PERFECT RECONSTRUCTION CANNOT SHOW WHETHER TWO CODEBOOKS ON A VAE'S LATENT SPACE AGREE

Anonymous authors
Paper under double-blind review

ABSTRACT

Many systems read the latent space of a variational autoencoder (VAE) as discrete codes, and rely on each latent being read back as the code it was written as. Shannon (1948) took such a shared codebook as given; inside a VAE nothing enforces it. We prove that no item of a standard training report can reveal whether the written and read codes agree on encoded data. We construct two VAEs that reconstruct every input perfectly, infer exactly, and match on every per-example ELBO and every code histogram, yet agreement is one in the first and zero in the second. We then show that the ELBO's own terms bound how far a measured agreement can move: the variational gap bounds the shift to exact inference, and the rate bounds the shift to generation from the prior. For Gaussian posteriors, a mean-field encoder pays an exact price in the gap that depends on its choice of axes. On probabilistic PCA the gap certifies a trained encoder in all 75 settings we test, at agreement between 0.76 and 0.97. On 240 trained VAEs, a floor set by the rate and the dataset size rules out rate-side certification in advance for 66 models, and none of them certifies. The results yield a simple protocol, Certify-or-Measure, for any pipeline that reads a VAE as codes.

1 INTRODUCTION

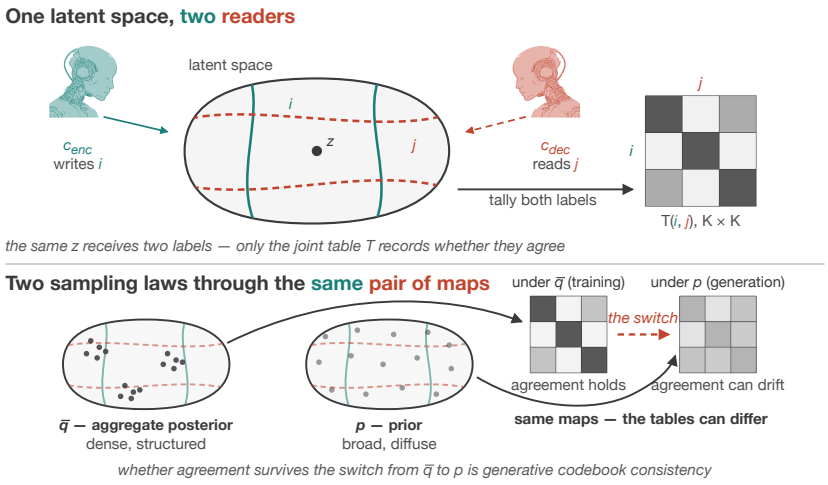

Figure 1: **A pipeline that reads a VAE's latent space as codes holds two codebooks on one space, and their agreement can change when the model generates.** Top: c_{enc} writes a latent z as code i, c_{dec} reads the same z as code j, and the joint table $T(i, j)$ counts the pairs; agreement is its diagonal mass. Bottom: the same maps under the aggregate posterior $\bar{q}$ of encoded data (agreement $\mathcal{A}_q$) and under the prior p used for generation ($\mathcal{A}_{\text{prior}}$). No item of a standard training report reveals $\mathcal{A}_q$ (Theorem 1); the rate bounds the drift from $\mathcal{A}_q$ to $\mathcal{A}_{\text{prior}}$, the failure of generative codebook consistency (Theorem 2b).

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.

Many systems read the latent space of a learned model as a small set of discrete codes. They cluster latents (Banerjee et al., 2005), quantize them into tokens (van den Oord et al., 2017) or for compression (Farvardin, 1990), and read latent units as interpretable factors (Higgins et al., 2017; Chen et al., 2018). Every such pipeline relies on a contract it rarely states: a latent written as code i is read back as code i. Shannon (1948) built communication theory on a codebook that transmitter and receiver share by assumption, and Weaver (1949) set the meaning of the symbols aside as a separate level. Inside one learned model, nothing guarantees that the codebook is shared. This paper asks whether it is, and whether anyone would notice if it were not.

A variational autoencoder (VAE) (Kingma & Welling, 2014; Rezende et al., 2014) is the natural place to ask. A pipeline that reads its latent space through K codes declares two *code maps* on that space before it writes anything. The encoder-side map c_{enc} gives the code a latent is written as; in practice it is fitted on encoder statistics. The decoder-side map c_{dec} gives the code the latent is read as; in practice it is a nearest-prototype rule on decoded outputs. Both are fixed partitions of the latent space, so the pipeline holds two codebooks on one space (Figure 1, top). Their *agreement* is the probability that one latent receives the same code from both. We write $\mathcal{A}_q$ for agreement on

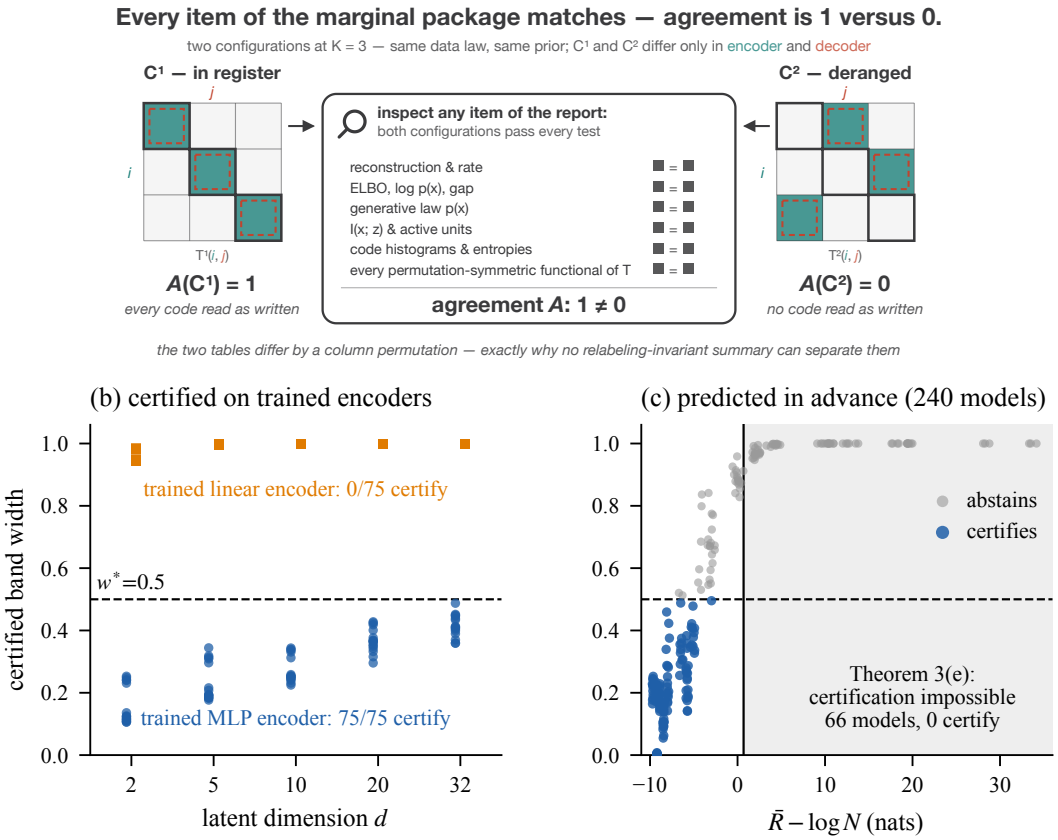

Figure 2: **No item of a standard training report tells the two configurations in (a) apart, yet the ELBO's own terms certify agreement on trained encoders (b) and predict in advance where no certificate can exist (c).** (a) Top: the witness pair of Theorem 1 at $K = 3$. Rows give the code a latent is written as (c_{enc}), columns the code it is read as (c_{dec}), and agreement is the mass on the outlined diagonal. Both configurations share the data, the prior and both code maps, reconstruct every input perfectly with zero variational gap, and match on every item of the marginal package (Definition 1). (b) Probabilistic PCA, where $\log p(x)$ is exact: width of the band that the variational gap places around each measured agreement (Theorem 2a), 15 cells per latent dimension (5 folds, 3 code maps). A band certifies when its width is at most $w^* = 1/2$. The trained mean-field MLP encoder certifies in every cell, and every band contains the agreement under the exact posterior (Appendix D.5). (c) Rate-side band width for 240 trained VAEs against $\bar{R} - \log N$, the average rate minus the log size of the audited split. Theorem 3(e) forbids certification in the shaded region, $\bar{R} - \log N > \log 2$ (Section 6.3).

encoded data and $\mathcal{A}_{\text{prior}}$ for agreement when the model generates from its prior (Figure 1, bottom). Training never asks for agreement. Practitioners watch a standard report instead: the evidence lower bound (ELBO) of each example with its reconstruction term and its *rate* (its KL divergence to the prior; Alemi et al., 2018), the log-likelihood, the information the latent carries about the input, active units, and how often each code is used. Reconstruction ties the decoder to the encoder, so a clean report looks like strong evidence that the two ends read one codebook.

In this work, we show that the report cannot settle the question, and that two quantities the ELBO already provides bound how far a measured agreement can move. Theorem 1 gives the negative result (Figure 2a). For every codebook size we construct two configurations that share the data, the prior and both code maps. Both reconstruct every input perfectly, and in both the *variational gap*, the amount by which the ELBO falls below $\log p(x)$, is zero. They match on every item of the standard report (Definition 1), example by example and code by code, and they generate the same data distribution, so no sample-quality score such as FID (Heusel et al., 2017) can separate them. Yet $\mathcal{A}_q$ is one in the first configuration and zero in the second. The joint table of written against read codes, kept in the pipeline's own indices, shows the difference at once.

The positive result is that the variational gap and the rate, read as *budgets* in nats, bound how far a measured agreement can move (Theorem 2). The gap bounds the distance between $\mathcal{A}_q$ and its value under exact inference; the rate bounds the distance between $\mathcal{A}_q$ and $\mathcal{A}_{\text{prior}}$. Both bounds are stated in the divergence between two coin flips, so they turn a budget into an interval, and a budget B allows no interval narrower than $1 - e^{-B}$ (Lemma 1). A budget must be a proved upper bound to be *licensed*; common plug-in estimates are lower bounds and are not (Lemma 2). For Gaussian posteriors, Theorem 3 gives the exact price that a mean-field encoder, which works coordinate by coordinate, pays in the gap: it depends on the choice of axes and vanishes in the posterior's own axes, while the certified quantities do not depend on the axes at all. The same theorem gives a floor: a latent cannot carry more than $\log N$ nats about the N examples over which the rate is averaged, so a rate above $\log N + \log 2$ rules out a rate-side band narrower than one half before any audit is run.

Trained models behave as the theory predicts (Section 6). On probabilistic PCA, where the likelihood is exact, the gap certifies a trained mean-field encoder in all 75 cells of a sweep over latent dimension, at agreement between 0.76 and 0.97, and every certified band contains the exact value (Figure 2b). A VAE trained on MNIST loses agreement when it generates, $\mathcal{A}_q = 0.770$ on encoded data against $\mathcal{A}_{\text{prior}} = 0.525$ under the prior over five seeds, and the same drift appears in probabilistic PCA where every quantity is exact. On the same probabilistic PCA models, a rotation of the latent axes moves the mean-field budget from 0.000 to 0.474 nats while the certified quantities stay fixed, as Theorem 3 predicts. On six SVHN models, a new licensed budget built from the rate is 2.5 to 3 times tighter than the best previously available one (Kolchinsky & Tracey, 2017). Across 240 trained VAEs on four datasets, the dataset floor rules out rate-side certification in advance for 66 models, and none of them certifies (Figure 2c).

Contributions.

- We prove that no item of a standard training report, including the per-example ELBO at perfect reconstruction and exact inference, can reveal whether a VAE's two code maps agree on encoded data (Theorem 1, Definition 1, Figure 2a).
- We show that the variational gap and the rate bound how far measured agreement can move under exact inference and under generation, prove that a budget B allows no band narrower than $1 - e^{-B}$, and state which budget estimates are licensed (Theorem 2, Lemmas 1 and 2).
- We price the certificate exactly for Gaussian posteriors under mean-field computation, and show that the rate and the dataset size decide in advance when rate-side certification is impossible (Theorem 3).
- We assemble the results into a protocol, Certify-or-Measure, and test both certificates on trained models: each fires only where the theory says it can, and none fires where the dataset floor rules it out (Sections 5.2 and 6, Figure 2b,c).

A clean training report does not show that the code a pipeline writes is the code it reads. Log the joint code table, and certify from the ELBO's own terms when their budgets are small enough.

2 SETUP: TWO CODE MAPS, ONE CHANNEL

Data $x \sim \mathcal{D}$; encoder $q(z \mid x)$ on $\mathcal{Z}$ with prior p; decoder $p(x \mid z)$; aggregate posterior $\bar{q} = \mathbb{E}_x q(\cdot \mid x)$. A *configuration* is $C = (\mathcal{D}, q, p(x \mid z), c_{\mathrm{enc}}, c_{\mathrm{dec}})$ where $c_{\mathrm{enc}}, c_{\mathrm{dec}} : \mathcal{Z} \to [K]$ are fixed measurable code maps. The maps are taken as given. In applications c_{enc} is fit on encoder statistics and c_{dec} is a nearest-prototype rule in the decoder's Bregman output geometry (Banerjee et al., 2005). Theorem 2 holds for any fixed pair, Theorem 1 constructs its own, and the experiments declare theirs. Each code map partitions $\mathcal{Z}$ into K labeled cells, a codebook on the latent space, so the pair gives two codebooks on one space.

For a sampling law μ on $\mathcal{Z}$ the pair $(c_{\mathrm{enc}}, c_{\mathrm{dec}})$ has joint table $T^{\mu}(i,j) = \mu(\{c_{\mathrm{enc}} = i\} \cap \{c_{\mathrm{dec}} = j\})$ (Figure 1). Agreement under μ is $\mathcal{A}[\mu] = \mu(\mathrm{Agr})$ with $\mathrm{Agr} = \{z : c_{\mathrm{enc}}(z) = c_{\mathrm{dec}}(z)\}$; we write $\mathcal{A}_q = \mathcal{A}[\bar{q}]$ (posterior sampling) and $\mathcal{A}_{\mathrm{prior}} = \mathcal{A}[p]$ (prior sampling), and $d_{\mathrm{bin}}(a\|b) = a \log \frac{a}{b} + (1-a) \log \frac{1-a}{1-b}$. A third reading appears in the experiments: $\mathcal{A}_{\mathrm{mean}}$, agreement of the *mean* reading (both maps evaluated at the posterior means $\mathbb{E}[z \mid x]$ rather than at draws).

Notation. $[K] = \{1, \ldots, K\}$; S is the number of posterior draws per datum when sampling is Monte Carlo. Two ELBO-derived scalars recur: the average rate $\bar{R} = \mathbb{E}_x D_{\mathrm{KL}}(q(\cdot \mid x)\|p)$ and the average variational gap $\bar{\Delta} = \mathbb{E}_x[\log p(x) - \mathrm{ELBO}(x)]$; $\bar{R}$ is logged in every VAE run and $\bar{\Delta}$ wherever $\log p(x)$ is tractable, and Section 4 reads them as *budgets*; an estimator is *licensed* when it is a certified upper bound on the quantity it replaces (Section 5.1), and the certified quantity is the *estimand*. For a configuration we write $\mathcal{A}(C) = \mathcal{A}[\bar{q}_C]$, and we apply $\mathcal{A}$ to a joint table where convenient, $\mathcal{A}(T) = \sum_i T(i,i)$, which is the same functional in table coordinates.

Definition 1 (Marginal package)**.** *The* marginal package $M(C)$ *of a configuration comprises: (m1) per-example reconstruction* $\mathbb{E}_{q(z|x)} \log p(x \mid z)$*; (m2) per-example rate* $D_{\mathrm{KL}}(q(\cdot \mid x)\|p)$*; (m3) per-example ELBO, marginal likelihood* $\log p(x)$*, variational gap* $\Delta(x) = \log p(x) - \mathrm{ELBO}(x)$*, and the generative law* $p(x)$ *as a distribution; (m4)* $I(x;z)$ *under the encoder joint and the per-dimension active-units statistics* $\mathrm{AU}_k = \mathrm{Var}_x \mathbb{E}[z_k \mid x]$*; (m5) encoder and decoder code histograms under* $\bar{q}$*, their entropies and occupied-code counts; (m6) every permutation-symmetric functional of the joint table* $T^{\bar{q}}$ *(every* F *with* $F(P_1 T P_2^{\top}) = F(T)$ *for permutation matrices* P_1, P_2*), including* $I(c_{\mathrm{enc}}; c_{\mathrm{dec}})$*, all Rényi mutual informations, joint entropy, both histograms as multisets, and the permutation-matched agreement* $\max_{\pi} \sum_i T(i, \pi(i))$*; (m7) prior-law agreement* $\mathcal{A}_{\mathrm{prior}}$.

Items (m1)–(m5) are what standard practice logs, and (m6) says that reporting the joint table up to relabeling does not help either. The name *marginal* is short and understates the package: reconstruction, the ELBO and $\log p(x)$ couple encoder to decoder, (m6) is a functional of the coupled table, and (m7) is agreement under the prior.

