446. This gap includes the activity threshold and grid resolution. Active-mode counts are nondecreasing in all three initializations. Figure 2(a,b) shows the analytic solution for the same spectrum and summarizes this numerical check.

Full-covariance dependence accounting. The posteriors in the component audit are parameterized by Cholesky factors, so we evaluate the general identity $C + \mathrm{TC}_{\mathrm{agg}} = C_{\mathrm{coord}} + \mathrm{TC}_{\mathrm{cond}}$. A secondary diagnostic fits a line to the three highest-information points of each path, corresponding to KL targets 8, 9, and 10. Using analytic Gaussian conditional TC, it gives mean aggregate-, conditional-, and net-TC slopes of 4.973, 0.0075, and 4.965, with run-clustered 95% intervals $[4.787, 5.160]$, $[-0.0022, 0.0190]$, and $[4.780, 5.157]$. The net point estimate therefore retains 99.85% of the aggregate slope. This three-point regression differs from the two-point cross-fitted $M \to H$ slope above. For conditional TC, the analytic Gaussian evaluation and the earlier posterior-draw density estimate have correlation 0.978 across checkpoints, with maximum absolute difference 0.0111 nat. Aggregate TC is evaluated from the aggregate density, rather than a Gaussian approximation to that density.

Numerical consistency. For the archived component values, the maximum residual in $V = D_{\mathrm{NLL}} + C + \mathrm{TC}_{\mathrm{agg}} + \mathrm{DWKL}$ is 1.27×10^{-6} nat across checkpoints. The corresponding paired-contrast and normalized-slope residuals are 1.70×10^{-6} nat and 5.43×10^{-6}, respectively. These checks assess numerical consistency of the decomposition.

External reconstruction likelihoods and scale. For dSprites, let p index the $64 \times 64 = 4{,}096$ pixels, let $x_p \in \{0, 1\}$ be the observed value, and let $\widehat{p}_p(z)$ be the decoder probability that pixel p equals one. The Bernoulli reconstruction term is

$$D_{\mathrm{NLL}}^{\mathrm{dSprites}} = \mathbb{E}_{X,Z}\left[-\sum_{p=1}^{4096}\{x_p \log \widehat{p}_p(Z) + (1 - x_p)\log(1 - \widehat{p}_p(Z))\}\right]. \tag{66}$$

For Cars3D, the factorized discretized-logistic decoder assigns a probability mass to each observed 8-bit channel value. $D_{\mathrm{NLL}}^{\mathrm{Cars3D}}$ is the expected negative log of those masses summed over $64 \times 64 \times 3 = 12{,}288$ channel values. Neither benchmark divides this quantity by the number of pixels or subtracts a baseline. A Cars3D value on the order of 10^4 nat/image therefore reflects the likelihood and per-image summation convention. Convergence must be assessed from training trajectories and component changes. The absolute loss magnitude cannot establish it. The different likelihoods also preclude a direct comparison with the dSprites loss. In both cases, the reported negative ELBO is exactly $D_{\mathrm{NLL}} + R_{\mathrm{KL}}$.

Dense dSprites rate sweep. The convolutional mean-field VAEs use a 10-dimensional latent space and the training loss $D_{\mathrm{NLL}} + \beta_R R_{\mathrm{KL}}$, without factor supervision. We varied β_R to sample operating points and ordered the analyses by the achieved quantity $R_{\mathrm{KL}} = \mathbb{E}_X D_{\mathrm{KL}}(q(z \mid X) \| p(z))$. The three pre-specified seeds are 6001, 6002, and 6003. Each seed used nine pre-specified weights, $\beta_R \in \{2.00, 1.75, 1.50, 1.375, 1.25, 1.00, 0.75, 0.50, 0.40\}$, and one additional weight, 0.625, selected using achieved KL rates without inspecting negative ELBO. The resulting sweep contains 30 operating points at epoch 300.

All checkpoints use the same ordered set of 32,768 image IDs and four posterior-noise draws. Because the KL is evaluated analytically, its value is identical across the four draws, which are not treated as independent observations. Objective closure satisfies $\max|-\mathrm{ELBO} - D_{\mathrm{NLL}} - R_{\mathrm{KL}}| = 7.629 \times 10^{-6}$

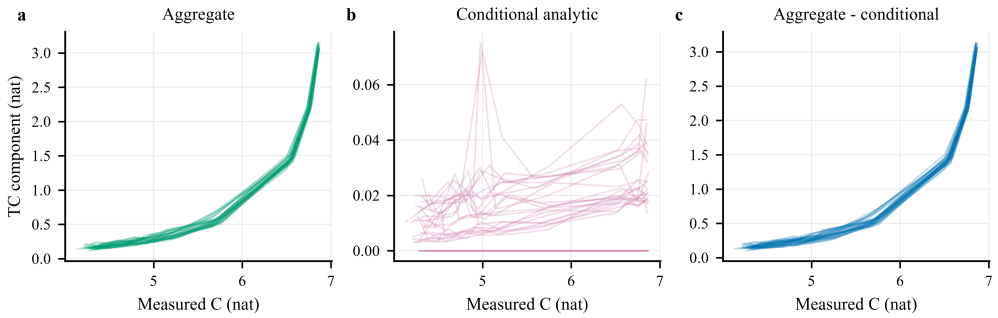


Figure 7: **Aggregate dependence dominates conditional dependence in the full-covariance sweep.** Aggregate TC grows sharply with measured C (left), conditional TC is much smaller (center, on a separate vertical scale), and aggregate minus conditional TC follows the aggregate trend (right). Thin curves are the 32 run–partition trajectories.

