TIMESTEP WEIGHTING IN FROZEN DIFFUSION CLASSIFIERS: GLOBAL PARTITIONS AND POINTWISE REACHABILITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Diffusion classifiers aggregate class-conditional denoising losses across timesteps, often using weights selected for accuracy. Yet evaluating particular weighting rules leaves open which decisions any shared weighting can express. We characterise this decision family for frozen scorers with nonnegative weights. Linear and mixed-integer programs compare a label-aware pointwise ceiling with the best shared weighting under the same strict-margin criterion, separating individually feasible decisions from jointly feasible ones. A positive gap occurs in all six audited loss tables, spanning two model families and two datasets. On Stable Diffusion 1.5, the gaps are exactly 11 and 10 images on the original Pet and CIFAR-10 tables, respectively, and 2–3 images on higher-accuracy CIFAR-10 noise blocks with Pillow (PIL) preprocessing; each table contains 100 images. Rank-preserving controls implicate cross-timestep class-order agreement in restricting reachability. A controlled timestep-support intervention changes pointwise reachability on Stable Diffusion. In an idealised Gaussian model with correctly specified means and class-shared covariances, we characterise Bayes-preserving metrics, prove spectral and cone obstructions, and derive an exact distance to the preserving set. The finite-table audit separates fitting shortfalls from shared-family incompatibility; the Gaussian results separately characterise population partition preservation under explicit assumptions.

1 INTRODUCTION

Text-to-image diffusion models can classify images by selecting the label with the lowest class-conditional denoising loss (Li et al., 2023; Clark & Jaini, 2023). This provides zero-shot classification without fitting a discriminative head and has motivated comparisons with the out-of-distribution behaviour of discriminative models (Jaini et al., 2024). Because the losses are evaluated at several noise levels, the resulting classifier depends on how they are weighted.

Existing work primarily evaluates weighting through accuracy and task adaptation. Learned weights improve accuracy by about one point and transfer across the datasets tested by Clark & Jaini (2023); later work learns model- and task-specific weights or optimises matching noise (Jeong et al., 2025; Wang & Chen, 2025). These studies report the performance of particular rules but do not identify the best performance attainable under one shared weight vector.

On a fixed table, this distinction separates fitting shortfalls from limitations of the shared weighting family. A small gain from reweighting may reflect an inadequate fitting procedure, a mismatch between its objective and accuracy, or a limitation of the decision family itself. In particular, two inputs can each admit weights that recover the correct label while requiring incompatible weight vectors. Better optimisation cannot resolve such a conflict within the shared family. We therefore ask which decisions timestep weighting can express, both individually and jointly.

Our analysis compares a label-aware pointwise ceiling, obtained by choosing weights separately for each input, with the best correct count under one shared weighting (Figure 1). Both use the same strict-margin criterion. Their difference measures the cost of sharing. The ceiling uses the true labels to select weights, so it serves as a diagnostic bound rather than a prediction rule for unseen inputs.

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

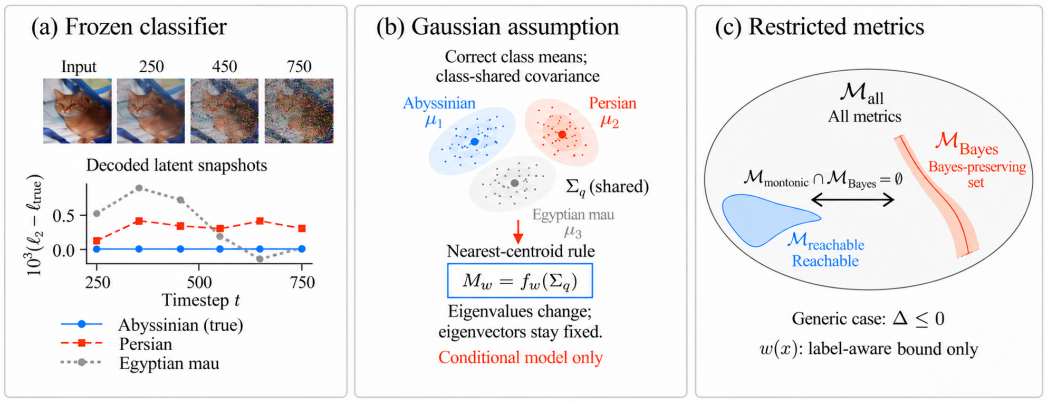

Figure 1: **Shared weighting and pointwise reachability.** **(a)** Oxford-IIIT Pet example (Parkhi et al., 2012): noise is added to the variational autoencoder (VAE) latent. The plotted contrasts $10^3[L(x, c, t) - L(x, y, t)]$, with true label y, change order across timesteps; one shared w aggregates losses across the dataset. **(b)** Separate Gaussian idealisation with correct means and class-shared covariance. **(c)** Schematic of the generic case in which the sets do not intersect (Section 5.2). Here $\mathcal{M}_{\text{monotonic}}$ denotes positive-definite metrics diagonal in the model covariance's fixed eigenbasis, with metric eigenvalues nonincreasing in the corresponding covariance eigenvalues; it contains $\mathcal{M}_{\text{reachable}}$, while $\mathcal{M}_{\text{Bayes}}$ preserves the Bayes regions. Allowing $w(x)$ gives a label-aware pointwise ceiling.