Scope. Invisibility concerns agreement in the pipeline's own indices under fixed maps, the quantity that matters when a pipeline writes code i and expects code i back. In the pair of Theorem 1(I) the code histograms are identical index by index, and the two joint tables are column permutations of each other, so reporting T both detects and repairs this pair. Proposition 1 gives a closed-form pair with equal histograms and equal $I(c_{\mathrm{enc}}; c_{\mathrm{dec}})$ whose tables are not relabelings of each other; Proposition 2 shows that matched summaries also leave the error structure free.

3 INVISIBILITY

Theorem 1 (Invisibility)**.** ***(I) One pair, whole package, shared data.*** *For every* $K \geq 2$ *there exist configurations* C^1, C^2 *with the same data law and prior, differing only in encoder and decoder, with* $M(C^1) = M(C^2)$*—every item of Definition 1, per example and in aggregate, to exact equality, including* $p(x)$ *as a distribution—while* $\mathcal{A}(C^1) = 1$ *and* $\mathcal{A}(C^2) = 0$.

(I′) Moment blindness with a sharp threshold. *For every* $K \geq 2$ *and* $m \geq 2$ *the pair in (I) can be built so that, in addition: (N1) per-example posterior means are identical example by example, so every functional of the means—per-dimension AU, their full covariance, any statistic—coincides; (N2) every aggregate-posterior mixed moment of degree below* m *in the coordinates that* c_{dec} *reads coincides; (N3) the threshold is sharp, with explicit degree-*m *cross moments separating the pair. The blindness order* m *is a design parameter independent of* K.

(II) Orbit form, for trained models. *Every prior-preserving bimeasurable bijection τ of $\mathcal{Z}$ leaves (m1)–(m3) and $I(x;z)$ exactly invariant for every x (and per-dimension AU when τ^{-1} acts as a translation on $\mathrm{supp}(\bar{q})$), while agreement transforms as $\mathcal{A}(C_\tau) = \bar{q}(\tau(\mathrm{Agr}))$; the full statement and proof are in Appendix A.5.*

The grid witness. Partition $[0,1)^2$ with a uniform prior into K^2 cells, with c_{enc} = row and c_{dec} = column. C^1 couples the data atoms to the diagonal cells and C^2 to a derangement σ of them (Figure 2a). Inference is exact in both configurations and every package item coincides. This includes $p(x)$, which is uniform over the same K^2 atoms, so no statistic of generated data, from two-sample tests to sample-quality scores such as FID and KID (Heusel et al., 2017; Bińkowski et al., 2018), can separate the pair. Yet agreement is 1 versus 0. Appendix A.1 gives the computation and shows that Gaussian emissions preserve the equality exactly.

Companions. Proposition 1 (Appendix A.4) carries the impossibility to non-uniform marginals: encoder and decoder histograms are equal to the same ν index by index, $I(c_{\mathrm{enc}}; c_{\mathrm{dec}})$ is matched exactly, and agreement differs (0 versus positive when $\max_i \nu_i \le \frac{1}{2}$), with exact feasible set $[\max(0, 2\max_i \nu_i - 1), 1]$. The grid pair still differs in one statistic outside the package, the per-example posterior means; the interleaved witness of Part (I′) removes that difference (Lemma 3, Appendix A.2). Appendix A collects all proofs, and Appendix A.3 states what the theorem does not claim and records the numerical verification of both witness constructions to machine precision.

4 Two-sided certification

The standard report cannot detect the failure, even at perfect reconstruction. Two scalars derived from the ELBO, the gap and the rate, can bound it, one each. Recall $\bar{\Delta} \ge 0$ and $\bar{R}$ from Section 2; write $p_\theta(\cdot \mid x)$ for the *exact* posterior of the model, $\bar{p}_\theta = \mathbb{E}_x\, p_\theta(\cdot \mid x)$ for its aggregate, and $\bar{\eta}_p = \mathbb{E}_x\, p_\theta(\{c_{\mathrm{enc}} \ne c_{\mathrm{dec}}\} \mid x)$ for the disagreement rate under exact-posterior sampling. Throughout this section assume $q(\cdot \mid x) \ll p_\theta(\cdot \mid x)$ for $\mathcal{D}$-a.e. x.

Theorem 2 (Two-sided certification)**.** ***(a) Reconstruction side.*** *For any fixed measurable $E \subseteq \mathcal{Z}$, $d_{\mathrm{bin}}\big(\bar{q}(E) \,\|\, \bar{p}_\theta(E)\big) \le \bar{\Delta}$; with $E = \{c_{\mathrm{enc}} \ne c_{\mathrm{dec}}\}$,*

$$d_{\mathrm{bin}}\big(1 - \mathcal{A}_q \,\|\, \bar{\eta}_p\big) \;\le\; \bar{\Delta},$$

and in fact $\bar{\Delta}$ decomposes as the sum of this binary-KL term, a Jensen term, and a within-cell term, all non-negative.

(b) Generative side. *For the same event,*

$$d_{\mathrm{bin}}\big(\mathcal{A}_q \,\|\, \mathcal{A}_{\mathrm{prior}}\big) \;\le\; D_{\mathrm{KL}}(\bar{q}\|p) \;\le\; \bar{R}.$$

Proof. Appendix B; (a) is the binary specialization of the post-processing decomposition of the gap, (b) is data processing plus the rate decomposition (Hoffman & Johnson, 2016).

Part (a) certifies the posterior-sampling audit and part (b) the generative one. Inverting (b) at an observed $\mathcal{A}_q$ and a budget $B \in \{D_{\mathrm{KL}}(\bar{q}\|p), \bar{R}\}$ yields a certified band $\{a : d_{\mathrm{bin}}(\mathcal{A}_q\|a) \le B\} \ni \mathcal{A}_{\mathrm{prior}}$. Its one-dimensional form loses nothing, because by Theorem 1(I) richer permutation-symmetric summaries cannot flag the failure anyway. All of its leverage comes from the budget, whose price and floor are the subject of Section 5.

5 The price of certification, and Certify-or-Measure

Theorem 2 is stated in exact quantities, but mean-field training computes rates coordinate-wise and certificates inherit the coordinate system. This section prices that inheritance exactly for Gaussian posteriors: exactly in linear-Gaussian models, and per example for the Gaussian approximation otherwise.

Let $\mathcal{Z} = \mathbb{R}^d$ and the posterior at x be $N(m_x, \Sigma)$ with precision $\Lambda = \Sigma^{-1} \succ 0$ and eigenvalues λ, condition number κ. For a basis $Q \in O(d)$ let $\mathcal{F}_Q = \{N(\mu, Q^\top D Q) : D \text{ diagonal} \succ 0\}$ be the mean-field family in basis Q, and $B_Q = \min_{q \in \mathcal{F}_Q} D_{\mathrm{KL}}(q\|N(m_x, \Sigma))$ the smallest certificate budget achievable in that basis.

Theorem 3 (The coordinate price of certification). *Write* $\Lambda_Q = Q\Lambda Q^\top$.

(a) Exact price and worst basis. $B_Q = \frac{1}{2}\log\big[\det \operatorname{diag}(\Lambda_Q)/\det \Lambda_Q\big]$, *attained at* $\mu^* = m_x$, $D^*_{ii} = 1/(\Lambda_Q)_{ii}$. *Moreover (i)* $B_Q = 0$ *iff* Λ_Q *is diagonal—exactly on the posterior eigenbasis, up to axis permutations, signs, and within-eigenspace rotations; (ii)* $\sup_Q B_Q = \frac{d}{2}\log\big(\mathrm{AM}(\lambda)/\mathrm{GM}(\lambda)\big)$, *finite and attained (Schur–Horn, Marshall et al., 2011); (iii) restricting the product family to Gaussians is without loss.*

(b) Estimand invariance under transport. *For* $R \in O(d)$, *transport the configuration:* $z \mapsto Rz$, *decoder* $g'(w) = g(R^\top w)$, *posterior* $N(Rm_x, R\Sigma R^\top)$, *prior* $R_\# p$, *and* transported *maps* $c'(w) = c(R^\top w)$. *Then the joint law of* $(c'_{\mathrm{enc}}, c'_{\mathrm{dec}})$ *under every transported sampling law equals the original joint law; in particular* $\mathcal{A}_q$ *under full-covariance posteriors,* $\mathcal{A}_{\mathrm{prior}}$, *and every functional of code-pair laws are exactly invariant.*

(c) The asymmetry. *For a fixed model,* $Q \mapsto B_Q$ *ranges over the closed interval* $\big[0, \frac{d}{2}\log(\mathrm{AM}/\mathrm{GM})\big]$ *while every transported estimand is constant; across models the price is unbounded (*$\kappa \to \infty$*). A mean-field-budget certificate is a property of the (model, basis) pair; the certified quantity is a property of the model. Reporting a mean-field KL budget without its basis is ill-posed; the price-minimizing basis is the posterior eigenbasis. The declared code map is the exact dual. Theorem 2 holds for* every *fixed measurable pair* $(c_{\mathrm{enc}}, c_{\mathrm{dec}})$ *at once, and a rate budget is a functional of* $\bar{q}$ *alone, so changing the map moves the certified* quantity *and leaves the* budget *fixed—the mirror of the basis, which moves the budget and leaves the quantity fixed. Agreement is a property of the (model, declared map) pair and the budget a property of the model; reporting either without the other is ill-posed, and Section 6.3 measures both halves on one checkpoint.*

(d) Informative floor. *Let* $y \in [L]$ *be any finite-alphabet label with* $y \perp z \mid x$, *and let* α *be the held-out accuracy of any predictor of* y *from the* sampled *code* $c_{\mathrm{enc}}(z)$, $z \sim q(\cdot \mid x)$. *Then with* $h_2(\cdot)$ *the binary entropy in nats and* $\gamma(\alpha) = H(y) - h_2(1-\alpha) - (1-\alpha)\log(L-1)$ *(Fano's inequality, Cover & Thomas, 2006),*

$$\bar{R} \;=\; D_{\mathrm{KL}}(\bar{q}\|p) + I(x;z) \;\geq\; I(x;z) \;\geq\; I(x; c_{\mathrm{enc}}(z)) \;\geq\; I(y; c_{\mathrm{enc}}(z)) \;\geq\; \gamma(\alpha),$$

and by Lemma 1 every band $\{a : d_{\mathrm{bin}}(\mathcal{A}\|a) \leq B\}$ *with* $B \geq \gamma$ *has width at least* $1 - e^{-\gamma}$. *At* $\gamma \geq 1$ *nat the width is* $\geq 1 - e^{-1} \approx 0.632$ *regardless of the observed agreement.*

(e) Dataset floor. *If the data law is empirical over* N *points—the same* N *points that define* $\bar{q} = \frac{1}{N}\sum_i q(\cdot \mid x_i)$ *and* $\bar{R}$*—then* $I(x;z) \leq H(x) = \log N$, *so* $D_{\mathrm{KL}}(\bar{q}\|p) \geq \bar{R} - \log N$*:* every *licensed budget obeys* $B \geq \bar{R} - \log N$, *the band has width at least* $1 - e^{-(\bar{R} - \log N)}$, *and certification at width* w^* *is* ***impossible*** *whenever* $\bar{R} > \log N - \log(1-w^*)$*—no estimator, however tight, can change this.*

The optimal mean-field Gaussian fit in (a) is classical (Bishop, 2006); what is new is reading its value as a *price paid in* Q and pairing it with the invariance statement (b).

Lemma 1 (Width). *For* $a \in [0,1)$ *and* $B > 0$ *let* $W(a,B)$ *be the width of* $\{x : d_{\mathrm{bin}}(a\|x) \leq B\}$. *Then* $W(a,B) \geq 1 - e^{-B}$, *with equality iff* $a = 0$.

Inverting it (proof in Appendix C) gives a *necessary* admission condition: width $\leq w^*$ requires $B \leq -\log(1-w^*)$ nats (0.693 at $w^* = 0.5$). It is not sufficient, because the width also depends on the observed agreement. Writing $B^*(a)$, the *admission budget*, for the largest budget whose band at agreement a has width $\leq w^*$: at $w^* = 0.5$,

$$B^*(0) = 0.693, \quad B^*(0.1) = 0.370, \quad B^*(0.25) = 0.195, \quad B^*(0.5) = 0.144,$$

symmetrically above $\frac{1}{2}$. The requirement is thus 1.9–4.8 times *stricter* than the 0.693 headline across this range, and it is $B \leq B^*(\mathcal{A}_q)$, not the headline, that the protocol of Section 5.2 admits on.

Appendix C proves (a)–(c), with the exact zero-set coset, a Kantorovich corollary, and the fact that B_Q depends only on $\mathrm{Corr}(\Lambda_Q)$; the supremum comes from the Schur–Horn characterization $\{\operatorname{diag}(Q\Lambda Q^\top) : Q \in O(d)\} = \{u : u \prec \lambda\}$. Three scope notes on (b), on refitting versus transporting maps, the exclusion of mean-field $\mathcal{A}_q$, and padding, are collected in Appendix C.3.

5.1 WHICH BUDGETS ARE LICENSED

Theorem 2(b) admits any $B \geq D_{\mathrm{KL}}(\bar{q}\|p)$, so the protocol's soundness rests entirely on the estimator being an *upper* bound. The two most natural tightenings are both wrong in the direction that matters.

Lemma 2 (Budget validity)**.** *Let f be any measurable coarsening of $\mathcal{Z}$ and $\tilde{q}$ the moment-matched Gaussian of $\bar{q}$. Then*

$$D_{\mathrm{KL}}(f_{\#}\bar{q}\|f_{\#}p) \;\leq\; D_{\mathrm{KL}}(\bar{q}\|p), \qquad D_{\mathrm{KL}}(\tilde{q}\|p) \;\leq\; D_{\mathrm{KL}}(\bar{q}\|p) \text{ when } p \text{ is Gaussian},$$

so binned/grid plug-ins (data processing) and moment-matched Gaussian plug-ins (maximum entropy at fixed moments) are lower *bounds and are not licensed budgets. By contrast $\bar{R}$ is licensed but loose by exactly $I(x;z)$; and for any certified lower bound $\hat{I} \leq I(x;z)$, $\bar{R} - \hat{I}$ is licensed.*

The last clause is the constructive one (proof in Appendix C.4). For a VAE, $\bar{q}$ is a finite Gaussian mixture with an exactly computable density, so $I(x;z) = \mathbb{E}_{i,\, z\sim q_i}[\log q_i(z) - \log \bar{q}(z)]$ is estimated *without bias*, unlike contrastive bounds (capped at $\log M$; van den Oord et al., 2018; Poole et al., 2019) or small-mixture plug-ins (floored at $\bar{R} - \log N'$). The statistic is clipped below at $-c$, which can only *raise* its mean, and by at most e^{-c} (Lemma 4); subtracting that amount from an empirical-Bernstein (Maurer & Pontil, 2009; Audibert et al., 2009) lower confidence bound on the clipped mean gives a certified $\hat{I}$, and the resulting mutual-information budget $B_{\mathrm{MI}} = \bar{R} - \hat{I}$ is licensed at level $1 - \delta_B$ (Appendix C.5). This is what moves the protocol off its abstention.

5.2 THE CERTIFY-OR-MEASURE PROTOCOL

Protocol: Certify-or-Measure

Inputs: an estimate $\hat{\mathcal{A}}_q$ of $\mathcal{A}_q$ with a $(1-\delta_q)$ confidence interval C_q; a budget B licensed at level $1-\delta_B$ ($\delta_B = 0$ for a logged $\bar{R}$); a pre-declared width threshold $w^* \in (0,1)$, failure probability δ, and draw count n. **(1)** Form the band $U = \bigcup_{a\in C_q}\{b : d_{\mathrm{bin}}(a\|b) \leq B\}$, an interval (Theorem 2b, Lemma 5). **(2)** If its width is $\leq w^*$ (for an exact $\mathcal{A}_q$, equivalently $B \leq B^*(\mathcal{A}_q)$, the admission budget of Section 5), **certify**: report U, which contains $\mathcal{A}_{\mathrm{prior}}$ with probability $\geq 1 - \delta_B - \delta_q$. **(3)** Else **abstain and measure**: draw $z_1, \ldots, z_n \sim p$, read both codes, report $\hat{\mathcal{A}} \pm \sqrt{\ln(2/\delta)/(2n)}$ (Dvoretzky et al., 1956), at one decoder pass per draw.

The protocol is sound in both branches: the certify branch by Theorem 2(b) and a union bound over the two Monte-Carlo failure events (budget and agreement estimate, both on the audited split; Appendix C.5), the measure branch by DKW. Its abstention region is a theorem, not an inconvenience. Under the rate budget, Theorem 3(d) *forces* abstention for informative codebooks. The tighter budget $D_{\mathrm{KL}}(\bar{q}\|p)$ carries no such floor, since a model whose aggregate posterior matches its prior certifies legitimately at any informativeness, so the protocol prefers it when it can be estimated. Coordinate inflation of a mean-field budget (Theorem 3a,c) and finite-sample resolution floors only widen bands, so certify decisions remain sound.

6 EXPERIMENTS

Section 6.1 shows the drift that Theorem 2(b) bounds; Sections 6.2 and 6.3 test the theory's predictions on trained models. Theorem 1 is an exact construction and is verified by script rather than by training. Uncertainty is mean $\pm$ standard error over seeds or folds with n stated, and every number is transcribed from the per-run files. Appendix D gives the full audit record, including the precommitted readings and every disclosed departure from them.