nat/image. Formal inference uses a continuous two-segment linear model with seed-specific intercepts and shared unrestricted slopes, weighted by the inverse paired-bootstrap variance in both the original fit and bootstrap refits. With 5,000 hierarchical-bootstrap draws, the pooled epoch-300 estimates are a breakpoint of 27.268 $[27.104, 29.934]$, a left slope of -0.340 $[-0.357, -0.169]$, and a right slope of 0.442 $[0.400, 0.596]$. The three run-specific estimates for (breakpoint, left slope, right slope) are $(29.279, -0.176, 0.592)$, $(29.955, -0.171, 0.595)$, and $(27.112, -0.326, 0.400)$. The lowest observed grid points occur at achieved rates 26.009, 25.799, and 26.283 nat/image and are reported separately from the fitted breakpoints.

All nine seed–epoch fits at epochs 225, 275, and 300 retain negative left and positive right slopes. Figure 4(c) applies Gaussian smoothing with bandwidth 0.75 nat/image and linear boundary extrapolation to the fitted curve and each bootstrap curve before taking pointwise quantiles. Breakpoint and slope inference uses the unsmoothed two-segment model. The earlier epochs are treated as temporal sensitivity analyses because the three sweeps retain measurable late-stage drift. With three top-level training seeds, population-level generalization of the bootstrap intervals remains limited.

Dense Cars3D rate sweep. Cars3D contains 17,568 rendered RGB images from 183 car identities, 24 azimuths, and four elevations (Reed et al., 2015). For each identity, 72/12/12 images enter the train/validation/test splits, yielding $13{,}176/2{,}196/2{,}196$ images. The 10-dimensional convolutional mean-field VAE is evaluated on the same ordered set of 2,196 test images with four fixed posterior-noise draws at every rate setting. The pre-specified seeds are 7401, 7402, and 7403. The ten training weights are

$$\begin{aligned}\beta_R \in \{&128, 51.399431, 12.995188, 4, 0.250376,\\ &0.147644, 0.088447, 0.056900, 0.042168, 0.031250\}.\end{aligned} \tag{67}$$

A repeated $\beta_R = 4$ checkpoint from an earlier stage was removed before analysis. The resulting 30 epoch-300 points span achieved $R_{\mathrm{KL}} = 10.769$–207.754 nat/image.

The Cars3D pooled curve in Fig. 4(d) uses the same weighted two-segment model and display smoothing as dSprites. Its breakpoint is 20.530 $[20.047, 20.530]$ nat/image, selected from 401 equally spaced candidates spanning $[10.867, 204.127]$ nat/image. The upper confidence limit matches the point estimate because bootstrap fits concentrate on that interior grid value. The pointwise band uses 5,000 bootstrap draws of seeds and paired image IDs. The numerical branch contrasts use three checkpoints fixed before this comparison: a low-rate model at $\beta_R = 128$, an intermediate-rate reference at $\beta_R = 12.995188$, and a high-rate model at $\beta_R = 0.056900$. Let $V_{s,L}, V_{s,M}, V_{s,H}$ be their mean negative ELBOs for seed $s \in \{1, 2, 3\}$. Define

$$\Delta_L = \frac{1}{3}\sum_{s=1}^{3}(V_{s,L} - V_{s,M}), \qquad \Delta_H = \frac{1}{3}\sum_{s=1}^{3}(V_{s,H} - V_{s,M}). \tag{68}$$

At epoch 300, $\Delta_L = 781.56$ nat/image with 95% CI $[739.29, 814.70]$, and $\Delta_H = 75.69$ nat/image with 95% CI $[38.53, 112.49]$. Both within-seed differences are positive in all three seeds. These

epoch-300 fixed-checkpoint intervals use paired image resampling within each sampled seed; the pooled-curve bootstrap shares one image resample across seeds. Both preserve pairing across rate settings and shared posterior draws. The reference is the fixed intermediate-rate checkpoint, not the minimum of a fitted smooth curve.

Only epoch-300 checkpoints enter the pooled panel. Longer epoch-500 diagnostics assess temporal stability and are kept separate from the common-horizon sweep. Because measurable rate drift remains, the Cars3D result describes this training horizon rather than a stationary optimum. The plotted values are divided by 10^3 only for tick formatting. All calculations use the original nat/image values.

