WHEN DOES CONDITIONAL VENDI MEASURE POSTERIOR DIVERSITY?

Anonymous authors
Paper under double-blind review

ABSTRACT

Conditional generation requires separating diversity within a condition from variation across conditions. We characterise the discrepancy between averaged conditional kernel entropy and joint-minus-conditioning entropy. For a finite condition catalogue, we obtain an attainable certificate determined by the conditioning kernel, an exact equality criterion, and sharp recovery limits when only the aggregate score is retained. Global spectral truncation creates a second obstruction: even with orthogonal condition features, its rank budget can erase or reverse conditional diversity. A conditional tail decomposition retains an explicit approximation bound. Exact controls isolate these mechanisms, and an audit of 1,200 released images finds 31 native-generator ranking reversals across 112 kernel and subset specifications. The initial correction certifies none of these reversals, while a post hoc refinement certifies all on the same banks. A prospective Qwen3 study with 4,096 stochastic outputs certifies a finite-bank temperature ordering, with no native ranking reversal. Population entropy rankings remain unresolved. The analysis identifies which conclusions aggregate scores support and gives an evaluation protocol based on direct grouping, explicit compression error and sampling uncertainty.

1 INTRODUCTION

A conditional generator can vary across prompts while always producing the same image at a fixed prompt. Its within-prompt diversity is zero. Evaluation should distinguish these sources of variation. In inverse problems, for observations $Y = \mathcal{A}(X) + \xi$, a sampler's conditional law Q_y aims to represent uncertainty about X given y.

Vendi measures diversity by exponentiating kernel spectral entropy (Friedman & Dieng, 2022). Two conditional constructions average entropy within the laws Q_y (Nguyen & Dieng, 2025) or subtract observation entropy from joint entropy (Sánchez Giraldo et al., 2014; Jalali et al., 2026). Both are well-defined, but overlapping condition features make them different quantities (Bach, 2022). Holding the output laws, weights and kernel fixed, changing only the condition representation can reverse a score's ranking of generators. Global spectral truncation can introduce a further reversal even when the condition labels are represented exactly. Figure 1 illustrates both mechanisms.

We determine when a joint-kernel conditional score supports a within-condition interpretation. The analysis separates conditioning distortion, spectral compression and sampling error. Its contributions are:

1. **Conditioning and recoverability.** We derive an attainable conditioning certificate, an exact equality condition and sharp limits on recovery from aggregate scores. A quantitative overlap bound supplies a population distortion witness. Sharpening and partition analysis identify the roles of condition resolution and repeated sampling.
2. **The global rank-budget obstruction.** We establish an exact law showing why global truncation fails to preserve conditional entropy. A Gaussian counterexample contradicts Jalali et al. (2026, Appendix B, Theorem 7). Conditional tail accounting preserves the decomposition with an explicit remainder.

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.

3. **Consequences for generator evaluation.** Exact controls and pretrained image and text studies demonstrate distortion and finite-bank ranking reversals. Prospective certification and an amended model-selection study test what the certificates resolve on new outputs, including their abstentions.

Direct grouped scoring computes the empirical target when labelled repeated outputs are available. The certificate quantifies the ambiguity of aggregate scores. All log-diversities are in nats.

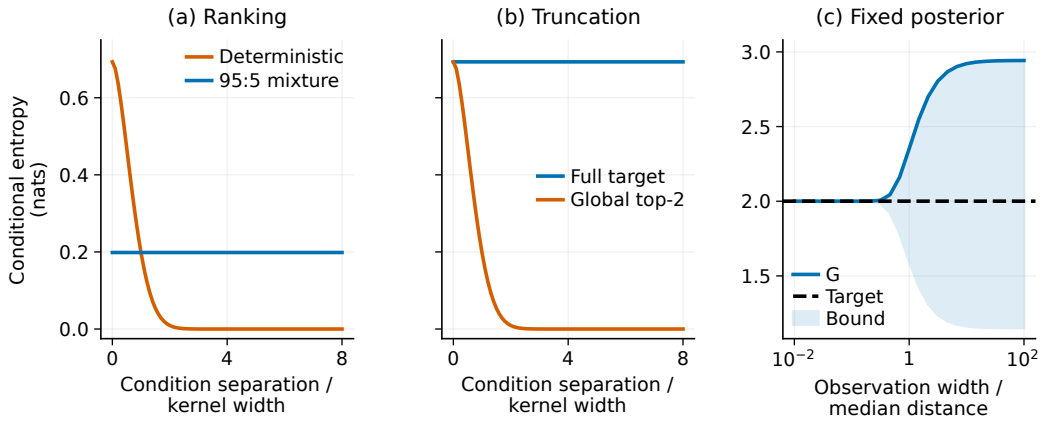

Figure 1: Conditioning overlap and global truncation can change the interpretation of a diversity score. (a) Changing only the conditioning kernel reverses the diversity ranking of deterministic condition-dependent outputs and a fixed 95:5 output mixture. (b) Global top-2 truncation erases conditional entropy although the full score and within-condition target both equal $\log 2$. (c) The exact posterior on a fixed digit dictionary is unchanged while the paired score varies with observation bandwidth. The shaded region is the deterministic interval $[\max(0, G - C), G]$, not a statistical confidence interval.

2 Targets and Related Work

We fix positive semidefinite kernels $k(x, x') = \langle \phi(x), \phi(x') \rangle$ and $\ell(y, z) = \langle \psi(y), \psi(z) \rangle$ for outputs and conditions. Both have unit diagonal, so their feature vectors have unit norm. Feature maps are strongly measurable into separable Hilbert spaces, the underlying spaces are standard Borel, and $y \mapsto Q_y$ is a measurable probability kernel. The conditional feature covariance and the two entropy targets are

$$\rho_y = \mathbb{E}_{X \sim Q_y} \phi(X)\phi(X)^*, \quad H_{\text{post}} = \mathbb{E}_Y S(\rho_Y), \quad G = S(\rho_{AB}) - S(\rho_B), \tag{1}$$

where $S(\rho) = -\operatorname{Tr}(\rho \log \rho)$, $\rho_{AB} = \mathbb{E}_Y[\rho_Y \otimes \psi(Y)\psi(Y)^*]$, $\rho_A = \mathbb{E}_Y \rho_Y$, and $\rho_B = \mathbb{E}_Y \psi(Y)\psi(Y)^*$. Each covariance is positive with trace one. Its eigenvalues describe how feature mass is distributed across orthogonal directions, and S is the Shannon entropy of that spectrum. We assume $S(\rho_A), S(\rho_B) < \infty$, which makes equation 1 finite. H_{post} averages diversity within the laws Q_y, whereas G subtracts the entropy of their condition representation. The notation H_{post} describes true posterior diversity only when $Q_y = P_{X|Y=y}$. Its exponential is the geometric mean of conditional Vendi scores.

For n equally weighted paired draws, G has the empirical counterpart

$$\widehat{G} = S((K \circ L)/n) - S(L/n), \qquad K_{ij} = k(x_i, x_j), \quad L_{ij} = \ell(y_i, y_j). \tag{2}$$

The Gram identity follows because a feature matrix and its adjoint have the same nonzero squared singular values. For non-unit output feature norms, keeping observation features unit norm, the covariance and Gram matrix must instead be divided by their trace. Global normalisation then changes conditional weights unless conditional expected squared norms agree. Appendix B makes this distinction explicit.

Position within existing work. Kernel covariance entropy and matrix-based conditional entropy predate Vendi (Sánchez Giraldo et al., 2014; Bach, 2022). Vendi and its order-q family apply these spectral quantities to diversity evaluation (Friedman & Dieng, 2022; Pasarkar & Dieng, 2023). The order-two score is inverse purity. Nguyen & Dieng (2025) use the difference between marginal and averaged conditional entropies for Vendi information gain. Jalali et al. (2026) develop joint-minus-conditioning Vendi and its truncated version for prompt-aware evaluation. We examine when the joint-minus-conditioning target can be interpreted as averaged conditional entropy, including after truncation.

Wu & Farnia (2026) study the same paired functional, its concavity and finite-sample underestimation, and entropy-constrained reweighting. They explicitly distinguish it from averaged conditional entropy. Underestimation of G does not imply underestimation of H_{post}. Our certificate quantifies this separate target discrepancy.

Finite-sample truncation is studied by Ospanov & Farnia (2025), and Farnia et al. (2026) establish diversity bias, monotonicity of expected log-Vendi, and concentration. We use these established statistical ideas with explicit population-target allowances. Likewise, per-condition diversity and the separation of diversity from consistency and realism are already used in conditional generation (Kannen et al., 2024; Astolfi et al., 2024). We build on these foundations to establish certificates and counterexamples for the conditional interpretation.

3 A CERTIFICATE FOR CONDITIONING DISTORTION

Overlapping condition features cannot distinguish labels as an orthogonal encoding would. Consequently, some variation between conditional laws remains in G. To quantify it, we use spectral inequalities for positive, trace-one matrices, called density operators in quantum information. Here they are ordinary feature covariances, and the argument requires no physical interpretation.

Theorem 3.1 (Sharp conditioning certificate)**.** *Under Section 2's assumptions,*

$$0 \le G - H_{\mathrm{post}} = S(\rho_A) - H_{\mathrm{post}} - I(A:B).$$

For m conditions with probabilities $p_y > 0$, let $L_{yz} = \ell(y,z)$ and $D_p = \mathrm{diag}(p)$. Then

$$0 \le G - H_{\mathrm{post}} \le C(p,L) := H(p) - S(D_p^{1/2} L D_p^{1/2}). \tag{3}$$

The upper endpoint is attained by mutually orthogonal conditional output supports. The lower endpoint by identical conditional states. Orthogonal conditioning features imply $G = H_{\mathrm{post}}$ for all output ensembles. Conversely, among arbitrary density-operator ensembles in an output space of dimension at least m, universal equality requires orthogonal conditioning features.

Here $I(A:B) = D(\rho_{AB} \| \rho_A \otimes \rho_B)$ is quantum mutual information. Data processing bounds it by $S(\rho_A) - H_{\mathrm{post}}$, proving the first inequality. The entropy-of-mixing upper bound proves equation 3. These are applications of established quantum-information inequalities (Schumacher & Westmoreland, 2000; Ruskai, 2002). Full proofs are in Appendix B. With an orthogonal classical label register Y, the discrepancy equals $I(A:Y \mid B)$ and the certificate equals $H(Y \mid B)$. For a fixed feature map, the converse requires the stated orthogonal output states to be realisable.

The certificate is computable before observing outputs. If orthogonal output ensembles are admissible, no uniformly smaller bound depending only on (p, L) is possible. It gives $H_{\mathrm{post}} \in [\max\{0, G - C\}, G]$, or equivalently $1 \le e^G / e^{H_{\mathrm{post}}} \le e^C$. Changing only the observation kernel can therefore change the reported score while leaving the target fixed. If $X = Y$ for two equally likely categories, the output features are orthogonal, and the absolute observation overlap is $c \in [0,1]$, then

$$H_{\mathrm{post}} = 0, \qquad G = C = \log 2 - h_2((1+c)/2), \tag{4}$$

where $h_2(u) = -u \log u - (1-u)\log(1-u)$. A deterministic generator can consequently receive any log-diversity from zero to $\log 2$.

Theorem 3.2 (Equality and a quantitative overlap obstruction)**.** *For the finite catalogue in Theorem 3.1, write* $\Delta = G - H_{\mathrm{post}}$, $c_{yz} = |L_{yz}|$ *and* $D_{yz} = \|\rho_y - \rho_z\|_1/2$. *Join* y, z *when* $c_{yz} > 0$. *Then* $\Delta = 0$ *exactly when the conditional states coincide within each connected component. For every pair,*

$$\Delta \geq 2\lambda(c_{yz})\frac{p_y p_z}{p_y + p_z} D_{yz}^2, \qquad \lambda(c) = 1 - \frac{1-c^2}{c}\,\mathrm{atanh}(c), \tag{5}$$

with $\lambda(0) = 0$ *and* $\lambda(1) = 1$. *The coefficient* $2\lambda(c)$ *cannot be increased uniformly over binary priors and output ensembles.*

Appendix B.2 derives the bound by a binary measurement and a scalar entropy-curvature inequality. It also permits sums over disjoint pairs. Equality specialises established quantum recovery theory (Hayden et al., 2004) to a finite catalogue of kernel density states. Different laws can have the same state. We make no priority claim for the coefficient. Sharpness is asymptotic at the uniform binary prior, rather than pointwise at each prior.

What can the aggregate score recover? Fix (p, L) and allow arbitrary conditional states in dimension at least m. Among rules observing only (G, p, L), the smallest worst-case absolute error for H_{post} is exactly $C(p, L)/2$. The rule $\max\{0, G - C/2\}$ attains it. For the lower bound, orthogonal pure conditional states give $(G, H_{\mathrm{post}}) = (C, 0)$, while identical states of entropy C give (C, C). These indistinguishable scalar observations also require an identification interval of width at least C. Moreover, a score difference greater than C guarantees the same target ordering, whereas every positive margin below C can be reversed. Appendix B.4 proves both sharp consequences. These limits apply when only (G, p, L) is retained. Conditional samples and full Gram blocks supply additional information.

Sharpening has a population guarantee and a sampling cost. For $m \geq 2$ and $\epsilon = \max_{y \neq z} |L_{yz}| < 1$, integer tensor powers give the valid kernel $L^{\circ a}$. Data processing and a relative-entropy bound yield

$$H_{\mathrm{post}} \leq G_{a+1} \leq G_a, \qquad G_a - H_{\mathrm{post}} \leq \log(1 + (m-1)\epsilon^{2a}). \tag{6}$$

A tolerance thus determines a sufficient power before observing outputs, for a fixed finite set of conditions. For a fixed paired sample, however, assuming $|\ell(y, z)| < 1$ for distinct observed values,

$$\widehat{G}_a \longrightarrow \sum_y \frac{n_y}{n} S(K_y/n_y), \tag{7}$$

where n_y is the number of repeats at y and K_y the output Gram block. If every observed y is distinct, the limit is zero. Indeed $L^{\circ a} \to I_n$ and $K \circ L^{\circ a} \to I_n$. Arbitrarily narrowing the kernel therefore collapses this estimator on a bank of distinct observations. Section 5 treats repeated sampling and controlled pooling as alternatives under explicit assumptions.

These limits distinguish three uses of the theory. Retained (G, p, L) support an identification interval. Retained joint data permit a resolution analysis through G_a. Labelled repeated outputs permit direct grouped entropy, which already computes the finite-bank target exactly and is our recommended estimator. Orthogonal labels yield the same answer with $C = 0$. A single condition also has $C = 0$.

4 Global truncation does not preserve conditioning

Conditioning distortion is present without approximation. Spectral truncation introduces a different error even when that distortion vanishes. Consider the redistributed top-t spectrum used by Ospanov & Farnia (2025); Jalali et al. (2026):

$$T_t(\lambda)_i = \lambda_i + \frac{1 - \sum_{j=1}^t \lambda_j}{t}, \quad 1 \leq i \leq t, \qquad S_t(\rho) = H(T_t(\lambda(\rho))), \tag{8}$$

where the decreasing eigenvalue sequence is padded with zeros. Define $G_t = S_t(\rho_{AB}) - S_t(\rho_B)$. This operation changes the joint and marginal spectra separately.

Proposition 4.1 (Exact rank-budget law)**.** *Let* Y *be uniform on* m *orthogonal conditioning states and every conditional output state equal* I_r/r. *For every integer* $t \geq 1$,

$$H_{\mathrm{post}} = G = \log r, \qquad G_t = \log\min(t, mr) - \log\min(t, m). \tag{9}$$

Consequently $G_t = 0$ *for* $t \leq m$. *Full recovery requires* $t \geq mr$ *when* $r > 1$.

The joint spectrum has mr equal eigenvalues and the conditioning spectrum has m. A global budget of t must represent both the condition labels and the variation within each condition. Thus $t \geq r$ suffices to retain each conditional entropy separately, but can still erase it after global compression. More generally, for uniform orthogonal conditions and $t \leq m$, $S_t(\rho_B) = \log t$ while $S_t(\rho_{AB}) \leq \log t$, so $G_t \leq 0$. It is strictly negative when the redistributed joint spectrum is nonuniform. For two equally likely conditions with output spectra (1) and $(1/2, 1/2)$ on disjoint supports, $G_2 = h_2(5/8) - \log 2 = -0.031584$, although $H_{\text{post}} = \frac{1}{2}\log 2 > 0$. A condition-independent deterministic law has $G_2 = H_{\text{post}} = 0$, so maximising G_2 strictly prefers that law in this comparison.

A finite counterexample with Gaussian kernels on both inputs. Consider the truncated interpretation in Jalali et al. (2026, Appendix B, Theorem 7). Let X and Y be independent and each uniform on $\{0, D\}$, with unit-bandwidth Gaussian kernels on both variables. Writing $c = e^{-D^2/2}$, the conditioning and output spectra are $((1+c)/2, (1-c)/2)$. The joint spectrum is their tensor product. Redistributing the omitted mass at $t = 2$ gives $(q, 1-q)$ with $q = 1/2 + (c + c^2)/4$, so

$$G_2 = h_2(q) - h_2((1+c)/2), \qquad \sum_y p_y S_2(\rho_y) = h_2((1+c)/2). \tag{10}$$

The discrepancy tends to $\log 2$ as D grows. The cited Eq. (16) gives $\Gamma = 16e^{-D^2}$, whereas its Theorem 7 bounds that discrepancy by $\sqrt{2\Gamma}\log(2/\Gamma)$, which tends to zero. At $D = 4$, the discrepancy is .693147082 nats and the claimed bound .026416531. Appendix C gives the spectra and assumption check. The error arises from globally truncating the weighted block spectrum before decomposing it by condition. For the untruncated independent-output law, $G = H_{\text{post}}$.

A decomposition that retains the omitted mass. For the leading t_y eigenvalues of ρ_y, write $\delta_y = 1 - \sum_{j \leq t_y} \lambda_{yj}$ and define

$$L_{t_y}(\rho_y) = \sum_{j \leq t_y} -\lambda_{yj} \log \lambda_{yj} - \delta_y \log \delta_y. \tag{11}$$

This retains the tail as one bin. Exactly,

$$S(\rho_y) - L_{t_y}(\rho_y) = \delta_y S(\lambda_{y,>t_y}/\delta_y) \leq \delta_y \log(r_y - t_y)$$

when $r_y > t_y$ is a valid finite rank bound. A zero tail contributes zero. Placing these compressed spectra in separate orthogonally labelled blocks gives joint entropy $H(p) + \sum_y p_y L_{t_y}(\rho_y)$. Subtracting $H(p)$ therefore preserves the desired conditional decomposition, with remainder at most $\sum_y p_y \delta_y \log(r_y - t_y)$. The construction requires conditional spectral blocks and controls their entropy approximation. For empirical blocks, population inference additionally requires a sampling allowance, developed next. The compressed blocks represent the conditional decomposition rather than the original cross-condition geometry.

5 Finite Samples and an Evaluation Protocol

For iid draws from a fixed conditional law, Appendix A proves almost-sure empirical entropy convergence to $S(\rho)$, including infinite entropy, under strongly measurable features and finite positive expected squared norm. At order one, finite population entropy suffices without an empirical tail assumption. Uniform finite-sample rates require additional structure. A law can place probability $1-\varepsilon$ on one feature direction and distribute ε over M unseen orthogonal directions, adding $h_2(\varepsilon)+\varepsilon \log M$ entropy while all n samples miss that tail with probability $(1-\varepsilon)^n$.

Certifying population distortion from a bounded witness. Fix a pair y, z and an effect $0 \leq M \leq I$ before evaluation, or select M on an independent pilot. For unit output features, $U(x) = \langle \phi(x), M\phi(x) \rangle \in [0, 1]$ and $D_{yz} \geq |\mathbb{E}_{Q_y} U - \mathbb{E}_{Q_z} U|$. From n_j iid evaluation draws at each $j \in \{y, z\}$, set

$$d_{\text{low}} = \max\{0, |\bar{U}_y - \bar{U}_z| - a_y - a_z\}, \qquad a_j = \sqrt{\log(4/\alpha)/(2n_j)}.$$

Hoeffding's inequality makes equation 5 valid with D_{yz} replaced by d_{low} with probability at least $1-\alpha$. This is a lower confidence bound on population distortion for the fixed catalogue. It supplies no entropy ranking, and fitting the effect on the evaluation draws requires further correction. Appendix B.2 gives the proof and simultaneous extension.

A population entropy interval requires additional structure. For unit features, $n \geq 2$ and a known population rank bound r, let $\widehat{H}_n = S(\rho_n)$, $B_{n,r} = \log(1 + (r-1)/n)$ and $a_{n,\alpha} = h_2(1/n)\sqrt{n\log(2/\alpha)/2}$. The interval $[\max(0, \widehat{H}_n - a_{n,\alpha}), \min(\log r, \widehat{H}_n + a_{n,\alpha} + B_{n,r})]$ covers $S(\rho)$ with probability at least $1-\alpha$. Appendix D proves the bias identity $S(\rho) - \mathbb{E}\widehat{H}_n = \mathbb{E}D(\rho_n\|\rho)$, its bound by $B_{n,r}$, and the concentration radius using established entropy identities and bounded differences (Schumacher & Westmoreland, 2000; Farnia et al., 2026). Observed Gram rank cannot replace r. A Gaussian kernel can have infinite population rank even on finite-dimensional inputs.