6.1 AGREEMENT DRIFTS WHEN A TRAINED VAE GENERATES

Figure 1 (bottom) sketches agreement that holds on encoded data and drifts when the model generates; a trained VAE shows this drift. We audit five convolutional VAEs trained on MNIST ($d = 10$, $K = 10$, one per seed). c_{enc} is a diagonal Gaussian mixture on encoder means and c_{dec} is the Bernoulli–Bregman nearest-prototype rule on decoded outputs, both in their own indices, read on 10,000 test images with $S = 16$ posterior and 20,000 prior draws. Agreement survives posterior sampling but degrades under the prior: $\mathcal{A}_q = 0.770 \pm 0.009$ versus $\mathcal{A}_{\mathrm{prior}} = 0.525 \pm 0.007$ ($n = 5$).

A task probe drops in step, from 0.700 ± 0.029 to 0.390 ± 0.019, and a label-free prototype control preserves the drop ($0.696 \pm 0.029 \to 0.375 \pm 0.027$), so the drift is not an artifact of the label map. The same drift appears in probabilistic PCA, where every quantity is exact and no estimator can be blamed: $\mathcal{A}[\bar{q}] - \mathcal{A}_{\text{prior}} = +0.128$ to $+0.338$ across three code maps at $d = 10$.

6.2 Certificates fire where the theory says they can

The reconstruction side certifies trained encoders at useful agreement. Since $\min_a B^*(a) = B^*(0.5) = 0.1438$, any budget up to 0.1438 certifies at every agreement. A trained mean-field MLP encoder on probabilistic PCA ($d = 10$, canonical basis) has $\bar{\Delta} = 0.088$ nats and certifies unconditionally. Swept over latent dimension, the same certificate fires in $75/75$ cells at $d \in \{2, 5, 10, 20, 32\}$ and agreement 0.76–0.97, against $0/75$ for a trained *linear* encoder, with widths growing $0.106 \to 0.487$ and crossing w^* just above $d = 32$, and every band contains the agreement under the exact posterior (Figure 2b, Appendix D.5). This side needs a certified *upper* bound on $\log p(x)$, which tractable models provide.

On the rate side, licensing decides. Lemma 2 predicts that unlicensed plug-in budgets fail, and they do: a grid plug-in certifies $20/20$ checkpoints it has no right to, and a moment-matched plug-in excludes the truth in $30/30$ conditions. The licensed mutual-information budget of Section 5.1 is sharply tighter than the best previously available licensed budget (Kolchinsky & Tracey, 2017): on six SVHN models, $B_{\text{MI}} = 0.519$–0.606 nats against 1.467–1.555. By Lemma 1 the previous budget clears the necessary condition on $0/6$ models: it admits no band narrower than 0.769, whatever the agreement and however the estimator is refined. Ours lowers that floor to ≤ 0.455 and clears the necessary condition for half width on $6/6$. On one of the six models, SVHN $d = 32$, a nonlinear convolutional VAE, the rate-side certificate fires (Appendix D.1). The binary-KL form matters: at that model's budget and point-estimate agreement, Pinsker gives width 0.553 and Bretagnolle–Huber 0.679, both above w^*, where binary KL gives 0.487.

6.3 The price and the floor, measured

A change of axes moves the budget and not the quantity. On linear-Gaussian models, where Theorem 3 is exact, a fixed DCT rotation of the same model moves the mean-field budget from 0.000 to 0.474 nats and voids the reconstruction-side certificate for all three map constructions, while the estimands move by at most 8×10^{-4}. The $O(d)$-closed full-covariance family is unaffected, and a trained mean-field MLP encoder pays a coordinate price too, $0.088 \to 0.673$ nats (Table 2, Appendix D.6).

A change of code map moves the quantity and not the budget. This is the dual in Theorem 3(c). Across 240 trained VAEs, changing the declared map on the same checkpoint moves $\mathcal{A}_q$ by a median 0.565 over four declared maps ($239/240$ above 0.1), while the rate budget, a functional of $\bar{q}$ alone, stays where it was. Agreement should therefore always be reported with its map.

The dataset floor predicts which models cannot be certified. We trained 240 fresh VAEs on FashionMNIST, MNIST, SVHN and CIFAR-10, up to the full training split, with $d \in \{16, 32, 64\}$, $K \in \{10, 32\}$, KL weights β up to 64 and three seeds, and fixed every reading beforehand (Appendix D.4). Theorem 3(e) forbids certification once $\bar{R} > \log N + \log 2$. That inequality holds for 66 of the 240 models, and none of them certifies (Figure 2c). Which side of the line a family falls on is not a matter of latent dimension: on FashionMNIST $\bar{R} - \log N$ reaches $+20.0$ nats and 36 models are excluded outright, while on SVHN it is negative throughout, -9.7 to -2.9, and none is. The floor is also attained, not merely respected: on six further FashionMNIST models (Appendix D.3), the measured $\hat{I}$ sits within 0.01 nats of $\log N$, so $D_{\text{KL}}(\bar{q} \| p) = \bar{R} - \log N$ to that accuracy.

7 Related work

Mismatched decoding (Scarlett et al., 2020) and noisy-channel quantization (Farvardin, 1990) name this failure mode for external channels; we move it inside one VAE, between its encoder side and its decoder side (VQ-VAE, van den Oord et al., 2017, is the constrained endpoint). Standard diagnostics all sit inside Definition 1: the ELBO and IWAE (Kingma & Welling, 2014; Rezende et al., 2014; Burda et al., 2016), rate–distortion decompositions (Alemi et al., 2018), disentanglement scores

(Higgins et al., 2017; Chen et al., 2018), information-theoretic summaries (Tishby et al., 2000; Cover & Thomas, 2006; Polyanskiy & Wu, 2025), and the mismatch between prior and aggregate posterior (Hoffman & Johnson, 2016; Tomczak & Welling, 2018; Bauer & Mnih, 2019). Some of these couple encoder and decoder, but only at the continuous level, so by Theorem 1 none determines agreement between the codes. Identifiability theory (Hyvärinen et al., 2019; Khemakhem et al., 2020; Locatello et al., 2019) states a related indeterminacy asymptotically, and Theorem 1(II) is its finite-population complement. Representation alignment (Moschella et al., 2023; Fumero et al., 2024; 2026; Huh et al., 2024) compares different models, whereas agreement is measured inside one model under its declared code maps. A prior preprint (Hayashi, 2026) already contains the coupled channel, Theorem 2(a) with its decomposition, and a narrower impossibility result over code histograms, their entropies, occupied codes and $I(x; z)$, and notes that a matched score can hide a codebook read entirely out of place. Definition 1 closes that case; the generative side, and with it the two-sided certificate, the price, the admission budget, the protocol and all experiments are new. Training-time consistency fixes (Cemgil et al., 2020; Wu et al., 2020) and detection of collapse on the marginal side (Dang et al., 2024) neither certify nor measure the coupled channel. Decoder-side map geometry follows Banerjee et al. (2005); Nielsen et al. (2010), and certify-or-abstain, standard in conformal and selective prediction (Vovk et al., 2005; Geifman & El-Yaniv, 2017), appears here with an exactly characterized abstention region.

8 Conclusion

A VAE can reconstruct every input perfectly, match its twin on every item of a standard report, and still read none of its codes back as written. The ELBO's own terms, read as budgets, bound how far a measured agreement can move, at a price and a floor the theory states in advance, and trained models behave as predicted. The reconstruction side needs a certified upper bound on $\log p(x)$, so outside tractable models the protocol runs on the rate side, where certifying a working codebook calls for a prior close to the aggregate posterior (Appendix D.6). The practical rule is short: log the joint code table in the pipeline's own indices, report agreement with its code map and a mean-field budget with its axes, and use Certify-or-Measure to decide whether to certify or to measure.

Reproducibility statement

All theorem-level claims are accompanied by numerical verification scripts with saved logs (grid witness at $K \in \{2, 3, 5, 8\}$; interleaved witness at all $(K, m) \in \{2, 3\} \times \{3, 4, 5\}$ with a radius-robustness check; closed-form-versus-numeric optimality, exact supremum, degenerate spectra, width-lemma bounds, and transport invariance for Theorem 3). The replication grid and the protocol run were preregistered (model sets, readings, and verdict criteria fixed before execution; the readings are reproduced verbatim in Figure 4); the three places where execution departed from the precommit are disclosed in Appendix D, with the sensitivity of every affected conclusion measured rather than asserted. Per-run CSVs, precommit documents, deviation disclosures, and the analysis ledger are included in the supplement, and every number in this paper is transcribed from those primary files.

Ethics statement

This work is a theoretical and empirical audit of representation consistency in generative models trained on standard public benchmarks (MNIST, FashionMNIST, SVHN, CIFAR-10); it involves no human subjects, no personal data, and no foreseeable direct dual-use concern. Its intended effect is to make silent failures of discrete latent interfaces measurable and disclosed rather than assumed away, which we view as ethically positive for downstream deployments that read codes from generative models.

AI use statement

Large language models were used as writing, coding and checking assistants throughout: drafting and revising prose; writing, refactoring and running analysis and verification scripts; maintaining the results ledgers; and adversarially checking derivations and stated numbers against the primary result files (one such check found the sign error in the clip correction disclosed in Appendix D). All

theorems, proofs, experimental designs, and conclusions were specified, verified, and approved by the authors; every number in the paper is transcribed from the primary logs, the verification script included in the supplement automatically re-checks most of them against those files, and the authors take full responsibility for the content.

REFERENCES

Alexander A. Alemi et al. Fixing a broken ELBO. In *International Conference on Machine Learning (ICML)*, 2018.

Jean-Yves Audibert, Rémi Munos, and Csaba Szepesvári. Exploration-exploitation tradeoff using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410(19):1876–1902, 2009.

Arindam Banerjee, Srujana Merugu, Inderjit S. Dhillon, and Joydeep Ghosh. Clustering with Bregman divergences. *Journal of Machine Learning Research*, 6:1705–1749, 2005.

Matthias Bauer and Andriy Mnih. Resampled priors for variational autoencoders. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2019.

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

Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

Yuri Burda, Roger Grosse, and Ruslan Salakhutdinov. Importance weighted autoencoders. In *International Conference on Learning Representations (ICLR)*, 2016.

A. Taylan Cemgil, Sumedh Ghaisas, Krishnamurthy Dvijotham, Sven Gowal, and Pushmeet Kohli. The autoencoding variational autoencoder. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Ricky T. Q. Chen, Xuechen Li, Roger B. Grosse, and David K. Duvenaud. Isolating sources of disentanglement in variational autoencoders. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley, 2nd edition, 2006.

Chris Cremer, Xuechen Li, and David Duvenaud. Inference suboptimality in variational autoencoders. In *International Conference on Machine Learning*, 2018.

H. Dang, T. Tran, T. Nguyen, and N. Ho. Beyond vanilla variational autoencoders: Detecting posterior collapse in conditional and hierarchical variational autoencoders. In *International Conference on Learning Representations (ICLR)*, 2024.

Aryeh Dvoretzky, Jack Kiefer, and Jacob Wolfowitz. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *Annals of Mathematical Statistics*, 27(3):642–669, 1956.

Nariman Farvardin. A study of vector quantization for noisy channels. *IEEE Transactions on Information Theory*, 36(4):799–809, 1990.

Marco Fumero, Marco Pegoraro, Valentino Maiorca, Francesco Locatello, and Emanuele Rodolà. Latent functional maps: A spectral framework for representation alignment. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Marco Fumero, Luca Moschella, Emanuele Rodolà, and Francesco Locatello. Navigating the latent space dynamics of neural models. *arXiv:2505.22785*, 2026.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, 2017.

Yusuke Hayashi. Lost and found in translation: Variational diagnostics for neural codebook channels. *arXiv:2605.18846*, 2026.

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

Irina Higgins et al. beta-VAE: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations (ICLR)*, 2017.

Matthew D. Hoffman and Matthew J. Johnson. ELBO surgery: yet another way to carve up the variational evidence lower bound. In *NeurIPS Workshop on Advances in Approximate Bayesian Inference*, 2016.

Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, 2nd edition, 2013.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. Position: The platonic representation hypothesis. In *International Conference on Machine Learning (ICML)*, 2024.

Aapo Hyvärinen, Hiroaki Sasaki, and Richard E. Turner. Nonlinear ICA using auxiliary variables and generalized contrastive learning. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2019.

Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvärinen. Variational autoencoders and nonlinear ICA: a unifying framework. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2020.

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

Artemy Kolchinsky and Brendan D. Tracey. Estimating mixture entropy with pairwise distances. *Entropy*, 19(7):361, 2017.

Gabriel Loaiza-Ganem and John P. Cunningham. The continuous Bernoulli: fixing a pervasive error in variational autoencoders. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Rätsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *International Conference on Machine Learning (ICML)*, 2019.

Albert W. Marshall, Ingram Olkin, and Barry C. Arnold. Inequalities: Theory of majorization and its applications. 2011.

Andreas Maurer and Massimiliano Pontil. Empirical Bernstein bounds and sample variance penalization. In *Conference on Learning Theory*, 2009.

Luca Moschella, Valentino Maiorca, Marco Fumero, Antonio Norelli, Francesco Locatello, and Emanuele Rodolà. Relative representations enable zero-shot latent space communication. In *International Conference on Learning Representations (ICLR)*, 2023.

Frank Nielsen, Jean-Daniel Boissonnat, and Richard Nock. Bregman Voronoi diagrams: properties, algorithms and applications. *Discrete & Computational Geometry*, 44(2):281–307, 2010.

Yury Polyanskiy and Yihong Wu. *Information Theory: From Coding to Learning*. Cambridge University Press, 2025.

Ben Poole, Sherjil Ozair, Aäron van den Oord, Alexander A. Alemi, and George Tucker. On variational bounds of mutual information. In *International Conference on Machine Learning*, 2019.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International Conference on Machine Learning (ICML)*, 2014.

Jonathan Scarlett, Albert Guillén i Fàbregas, Anelia Somekh-Baruch, and Alfonso Martinez. Information-theoretic foundations of mismatched decoding. *Foundations and Trends in Communications and Information Theory*, 17(2–3):149–401, 2020.

Claude E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27 (3):379–423, 1948. doi: 10.1002/j.1538-7305.1948.tb01338.x.

Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method. *arXiv:physics/0004057*, 2000.

Jakub M. Tomczak and Max Welling. VAE with a VampPrior. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2018.

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005.

Warren Weaver. Recent contributions to the mathematical theory of communication. In *The Mathematical Theory of Communication*. University of Illinois Press, Urbana, IL, 1949.

Chen Wu, Prince Zizhuang Wang, and William Yang Wang. On the encoder–decoder incompatibility in variational text modeling and beyond. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 3449–3464, 2020.

A Proofs for Theorem 1

A.1 The grid witness, in full

Fix $K \geq 2$ and a derangement σ of $[K]$. Latent space $\mathcal{Z} = [0,1)^2$ with uniform (Lebesgue) prior; cells $G_{ij} = [\frac{i-1}{K}, \frac{i}{K}) \times [\frac{j-1}{K}, \frac{j}{K})$, each of prior mass $1/K^2$; fixed maps $c_{\mathrm{enc}}(z) = \lceil Kz_1 \rceil$, $c_{\mathrm{dec}}(z) = \lceil Kz_2 \rceil$. Data $\mathcal{X} = \{v_1, \ldots, v_K\} \cup J$ with J a pool of $K^2 - K$ distinct junk atoms; $\mathcal{D} = \mathrm{Unif}\{v_1, \ldots, v_K\}$. Configuration C^1: the decoder emits v_i deterministically on G_{ii} and assigns the junk atoms bijectively to the off-diagonal cells; encoder $q^1(\cdot \mid v_i) = \mathrm{Unif}(G_{ii})$. Configuration C^2: the decoder emits v_i on $G_{i\sigma(i)}$ and the same junk pool bijectively on the complementary cells; $q^2(\cdot \mid v_i) = \mathrm{Unif}(G_{i\sigma(i)})$.

(m1)–(m3). For $x = v_i$ the posterior is uniform on the unique cell where the decoder emits v_i, so reconstruction $\mathbb{E}_q \log p(x \mid z) = \log 1 = 0$; rate $D_{\mathrm{KL}}(\mathrm{Unif}(\text{cell}) \| \mathrm{Unif}([0,1)^2)) = \log K^2$; ELBO $= -\log K^2$. The marginal likelihood is the prior mass of the emitting cell: $p(v_i) = 1/K^2$, so $\log p(v_i) = \mathrm{ELBO}(v_i)$ and the gap is 0: inference is exact ($q(\cdot \mid v_i)$ *is* the true posterior, uniform on the emitting cell). All values coincide across C^1, C^2. As a distribution, $p(x)$ is uniform over the same K^2 atoms (each cell emits one atom of mass $1/K^2$) in both configurations.

(m4). $\bar{q}$ has density K on the K occupied cells, so $h(\bar{q}) = -\log K$, while each $q(\cdot \mid v_i)$ is uniform on one cell of area K^{-2}, so $\mathbb{E}_x h(q(\cdot \mid x)) = -\log K^2$; hence $I(x; z) = h(\bar{q}) - \mathbb{E}_x h(q(\cdot \mid x)) = \log K$ in both. Posterior means are cell centers; coordinate 1 means form the multiset $\{(i - \frac{1}{2})/K\}$ in both; coordinate 2 means form $\{(i - \frac{1}{2})/K\}$ versus $\{(\sigma(i) - \frac{1}{2})/K\}$—equal as multisets since σ is a bijection—so each per-dimension AU_k coincides. (The *cross-covariance* of the mean pairs differs; that leak is what Part I′ removes.)