We find a positive gap on all six audited loss tables. On Stable Diffusion 1.5 (SD 1.5; Rombach et al., 2022), the gaps are exactly 11 and 10 images out of 100 on Pet and CIFAR-10, and 8 and 7 on fresh images and a denser timestep grid. PixArt-α (Chen et al., 2024b) also shows the gap, but its timestep-support effect misses the preset threshold on Pet and on both CIFAR-10 noise-draw halves. Controls that keep each timestep's class ranking leave the fraction of images with one reachable label unchanged and reachable-set sizes close to measured values, implicating cross-timestep class-order agreement as a proximate restriction.

We then study a complementary population question in an idealised Gaussian model: when can weighting preserve the Bayes decision regions? In this setting, weighting changes the induced scorer metric's eigenvalues but fixes its eigenvectors. Characterising the metrics that preserve the decision regions makes this explicit and yields conditions under which no weighting can recover them.

Contributions.

1. **Limits of shared weighting.** We combine standard linear and mixed-integer linear programs (LP and MILP; cf. Adem & Gochet, 2004) with rational verification to determine label-aware pointwise ceilings and shared-weight optima exactly on four audited SD 1.5 tables.
2. **Geometry of decision preservation.** Under the Gaussian model, Theorem 1 characterises Bayes-preserving metrics. We identify spectral and nonnegative-cone obstructions and quantify separation from the preserving set (exact distance and dual certificate in Appendix O).
3. **Empirical evidence and scope.** Rank and support interventions examine what restricts reachability; tests across panels, grids, preprocessing, and model families assess which findings persist (Section 4).

2 Related Work

Diffusion and generative classifiers. Conditional denoising losses enable zero-shot classification (Li et al., 2023; Clark & Jaini, 2023), with behaviour distinct from discriminative models (Ng & Jordan, 2001; Jaini et al., 2024). Weighting enters both diffusion training and classification: training

weights nonincreasing in log signal-to-noise ratio admit an evidence lower bound (ELBO) interpretation under Gaussian data augmentation (Kingma & Gao, 2023), while inference accuracy can be sensitive to a mismatch between training and test-time weights (Chen et al., 2024a). Adaptation methods learn weights (Jeong et al., 2025), optimise matching noise (Wang & Chen, 2025), or adapt prompts and select timesteps (Yoon et al., 2025). Reweighting gains are not universal: Jeong et al. (2025) find no significant SD 1.5/2.0 benefit in their low-shot compositional setting. We hold the scorer fixed and ask what weighting alone can express, irrespective of the adaptation procedure.

Our partition question also relates to decision preservation under misspecification (Domingos & Pazzani, 1997) and covariance-restricted discriminants (Flury, 1984; Flury & Schmid, 1992). We study the spectral metric family induced by one fixed, possibly misspecified covariance.

Reachability and metric learning. Metric learning optimises a Mahalanobis metric (Weinberger et al., 2005); our scorer restricts the available metric family. For one input, each class is a point $\ell_c = (L(x, c, t))_t$ in timestep-loss space, and a class is pointwise reachable at zero margin exactly when some nonnegative w satisfies $w^\top(\ell_j - \ell_c) > 0$ for every $j \neq c$, as in weighted-sum scalarisation (Bazgan et al., 2022). Our additional constraint is that the same w must serve every input.

Gaussian and linear score idealisations. Gaussian and linearised score approximations can describe aspects of learned diffusion models, especially away from the lowest noise levels (Wang & Vastola, 2024; Li et al., 2024); linear analyses also reproduce some behaviour of nonlinear classifier-free guidance (Li et al., 2025). Section 5 states our additional assumptions and examines their empirical implications. Gaussian spectral analysis for schedule design (Benita et al., 2025) concerns generated spectra, whereas we study classification regions under a fixed scorer.