Relation to the rate–distortion trade-off. On an ideal differentiable optimum frontier $D_{\mathrm{NLL}}(R_{\mathrm{KL}})$, an interior solution of Eq. (18) satisfies $D'_{\mathrm{NLL}}(R_{\mathrm{KL}}) = -\beta_R$. The derivative of the evaluated unweighted objective is then $V'(R_{\mathrm{KL}}) = 1 - \beta_R$. This classical rate–distortion trade-off permits a reversal of the unweighted objective across KL weights (Alemi et al., 2018). The external sweeps test whether the response recurs in image VAEs at a common training horizon. The controlled component audit quantifies the aggregate-TC contribution to the reversal.

E Gaussian Noise Interventions on the Controlled Models

E.1 Design, evaluation distribution, and comparison with trained models

The Gaussian noise intervention used the same eight runs (5201–5208) and four prescribed conditions as the original controlled sweep. For each run–condition pair, we selected the tenth model, trained at target KL 10 and having the highest measured information in its sweep, and held its encoder mean map, decoder, prior, and likelihood fixed. Write $u_i = f(x_i)$ for the frozen mean of the ith input. The evaluation distribution assigns probability $1/N$ to each of the same $N = 1{,}024$ ordered test inputs. The intervened encoder is

$$q_t(z \mid x_i) = \mathcal{N}(z; u_i, t^{-1} I_m), \qquad z_{i\ell}(t) = u_i + t^{-1/2}\varepsilon_{i\ell}, \tag{69}$$

where $m = 6$, $\ell \in \{1, \ldots, L\}$ indexes $L = 4$ fixed independent standard Gaussian draws per input, and the same draws are reused across noise levels and objective terms. The intervention replaces each selected model's encoder covariance with isotropic Gaussian noise, including when the trained encoder has full covariance. The four training-condition names identify the selected models. Every intervention uses the same isotropic noise law.

Each intervention used a fixed initial grid of seven noise precisions, recorded for every run in `noise_points_frozen_v4.csv` and shared across its four conditions. Denoting the largest initial precision by $t_{\max}$, we added three levels: $1.5t_{\max}$, $2t_{\max}$, and $3t_{\max}$. We retained all ten levels on all 32 paths, giving 320 evaluation points. Network calculations used 32-bit floating-point arithmetic (FP32), while mixture densities, estimation errors, and aggregate statistics used 64-bit floating-point arithmetic (FP64). No optimization was performed during the intervention. The independent derivative audit evaluated joint and marginal densities at the paired endpoints specified below.

The intervention fixes each selected model's mean map and replaces its covariance by $t^{-1} I_m$, whereas the original sweep separately optimizes the encoder at each target.

Under the finite evaluation distribution, $C(t) \le \log N = 6.9315$ nat. The actual ceiling is $H(U)$, equal to $\log N$ when all reference means are distinct. It is approached only as precision tends to infinity, whereas every global minimizer in Proposition 2 has finite precision (Appendix B.4). With $s = \log t$, small values of dC/ds near this bound make derivative ratios sensitive to Monte Carlo error. The continuous Gaussian high-rate limit in Appendix C.5 is not an asymptotic model for this finite mixture.

Finite reference support and population dependence. The empirical mixture corresponds to a signal law with finite support, rather than an estimate of a continuous population density. To see why the distinction matters at high precision, let J be the uniformly distributed reference index and write $U = u_J$. If every coordinate $u_{J,j}$ uniquely identifies J, then vanishing noise allows each coordinate,

as well as the joint code, to recover the same index. Finite-alphabet decoding and Eq. (11) give

$$C(t) \longrightarrow \log N, \qquad \mathrm{TC}_{\mathrm{agg}}(t) \longrightarrow (m-1)\log N \qquad (t \to \infty).$$

For example, independent continuous population coordinates have distinct sampled values almost surely. Their finite empirical joint law can therefore display this limiting dependence even though the population coordinates are independent. More generally, the finite-signal limit is $\sum_j H(U_j) - H(U)$, where H denotes discrete entropy. This does not alter the exact dependence-growth identity for the specified empirical law. It does mean that a sharp high-precision increase alone cannot identify dependence in the population representation. The present audit fixes $N = 1{,}024$ and does not measure sensitivity to reference-set size. Its conditional Monte Carlo intervals describe noise-draw uncertainty, not this distributional sensitivity.

E.2 Two separate calculations of the dependence-growth rate

Estimation-error calculation. The aggregate density is the full empirical mixture

$$q_t(z) = \frac{1}{N}\sum_{r=1}^{N} \mathcal{N}(z; u_r, t^{-1} I_m). \tag{70}$$

The index r runs over all reference means, including the queried input. This is the exact mixture for the specified empirical distribution, rather than a leave-one-out density estimate. Let $w_r(z,t)$ be the posterior probability of reference component r given the complete code. For coordinate j, let $w_{r,j}(z_j, t)$ be the corresponding probability using only that coordinate:

$$\begin{aligned} w_r(z,t) &= \frac{\exp(-t\|z-u_r\|^2/2)}{\sum_a \exp(-t\|z-u_a\|^2/2)}, \\ w_{r,j}(z_j,t) &= \frac{\exp[-t(z_j-u_{r,j})^2/2]}{\sum_a \exp[-t(z_j-u_{a,j})^2/2]}. \end{aligned} \tag{71}$$