(m5)–(m7). The joint tables under $\bar{q}$ are $T^1 = \frac{1}{K}\mathrm{Id}$ and $T^2 = \frac{1}{K}P_\sigma$: both have uniform row and column sums (identical histograms, entropies, occupied counts) and are related by a column permutation, $T^2 = T^1 P_\sigma$, so *every* F with $F(P_1 T P_2^\top) = F(T)$ coincides on them—this is the exact reason (m6) cannot separate the pair, and it is stronger than equality of entry multisets, which alone would not suffice. In particular $I(c_{\mathrm{enc}}; c_{\mathrm{dec}}) = \log K$ in both and the permutation-matched agreement is 1 in both. Under the prior the code pair is uniform on $[K]^2$, so $\mathcal{A}_{\mathrm{prior}} = 1/K$ in both. Yet $\mathcal{A}(C^1) = \bar{q}^1(\bigcup_i G_{ii}) = 1$ and $\mathcal{A}(C^2) = 0$. □

Smoothing remark. Embed $\mathcal{X}$ mutually equidistantly (one-hot in $\mathbb{R}^{K^2}$ suffices) and replace deterministic emission by $p(x \mid z) = N(x; x_{\mathrm{cell}}(z), \varepsilon^2 I)$. By equidistance, the likelihood of any atom under any non-emitting cell is a shared constant, and the entire package equality is preserved exactly for every $\varepsilon > 0$, with three shared shifts: reconstruction becomes a common constant, ELBO and $\log p(x)$ shift with it, and the common gap becomes positive; the rate is unchanged. The equality is also exact under a generic (non-equidistant) embedding: the encoders are unchanged, and in both configurations every cell emits one of the same K^2 atoms with weight $1/K^2$, so $p(x) = K^{-2} \sum_{\text{cells}} N(x; x_{\mathrm{cell}}, \varepsilon^2 I)$ is the same distribution, reconstruction is the same constant $\log N(v_i; v_i, \varepsilon^2 I)$, and the rate, ELBO, $\log p(x)$ and gap coincide per example; (m4)–(m7) do not involve the emission. Equidistance only makes the shared values explicit.

A.2 The interleaved product witness

Overview (sketch of Part I′). The grid pair still leaks through one non-package statistic: per-example posterior *means* encode both codes. The repair takes $\mathcal{Z} = \mathbb{R}^2 \times \mathbb{R}^2$ with a product of uniform annuli: c_{enc} reads a coarse angular sector (K of them) of plane A, and c_{dec} an m-fold rotationally symmetric *interleaved class* of plane B. Classes are invariant under rotation by $2\pi/m$, so their moment tensors of degree $< m$ are isotropic and class-independent: the pairing σ is invisible to *all* aggregate moments of B-degree $< m$, and per-example means are identical outright. At degree exactly m an m-th harmonic carries the class index, and explicit cross moments separate the pair (numerically at $K = 3$, $m = 3$: the degree-3 moment $\mathbb{E}[z_{\mathsf{A},2} z_{\mathsf{B},1}^2 z_{\mathsf{B},2}]$ equals 0.1040 versus -0.0520)—as something must, since agreement is a $\bar{q}$-functional.

The engine is a harmonic-blindness lemma (Lemma 3, stated and proved next), a one-line consequence of the circular-harmonic expansion: rotation multiplies the k-th harmonic by $e^{ik\omega}$, and invariance under rotation by $2\pi/m$ kills every harmonic with $m \nmid k$, leaving only the rotation-invariant $k = 0$ term below degree m. Part (II) is proved in Appendix A.5.

Lemma 3 (Harmonic blindness)**.** *Let μ_0 be a distribution on $\mathbb{R}^2$ invariant under rotation by $2\pi/m$, and $\mu_j = (R_{\theta_j})_{\#}\mu_0$ its rotations. Then all moment tensors of μ_j of total degree $t < m$ are independent of j; if $\mathbb{E}_{\mu_0}[r^m e^{im\varphi}] \neq 0$, the degree-$m$ tensors depend on j through the phase $e^{im\theta_j}$.*

Proof of Lemma 3. In polar form a degree-t monomial moment is a linear combination of circular harmonics $\mathbb{E}[r^t e^{ik\varphi}]$ with $|k| \leq t$ and $k \equiv t \pmod 2$. Rotation by ω multiplies the k-th harmonic by $e^{ik\omega}$; invariance of μ_0 under $\omega = 2\pi/m$ forces $\mathbb{E}_{\mu_0}[r^t e^{ik\varphi}] = 0$ unless $m \mid k$. For $t < m$ only $k = 0$ survives, and $k = 0$ harmonics are rotation-invariant, hence equal for all μ_j. At $t = m$ the $k = \pm m$ harmonics survive and transform by $e^{\pm im\theta_j}$. □

Construction. Take $\mathcal{Z} = \mathbb{R}^2 \times \mathbb{R}^2$ with a product of uniform annuli. c_{enc} reads the coarse angular sector (K of them) of plane A. On plane B, partition into Km fine sectors and let class j be the union of the m sectors congruent to j mod K—an m-fold rotationally symmetric set of zero mean; c_{dec} reads the class. The two configurations couple sectors to classes by the identity and by the derangement σ, exactly as the grid couples rows to columns. The package items repeat the grid computation verbatim, with joint cells $O_{ij} = \mathrm{sector}_{\mathsf{A}}(i) \times \mathrm{class}_{\mathsf{B}}(j)$ (prior mass $1/K^2$ each; the cells are congruent, related by rotations of the two planes, so rates and masses match). For the new items:

(N1). $\mathbb{E}[z \mid v_i] = (\mu_{\mathsf{A}}(i), \mu_{\mathsf{B}}(\mathrm{class}))$ where μ_{B} of any class vanishes: a class is invariant under rotation by $2\pi/m$ ($m \geq 2$), and the only vector fixed by a nontrivial rotation is 0. Hence $\mathbb{E}[z \mid v_i] = (\mu_{\mathsf{A}}(i), 0, 0)$ in both configurations, example by example.

(N2). A mixed moment factorizes over the product cell: $\mathbb{E}_{\bar{q}}[z_{\mathsf{A}}^a z_{\mathsf{B}}^b] = \frac{1}{K} \sum_i \mathbb{E}[z_{\mathsf{A}}^a \mid \mathrm{sector}_{\mathsf{A}}(i)] \cdot \mathbb{E}[z_{\mathsf{B}}^b \mid \mathrm{class}(c(i))]$ with $c = \mathrm{id}$ or σ. By Lemma 3, for $|b| < m$ the class-conditional factor is class-independent, so the sum is identical for both couplings.

(N3). At $|b| = m$ the $k = \pm m$ harmonic of the class survives and carries the class phase $e^{im\theta_j}$ with $\theta_j = 2\pi j/(Km)$, i.e. the K-th root of unity $e^{2\pi ij/K}$: the factor now depends on j, and pairing it with any class-distinguishing A-factor produces an explicit separating moment; the verification log exhibits one for every (K, m) in the grid. Since $\mathcal{A}$ is a $\bar{q}$-functional, *some* statistic had to separate the pair; the construction pins the minimal polynomial order at exactly m. □

A.3 Scope of Theorem 1

The theorem does not claim the following. (i) Part I matches the full package at uniform occupied masses; the full-package question at non-uniform masses is open—Proposition 1 gives the sharp *channel-level* statement there, and when $\max_i \nu_i > \frac{1}{2}$ zero agreement is infeasible ($\mathcal{A} \geq 2\max_i \nu_i - 1$), leaving strict separation only. (ii) The maps are fixed; identifiability for maps refit per configuration is outside the theorem, and the witnesses are constructions— no genericity is claimed: we exhibit constructions, not a sampling argument about trained models. (iii) In Proposition 2 the $K = 3$ witness matches (agreement, marginals, entropies, MI) but is separated by, e.g., its maximal entry (0.2333 vs. 0.3070) and Frobenius norm (0.4397 vs. 0.4515): the claim is that the scalar summaries the literature reports do not determine structure, not that no table functional does—the table's placement does, which is why reporting T is the recommendation. (iv) All constructions are verified numerically to machine precision (grid witness $K \in \{2, 3, 5, 8\}$; interleaved witness for all $(K, m) \in \{2, 3\} \times \{3, 4, 5\}$ plus a radius-robustness check); scripts and logs are in the supplement.

A.4 Non-uniform marginals with matched mutual information

Proposition 1 (Non-uniform marginals)**.** *For every probability vector ν with at least two positive entries there exist coupled channels with index-wise identical encoder and decoder marginals ν and exactly matched $I(c_{\mathrm{enc}}; c_{\mathrm{dec}})$ whose agreements differ; when $\max_i \nu_i \leq \frac{1}{2}$ the pair can be taken with $\mathcal{A} = 0$ versus $\mathcal{A} > \sum_i \nu_i^2$. The feasible set of $\mathcal{A}$ over all couplings with both marginals ν is exactly $[\max(0,\ 2\max_i \nu_i - 1),\ 1]$.*

Let $\nu = (\nu_1, \ldots, \nu_K)$ have at least two positive entries. Define the independence coupling $\Pi = \nu\nu^\top$ ($\mathcal{A}(\Pi) = \sum_i \nu_i^2 =: A_0$), the identity coupling $P_{\mathrm{high}} = \mathrm{diag}(\nu)$ ($\mathcal{A} = 1$), and any coupling P_{low} with both marginals ν attaining the minimal diagonal mass $A_{\min} = \max(0, 2\max_i \nu_i - 1)$. Existence of P_{low}: for every coupling, $P(i,i) \geq \nu_i - \sum_{j \neq i} P(i,j) \geq 2\nu_i - 1$ and at most one index can exceed $\frac{1}{2}$, so $\mathcal{A} \geq A_{\min}$ always; attainability for $\max_i \nu_i \leq \frac{1}{2}$ is Hall/max-flow feasibility of a zero-diagonal doubly-ν table (feasible iff $\nu_i \leq 1 - \nu_i$ for all i), and for $\max_i \nu_i > \frac{1}{2}$ by an explicit greedy construction. This proves the exact feasible interval $[\max(0, 2\max_i \nu_i - 1), 1]$ claimed in Proposition 1.

Consider the segments $P_t = (1-t)\Pi + tP_{\mathrm{high}}$ and $\bar{P}_s = (1-s)\Pi + sP_{\mathrm{low}}$. (i) Every $P_t, \bar{P}_s$ has both marginals index-wise equal to ν. (ii) With marginals fixed, $I = 2H(\nu) - H(\text{joint})$; joint entropy is strictly concave and each path is affine and non-constant, so I is convex, continuous, vanishes only at Π, and is strictly increasing along each segment, ranging over $[0, H(\nu)]$ on $\{P_t\}$ and $[0, I(P_{\mathrm{low}})]$ on $\{\bar{P}_s\}$ with $I(P_{\mathrm{low}}) > 0$. (iii) Hence for every $m \in (0, I(P_{\mathrm{low}})]$ there are unique t^*, s^* with $I(P_{t^*}) = I(\bar{P}_{s^*}) = m$, and since $\mathcal{A}$ is affine along each path with slopes $1 - A_0 > 0$ and $A_{\min} - A_0 < 0$, $\mathcal{A}(P_{t^*}) > A_0 > \mathcal{A}(\bar{P}_{s^*})$, with gap $t^*(1 - A_0) + s^*(A_0 - A_{\min}) > 0$. When $\max_i \nu_i \leq \frac{1}{2}$, taking P_{low} with zero diagonal and $m = I(P_{\mathrm{low}})$ gives the zero-versus-positive form with $\mathcal{A}(P_{t^*}) > A_0 = \sum_i \nu_i^2$. □

Scope discipline. For $m > I(P_{\mathrm{low}})$ no claim is made: for $K = 2$ and skewed ν, marginals plus MI can *determine* $\mathcal{A}$ uniquely on part of the range, so the matched-MI window is genuinely $(0, I(P_{\mathrm{low}})]$; the window is quantitatively small for very skewed ν—that is the coupling floor at work, and it is consistent with, not contrary to, the certification narrative. The pair below settles whether the marginals can be made to match *exactly*, and not only after relabeling, at exactly equal mutual information, and settles it with an explicit table rather than an existence claim.

An explicit instance: marginals matched exactly, no relabeling. Take $K = 3$ and $\nu = (\frac{1}{2}, 0.3, 0.2)$. Row 1 carries mass $\frac{1}{2}$ while columns $2, 3$ have total capacity $\frac{1}{2}$, so the zero-diagonal coupling with both marginals ν is *unique*; solving $I(P_t) = I(P_{\mathrm{low}})$ along $P_t = (1-t)\nu\nu^\top + t\,\mathrm{diag}(\nu)$ gives $t^* = 0.866779$:

$$P_{\mathrm{low}} = \begin{pmatrix} 0 & 0.3 & 0.2 \\ 0.3 & 0 & 0 \\ 0.2 & 0 & 0 \end{pmatrix}, \qquad P_{t^*} = \begin{pmatrix} 0.466695 & 0.019983 & 0.013322 \\ 0.019983 & 0.272024 & 0.007993 \\ 0.013322 & 0.007993 & 0.178685 \end{pmatrix}.$$

Both tables have encoder *and* decoder marginals index-wise equal to ν, and both have mutual information exactly $\log 2 = 0.693147$ nats: whenever $\nu_1 = \frac{1}{2}$ the off-diagonal terms cancel in

$I = 2H(\nu) - H(P)$, leaving $I(P_{\text{low}}) = \log 2$ in closed form. Yet $\mathcal{A}(P_{\text{low}}) = 0$ exactly while $\mathcal{A}(P_{t^*}) = 0.917403$, against $A_0 = \sum_i \nu_i^2 = 0.38$ for the independent coupling. The two tables are not relabelings of one another—their entry multisets differ—so here the marginals match exactly, and not only up to relabeling. This is a channel-level statement about two joint tables, which is the level at which the question was posed; the full-package version at non-uniform masses remains open (Appendix A.3).

A.5 Orbit form

Full statement of Theorem 1(II). Let τ be a prior-preserving bimeasurable bijection of $\mathcal{Z}$ and $C_\tau = (\mathcal{D}, \tau_{\#}^{-1}q, p(x \mid \tau\,\cdot), c_{\text{enc}}, c_{\text{dec}})$. Then (m1)–(m3) and $I(x;z)$ are exactly invariant for every x; the per-dimension AU is additionally invariant whenever τ^{-1} restricted to $\text{supp}(\bar{q})$ is a translation. Agreement transforms as $\mathcal{A}(C_\tau) = \bar{q}(\tau(\text{Agr}))$ and changes iff τ effects a net transport of $\bar{q}$-mass across the disagreement boundary.