Spectral cones and partition preservation. The conic hull of the generating curve is a generalised-Stieltjes moment cone (Widder, 1938; Karlin & Studden, 1966; Krein & Nudelman, 1977; Sokal, 2010), and spectral reweighting cannot rotate eigenvectors (Ledoit & Péché, 2011; Bun et al., 2017). Dimension arguments in convex and semidefinite geometry are classical (Barvinok, 2002; Pataki, 1998). Our analysis identifies the metrics that preserve a labelled nearest-centroid partition and intersects this set with the restricted spectral cone, with the required transversality proved in Theorem 14. Appendix A compares this intersection problem with invariant design (Pukelsheim, 2006), Delaunay-type cones (Schürmann, 2009; Erdahl & Rybnikov, 2002), power-diagram realisability (Aurenhammer, 1987; Brieden et al., 2017), and calibration dimension (Ramaswamy & Agarwal, 2016).

3 Problem setup: global and pointwise reachability

Model-free formulation. A frozen diffusion classifier scores label c on x by $\sum_t w_t\, L(x, c, t)$, the weighted class-conditional denoising loss, and predicts $\arg\min_c \sum_t w_t\, L(x, c, t)$ (Li et al., 2023). The weighting enters linearly, so the attainable decisions on an input are determined by the finite vectors $L(x, \cdot, t)$, without assumptions on the model that produced them. We call c *pointwise reachable* at x if some $w = w(x) \geq 0$ makes it the strict $\arg\min$; deciding this is a linear feasibility problem, not a search over candidate weights (Equation 1).

Which labels can weighting recover? For one image, let $L_{tc} = L(x, c, t)$ index T timesteps and C candidate classes. We maximise label c's smallest loss advantage over its competitors, with weights in $\Delta_T = \{w \geq 0 : \sum_t w_t = 1\}$:

$$\gamma_c(x) = \max_{w \in \Delta_T,\, \gamma} \gamma \quad \text{s.t.} \quad \sum_t w_t \frac{L_{tj} - L_{tc}}{s(x)} \geq \gamma \quad \text{for every } j \neq c. \tag{1}$$

The positive scale $s(x)$ is 1 for raw margins or the registered per-image reference for normalised margins. A label is reachable if its best margin exceeds τ; equality does not count. Thus $\mathcal{R}_\tau(x) := \{c : \gamma_c(x) > \tau\}$ is the reachable set, and an image is *invariant* at this threshold if the set has one label. Normalised shared-gap analyses use $\tau = 10^{-6}$ unless stated.

Can one weighting recover them jointly? The label-aware pointwise ceiling $O_\tau = \sum_i \mathbf{1}\{y_i \in \mathcal{R}_\tau(x_i)\}$ counts recoverable true labels when each image may choose its own weights. The shared

optimum $G^\star_\tau$ counts the most true labels that clear the same margin under one common w. Their difference is the sharing gap. We bound $G^\star_\tau$ by $[L, U]$: a verified weighting gives the lower bound, and a mixed-integer program gives the upper bound. The latter relaxes $> \tau$ to $\geq \tau$, so proving that K correct decisions are infeasible establishes $G^\star_\tau \leq K - 1$.

Verifying optimality. For exact results, we treat stored contrasts and τ as exact rational numbers and verify matching bounds with checkable certificates (Cheung et al., 2017) (Appendix G). This establishes exact pointwise ceilings and shared-weight optima on the four audited SD 1.5 tables; the PixArt-α rows remain solver-grade brackets. Certification concerns stored contrasts, not expected losses; the LP models, VIPR (verifying integer programming results) certificates, pointwise witnesses and a verifier are released (Appendix Q).

4 Empirical reachability in frozen diffusion classifiers

We test joint feasibility, intervene on timestep support, and assess robustness and transfer. Each table uses T timesteps and R noise draws per timestep: T determines the weighting coordinates, whereas R controls Monte Carlo precision. Table 1 summarises the six comparisons. Protocols, margins, registration status, and verification records are documented in Appendix Q (Table 29).

4.1 Does shared weighting attain the pointwise ceiling?

Testing joint feasibility. On the Pet table, 97 of 100 true labels can each be recovered by some weighting, but any one shared weighting recovers at most 86. Thus at least 11 individually recoverable decisions must be lost when weights are shared. CIFAR-10 gives 91 versus 81, a gap of 10. Both optima are exact (Table 1). Thus, the gaps reflect a limitation of shared nonnegative weighting on these fixed loss tables.

Which loss properties restrict reachability? The $T = 6$ SD 1.5 controls separate class order from loss spacing. Equally spacing losses while preserving each row's class order and range leaves reachable-set sizes close to their measured values and the invariant fractions unchanged. Breaking those orderings while preserving each row's loss multiset instead raises the mean reachable count from 2.55 to 15.76 on Pet and from 3.09 to 7.05 on CIFAR-10, with no invariant images. Class order thus largely accounts for the small reachable sets in this control; other spacing effects are not excluded (Appendix M).