The sums over a and r have N terms. The exact conditional means under this mixture are $\widehat{u}(z,t) = \sum_r w_r(z,t) u_r$ and $\widehat{u}_j(z_j,t) = \sum_r w_{r,j}(z_j,t) u_{r,j}$. At each $z_{i\ell}(t)$, we compute the conditional variance $\sum_r w_r \|u_r\|^2 - \|\widehat{u}\|^2$ and the sum of marginal conditional variances $\sum_j (\sum_r w_{r,j} u_{r,j}^2 - \widehat{u}_j^2)$. Averaging over i, ℓ estimates M_{joint} and M_{sep}, whose expectations equal the squared-error definitions in Eq. (28). Only the outer Gaussian-noise expectation is sampled. The conditional means do not require a learned regressor or a Gaussian approximation to the aggregate.

Recall $s = \log t$, and let a dot denote differentiation with respect to s. Set $T = \mathrm{TC}_{\mathrm{agg}}$ and $W = \mathrm{DWKL}$. By Eq. (30), the estimation-error predictions are

$$\dot{C}_{\mathrm{MMSE}} = \frac{t}{2} M_{\mathrm{joint}}, \qquad \dot{T}_{\mathrm{MMSE}} = \frac{t}{2}(M_{\mathrm{sep}} - M_{\mathrm{joint}}), \qquad \rho_{\mathrm{MMSE}} = \frac{\dot{T}_{\mathrm{MMSE}}}{\dot{C}_{\mathrm{MMSE}}}. \tag{72}$$

Sample estimates replace the expectations in the audit. These MMSEs measure recovery of the latent signal U, not reconstruction of X by the decoder.

Direct density calculation. Independently, at every paired endpoint we evaluate

$$\begin{aligned} \widehat{C}(s) &= \frac{1}{NL}\sum_{i,\ell} \log \frac{q_{e^s}(z_{i\ell}(e^s) \mid x_i)}{q_{e^s}(z_{i\ell}(e^s))}, \\ \widehat{T}(s) &= \frac{1}{NL}\sum_{i,\ell} \log \frac{q_{e^s}(z_{i\ell}(e^s))}{\prod_j q_{e^s}(z_{i\ell,j}(e^s))}. \end{aligned} \tag{73}$$

Here $q_t(z_j)$ is the univariate marginal mixture of Eq. (70). The calculation uses its joint and marginal log densities directly. In particular, it never constructs T as $R_{\mathrm{KL}} - C - \mathrm{DWKL}$.

We evaluated the first, seventh, and tenth precision levels (zero-based indices 0, 6, and 9) for every path, giving 96 paired checks. With $h = 0.001$, we used the same inputs and Gaussian draws at $s-h$ and $s+h$. For $A \in \{C, T\}$, define

$$\dot{A}_{\mathrm{FD}} = \frac{\widehat{A}(s+h) - \widehat{A}(s-h)}{2h}, \qquad \rho_{\mathrm{FD}} = \frac{\dot{T}_{\mathrm{FD}}}{\dot{C}_{\mathrm{FD}}}. \tag{74}$$

This uses 192 endpoint evaluations. The ratio of the two global derivatives is compared with ρ_{MMSE} in Fig. 3(a). Here, "independent calculations" refers to different estimator constructions with shared inputs and draws, not statistically independent estimates.

The median absolute ratio error is 0.0161. All 96 ratios are reported, including the larger discrepancies at the highest precision (Table 2). A nonzero denominator permits a numerical ratio but does not ensure stability when $\dot{C}$ is small. The maximum residual in $\widehat{C}+\widehat{T}+\widehat{W} = \widehat{R}_{\mathrm{MC}}$ is 3.55×10^{-15}, where $\widehat{W}$ and $\widehat{R}_{\mathrm{MC}}$ average $\log[\prod_j q_t(Z_{t,j})/p(Z_t)]$ and $\log[q_t(Z_t \mid X)/p(Z_t)]$, respectively, over the same draws. The plotted objective instead combines sampled reconstruction with analytic Gaussian KL; it need not equal the sampled component sum exactly.

E.3 Actual decoder response and full-objective turning points

To locate the objective minimum along each Gaussian noise intervention, we measure how the fixed model's reconstruction loss changes with noise precision. Let $\ell_\theta(x,z) = -\log p_\theta(x \mid z)$, with θ held fixed. Define $D = D_{\mathrm{NLL}}$ and $R = R_{\mathrm{KL}}$ along the intervention. Reparameterization gives the reconstruction derivative (Kingma and Welling, 2014; Rezende et al., 2014)

$$\dot{D}_{\mathrm{dec}} = -\frac{1}{2}\mathbb{E}\left[\nabla_z \ell_\theta(X, Z_t)^\top (Z_t - U)\right]. \tag{75}$$

We keep the decoder weights fixed and differentiate its loss with respect to the latent input. This measures how less encoder noise changes the fixed decoder's reconstruction loss. It does not assume that the decoder attains the Gaussian rate–distortion optimum.