Proof. For each x the pair (encoder pushforward, decoder pullback) is a change of variables: reconstruction integrands compose to the original, D_{KL} is invariant under bimeasurable bijections applied to both arguments (prior preserved), so (m1)–(m3) are pointwise invariant, and $I(x;z)$ is invariant as a functional of the joint law. The aggregate posterior transforms as $\bar{q}_\tau = \tau_{\#}^{-1}\bar{q}$, so

$$\mathcal{A}(C_\tau) = \bar{q}_\tau(\text{Agr}) = \bar{q}(\tau(\text{Agr})),$$

which differs from $\mathcal{A} = \bar{q}(\text{Agr})$ iff τ transports net $\bar{q}$-mass across the boundary of Agr. (The orientation matters: writing $\tau^{-1}(\text{Agr})$ here is wrong, though undetectable when τ is an involution; the verification suite checks a non-involutive quarter-shift.) For AU: the latent of C_τ is $\tau^{-1}(z)$ with $z \sim q(\cdot \mid x)$, so if $\tau^{-1}|_{\text{supp}(\bar{q})}$ is a translation, posterior means shift by a constant and every $\text{Var}_x\mathbb{E}[z_k \mid x]$ is unchanged. Without the hypothesis AU can move: on the uniform prior on $[0,1)$, swapping $[0, \frac{1}{4})$ with $[\frac{1}{4}, \frac{1}{2})$ is prior-preserving and moves $\text{Var}_x\mathbb{E}[z \mid x]$ from 0.0625 to 0.015625 in a two-example configuration for which all other items are invariant. □

A.6 Structure beyond agreement

Proposition 2 (Structure)**.** *For every $K \geq 3$ there exist coupled channels with identical agreement, identical uniform marginals on both sides, and exactly matched mutual information (hence matched joint, conditional, and marginal entropies) whose error structures differ qualitatively (cyclic/bijective versus many-to-one); for $K \geq 4$ a derangement family matches* all *permutation-invariant scalars while pairwise and cyclic confusion structure still differ.*

For $K = 3$, the two tables (row $= c_{\text{enc}}$, column $= c_{\text{dec}}$)

$$P_{\text{cyc}} = \begin{pmatrix} 7/30 & 1/10 & 0 \\ 0 & 7/30 & 1/10 \\ 1/10 & 0 & 7/30 \end{pmatrix}, \qquad P_{\text{mrg}} = \begin{pmatrix} 0.183333 & 0.026293 & 0.123707 \\ 0.026293 & 0.307040 & 0 \\ 0.123707 & 0 & 0.209627 \end{pmatrix}$$

have uniform marginals on both sides, equal trace (agreement), and mutual information matched to 5×10^{-6} (the second table's off-diagonal entries are tuned to match MI exactly in the limit), yet their error structures differ qualitatively: cyclic single-successor confusion versus asymmetric merge into code 2. They are separated by, e.g., the maximal entry (0.2333 vs 0.3070) and the Frobenius norm (0.4397 vs 0.4515)—consistent with the scope of Proposition 2: what is matched is (agreement, marginals, entropies, MI), i.e. the scalar summaries commonly reported. For $K \geq 4$, take two derangements of different cycle type (e.g. a product of transpositions versus a single K-cycle): the tables $\frac{1}{K}P_\sigma$, $\frac{1}{K}P_{\sigma'}$ satisfy $P_{\sigma'} = P_\sigma\, P_{\sigma^{-1}\sigma'}$, i.e. they differ by a column permutation, hence match *all* permutation-invariant scalars, while pairwise versus cyclic confusion structure differs. □

B Proof of Theorem 2

(a). Write $\Delta(x) = D_{\text{KL}}(q(\cdot \mid x) \| p_\theta(\cdot \mid x))$; the ELBO/evidence identity gives $\bar{\Delta} = \mathbb{E}_x\Delta(x)$. For measurable $f : \mathcal{Z} \to \mathcal{Y}$ ($\mathcal{Y}$ standard Borel), the KL chain rule under disintegration gives, per example,

$$\Delta(x) = D_{\text{KL}}\big(f_{\#}q(\cdot \mid x) \,\|\, f_{\#}p_\theta(\cdot \mid x)\big) + \rho(x; f), \qquad \rho(x; f) \geq 0,$$

where ρ integrates the fiber-conditional KLs. Averaging over x and applying joint convexity of KL to the mixtures $\bar{q}_f = \mathbb{E}_x f_\# q(\cdot \mid x)$, $\bar{p}_f = \mathbb{E}_x f_\# p_\theta(\cdot \mid x)$ yields $\bar{\Delta} = D_{\mathrm{KL}}(\bar{q}_f \| \bar{p}_f) + \mathcal{J}_f + \bar{\rho}_f$ with $\mathcal{J}_f, \bar{\rho}_f \geq 0$ the Jensen and fiber residuals. With $f = \mathbf{1}_E$ the code-level term is the binary KL $d_{\mathrm{bin}}(\bar{q}(E) \| \bar{p}_\theta(E))$, giving the display and the three-term decomposition; equality conditions (f sufficient; mixture degeneracy) follow from the two inequalities used.

(b). The data-processing inequality for the map $z \mapsto \mathbf{1}_{\mathrm{Agr}}(z)$ gives $d_{\mathrm{bin}}(\bar{q}(\mathrm{Agr}) \| p(\mathrm{Agr})) \leq D_{\mathrm{KL}}(\bar{q} \| p)$. The second inequality is the standard rate decomposition $\bar{R} = D_{\mathrm{KL}}(\bar{q} \| p) + I(x; z)$ (Hoffman & Johnson, 2016) with $I(x; z) \geq 0$. □

C Proofs for Theorem 3

Proof of Lemma 1. Feasible-point constructions. For the upper endpoint, $x_0 = 1 - (1-a)e^{-B} \geq a$ gives $d_{\mathrm{bin}}(a \| x_0) = a \log \frac{a}{x_0} + (1-a)B \leq B$, so $x_+ \geq 1 - (1-a)e^{-B}$. For the lower endpoint, $x_1 = ae^{-B/a} \leq a$ gives $d_{\mathrm{bin}}(a \| x_1) = B + (1-a) \log \frac{1-a}{1-x_1} \leq B$, so $x_- \leq ae^{-B/a}$. Hence $W \geq (1 - e^{-B}) + a(e^{-B} - e^{-B/a}) \geq 1 - e^{-B}$ since $B/a \geq B$; the added term is strictly positive for $a \in (0, 1)$. At $a = 0$, $d_{\mathrm{bin}}(0 \| x) = -\log(1-x)$, so $x_+ = 1 - e^{-B}$ and $x_- = 0$ exactly. □

C.1 (a): exact price, zero set, supremum

With $q = N(\mu, Q^\top D Q)$ and posterior $N(m, \Sigma)$, $\Lambda = \Sigma^{-1}$:

$$D_{\mathrm{KL}}(q \| p(\cdot \mid x)) = \tfrac{1}{2}\Big[\mathrm{tr}(\Lambda_Q D) + (\mu - m)^\top \Lambda (\mu - m) - d - \textstyle\sum_i \log D_{ii} - \log\det \Lambda_Q\Big]$$

(using $\det \Lambda_Q = \det \Lambda = 1/\det \Sigma$ for orthogonal Q, so the last term is the usual $+\log\det\Sigma$); the mean term is minimized at $\mu^* = m$, and the trace and log-det terms decouple over i: $\frac{1}{2}[(\Lambda_Q)_{ii} D_{ii} - \log D_{ii}]$ is strictly convex in $\log D_{ii}$ with unique minimum $D^*_{ii} = 1/(\Lambda_Q)_{ii}$. Substituting gives $B_Q = \frac{1}{2}[\log\det \mathrm{diag}(\Lambda_Q) - \log\det\Lambda]$. Hadamard's inequality (Horn & Johnson, 2013, Thm. 7.8.1) gives $\det\mathrm{diag}(M) \geq \det M$ for $M \succ 0$ with equality iff M is diagonal, so $B_Q \geq 0$ with equality iff Λ_Q is diagonal. For the zero set, write $\Lambda = U \Lambda_{\mathrm{d}} U^\top$ and $\Xi = QU$: $\Lambda_Q = \Xi \Lambda_{\mathrm{d}} \Xi^\top$ is diagonal iff Ξ maps each eigen-coordinate-subspace of Λ_{d} onto a coordinate subspace, i.e. $\Xi = PV$ with P a permutation matrix and $V \in \mathcal{O}_{\mathrm{eig}}$ a within-eigenspace orthogonal map (diagonal signs when the spectrum is simple); hence $Q \in \{\text{signed permutations}\} \cdot \mathcal{O}_{\mathrm{eig}} \cdot U^\top$ exactly. For the supremum: by Schur–Horn, $\{\mathrm{diag}(Q \Lambda Q^\top) : Q \in O(d)\} = \{u : u \prec \lambda\}$; $\prod_i u_i$ is Schur-concave, and the uniform vector $\bar{\lambda}\mathbf{1}$ (with $\bar{\lambda}$ the arithmetic mean) is majorized by every u of equal trace, so $\max \prod u_i = \bar{\lambda}^d$, attained; hence $\sup_Q B_Q = \frac{d}{2}\log(\mathrm{AM}(\lambda)/\mathrm{GM}(\lambda))$. The Kantorovich corollary follows from $\mathrm{GM}(\lambda)^{-1} = \mathrm{GM}(1/\lambda) \leq \mathrm{AM}(1/\lambda)$ and the scalar Kantorovich bound $\mathrm{AM}(\lambda)\,\mathrm{AM}(1/\lambda) \leq (\kappa+1)^2/(4\kappa)$ for spectra in $[\lambda_{\min}, \lambda_{\max}]$. (iii) For any product measure $q = \otimes_i q_i$ in basis Q, $-\mathbb{E}_q \log p(\cdot \mid x)$ depends on q only through per-coordinate first and second moments (the Gaussian log-density is quadratic and cross terms factorize under a product), and at fixed variance the entropy term is maximized by the Gaussian; hence the Gaussian product with matched moments has no larger KL, and restricting to Gaussians is without loss. (iv) $\det\mathrm{diag}(\Lambda_Q)/\det\Lambda_Q = 1/\det\mathrm{Corr}(\Lambda_Q)$ is invariant under diagonal rescaling, and means canceled at $\mu^* = m$. □

C.2 (b): transport invariance

For any sampling law μ_z on the original space, the transported configuration samples $R_\# \mu_z$ and reads transported maps $c'(w) = c(R^\top w)$: the code pair is $(c_{\mathrm{enc}}(R^\top R z), c_{\mathrm{dec}}(R^\top R z)) = (c_{\mathrm{enc}}(z), c_{\mathrm{dec}}(z))$ in law, so every functional of code-pair laws under model-defined sampling laws (posterior, aggregate posterior, prior) is invariant. Decoder-side *constructed* maps transport coherently: with $g'(w) = g(R^\top w)$, decoded prototypes transport as $R\mu_k$ and the Bregman divergence matrix is unchanged. For *refit* maps, the class-mean and trace-regularized Mahalanobis fitting operators commute with the $O(d)$ action in exact arithmetic (the regularizer is conjugation-invariant), giving exact equivariance; k-means refitting is equivariant in distribution only (initialization and tie-breaking are basis-dependent), and padded embeddings are not $O(d)$ actions and require a separate zero-block argument, so both are scoped out of the theorem and handled empirically. □

C.3 Scope notes for Theorem 3(b)

Three scope notes are load-bearing. (b) speaks of *transported* maps: under *refitting*, the class-mean and regularized-Mahalanobis constructions stay exactly equivariant while k-means is equivariant only in distribution (stress-tested, not proved). $\mathcal{A}_q$ under a *mean-field* posterior is deliberately excluded, since the transported mean-field family is no longer mean-field—that is (a)'s content. And *padding* embeddings are not $O(d)$ actions; the padded stress case is labelled accordingly.

C.4 Proof of Lemma 2

The first inequality is the data-processing inequality for KL under the measurable map f. For the second: since $\tilde{q}$ matches the first two moments of $\bar{q}$ and $\log p$ is quadratic when p is Gaussian, $\mathbb{E}_{\tilde{q}}[\log p] = \mathbb{E}_{\bar{q}}[\log p]$, while the differential entropy satisfies $h(\tilde{q}) \geq h(\bar{q})$ by maximum entropy at fixed second moments; writing $D_{\mathrm{KL}}(\cdot\|p) = -h(\cdot) - \mathbb{E}_{\cdot}[\log p]$ gives $D_{\mathrm{KL}}(\tilde{q}\|p) \leq D_{\mathrm{KL}}(\bar{q}\|p)$. The last two clauses are the rate decomposition $\bar{R} = D_{\mathrm{KL}}(\bar{q}\|p) + I(x;z)$ (Hoffman & Johnson, 2016): $\bar{R}$ is an upper bound loose by exactly $I(x;z)$, and $\bar{R} - \hat{I} \geq D_{\mathrm{KL}}(\bar{q}\|p)$ whenever $\hat{I} \leq I(x;z)$. □

C.5 Two finite-sample lemmas behind the licensed MI budget

Throughout the audit the reference distribution is the test split: $\bar{q}$ is the finite mixture $\frac{1}{N}\sum_{i=1}^{N} q_i$ over it (the mixture whose density the MI budget evaluates exactly), $\bar{R}$ its mean rate, and $\mathcal{A}_q = \frac{1}{N}\sum_i q_i(\mathrm{Agr})$; the only randomness is Monte Carlo. Theorem 2(b) holds for any reference distribution provided $\mathcal{A}_q$, $\bar{q}$ and $\bar{R}$ are taken under the same one, and the certified quantity $\mathcal{A}_{\mathrm{prior}} = p(\mathrm{Agr})$ does not depend on it.

Lemma 4 (Clip tail)**.** *Let $q_1, \ldots, q_N$ be probability densities on $\mathcal{Z}$, $\bar{q} = \frac{1}{N}\sum_i q_i$, and $T = \log q_i(z) - \log \bar{q}(z)$ with $i \sim \mathrm{Unif}[N]$, $z \sim q_i$. Then $T \leq \log N$ almost surely, $\mathbb{E}[T] = I(i;z)$, and for every $c > 0$ the clipped statistic $T_c = \max(T, -c)$ satisfies*

$$I(i;z) \;\leq\; \mathbb{E}[T_c] \;\leq\; I(i;z) + e^{-c}.$$

Proof. $\bar{q} \geq q_i/N$ pointwise gives $T \leq \log N$, and $\mathbb{E}[T] = I(i;z)$ is the mixture form of mutual information. Since $T_c = T + (-c - T)_+$, the lower bound is immediate. For the upper bound, $\mathbb{E}[e^{-T}] = \frac{1}{N}\sum_i \int_{\{q_i>0\}} \bar{q} \leq 1$, so Markov's inequality gives $\Pr(-T > u) = \Pr(e^{-T} > e^u) \leq e^{-u}$, and

$$\mathbb{E}[(-c-T)_+] = \int_0^\infty \Pr(-T > c+s)\, ds \;\leq\; \int_0^\infty e^{-(c+s)}\, ds \;=\; e^{-c}. \qquad \square$$

Consequently, if L_c is a $(1-\delta_B)$ lower confidence bound on $\mathbb{E}[T_c]$ from M i.i.d. draws of $T_c \in [-c, \log N]$, then $\hat{I} = \max\{0,\, L_c - e^{-c}\}$ satisfies $\hat{I} \leq I(i;z)$ with probability $\geq 1 - \delta_B$, and Lemma 2 licenses $B_{\mathrm{MI}} = \bar{R} - \hat{I}$ at that level. We use the empirical-Bernstein inequality in the form of Audibert et al. (2009), $L_c = \bar{T}_c - \sqrt{2V_M \ln(3/\delta_B)/M} - 3(c + \log N)\ln(3/\delta_B)/M$ with V_M the unbiased sample variance (which only loosens the bound); the inequality is two-sided at level δ_B and therefore conservative one-sided. At $c = 5$ the correction is $e^{-5} = 0.0067$ nats.

Lemma 5 (Certifying with an estimated agreement)**.** *Let $\hat{\mathcal{A}}_q$ average S posterior draws at each of the N test points, with independent noise for every (point, draw) pair, so that it is the mean of NS independent $\{0,1\}$-valued variables whose means average to $\mathcal{A}_q$; then $C_q = [a_-, a_+] = \hat{\mathcal{A}}_q \pm \sqrt{\ln(2/\delta_q)/(2NS)}$ contains $\mathcal{A}_q$ with probability $\geq 1 - \delta_q$ (Hoeffding). If $B \geq D_{\mathrm{KL}}(\bar{q}\|p)$ with probability $\geq 1 - \delta_B$, then*

$$U = \bigcup_{a \in C_q} \{b : d_{\mathrm{bin}}(a\|b) \leq B\} = \big[\, b_-(a_-),\, b_+(a_+) \big],$$

where $b_-(a) \leq a \leq b_+(a)$ are the two roots of $d_{\mathrm{bin}}(a\|b) = B$, contains $\mathcal{A}_{\mathrm{prior}}$ with probability $\geq 1 - \delta_B - \delta_q$.

Proof. On the event that $\mathcal{A}_q \in C_q$ and $D_{\mathrm{KL}}(\bar{q}\|p) \leq B$, of probability $\geq 1 - \delta_B - \delta_q$ by a union bound, Theorem 2(b) gives $d_{\mathrm{bin}}(\mathcal{A}_q\|\mathcal{A}_{\mathrm{prior}}) \leq B$, so $\mathcal{A}_{\mathrm{prior}}$ lies in the band at $a = \mathcal{A}_q \in C_q$ and

hence in U. Each band is an interval containing a, and both of its endpoints are non-decreasing in a: $\partial_a d_{\mathrm{bin}}(a\|b) = \log \frac{a(1-b)}{b(1-a)}$ has the sign of $a - b$, so raising a raises d_{bin} at every $b < a$ (the lower root moves up) and lowers it at every $b > a$ (the upper root moves up); hence the union of the bands over an interval of a is the interval stated. □

D Experimental details and disclosures

D.1 The full audit record

This subsection reproduces the complete account of the audit, including the preregistered replication, the failure modes the audit detected in our own pipeline, and the rate-side certificates at scale, of which the main text reports the highlights.

What the audit detects. Four failure modes occur in the models below, each named by the reading that found it. (i) *Generative drift*: $\mathcal{A}_q = 0.770 \to \mathcal{A}_{\mathrm{prior}} = 0.525$ on MNIST with a task probe and a label-free prototype control falling in step (both scored by the same external classifier, so the second is a sensitivity analysis rather than independent validation), and $\mathcal{A}[\bar{q}] - \mathcal{A}_{\mathrm{prior}} = +0.128$ to $+0.338$ on PPCA, where every quantity is exact and no estimator can be blamed. (ii) *Sampled-channel collapse*: one encoder map read at posterior *means* occupies all ten codes in $12/12$ replication models (2.20–2.28 of a possible 2.30 nats), while read on *draws* it occupies ≤ 5 codes in $11/12$ and ≤ 3 in $10/12$—the code is used, and the coupled channel is still gone. (iii) *A map-fitting artifact in our own audit*, $\mathcal{A}_{\mathrm{mean}} - \mathcal{A}_q$ up to $+0.52$, removed in $6/6$ by the pre-declared control. (iv) *An unlicensed budget*, certifying $20/20$ checkpoints it has no right to and excluding the truth in $30/30$, as Lemma 2 predicts. The first two are a property of the trained models, (ii) of the (model, declared map) pair—the pre-declared control restores $10/10$ codes in $12/12$—and (iii) and (iv) the auditor's; telling them apart is what the audit buys.