A label-blind comparison at $T = 21$ selects the $|\mathcal{R}_\tau(x)|$ classes with smallest uniformly pooled losses. These lists contain the true label on $99/90$ Pet/CIFAR-10 images, versus $97/91$ for the reachable sets. Their sizes come from the LP, so this is a comparison at matched set size. Replacing losses with class midranks also leaves a positive solver lower bound on the shared gap in all six follow-up tables (Appendix N).

Table 1: **Shared-weight limits** ($n = 100$ images each; strict margin $\tau = 10^{-6}$, or 10^{-12} for the PixArt-α rows, on per-image max-normalised contrasts). A: uniform-weighting accuracy; O_τ: label-aware pointwise ceiling, an upper bound on the shared strict count; $G^\star_\tau$: best single shared weighting. Exact rows have matching bounds verified in rational arithmetic; solver rows are solver-grade brackets with floating-point upper bounds.

Model	Dataset	Panel	T	R	A	O_τ	$G^\star_\tau$	$O_\tau - G^\star_\tau$	Grade
SD 1.5	Pet	original	21	64	76	97	86	11	exact
SD 1.5	CIFAR-10	original	21	64	72	91	81	10	exact
SD 1.5	CIFAR-10	fresh, zero overlap (C)	21	64	78	97	89	8	exact
SD 1.5	CIFAR-10	C panel, dense grid (E)	101	32	80	99	92	7	exact
PixArt-α	Pet	original	6	32	44	78	$[48, 52]$	$[26, 30]$	solver
PixArt-α	CIFAR-10	original	6	32	46	78	$[55, 57]$	$[21, 23]$	solver

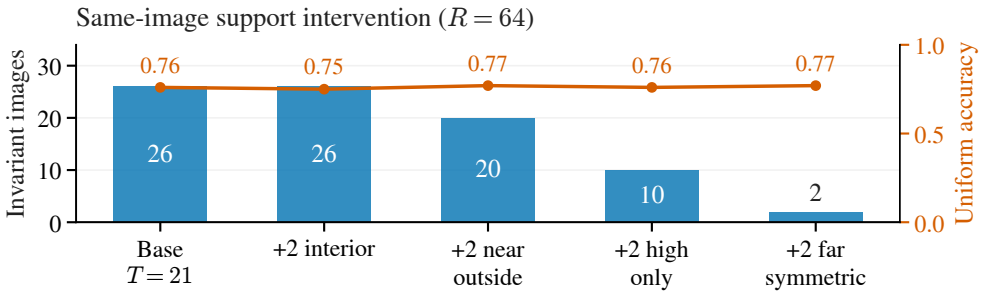

Figure 2: **Support dependence on fixed loss tables.** Fixed-design same-image Pet intervention at R=64: the T=21 base is compared with four count-matched arms adding interior, near-outside, high-only, and far-symmetric timestep pairs.

4.2 Does timestep support change reachability?

Separating support from the number of weights. We vary timestep location while fixing the number added, the 100 Pet images, and $R = 64$ (Figure 2). The analysis plan was fixed during measurement, before examining intervention results. Adding interior, near-outside, high-only, or far-symmetric pairs to the $T = 21$ base leaves $26, 20, 10, 2$ invariant images, respectively, while uniform accuracies stay at 0.75–0.77. The registered support-range contrast persists across four LP cutoffs, after excluding overlapping images, and in both $R = 32$ blocks (Appendix M). Thus where timesteps are added can change pointwise reachability even when accuracy barely changes.

Probe precision is distinct from support. $R^\star$ is the first tested per-block draw count whose 95% Wilson lower bound on decision agreement between trial blocks reaches 0.90; R_{sat} is the first after which accuracy stays within two points of its largest-R value. The agreement criterion is not met at the saturation budget: $R^\star = 32$ whereas $R_{\text{sat}} = 8$ on Pet; on CIFAR-10, $R^\star > 32$ while $R_{\text{sat}} = 16$. Thus accuracy saturation does not ensure stable per-image decisions on these panels. This ordering does not determine an optimal budget allocation: the registered allocation forecast fails on a larger nested sample (Figure 6; Appendices D and E).

4.3 How robust is the gap across models and protocols?

Cross-family replication. To test whether the findings extend beyond SD 1.5, we evaluate frozen PixArt-α tables (Chen et al., 2024b). Pet gives $O_\tau = 78$, $[L, U] = [48, 52]$ and CIFAR-10 gives 78, $[55, 57]$; both noise halves and four cutoffs retain the shared gap. The support-range effect is less robust: Pet's change is 5 against a pre-specified threshold of 10, while CIFAR-10 reaches its threshold only on the full table and fails on both noise halves. Thus the shared gap persists under the registered checks, but the support-range effect does not (Appendix K).