For completeness, W is differentiated directly from its marginal mixtures. At coordinate j, define $a_j(z_j,t) = \partial_{z_j} \log q_t(z_j)$ and $b_j(z_j,t) = \partial_s \log q_{e^s}(z_j)$ with z_j fixed. The weights above give

$$\begin{aligned} a_j(z_j,t) &= -t\sum_r w_{r,j}(z_j,t)(z_j - u_{r,j}), \\ b_j(z_j,t) &= \frac{1}{2} - \frac{t}{2}\sum_r w_{r,j}(z_j,t)(z_j - u_{r,j})^2. \end{aligned} \tag{76}$$

Writing $v_j = -(Z_{t,j} - U_j)/2$ for the latent velocity, and using the fixed standard normal prior, we obtain

$$\dot{W} = \mathbb{E}\sum_j \left[b_j(Z_{t,j},t) + \{a_j(Z_{t,j},t) + Z_{t,j}\}v_j\right]. \tag{77}$$

Thus the marginal derivative is not obtained as a remainder after subtracting $\dot{C}$ and $\dot{T}$ from the full KL derivative.

Derivative-based localization uses

$$g = \dot{D}_{\mathrm{dec}} + \dot{W} + \dot{C}_{\mathrm{MMSE}} + \dot{T}_{\mathrm{MMSE}}, \qquad \frac{dV}{dC} = \frac{g}{\dot{C}_{\mathrm{MMSE}}}. \tag{78}$$

We obtained the root estimate by linear interpolation of the negative-to-positive crossing of g in $s = \log t$, interpolated C on the same segment, and fixed the estimate before examining the loss minimum. Since $\dot{C} > 0$, this is the zero prescribed by Eq. (16). A separate finite-difference check of the full objective gave mean absolute derivative discrepancy 0.00184 and maximum discrepancy 0.00507 at the 96 audited positions. We also test whether the optimal Gaussian benefit can substitute for the actual decoder response in minimum localization. This surrogate sets $C_{\mathrm{loc}} = C$ and replaces the measured benefit $-\dot{D}_{\mathrm{dec}}/\dot{C}_{\mathrm{MMSE}}$ by $B_{\mathrm{G}}(C_{\mathrm{loc}})$. This tests whether the local optimal-coding rule transfers to the global Gaussian noise intervention, including both the identification $C_{\mathrm{loc}} = C$ and the substitution of the optimal reconstruction benefit for the measured decoder response.

For this surrogate, let y_i be the source vector of reference input i and set $D_\nu = \mathrm{diag}(\nu_1,\dots,\nu_6)$ using the factor variances in Appendix D. We use the six eigenvalues λ_{ij} of

$$D_\nu^{1/2} J_{F_0/a}(y_i)^\top J_{F_0/a}(y_i) D_\nu^{1/2}, \qquad i = 1,\dots,N.$$

The shared threshold solves $N^{-1}\sum_{i,j}\frac{1}{2}[\log(\lambda_{ij}/\tau)]_+ = C$, with equal reference weights. This surrogate uses the fixed factor covariance rather than estimated neighborhood covariances.

Table 2: Direct validation of the dependence-growth calculation. Derivatives are with respect to $s = \log t$. Each precision level contains 32 checks. MAE denotes mean absolute error; signed bias is the mean difference (MMSE minus finite difference).

Quantity	Checks	MAE	Signed bias
$\dot{C}$	96	0.0037	0.0023
$\dot{T}$	96	0.0047	−0.0008
$\rho = dT/dC$, all levels	96	0.4090	0.3030
Ratio by precision level	**Checks**	**MAE**	**Maximum absolute error**
First level	32	0.0033	0.0055
Seventh level	32	0.0198	0.0569
Tenth level	32	1.2039	3.3555

Discrete minima and path comparison. Let C_P denote the information at the lowest-loss model in the original trained-model trajectory, and let C_N denote the information at the lowest-loss sampled point of its paired Gaussian noise intervention. Let $\widehat{C}_N$ denote the intervention's derivative-based root estimate. Original checkpoints are ordered by measured C, with ties in loss broken toward lower C. Intervention points are ordered by t, with loss ties broken toward lower C and then t. Each selected minimum has a sampled neighbor on both sides in its respective ordered sequence. The two immediately neighboring samples define a discretization range around each minimum. These ranges are not confidence intervals.

Across the Gaussian noise interventions, all 32 derivative-based root estimates lie inside their intervention neighboring-sample ranges, with a mean absolute error of 0.0967 nat relative to C_N. The geometric surrogate's mean absolute error to C_N is 0.8344 nat, with none of its 32 predictions inside these ranges. This comparison supports using the actual decoder response to localize minima along the fixed-model noise family.

Across the paired trained and intervened paths, the neighboring ranges overlap in all 32 cases, and every root estimate also lies within its corresponding trained-path range. The mean absolute shift $|C_N - C_P|$ is 0.1459 nat, with an eight-run cluster bootstrap 95% interval of $[0.0949, 0.1924]$. The mean absolute root error relative to C_P is 0.1242 nat. Minimum locations therefore agree closely across the paired paths in this benchmark.