Generative codebook consistency: a case study and a preregistered replication. Case study (MNIST, $d = 10$). Convolutional VAEs (5 seeds, $K = 10$); c_{enc} a diagonal Gaussian mixture on encoder means, c_{dec} the Bernoulli–Bregman nearest-prototype rule on decoded outputs (natural indices, no relabeling); 10,000 test images, $S = 16$ posterior and 20,000 prior draws. Agreement survives posterior sampling but degrades under the prior: $\mathcal{A}_q = 0.770 \pm 0.009$ versus $\mathcal{A}_{\mathrm{prior}} = 0.525 \pm 0.007$ ($n = 5$), with a task probe (external classifier, accuracy 0.926, scored on decoded samples against the sampled code's majority label) dropping in step (0.700 ± 0.029 versus 0.390 ± 0.019) and a label-free prototype control preserving the drop ($0.696 \pm 0.029 \to 0.375 \pm 0.027$)—not a label-map artifact—while mean and sampled readings agree here ($\mathcal{A}_{\mathrm{mean}} = 0.771 \pm 0.009 \approx \mathcal{A}_q$).

Preregistered replication: a null, reported in full. We fixed, before any run, a 2×2 grid (FashionMNIST and SVHN, $d \in \{16, 32\}$, 3 seeds each) and the criterion: a cell replicates if $\mathcal{A}_q - \mathcal{A}_{\mathrm{prior}} > 0.05$ with the same sign in $3/3$ seeds *and* the task probe drops in $3/3$; the phenomenon generalizes if $\geq 3/4$ cells replicate. **Zero of four cells replicated** (Table 3); at FashionMNIST $d = 32$ the direction *reverses* in $3/3$, co-occurring with sampled-channel collapse (1–2 occupied sampled codes, label accuracy ≈ 0.10). Under the declared map the sampled channel is degenerate in $11/12$ of these models, so most cells never reproduced the case study's precondition—and the scaled grid below shows that degeneracy, and the null itself, belong to the declared map rather than to the trained models.

Mean and sampled readings of the same map also diverge on FashionMNIST in $6/6$ models, which the pre-declared refit control removes ($0/6$ survive, $10/10$ codes restored in all 12; Appendix D).

The coordinate price, on models. On linear-Gaussian models, where Theorem 3 is exact, a fixed DCT rotation of the same model moves the mean-field budget from 0.000 to 0.474 nats and voids the reconstruction-side certificate for all three map constructions, while the estimands move by at most 8×10^{-4} and the $O(d)$-closed full-covariance family is unaffected—Theorem 3(a) and (b), measured separately on the two sampling laws. *Trained* mean-field encoders pay the same price, unequally. The same models reproduce the case study's generative drift where every quantity is exact ($\mathcal{A}[\bar{q}] - \mathcal{A}_{\mathrm{prior}} = +0.128$ to $+0.338$), and the Theorem 2(b) band contains $\mathcal{A}_{\mathrm{prior}}$ in $30/30$

conditions while being full width at the licensed budget. Table 2, the trained-encoder arms and the rest of the numbers are in Appendix D.

Certify-or-Measure, end to end. We run the protocol ($w^* = 0.5$, $\delta = 0.05$, $n = 20{,}000$) on 52 trained models—20 MNIST screening checkpoints, the 12 models of the preregistered replication above, and 20 two-dimensional checkpoints added under a disclosed deviation (Appendix D)—always with a licensed budget, admitting on the exact criterion $B \le B^*(\mathcal{A}_q)$.

Under the loosest licensed budget, $\bar{R}$, nothing certifies. All 52 models abstain and measure. Where Theorem 3(e) binds, the impossibility is *attained* rather than merely bounded: on the six FashionMNIST models the measured $\hat{I}$ sits within 0.01 nats of $\log N$, so $D_{\mathrm{KL}}(\bar{q}\|p) = \bar{R} - \log N$ to that accuracy. The rest abstain on $B^*(\mathcal{A}_q)$ (Figure 3, Appendix D.3).

With the licensed MI budget of Section 5.1, it fires. On the six SVHN models (the only $d \ge 10$ family the floor leaves open): $B_{\mathrm{MI}} = 0.519$–0.606 nats against 1.467–1.555 for the best previously available licensed budget (Kolchinsky & Tracey, 2017), with residual over an unbiased point estimate of the truth of 0.058–0.061 nats—the concentration term plus the 0.007-nat clip correction of Lemma 4. Calling this a 2.5–3.0× tightening understates it. By Lemma 1 no band is narrower than $1 - e^{-B}$, so admission at w^*=0.5 requires $B \le \log 2 = 0.693$ *whatever* the agreement: at 1.467–1.555 nats the previous budget admits nothing narrower than 0.769, on $0/6$ models, for every estimator refinement, whereas ours lowers that floor to ≤ 0.455 and clears the necessary condition on $6/6$. The improvement is categorical, not quantitative—though necessary is not sufficient, and five of the six still abstain. The binary-KL form is load-bearing too: at the same budget and agreement, Pinsker gives width 0.553 and Bretagnolle–Huber 0.679, both abstaining where binary KL gives 0.487 and certifies. One model certifies: SVHN $d = 32$, seed 2, $B_{\mathrm{MI}} = 0.565 \le B^*(\hat{\mathcal{A}}_q{=}0.0216) = 0.589$; over the Hoeffding interval $\hat{\mathcal{A}}_q \pm 0.0021$ ($NS = 416{,}512$ independent draws) the certified band is $[0.000,\ 0.491]$ (width $0.491 \le w^*$, coverage $1 - \delta_B - \delta_q = 0.90$), which contains the realized $\mathcal{A}_{\mathrm{prior}} = 0.0349$—the first non-vacuous *rate-side* certificate above $d = 2$ in this line of work, on a nonlinear convolutional VAE. It survives Bonferroni over the six models (0.499) and the per-point empirical-Bernstein interval (0.491), abstaining only under per-point Hoeffding (0.503; Appendix D). The flip side, stated plainly: it fires where the channel is most degenerate—two-percent agreement, where B^* is loosest—certifying a bounded *failure* of consistency; three of the other five abstain at agreement 0.28–0.33, where $B^* = 0.16$–0.18 is three times stricter, and two, at agreement 0.03–0.06, abstain on the budget itself ($B_{\mathrm{MI}} = 0.551$, 0.606 against $B^* = 0.460$, 0.551).

Counting both branches. The scoreboard above is the rate side alone. Since $\min_a B^*(a) = B^*(0.5) = 0.144$, any budget below 0.144 certifies at *every* agreement, and the gap side clears it: the *trained* MLP encoder of Section 6.3 at $\bar{\Delta} = 0.088$ nats certifies unconditionally, and swept over native rank the same certificate fires in $75/75$ cells at $d \in \{2, \ldots, 32\}$ against $0/75$ for a trained *linear* encoder, with widths growing $0.106 \to 0.487$ and crossing w^* just above $d = 32$—the high-dimensional limit of this branch, located rather than conceded (Appendix D.5). Each branch fires where the theory says it can, and the protocol is run on the rate side for a licensing reason: Theorem 2(a) needs a certified *upper* bound on $\bar{\Delta}$, hence on $\log p(x)$, and IWAE bounds it from *below*—the gap-side instance of the direction Lemma 2 rules out.

The analyst's half of the verdict. The budget is map-free and the estimand is not, so admission can flip with the declared map—and on our own positive it does. Under the pre-declared refit control (Appendix D.6) (same checkpoint, same decoder rule, code map refitted on sampled rather than mean latents) the certifying model's $\hat{\mathcal{A}}_q$ moves $0.0216 \to 0.2996$, so B^* moves $0.589 \to 0.174$ and the same $B_{\mathrm{MI}} = 0.565$ gives a band of width 0.773: abstain. No reported band is wrong—soundness never used the map—but the admission did. Across the 240 models of Section 6.3 the same choice moves $\mathcal{A}_q$ by a median 0.565 over four declared maps ($239/240$ above 0.1), and the case study's posterior-to-prior drop appears in 2 of 80 cells when the map is fitted on posterior *means*, 28 on posterior *draws* and 0 on *prior* draws—largest where the prior sits just outside the fitting law and vanishing where the prior *is* that law, so the locked $0/4$ used the law under which the effect is rarest—Theorem 3(c)'s asymmetry, on its other axis.

Lemma 2 is not hypothetical: two unlicensed plug-ins certify $20/20$ and exclude the truth $30/30$ respectively (Appendix D), and the data caught only one. The protocol thus does what it claims—

certify where a licensed budget is small enough, measure elsewhere, and (Theorem 3d,e) say in advance which will happen—and w^* was not tuned.

The frontier, at scale. We re-ran the audit on 240 fresh models—FashionMNIST, MNIST, SVHN and CIFAR-10, up to the full training split, $d \in \{16, \ldots, 64\}$, $K \in \{10, 32\}$, KL weights β up to 64, three seeds—with every reading fixed beforehand (Appendix D.4).

The dataset floor is predictive, and it is the quantity that matters. Theorem 3(e) forbids certification once $\bar{R} > \log N + \log 2$. That inequality holds for 66 of the 240 models and *none* of them certifies; the precommit had registered any such certificate in advance as an instrument defect, not a refutation. Which side of the line a family falls on is not a matter of latent dimension: on FashionMNIST $\bar{R} - \log N$ reaches $+20.0$ nats and 36 models are excluded outright, while on SVHN it is negative throughout, -9.7 to -2.9, and none is.

The protocol certifies broken codebooks, not working ones. Overall 131 of the 240 configurations certify, against 1 of the 6 SVHN models the licensed budget was re-audited on above—so that rate was a property of the families audited. But 129 of the 131 certificates fire at $\mathcal{A}_q \leq 0.24$ (median 0.064), and no certificate clears its own split's base rate by more than 0.007: $\gamma(\alpha) \leq 2.2 \times 10^{-4}$ nats on every certificate, so Theorem 3(d)'s floor is inactive wherever the branch fires. That is the honest shape of the positive result, and Lemma 1 says it could not be otherwise: B^* is largest exactly where agreement is extreme. What a *working* codebook would require of $D_{\mathrm{KL}}(\bar{q}\|p)$ is quantified in Appendix D.6.

D.2 Two floors, and which one binds

Which of the two impossibilities binds, on the five audited MNIST checkpoints; Section 6.3 tests the same partition on 240 models.

Two floors, and which one binds. On the five audited MNIST checkpoints (the case-study models, not the collapsed replication models of Table 4) we measure α per Theorem 3(d)—held-out accuracy of the majority-label predictor on *sampled* codes against *true* labels—and find $\alpha \in [0.62, 0.78]$, so $\gamma(\alpha) \in [0.81, 1.28]$ nats, the chain $\bar{R} \geq \gamma(\alpha)$ holding $5/5$ ($\bar{R} \approx$ 27–28), and a floor on any rate-budget band of $1 - e^{-\gamma} \in [0.55, 0.72]$. The two impossibilities are not duplicates but a partition, and here (e) is the operative one: $\bar{R} - \log N = 17.99$–18.72 nats, 14.1–23.2 times γ, so the band cannot beat $1 - 10^{-8}$. Theorem 3(d) binds in the complementary window $\gamma \leq \bar{R} \leq \log N + \gamma$—low-rate but still informative codes—which no model audited here occupies. Either way abstention is the *correct* output.

D.3 The 52-model audit under the rate budget, in full

The measurement side of the 52-model audit in Appendix D.1, with the intervals and coverage counts the body compresses.

All 52 models abstain and measure (Figure 3, Appendix D): DKW intervals at ± 0.0096, containing an independently recomputed $\mathcal{A}_{\mathrm{prior}}$ $20/20$ (screening) and agreeing with freshly seeded draws within the combined width $12/12$ (replication). Theorem 3(e) explains most of it: on the six FashionMNIST models $\bar{R} - \log N = 18.2$–20.2 nats, so certification there is impossible for *any* estimator; nor is the floor loose—the measured $\hat{I} = 9.201$–9.208 against $\log N = 9.2103$ sits within 0.01 nats of maximum, so $D_{\mathrm{KL}}(\bar{q}\|p) = \bar{R} - \log N$ to within 0.01 nats.

D.4 The scaled frontier grid

What was fixed before the grid ran. The precommit (shipped as `PRECOMMIT_SCALE.md`) fixes the instrument at the values of Section 6.1—$S = 16$, 20,000 prior draws, the diagonal-GMM encoder map with `reg_covar`$= 10^{-4}$, the decoder-side Bregman prototype map, Adam at 10^{-3}, batch 256, 40 epochs, the licensed budget with $M = 20{,}000$, $c = 5$, $\delta = 0.05$, and the seed offsets $5000/41{,}000/53{,}000/51{,}000/91{,}000 + 100d + s$—and varies the dataset, the training size, d, K and β, with β the only training setting moved away from the paper's value. Stage 1 (FashionMNIST and SVHN, $d \in \{16, 32\}$, $K = 10$, six values of β, three seeds; 144 configurations) ran in full. Stage 2 (CIFAR-10 and MNIST at full size) was planned at 288 configurations ($d \in \{8, 16, 32, 64\}$, $K \in$

$\{10, 32, 128\}$, $\beta \in \{1, 4, 16, 64\}$, three seeds) and is incomplete: 96 ran before the deadline (CIFAR-10 with $d \in \{16, 32, 64\}$ and $K \in \{10, 32\}$; MNIST with $d \in \{16, 32\}$ and $K = 10$), and the refined β sweep planned as Stage 3 was not run. As the precommit requires, every completed row is reported. Six readings were written down first. One of them, (P2), asked whether the certify/abstain boundary coincides with $B_{\mathrm{MI}} = B^*(\mathcal{A}_q)$; it does by construction, since the decision is computed from the band that curve defines, so it carries no information and we record only that it was registered. The other five follow. **(P1)** No configuration with $\bar{R} > \log N + \log 2$ may certify; a violation is an instrument bug, not a result. **(P3)** If every certifying model has α within 0.02 of chance, certification is bought only with uninformative codes and we report that as the frontier; if some certifying model has $\alpha \geq 2/L$, the certify branch reaches useful codes and we report the region. **(P4)** We report the spread of $\mathcal{A}_q$ over the four declared maps as an interval and make no claim that one map is correct. **(P5)** We report the posterior-to-prior gap under each of the three fitting laws, stratified by α, as a rate across a new population. Both directions of every reading were publishable when they were written.

Governance. These are 240 *new* models across four datasets, reported as a separate population. They do not re-open the locked 2×2 replication: that verdict was fixed on the declared map and stands at $0/4$. The grid is also not a bit-level reproduction of the original runs—the training loop is GPU-resident and `sklearn`'s EM reaches a different optimum—so on the paper's own configuration (FashionMNIST, d=16, β=1, seed 0, 15,000 images) the grid gives $\bar{R} = 26.8$ against 27.8, $\alpha = 0.184$ against 0.229, $\gamma = 0.033$ against 0.071 and $\mathcal{A}_q = 0.154$ against 0.251. Within the grid every configuration is deterministic: a same-seed rerun reproduces all 65 substantive columns bit for bit after fixing cuDNN, the shuffling generator and `sklearn`'s thread count, and eighteen configurations that a double-writer accident ran twice agreed on every substantive column. Two instrument defects found after the first pass—a probe classifier fitted on the wrong set, and that double-writer accident—are recorded as a dated addendum to the precommit; the affected rows were discarded and recomputed. The probe is the declared one, fitted once per dataset on the whole training split, with test accuracy 0.9264 (MNIST), 0.8428 (FashionMNIST), 0.6145 (SVHN) and 0.4455 (CIFAR-10); on CIFAR-10 it is weak enough that we draw no probe-dependent conclusion there.

Results against those readings. **(P1)** Per dataset, the configurations above $\log N + \log 2$ and the certificates among them are $36/72$ and 0 (FashionMNIST, threshold 9.9035 nats), $0/72$ and 0 (SVHN, threshold 10.8602), $12/24$ and 0 (MNIST, 9.9035) and $18/72$ and 0 (CIFAR-10, 9.9035): 66 excluded in advance, 0 violations. **(P3)** Certificates number 13, 58, 6 and 54, and 131 in all; the highest α any certificate reaches is 0.1051 ($\beta = 32$, $\mathcal{A}_q = 0.056$), 0.1959 (26 tied certificates at $\beta \in \{2, \ldots, 32\}$, $\mathcal{A}_q = 0.026$–0.172), 0.1003 ($\beta = 64$, $\mathcal{A}_q = 0.067$) and 0.1063 ($\beta = 4$, $\mathcal{A}_q = 0.073$), against majority-class rates of 0.1000, 0.1959, 0.1135 and 0.1000 for the same splits—so no certified code beats its own split's base rate by more than 0.007, $\gamma \leq 2.2 \times 10^{-4}$ nats throughout, and the $2/L$ bar registered as "useful" is unreached everywhere. The first registered reading therefore holds on FashionMNIST, MNIST and CIFAR-10 in the strict form (within 0.02 of base rate) and on SVHN in the form that matters (at the base rate exactly). **(P4)** Map-sensitivity spread has median 0.565 over the 240 models, exceeds 0.1 in 239 of them, and reaches 0.921. **(P5)** Over the 80 cells, the posterior-to-prior gap clears 0.05 with the same sign in $3/3$ seeds in 2 cells under the mean-fitted map, 28 under the posterior-draw-fitted map and 0 under the prior-draw-fitted map; the probe half of the criterion is available for the declared map only, which is an asymmetry in this comparison. **(P6)** Positive rate $131/240 = 0.546$, and 129 of the 131 at $\mathcal{A}_q \leq 0.24$. Per β, no configuration certifies at $\beta = 1$; FashionMNIST certifies $1/12$ at $\beta = 16$ and $12/12$ at $\beta = 32$, SVHN $10/12$ at $\beta = 2$ and $12/12$ at every larger β, MNIST $6/6$ at $\beta = 64$ only, and CIFAR-10 $18/18$ at every $\beta \geq 4$. Per-configuration rows are in `s1_*.csv` and `s2_*.csv`; the readings are recomputed by `analyze_scale.py`.