The interval can be too wide to resolve a comparison: for $n = 16, r = 384, \alpha = .05$, the fluctuation radius is 1.270 nats and the bias allowance 3.216 nats. Such a result calls for abstention, more samples, or additional structure. For purity, an alternative is the unbiased order-two U-statistic $\widehat{u} = \{n(n-1)\}^{-1}\sum_{i\neq j}|k(X_i, X_j)|^2$, for which $\Pr(|\widehat{u} - \operatorname{Tr}\rho^2| > b) \leq 2e^{-2\lfloor n/2\rfloor b^2}$. Its reciprocal and negative logarithm are generally biased. Purity addresses a different spectral summary, so its concentration does not settle an order-one ranking. Posterior fidelity remains a separate question.

Resolving continuous conditions requires samples within each resolution cell. A finite partition $Z = q(Y)$ pools conditional states into $\bar{\rho}_z = \mathbb{E}[\rho_Y \mid Z = z]$. In a common known r-dimensional feature space with $r \geq 2$, if all within-cell trace distances are at most δ, entropy continuity gives

$$0 \leq H_Z - H_{\text{post}} \leq F_r(\delta), \qquad F_r(\delta) = h_2(t) + t\log(r-1), \quad t = \min\{\delta, 1 - 1/r\}.$$

For $r = 1$ all these entropies vanish. A smooth encoding of the cell labels adds at most $C(p, L)$. Orthogonal labels remove that additional error. For m equally weighted cells and $n \geq 2$ independent draws from each cell's conditional mixture, $\hat{H}_Z$ has downward bias at most $\log(1 + (r-1)/n)$ and concentration radius $h_2(1/n)\sqrt{n\log(2/\alpha)/(2m)}$. Refining cells at fixed total budget $N = mn$ can reduce pooling error while increasing the bias allowance. Appendix D.3 proves the combined population interval and a consistent refinement regime under a known continuity modulus. The analysis uses standard entropy continuity (Audenaert, 2007) and concentration, with known finite feature dimension, cell weights and the stated sampling design.

Operational consequence. Fix the catalogue, condition weights and output kernel before comparing samplers. Generate repeated independent runs at identical observations. Compute blockwise entropy, same-feature purity and leading-eigenvalue baselines, and an independent fidelity diagnostic. If a joint-minus-conditioning score is used, report $C(p, L)$ alongside it. If spectra are compressed, report the retained tail mass and its approximation allowance per block. For population rankings, combine valid per-condition intervals by their fixed weights, allocating the total failure probability across conditions and samplers. Assert an ordering only when the resulting intervals are disjoint. Appendix D gives an error-spending construction for repeated sampling decisions. When the sampling or rank requirements cannot be justified, or intervals overlap, leave the population comparison unresolved.

6 Experiments with identifiable conditional targets

We first separate conditioning, compression and sampling errors using exact categorical operators and a finite image dictionary. Independent sampling controls then test ensemble bias and the distortion witness. Appendices E, F and F.4 give the complete designs and results.

Exact conditioning and truncation controls. Figure 1a changes only Gaussian conditioning overlap for two equally weighted categories. Deterministic condition-dependent outputs have zero conditional entropy but reverse their paired-score ordering against an observation-independent 95:5 mixture with entropy .198515. Panel b instead uses identical uniform two-output laws. Their full score is $\log 2$, while global top-2 truncation approaches zero as conditions separate. Direct eigendecompositions verify equation 9 on the released grid and the negative example in Section 4.

Table 1: Spectral approximation at equal retained-head budgets for the exact digit target $H_{\text{post}} = 2.000490$ with orthogonal condition features. The global joint spectrum retains T heads. Each of six blocks retains $T/6$. Tail aggregation gives a global interval using joint rank 240. Blocks need six tail-mass scalars. Global summaries need one tail mass and the marginal spectrum, here the shared condition weights.

Heads T	Global G_T	Blockwise interval	Global tail interval width
24	1.259756	$[1.411641, 2.298545]$	1.284051
48	1.722416	$[1.774628, 2.097903]$	0.442175
96	1.956890	$[1.960244, 2.010342]$	0.052850
144	1.995353	$[1.992213, 2.001694]$	0.004472
192	2.000142	$[1.999397, 2.000672]$	0.000214
240	2.000490	$[2.000490, 2.000490]$	0.000000

A fixed inverse problem on real digit images. A uniform prior on 40 real digit images, four fixed noisy pixel measurements and six fixed observations give exact finite posteriors (Alpaydin & Kaynak, 1998; Pedregosa et al., 2011). The target averages over these six observations. Under a fixed Gaussian output kernel, sweeping only observation bandwidth changes G from 2.000490 to 2.942288 nats although $H_{\text{post}} = 2.000490$ throughout (Figure 1c). At the median-distance bandwidth, deterministic MAP scores .584112, above a fixed two-image mixture's .301294 despite having zero within-condition entropy. A normalised linear output kernel also reverses this comparison (.347610 versus .071186). The fixed mixture serves only as a diversity control. Appendix E gives both complete sweeps and the observation design.

The certificate changes the evaluation decision. From the median-bandwidth kernel, the smallest integer a with $C(p, L^{\circ a}) \leq .05$ is 8, selected without outputs. Its $C = .038088$ gives intervals $[0, .010799]$ for MAP and $[.263205, .301294]$ for the pair, certifying the correct diversity ordering. The posterior interval $[1.973912, 2.012000]$ contains its exact target. The linear output kernel also certifies the ordering. Appendix E compares other tolerances and the more conservative overlap-only bound. A post hoc stress test retains all 45 mask, noise and seed settings: all reverse at the median bandwidth and all certify after sharpening. These coupled sensitivity checks test the mechanism within the same dictionary (Appendix F).

Compression and sampling have distinct effects. At $T = 96$ in Table 1, the global estimate has absolute error .043600, compared with .015198 for the blockwise interval midpoint. Global redistribution is more accurate at $T = 192$, and its tail-bin interval can be narrower at large budgets. Blockwise compression thus preserves the conditional decomposition and local tail accounting without uniformly improving accuracy. Appendix F provides both output kernels and complete budget accounting. A separate multinomial sampling experiment uses 40 independent repetitions of the six-condition design at each of seven budgets. At 16 draws per condition, the Gaussian-kernel grouped entropy has mean 1.611207 and standard deviation .089959. At 512 draws, these are 1.981323 and .016461. This downward sampling bias persists even when the conditioning target is evaluated directly.

For continuous conditioning we use $Y \sim \text{Unif}[-1, 1]$, $X \mid Y = y \sim \mathcal{N}(\sin(2y), .5^2)$ and an output Gaussian kernel of bandwidth one. Its within-condition entropy is analytically .553303. Kernel sharpening sends the score from 128 distinct paired observations to zero. With eight fixed observations and 16 draws each, it approaches .493462, exactly the groupwise empirical target. The residual difference from the population entropy illustrates why the sampling-design limit in equation 7 does not remove small-ensemble bias.

A population distortion witness on noncommuting states. Designed controls use full-rank states in dimension 32, including identical-state and zero-overlap nulls, over 36 settings with 2,000 repetitions each. At rotation $\pi/3$, overlap .8 and 256 samples per condition, the true gap is .214631 nats and the median lower 95% confidence bound is .065027. All bounds cover in these runs. Appendix F.4 includes weak-signal failures to detect positive distortion. These controls fix the witness

Table 2: Actual GenEval output banks: four images per prompt, 100 prompts. Each output kernel is a normalised feature dot product. Conditioning is the median-bandwidth text Gaussian kernel. $H = \hat{H}_{\text{group}}$. G is the unmodified score. G_a uses prompt-only power selection. The same $a = 10$ and $C = .040583$ apply to both representations and all systems. Human scores are mean caption fit on the original 1 to 4 scale, not diversity judgments or aesthetic ratings.

	CLIP			DINOv2			Caption
System	Direct H	G	G_a	Direct H	G	G_a	fit
SD 2.1	.437	1.400	.451	.891	2.877	.919	3.344
IF-XL	.414	1.441	.428	.825	2.849	.853	3.546
CLIP retrieval	.621	1.646	.638	1.028	3.068	1.061	3.104

from the known model before sampling. Retrospective learned witnesses on Qwen give much smaller positive values and carry no confirmatory coverage claim after prior bank inspection (Appendix F.5).

7 Audits of pretrained conditional generators

We test whether the mechanisms above change the evaluation of pretrained generators. Direct grouped entropy $\hat{H}_{\text{group}}$ supplies the exact within-condition target for each observed bank. Identification intervals concern that target. We distinguish initial analyses, the post hoc GenEval resolution refinement, and later studies on fresh outputs.

Released diffusion outputs with matched prompts. The GenEval human-study release (Ghosh et al., 2023) contains four images per prompt for 100 matched prompts from Stable Diffusion 2.1 and IF-XL, together with CLIP retrieval from LAION-5B as a comparator. We use all 1,200 images from the matched release. Normalised CLIP ViT-B/32 (Radford et al., 2021) and DINOv2-small (Oquab et al., 2024) features define linear and Gaussian output kernels. Conditioning uses CLIP text cosine or Gaussian kernels. Appendix G gives the kernel grid and task groups. Exact source versions are supplied with the code.

Table 2 shows a native-generator ordering reversal at a conventional bandwidth: CLIP grouped entropy gives $.436783 > .414097$ for SD 2.1 versus IF-XL, while the paired score gives $1.399684 < 1.441155$. The default DINOv2 comparison preserves its ordering. Seven of the 16 whole-bank specifications reverse. Across all seven tag groups, 31 of 112 native-generator specifications reverse. These correlated specifications measure sensitivity within the same banks. The only whole-bank DINOv2 reversal has a small margin, -2.86×10^{-5} nats, under its Gaussian output kernel and widest conditioning bandwidth. CLIP prompt-resampling ranges for the grouped difference include zero.

At the initial tolerance $\eta = .05$, input-only sharpening corrects all 31 reversed point orderings, but *certifies none of those 31 comparisons*. For the main CLIP comparison, power 10 leaves a score gap of $.022661$ nats, smaller than the shared allowance $C = .040583$, so the identification intervals overlap. The input-only tolerance controls conditioning ambiguity but cannot anticipate the output-score gap. Global truncation causes a separate error: with orthogonal prompt labels, retaining 100 global heads gives log-scores $-.000849$ for both diffusion systems, versus grouped targets $.436783$ and $.414097$. All budgets, including complete spectra, are retained.

A post hoc resolution analysis uses tolerances $.05, .01, .001$ across all 112 native-generator settings. At .01, 30 of the 31 reversed comparisons have disjoint intervals. At .001, all 31 do. These results supplement the original 0-of-31 outcome. The algebraic bound holds for every admissible kernel, so fixed-bank certificates require no statistical multiplicity correction. Their post hoc selection still precludes a claim of prospective performance or inference about unseen outputs.

Amended fresh-image decision illustration. After a stopped pilot, a declared amendment generates 384 images on 24 fresh fixed prompts, with two four-sample banks per model (Appendix I). On bank A, direct scoring and input-only sharpening at $\eta = .01$ choose SD 2.1, while the native score chooses IF-XL. Sharpening certifies this choice on A. Frozen before B feature analysis, the choices yield a $.007782$-nat direct CLIP-diversity gain on B. This is one aggregate choice per rule, with the same

Table 3: Qwen3-0.6B banks with fixed CLIP output features and median-bandwidth Gaussian prompt conditioning. Initial: 16 prompts, eight samples per temperature, $a = 6$, $C = .031825$. Confirmation: 32 new prompts, 64 samples per temperature, $a = 9$, $C = .005208$. Greedy rows use actual deterministic completions. Temperatures do not define a quality ordering.

Study	Decoding	Direct H	G	G_a
Initial	Temperature .7	.356	.959	.373
	Temperature 1.1	.407	.994	.423
	Greedy control	0	.621	.017
Confirmation	Temperature .7	.513	1.274	.516
	Temperature 1.1	.623	1.356	.625
	Greedy control	0	.850	.003

gain from direct scoring. On B, sharpening also corrects the native point reversal, but its certificate abstains and the population difference intervals include zero. The observed gain concerns feature diversity. Image quality and repeated-study policy reliability remain untested.

Fresh inference and prospective confirmation. The initial Qwen3-0.6B study (Qwen Team, 2025a;b) uses 16 prompts, eight completions per temperature and 16 greedy controls. All 24 declared temperature comparisons preserve their ordering. The main CLIP score inflates the zero-entropy greedy target by .621465 nats. Input-only sharpening at $\eta = .05$ certifies the existing bank ordering (Table 3). Appendix G gives all specifications, intervals and termination counts.

The prospective follow-up fixes $\eta = .01$ and one temperature contrast before generation on 32 new prompts (Appendix H). It produces 4,096 stochastic completions, 64 per temperature and prompt, plus 32 greedy controls. Prompt-only selection chooses power 9. The direct difference is .109649 nats, and the selected interval $[.104373, .114789]$ certifies the finite-bank direction. Neither score reverses the temperature ranking. Both disjoint 32-sample halves certify it. The greedy score falls from .849561 to .002853, with direct target zero.

The simultaneous 95% population difference interval $[-2.422800, 2.642099]$ remains uninformative under independent sampling and known feature rank at most 512. Its target averages over the 32 fixed prompts. Larger banks and prospective tolerance selection therefore strengthen the finite-bank evidence without resolving population ranking.

The published full-score implementation reproduces the GenEval reversals (Appendix J), locating the discrepancy in the statistic itself. Direct scoring is 24 to 31 times faster in our same-feature CPU benchmark (Appendix G.5). The certificate's role is to determine whether the smooth score supports the same within-condition conclusion. GenEval's caption-fit ratings assess a separate property: the model with higher diversity need not have better prompt alignment.

8 Discussion and Limitations

A joint-minus-conditioning score retains variation between conditional states when their condition features overlap. Our certificate quantifies this ambiguity, the equality criterion identifies when it vanishes, and the recovery limits show what an aggregate score can determine. Global truncation introduces a separate rank-budget obstruction even with exact condition labels. These mechanisms explain why a score can favour a deterministic generator or reverse a within-condition diversity ranking.

The evaluation protocol follows the available information. Use direct grouped entropy for labelled repeated outputs, a conditioning certificate when interpreting a joint-kernel score, and conditional tail accounting when compressing spectra. Population claims require a sampling guarantee in addition to these deterministic calculations. Our pretrained studies establish finite-bank effects, including a small held-out selection gain shared by direct scoring, while population rankings remain unresolved. Continuous pooling requires known finite feature dimension and a continuity modulus. Finally, every entropy result is relative to the chosen representation. Posterior fidelity and generation quality require separate evaluation.

AI USE STATEMENT

The author used OpenAI ChatGPT and Codex for literature retrieval, challenging research ideas, mathematical formulation and proof checking, coding, manuscript drafting. The author independently derived the proofs presented in the appendices, produced the numerical analyses using a separate implementation, verified data provenance and hashes, and reviewed every claim. The generative models evaluated in this work produced the identified image and text samples. AI-assisted code also generated the declared synthetic controls. No experimental results were fabricated or selectively altered. AI systems are not authors and cannot be held accountable. The author takes full responsibility for the content, claims, code and remaining errors.

REPRODUCIBILITY STATEMENT

The manuscript appendices contain complete proofs and experimental protocols. The accompanying anonymous code archive is organised as a reproducibility repository, with a single CPU entry point for numerical replay from the supplied feature arrays and exact-control inputs. It includes every reported outcome, frozen analysis plans and source snapshots. Its provenance manifest records exact dataset and checkpoint revisions, checksums, software environments and compute accounting. Separate instructions cover original generation and feature extraction, which require model downloads and, for the image studies, GPU resources. The fresh-image archive supplies features and metadata rather than image pixels. Appendix G distinguishes GenEval's initial analysis from the post hoc refinement. Appendices H and I retain the Qwen confirmation and the stopped image pilot with its subsequent amendment. No outcome-dependent exclusion is introduced by this presentation.

REFERENCES

E. Alpaydin and C. Kaynak. Optical recognition of handwritten digits, 1998. URL https://doi.org/10.24432/C50P49. Dataset, CC BY 4.0.

Pietro Astolfi, Marlène Careil, Melissa Hall, Oscar Mañas, Matthew Muckley, Jakob Verbeek, Adriana Romero Soriano, and Michal Drozdzal. Consistency-diversity-realism pareto fronts of conditional image generative models, 2024. URL https://arxiv.org/abs/2406.10429.

Koenraad M. R. Audenaert. A sharp continuity estimate for the von Neumann entropy. *Journal of Physics A: Mathematical and Theoretical*, 40(28):8127–8136, 2007. doi: 10.1088/1751-8113/40/28/S18. URL https://arxiv.org/abs/quant-ph/0610146.

Francis Bach. Information theory with kernel methods, 2022. URL https://arxiv.org/abs/2202.08545.

Farzan Farnia, Mohammad Jalali, and Azim Ospanov. Exposing diversity bias in deep generative models: Statistical origins and correction of diversity error, 2026. URL https://arxiv.org/abs/2602.14682.

Dan Friedman and Adji Bousso Dieng. The vendi score: A diversity evaluation metric for machine learning, 2022. URL https://arxiv.org/abs/2210.02410.

Dhruba Ghosh, Hanna Hajishirzi, and Ludwig Schmidt. GenEval: An object-focused framework for evaluating text-to-image alignment, 2023. URL https://arxiv.org/abs/2310.11513.

Patrick Hayden, Richard Jozsa, Dénes Petz, and Andreas Winter. Structure of states which satisfy strong subadditivity of quantum entropy with equality. *Communications in Mathematical Physics*, 246:359–374, 2004. doi: 10.1007/s00220-004-1049-z. URL https://arxiv.org/abs/quant-ph/0304007.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. doi: 10.1080/01621459.1963.10500830.

Mohammad Jalali, Azim Ospanov, Amin Gohari, and Farzan Farnia. Conditional vendi score: Prompt-aware diversity evaluation for generative ai models and llms. In *Proceedings of the 29th International Conference on Artificial Intelligence and Statistics*, volume 300 of *Proceedings of*

Machine Learning Research, pp. 1369–1377. PMLR, 2026. URL https://proceedings.mlr.press/v300/jalali26a.html.

Nithish Kannen, Arif Ahmad, Marco Andreetto, Vinodkumar Prabhakaran, Utsav Prabhu, Adji Bousso Dieng, Pushpak Bhattacharyya, and Shachi Dave. Beyond aesthetics: Cultural competence in text-to-image models, 2024. URL https://arxiv.org/abs/2407.06863.

Masato Koashi and Nobuyuki Imoto. Operations that do not disturb partially known quantum states. *Physical Review A*, 66:022318, 2002. doi: 10.1103/PhysRevA.66.022318. URL https://arxiv.org/abs/quant-ph/0101144.

Quan Nguyen and Adji Bousso Dieng. Vendi information gain: An alternative to mutual information for science and machine learning, 2025. URL https://arxiv.org/abs/2505.09007.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision, 2024. URL https://arxiv.org/abs/2304.07193.

Azim Ospanov and Farzan Farnia. Do vendi scores converge with finite samples? truncated vendi score for finite-sample convergence guarantees. In *Proceedings of the Forty-first Conference on Uncertainty in Artificial Intelligence*, volume 286 of *Proceedings of Machine Learning Research*, pp. 3272–3299. PMLR, 2025. URL https://proceedings.mlr.press/v286/ospanov25a.html.

Amey P. Pasarkar and Adji Bousso Dieng. Cousins of the vendi score: A family of similarity-based diversity metrics for science and machine learning, 2023. URL https://arxiv.org/abs/2310.12952.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. URL https://jmlr.org/papers/v12/pedregosa11a.html.

Qwen Team. Qwen3-0.6B: Official model card and checkpoint. Hugging Face, 2025a. URL https://huggingface.co/Qwen/Qwen3-0.6B. Accessed 10 September 2026.

Qwen Team. Qwen3 technical report, 2025b. URL https://arxiv.org/abs/2505.09388.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 8748–8763. PMLR, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Mary Beth Ruskai. Inequalities for quantum entropy: A review with conditions for equality. *Journal of Mathematical Physics*, 43:4358–4375, 2002. doi: 10.1063/1.1497701. URL https://arxiv.org/abs/quant-ph/0205064. Erratum: volume 46, 019901 (2005).

Luis G. Sánchez Giraldo, Murali Rao, and José C. Príncipe. Measures of entropy from data using infinitely divisible kernels, 2014. URL https://arxiv.org/abs/1211.2459. arXiv version 3. First submitted in 2012.

Benjamin Schumacher and Michael D. Westmoreland. Relative entropy in quantum information theory, 2000. URL https://arxiv.org/abs/quant-ph/0004045.

Youqi Wu and Farzan Farnia. Do large language models capture the diversity in their training data?, 2026. URL https://arxiv.org/abs/2609.02275. Version 1, 2 September 2026.

Guide to the appendices. Appendix B gives the conditioning certificate, equality criterion and aggregate-score recovery limits. Appendix C checks the published truncation theorem against a finite Gaussian example. Appendix D separates population sampling error from conditioning error and treats continuous conditions. Complete generator results and analysis histories are in Appendices G–I.

A Operator assumptions and consistency

A.1 Standing assumptions and normalisation

Let $\mathcal{H}$ be a separable real or complex Hilbert space and let $\phi : \mathcal{X} \to \mathcal{H}$ be strongly measurable. The kernel is $k(x, x') = \langle \phi(x), \phi(x') \rangle$. We assume

$$0 < \tau := \mathbb{E}\|\phi(X)\|^2 < \infty. \tag{12}$$

The rank-one operator $\phi(X)\phi(X)^*$ is strongly measurable in the separable Banach space $\mathcal{S}_1(\mathcal{H})$ and has trace norm $\|\phi(X)\|^2$. Consequently,

$$C = \mathbb{E}[\phi(X)\phi(X)^*], \qquad \rho = C/\tau$$

are well-defined Bochner integrals, $C \geq 0$, $\operatorname{Tr} C = \tau$, and ρ is a density operator. Strong measurability cannot in general be replaced merely by measurability of the scalar kernel and its diagonal. For conditional distributions, $\mathcal{X}$ and $\mathcal{Y}$ are standard Borel spaces. Statements about $P_{X|Y=y}$ refer to a regular conditional law and hold for P_Y-almost every admissible y.