Panel, grid, and preprocessing sensitivity. To assess dependence on the sampled images and timestep discretisation, a zero-overlap CIFAR-10 panel retains a gap of exactly 8; the denser $T = 101$ grid on that same panel retains exactly 7. The gap is thus not confined to the original panel or coarse grid (Table 10, Appendix J). With the wider 21-point timestep grid over $[50, 950]$, ℓ_1 loss and the original tensor preprocessing, accuracy is 79% and the gap persists: $O_\tau = 98$ against $[87, 92]$ (row F, Table 29). The paired comparison with Pillow (PIL) preprocessing reaches 89% accuracy with an exact gap of 3. On the same images, two fresh $R = 96$ noise blocks A and B give accuracies 90% and 89% and exact gaps 2 and 3, respectively, after re-optimising both endpoints (Table 23, Appendix N.4). These PIL tables have higher accuracy and smaller certified sharing gaps than the tensor-preprocessed comparison. A three-image conflict makes the original PIL table's incompatible requirements concrete (Appendix N.5).

Transfer of a fitted weighting. A separate held-out evaluation tests whether one validation-selected vector improves accuracy on 4,500 unused CIFAR-10 indices (Appendix G): the gain is 1.09 points (one-sided 95% lower bound $+0.44$, $p = 0.0030$ under within-class exchangeability).

The gain is similar in magnitude to Clark & Jaini (2023), whereas Jeong et al. (2025) find no significant SD 1.5 benefit in a low-shot compositional setting; protocols differ, and the in-sample $G^\star_\tau - A$ of 9–12 points does not predict held-out improvement. In a separate same-image noise-transfer check, label-aware per-image weights selected on one half remain strictly correct on 88–99% of source-reachable SD images (69–81% on PixArt-α; Appendix N), so most source-selected witnesses survive the other noise half.

Can improved fitting or allocation recover the difference? Additional analyses (Appendix N) distinguish three limitations. Cross-entropy fits remain below the shared optimum, but their shortfall varies with logit scale (Table 26); it is not a measure of the representational gap. Giving four label-blind image clusters separate weightings beats one shared weighting in only one of 12 noise-half transfers, by one image, so the pointwise ceiling does not imply an effective conditional policy. Finally, a 500-timestep measurement gives 78% accuracy, and budget-matched class pruning gives 79%, equal to uniform allocation, so these allocation procedures do not improve accuracy on this panel (Appendix N.3).

5 Conditional Gaussian geometry

We now ask when nonnegative timestep weighting can preserve the Bayes decision regions in an idealised Gaussian population model. The argument compares two sets of metrics: those that weighting can produce, and those that preserve the regions. Their intersection determines whether preservation is possible; Section 5.3 then examines the model's empirical applicability.

Model and assumptions. Let $x \in \mathbb{R}^d$, $d \geq 2$, with $C \geq 2$ classes. The true law is $x \mid c \sim \mathcal{N}(\nu_c, \Sigma_p)$, with equal priors, distinct means and $\Sigma_p \succ 0$, as in linear discriminant analysis (LDA; Hastie et al., 2009, Chapter 4). The Gaussian score model uses the correct means but a possibly incorrect, class-shared covariance $\Sigma_q \succ 0$; its denoiser is the exact optimum for that Gaussian law. Equal priors remove prior offsets from the boundaries, while correct means isolate covariance misspecification. The optimal denoiser is affine, with the same linear part and loss trace term for every class. Comparing expected losses therefore reduces to comparing distances to the means under one metric (Appendix H). Gaussian score approximations motivate this idealisation (Wang & Vastola, 2024; Li et al., 2024).

Weighting acts on eigenvalues only. In the original input coordinates, write $\Sigma_q = Q\,\mathrm{diag}(\lambda_1, \ldots, \lambda_d)Q^\top$, where Q is orthogonal and $\lambda_1, \ldots, \lambda_d > 0$ are the model covariance's eigenvalues. In variance-exploding form, timestep t adds Gaussian noise of standard deviation $\sigma_t > 0$ to the clean input; σ_t is fixed by the schedule. For nonnegative $w \neq 0$, the noise-expected squared losses induce the metric (Appendix H)

$$M_w = f_w(\Sigma_q) = Q\,\mathrm{diag}(n)Q^\top, \qquad n := \sum_{t=1}^{T} w_t\, v(\sigma_t) \in \mathcal{C}_T, \qquad v(\sigma)_k := \frac{\sigma^2}{(\sigma^2 + \lambda_k)^2}, \tag{2}$$

$$\mathcal{C}_T := \mathrm{cone}\{v(\sigma_t) : 1 \leq t \leq T\} \subseteq \mathcal{C} := \mathrm{cone}\{v(\sigma) : \sigma > 0\}.$$

Here $f_w(\lambda) := \sum_t w_t \sigma_t^2/(\sigma_t^2+\lambda)^2$ is the scalar spectral response, n contains the metric eigenvalues, $v(\sigma)$ is one noise scale's contribution, and cone denotes all nonnegative combinations. Weighting changes the eigenvalues within $\mathcal{C}_T$, while fixing the eigenvectors Q. The larger cone $\mathcal{C}$ allows every positive noise scale; failure there rules out any finite schedule. Positive rescaling leaves decisions unchanged, so simplex-normalised weights suffice.