How the families reach the certifying region. On FashionMNIST, MNIST and CIFAR-10 the rate is far above $\log N$ at $\beta = 1$ and admission has to be bought by raising β until the code collapses; the four FashionMNIST models with $\alpha \geq 0.12$ all sit at $\beta = 1$ with $\bar{R} \approx 28$ nats and certify $0/4$. On SVHN the rate is below $\log N$ at every β, so admission is available throughout—but the codes are collapsed there too, occupying one to five of $K = 10$ sampled codes, and the highest α a certifying model reaches is exactly the split's majority-class frequency. The two routes differ; the destination does not.

Claim D: code-addressed generation. The interface is: name a class c, look up the codes whose declared encoder-side semantics is c, draw from that component of the fitted map, decode, and let an external classifier—trained on data, never on codes—judge. The reading we registered was the Spearman correlation of that accuracy with $\mathcal{A}_q$, with $\mathcal{A}_{\text{prior}}$ and with five items of Definition 1, over the whole grid, with both outcomes publishable. *The registered analysis came out negative and we report it as such*: on FashionMNIST no predictor reaches $|\rho| = 0.27$—$\mathcal{A}_{\text{prior}}$ -0.08, $\mathcal{A}_q$ -0.20, the ELBO $+0.10$, reconstruction $+0.09$, rate $+0.05$, $I(x;z)$ $+0.06$, occupied codes $+0.22$, usage entropy $+0.26$—so on this population the coupled quantity is not needed to predict this interface, and neither is anything else. That weakens the motivation for this interface.

Two facts about the interface do survive, and they are not correlations. First, only 2 of the 72 FashionMNIST configurations can address all ten classes at all (mean coverage 0.811): the declared map's majority labels cover a strict subset, so the interface silently refuses classes while the code histogram it is built from looks healthy. Second, routing on the encoder's declared semantics is worse than routing on the decoder's own read of the *same* draw in $72/72$ configurations (mean -0.096, worst -0.240). That second gap is the coupling by construction, and no item of Definition 1 reports it. Its measurement degenerates when the map can address only one class—then every code carries the same majority label and the decoder-routed accuracy is 1 by identity—so we compute it only where coverage exceeds one class, which is all 72 FashionMNIST cells and excludes 23 of the SVHN cells.

Per dataset, as the precommit requires. FashionMNIST: no predictor reaches $|\rho| = 0.27$. SVHN: none reaches 0.13. MNIST ($n = 24$): every predictor does, and $I(x;z)$ $(+0.53)$ and the ELBO $(+0.51)$ lead $\mathcal{A}_{\text{prior}}$ $(+0.49)$ and $\mathcal{A}_q$ $(+0.39)$. CIFAR-10: the rate term $(+0.37)$, the ELBO $(+0.36)$ and $I(x;z)$ $(+0.36)$ lead while $\mathcal{A}_{\text{prior}}$ and $\mathcal{A}_q$ sit at -0.02; the probe there is weak (0.4455 test accuracy), which is a reason to discount that family but not to drop it from the report. Agreement never leads the marginal diagnostics by more than 0.02 on any of the four, and on the two families where anything predicts the interface a marginal diagnostic leads. The registered second branch therefore fires: for this interface the coupled quantity is not needed, and we report it as such.

Post-hoc, and not evidence: restricting to the twelve $\beta = 1$ cells, the only ones whose codes carry label accuracy above chance, $\mathcal{A}_{\text{prior}}$ predicts the interface at $\rho = +0.78$ against the ELBO's $+0.06$. The subset was not registered, $n = 12$, and we report the contrast only because it names the experiment a properly powered test would run—sweep models at fixed, non-trivial α—not because it establishes anything. We have tested one interface; retrieval, latent editing and rate–distortion compression are untested.

D.5 The gap-side certificate swept over native rank

The stress suite that produced Table 2 also fits PPCA at native ranks $d \in \{2, 5, 10, 20, 32\}$ (5 folds $\times$ 3 map constructions per rank) with two amortized encoders, and audits the reconstruction-side certificate of Theorem 2(a) in every cell. Table 1 reports it. The trained mean-field MLP encoder certifies in $75/75$ cells, every band containing the estimand; the trained *linear* encoder certifies in $0/75$. The separation is not a threshold artifact: the two budgets differ by one to two orders of magnitude at every rank ($\bar{\Delta} = 0.040$–0.230 against 2.62–18.08 nats). Band width grows monotonically with d and reaches 0.487 at $d = 32$, so on this family the branch runs out just above the rank we audited—the high-dimensional limit, located rather than asserted.

Two disclosures that belong with Table 1. First, the criterion. The precommit of that suite locked a *stricter* admission conjunction for its own utility claim—exact test $\bar{\Delta} \leq 0.05$ nats, Monte-Carlo-expanded certified lower endpoint ≥ 0.80, certified degradation ≤ 0.10, and an independent-RNG overlap—and recorded those gates as failing $0/5$ at canonical $d = 10$ and at padded $D = 32$. We therefore report both readings. Under this paper's declared $w^* = 0.5$ the certificate fires $75/75$. Under the stricter locked conjunction the picture at $d = 2$ splits by declared map, and the split is the paper's own thesis. On the unsupervised k-means and natural-Euclidean maps every certificate clause passes $5/5$ ($\bar{\Delta} = 0.0395$–0.0473, lower endpoint 0.8531–0.8875, degradation 0.0797–0.0995, independent overlap $5/5$) and the model is rejected only by that suite's nondegeneracy *screen*, which demands $\geq 5\%$ disagreement while these channels disagree on 3.33–4.91% of draws—the certificate passed and the screen rejected the model for being too healthy to be interesting. On the Mahalanobis map, the one that *does* clear the screen $5/5$, the same model fails the lower-endpoint and degrada-

	trained MLP encoder			trained linear encoder	
d	$\bar{\Delta}$ (nats)	band width	certify / cover	$\bar{\Delta}$ (nats)	certify
2	0.0395–0.0473	0.106–0.253	15/15 / 15/15	2.620–2.653	0/15
5	0.0635–0.0877	0.177–0.344	15/15 / 15/15	4.586–4.630	0/15
10	0.0773–0.0959	0.225–0.344	15/15 / 15/15	6.851–6.928	0/15
20	0.1172–0.1654	0.296–0.428	15/15 / 15/15	10.838–10.994	0/15
32	0.1483–0.2296	0.359–0.487	15/15 / 15/15	17.491–18.083	0/15

Table 1: **The gap-side certificate separates a trained mean-field MLP encoder from a trained linear one at every native rank from 2 to 32—75/75 against 0/75—and its width locates where the branch expires.** PPCA, 5 folds × 3 map constructions per rank; ranges are over the 15 cells. "certify" counts cells with band width $\leq w^* = 0.5$; "cover" counts cells whose band contains the estimand $\mathcal{A}[\bar{p}_\theta]$.

tion clauses 0/5 (0.6228–0.6473 and 0.1416–0.1550). The analyst's map decides both whether the channel looks degenerate and whether the certificate passes, on one checkpoint, at one rank. At $d \geq 5$ the clauses themselves fail as $\bar{\Delta}$ grows past 0.05 nats. The native ranks were pre-declared in that precommit as secondary controls and carry no prospective pass gate; we report them as a sweep, not as a gated claim. Second, the licensing. The budget here is $\bar{\Delta}$ itself, and IWAE bounds $\log p(x)$ from *below*, so outside models with tractable or upper-boundable evidence there is no licensed route to this branch—the gap-side instance of Lemma 2. That is why the protocol of Section 5.2 is run on the rate side.

D.6 The coordinate price, on models, in full

The full version of Section 6.3, whose body text states the conclusions and defers the table and the trained-encoder arms to here.

Linear-Gaussian models (probabilistic PCA, PPCA), $d = 10$, 5 folds, 10,000 test points, $S = 64$—the regime where Theorem 3 is exact. We compare the canonical basis against a fixed discrete-cosine-transform (DCT) rotation of the same model (likelihood, decoder, and all three map constructions transported). Theorem 3(b) predicts estimands do not move, (a) that the mean-field certificate pays; both hold, measured on the two sampling laws separately.

Reconstruction side (Table 2a). The estimand is the exact-posterior agreement $\mathcal{A}[\bar{p}_\theta] = 1 - \bar{\eta}_p$, the budget the mean-field gap B_Q—Theorem 2(a) and Theorem 3(a) currency. The budget rises from 0.000 to 0.474 nats under the rotation and the certificate collapses, for all three maps, while the estimand moves $\leq 5 \times 10^{-4}$. Two controls: the *full-covariance* family—closed under $O(d)$, hence inside Theorem 3(b)—certifies identically in both bases; and *trained* mean-field encoders pay too, unequally (an MLP encoder goes $0.088 \to 0.673$ nats, while a linear one is dominated by amortization error (Cremer et al., 2018) in *both* bases, $6.897 \to 7.476$ nats, so there the rotation is not what destroys the certificate).

Generative side (Table 2b). Theorem 2(b)'s estimand is agreement under *prior* draws, so we measure it directly (200,000 draws from $N(0, I_{10})$ per condition). It is basis-invariant to $\leq 8 \times 10^{-4}$, matching Theorem 3(b); PPCA shows the *same* generative drift as the MNIST case study in a model where every quantity is exact ($\mathcal{A}[\bar{q}] - \mathcal{A}_{\text{prior}} = +0.128, +0.156, +0.338$ across maps); and the certificate is sound but vacuous—at the licensed budget $\bar{R} \approx 20.6$–21.1 nats the band contains $\mathcal{A}_{\text{prior}}$ in 30/30 conditions with width 1.000 throughout.

Pipelines. Case study and replication: convolutional VAEs (1×28×28 for MNIST/FashionMNIST; a 3×32×32 variant for SVHN); $\beta = 1$; declared likelihoods: Bernoulli-family for near-binary images (continuous-Bernoulli for FashionMNIST, Loaiza-Ganem & Cunningham, 2019) and Gaussian with fixed $\sigma = 1$ for SVHN (declared before the runs, not tuned); $K = 10$; encoder map = diagonal GMM on encoder means of a stratified 15,000-image train subset (per seed); decoder map = Bregman nearest prototype in decoder-output space (Type-1 Bernoulli KL, or squared Euclidean for the Gaussian case); natural indices; full test splits; $S = 16$; 20,000 prior draws. PPCA suite: closed-form posteriors, 5 disjoint folds, three map constructions (natural-Euclidean class means, regularized Mahalanobis, unsupervised k-means), mean-field families in the canonical and DCT bases plus the full-covariance control, $S = 64$.

(a) map (canonical → DCT)	$\mathcal{A}_q$	$\mathcal{A}[\bar{p}_\theta]$	budget B_Q	band width
natural-Euclidean	$0.9016 \to 0.9013$	$0.9019 \to 0.9017$	$0.000 \to 0.474$	$0.000 \to 0.559$
Mahalanobis	$0.7869 \to 0.7864$	$0.7870 \to 0.7869$	$0.000 \to 0.474$	$0.000 \to 0.677$
k-means	$0.9076 \to 0.9074$	$0.9076 \to 0.9072$	$0.000 \to 0.474$	$0.000 \to 0.551$
k-means, full-cov. q	$0.9073 \to 0.9074$	$0.9074 \to 0.9074$	$0.000 \to 0.000$	$0.000 \to 0.000$
(b) map (canonical → DCT)	$\mathcal{A}[\bar{q}]$	$\mathcal{A}_{\text{prior}}$	drift	budget $\bar{R}$
natural-Euclidean	$0.9014 \to 0.9006$	$0.7738 \to 0.7731$	$+0.128$	$21.1 \to 20.6$
Mahalanobis	$0.7863 \to 0.7855$	$0.4484 \to 0.4491$	$+0.338$	$21.1 \to 20.6$
k-means	$0.9076 \to 0.9072$	$0.7512 \to 0.7518$	$+0.156$	$21.1 \to 20.6$

Table 2: **A change of basis alone moves the mean-field budget to 0.474 nats and voids the reconstruction-side certificate for every map, while the estimands move by at most 8×10^{-4} and the $O(d)$-closed full-covariance family is unaffected—the asymmetry of Theorem 3(c).** PPCA models, means over 5 folds. (a) *Reconstruction side* (optimal diagonal q unless noted): the estimand moves $\leq 5 \times 10^{-4}$. (b) *Generative side*: $\mathcal{A}_{\text{prior}}$ from 200,000 true prior draws (s.e. ≤ 0.002); the estimand is invariant under the rotation to $\leq 8\times10^{-4}$ (Theorem 3b), generative drift $\mathcal{A}[\bar{q}] - \mathcal{A}_{\text{prior}}$ is large in an exactly-solvable model, and the Theorem 2(b) band contains $\mathcal{A}_{\text{prior}}$ in $30/30$ conditions but is full-width at the licensed budget.

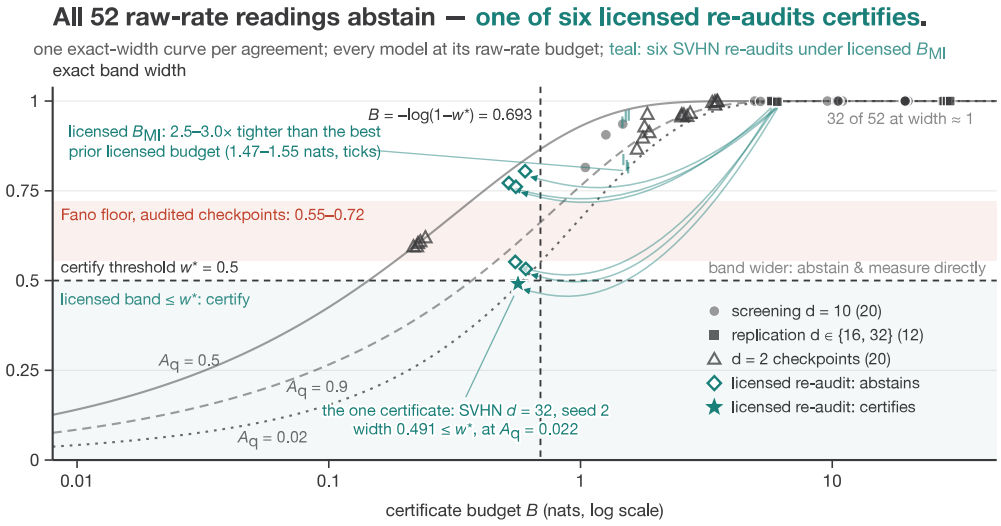

Figure 3: When the certificate can fire. Curves: exact band width, a function of budget *and* observed agreement; markers: all 52 protocol models, every one abstaining (the 20 MNIST screening models are placed at the direct estimate of $D_{\text{KL}}(\bar{q}\|p)$ from their audit; at their $\bar{R}$ budgets their widths are ≥ 0.999); certifying needs a point below w^*. The vertical line is Lemma 1's *necessary* condition $B \leq -\log(1-w^*)$—not sufficient: five $d = 2$ checkpoints clear it and still abstain, agreement near $\frac{1}{2}$ where the band is widest. Shaded strip: Fano floor $1 - e^{-\gamma}$ on the audited checkpoints (Theorem 3d). Teal: the six SVHN models re-audited under the licensed MI budget of Section 5.1—2.5–3.0× tighter than the best previously available licensed budget (ticks; $\bar{R}$ minus the Kolchinsky–Tracey bound on $I(x;z)$, Kolchinsky & Tracey, 2017)—of which exactly one (star) certifies; the six widths are those of the certified bands over the $\hat{\mathcal{A}}_q$ interval of Lemma 5.