E.4 TC COMPONENT ATTRIBUTION AND UNCERTAINTY

Point and interval attribution. The primary evaluation point is the next sampled precision above each observed intervention minimum. It is selected independently of the derivative-based root estimate. Differentiating the full KL in Eq. (34) with respect to $s = \log t$ gives $\dot{R} = m(1 - t^{-1})/2$. The complete derivative can therefore also be computed directly as $\dot{D}_{\mathrm{dec}} + \dot{R}$. Decomposing it through I–MMSE identifies how much of that derivative is attributable to TC, rather than providing a computational shortcut for finding the minimum. Define

$$g_{\mathrm{dir}} = \dot{D}_{\mathrm{dec}} + \dot{R}, \qquad g_{-T} = g_{\mathrm{dir}} - \dot{T}_{\mathrm{MMSE}}. \tag{79}$$

All 32 primary points have $g_{\mathrm{dir}} > 0$ and $g_{-T} < 0$. The pointwise 95% conditional Monte Carlo intervals remain above zero for g_{dir} and below zero for g_{-T}. The mean TC margin $-g_{-T}$ is 1.6547 nat per unit s, with an eight-run cluster interval of $[1.5676, 1.7666]$. All 73 available right-branch points satisfy the same point-estimate sign criterion. This subtracts a measured component on fixed models. It does not describe retraining a model without a TC penalty.

As a separate finite-interval check, the minimum-to-next-right contrast satisfies $\Delta(V - T) < 0 < \Delta V$ for all 32 interventions and all 32 original trained-model trajectories. The latter is an additional local application of Eq. (17). It is distinct from the cross-fitted $M \to H$ comparison in Section 3.4, whose reference is selected on the other seven runs.

Conditional Monte Carlo intervals. The 1,024 inputs are fixed, and the Gaussian draws are independent across input–draw pairs. For any vector of per-draw quantities used in a comparison, let

S_i be its unbiased sample covariance over the four draws at input i. An unbiased estimator of the covariance of its grand mean is

$$\Omega = \frac{1}{N^2 L} \sum_{i=1}^{N} S_i. \tag{80}$$

Paired endpoint differences are formed at matching input–draw indices before this calculation. Between-input variation is excluded because the target is conditional noise uncertainty at the fixed empirical distribution. Ratio intervals use a joint delta method on global means, retaining numerator–denominator covariance. Comparing the two ratios also retains covariance between the estimation-error and density calculations. Ratios are not averaged over per-input ratios. The normal 95% intervals in Fig. 3(a) are pointwise and do not imply simultaneous coverage for all 96 points.

Variation across runs. Continuous summaries across paths use 2,000 bootstrap replicates with random seed 20260908. Resampling uses the eight training runs as independent units, retaining all four conditions within each run; 95% intervals are the 2.5th and 97.5th bootstrap percentiles. Counts such as 32/32 describe the evaluated trajectories.

F Nonlinear Validation of the Allocation Rankings

This appendix expands the targeted check in Section 3.2. The experiment tests the allocation ranking suggested by Lemma 1: does combining local factor variation and observation sensitivity predict the relative distribution of retained factor information better than either ingredient alone?

Data, model, and measurement. The known generator F maps generative variables $Y \in \mathbb{R}^6$ to observations $X \in \mathbb{R}^8$. The VAE reconstructs X using a six-dimensional code and the fixed decoder mean $g_\theta = F$; only its encoder is optimized. After training, a separate predictor h_j estimates factor Y_j from the encoder mean $\bar{Z} = \mathbb{E}[Z \mid X]$, with the VAE held fixed:

$$\underbrace{Y \xrightarrow{F} X}_{\text{data generation}}, \qquad \underbrace{X \xrightarrow{q_\phi} Z \xrightarrow{g_\theta} \widehat{X}}_{\text{VAE reconstruction}}, \qquad \underbrace{\bar{Z} \xrightarrow{h_j} \widehat{Y}_j}_{\text{factor readout}}.$$

The first five source coordinates are evaluation targets; the sixth selects the local region. The geometry matrices below define allocation scores, which are compared with two empirical retention readouts.

Generator and local regions. Draw $Y_6 \sim \mathcal{N}(0,1)$ and assign patch $p \in \{1,\dots,6\}$ using the five boundaries

$$(-0.96742157, -0.43072730, 0, 0.43072730, 0.96742157).$$

Conditional on p, the first five coordinates are independent, centered Gaussians whose variances are the cyclic right shift by $p-1$ of $(4, 2.4, 1.4, 0.85, 0.50)$. Each patch thus defines a region in generative-factor space. For the sensitivities $b_{p,j}$ in Table 3, the generator and observation noise are

$$\begin{aligned} F(y) &= \frac{1}{s_0}\left(\sqrt{b_{p,1}}y_1 \quad \cdots \quad \sqrt{b_{p,6}}y_6 \quad 0.05y_1y_2 \quad 0.05y_3y_4\right)^\top, \\ X &= F(Y) + \frac{\varepsilon}{s_0}, \qquad \varepsilon \sim \mathcal{N}(0, 0.1 I_8), \end{aligned} \tag{81}$$

where $s_0 = 1.20945203524494$ is a fixed common scale. The decoder uses the same scaling, and the Gaussian likelihood variance is $\sigma_x^2 = 0.1/s_0^2$. Independent training, validation, and test sets contain 60,000, 10,000, and 20,000 observations, generated with seeds 260916, 260917, and 260918, respectively.