To compare with Bayes, whiten the input: $z := \Sigma_p^{-1/2}x$, $\mu_c := \Sigma_p^{-1/2}\nu_c$, and $B_w := \Sigma_p^{1/2} M_w \Sigma_p^{1/2}$. Then

$$(x - \nu_c)^\top M_w (x - \nu_c) = (z - \mu_c)^\top B_w (z - \mu_c).$$

Classification minimises this quadratic form over c; the true covariance is now I, so Bayes uses $B = I$. With means and covariances fixed, can varying w preserve the Bayes regions?

Which changes leave the regions intact? First separate directions that distinguish class means from directions they share. Write

$$S := \mathrm{span}\{\mu_i - \mu_j : 1 \leq i < j \leq C\}, \qquad k := \dim S, \qquad r_0 := d - k.$$

Thus S contains all differences between means; in its orthogonal complement $S^\perp$, of dimension r_0, every mean has the same component. A metric can change distances within $S^\perp$ without changing the winning class, provided it does not couple this space to S.

Within S, the boundaries impose stronger restrictions. Call (i, j) *active* when the Euclidean cells of μ_i and μ_j share a $(d-1)$-dimensional facet. Its normal is $\delta_{ij} := \mu_i - \mu_j$. Preserving that facet requires δ_{ij} to be a metric eigenvector; non-orthogonal eigenvectors of a symmetric metric must share an eigenvalue. Join active contrasts with nonzero inner product into a graph. Its m connected components span mutually orthogonal blocks $S_1, \ldots, S_m$, with $S = \bigoplus_a S_a$ (Lemma 13, Appendix R). A triangle has one block with a common in-plane scale; a square has two blocks, allowing independent horizontal and vertical scaling.

These observations define the candidate preserving metrics:

$$\mathcal{E} := \left\{ \bigoplus_{a=1}^{m} \alpha_a I_{S_a} \oplus B_0 : \ \alpha_a > 0, \ B_0 \succ 0 \text{ on } S^\perp \right\}.$$

Here I_{S_a} is the identity on block S_a, α_a its positive scale, and B_0 any positive-definite metric on the shared directions, omitted when $r_0 = 0$. The direct sum forbids coupling between blocks. Figure 3 compares this target with the reachable cone.

5.1 Bayes-preserving metrics

Theorem 1 (The Bayes-preserving set)**.** *For any whitened metric $B \succ 0$ (in particular B_w), the nearest-centroid rule under B agrees almost everywhere with the Bayes rule if and only if $B \in \mathcal{E}$; we call such metrics* Bayes-preserving.

Proof. For necessity, an active pair's common facet is $(d-1)$-dimensional and lies in both bisectors, forcing $B\delta_{ij}$ parallel to δ_{ij}; symmetry of B then equates the factor across non-orthogonal active contrasts. For sufficiency, positively scaling a half-space normal leaves the half-space unchanged, and the non-redundant constraints are exactly the active ones. See Appendix R. □

Preservation is thus weaker than recovering the Bayes metric I: a metric may rescale each contrast block and choose any metric on $S^\perp$, but may not couple them; inactive pairs add no condition.