For independent and identically distributed observations $X_1, X_2, \ldots$, define

$$C_n = \frac{1}{n}\sum_{i=1}^{n} \phi(X_i)\phi(X_i)^*, \quad \tau_n = \operatorname{Tr} C_n, \quad \rho_n = C_n/\tau_n$$

whenever $\tau_n > 0$. This last event holds eventually almost surely. For finite features $z_i \in \mathbb{R}^d$, use $\widehat{\rho} = (\sum_i z_i z_i^*)/\sum_i \|z_i\|^2$. The divisor n suffices only for unit-norm features. The same trace normalisation is required for Nyström features of nonconstant norm.

Theorem A.1 (Strong consistency at every entropy value)**.** *Under equation 12, $\|\rho_n - \rho\|_1 \to 0$ almost surely and*

$$S(\rho_n) \longrightarrow S(\rho) \quad \textit{almost surely}$$

in $[0, \infty]$, where $S(\rho) = -\operatorname{Tr}(\rho \log \rho)$. In particular, finite population entropy is sufficient for almost-sure consistency of $\exp S(\rho_n)$. No separate condition on empirical entropy tails is required.

Proof. The strong law for integrable random variables in a separable Banach space gives $\|C_n - C\|_1 \to 0$ almost surely. Hence $\tau_n \to \tau > 0$ and

$$\|\rho_n - \rho\|_1 \leq \frac{\|C_n - C\|_1}{\tau_n} + \frac{|\tau_n - \tau|}{\tau_n} \longrightarrow 0.$$

Let (λ_j, e_j) enumerate the strictly positive eigenpairs of ρ and let Q project onto its kernel. Since $\mathbb{E}\|Q\phi(X)\|^2 = \operatorname{Tr}(QC) = 0$, $\phi(X_i)$ belongs to the support of ρ simultaneously for all i outside a single null set. Define the nonnegative measurable quadratic form

$$g(X) = \sum_j (-\log \lambda_j)|\langle e_j, \phi(X)\rangle|^2.$$

This increasing spectral sum defines the quadratic form without requiring $\phi(X)$ to belong to the operator domain of the unbounded operator $-\log \rho$. Tonelli's theorem gives $\mathbb{E}g(X) = \tau \sum_j -\lambda_j \log \lambda_j = \tau S(\rho)$. If $S(\rho) < \infty$, the scalar strong law therefore implies

$$\operatorname{Tr}[\rho_n(-\log \rho)] = \frac{n^{-1}\sum_{i=1}^{n} g(X_i)}{\tau_n} \longrightarrow S(\rho) \quad \text{almost surely.}$$

Every ρ_n has finite rank and finite entropy. Its displayed cross entropy is finite almost surely. Nonnegativity of quantum relative entropy then gives $S(\rho_n) \leq \operatorname{Tr}[\rho_n(-\log \rho)]$, and thus $\limsup_n S(\rho_n) \leq S(\rho)$.

For the lower bound, trace-norm convergence implies coordinatewise convergence of the ordered eigenvalues. Applying Fatou's lemma to the nonnegative function $h(u) = -u \log u$, $h(0) = 0$, yields $S(\rho) \le \liminf_n S(\rho_n)$. This establishes the finite-entropy result. If $S(\rho) = \infty$, the same Fatou inequality forces $S(\rho_n) \to \infty$ in the extended-real sense. The cross-entropy argument is not needed. □

Remark A.2 (Scope of the consistency result). The theorem concerns iid averages from one fixed law. Entropy can still be discontinuous under arbitrary trace-norm perturbations, and no distribution-uniform convergence rate follows. Kernel approximation, changing feature maps, dependent sampler output and changing solver laws require additional assumptions. At a fixed observation y, iid draws from $P_{X|Y=y}$ or a specified solver law Q_y converge to their respective entropy targets. Identifying the latter with posterior entropy requires $Q_y = P_{X|Y=y}$.

Proposition A.3 (Hilbert-Schmidt continuity for every $q > 1$)**.** *For any density operators ρ, σ and any $q > 1$, writing $\delta = \|\rho - \sigma\|_2$,*

$$\big|\|\rho\|_q - \|\sigma\|_q\big| \le \|\rho - \sigma\|_q \le \begin{cases} 2^{(2-q)/q}\delta^{2(q-1)/q}, & 1 < q < 2, \\ \delta, & q \ge 2. \end{cases}$$

It follows that Hilbert-Schmidt convergence implies convergence of $H_q(\rho) = (1-q)^{-1} \log \mathrm{Tr}(\rho^q)$ and $\mathrm{VS}_q(\rho) = \exp H_q(\rho)$ for every $q > 1$.

Proof. For $q \ge 2$, monotonicity of Schatten norms gives $\|\rho - \sigma\|_q \le \|\rho - \sigma\|_2$. For $1 < q < 2$, Schatten interpolation and $\|\rho - \sigma\|_1 \le 2$ give

$$\|\rho - \sigma\|_q \le \|\rho - \sigma\|_1^{(2-q)/q} \|\rho - \sigma\|_2^{2(q-1)/q}.$$

The first inequality is the reverse triangle inequality. Every fixed density operator has $\|\rho\|_q > 0$, so continuity of $\log u$ and $u^{q/(1-q)}$ at this positive number proves the claim. More quantitatively,

$$|H_q(\rho) - H_q(\sigma)| \le \frac{q}{q-1} \frac{\|\rho - \sigma\|_q}{\min(\|\rho\|_q, \|\sigma\|_q)}.$$

The denominator explains why a uniform entropy-error constant does not follow from this argument when effective dimension is unrestricted. □

Remark A.4 (The boundary $q = 1$ and orders below one). Order two is special because its score is inverse purity and can be estimated from kernel pairs without an eigendecomposition. It is not the only Hilbert-Schmidt-continuous order. For $q < 1$, $\mathrm{Tr}(\rho^q)$ may be infinite even if $S(\rho) < \infty$. Finite entropy guarantees the right-hand limit $H_q(\rho) \to S(\rho)$ as $q \downarrow 1$, but not a two-sided limit. A sufficient additional condition for a two-sided finite limit is $\mathrm{Tr}(\rho^{q_0}) < \infty$ for some $q_0 < 1$.

B Conditioning comparison and resolution

B.1 Two conditional targets and a sharp comparison

Assume standard Borel $\mathcal{X}, \mathcal{Y}$ and strongly measurable unit-norm feature maps $\phi : \mathcal{X} \to \mathcal{H}_A$ and $\psi : \mathcal{Y} \to \mathcal{H}_B$ into separable Hilbert spaces. Let $y \mapsto Q_y$ be a measurable probability kernel. Write

$$\rho_y = \mathbb{E}[\phi(X)\phi(X)^* \mid Y = y], \quad \sigma_y = \psi(y)\psi(y)^*, \quad \rho_{AB} = \mathbb{E}_Y[\rho_Y \otimes \sigma_Y],$$

with marginals $\rho_A = \mathbb{E}\rho_Y$ and $\rho_B = \mathbb{E}\sigma_Y$. We assume $S(\rho_A), S(\rho_B) < \infty$. The two finite quantities of interest are

$$H_{\mathrm{post}} = \mathbb{E}_Y S(\rho_Y), \qquad H_{\mathrm{Gram}} = S(\rho_{AB}) - S(\rho_B).$$

Concavity bounds H_{post} by $S(\rho_A)$ and subadditivity bounds $S(\rho_{AB})$ by $S(\rho_A) + S(\rho_B)$, ensuring that both entropy differences are well defined. The second quantity is the population target of the order-one joint-minus-marginal Gram construction of Sánchez Giraldo et al. (2014) and can differ from the first quantity. Bach (2022) already noted this failure of the ordinary conditional chain rule. The comparison below quantifies its direction and, for categorical conditioning, supplies a sharp observable certificate.

Theorem B.1 (Conditioning representation gap)**.** *Under the preceding assumptions,*

$$0 \le H_{\mathrm{Gram}} - H_{\mathrm{post}} = \big\{S(\rho_A) - \mathbb{E}S(\rho_Y)\big\} - I(A:B) \le S(\rho_A) - H_{\mathrm{post}}, \tag{13}$$

where $I(A:B) = D(\rho_{AB}\|\rho_A \otimes \rho_B)$*. If* Y *takes* m *values with probabilities* $p_y > 0$*, then*

$$0 \le H_{\mathrm{Gram}} - H_{\mathrm{post}} \le H(p) - S(\rho_B) =: C(p,\psi). \tag{14}$$

The upper endpoint is attained whenever the supports of ρ_y *are pairwise orthogonal, and the lower endpoint is attained when all* ρ_y *coincide. In particular, the certificate is optimal over all finite-entropy density-operator ensembles on* $\mathcal{H}_A$ *when* $\dim \mathcal{H}_A \ge m$*. Orthogonal observation features give equality* $H_{\mathrm{Gram}} = H_{\mathrm{post}}$ *for every posterior ensemble. Conversely, if* $\dim \mathcal{H}_A \ge m$*, equality for every finite-entropy density-operator ensemble on* $\mathcal{H}_A$ *implies that the observation feature vectors are pairwise orthogonal. If one instead quantifies only over laws under a fixed feature map* ϕ*, this converse requires that* ϕ *can realize* m *pairwise orthogonal unit states.*

Proof. Joint convexity of relative entropy, including its integral version, gives

$$\begin{aligned} I(A:B) &= D(\mathbb{E}[\rho_Y \otimes \sigma_Y]\|\mathbb{E}[\rho_A \otimes \sigma_Y]) \\ &\le \mathbb{E}D(\rho_Y \otimes \sigma_Y\|\rho_A \otimes \sigma_Y) = \mathbb{E}D(\rho_Y\|\rho_A) \\ &= S(\rho_A) - \mathbb{E}S(\rho_Y). \end{aligned}$$

These standard quantum-information facts are consequences of strong subadditivity. See Ruskai (2002). The integral inequality follows by finite measurable approximation and lower semicontinuity of relative entropy, or directly from data processing for the classical preparation channel $y \mapsto \sigma_y$. Finite marginal entropies justify the last identity and ensure the quantities are finite. Subtracting from $S(\rho_A)$ proves the lower bound and the identity in equation 13. Nonnegativity of $I(A:B)$ proves its upper bound.

For finite Y, the upper entropy-of-mixing inequality and purity of σ_y imply

$$S\left(\sum_y p_y \rho_y \otimes \sigma_y\right) \le H(p) + \sum_y p_y S(\rho_y \otimes \sigma_y) = H(p) + \sum_y p_y S(\rho_y).$$

Subtract $S(\rho_B) + H_{\mathrm{post}}$ to obtain equation 14. If the supports of ρ_y are orthogonal, the product states have orthogonal supports, and the entropy of this direct sum is exactly $H(p) + \sum_y p_y S(\rho_y)$. If $\rho_y = \rho_0$ for all y, then $\rho_{AB} = \rho_0 \otimes \rho_B$, giving zero gap. If the $\psi(y)$ are orthogonal, direct-sum entropy gives $S(\rho_{AB}) = H(p) + H_{\mathrm{post}}$ and $S(\rho_B) = H(p)$.

For the converse choose mutually orthogonal pure ρ_y. Their gap is $H(p) - S(\rho_B)$. Let $L_{yz} = \langle\psi(y), \psi(z)\rangle$, $D_p = \mathrm{diag}(p)$ and $G = D_p^{1/2} L D_p^{1/2}$. The nonzero spectra of G and ρ_B agree and $G_{yy} = p_y$, so

$$H(p) - S(\rho_B) = D(G\|D_p).$$

This vanishes exactly when $G = D_p$, which, since $p_y > 0$, is equivalent to $L_{yz} = 0$ for all $y \ne z$. □

Remark B.2 (Information interpretation). For finite Y, introduce the orthogonal classical label register and $\omega_{YAB} = \sum_y p_y |y\rangle\langle y| \otimes \rho_y \otimes \sigma_y$. Direct-sum entropy gives $H_{\mathrm{Gram}} - H_{\mathrm{post}} = I(A:Y \mid B)_\omega$ and $C(p,\psi) = H(Y \mid B)_\omega$. Thus the gap measures information about posterior features lost when the observation is represented by a possibly overlapping kernel feature state. It can occur even with exact posterior sampling: changing only the observation kernel can change H_{Gram} while every posterior law and H_{post} remain fixed.

Corollary B.3 (A computable overlap bound)**.** *For the categorical setting, let* $\epsilon = \max_{y \ne z} |L_{yz}|$*. Then*

$$C(p,\psi) = D(G\|D_p) \le \log\left(\sum_z p_z \sum_y |L_{yz}|^2\right) \le \log\big(1 + (m-1)\epsilon^2\big). \tag{15}$$

The exact certificate is evaluable from the label probabilities and the $m \times m$ *observation Gram matrix, without posterior samples.*

Proof. For any finite-dimensional states R, T, with $T > 0$, spectral decompositions $R = \sum_i r_i u_i u_i^*$ and $T = \sum_j t_j v_j v_j^*$ and Jensen's inequality for log give

$$D(R\|T) = \sum_{i,j} r_i |\langle u_i, v_j\rangle|^2 \log(r_i/t_j)$$
$$\leq \log\left(\sum_{i,j} \frac{r_i^2}{t_j} |\langle u_i, v_j\rangle|^2\right) = \log \mathrm{Tr}(R^2 T^{-1}).$$

Terms with $r_i = 0$ have zero weight. Apply this to (G, D_p) and use

$$\mathrm{Tr}(G^2 D_p^{-1}) = \sum_z p_z \sum_y |L_{yz}|^2 \leq 1 + (m-1)\epsilon^2.$$

□

Example B.4 (Exact binary discrepancy)**.** Let Y be uniform on two categories, let $X = Y$, and use orthogonal unit features for X. Choose unit observation features with overlap $c \in [0, 1]$. Then every posterior is a point mass, $H_{\mathrm{post}} = 0$, whereas

$$H_{\mathrm{Gram}} = \log 2 - h_2((1+c)/2), \qquad h_2(t) = -t \log t - (1-t)\log(1-t).$$

This discrepancy ranges from zero at $c = 0$ to $\log 2$ at $c = 1$ and attains the certificate exactly. It remains present with unlimited paired data, since it is a difference of population targets.

Remark B.5 (Nonconstant feature norms). If only $0 < \tau = \mathbb{E}\|\phi(X)\|^2 < \infty$ is assumed, let $\tau_y = \mathbb{E}[\|\phi(X)\|^2 \mid Y = y]$ and normalize each nonzero conditional operator. The normalized joint operator then uses the tilted observation law $d\widetilde{P}_Y(y) = (\tau_y/\tau) dP_Y(y)$: $\rho_{AB} = \int \rho_y \otimes \sigma_y \, d\widetilde{P}_Y(y)$. All comparisons above hold with this same tilted law throughout. Using the original observation weights for conditional entropy while normalizing the joint operator globally generally destroys the identity. Unit-diagonal kernels avoid this change of estimand.

B.2 Output separation, exact equality, and a population witness

For a finite catalogue, write

$$\Delta = G - H_{\mathrm{post}} = I(A : Y \mid B), \qquad c_{yz} = |\langle \psi(y), \psi(z)\rangle|, \qquad D_{yz} = \tfrac{1}{2}\|\rho_y - \rho_z\|_1.$$

Define the continuous function

$$\lambda(c) = 1 - \frac{1-c^2}{c} \operatorname{atanh} c \quad (0 < c < 1), \qquad \lambda(0) = 0, \quad \lambda(1) = 1. \tag{16}$$

The finite-catalogue assumptions and finite marginal entropy are those of Appendix B. No finite-dimensional output-space assumption is needed below.

Theorem B.6 (Output-dependent conditioning obstruction)**.** *For every pair of distinct positive-probability conditions,*

$$\Delta \geq 2\lambda(c_{yz}) \frac{p_y p_z}{p_y + p_z} D_{yz}^2. \tag{17}$$

The right sides may be summed over any collection of disjoint pairs. For a binary catalogue, the coefficient $2\lambda(c)$ multiplying $p_y p_z D_{yz}^2$ cannot be increased uniformly over priors and output ensembles. Moreover, $\Delta = 0$ if and only if the conditional states are constant on each connected component of the graph with edges $c_{yz} > 0$.

Proof. First restrict to two conditions with normalised probabilities $q, 1-q$ and overlap c. For $h(q) = -q \log q - (1-q)\log(1-q)$ put

$$r = \sqrt{1 - 4(1-c^2)q(1-q)}, \qquad f_c(q) = h(q) - h((1+r)/2).$$

The eigenvalues of $q\sigma_y + (1-q)\sigma_z$ are $(1 \pm r)/2$, so $f_c(q) = H(Y \mid B)$. The function $f_c - \lambda(c)h$ is concave. To see this, for $0 < c < 1$ and $0 < q < 1$ direct differentiation gives

$$[-f_c''(q)]q(1-q) = \frac{c^2}{r^2}\lambda(r). \tag{18}$$

The series

$$\frac{\lambda(t)}{t^2} = \sum_{n\geq 0} \frac{2t^{2n}}{(2n+1)(2n+3)}$$

is increasing on $[0,1)$, and $r \geq c$. Thus $-f_c''(q) \geq \lambda(c)/[q(1-q)] = -\lambda(c)h''(q)$. Continuity handles prior endpoints. For $c=0$, $f_c = 0$. For $c=1$, $f_c = h$.

Apply a two-outcome measurement to A, using the positive spectral projection of $\rho_y - \rho_z$ as its first effect, and call its outcome Z. This gives Bernoulli probabilities a, b with $|a-b| = D_{yz}$. Conditional mutual information cannot increase under this local operation. If $w_z = \Pr(Z=z)$ and $q_z = \Pr(Y=y \mid Z=z)$, then

$$\begin{aligned} I(A:Y \mid B) &\geq I(Z:Y \mid B) \\ &= f_c(q) - \sum_z w_z f_c(q_z) \\ &\geq \lambda(c)\left[h(q) - \sum_z w_z h(q_z)\right] = \lambda(c) I(Z:Y). \end{aligned}$$

The last inequality uses concavity of $f_c - \lambda(c)h$ and $\sum_z w_z q_z = q$. Zero-probability outcomes are omitted. Writing $v = qa + (1-q)b$, Pinsker's inequality in nats gives

$$\begin{aligned} I(Z:Y) &= qD(\mathrm{Ber}(a)\|\,\mathrm{Ber}(v)) + (1-q)D(\mathrm{Ber}(b)\|\,\mathrm{Ber}(v)) \\ &\geq 2q(a-v)^2 + 2(1-q)(b-v)^2 \\ &= 2q(1-q)(a-b)^2. \end{aligned}$$

The case $v \in \{0,1\}$ forces $a=b$ and is trivial.

For a general catalogue, let T record the chosen pair as one block and every other condition as a singleton. The chain rule implies

$$I(A:Y \mid B) = I(A:T \mid B) + I(A:Y \mid B,T) \geq (p_y + p_z) I(A:Y \mid B, T = \{y,z\}).$$

Apply the binary result with $q = p_y/(p_y+p_z)$ to obtain equation 17. A partition into disjoint pairs and singletons proves the summed bound. These arguments involve only a binary output measurement and finite classical and conditioning registers, so they apply to separable output spaces under the stated finite-entropy assumptions.

For sharpness, take $q = 1/2$ and the commuting qubit states

$$\rho_0 = \mathrm{diag}(1/2+\varepsilon, 1/2-\varepsilon), \qquad \rho_1 = \mathrm{diag}(1/2-\varepsilon, 1/2+\varepsilon).$$

Then $D_{01} = 2\varepsilon$, and measuring the common eigenbasis loses no information. Consequently

$$\Delta = f_c(1/2) - f_c(1/2+\varepsilon) = 2\lambda(c)\varepsilon^2 + o(\varepsilon^2),$$

whereas $q(1-q)D_{01}^2 = \varepsilon^2$. This establishes asymptotic sharpness of the uniform coefficient, not pointwise optimality at every prior or finite separation.

Since $\lambda(c) > 0$ for $c > 0$, zero gap forces $D_{yz} = 0$ on every edge, and hence equality of states along paths. Conversely, let the components be C_j, their masses q_j, their common output states τ_j, and $B_j = q_j^{-1} \sum_{y \in C_j} p_y \sigma_y$. Different components have orthogonal conditioning spans, giving

$$\rho_B = \bigoplus_j q_j B_j, \qquad \rho_{AB} = \bigoplus_j q_j (\tau_j \otimes B_j).$$

Direct-sum entropy yields $G = \sum_j q_j S(\tau_j) = H_{\mathrm{post}}$. □

Remark B.7 (Relation to established equality theory). The graph condition specialises quantum Markov equality and the principle that one cannot obtain information about nonorthogonal pure states while preserving them. In finite dimensions, Petz recovery gives another short proof: $I(A:Y \mid B) = 0$ implies a channel $\mathcal{R}: B \to AB$ with $\mathcal{R}(\sigma_y) = \rho_y \otimes \sigma_y$ for every positive-probability label (Hayden et al., 2004). Root-fidelity monotonicity then gives $c_{yz} \leq c_{yz} F(\rho_y, \rho_z)$, forcing $\rho_y = \rho_z$ when $c_{yz} > 0$. The general invariant-state decomposition is due to Koashi & Imoto (2002). Our use is an explicit pure-conditioning specialisation. The proof above supplies the quantitative coefficient directly.