Local matrices and factor scores. Let $\mathcal{J}_p$ index all test-set source vectors in patch p, let $\bar{y}_p$ be their mean, and let $\breve{y}_p$ be their coordinatewise median. The matrices are

$$\begin{aligned} \Sigma_p &= \frac{1}{|\mathcal{J}_p| - 1} \sum_{i \in \mathcal{J}_p} (y_i - \bar{y}_p)(y_i - \bar{y}_p)^\top, \\ G_p &= J_F(\breve{y}_p)^\top J_F(\breve{y}_p), \qquad S_p = \Sigma_p^{1/2} G_p \Sigma_p^{1/2}. \end{aligned} \tag{82}$$

Table 3: Patch-specific squared linear sensitivities $b_{p,j}$ in Eq. (81). The generator multiplies coordinate j by $\sqrt{b_{p,j}}/s_0$.

Patch p	$b_{p,1}$	$b_{p,2}$	$b_{p,3}$	$b_{p,4}$	$b_{p,5}$	$b_{p,6}$
1	0.378945	0.988387	0.683018	1.730824	2.156672	3.594958
2	2.615387	0.453734	0.718746	0.285986	2.409575	1.416035
3	1.406622	0.995416	0.155272	0.557204	0.852518	0.621367
4	0.655075	1.253343	1.421046	0.252104	0.287827	0.960871
5	5.863425	1.292756	2.170491	6.993115	0.884654	1.674656
6	0.367187	0.827264	0.392127	0.778333	2.290802	0.686903

All three matrices are 6×6. The generator is smooth within each patch and discontinuous at patch boundaries. The 8×6 Jacobian differentiates the scaled generator with respect to all six source coordinates at the patch center, holding the patch assignment fixed locally. Thus G_p uses one Jacobian per patch. The numerical square root floors covariance eigenvalues at 10^{-10}; all observed eigenvalues exceed this floor.

Write $S_p v_{p,k} = \lambda_{p,k} v_{p,k}$ with orthonormal eigenvectors. The Hungarian algorithm chooses a permutation π_p maximizing $\sum_{j=1}^{6} |(v_{p,\pi_p(j)})_j|$. Matching uses the components in the covariance-standardized coordinates of S_p, without mapping directions back by $\Sigma_p^{1/2}$. The strengths for the three predictors are

$$a^S_{p,j} = \lambda_{p,\pi_p(j)}, \qquad a^G_{p,j} = (G_p)_{jj}, \qquad a^{\Sigma_{\mathrm{loc}}}_{p,j} = (\Sigma_p)_{jj}. \tag{83}$$

Only S_p uses eigendirection matching; the two component predictors use coordinatewise diagonal entries. Matching is performed over all six coordinates before retaining the five evaluation targets. It assigns factor labels to the scores without identifying learned latent coordinates with eigenmodes.

Evaluation grid and allocation scores. Eight training seeds (5401–5408) each yield six checkpoints with target KL values $0.20, 0.35, 0.55, 0.80, 1.50, 4.20$. For this allocation study, checkpoints minimize the validation capacity-penalized objective used in training. Each comparison unit is one checkpoint, patch, and target factor: $6 \times 6 \times 5 = 180$ units per seed and $1{,}440$ in total. Let $R = R_{\mathrm{KL}}$ be the checkpoint's achieved mean analytic encoder KL on the validation set. Across the 30 patch–target pairs, normalize the component strengths by

$$\widetilde{a}^P_{p,j} = a^P_{p,j} \frac{\mathrm{median}_{p,j}(a^S_{p,j})}{\mathrm{median}_{p,j}(a^P_{p,j})}, \qquad P \in \{G, \Sigma_{\mathrm{loc}}\}, \qquad \widetilde{a}^S = a^S.$$

All three medians are strictly positive. The allocation scores are

$$c^P_{s,R,p,j} = \frac{1}{2}\left[\log \frac{\widetilde{a}^P_{p,j}}{\tau_S(R)}\right]_+, \qquad \frac{1}{6}\sum_{p=1}^{6}\sum_{j=1}^{5} c^S_{s,R,p,j} = R, \tag{84}$$

with score zero at zero strength. No floor is applied to the strengths. The common threshold $\tau_S(R)$ is found by 120 bisection steps on $[10^{-10}, 2\max_{p,j} a^S_{p,j}]$. Only the S scores satisfy the budget equality. This operational KL calibration evaluates local allocation rankings; the scores are not measured mutual information for individual factors.