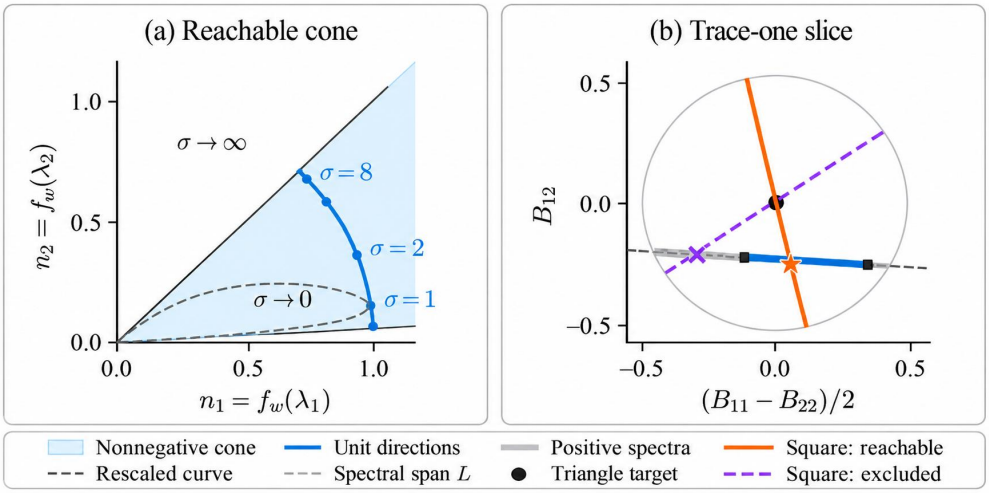

Figure 3: **A continuous spectral cone meets a thin partition-preserving target.** Here $d = 2$, Σ_q has eigenvalues $(1, 4)$, and Σ_p has eigenvalues $(1, 3)$ tilted $30°$ from the Σ_q eigenbasis. **(a)** A fixed schedule reaches only the cone over its finitely many sampled directions. **(b)** The triangle target misses the spectral family introduced in Section 5.2; each square target crosses that family, but only one crossing lies inside the nonnegative cone (Appendix B).

How restrictive is preservation? For dimension counting, take the linear span $\mathfrak{E} := \operatorname{span} \mathcal{E}$. It has m block scales and $\binom{r_0+1}{2}$ free entries in B_0. Within the space $\mathrm{Sym}(d)$ of real symmetric matrices, its codimension is therefore

$$\operatorname{codim} \mathfrak{E} = kd - \binom{k}{2} - m. \tag{3}$$

This counts the independent linear constraints on a preserving metric, in addition to positive definiteness. The next step asks whether the scorer's restricted family can satisfy them. Appendix C gives dimension fractions and examples at Contrastive Language–Image Pre-training (CLIP; Radford et al., 2021) and SD 1.5 latent scales (Table 3).

5.2 Spectral and conic obstructions

We first allow arbitrary eigenvalues while keeping the scorer's eigenvectors fixed. This relaxes both nonnegative weighting and the finite schedule: impossibility here also rules out the original family. Suppose Σ_q has distinct eigenvalues. Every reachable whitened metric then lies in the d-dimensional space $L = \{\Sigma_p^{1/2} Q \operatorname{diag}(n) Q^\top \Sigma_p^{1/2} : n \in \mathbb{R}^d\}$. Preservation requires a nonzero intersection with $\mathfrak{E}$, since zero is not positive definite. For subspaces in general position, $\dim(L \cap \mathfrak{E}) = \max(0, \Delta)$, where

$$\Delta \;=\; d + \dim \mathfrak{E} - \dim \mathrm{Sym}(d) \;=\; d(1-k) + \binom{k}{2} + m. \tag{4}$$

Thus $\Delta \le 0$ rules out preservation in general position; it does not measure how far the sets are apart. For this structured family, general position is proved outside the zero set of a nonzero polynomial $\mathcal{G}$ in $H = \Sigma_p^{1/2} Q$ at fixed whitened means. This is the meaning of "generic" below. The exceptional set has measure zero within each eigenvalue-multiplicity pattern and for almost every raw configuration (Theorem 14, Appendix R.3).

Corollary 2 (Generic impossibility)**.** *Over all admissible $1 \le m \le k \le d$, $\Delta > 0$ holds only when $k = 1$ (collinear class means) or for the single additional parameter triple $(d, k, m) = (2, 2, 2)$. Consequently, for $k \ge 2$, $d \ge 3$, and (Σ_p, Σ_q) outside $\{\mathcal{G} = 0\}$, no metric in L is Bayes-preserving.*

Proof. Appendix R gives the closed-form argument for arbitrary d; the verification script only cross-checks all 37,820 admissible triples with $1 \le m \le k \le d \le 60$. □

Outside $k = 1$ and $(d, k, m) = (2, 2, 2)$, $\Delta \le 0$, so generic covariance pairs fail even this relaxed test and no choice of timesteps or nonnegative weights preserves the Bayes regions. In this regime, $\Sigma_q = \Sigma_p$ lies in the exceptional set. Repeated eigenvalues or a finite schedule only shrink the family (Appendix R.3). For any scalar function g, the whitened matrix $\Sigma_p^{1/2} g(\Sigma_q) \Sigma_p^{1/2}$ also lies in L; nonnegativity enters below. Write $[A, B] = AB - BA$.