Corollary B.8 (A bounded-output population witness)**.** *Fix a pair y, z and an effect $0 \le M \le I$ independently of the evaluation outputs, and define $U(x) = \langle \phi(x), M\phi(x) \rangle \in [0,1]$. Draw $n_y, n_z \ge 1$ independent outputs from their specified conditional laws. For $0 < \alpha < 1$, let*

$$a_j = \sqrt{\frac{\log(4/\alpha)}{2n_j}}, \qquad d_{\text{low}} = \left[|\overline{U}_y - \overline{U}_z| - a_y - a_z \right]_+.$$

With probability at least $1 - \alpha$,

$$\Delta \ge 2\lambda(c_{yz}) \frac{p_y p_z}{p_y + p_z} d_{\text{low}}^2. \tag{19}$$

For J fixed pair/effect witnesses, replacing $\log(4/\alpha)$ by $\log(4J/\alpha)$ makes all their bounds simultaneous. Their maximum or a sum over selected disjoint pairs may then be reported.

Proof. Writing $\mu_j = \mathbb{E}U(X) = \mathrm{Tr}(M\rho_j)$, the variational formula for trace distance gives $D_{yz} \ge |\mu_y - \mu_z|$. Hoeffding's inequality (Hoeffding, 1963) bounds each event $|\overline{U}_j - \mu_j| > a_j$ by $\alpha/2$. On their joint complement, the triangle inequality gives $D_{yz} \ge d_{\text{low}}$. Apply Theorem B.6. For J witnesses, a union bound over their $2J$ means proves the simultaneous claim. Shared samples across witnesses do not require independence between these events. □

The witness may be chosen on an independent pilot, with validity conditional on that pilot and therefore also unconditional. Choosing an arbitrary effect from the evaluation outputs or using outcome-dependent stopping requires an additional argument. The feature map, weights and overlaps must be fixed. The weights define the population target and can differ from the stratified sample fractions. The certificate concerns population representation distortion for this catalogue. Population entropy, rankings, semantic fidelity and generalisation to unseen conditions require separate evidence.

If several effects target one pair, take the maximum valid bound for that pair before summing over disjoint pairs. Overlapping pair bounds cannot simply be added. Zero-probability labels are removed before constructing the graph. An estimated weight or overlap requires its own conservative allowance.

B.3 Observation resolution and sampling

For $m = 1$, the conditioning certificate is zero and no sharpening is needed. In the following theorem assume $m \ge 2$.

Theorem B.9 (Tensor sharpening of categorical conditioning)**.** *In the finite categorical setting of Theorem B.1, assume $\epsilon = \max_{y \ne z} |\langle \psi(y), \psi(z) \rangle| < 1$. For each integer $t \ge 1$, replace the observation feature by $\psi_t(y) = \psi(y)^{\otimes t}$, equivalently its kernel by $\ell_t(y,z) = \ell(y,z)^t$. Let $H_{\text{Gram},t}$ denote the resulting conditional entropy. Then*

$$H_{\text{post}} \le H_{\text{Gram},t+1} \le H_{\text{Gram},t}, \qquad 0 \le H_{\text{Gram},t} - H_{\text{post}} \le \log\left(1 + (m-1)\epsilon^{2t}\right).$$

In particular $H_{\text{Gram},t} \downarrow H_{\text{post}}$. For $m > 1$ and a desired bias tolerance $\eta > 0$, any integer $t \ge 1$ satisfying

$$t \ge \frac{\log((m-1)/(e^\eta - 1))}{2\log(1/\epsilon)}$$

suffices when $0 < \epsilon < 1$. $\epsilon = 0$ already has zero bias.

Proof. Tracing out one tensor factor maps the joint state on AB^{t+1} to that on AB^t, and similarly for its product marginals. Data processing gives $I(A : B^{t+1}) \ge I(A : B^t)$. Since $H_{\text{Gram},t} = S(\rho_A) - I(A : B^t)$, the sequence is nonincreasing. Theorem B.1 gives its lower bound. The observation Gram overlaps are L_{yz}^t, so Corollary B.3 gives the stated error bound. Its limit is zero. Solving $(m-1)\epsilon^{2t} \le e^\eta - 1$ yields the sufficient choice of t. □

Proposition B.10 (Sharpening a fixed paired sample)**.** *Let $(x_i, y_i)_{i=1}^n$ be a fixed paired sample, let $K_{ij} = k(x_i, x_j)$ and $L_{ij} = \ell(y_i, y_j)$ be unit-diagonal PSD Gram matrices, and assume $|\ell(y,z)| < 1$*

for distinct observed labels. Write $n_y = \#\{i : y_i = y\}$ and let K_y be the corresponding principal submatrix. For integer $t \to \infty$,

$$S\left(\frac{K \circ L^{\circ t}}{n}\right) - S\left(\frac{L^{\circ t}}{n}\right) \longrightarrow \sum_{y:n_y>0} \frac{n_y}{n} S(K_y/n_y). \tag{20}$$

If every y_i occurs once, the limit is zero for every K.

Proof. The Schur product theorem makes all displayed matrices PSD, and their diagonal entries are $1/n$, so their traces are one. Order samples by their observed labels. Entrywise, $L^{\circ t}$ tends to the block diagonal matrix with blocks $\mathbf{1}_{n_y}\mathbf{1}_{n_y}^*$, while $K \circ L^{\circ t}$ tends to the block diagonal matrix with blocks K_y. In fixed dimension, entropy is continuous. The normalized second limit has one nonzero eigenvalue n_y/n per block, and hence entropy $H((n_y/n)_y)$. The first has entropy $H((n_y/n)_y) + \sum_y (n_y/n) S(K_y/n_y)$ by the entropy formula for direct sums. Subtraction gives equation 20. Singleton blocks have entropy zero. □

Remark B.11 (Population and fixed-sample limits). For categorical observations, repeated samples per category allow equation 20 to estimate the averaged posterior target as those sample counts grow. For a nonatomic observation law, paired observations are almost surely unique. Sharpening the observation kernel at fixed n then forces the empirical conditional entropy to zero, even if the population conditional entropy is positive. Continuous conditioning therefore requires either repeated draws at fixed observations or an analysis linking smoothing to sample size. The finite-category population bound alone cannot justify an arbitrarily narrow kernel on distinct observations.

B.4 What the aggregate score alone can determine

Direct within-condition scoring uses the conditional blocks, whereas an aggregate joint-minus-conditioning score discards them. The following result quantifies the resulting information loss. The full joint Gram matrix and output marginal may retain further information.

Proposition B.12 (Exact recovery limit from the aggregate score)**.** *Fix $m \geq 2$, strictly positive weights p, and a unit-diagonal positive semidefinite conditioning Gram matrix L. Let $C = H(p) - S(D_p^{1/2} L D_p^{1/2})$. Consider all conditional density-operator ensembles in a common output space of dimension $d \geq m$, with finite entropies. If a recovery rule observes only (G, p, L), then*

$$\inf_T \sup_{\{\rho_y\}} \left| T(G, p, L) - \sum_y p_y S(\rho_y) \right| = \frac{C}{2}.$$

The rule $T(G, p, L) = \max\{0, G - C/2\}$ attains this error. If $C > 0$, two ensembles have the same (G, p, L), with $G = C$, but targets 0 and C. Consequently any interval required to contain every compatible target at this observation must have width at least C.

Proof. Write $H = \sum_y p_y S(\rho_y)$. The conditioning certificate gives $H \leq G \leq H + C$, so the un-clipped midpoint $G - C/2$ has error at most $C/2$. Clipping its value to $[0, \infty)$ cannot increase error because $H \geq 0$.

For the lower bound, fix orthonormal vectors $e_1, \ldots, e_m$ and first take $\rho_y = e_y e_y^*$. Then $H = 0$ and $G = C$. Since $0 \leq C \leq H(p) \leq \log m$, there is a density matrix σ on their span with $S(\sigma) = C$. For example, the continuous family $(1-u)e_1 e_1^* + uI_m/m$, $0 \leq u \leq 1$, has entropy ranging from 0 to $\log m$. In the second ensemble let $\rho_y = \sigma$ for every y. Its joint state is $\sigma \otimes \rho_B$, so $H = G = C$. At the common observation $(G, p, L) = (C, p, L)$ every scalar estimate a has $\max\{|a|, |a - C|\} \geq C/2$. An interval containing both compatible targets has width at least C. When $C = 0$ the certificate already gives $G = H$, completing the proof. □

The output feature map may be held fixed throughout this construction: it suffices that it can realise the m orthogonal unit states, since σ can be chosen as a mixture of those same states. If a fixed feature map cannot realise them, the upper bound still holds but the lower construction need not be admissible. The deterministic recovery limit applies to the specified aggregate summary. Access to conditional samples, grouped Gram blocks, the complete joint operator or additional spectral information changes the recovery problem.

Corollary B.13 (Sharp margin for universal ranking)**.** *Under the same unrestricted-ensemble premise, a conditioning encoding preserves all strict within-condition entropy rankings if and only if $L = I_m$. More precisely, for two ensembles with common (p, L),*

$$(G_A - G_B) - C \le H_A - H_B \le (G_A - G_B) + C.$$

Thus $G_A - G_B > C$ guarantees $H_A > H_B$. If $C > 0$, every positive score margin $b < C$ is compatible with a strictly reversed target ordering. At margin $b = C$ a target tie is possible.

Proof. Subtract $G_j = H_j + d_j$, where $0 \le d_j \le C$, to obtain the displayed bound. To show sharpness take ensemble A to be the orthogonal pure ensemble in Proposition B.12, and let ensemble B be condition independent with entropy $C - b$. Then $G_A = C$, $H_A = 0$, and $G_B = H_B = C - b$. For $0 < b < C$ the score margin is b, while $H_A < H_B$. For $b = C$ both targets are zero. Finally, $C = D(D_p^{1/2} L D_p^{1/2} \| D_p)$ vanishes exactly when $L = I_m$. □

Deterministic certificates and exploratory resolution choices. For a fixed bank, the empirical conditional states are density matrices, so the conditioning certificate holds simultaneously for every valid conditioning Gram matrix and every integer tensor power. Selecting a tighter tolerance after inspecting that bank does not require a statistical multiple-testing adjustment for this algebraic statement. Such a selection is exploratory and supports conclusions about the observed bank. Independent confirmation on unseen outputs remains a separate task, and the originally specified tolerance and its abstentions must remain reported. For population confidence intervals obtained from sampled outputs, sampling uncertainty and any statistical selection must be handled separately.

C Global truncation and conditional compression

C.1 Global truncation does not commute with conditioning

For a density operator R with eigenvalues $\lambda_1 \ge \lambda_2 \ge \cdots \ge 0$, padded by zeros if necessary, define the redistributed top-t spectrum and its entropy by

$$\lambda_j^{[t]} = \lambda_j + \frac{1 - \sum_{a=1}^{t} \lambda_a}{t}, \quad 1 \le j \le t, \qquad H^{[t]}(R) = -\sum_{j=1}^{t} \lambda_j^{[t]} \log \lambda_j^{[t]}.$$

This is the convention in Eq. (13) of Jalali et al. (2026), restated explicitly in their Eq. (23). It differs both from discarding the tail without renormalization and from renormalizing the retained eigenvalues proportionally. Our counterexample below holds under the proportional convention as well.

Proposition C.1 (An exact rank-budget law)**.** *Let Y be uniform on $m \ge 2$ categories represented by orthogonal unit features. Given each category, let the posterior operator have $r \ge 2$ equal nonzero eigenvalues. Then $H_{\text{post}} = \log r$, whereas the globally truncated joint-minus-prompt entropy is*

$$H^{[t]}(\rho_{AB}) - H^{[t]}(\rho_B) = \log \min(t, mr) - \log \min(t, m). \tag{21}$$

In particular, it is zero for every $t \le m$, and it equals the full posterior entropy only when $t \ge mr$. By contrast, truncating each posterior separately at $t \ge r$ already preserves its entropy exactly. The formula also holds for proportional renormalization of top-t spectra.

Proof. Orthogonal observation features make ρ_{AB} a direct sum of m blocks $(1/m)\rho_y$, each with r eigenvalues $1/(mr)$. Its spectrum is uniform on mr coordinates. The prompt operator has m eigenvalues $1/m$. For a uniform spectrum of rank d, redistribution of its omitted mass over the first $t < d$ coordinates gives t entries $1/t$, hence entropy $\log t$. For $t \ge d$, the omitted mass is zero and the entropy remains $\log d$. Proportional renormalization produces the same vectors in this uniform case. Subtraction proves the formula. □

Corollary C.2 (Negative truncated conditional entropies)**.** *Let Y be uniform on m orthogonally represented categories, with any posterior density operators. For every $1 \le t \le m$,*

$$H^{[t]}(\rho_{AB}) - H^{[t]}(\rho_B) \le 0.$$

The inequality is strict if the first t eigenvalues of ρ_{AB} are not all equal. Thus its exponent can be less than one, although the untruncated conditional score is at least one.

Proof. The prompt spectrum is uniform on m coordinates, so its redistributed top-t entropy is $\log t$. The joint redistributed spectrum is a probability vector on t coordinates and has entropy at most $\log t$, with equality exactly for the uniform vector. Adding the same scalar to each retained eigenvalue makes that vector uniform exactly when the retained eigenvalues were equal. The full joint state is separable and Theorem B.1 ensures its conditional entropy is nonnegative. □

Example C.3 (A finite Gaussian counterexample to truncated aggregation)**.** Let Y be uniform on two point prompts $0, D \in \mathbb{R}$ and let X be an independent copy. Use unit-bandwidth Gaussian kernels on both variables, $k(u,v) = \ell(u,v) = \exp(-(u-v)^2/2)$, and set $c = e^{-D^2/2}$. Both feature maps have unit norm. The conditioning mixture has two components with weights $1/2$, means 0, D, and zero within-component variances. The spectra are

$$\lambda(\rho_B) = \lambda(\rho_y) = ((1+c)/2, (1-c)/2),$$
$$\lambda(\rho_{AB}) = ((1+c)^2/4, (1-c^2)/4, (1-c^2)/4, (1-c)^2/4).$$

At $t = 2$, the redistributed joint spectrum is $(q, 1-q)$ with $q = 1/2 + (c+c^2)/4$. The conditioning and conditional output spectra already have rank two. Hence

$$G_2 = h_2(q) - h_2((1+c)/2), \qquad \sum_{g=1}^{2} \tfrac{1}{2} H^{[2]}(X \mid G = g) = h_2((1+c)/2).$$

Here G in $X \mid G = g$ denotes the mixture-component label, following the cited theorem. The absolute aggregation discrepancy is

$$\left| h_2\left(\tfrac{1}{2} + \tfrac{c+c^2}{4}\right) - 2h_2\left(\tfrac{1+c}{2}\right) \right| \longrightarrow \log 2.$$

Appendix B, Theorem 7, Eq. (19), of the final published Jalali et al. (2026) bounds this discrepancy by $\sqrt{2\Gamma}\log(2/\Gamma)$. Their Eq. (16) defines

$$\Gamma = 32 \sum_{i=1}^{m} \omega_i \left[\frac{\sigma_i^2}{\sigma^2} + (i-1) \sum_{j<i} \exp\left(-\frac{\|\mu_i - \mu_j\|^2}{\sigma^2} \right) \right].$$

With $m = 2$, $\omega_1 = \omega_2 = 1/2$, $\sigma_1^2 = \sigma_2^2 = 0$ and $\sigma = 1$, only the $i = 2, j = 1$ term remains, giving $\Gamma = 16e^{-D^2}$. At $D = 4$, the actual discrepancy is 0.693147082 nats while the claimed upper bound is 0.026416531 nats. Thus both kernels are Gaussian at a fixed positive bandwidth, with finite support and finite entropy throughout.

The failing step is the final equality in the published proof. Globally truncating the stacked spectrum of weighted blocks differs from truncating each block separately and averaging its entropy. The rank-budget law above isolates this failure. The counterexample concerns the truncated aggregation statement, while the untruncated identity remains valid for this independent-output law.

For an even simpler spectral calculation, retain the same Gaussian prompt law and replace the output features by two equally likely orthogonal unit states, independent of Y. The joint spectrum becomes $((1+c)/4, (1+c)/4, (1-c)/4, (1-c)/4)$, whose redistributed top-two spectrum is $(1/2, 1/2)$. The aggregation discrepancy is then $h_2((1+c)/2) = 0.693147124$ nats at $D = 4$, against the same upper bound. This orthogonal-output variant also fails under proportional renormalisation of the retained spectrum. The Gaussian-on-both-inputs calculation above uses the published equal-redistribution convention.

Proposition C.4 (Blockwise tail aggregation preserves conditioning)**.** *For each category y, choose $t_y \in \{0, 1, \ldots\}$ and let $\lambda_{y,j}$ be its posterior eigenvalues. Write $\delta_y = \sum_{j>t_y} \lambda_{y,j}$ and*

$$L_{t_y}(\rho_y) = \sum_{j=1}^{t_y} h(\lambda_{y,j}) + h(\delta_y), \qquad h(a) = -a\log a, \quad h(0) = 0.$$

This quantity retains the leading eigenvalues and aggregates the remaining mass into one category-specific tail state. If $\overline{\rho}_y$ is the diagonal density matrix with entries $(\lambda_{y,1}, \ldots, \lambda_{y,t_y}, \delta_y)$, then for orthogonal category features

$$S\left(\bigoplus_y p_y \overline{\rho}_y \right) - H(p) = \sum_y p_y L_{t_y}(\rho_y). \tag{22}$$

The exact posterior-entropy loss is

$$H_{\mathrm{post}} - \sum_y p_y L_{t_y}(\rho_y) = \sum_{y:\delta_y>0} p_y \delta_y S\big((\lambda_{y,j}/\delta_y)_{j>t_y}\big)\,. \tag{23}$$

If each posterior operator has rank at most $r_y < \infty$, the loss is at most

$$\sum_{y:\delta_y>0} p_y \delta_y \log(r_y - t_y).$$

Terms with no tail have loss zero. For nonorthogonal observation features, forming $\sum_y p_y \overline{\rho}_y \otimes \sigma_y$ gives the same representation-gap certificate as Theorem B.1, with target $\sum_y p_y L_{t_y}(\rho_y)$.

Proof. Entropy of a direct sum of normalized blocks gives equation 22. For any $\delta_y > 0$,

$$\sum_{j>t_y} h(\lambda_{y,j}) = h(\delta_y) + \delta_y S\big((\lambda_{y,j}/\delta_y)_{j>t_y}\big)\,.$$

Subtract the aggregated-tail entropy and average over y to obtain equation 23. A normalized tail supported on at most $r_y - t_y$ entries has entropy at most $\log(r_y - t_y)$. Finally, each $\overline{\rho}_y$ is a density operator, so the conditioning-gap theorem applies directly, after embedding them in a common Hilbert space if their dimensions differ. □

Remark C.5 (Scope of the remedy). Blockwise tail aggregation is the elementary entropy grouping identity, not a new quantum entropy construction. Its useful role here is to preserve the intended conditional target and expose the approximation error. The matrices $\overline{\rho}_y$ are auxiliary spectral summaries. Their chosen common diagonal basis does not preserve the original alignment of posterior eigenspaces across categories. The nonorthogonal certificate concerns the auxiliary joint state and does not bound its difference from the original joint RKHS state. The computation requires posterior spectra grouped by observation or category. With unbounded rank, discarded mass alone cannot bound the discarded entropy. A rank or spectral-tail assumption is essential. Neither this compression identity nor a separation bound controls error from estimating each posterior spectrum with a small number of samples.

D FINITE-SAMPLE POPULATION CERTIFICATES

D.1 BIAS-AWARE CERTIFICATES FOR A FIXED CONDITIONAL OUTPUT LAW

Fix an observation y, a solver, and a kernel before drawing the evaluation samples. Write P for this conditional output law. Throughout this subsection, ϕ is a strongly measurable map into a separable Hilbert space with $\|\phi(x)\| = 1$, $k(x,x') = \langle \phi(x), \phi(x')\rangle$, and

$$Z_i = \phi(X_i) \otimes \phi(X_i), \qquad \rho = \mathbb{E} Z_i, \qquad \rho_n = \frac{1}{n}\sum_{i=1}^n Z_i, \qquad X_i \overset{\mathrm{iid}}{\sim} P.$$

Set $S(\rho) = -\operatorname{Tr}(\rho \log \rho)$, and use natural logarithms. The iid requirement concerns independent complete sampler runs. Consecutive states of a Markov chain generally require a separate dependence analysis.

Proposition D.1 (Finite-sample entropy bias)**.** *Suppose $S(\rho) < \infty$. Then*

$$b_n := S(\rho) - \mathbb{E} S(\rho_n) = \mathbb{E} D(\rho_n \| \rho) \geq 0, \qquad b_{n+1} \leq b_n. \tag{24}$$

If $\operatorname{rank}(\rho) \leq r < \infty$, then

$$0 \leq b_n \leq B_{n,r} := \log\left(1 + \frac{r-1}{n}\right). \tag{25}$$

For the empirical entropy with its tail dropped, without redistribution, $S_{\mathrm{head},m}(\rho_n) = \sum_{j\leq m} -\lambda_j(\rho_n)\log\lambda_j(\rho_n)$, $S(\rho) - \mathbb{E} S_{\mathrm{head},m}(\rho_n) \geq b_n$. Thus dropping the entropy tail cannot reduce its downward bias for the full population entropy target. This statement does not concern redistributed truncation. Dropping the tail may change variance, computational cost, or the chosen target.