Per-seed replication data. Table 4 gives every model of the replication grid. Two regime facts belong with it. First, the declared SVHN likelihood (Gaussian, σ fixed to 1) is a weak reconstruction term on these data: the resulting rates are $\bar{R} \approx 5.7$–6.1 nats against 27–29 nats on FashionMNIST, i.e. the SVHN cells sit in an effectively high-β regime with near-collapsed sampled channels (1–3 occupied sampled codes). We report this as a consequence of a preregistered choice, not as a tuned outcome; a continuous-Bernoulli or learned-σ variant would be a *new* protocol under a new precommit, not a repair of this one. Second, sampled-code label accuracy on the d=32 cells is at or near chance ($\alpha \approx 0.10$ for ten classes), so the Fano floor $\gamma(\alpha)$ is vacuous there—the codebooks in

cell	$\mathcal{A}_q - \mathcal{A}_{\text{prior}}$ per seed	task probe post→prior	verdict
Fashion d=16	$+0.109$ / $+0.034$ / $+0.030$	falls $2/3$, rises $1/3$	null
SVHN d=16	$+0.033$ / -0.001 / $+0.068$	rises $3/3$	null
Fashion d=32	-0.035 / -0.033 / -0.065	rises $3/3$	null (reversed)
SVHN d=32	-0.010 / $+0.027$ / -0.013	rises $3/3$	null

Table 3: Preregistered replication of the case-study phenomenon: all four cells fail the pre-declared criterion (gap > 0.05 of one sign *and* probe drop, each in $3/3$ seeds). SVHN probe columns use the precommitted 10-epoch evaluator (disclosed deviation, this appendix); verdicts are identical under either evaluator. Per-seed values: Table 4.

data	d	seed	$\bar{R}$	$\mathcal{A}_q$	$\mathcal{A}_{\text{prior}}$	$\mathcal{A}_{\text{mean}}$	occ.	α	$\gamma(\alpha)$
Fashion	16	0	27.84	0.251	0.142	0.631	10	0.229	0.071
Fashion	16	1	28.00	0.158	0.124	0.615	2	0.136	0.006
Fashion	16	2	27.46	0.158	0.128	0.680	5	0.171	0.024
Fashion	32	0	27.70	0.098	0.132	0.467	1	0.100	≤ 0
Fashion	32	1	28.56	0.126	0.158	0.524	2	0.105	0.000
Fashion	32	2	29.41	0.170	0.235	0.551	1	0.100	≤ 0
SVHN	16	0	6.07	0.334	0.301	0.142	2	0.179	≤ 0
SVHN	16	1	5.91	0.062	0.063	0.168	3	0.069	≤ 0
SVHN	16	2	5.98	0.333	0.265	0.098	1	0.067	≤ 0
SVHN	32	0	5.77	0.032	0.043	0.160	2	0.194	≤ 0
SVHN	32	1	5.72	0.282	0.255	0.211	1	0.067	≤ 0
SVHN	32	2	6.05	0.022	0.035	0.163	1	0.078	≤ 0

Table 4: **In $9/12$ models the sampled codes carry no usable label information at all, and the case study's precondition—an informative, non-degenerate coupled channel—was never reproduced here**; the γ column is the sharpest summary of the regime. Every model of the preregistered replication grid (no selection). $\mathcal{A}_{\text{mean}}$ is the mean reading; "occ." counts occupied sampled codes above 1% mass; α is held-out accuracy of the majority-label predictor on *sampled* codes and $\gamma(\alpha)$ the Fano floor it implies. Task-probe values, which depend on the external evaluator, are in the ledger for both evaluators.

those cells carry almost no label information, which is itself the reason the task-probe clause cannot be met.

The padded arm (outside the $O(d)$ statement). Padding the latent space from $d = 10$ to 32 before rotating is not an $O(d)$ action, and Theorem 3(b) does not cover it. Empirically the pad-only arm behaves like the canonical one (budget 0.000, band width 0.000, $\mathcal{A}[\bar{p}_\theta] = 0.9075$), while pad-then-rotate drives the mean-field budget to 29.4 nats and the band to full width, with $\mathcal{A}[\bar{p}_\theta] = 0.9074$ still unmoved; the full-covariance control certifies in every one of these arms (budget 0.000). Prior-draw agreement was not measured on the padded arms, so this arm speaks to Theorem 2(a) only. We report the padded arm as a labelled stress case, not as evidence for the theorem.

Preregistration and its three disclosed deviations. The replication grid and the protocol run were specified in a precommit document—model sets, readings, verdict criteria—before execution (reproduced verbatim in Figure 4). Three departures are disclosed here: two found afterwards by an adversarial audit of this manuscript, and one introduced by that audit's own correction.

(D1) Evaluator epochs. The SVHN external evaluator was trained for 12 epochs where the precommit specified 10. We retrained it at exactly 10 epochs and recomputed every classifier-dependent reading on the saved checkpoints. All four cell verdicts are unchanged under either evaluator: the agreement-gap clause is classifier-independent and fails on its own in all four cells, and the probe clause also fails at 10 epochs, with the probe *rising* post→prior in $6/6$ SVHN models. Table 3 quotes the 10-epoch (precommitted) readings. Absolute probe values are evaluator-sensitive and are never quoted here as evaluator-free quantities.

(D2) Protocol population. The precommit fixed the Certify-or-Measure inputs as the two screening families (jointly, the 20 MNIST screening checkpoints of Appendix D.1) plus the replication models. The first run omitted the replication models; we have since run the protocol on all twelve of them (reported in Appendix D.3: abstain $12/12$, fresh-draw consistency $12/12$). Separately, the 20

`PRECOMMIT_W2.md`: precommitted readings for the replication grid (verbatim)

```
- R-repro: the posterior-prior contrast reproduces in a cell if A_q - A_prior > 0.05
  with the same sign in 3/3 seeds AND the task probe drops under the prior arm with
  the same sign in 3/3 seeds. Report per-cell verdicts; the flagship claim survives if
  >= 3 of 4 (dataset x d) cells reproduce.
- R-null: cells that do not reproduce are reported as nulls, verbatim, and scoped into
  the paper (the finding is then dataset/dimension-conditional).
- Mean-vs-posterior arm is expected null (as at MNIST d=10); reported either way.
- ALL cells and seeds reported; numbers only from per_model CSVs.
```

`PRECOMMIT_LICENSED_MI.md`: fixed inputs and readings for the licensed MI budget (verbatim)

```
## Fixed before running (no tuning afterwards)
- M = 20000 samples, c = 5 nats, delta = 0.05, w* = 0.5.
- Models: all 12 E1-ext checkpoints; SVHN reported first (the open family).
- Sampling seed 91_000 + 100*d + seed.
- A_q is read from e1ext_per_model.csv; the admission threshold is the exact
  B*(A_q), **not** the 0.693 limit.
## Readings, fixed before running
- R-certify: a model certifies iff B_MI <= B*(A_q). Report every model.
- R-progress: if no model certifies but B_MI <= 0.8 on some model, the
  estimator route is working and the next step is a tighter concentration
  (more samples) or a better clip, not a new method.
- R-stop: if B_MI > 1.4 everywhere -- i.e. no better than the
  Kolchinsky-Tracey bound already measured -- the MI-lower-bound route is
  abandoned and pillar A stops. The paper then reports the impossibility
  framing (plan risk 1).
- Whatever happens, the split-half plug-in estimate of the truth is reported
  alongside, labelled as an estimate and never as a budget.
- w* is not re-tuned under any outcome (lessons #16).
```

Figure 4: The precommitted readings, reproduced verbatim from the two precommit documents in the supplement (non-ASCII glyphs transliterated: $\geq$, δ, $\times$, $\cdot$, dashes; Markdown backticks, blank lines and the bold markers around the reading names removed). The replication readings (top) fixed the verdict criteria of the replication in Appendix D.1 before the runs; the licensed-budget readings (bottom) fixed M, c, δ, w^*, the seeds, and the certify/progress/stop readings before the run of Section 6.2. The clip-tail correction and the $\hat{\mathcal{A}}_q$ interval of Appendix C.5 changed none of these inputs; their dated addendum sits in the same document (deviation D3).

two-dimensional checkpoints were *not* in the precommit and are labelled in the text as a post-hoc extension; under licensed budgets they all abstain, and they serve to probe the certify boundary and to demonstrate Lemma 2's unlicensed failure modes. Under the budgets fixed in the original precommit, the preregistered population abstains everywhere; the one certificate that fires (Section 6.2) does so on a preregistered model but under the licensed MI budget, whose estimator, clip, seeds, and readings were fixed in a separate, later precommit before that run. We consider this the honest reading and have not restated either as an original preregistered result.

(D3) The agreement interval. The precommit's certify reading compared B_{MI} with B^* at the point estimate $\hat{\mathcal{A}}_q$ read from the evaluation run. The interval treatment of Lemma 5—$\delta_q = 0.05$, Hoeffding over the NS posterior draws as the primary interval, the reading under which the test split is the reference distribution—was added after the run, on 2026-09-08, in the same dated addendum as the clip correction; the per-point variants were computed at the same time and are reported above, including the one under which the model abstains.

A further minor issue (one verification script overwrote its output file instead of appending) was fixed without data loss. All are recorded in the results ledger shipped with the supplement.

The mean-vs-sampled divergence and its control, in full. The replication runs show the *mean* and *sampled* readings of the same encoder map diverging on FashionMNIST in $6/6$ models ($\mathcal{A}_{\mathrm{mean}} - \mathcal{A}_q \in [+0.37, +0.52]$), where at MNIST $d = 10$ they agreed. Under the pre-declared control (refit the same family on *sampled* latents, change nothing else) the divergence collapses to $[+0.002, +0.012]$ on FashionMNIST—$0/6$ survive—and to $[+0.007, +0.119]$ on SVHN: a map-fitting artifact. The control also removes the sampled-channel collapse (all 12 models then use $10/10$ codes), so the degeneracy confounding the null belongs to the *declared map construction*, not the trained models—that does not revive the null, but it names what a revised protocol must change. Under the mean-fitted map, mean-side code usage never collapses ($10/10$ occupied codes in all 12 models, usage entropy 2.20–2.28 nats) while sampled-side usage collapses in $11/12$ models to at most five codes ($10/12$ to at most three). The control refits the same mixture family, with the same K, initialization seed and regularizer, on one sampled latent per training example (seed

series $71{,}000 + 100d + s$), leaves the checkpoint, the decoder *rule*, the test split and the test draws untouched—the decoder prototypes are re-derived from the refitted mixture's means, exactly as the main run derives them, so one analyst choice moves both readers—and recomputes both readings. Per-model results are in the supplement; the summary is that the divergence collapses to ≤ 0.012 on all six FashionMNIST models and to ≤ 0.119 on SVHN, and that every one of the 12 models then uses all 10 codes on sampled latents. We had pre-declared that survival in $\geq 5/6$ FashionMNIST models would keep the finding; $0/6$ survive, so the finding is downgraded to a statement about map transfer.

Two live demonstrations that budget validity is load-bearing. A grid plug-in of $D_{\mathrm{KL}}(\bar{q}\|p)$ certifies all 20 $d = 2$ checkpoints with band widths 0.061–0.493, and every band happens to contain the realized $\mathcal{A}_{\mathrm{prior}}$; under the licensed $\bar{R}$ the same checkpoints certify $0/20$. A binned KL is a data-processing lower bound, so that budget was never licensed and the agreement of its bands with the truth was luck. A moment-matched Gaussian plug-in on the PPCA panel gives bands of width 0.111–0.181 that *exclude* the realized $\mathcal{A}_{\mathrm{prior}}$ in $30/30$ conditions; there the error was visible in the data. Both are instances of Lemma 2.

The admissible agreement window, and where the shortfall lives. Because B^* is strictly decreasing on $[0, 1/2]$, a budget caps the agreement it can ever admit, and the cap is computable. At the certifying model's $B_{\mathrm{MI}} = 0.565$ the protocol can admit only $\mathcal{A}_q \leq 0.0281$ or, symmetrically, $\mathcal{A}_q \geq 0.9719$; at the best of the six, 0.519, only $\mathcal{A}_q \leq 0.0420$. Certifying the case study's own $\mathcal{A}_q = 0.770$ would need $B \leq 0.2073$, and $\mathcal{A}_q = 0.5$ needs $B \leq 0.1438$. The residual between B_{MI} and an unbiased point estimate is 0.058–0.061 nats, so even a perfect estimator of $I(x; z)$—budget $0.565 - 0.058 = 0.507$—would admit only $\mathcal{A}_q \leq 0.0459$. The remaining gap is therefore not estimator-side: rate-budget certification of generative codebook consistency at useful agreement needs aggregate-posterior mismatch below roughly 0.14–0.21 nats, which is a target for the prior-flexibility line (Tomczak & Welling, 2018; Bauer & Mnih, 2019) rather than for a sharper bound on $I(x; z)$. The gap side is not subject to this cap: since $\min_a B^*(a) = B^*(0.5) = 0.1438$, the trained MLP encoder of Section 6.3 at $\bar{\Delta} = 0.088$ nats certifies at every agreement, with band $[0.7335, 0.9815]$ (width 0.2480) at its observed $\mathcal{A}_q = 0.9016$ and width at most 0.4017 over all agreements.

The licensed MI budget. $M = 20{,}000$ samples, statistic clipped below at $c = 5$ nats, $\delta_B = 0.05$, sampling seeds fixed before the run together with the readings. The lower clip raises the mean of the statistic by at most $e^{-5} = 0.0067$ nats (Lemma 4), which is subtracted from the empirical-Bernstein lower bound; in the samples the clip binds on at most 3 of the $20{,}000$ draws per model and the empirical tail mass is below 10^{-4} nats, so the analytic correction is conservative. The clip and the empirical-Bernstein constant are the only free choices; the concentration term plus the correction is 0.058–0.061 nats on the SVHN models, i.e. the whole residual between B_{MI} and an unbiased point estimate of the truth. The precommit stated that the clip *lowers* the mean; the sign error was found by an adversarial audit after the run, is recorded as a dated addendum to the precommit, and its correction raised every budget by 0.0067 nats without changing any verdict. The precommit also promised the split-half plug-in estimate of $D_{\mathrm{KL}}(\bar{q}\|p)$ alongside: it is 0.244–0.428 nats on the six SVHN models (`kl_plugin_split.csv`), below the point estimates $\bar{R} - \bar{T}_c = 0.459$–$0.548$ nats to which the residual above refers; it evaluates the log-density of a random half of the mixture, which by Jensen underestimates $\log \bar{q}$ on average, so it is biased downward and is quoted as a lower reference, never as a budget.

The agreement estimate. $\hat{\mathcal{A}}_q$ averages $S = 16$ posterior draws at each of the $N = 26{,}032$ test points, $NS = 416{,}512$ independent indicators; the Hoeffding interval of Lemma 5 is ± 0.0021 at $\delta_q = 0.05$, and the certified band is the union over it, width 0.491 against 0.487 at the point estimate. Sensitivity: if the test points are instead regarded as an i.i.d. sample, so that the draws sharing a point are dependent through it, the empirical-Bernstein interval on the N per-point means (Audibert et al., 2009) is ± 0.0022 and the band still certifies (width 0.491); Hoeffding on the per-point means (± 0.0084) gives width 0.503 and would abstain. These variants change only the agreement interval; the budget remains the test-split quantity, so under the i.i.d. reading they gauge the verdict's sensitivity to the interval rather than constitute certificates. At $\delta_B = \delta_q = 0.025$ (coverage 0.95) the width is 0.494; under a Bonferroni correction over the six SVHN models (family-wise 0.90) it is 0.499—a

certificate with no margin to spare, which is the honest reading of a two-percent-agreement channel. All variants are tabulated in the supplement (`aq_interval.csv`).

The comparison baseline—the best previously available licensed budget, 1.467–1.555 nats—is $\bar{R}$ minus the Kolchinsky–Tracey Bhattacharyya lower bound on $I(x;z)$, computed in closed form for the Gaussian-mixture $\bar{q}$ (Kolchinsky & Tracey, 2017) and licensed by the same rate decomposition (Lemma 2); the closed-form $\log(1+\chi^2(\bar{q}\|p))$ budget was also computed and is looser on every SVHN model (8.1–15.6 nats).

Attribution of measured quantities. Sampled-code label accuracies use held-out true labels with train-side majority votes only (the Fano chain of Theorem 3(d) requires the label, not a code-derived proxy); plug-in KL budgets are used only in the conservative direction established by Lemma 2; and per-map, per-arm ranges are reported without selection—including the reversed-direction cells and the mixed-sign SVHN divergence that accompany the FashionMNIST findings.

Relation to the prior preprint, and attribution of Theorem 2. The coupled channel, the post-processing decomposition and the gap-side identity and bound of Theorem 2(a), and a marginal-impossibility result, appear in a prior preprint (Hayashi, 2026) as its Lemma 7 and Corollary 8 and its Proposition 4. That Proposition matches the two code histograms, their entropies, the occupied-code counts and $I(x;z)$; its Remark 5 records that a Hungarian-matched score can hide the deranged case. Relative to it, the following are new here. (i) Definition 1 closes the hole Remark 5 names and widens the package well past it: the matched pair here also agrees on *every* permutation-symmetric functional of the joint code table—including the permutation-matched agreement itself, so no relabeling-invariant summary separates the pair—and on the *per-example* reconstruction, rate, ELBO, $\log p(x)$ and variational gap, on the generative law $p(x)$ as a distribution, on the per-dimension active-unit statistics, on $\mathcal{A}_{\mathrm{prior}}$, and (item (I′)) on all latent moments below a designable order. (ii) Theorem 2(b), the rate-side bound, and with it the two-sidedness, is new. (iii) Theorem 3 in all of its parts, Lemma 1, Lemma 2 and the licensed mutual-information budget, the admission budget $B^*(a)$ and the Certify-or-Measure protocol have no counterpart in the preprint, and neither does any audit reported here. Two model populations are shared with it: the five MNIST models of the case study (Section 6.1), which are also the $\beta = 1$ arm of the 20 screening models, and the 20 two-dimensional checkpoints of Section 6.2 are trained models audited in the preprint. For the two-dimensional checkpoints, the encoder-side agreement and the rate are taken from the preprint's audit tables, the agreement re-verified here by recomputation to a relative error below 10^{-7}; every other quantity reported for these models is computed here. The derivations here are self-contained.