Readout 1: Gaussian canonical-correlation information proxy. For each checkpoint and patch, draw two posterior codes $z \in \mathbb{R}^6$ for each of $1{,}024$ test inputs. The same input indices, selected without replacement using seed 9102, are used across checkpoints. Each target value y_j is repeated for its two draws, giving $2{,}048$ paired observations. Let $\widehat{\Sigma}_{zz}$, $\widehat{\Sigma}_{y_j z} = \widehat{\Sigma}_{z y_j}^{\top}$, and $\widehat{\Sigma}_{y_j y_j}$ be their mean-centered sample covariances, each with denominator $2{,}047$. The implemented statistic is

$$\begin{aligned} r^2_{s,R,p,j} &= \mathrm{clip}_{[0,\,1-10^{-10}]}\left(\frac{\widehat{\Sigma}_{y_j z}(\widehat{\Sigma}_{zz} + 10^{-8} I_6)^{-1}\widehat{\Sigma}_{z y_j}}{\max\{\widehat{\Sigma}_{y_j y_j}, 10^{-8}\}}\right), \\ \widehat{I}_{s,R,p,j} &= -\frac{1}{2}\log\!\left(1 - r^2_{s,R,p,j}\right). \end{aligned} \tag{85}$$

All target variances are positive. The full code is used because information about a factor can be distributed across latent coordinates. Without regularization or clipping, the transform equals mutual information for jointly Gaussian variables; here it provides a covariance-based information proxy.

Readout 2: test-set factor-coordinate decoding. For each patch and target, ridge regression fits y_j from the six-dimensional encoder mean using $2{,}048$ training inputs, selected without replacement with seed 9101, and evaluates on the same $1{,}024$ test inputs used above. Each input coordinate is centered by its fitting-set mean and divided by its fitting-set standard deviation (computed with divisor $2{,}048$), floored at 10^{-6}. For the standardized mean $\widetilde{z}_i$, the fitted coefficients minimize

$$\sum_{i \in \mathcal{I}_{\mathrm{fit}}} (y_{i,j} - b - w^\top \widetilde{z}_i)^2 + 0.01\|w\|^2,$$

so the penalty applies to the sum of squared errors and leaves the intercept b unpenalized. With $N_{\mathrm{test}} = 1{,}024$ and test-set target mean $\bar{y}_{\mathrm{test},j}$, the evaluation statistic is

$$R^2_{\mathrm{probe},s,R,p,j} = 1 - \frac{\sum_{i \in \mathcal{I}_{\mathrm{test}}} (\widehat{y}_{i,j} - y_{i,j})^2}{\max\{\sum_{i \in \mathcal{I}_{\mathrm{test}}} (y_{i,j} - \bar{y}_{\mathrm{test},j})^2,\ N_{\mathrm{test}} 10^{-12}\}}. \tag{86}$$

The variance floor is inactive for every evaluated target. Geometry is estimated from the test-set factor support before model evaluation, so this readout measures prediction on samples held out from probe fitting within known geometric regions. The unit-level probe values range approximately from 0.792 to 0.992.

Within-seed rank agreement. For readout $T \in \{\widehat{I}, R^2_{\mathrm{probe}}\}$ and predictor $P \in \{S, G, \Sigma_{\mathrm{loc}}\}$, define

$$\rho_s^{P,T} = \mathrm{Spearman}\big(\{c^P_{s,R,p,j}\}_{R,p,j}, \{T_{s,R,p,j}\}_{R,p,j}\big), \qquad \bar{\rho}^{P,T} = \frac{1}{8}\sum_{s=1}^{8} \rho_s^{P,T}. \tag{87}$$

Both vectors contain the same 180 comparison units. Units are unweighted, tied values receive average ranks, and correlations are computed within each seed before averaging.

Table 4: Mean within-seed Spearman correlations between allocation scores and the two factor-retention readouts.

Allocation predictor	Rank correlation with Gaussian-CCA proxy	Rank correlation with test-set probe R^2
Joint $S = \Sigma_{\mathrm{loc}}^{1/2} G \Sigma_{\mathrm{loc}}^{1/2}$	$\mathbf{0.879}$	$\mathbf{0.493}$
Metric G alone	0.689	0.249
Covariance Σ_{loc} alone	0.631	0.149

Interpretation of the ranking evidence. Each within-seed correlation spans six rates, six regions, and five targets, so common rate trends can contribute to agreement. On this across-rate task, S improves mean Gaussian-CCA rank agreement by 0.190 over G and 0.248 over Σ_{loc}. The corresponding probe-ranking gains are 0.244 and 0.344. For the probe readout, the gain over G is positive in all eight seeds, with a 95% interval of $[0.240, 0.248]$. This percentile interval uses $4{,}000$ bootstrap resamples of the eight paired seed-level differences $\rho_s^{S,R^2_{\mathrm{probe}}} - \rho_s^{G,R^2_{\mathrm{probe}}}$, with random seed 20260816. These gains support combining local variation and observation sensitivity to predict relative retention across the evaluated combinations.