Proof. Almost surely Z_i is supported on the support of ρ: if Q is the orthogonal projection onto its kernel, then $\mathbb{E}\|Q\phi(X)\|^2 = \mathrm{Tr}(Q\rho) = 0$. Define the nonnegative quadratic form $g(X) = \langle \phi(X), -\log \rho\, \phi(X)\rangle$ by its increasing spectral sum on this support. Tonelli's theorem gives $\mathbb{E}g(X) = S(\rho) < \infty$. The relative entropy identity $D(\rho_n\|\rho) = n^{-1}\sum_i g(X_i) - S(\rho_n)$ has finite terms almost surely, and taking expectations proves the equality and its nonnegativity. The average of the $n+1$ leave-one-out empirical operators, each formed from n observations, is ρ_{n+1}. Entropy concavity and exchangeability therefore imply $\mathbb{E}S(\rho_{n+1}) \geq \mathbb{E}S(\rho_n)$.

For the finite-rank bound, restrict to the support of ρ, and write $s = \mathrm{rank}(\rho) \leq r$. For any density matrix σ on this space,

$$D(\sigma\|\rho) \leq \log \mathrm{Tr}(\sigma^2\rho^{-1}).$$

For completeness, if $\sigma = \sum_i a_i u_i \otimes u_i$ and $\rho = \sum_j \lambda_j e_j \otimes e_j$, the weights $a_i|\langle u_i, e_j\rangle|^2$ sum to one. Applying scalar Jensen to the logarithm of a_i/λ_j proves this inequality, with zero a_i omitted. Since $Z_i^2 = Z_i$, independence gives

$$\begin{aligned}\mathbb{E}\,\mathrm{Tr}(\rho_n^2\rho^{-1}) &= \frac{1}{n}\mathrm{Tr}(\rho\rho^{-1}) + \frac{n-1}{n}\mathrm{Tr}(\rho^2\rho^{-1}) \\ &= 1 + \frac{s-1}{n}.\end{aligned}$$

A second application of Jensen proves equation 25. The truncation statement follows from $S_{\mathrm{head},m}(\rho_n) \leq S(\rho_n)$. □

Expected-entropy monotonicity is already established for log-Vendi by Farnia et al. (2026, Proposition 1), and the underlying finite-sample bias is familiar from classical entropy estimation. The proposition above collects the facts needed to distinguish a population certificate from a confidence interval around a biased finite-sample expectation.

Proposition D.2 (A population entropy interval with known rank bound)**.** *Suppose a deterministic bound $r \geq \mathrm{rank}(\rho)$ is known, and let $n \geq 2$, $\delta \in (0,1)$,*

$$h_2(t) = -t\log t - (1-t)\log(1-t), \qquad a_n(\delta) = h_2(1/n)\sqrt{\frac{n}{2}\log\frac{2}{\delta}}.$$

With probability at least $1-\delta$, the population log-Vendi belongs to

$$I_n^{(1)}(\delta, r) = \left[\max\{0, S(\rho_n) - a_n(\delta)\},\quad \min\{\log r, S(\rho_n) + a_n(\delta) + B_{n,r}\}\right]. \tag{26}$$

Exponentiating the endpoints gives an interval for population Vendi. The rank bound must concern the population feature space. Replacing r by the observed Gram-matrix rank is invalid.

Proof. On replacing one observation, write the original and modified operators as $(1-1/n)\tau + (1/n)Z$ and $(1-1/n)\tau + (1/n)Z'$, where Z, Z' are pure states and τ is shared. The mixture-entropy bounds place both entropies in $[(1-1/n)S(\tau), (1-1/n)S(\tau) + h_2(1/n)]$. Thus the exact uniform bounded-difference constant is at most $h_2(1/n)$. Equality is attained by replacing one of n identical unit vectors by an orthogonal vector, provided that this pair is admissible under the output feature map. McDiarmid's inequality gives $|S(\rho_n) - \mathbb{E}S(\rho_n)| \leq a_n(\delta)$ with the stated probability. Combine this event with $0 \leq b_n \leq B_{n,r}$ and $0 \leq S(\rho) \leq \log r$. □

The rank-based allowance can be large. A repeated-subsampling standard error measures fluctuation around a finite-sample expectation, leaving the bias b_n unresolved. For $n = 16$, $r = 384$, and $\delta = 0.05$, $a_n(\delta) \simeq 1.270$ and $B_{n,r} \simeq 3.216$ nats. A precise population-entropy claim from such a budget therefore needs additional spectral information or stronger data-dependent theory.

Proposition D.3 (Purity and order-two intervals)**.** *Let $p = \mathrm{Tr}(\rho^2) = \mathbb{E}|k(X, X')|^2$ with independent $X, X' \sim P$. For $n \geq 2$, define*

$$\widehat{p}_n = \frac{1}{n(n-1)}\sum_{i\neq j}|k(X_i, X_j)|^2, \qquad m = \lfloor n/2 \rfloor.$$

Then $\mathbb{E}\widehat{p}_n = p$, and

$$\mathbb{P}(|\widehat{p}_n - p| > t) \leq 2e^{-2mt^2}. \tag{27}$$

Consequently, if $u = \sqrt{\log(2/\delta)/(2m)}$, $\ell = \max\{0, \widehat{p}_n - u\}$, *and* $v = \min\{1, \widehat{p}_n + u\}$, *then*

$$p \in [\ell, v], \quad H_2(\rho) \in [-\log v, -\log \ell], \quad \mathrm{VS}_2(\rho) \in [1/v, 1/\ell]$$

with probability at least $1 - \delta$, *interpreting* $-\log 0 = 1/0 = \infty$. *If a population rank bound* r *is known, replace* ℓ *by* $\max\{1/r, \ell\}$. *These are intervals for order-two diversity, not two-sided intervals for full entropy. Nevertheless* $S(\rho) \geq H_2(\rho)$ *gives the certified lower bound* $S(\rho) \geq -\log v$.

Proof. The PSD and unit-diagonal assumptions imply $|k(x, x')|^2 \in [0, 1]$. The stated purity identity follows by expanding $\mathrm{Tr}(\rho^2)$. For any permutation of the observations, the mean over its first m disjoint pairs averages independent $[0, 1]$ variables of mean p. Averaging this paired statistic over all permutations equals the U-statistic $\widehat{p}_n$. Jensen's inequality for the exponential, followed by Hoeffding's lemma for each paired statistic, yields the same moment generating function bound as m independent bounded observations. Chernoff's argument gives equation 27. The remaining claims follow from monotonicity of the transformations, $1/r \leq p \leq 1$, and the standard Rényi entropy ordering. □

The diagonal-inclusive empirical purity satisfies the exact identity

$$\mathrm{Tr}(\rho_n^2) = \frac{1}{n} + \frac{n-1}{n}\widehat{p}_n, \qquad \mathbb{E}\,\mathrm{Tr}(\rho_n^2) = p + \frac{1-p}{n}.$$

Subtracting this known diagonal bias estimates purity without bias. Taking its reciprocal or logarithm does not yield an unbiased Vendi or entropy estimator, and a zero estimate must not be silently replaced by an arbitrary positive constant.

Corollary D.4 (Comparisons and valid adaptive stopping). *Consider* $L < \infty$ *predeclared conditional-law/kernel targets and predeclared sample sizes* n_t, $t \geq 1$. *For each target* j, *at stage* t *use failure probability* $\delta_{j,t} = 6\delta/(\pi^2 L t^2)$ *in the desired interval above. Then every resulting interval covers its population target simultaneously, at all stages, with probability at least* $1 - \delta$. *In particular, stopping at the first stage where two entropy intervals are disjoint certifies their ordering with this same familywise confidence. Stopping when an interval has width at most* 2η *gives an estimate with absolute error at most* η *by taking its midpoint. If intervals still overlap at the declared maximum budget, the comparison is unresolved. These conclusions also hold conditionally on fixed observed inputs* y. *They concern* $Q_a(\cdot \mid y)$, *not the unknown true posterior unless the solver is exact.*

Proof. A union bound gives total failure probability at most $\sum_{j=1}^{L} \sum_{t=1}^{\infty} 6\delta/(\pi^2 L t^2) = \delta$. No independence between targets or nested stages is needed for this union bound. Independence of the draws within each evaluated target, required by its fixed-stage interval, is still essential. □

D.2 Why finite entropy alone cannot certify a finite budget

Proposition D.5 (Rare-tail obstruction). *On* $\mathbb{N}_0$ *with the unit-diagonal identity kernel, let* $P_0 = \delta_0$ *and*

$$P_{\varepsilon,M} = (1-\varepsilon)\delta_0 + \frac{\varepsilon}{M}\sum_{j=1}^{M}\delta_j, \qquad 0 < \varepsilon < 1, \quad M \in \mathbb{N}.$$

Both distributions have finite entropy, but

$$S(P_0) = 0, \qquad S(P_{\varepsilon,M}) = h_2(\varepsilon) + \varepsilon \log M,$$
$$p(P_0) = 1, \qquad p(P_{\varepsilon,M}) = (1-\varepsilon)^2 + \varepsilon^2/M.$$

The total-variation distance between their n*-sample laws is* $1 - (1-\varepsilon)^n$. *If their entropy gap exceeds* 2η, *every (possibly randomized) estimator* T_n *satisfies*

$$\max_{P \in \{P_0, P_{\varepsilon,M}\}} P\big(|T_n - S(P)| > \eta\big) \geq \frac{(1-\varepsilon)^n}{2}.$$

Thus no fixed finite sample size gives a uniform finite-error guarantee with failure probability below $1/2$ *over all finite-entropy laws and unit-diagonal kernels with no common rank or tail restriction. Moreover, no deterministic uniformly valid population-entropy upper confidence bound can be finite on the all-zero sample over this class.*

Proof. For the identity kernel, the covariance operator is diagonal with the probability masses as eigenvalues, giving the displayed entropy and purity. The P_0 sample is always $(0, \dots, 0)$. Its probability under the alternative is $(1-\varepsilon)^n$, proving the total-variation formula. The two events of estimation within η correspond to disjoint parameter intervals. The standard two-point testing bound therefore gives the claimed lower bound. Equivalently, thresholding T_n at the midpoint yields a test whose error probabilities sum to at least one minus total variation. For any fixed n, η, take ε arbitrarily small and then choose finite M sufficiently large to make the entropy gap exceed 2η. This makes the lower bound arbitrarily close to $1/2$. Finally, if an upper bound takes a finite value U on the zero sample, choose ε such that $(1-\varepsilon)^n > \delta$ and finite M such that $S(P_{\varepsilon,M}) > U$. Its failure probability then exceeds δ. □

For the same alternatives, $H_2(P_{\varepsilon,M}) \le -2\log(1-\varepsilon)$ remains small regardless of M. This separates the sensitivity of full entropy to rare populations from the bounded-kernel estimability of purity. Almost-sure plug-in consistency still holds for each fixed finite-entropy law. A known finite-dimensional feature map imposes a common rank bound and therefore lies outside this unrestricted class.

D.3 Continuous conditions: what a finite partition changes

For continuous observations, a finite partition gives an explicit resolution at which nearby conditions are pooled. We bound the resulting error using established finite-dimensional entropy continuity and concentration tools.

Let Q_y denote the law actually sampled at y, with unit features in a common known r-dimensional space, $2 \le r < \infty$. Write $\rho_y = \int \phi(x)\phi(x)^* Q_y(dx)$ and $H_Y = \int S(\rho_y) P_Y(dy)$. If Q_y is the true posterior this is posterior diversity. The statements equally apply to a generator's own conditional law. Fix a measurable partition $Z = q(Y) \in \{1, \dots, m\}$, with known nonzero probabilities p_z, and set

$$\bar{\rho}_z = \mathbb{E}[\rho_Y \mid Z = z], \qquad H_Z = \sum_z p_z S(\bar{\rho}_z).$$

The state $\bar{\rho}_z$ mixes the conditional states within a cell. Taking its entropy differs from evaluating a representative condition or averaging the individual conditional entropies.

Proposition D.6 (Coarsening and representation have separate costs)**.** *Assume that, in each cell, outside a conditional null set, $\frac{1}{2}\|\rho_y - \rho_{y'}\|_1 \le \delta$. Define*

$$F_r(\delta) = h_2(t) + t\log(r-1), \qquad t = \min\{\delta, 1 - 1/r\}.$$

Then

$$0 \le H_Z - H_Y \le F_r(\delta). \tag{28}$$

If the finite cell labels are subsequently encoded by a unit-diagonal positive-semidefinite Gram matrix L, their joint-minus-label score obeys

$$0 \le G_Z - H_Y \le F_r(\delta) + C(p, L). \tag{29}$$

The orthogonal label kernel removes the second cost exactly.

Proof. Entropy concavity gives the lower bound in each cell. Convexity of the trace norm gives $\frac{1}{2}\|\bar{\rho}_z - \rho_y\|_1 \le \delta$ almost everywhere. The Audenaert entropy-continuity inequality (Audenaert, 2007) bounds the entropy difference by $h_2(u) + u\log(r-1)$ at the actual trace distance u. This expression increases up to $1 - 1/r$, where it equals $\log r$. Using the trivial entropy range after that point proves the saturated bound $F_r(\delta)$. Integrate over the cells. Theorem 3.1 applied to the finite ensemble $(p_z, \bar{\rho}_z)$ gives $0 \le G_Z - H_Z \le C(p, L)$, proving the second claim. □

A known modulus of continuity makes this bound operational. If $\frac{1}{2}\|\rho_y - \rho_{y'}\|_1 \le \omega(d(y, y'))$ and all cells have diameter at most h, take $\delta = \omega(h)$. Without a justified modulus, small distances between observations alone do not certify small coarsening error. A Gaussian output kernel generally has infinite population rank. Finite input dimension does not justify the common r-dimensional premise.

D.4 A SIMULTANEOUS RESOLUTION AND SAMPLE-ALLOCATION BOUND

For each cell, independently draw $n_z \geq 2$ outputs from $\bar{Q}_z = \int Q_y P_Y(dy \mid Z = z)$ and construct $\hat{\rho}_z$. One implementation draws a fresh Y in the cell before each independent generator run. Holding one representative y_z fixed instead samples Q_{y_z}, a different target whose comparison with the mixture requires another bound. Known p_z are part of the design. Plugging in random frequencies adds a separate weighting error. Set

$$\hat{H}_Z = \sum_z p_z S(\hat{\rho}_z),$$
$$B = \sum_z p_z \log\left(1 + \frac{r-1}{n_z}\right),$$
$$A = \left[\frac{1}{2}\log\frac{2}{\alpha}\sum_z p_z^2 n_z h_2(1/n_z)^2\right]^{1/2}.$$

Proposition D.7 (Grouped population interval)**.** *Under the preceding design and Proposition D.6's assumptions, with probability at least* $1-\alpha$*,*

$$H_Y \in \left[\max\{0, \hat{H}_Z - A - F_r(\delta)\}, \quad \min\{\log r, \hat{H}_Z + A + B\}\right]. \tag{30}$$

For a fixed finite catalogue without coarsening, set $F_r(\delta) = 0$*. For* m *equally weighted cells with* n *independent draws each and* $N = mn$*,*

$$B = \log\left(1 + \frac{(r-1)m}{N}\right), \qquad A = h_2(1/n)\sqrt{\frac{n}{2m}\log\frac{2}{\alpha}}. \tag{31}$$

Proof. The bias bound in Proposition D.1 applied to each cell gives $0 \leq H_Z - \mathbb{E}\hat{H}_Z \leq B$. Replacing one output in cell z changes its entropy by at most $h_2(1/n_z)$, by the shared-mixture argument in Proposition D.2. Its contribution to the weighted estimator changes by at most $p_z h_2(1/n_z)$. McDiarmid's inequality over all independent output draws therefore gives $\Pr(|\hat{H}_Z - \mathbb{E}\hat{H}_Z| > A) \leq \alpha$. Combining this event with the one-sided bias bound and $H_Z - F_r(\delta) \leq H_Y \leq H_Z$ gives the interval. Substitution gives the balanced formulas. □

At fixed N, making m large reduces cell diameter but increases the bias allowance through m/N. This is an explicit resolution and sample-budget allocation tradeoff, even when label encoding itself has zero certificate. For balanced partitions with diameter $h_N \to 0$, $m_N \to \infty$, $n_N = N/m_N \to \infty$, fixed r, and $\omega(h_N) \to 0$, all three allowances vanish. Indeed $h_2(1/n) \leq (1 + \log n)/n$ gives $A \leq (1 + \log n)\sqrt{\log(2/\alpha)/(2N)}$. Thus the estimator is consistent for H_Y under this stratified design. Taking a summable sequence $\alpha_N = N^{-2}$ also gives almost-sure convergence by Borel-Cantelli. This establishes a sufficient regime. It leaves minimax rates, adaptive bandwidth selection and infinite-rank continuity outside the scope of this bound.

D.5 WHAT THE DIRECT GROUPED BASELINE COSTS

With m conditions and n outputs per condition, direct grouped entropy requires m Gram matrices of size $n \times n$. Dense construction stores $O(mn^2)$ entries if all blocks are retained, or $O(n^2)$ when processed sequentially. Dense full eigendecomposition costs $O(mn^3)$. A dense paired score on all $N = mn$ outputs uses an $N \times N$ matrix, requiring $O(m^2n^2)$ storage and $O(m^3n^3)$ eigendecomposition. Actual runtimes depend on the implementation as well as these dense arithmetic counts. Feature-space covariance methods, matrix-vector oracles, low-rank approximations and block structure can alter both costs. In particular, an implementation using orthogonal labels should exploit the block structure rather than run a dense global solver. The compression experiments compare retained head counts and remainder information, with accuracy advantages varying across settings.

D.6 An exact check of the resolution tradeoff

Let $Y \sim \mathrm{Unif}[0,1]$ and $X \mid Y = y \sim \mathrm{Bernoulli}(.1 + .8y)$, with the two orthogonal output features. The target is $H_Y = \int_0^1 h_2(.1 + .8y)\, dy = .5778952084$. Partition into m equal intervals and draw $n = N/m$ independent samples from each conditional mixture. Its Bernoulli parameter is exactly the midpoint value, so binomial sums compute the empirical entropy expectation and variance without Monte Carlo. Here $r = 2$ and $\delta = .8/m$. Table 4 illustrates the two opposing biases. The release

Table 4: Exact expectations at total budget $N = 4096$. Pooling bias is $H_Z - H_Y \geq 0$. Sampling bias is $H_Z - \mathbb{E}\hat{H}_Z \geq 0$. Their cancellation is not evidence that either error is absent.

Cells m	Draws/cell n	Pooling bias	Sampling bias	Net expected error
2	2048	.032969	.000244	.032725
8	512	.002262	.000978	.001284
32	128	.000143	.003930	-.003787
128	32	.000009	.016065	-.016056

retains all 21 designs from $N \in \{1024, 4096, 16384\}$ and $m \in \{2, 4, 8, 16, 32, 64, 128\}$, including every allowance and the exact expectation. All numerical biases satisfy their stated analytic bounds. This controlled continuous-condition model checks the interpretation and constants. Its practical benefit on generator data remains to be tested.

E Conditioning experiments

E.1 Exact finite-law calculation and implementation

All conditioning experiments use natural logarithms, unit-diagonal positive semidefinite kernels, and a specified finite set of condition weights. For conditions $y_1, \ldots, y_m$ and output atoms $x_1, \ldots, x_d$, let $p_j = \Pr(Y = y_j)$, $Q_{ja} = Q_{y_j}(\{x_a\})$, and $K_{ab} = k(x_a, x_b)$, $L_{ij} = \ell(y_i, y_j)$. Write $D(v) = \mathrm{diag}(\sqrt{v})$ and $w_{ja} = p_j Q_{ja}$. We compute

$$H_{\mathrm{post}} = \sum_{j=1}^{m} p_j S\big(D(Q_{j,:})KD(Q_{j,:})\big), \tag{32}$$

$$G = S\big(D(w)(L \otimes K)D(w)\big) - S\big(D(p)LD(p)\big). \tag{33}$$

The vector w in equation 33 is ordered first by condition, then by atom. These are exact covariance spectra for the declared finite joint law. They require no posterior Monte Carlo approximation. The experiment checks $0 \leq G - H_{\mathrm{post}} \leq H(p) - S(D(p)LD(p))$ throughout. The resulting interval

$$[\max\{0, G - H(p) + S(D(p)LD(p))\},\ G]$$

is a deterministic identification interval for H_{post} under that law, not a statistical confidence interval.

The implementation uses double-precision symmetric eigendecomposition, checks that the minimum eigenvalue is at least -5×10^{-11} and that the trace differs from one by less than 5×10^{-10}, and clips negative roundoff eigenvalues to zero. The largest joint matrix is 240×240. All computation runs on CPU with numerical libraries limited to one thread. The master seed is 20270917. We compute full spectra for validation, including when subsequently evaluating a truncation rule. The experiments do not benchmark a partial eigensolver or claim a computational speedup.

E.2 Categorical equality cases, ranking reversal, and truncation

We use two equally weighted conditions at scalar locations 0 and D, with Gaussian observation kernel of bandwidth one. Their feature overlap is $c = \exp(-D^2/2)$. The sweep is $D \in \{0, 0.1, \ldots, 8\}$, and the output kernel is the identity on the specified output atoms.

First, the condition-dependent deterministic law $X = Y$ has $H_{\mathrm{post}} = 0$ and

$$G = \log 2 - h_2((1 + c)/2),$$

where $h_2(u) = -u \log u - (1-u)\log(1-u)$. This saturates the conditioning-gap upper bound. In contrast, using the same $(0.95, 0.05)$ output distribution at both conditions gives $H_{\text{post}} = G = h_2(0.95) = 0.19851524$, independent of the observation kernel. Their ordering by G therefore reverses when only that kernel changes. The two controlled laws isolate diversity while leaving posterior fidelity outside the comparison.

Next, the same uniform two-atom output law at both conditions has $H_{\text{post}} = G = \log 2$ for every D, whereas the redistributed top-2 difference is $G_2 = \log 2 - h_2((1+c)/2)$. Thus sharpening the observation kernel removes the apparent diversity of this globally truncated statistic, even though within-condition top-2 entropy is already exact. The code implements the uniform redistribution in equation 8, rather than division of the retained eigenvalues by their sum.

Finally, consider conditional probabilities $(1, 0, 0)$ and $(0, 1/2, 1/2)$ on three orthogonal output atoms. At orthogonal conditions, the joint spectrum is $(1/2, 1/4, 1/4)$. The shifted top-2 spectrum is $(5/8, 3/8)$, so

$$H_{\text{post}} = \tfrac{1}{2}\log 2 = 0.34657359, \qquad G_2 = h_2(5/8) - \log 2 = -0.03158394.$$

The negative truncated log-score is a property of the exact spectra, not a numerical violation of the nonnegativity of the untruncated G.

We independently check equation 9 for $m, r \in \{2, 4, 8\}$ and truncation ranks selected from $\{1, 2, 3, 4, 6, 8, 12, 16, 24, 32, 48, 64, m, r, mr\}$ with $t \le mr$. For example, $m = r = 8$ gives $G_8 = 0$ and $G_{32} = \log 4$, while $H_{\text{post}} = \log 8$. Full joint rank 64 recovers the target. An additional tensor-power experiment uses four equally weighted conditions, initial off-diagonal overlap 0.8, and four independently generated $\text{Dirichlet}(1, 1, 1, 1, 1)$ output probability vectors with seed 20270918. The full G decreases toward the unchanged H_{post} as the observation Gram matrix is raised entry-wise to powers $1, \ldots, 16, 24, 32, 48, 64, 96, 128$. All computed gaps obey the certificate. The exact categorical and rank-grid values are included in `experiments/conditioning/results/categorical_kernel_and_truncation.csv`, `truncation_rank_grid.csv`, and `finite_category_tensor_powers.csv` in the same results directory.

E.3 Continuous observations: an analytic population reference

The continuous law is

$$Y \sim \text{Uniform}[-1, 1], \qquad X \mid Y = y \sim \mathcal{N}(\sin(2y), s^2), \qquad s = 0.5,$$

with output kernel $k(x, z) = \exp(-(x-z)^2/(2a^2))$, $a = 1$. Translation invariance makes the conditional covariance spectrum independent of y. Its entropy is available analytically, so no finite-sample reference estimate is used.

Derivation of the reference. For $P = \mathcal{N}(0, s^2)$ with density p_s, the kernel integral operator on $L^2(P)$ is unitarily equivalent, under $f \mapsto \sqrt{p_s} f$, to the operator on $L^2(\mathbb{R})$ with kernel

$$\widetilde{k}(x, z) = \frac{1}{\sqrt{2\pi} s} \exp\left[-\left(\frac{1}{2a^2} + \frac{1}{4s^2} \right)(x^2 + z^2) + \frac{xz}{a^2} \right]. \tag{34}$$

Set

$$u = \sqrt{1 + 4s^2/a^2}, \qquad b = \frac{u-1}{u+1}, \qquad \omega = \frac{u}{2s^2}.$$

The Mehler identity for orthonormal Hermite functions $\{\varphi_j^\omega\}_{j \ge 0}$ is

$$\sum_{j=0}^{\infty} b^j \varphi_j^\omega(x) \varphi_j^\omega(z) = \sqrt{\frac{\omega}{\pi(1-b^2)}} \exp\left[-\frac{\omega\{(1+b^2)(x^2+z^2) - 4bxz\}}{2(1-b^2)} \right].$$

Multiplication by $1 - b$ and substitution of u, b, ω gives equation 34, including its prefactor. Consequently the eigenvalues are $(1-b)b^j$, $j \ge 0$. Their sum is one, as required by the unit diagonal. Summing the geometric entropy series gives

$$H_{\text{post}} = -\log(1-b) - \frac{b}{1-b}\log b = (\nu+1)\log(\nu+1) - \nu \log \nu, \quad \nu = \frac{u-1}{2}. \tag{35}$$

For $s = 0.5, a = 1$, this is 0.55330330 nats. A nonzero Gaussian mean merely translates the integral operator and leaves this spectrum unchanged.

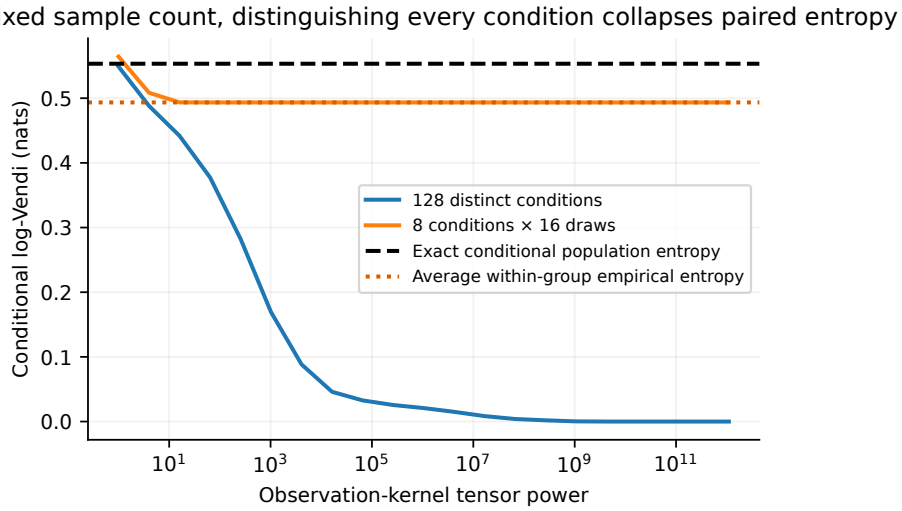

Figure 2: At fixed total sample count, distinguishing every observation by kernel sharpening drives the paired entropy to zero. Repeated observations instead recover the average within-group empirical entropy. The dashed population reference is the analytic Gaussian/RBF value in equation 35. It is not estimated from these samples.

Sampling designs and kernel sharpening. Using seed 20270919, the first design draws 128 independent (Y, X) pairs, whose observations are almost surely distinct. The second design fixes eight equally spaced observations in $[-0.9, 0.9]$ and draws 16 independent conditional outputs at each observation. Although the second design changes the condition distribution, its population average conditional entropy is the same because equation 35 does not depend on y. Both use the observation Gaussian kernel with initial bandwidth 0.4, raised entrywise to powers 2^j for $j = 0, 2, \ldots, 40$.

At fixed sample count, $\widehat{G}$ in the distinct-observation design converges to zero. In the repeated-condition design, it converges to 0.49346242, agreeing with the average within-group empirical entropy to 1.2×10^{-14}. This latter value remains below the analytic population target. Recovering the empirical target leaves its sampling error unchanged. The fixed-sample experiment isolates the sampling design and order of limits, without estimating an asymptotic convergence rate. The exact generated observations and outputs are supplied in `experiments/conditioning/results/continuous_sampling_arrays.npz`.

E.4 Exact inverse problem on a finite digit-image dictionary

We select the first four images of each digit label, in the order returned by `sklearn.datasets.load_digits`, and concatenate the labels in order $0, \ldots, 9$. These 40 actual 8×8 handwritten digit images come from the UCI Optical Recognition of Handwritten Digits dataset (Alpaydin & Kaynak, 1998; Pedregosa et al., 2011). Pixel values are divided by 16, giving vectors $x_a \in [0, 1]^{64}$. The prior is uniform on these 40 atoms. There is no training stage, learned representation, or fitted generative model.

The forward operator A retains the same four pixels in every image: zero-based row-major indices $(18, 21, 34, 37)$. With $\sigma = 0.35$, the declared observation model is

$$Y = AX + \sigma\varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, I_4).$$

The six fixed ground-truth dictionary indices are $(0, 7, 14, 21, 28, 35)$, and seed 20270920 generates their independent Gaussian noise vectors. For each resulting observation y_j, the conditional law is computed exactly:

$$Q_{ja} = \Pr(X = x_a \mid Y = y_j) = \frac{\exp(-\|y_j - Ax_a\|^2/(2\sigma^2))}{\sum_{b=1}^{40} \exp(-\|y_j - Ax_b\|^2/(2\sigma^2))}. \tag{36}$$

Output kernel	Specified output law	H_{post}	Range of G
Gaussian	Exact dictionary posterior	2.000490	2.000490 to 2.942288
Gaussian	Deterministic MAP fixture	0	0 to 1.531848
Gaussian	Fixed two-image fixture	0.301294	0.301294 to 0.301294
Normalized linear	Exact dictionary posterior	0.937900	0.937900 to 1.370238
Normalized linear	Deterministic MAP fixture	0	0 to 0.950754
Normalized linear	Fixed two-image fixture	0.071186	0.071186 to 0.071186

Table 5: Exact fixed-law results across the observation-bandwidth sweep. The interval in the last column describes changing the observation kernel, not sampling uncertainty. All values are log-diversities in nats.

We evaluate these six conditionals using $p_j = 1/6$ in equation 32 to equation 33. These weights specify a finite *observation-bank design average*, distinct from the $1/40$ prior weights over image atoms and from integration over the continuous observation marginal of the Gaussian forward model.

Every law and metric uses identical dictionary pixels, observations, masks, and condition weights. Two output kernels are fixed before computing the metric sweep:

$$k_{\mathrm{RBF}}(x,z) = \exp(-\|x-z\|^2/(2a_x^2)), \qquad a_x = \tfrac{1}{2}\,\mathrm{median}_{a<b}\,\|x_a - x_b\| = 1.53554880,$$
$$k_{\mathrm{lin}}(x,z) = \frac{\langle x,z\rangle}{\|x\|\|z\|}.$$

The Gaussian observation kernel uses bandwidth $a_y = r\,\mathrm{median}_{i<j}\,\|y_i - y_j\|$, where the base median is 1.40419920 and r takes 25 log-spaced values from 0.01 to 100. The output kernels and posterior weights remain fixed as r changes. Because the law has 40 output atoms, each conditional operator has a known rank bound of 40 under either kernel. No infinite-rank tail assumption is needed for the certificates below.

We also define two controlled sampler laws on the same observations: deterministic MAP selects an atom with largest probability in equation 36, while the condition-independent fixture always outputs the same 50:50 mixture of the closest two dictionary images (indices 17 and 19). This fixed mixture tests diversity ranking while making no attempt to approximate the posterior or improve reconstruction quality.

Table 5 and Figure 3 show that the observation kernel changes G while the intended target remains fixed. The phenomenon also occurs with finite-dimensional normalized linear features. At the narrowest observation bandwidth, global redistributed top-2 and top-4 differences for the exact Gaussian-kernel posterior are -0.000531 and -0.002246. Global rank 8 gives 0.275526, whereas the target is 2.000490 and the average within-condition redistributed top-8 entropy is 1.705955. The untruncated value at that bandwidth agrees with the target to roundoff.

E.5 Selecting a useful sharpening power from observations alone

The median-distance observation bandwidth gives $G = 2.351500$ for the exact Gaussian-kernel posterior, versus its unchanged target $H_{\mathrm{post}} = 2.000490$. At that same bandwidth, the deterministic MAP fixture scores 0.584112, above the fixed two-image fixture's 0.301294. Thus the ranking problem is already present at this conventional bandwidth choice, without using an extreme endpoint of the sweep.

We turn the observation certificate into a selection rule. Let L denote the observation Gram matrix at median bandwidth and define

$$C(a) = \log 6 - S(L^{\circ a}/6), \qquad a_{\mathrm{exact}}(\eta) = \min\{a \in \mathbb{N} : a \geq 1,\ C(a) \leq \eta\}.$$

This choice uses only the six observations and their fixed weights. In the implementation, all selected powers are frozen before any output-state metric is evaluated. We compare it with the sufficient overlap-bound rule

$$a_\epsilon(\eta) = \max\left\{1, \left\lceil \frac{\log((e^\eta - 1)/5)}{2\log\epsilon} \right\rceil\right\}, \qquad \epsilon = \max_{i\neq j} |L_{ij}| = 0.89125222,$$

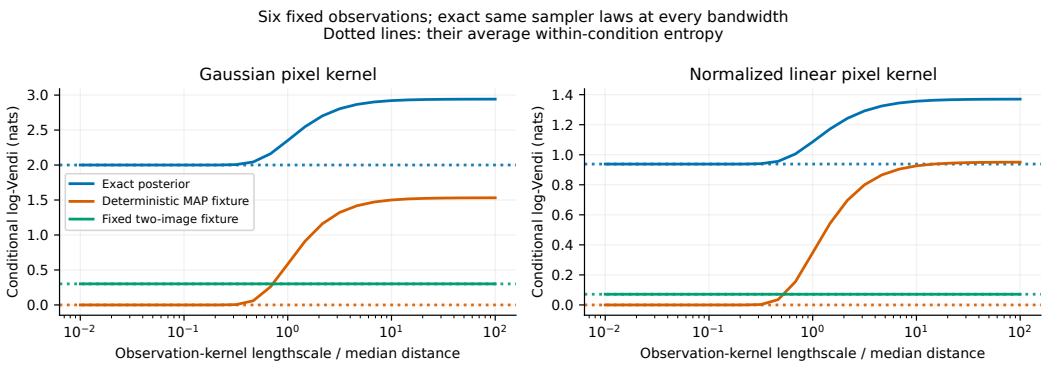

Figure 3: Conditioning-kernel dependence under two fixed output kernels. Solid curves are G and dotted lines are the corresponding unchanged H_{post}. The deterministic MAP and fixed-mixture curves cross as only the observation bandwidth varies. The image dictionary and all six observations are identical for every curve.

Tolerance η	$a_{\mathrm{exact}}(\eta)$	$C(a_{\mathrm{exact}})$	$a_{\epsilon}(\eta)$
0.10	6	0.072861	17
0.05	8	0.038088	20
0.01	13	0.009656	27

Table 6: Sharpening powers selected entirely from the observation kernel. The exact small-matrix certificate requires fewer powers than its worst-overlap sufficient bound. Neither rule consults an output law.

which ensures $\log(1 + 5\epsilon^{2a}) \leq \eta$. The exact conditioning certificate at the initial power one is 0.755421.

At tolerance 0.05, power eight reduces the exact posterior's paired entropy to 2.012000. Its certified target interval $[1.973912, 2.012000]$ contains the unchanged 2.000490. More decisively, the MAP target interval is $[0, 0.010799]$, whereas the two-image fixture's interval is $[0.263205, 0.301294]$. Their separation certifies the correct within-condition diversity ordering. The same observation-selected power also certifies this ordering with the normalized linear output kernel. This identifies the ordering of the declared finite-bank laws. Posterior fidelity and Monte Carlo error remain separate questions. Full evaluations, selection rules, and the observation-Gram hash are recorded in `experiments/conditioning/results/digits_certificate_selected_sharpening.csv` and the corresponding JSON file.

E.6 Blockwise approximation certificates and sample-budget controls

For each of the six exact conditional states, we retain its leading t eigenvalues and tail mass δ_j. The lower endpoint is the mean of $L_t(\rho_{y_j})$ from equation 11. The upper endpoint adds $\frac{1}{6}\sum_j \delta_j \log(40-t)$ when $t < 40$. These endpoints use the known finite dictionary rank bound and certify spectral approximation of the exact conditional law. They remain unchanged throughout the observation bandwidth sweep because the conditional laws are fixed. Sampling error for an unknown population operator requires an additional bound.

For the Gaussian output kernel, whose exact target is 2.00049041, Table 7 gives the resulting intervals. At $t = 16$ per condition, retaining 96 eigenvalues in total gives $[1.960244, 2.010342]$. At $t = 24$, retaining 144 gives $[1.992213, 2.001694]$. The observed narrow intervals follow from the measured tail masses in this dictionary, without an assumed general spectral decay.

We separately quantify finite-ensemble error by drawing multinomial counts from each exact conditional vector $Q_{j,:}$. For each $n \in \{8, 16, 32, 64, 128, 256, 512\}$ samples per condition, we generate 40 independent repetitions with seed 20270921. Both output kernels use the same saved empirical

Rank per condition	Total retained	Mean tail mass	Lower bound	Upper bound
8	48	0.09327735	1.774628	2.097903
16	96	0.01576371	1.960244	2.010342
24	144	0.00341951	1.992213	2.001694
32	192	0.00061313	1.999397	2.000672

Table 7: Blockwise retained-tail certificates for the exact Gaussian-kernel digit posterior. The target $H_{\mathrm{post}} = 2.00049041$ lies in every interval. These are deterministic spectral approximation bounds.

counts in each repetition. The estimator averages the six untruncated empirical entropies, using weights $1/6$. It involves no observation kernel or spectral truncation. Consequently this control isolates posterior sampling error from the previous two effects.

For the Gaussian output kernel, the mean estimates are 1.611207 at $n = 16$, 1.878643 at $n = 64$, and 1.981323 at $n = 512$, versus the exact target 2.00049041. Their standard deviations are respectively 0.089959, 0.049921, and 0.016461. These are standard deviations across 40 independent posterior-ensemble repetitions for a fixed six-observation bank, not standard errors or variability across observation banks. Raw counts, every repetition, and both kernels' summaries are distributed with the experiment code.

E.7 Data provenance and reproducibility files

The digit images are a subset of the 1,797-image test portion of the UCI Optical Recognition of Handwritten Digits dataset bundled with scikit-learn (Alpaydin & Kaynak, 1998; Pedregosa et al., 2011). The original dataset is credited to E. Alpaydin and C. Kaynak and is distributed by UCI under the Creative Commons Attribution 4.0 International licence. Our package includes only the 40 selected images, scaled by $1/16$, together with generated noise, observations, and derived matrices. This finite-dictionary exercise is not evidence about a learned natural-image posterior, a diffusion sampler, or a clinical imaging application.

The code archive includes the 40 selected image vectors, original dataset indices and labels, measurement mask, noise, observations, exact posterior weights and both output kernels. Its provenance manifest records checksums and the precise array-hashing convention. The experiment script regenerates every table and figure from these inputs, with its dependencies pinned in the repository.

Diversity and fidelity. For a stationary output kernel, translating all outputs preserves the entire within-condition Gram spectrum. Spectral diversity therefore cannot establish agreement with a specified posterior or alignment with a text prompt by itself. The target is variation in the stated feature representation, and the audit requires a separate fidelity diagnostic.

F Additional evaluation checks

F.1 Observation-design stress suite

The robustness suite was designed after the initial diagnostic experiment. Its declared grid is the Cartesian product of three masks, noise standard deviations $\{.15, .35, .6\}$, and seeds $\{101, 102, 103, 104, 105\}$. Each setting has six fixed observation instances, giving 45 settings and 270 condition instances. The dictionary and Gaussian output kernel are unchanged. The masks are the original pixel indices $[18, 21, 34, 37]$, alternative central pixels $[19, 20, 35, 36]$, and four uniformly selected indices fixed by seed 20270923. All indices are zero-based in row-major order. The exact selected indices, truth atoms, noise arrays, observations and hashes are included in the experiment directory.

For each setting, we compute the exact finite-dictionary posterior, its deterministic MAP law, and the observation-independent uniform pair of dictionary images with indices $[17, 19]$. This pair and the output bandwidth 1.5355488045 are fixed throughout. The pair's entropy is $.3012938546$ nats. We start from the median-distance observation kernel and select the smallest positive integer power whose exact certificate is at most $.05$ nats. Selection uses only condition probabilities and the

observation kernel. The experiment records the selected power, each exact within-condition entropy, each joint-minus-observation score, and the identification interval for every fixture.

The initial score places MAP above the pair in all 45 settings. Selection reverses this erroneous diversity ordering and certifies the pair above MAP in all settings. Powers range from 5 to 41, and the achieved certificate ranges from .0286876305 to .0497760109. The minimum separation between the pair's lower endpoint and MAP's upper endpoint is .2030493603 nats. There are no failed selections or numerical bound violations. This is consistent with the theorem: MAP has zero conditional entropy, so the common certificate and the pair's fixed positive entropy permit a correct ordering at a sufficiently small tolerance.

Every declared setting is retained. For each seed, the same standard-normal noise array is reused across masks and noise levels, coupling the 45 settings. We therefore report descriptive counts without binomial confidence intervals. The suite tests sensitivity to the observation design within the fixed dictionary. Frequency estimates beyond this construction, trained-generator comparisons and posterior-quality evaluation require separate experiments.

The portable directory `experiments/conditioning_robustness/` contains the declared design, saved inputs, all outcomes, a provenance manifest and the executable script.

F.2 Equal retained-head budgets and a global tail comparator

The budget comparison uses the same six exact conditional laws and both fixed output kernels. For $T \in \{24, 48, 96, 144, 192, 240\}$, global redistribution keeps the leading T joint eigenvalues and subtracts the exact observation entropy. Blockwise compression keeps $T/6$ eigenvalues per condition. The marginal rank is six, so retaining it exactly avoids an additional confound. We evaluate orthogonal observation features, where $G = H_{\mathrm{post}}$, and the original median-bandwidth features, where approximation to G and discrepancy from H_{post} must be reported separately.

Retained eigenvalue counts are equal. In addition, the global spectral summary stores one joint tail mass and six marginal eigenvalues, giving $T + 7$ scalar slots. The blockwise summary stores six tail masses, giving $T + 6$ slots. Condition weights, rank bounds and kernel inputs are common and excluded from these generic counts. Under orthogonal conditioning the marginal spectrum is the already-shared condition-probability vector, so a specialised global implementation needs only $T + 1$ additional spectral scalars. A zero tail slot is counted even at full rank. All spectra are computed fully for validation. The comparison measures retained spectral information, while runtime, memory use and partial-eigensolver complexity depend on the implementation.

Tail aggregation is also available globally. If the joint rank is at most $R = 240$, the retained-head entropy plus one tail bin gives a lower bound on $S(\rho_{AB})$, with remainder at most $\delta \log(R - T)$. Subtracting exact $S(\rho_B)$ yields an interval for G. This is also an interval for H_{post} when observation features are orthogonal. For nonorthogonal features it retains the conditioning discrepancy and must be combined with its certificate before targeting H_{post}. Blockwise summaries directly target the conditional average and allow condition-specific ranks and tail budgets.

The more accurate approximation depends on the budget. In the Gaussian-output, orthogonal-conditioning case, the blockwise midpoint error at $T = 96$ is .015198, below the global redistributed error .043600. At $T = 192$, global redistribution has error .000349, below the blockwise midpoint error .000456. Global tail aggregation also gives a narrower interval than equal-allocation blockwise aggregation at $T = 144$ and 192. The benefit of blockwise compression here is its direct conditional decomposition and explicit error accounting, with no uniform accuracy or efficiency advantage.

The full 24-row table, both error targets, all interval endpoints, saved validation spectra and input hashes are in `experiments/truncation_budget/`. Every interval was checked against the exact operator target, and the full-budget results recover the corresponding entropy to numerical precision.

F.3 A conservative staged comparison with a known population rank

This categorical control evaluates the interval protocol in Proposition D.2 and Corollary D.4, including the cost of its conservative concentration bound. We fix the laws $P = (1/2, 1/2)$ and $Q = (0.95, 0.05)$, with a feature map consisting of two orthogonal unit vectors. The population rank

Draws per target	Interval for S_P	Interval for S_Q	Decision
16	$[0, 0.693147]$	$[0, 0.693147]$	Unresolved
64	$[0, 0.693147]$	$[0, 0.693147]$	Unresolved
256	$[0, 0.693147]$	$[0, 0.693147]$	Unresolved
1 024	$[0.208350, 0.693147]$	$[0, 0.672028]$	Unresolved
4 096	$[0.400095, 0.693147]$	$[0, 0.487534]$	Unresolved
16 384	$[0.521066, 0.693147]$	$[0.026634, 0.370745]$	$S_P > S_Q$ certified
65 536	$[0.594237, 0.693147]$	$[0.100381, 0.298208]$	$S_P > S_Q$ certified
262 144	$[0.637187, 0.693147]$	$[0.141905, 0.253821]$	$S_P > S_Q$ certified

Table 8: Conservative population-entropy intervals with one 0.05 error budget shared across two targets and all stages. The first disjoint stage requires 16 384 independent draws per target. Endpoints are in nats.

bound $r = 2$ is therefore known independently of the samples. Their exact entropies are

$$S_P = \log 2 = 0.69314718, \qquad S_Q = h_2(0.95) = 0.19851524, \qquad S_P - S_Q = 0.49463194.$$

These known values audit the results but are not supplied to either interval construction or the stopping decision.

The predeclared sample counts per target are $16, 64, 256, 1024, 4096, 16384, 65536, 262144$. With fixed seed 20270930, we generate independent multinomial increments for each target at every stage and accumulate category counts. This produces the same sufficient statistics as nested prefixes of independent category draws, without storing large sample arrays or Gram matrices. For overall failure budget $\delta = 0.05$ and $L = 2$ targets, target j at stage t receives $\delta_{j,t} = 6\delta/(\pi^2 L t^2)$. The sum over both targets and all positive-integer stages is 0.05. Thus the simultaneous coverage and valid first-disjoint-interval stopping guarantee follow from Corollary D.4. Independent stages are not assumed.

The experiment calls the distributed `entropy_population_interval` function on the empirical entropy, sample count, known rank bound, and allocated failure probability, and calls `stage_error` for error spending. At $n = 16$, both sufficient-statistic intervals were independently checked against the explicit 16×16 Gram-matrix implementation. The population rank bound is never replaced by the observed sample rank.

Table 8 shows that the comparison remains unresolved at 4096 draws per target. It is first certified at stage six, with 16 384 draws per target and 32 768 total. Every displayed interval covers its known target in this one fixed-seed run, which checks the numerical implementation. Coverage is guaranteed by the theorem and error allocation, rather than estimated from this single run. The large sample requirement exposes the conservatism of the supplied bound, even for a rank-two law with a substantial entropy gap.

The script and all results are provided in `experiments/sampling_certificate/`. Files include every interval and allocated failure probability, raw multinomial increments and cumulative counts, the first disjoint stage, and SHA-256 hashes of the script and both imported public implementation modules. Running `python experiments/sampling_certificate/run_sampling_certificate.py` from the project root reproduces the table.

F.4 Designed controls for the population distortion witness

The experiment evaluates Corollary B.8 on known density states, so the true population gap is available without entropy estimation. Let $d = 32$, $s = .8$, $v_0 = e_0$ and $v_1 = \cos(\theta)e_0 + \sin(\theta)e_1$. Define

$$\rho_j = s v_j v_j^* + (1 - s) I_d/d, \qquad \theta \in \{0, \pi/6, \pi/3\}.$$

The two conditions have equal weights. Each state is realised by a law assigning mass $.8$ to v_j and mass $.2/32$ to each standard basis vector. These are 33 labelled atoms, with coincident feature atoms permitted. The states have rank 32 and trace distance $.8 \sin\theta$. For nonzero angles they do not commute, with commutator Frobenius norm $.391918$, while their difference has rank two.

The conditioning vectors are $(1, 0)$ and $(c, \sqrt{1-c^2})$ for $c \in \{0, .2, .8, 1\}$. The witness M is the positive spectral projection of the known population difference $\rho_0 - \rho_1$, computed before sampling. It attains the trace distance. At $\theta = 0$, $M = 0$. This control uses the oracle effect throughout. At each of $n \in \{64, 256, 512\}$ independent outputs per condition, we draw 2,000 independent multinomial repetitions. The complete grid has 36 settings and 72,000 repetitions. Nominal 95% coverage applies separately to each setting, with no simultaneous guarantee over the grid.

Table 9: All population-witness settings. Each row includes three sample sizes. $L_\star = \lambda(c)D_{01}^2/2$ is the analytic bound and $\widetilde{L}_n$ the median sampled lower confidence bound. P_{64} is the fraction of positive bounds at $n = 64$. At $n = 256$ and 512 that fraction is one for every nonzero-gap row and zero for every null row. All 2,000 repetitions cover in every cell. Tiny numerical null residues are displayed as zero.

θ	c	Δ	$L_\star$	$\widetilde{L}_{64}$	$\widetilde{L}_{256}$	$\widetilde{L}_{512}$	P_{64}
0	0	0	0	0	0	0	0
0	.2	0	0	0	0	0	0
0	.8	0	0	0	0	0	0
0	1	0	0	0	0	0	0
$\pi/6$	0	0	0	0	0	0	0
$\pi/6$	.2	.003799	.002151	.000014	.000629	.000973	.7815
$\pi/6$	.8	.076750	.040450	.000204	.011620	.018370	.7755
$\pi/6$	1	.177021	.080000	.000521	.023191	.036331	.7835
$\pi/3$	0	0	0	0	0	0	0
$\pi/3$	.2	.011364	.006452	.001386	.003475	.004248	1
$\pi/3$	.8	.214631	.121350	.026238	.065027	.079892	1
$\pi/3$	1	.423651	.240000	.052230	.129138	.158072	1

Eighteen cells force a zero certificate through identical states or orthogonal conditioning. All 18 nontrivial cells also cover in 2,000 of 2,000 repetitions. The lower endpoint of the two-sided 95% Wilson interval for each such coverage frequency is .998083. These frequencies check the implementation on the designed laws, while the theorem supplies the coverage guarantee. At $\theta = \pi/6$ and $n = 64$, positive distortion goes undetected in 21.65 to 22.45 percent of repetitions. All failures are retained. The illustrative $\theta = \pi/3, c = .8, n = 256$ cell was not designated as a primary contrast in the design.

The random seed is 20260925. The code obtains population entropies by symmetric eigendecomposition, discards eigenvalues below 10^{-13} and allows 2×10^{-12} for numerical comparisons. The lower bounds use the exact stated overlap and weights. The saved design, executable, all 36 result rows and validation record accompany the release. The design hash identifies the specification but does not independently establish chronology. This experiment evaluates a distortion certificate on known laws. It leaves pretrained-generator performance, population entropy, rankings and semantic fidelity untested.

F.5 Retrospective learned witnesses on the Qwen bank

This feasibility check reuses the already analysed confirmation banks. We greedily match the 32 fixed prompts by descending absolute native-kernel overlap, with ties broken by prompt indices and no repeated vertices. The resulting 16 pairs use input information only. For each temperature and pair, we learn M as the positive spectral projector of the difference between the two empirical covariance operators on half A, comprising draws 0 to 31. We evaluate it only on half B, comprising draws 32 to 63, using the fixed CLIP output features. The projector is computed by reduced QR and eigendecomposition with eigenvalue cutoff 10^{-12}. The computation uses this single kernel, witness procedure and split, with no new generated outputs.

The formula allocates $\alpha = .05/32$ to each of 32 pair-temperature tests. With 32 evaluation draws per prompt, the summed Hoeffding radius is .700346. For a procedure fixed independently of its evaluation data, and under iid conditional sampling, Corollary B.8 would give simultaneous 95% coverage and permit each matching sum. Here the design was fixed before this computation but after the banks had been analysed. Earlier adaptive choices are outside the nominal calculation, so we treat these values as retrospective feasibility evidence.

Table 10: All 32 retrospective learned-witness values, shown by prompt pair and temperature. Indices refer to the fixed zero-based prompt list. $L_{.7}$ and $L_{1.1}$ are the nominal pairwise expressions in equation 19, multiplied by 10^6 for readability. All zero values are retained.

Prompt y	Prompt z	Overlap	$10^6 L_{.7}$	$10^6 L_{1.1}$
3	16	0.776278	0.000000	0.000000
11	20	0.766021	0.000000	0.000000
9	15	0.748529	0.000000	0.000000
1	18	0.742620	0.000000	0.000000
5	12	0.723720	0.000000	0.000000
6	28	0.707779	0.000000	0.000000
22	25	0.706500	0.000000	0.000000
7	30	0.700762	0.000000	0.000000
10	27	0.694693	0.000000	0.000000
8	13	0.678444	0.545636	0.000000
4	23	0.660199	14.064242	8.046917
0	31	0.655324	0.000000	0.000000
17	29	0.648394	0.264085	0.000000
14	26	0.634765	6.261575	0.000000
19	21	0.577178	62.516365	10.339625
2	24	0.472597	61.556479	32.410702
Matching sum			145.208382	50.797244

Six of 16 pairwise values are positive at temperature .7 and three at temperature 1.1. The sums are .000145208382 and .000050797244 nats, respectively. The learned witness is computable, but these small values do not usefully estimate the full gap. A zero value leaves open whether distortion vanishes, and comparing lower bounds cannot order the true distortions across temperatures. The release retains both source hashes, all effects, 2,048 evaluation witness values and every pairwise result. Population entropy rankings and semantic fidelity remain unassessed by this retrospective check.

G Pretrained-generator protocols and complete outcomes

G.1 Scope and fixed empirical targets

For each system and prompt y, the bank contains n_y outputs with normalised kernel diagonal. We compute

$$\widehat{H}_{\text{group}} = \sum_y p_y S(K_y/n_y), \qquad \widehat{G} = S(J) - S(\sqrt{p} L \sqrt{p}),$$

where J is the trace-one, sample-weighted joint Gram matrix: sample (y, j) has weight p_y/n_y. Its $(yj, z\ell)$ entry is $\sqrt{p_y p_z/(n_y n_z)}\, k(x_{yj}, x_{z\ell}) L_{yz}$. All reported banks use uniform prompt weights and equal group sizes. The full grouped score and the orthogonal prompt-label baseline agree numerically. The interval $[\max(0, \widehat{G} - C), \widehat{G}]$ contains this finite-bank target for every admissible output kernel. It is not a sampling confidence interval. Choosing the output RBF bandwidth from a pooled bank changes the empirical similarity definition. Those cases are explicitly conditional on that bank.

G.2 Released GenEval images: design and provenance

The original GenEval human-study release (Ghosh et al., 2023) supplies 100 common prompt IDs with four images per system. We retain their complete intersection across Stable Diffusion 2.1, IF-XL and CLIP retrieval, determined from the original annotation records without selecting on ratings or diversity. The public directory name `sdv2` refers to the Stable Diffusion 2.1 human-study system described in GenEval. CLIP retrieval is the benchmark's LAION-5B retrieval comparator, not generated data.

Table 11: Post hoc resolution follow-up over every SD 2.1 versus IF-XL specification. Each tag contributes 16 settings. A certificate means disjoint deterministic intervals for the fixed empirical banks. The tolerance grid was fixed after inspecting the primary .05 results. It does not replace those results.

	Raw	Originally reversed comparisons certified at tolerance		
Group	reversals	.05	.01	.001
Whole bank	7	0	7	7
Colour attribute	2	0	2	2
Colours	4	0	4	4
Counting	6	0	6	6
Position	6	0	5	6
Single object	0	0	0	0
Two objects	6	0	6	6
Total	31	0	30	31

The dataset and original annotations are from the public GenEval release. Exact revisions, source paths and image checksums are recorded in the accompanying provenance manifest. Downloaded images were verified against the upstream records. Four malformed IF-XL annotation records contain unquoted line breaks in a free-text object description. Joining each literal continuation to its preceding record recovers the expected 22 fields and all 2,000 judgments for that system. The original files and exact repairs are retained. No rating is imputed or removed. The final total is 6,000 judgments, five per image.

The upstream field `Answer.task-quality` asks caption fit on a 1 to 4 scale, as verified against question 5 of the original human-study form. It is not an aesthetic or human-diversity score. Mean caption fit is 3.344 for SD 2.1, 3.546 for IF-XL and 3.1035 for retrieval. These annotations do not define a desired diversity ranking.

We use the projected 512-dimensional CLIP ViT-B/32 image representation and the 384-dimensional DINOv2-small final CLS representation. Each vector is normalised to Euclidean unit norm. The same CLIP checkpoint encodes prompts. The repository supplies both original and normalised vectors and pins the encoder versions and preprocessing configuration.

The fixed output specifications are a linear kernel and a Gaussian kernel in each image representation. Gaussian bandwidths are the median pooled pairwise feature distance across all three systems: $.925524$ for CLIP and 1.388048 for DINOv2. Prompt kernels are text cosine similarity and Gaussian kernels at $.5, 1, 2$ times the median prompt distance $.920798$. A single prompt kernel and its input-selected power are shared across systems and output representations within each tag group. No bandwidth is selected to cause a reversal. The original six task tags, plus the whole bank, are all retained.

Complete comparisons and negative outcomes. The 4 output specifications, 4 prompt kernels, 7 groups and 3 system pairs give 336 primary comparisons. There are 31 raw point-ranking reversals, all in the 112 SD 2.1 versus IF-XL comparisons. 7 occur in the 16 whole-bank native-generator comparisons. There is no reversal involving retrieval. These are correlated specifications of shared data. At the primary tolerance $.05$, the selected powers remove every reversed point ordering but certify none of those 31 comparisons. Of all 112 native-generator settings, 52 have disjoint selected intervals. These 52 had already agreed in raw point ordering. No certified direction contradicts the exact finite-bank target.

The exploratory follow-up fixes the grid $.05, .01, .001$ for all 112 native-generator settings before computing that follow-up. At .01, 111 of 112 settings have disjoint intervals, including 30 of the 31 originally reversed cases. At .001, all 112 do, including all 31 reversals (Table 11). The main CLIP linear/median-prompt-kernel case explains the change: at .05, the selected gap $.022661$ is smaller than $C = .040583$. At .01, power 16 gives $C = .009066$ and disjoint intervals $[.429958, .439024]$ for SD 2.1 and $[.407185, .416250]$ for IF-XL. Finer resolution bounds the conditioning ambiguity

more tightly on the same samples. The population sampling error remains outside these deterministic intervals.

The median-kernel DINOv2 comparison in Table 2 does not reverse. Its only whole-bank reversal uses the Gaussian output kernel and widest prompt bandwidth, with a small raw margin of approximately -2.86×10^{-5} nats. Reversal counts therefore include effects of quite different magnitudes. For the CLIP linear grouped SD 2.1 versus IF-XL difference .022686, 2,000 paired prompt resamples give the 2.5th to 97.5th percentile range $[-.005907, .049167]$. This describes resampling of the selected prompt bank and supplies no coverage guarantee for unseen prompts or population conditional entropy.

Sampling, compression and simpler baselines. For each output specification and system, the release enumerates all common subsets of one, two, three and four sample IDs per prompt: $4, 6, 4, 1$ subsets respectively, giving 180 finite-bank sensitivity rows. In particular, one retained image has zero within-prompt entropy by construction while a smooth-condition score can remain positive. No subsample is treated as an independent generator rerun. The release also retains every declared global spectral budget, with the full spectrum and exact-label conditioning as controls, and per-prompt same-feature purity, leading eigenvalue and dispersion. These baselines use the same representation as entropy. Direct groupwise entropy is the simplest exact baseline when prompt identities and repeated outputs are available. Sharpening does not claim a computational advantage over it.

G.3 Fresh Qwen3 inference: complete design and outcomes

We use Qwen3-0.6B with the checkpoint pinned in the provenance manifest. The design file, created before analysed generations, specifies 16 descriptive prompts, eight sampled outputs per prompt and temperatures .7 and 1.1. Each prompt requests two concise sentences describing a fixed scene. The full verbatim list is released. Sampling uses top-$p = .95$, top-$k = 20$, thinking disabled, float32 CPU inference and at most 40 continuation tokens. The batch seed is $13001 + 1000t + j$, where $t \in \{0, 1\}$ indexes temperature and $j \in \{0, \ldots, 7\}$ indexes the replicate. Every batch has the same 16-prompt order. An unrelated compute pilot asks for the numbers one to five and retains the full predeclared design before any analysed outputs are generated. Generation takes 402.8 seconds using two CPU threads in the recorded environment.

We additionally produce one greedy continuation per prompt and repeat it eight times to form a labelled zero-diversity control. This creates 16 unique greedy outputs, not 128 independent greedy draws. Every stochastic output is retained, including all token-limit events: 9 of 128 outputs at temperature .7 and 6 of 128 at temperature 1.1 reach the 40-token cap without EOS. There are no empty outputs. All 128 texts at each temperature are distinct. All greedy outputs terminate naturally. Raw tokens, decoded text, seeds and termination flags accompany the source and checkpoint hashes.

CLIP text features use the same fixed official checkpoint as the image experiment. The three output specifications are CLIP cosine, CLIP Gaussian with the pooled-bank median bandwidth .884294, and character 3 to 5 gram TF-IDF cosine. TF-IDF is fitted to the pooled realised text bank and is a non-pretrained representation sensitivity check. Prompt features use five Gaussian bandwidths $.25, .5, 1, 2, 4$ times their median distance .690618, plus constant and exact-label kernels. The smallest integer power of the median prompt kernel with $C \leq .05$ is included as an eighth specification: $a = 6$, $C = .031825$. These choices yield 72 score rows, all satisfying the conditioning bound, and 24 natural-temperature comparisons. All 24 preserve the grouped ordering. All 72 pair comparisons, including greedy comparisons, also preserve it.

The raw CLIP cosine intervals are $[0, .959457]$ and $[0, .993842]$. Power 6 gives disjoint intervals $[.340779, .372604]$ and $[.391059, .422884]$, certifying the existing empirical temperature ordering. The constructed greedy bank has exact effective diversity $\exp H = 1$. The default smooth score gives $\exp G = 1.8617$, reduced to 1.0168 after sharpening. This control concerns deterministic decoding, separate from stochastic Qwen3 generation. Output quality was not assessed by human raters, so the temperature ordering is interpreted only as feature diversity.

Table 12: All Qwen3 output representations at the median prompt bandwidth. Raw conditioning has $C = 1.163296$. Input-selected power 6 has $C = .031825$ for every row. All values are finite-bank log-diversities in nats.

Output kernel	Decoding	H	G	G_6
CLIP cosine	.7	.356280	.959457	.372604
	1.1	.406892	.993842	.422884
	Greedy	0	.621465	.016626
CLIP Gaussian	.7	.413242	1.020833	.429886
	1.1	.470923	1.064660	.487252
	Greedy	0	.629036	.017149
Character TF-IDF	.7	1.731152	2.876061	1.762658
	1.1	1.900455	3.046452	1.931993
	Greedy	0	1.157586	.031750

G.4 Reproduction and remaining empirical scope

The repository provides CPU replay from the supplied features, complete comparison tables and the original text continuations. Generation and feature extraction have separate instructions, with pinned sources and the original annotation repairs. Repeating image extraction requires downloading the 1,200 source images and encoder weights. The recorded analysis plans distinguish the initial results from the later GenEval resolution refinement.

The image banks use 2023 diffusion releases, while the text experiment uses newly generated outputs from the 2025 Qwen3-0.6B model. The follow-up in Appendix H increases the text ensemble size to 64 on new prompts, and the amended study in Appendix I examines one held-out image-bank selection. Together these studies test finite-bank distortion in specified representations. They leave population entropy ranking, posterior calibration, broader generator coverage and generation-quality gains unresolved. CLIP and DINOv2 expose representation sensitivity, which is visible in the different margins and preserved as well as reversed orderings.

G.5 Direct scoring under the same information budget

All evaluated banks retain prompt labels and repeated outputs. Direct grouped entropy therefore uses exactly the same observations and feature arrays as the smooth conditional score. It computes the empirical target with no conditioning error. An orthogonal-label joint computation gives the same answer by the direct-sum entropy identity, but explicitly constructing its dense matrix is unnecessary. Neither approach removes within-prompt sampling bias relative to a generator population.

We conducted an exploratory CPU cost audit on all three image systems, both representations, and both declared output kernels. Each bank contains 100 groups of four outputs. Twenty-one timed repetitions follow a warm-up, using one CPU thread. Hardware and software versions are recorded with the benchmark. The timing starts from the same normalised features. It excludes image generation, feature extraction, shared bandwidth definition, and row gathering equally. Prompt-kernel preparation and power selection cost 6.05 ms once and are recorded separately. The comparison below reports medians in milliseconds for the linear output kernel.

Features	System	Direct groups	Dense G	Dense G_{10}
CLIP	SD 2.1	.305	8.269	8.014
CLIP	IF-XL	.305	8.328	8.105
CLIP	Retrieval	.343	8.840	8.355
DINOv2	SD 2.1	.269	8.016	7.796
DINOv2	IF-XL	.263	7.936	7.803
DINOv2	Retrieval	.258	7.811	7.628

Across all twelve representation, output-kernel and system settings, the dense raw-to-direct runtime ratio ranges from 24.26 to 31.30. The selected-power ratio ranges from 23.53 to 29.89. These implementation-specific CPU measurements favour direct scoring when groups are available. The full

timing distribution and exact-label equivalence checks accompany the code. The certificate addresses the interpretation of the smooth statistic.

Independent numerical reanalysis. A separate implementation recomputed all 336 image score rows, all 336 resolution-follow-up settings, and all 72 text score rows from the supplied feature arrays. Maximum absolute discrepancies from the recorded results were 3.6×10^{-15} for images and 5.5×10^{-14} for text. Direct grouping agreed with orthogonal-label scoring to 1.3×10^{-15}. The primary CLIP prompt bootstrap reproduced the descriptive range $[-.005907, .049167]$. This is numerical replication from fixed features, not a repeat of raw-image extraction or language-model inference. The stored image grid was recorded before scoring, but it was not an externally timestamped preregistration. The correlated grid and tag comparisons do not estimate a population frequency of ranking reversals.

H Prospective confirmation on fresh output banks

H.1 Design fixed before generation

The confirmation uses 32 new descriptive prompts with no verbatim overlap with the 16 prompts of the initial Qwen study. The design was recorded on 19 September 2026. Its source and protocol hashes were frozen before the run on 22 September 2026. This is a prospective design with local provenance records, not an externally registered preregistration. The earlier GenEval tolerance follow-up remains post hoc.

The generator and encoder are the same pinned Qwen3-0.6B and CLIP ViT-B/32 checkpoints as in Appendix G. Their full revisions are also recorded in the confirmation protocol. All CLIP vectors are normalised, and the output kernel is their inner product. It has feature dimension 512 and no bandwidth fitted to the generated outputs. Generation uses float32 precision, eager attention, thinking disabled, top-$p = .95$, top-$k = 20$, a cap of 64 new tokens, and batches of 16 continuations. For each prompt, it generates 64 stochastic outputs at each temperature .7 and 1.1. Distinct batch seeds are deterministically derived from the frozen identifiers. A separately generated greedy completion at each prompt supplies a deterministic control. Repeating a greedy output for algebraic comparison does not create independent samples.

The conditioning kernel is Gaussian on normalised prompt features. Its bandwidth is the median nonzero pairwise prompt distance. Before any study generation, the procedure selects the smallest positive integer power giving $C \leq .01$. The selected bandwidth is .723690, the native certificate is 1.536870, and power 9 gives $C = .0052082007$. No study outputs enter this selection. The numerical implementation uses a binary search over powers up to 65,536. Failure to reach the fixed tolerance would have been retained as a failed selection.

The primary bank uses all 64 samples per temperature and prompt. The only primary ranking contrast is temperature 1.1 minus temperature .7. Primary outcomes are each temperature's conditioning error $G - H$, certification of that contrast, and the runtime of direct and joint scoring on the same features. Secondary banks use the first eight samples and two disjoint halves, A with indices 0 to 31 and B with indices 32 to 63. These are repeated samples at the same fixed prompts. The halves are disjoint, while the first-eight and all-64 banks overlap. All declared comparisons are retained regardless of direction or certification status.

H.2 Complete finite-bank results

The run contains all 288 expected generation batches and 4,128 outputs: 4,096 stochastic continuations and 32 greedy controls. Neither GPU worker reused prior outputs. The recorded diagnostics show zero empty decoded outputs, zero CLIP truncations and zero generations without an end-of-sequence token. These checks describe termination and preprocessing, not semantic quality or instruction adherence.

Every native and selected comparison preserves the direct temperature ordering, and every selected identification interval excludes zero. The prospective tolerance therefore permits identification even though the native ranking was already correct. At 64 samples, the native score overstates the two direct entropy targets by .760742 and .733045 nats. The selected errors are .002565 and .002496

Table 13: Prospectively specified banks and the sole temperature contrast. Δ is the value at temperature 1.1 minus the value at .7. The last column bounds ΔH for the realised bank. It is not a confidence interval for generator-population entropy.

Bank	Direct ΔH	Native ΔG	Selected ΔG_9	Identification interval
All 64, primary	.109649	.081953	.109581	$[.104373, .114789]$
First 8	.074861	.057806	.074848	$[.069640, .080056]$
Half A, 32	.093180	.071389	.093124	$[.087916, .098333]$
Half B, 32	.099972	.075209	.099919	$[.094711, .105127]$

nats. The greedy control has direct entropy zero, native score .849561 and selected score .002853. These comparisons assess feature diversity, with output quality left unevaluated.

Increasing the bank size changes the empirical target. In the first-eight bank, the direct entropies are .366288 and .441148, compared with .513350 and .622999 for all 64 outputs. This material sample-budget sensitivity falls outside the deterministic identification bound, which controls conditioning error at each bank size. Population inference requires the additional calculation below.

H.3 Population uncertainty

For $m = 32$ fixed prompts, $n = 64$ independent stochastic samples per prompt and known feature rank at most $d = 512$, the rank-bias and bounded-difference argument in Appendix D gives

$$r_\alpha = h_2(1/n)\sqrt{\frac{n}{2m}\log\frac{2}{\alpha}}, \qquad b_{n,d} = \log\left(1 + \frac{d-1}{n}\right),$$

and the interval

$$\left[\max(0, \widehat{H}_{\text{group}} - r_\alpha),\ \min(\log d, \widehat{H}_{\text{group}} + r_\alpha + b_{n,d})\right].$$

Each temperature uses $\alpha = .025$. A union bound therefore gives simultaneous coverage at least .95 under the stated sampling assumptions. No effective empirical rank is substituted for the known dimension. For the two temperatures, respectively, the intervals are $[.344869, 2.877318]$ and $[.454518, 2.986968]$. Subtraction yields

$$[-2.422800, 2.642099]$$

for the difference in average population feature entropy. It includes zero. The target averages over the fixed 32 prompts and supports no inference to a population of unseen prompts. The calculation assumes independent draws from each specified generation law. Recorded seeds and reproducible hardware do not themselves prove that assumption. Greedy duplicates are excluded from this calculation.

H.4 Resources and the direct-scoring comparison

The study ran on two NVIDIA A40 GPUs. A separate compute pilot discarded its 80 counting-prompt outputs. The completed job occupied both GPUs for 254 seconds, giving .141111 allocated GPU-hours. This allocation includes idle time within the job and excludes queue time and separate preparation. The repository records the software environment, individual process timings, peak device memory and scheduler accounting.

The prospectively specified runtime outcome uses one timed pass per bank. For the primary temperature .7 bank, direct grouped entropy takes 12.580 ms, compared with 389.930 ms for native joint entropy and 366.785 ms for selected joint entropy. For temperature 1.1, the corresponding times are 7.012, 372.270 and 368.950 ms. All start from the same features. The joint methods construct matrices of dimension 2,048, whereas direct scoring diagonalises separate 64-sample group matrices. This implementation comparison favours direct scoring. A single pass supplies no runtime uncertainty estimate. Appendix G.5 reports a separate repeated image benchmark.

H.5 Provenance, reproduction and interpretation

The repository retains the original protocol and subsequent completion records, together with the prompts, generated text, tokens, features and complete analysis tables. A separate implementation recomputed the conditioning certificate, every bank score, the greedy control and the population intervals from the archived arrays. Its maximum absolute discrepancy was below 2.3×10^{-15}. This check covers numerical replay from fixed features. Model inference and text-to-feature extraction were not repeated independently.

The experiment confirms finite-bank identification at a tighter, prospectively fixed tolerance on new data from one small pretrained model, one representation and 32 chosen prompts. The native temperature ranking stays unchanged and the population ordering remains unresolved. Direct grouped scoring computes the observed target exactly. Generation quality and Bayesian posterior fidelity are outside this design, and the new study leaves GenEval's original abstentions and post hoc follow-up unchanged.

I Amended fresh-image decision illustration

I.1 Protocol history and the amended question

This study asks whether an aggregate model choice from one fresh image bank transfers to a second bank under a fixed CLIP-diversity target. The SD 2.1 versus IF-XL pair was chosen because it already exhibited a GenEval reversal. The new protocol specifies its own generation settings, with historical source differences recorded below.

The original plan used 24 pilot prompts with four images per model, giving 192 images. A prespecified variance and resource rule required 102 confirmation prompts, exceeding the largest candidate panel of 96. Candidate allocations were also subject to time and GPU-hour caps. The recorded outcome was `stop_pilot_only`. Before the amendment, the pilot direct SD 2.1 minus IF-XL mean of .050131 nats and standard deviation .108992 had been inspected. Sixteen prompt contrasts were positive and eight negative. Only the standard deviation entered the planning rule. Native pilot scores were not computed. The heuristic concerned a mean four-sample direct contrast and gave no power guarantee for the later selection-policy gain.

The subsequent amendment retains this stopped outcome and declares a smaller finite-bank illustration. It uses the first 24 prompts of the original locked, unused 96-prompt confirmation panel, with four images per prompt, two models and two banks, giving 384 new images. No pilot images enter these banks. The prompts, original A/B seeds, models, kernels, tolerance and decision rules are unchanged. No A/B outputs had been seen when the amendment was locked on 25 September 2026. This is a prospective amendment after observed pilot results, not completion of the original confirmation design or an externally registered preregistration.

I.2 Fixed prompts, generation and scoring

Prompt selection excludes the 100 previously analysed GenEval image prompts by exact text. Within each of the six task tags, prompts are sorted by a fixed hash and the tags are interleaved. The pilot uses the first 24 prompts. The amendment uses the next 24, with four prompts per tag and uniform weights. The full prompt list and hashes are retained. The target concerns these 24 fixed prompts, not a distribution of unseen prompts.

SD 2.1 uses a pinned community mirror adopted before the pilot because the original repository was unavailable. Architecture and prediction configuration were checked, but full historical pipeline, tokenizer and configuration byte equivalence is not established. The IF pipeline combines DeepFloyd IF-I-XL, IF-II-L and the Stable Diffusion x4 upscaler. The provenance manifest records the exact sources and checkpoints.

Generation uses half precision, model-bundled schedulers, CPU offload and batch size one. The repository pins the Diffusers and PyTorch environment. SD 2.1 uses 50 steps, guidance 7.5 and 768×768 pixels. The three IF stages use 100, 50 and 75 steps with guidance 7, 4 and 9, producing 64×64, 256×256 and 1024×1024 outputs. The last two noise levels are 250 and 100. Each image has a separately seeded CPU generator, consumed sequentially by the IF stages. There are no

prompt additions, quality retries, content exclusions, replacement models or adaptive extra banks. All returned images and flags are retained. Intermediate tensor-stage flags can be null and are not represented as separate safety checks.

The fixed CLIP ViT-B/32 checkpoint from Appendix G encodes images in float32 on CPU, followed by normalisation in float64. The output kernel is the inner product of the 512-dimensional unit vectors. The native prompt kernel is Gaussian on normalised CLIP text features with median nonzero prompt distance $\sigma = .905071$. It is the earlier declared specification, not a universal default of the published metric. Before pilot generation, prompt-only selection chooses the smallest integer power with $C \leq \eta = .01$. Power 10 gives $C = .0082346235$, compared with the native $C = 1.4142102084$. Both banks retain these same kernels. No truncation is applied.

Direct scoring averages the four-image group entropies, as in Appendix G. The native and selected scores use the corresponding full joint matrices. On bank A, the direct and native rules each select the model with the larger aggregate score. Differences within 10^{-8} nats are ties, resolved to SD 2.1. The selected rule chooses only when its deterministic identification intervals separate by more than 10^{-8}, otherwise it abstains. Both banks' images are generated before selection on A. The A choices are saved before any B feature extraction or scoring, so the recorded separation concerns feature access and analysis. The sole primary outcome is

$$\widehat{H}_B(\widehat{m}_{\mathrm{direct,A}}) - \widehat{H}_B(\widehat{m}_{\mathrm{native,A}}).$$

Each rule makes one choice over the full panel, so the study supplies one policy comparison rather than 24 policy trials. The protocol retains zero or negative gains and assigns no numerical utility to abstention.

I.3 Complete bank outcomes and held-out gain

All 384 planned images are accounted for, with feature-array shape $24 \times 4 \times 512$ for each model and bank. Table 14 gives the three scores. Table 15 separates point orderings from finite-bank certification.

Table 14: Fresh-image scores in nats. Both banks use the same fixed 24 prompts, four separately seeded images per model and prompt, and input-selected power 10. Direct H is the exact grouped target for the realised bank.

Bank	Model	Direct H	Native G	Selected G_{10}
A	SD 2.1	.451850	1.153290	.455529
	IF-XL	.440489	1.218684	.444793
B	SD 2.1	.458617	1.152787	.462173
	IF-XL	.450834	1.218207	.454937

Table 15: SD 2.1 minus IF-XL contrasts. The final column bounds the direct difference for each realised bank and is not a population confidence interval. The selected rule certifies SD 2.1 on A and abstains on B.

Bank	Direct ΔH	Native ΔG	Selected ΔG_{10}	Identification interval
A	.011361	-.065394	.010736	$[.002502, .018971]$
B	.007782	-.065420	.007236	$[-.000999, .015471]$

On A, direct scoring selects SD 2.1 and the native score selects IF-XL. The selected intervals are $[.447295, .455529]$ for SD 2.1 and $[.436558, .444793]$ for IF-XL, certifying the direct direction. The frozen direct and certified choices both obtain $H_B = .4586165512$, while the frozen native choice obtains .4508344667. The primary held-out gain is $+.0077820845$ nats. Direct scoring and sharpening thus achieve the same observed gain. This single comparison supports neither a population-regret estimate nor a confidence interval for policy performance.

Bank B again reverses the native ordering and corrects it after sharpening. Its selected intervals, $[.453939, .462173]$ and $[.446703, .454937]$, nevertheless overlap. The difference interval

$[-.0009985931, .0154706539]$ requires abstention on B. The primary outcome still evaluates the frozen A choice, while the separate B certification remains unresolved.

The duplicated-image control repeats one observed image four times per prompt. Its direct entropy is zero by construction, up to 1.1×10^{-16} numerical error. Across the two models and banks, native scores range from .772172 to .861487, while selected scores range from .004316 to .005111. These constructed duplicates require no extra generator draws. Mean CLIP prompt-image cosines are .316052 and .319004 on A, and .316159 and .320935 on B, for SD 2.1 and IF-XL respectively. This descriptive alignment diagnostic supplies no human quality assessment or evidence of an image-quality gain.

I.4 Population uncertainty and computational scope

The same rank-bias and bounded-difference calculation as in Appendix H uses $m = 24$, $n = 4$, known feature rank at most 512 and $\alpha = .0125$ for each of the four model-bank entropy targets. A union bound gives simultaneous coverage at least .95 under ideal independent draws across prompt and replicate indices within each model-bank target. No empirical rank replaces the known dimension. The SD 2.1 and IF-XL intervals are $[.086146, 5.675427]$ and $[.074785, 5.664066]$ on A, and $[.092912, 5.682193]$ and $[.085130, 5.674411]$ on B. The resulting model-difference intervals are

$$A : [-5.577920, 5.600642], \qquad B : [-5.581499, 5.597063].$$

Both include zero. They concern average feature entropy over the fixed prompts, not a prompt population or confidence intervals for the selection-policy gain. Distinct recorded seeds support reproducibility but do not by themselves establish the ideal sampling assumption.

The supplied timing records use the same cached normalised features, warm up all methods and alternate their order over five repetitions. Generation and encoding are excluded equally. These small dense CPU timings describe the released implementations and make no universal speed claim. Direct grouped scoring already computes the finite-bank target exactly with the same information.

I.5 Computational cost and reproducibility

The amended study used 5.145556 allocated GPU-hours on two NVIDIA A40 GPUs. Including the original pilot gives 7.81 GPU-hours. Preparation used 4.626667 allocated CPU-hours. These totals include idle allocation and count each job once. Complete scheduler records, including the failed startup, are retained with the code.

The repository retains the original stopped plan, amendment, selection records, per-image metadata and feature arrays. A separate numerical implementation checks 306 reported values, with maximum absolute discrepancy below 5.8×10^{-15}. The exported archive omits image pixels, so this replay verifies feature-level calculations and metadata consistency. It relies on the recorded image verification and does not repeat generation or encoding. Local records document the analysis sequence, which was not externally preregistered.

The fresh banks show one model-selection consequence under the fixed CLIP representation. Population ordering, repeated-study policy reliability, human image quality, Bayesian posterior fidelity and broader generator behaviour remain open. The stopped original design, GenEval's primary 0-of-31 certification result and later post hoc refinement, and Qwen's absence of native temperature reversals remain part of the evidence.

J Replication with the published implementation

J.1 Source and evaluated interface

We replay the released implementation of Jalali et al. (2026) on the same cached features, using unchanged tracked source and float64 CPU arithmetic. Eight interface tests pass, and the replay takes 85 seconds. The repository supplies the exact source snapshot, environment and a logger-only import adapter that leaves the score calculation unchanged.

The high-level `ConditionalEvaluation.conditional_entropy` returns five log quantities, including $G = S((K \circ L)/n) - S(L/n)$. Exponentiation gives the score in Jalali et al. (2026,

Eq. 6) and preserves ordering. The exact path accepts different output and prompt feature dimensions. Its defaults are the Gaussian kernel and full Vendi, with `top_k=None`. Gaussian features are not automatically normalised. Cosine kernels normalise their input rows. A bandwidth or similarity function must be provided, with no automatic numerical bandwidth default.

For GenEval we call the unchanged high-level method directly, using cosine kernels on both modalities or Gaussian kernels on both. The latter uses our fixed pooled-bank medians, .9255238147 for output features and .9207977100 for prompt features, rather than claiming to reproduce the authors' experimental bandwidths. The high-level method shares one kernel-type argument across modalities. Qwen's locked linear-output and Gaussian-prompt specification therefore composes unchanged published `compute_entropy` calls on those Gram matrices.

J.2 Full-spectrum replay

All 24 audited full-score evaluations agree with the saved independent calculations to at most 9.86×10^{-14} nats. They comprise six high-level GenEval calls, 16 Qwen scores across the primary and three secondary banks, and two greedy-control entries. The six GenEval calls alone agree within 4.0×10^{-15} nats. Table 16 gives every GenEval pairwise comparison for these two specifications. This replay tests the implementation on the existing banks.

Table 16: Published high-level GenEval replay on cached CLIP features. Differences are the first system minus the second. G is the full published score in log units and H is direct grouped entropy.

Kernels	Systems	G difference	H difference	Reversed
Cosine	SD 2.1 / IF-XL	-.03907652	+.02268619	Yes
Cosine	SD 2.1 / retrieval	-.25069246	-.18459643	No
Cosine	IF-XL / retrieval	-.21161593	-.20728262	No
Gaussian	SD 2.1 / IF-XL	-.03107827	+.02227231	Yes
Gaussian	SD 2.1 / retrieval	-.21888691	-.16944333	No
Gaussian	IF-XL / retrieval	-.18780864	-.19171564	No

On Qwen's primary confirmation bank, the temperature-1.1 minus temperature-.7 difference is $+.08195270$ for the full published score and $+.10964934$ for grouped entropy. Thus the full score preserves the temperature ordering. Its gaps from grouped entropy are .76074177 and .73304513, reduced to .00256497 and .00249646 by the selected prompt kernel. The deterministic greedy control has native score .84956146 and grouped entropy zero. Agreement between implementations locates this dependence on prompt resolution in the statistic itself.

J.3 Optional truncation and implementation details

The published source implements additive tail redistribution as in equation 8. However, its joint-entropy call forwards `feature_dim` but omits `top_k`, while both marginal calls forward it. With $X = T = I_8$, cosine kernels, `computation_method=truncated-vendi`, `top_k=2` and the default `feature_dim=3000`, this argument mismatch gives $\log 4$ instead of the common-budget score zero. Setting both `feature_dim=t` and `top_k=t` avoids the mismatch. Across 20 cached evaluations, this aligned-budget route agrees with the paper's formula within 4.98×10^{-14} nats. The full-score reversals above do not use this optional path.

With Qwen's native, unsharpened prompt kernel and correctly aligned diagnostic budgets $t = 32$ and 100, the primary temperature difference is respectively $-.00467838$ and $+.02978453$, versus $+.08195270$ at full spectrum. The low-budget reversal persists after removing the argument mismatch. These post hoc diagnostic budgets differ from both the source's full-spectrum default and the paper's reported $t = 10000$ in Appendix E.1. The paper's fixed-t consistency statement concerns truncated population entropy, whose target may differ from full entropy.

The replay evaluates the released score computation on cached banks. The published bandwidth-selection and diffusion-guidance experiments additionally require a generation and feature-extraction pipeline absent from the source release. Interface checks, adapters and the source snapshot with its MIT licence are supplied with the code.