Corollary 3. *If $m = 1$, $r_0 = 0$ and $[\Sigma_p^{-1}, \Sigma_q] \ne 0$, then no M_w is Bayes-preserving.*

Proof. Every M_w commutes with Σ_q, whereas $\mathfrak{E} = \operatorname{span}\{I\}$ and Theorem 1 require $M_w \propto \Sigma_p^{-1}$, which the premise $[\Sigma_p^{-1}, \Sigma_q] \ne 0$ rules out. This is the same fixed-eigenbasis obstruction discussed for rotationally invariant estimators (Bun et al., 2017); the Bayes-preserving formulation is ours. □

Corollary 3 isolates eigenvector rigidity. Its conditions require $k = d$, hence $C \ge d + 1$ classes; preservation would then need $M_w \propto \Sigma_p^{-1}$, but weighting cannot rotate the eigenvectors of Σ_q onto those of Σ_p. Corollary 2 covers generic cases with $k \ge 2$, $d \ge 3$; neither result measures separation.

The cone obstruction. A positive-definite preserving metric in the linear relaxation need not be reachable by nonnegative weighting. This occurs for square rotations below and on an open planar binary set ($k = 1$, Appendix T).

Proposition 4 (Nonnegative-cone obstruction)**.** *There exist fixed $\Sigma_p, \Sigma_q \succ 0$ in $d = 2$, with distinct eigenvalues of Σ_q, and a nonempty open interval $I \subset (0, \pi/2)$ such that, for every $\psi \in I$, a unit square of whitened means rotated by ψ has $\Delta = 1$ and admits a positive-definite Bayes-preserving metric in L, but none reachable by nonnegative weighting over arbitrary positive noise scales.*

Proof. Appendix R fixes a covariance pair and certifies a ratio in $(0, 1)$; continuity gives the interval. □

For $d = 2$ and $\lambda_1 < \lambda_2$, the cone $\mathcal{C}$ admits exactly $n_1/n_2 \in (1, \mathrm{cond}(\Sigma_q)^2)$, where $\mathrm{cond}(\Sigma_q) = \lambda_2/\lambda_1$; a finite grid may narrow this interval. The larger-variance eigendirection receives less metric weight. Increasing anisotropy raises only the upper endpoint, leaving ratios below 1 excluded (Appendix B).

Appendix O gives the exact Frobenius distance $\delta_\star$ from the closed hull of trace-normalised reachable whitened metrics to the preserving set, with a dual certificate. Each metric's excess population 0–1 risk on that hull is bounded above and below by instance-dependent constants times its squared distance to the preserving set. The constants are unevaluated; the risk infimum over finite schedules is of order $\delta_\star^2$, and a prescribed grid can do worse (Appendix P).

5.3 Empirical assessment of the Gaussian model

We assess the idealisation at three levels: loss profiles, decisions, and denoiser structure.

Loss profiles. We compare spectral responses with other smooth functions. On $4{,}957$ CIFAR-10 images, a frozen spectral basis has smaller held-out residuals than monomial and power-law controls on 55.5% and 56.5% of images, respectively. Median relative residuals remain near 0.77 for all three. The fixed basis misses its registered transfer criterion at two further checkpoints (Appendices F, I).

Decisions. We test whether reconstructed losses preserve individual recoverability. Per-image spectral fits agree with measured true-class reachability on $88\%/94\%$ of Pet/CIFAR-10 images, below the $98\%/96\%$ agreement between measured noise halves. These fits do not impose a shared covariance, and other smooth families also yield positive gaps (Appendices F, U).

Denoiser structure. We check necessary relations imposed by the Gaussian model. Local Jacobian checks meet prespecified criteria, but an operator probe is numerically inconclusive. Class-output contrasts violate the exact relations at one 32-bit-float (fp32) operating point. A separate 32-dimensional surrogate fitted on source images predicts held-out contrasts poorly; its frozen witnesses certify no conflict on target images (Appendices L, X, Z). Together, these tests do not validate the Gaussian assumptions for the measured denoisers.

6 Conclusion and limitations

All six audited loss tables exhibit a positive gap between the pointwise ceiling and the best shared nonnegative weighting. This gap differs from a fitted rule's shortfall from the shared optimum. Separately, the Gaussian analysis characterises Bayes-preserving metrics and obstructions to reaching them. Partition mismatch and finite-table sharing cost are logically independent even for expected Gaussian losses (Appendix S); whether the Gaussian geometry explains the measured gaps remains open.

Empirical magnitudes depend on the protocol. Uniform accuracy is 72–80% on the four audited SD 1.5 tables, versus 87.5% (Pet) and 85.5% (CIFAR-10) in Li et al. (2023); the protocols differ. Reachability varies with noise draws, but fresh-noise gaps remain positive under the original and higher-accuracy PIL protocols (Appendices N.1, N.4, Y). On a fresh 500-image PIL panel, uniform accuracy is 87.8% and the gap bounds are $[35, 35]$, still solver-grade without an exact upper certificate; the frozen weighting gains 0.0 points (Appendix N.4), in a transfer test separate from Section 4. PixArt-α also retains floating-point bounds, and its support-range effect misses the registered threshold on Pet and both CIFAR-10 noise halves. Population and broader checkpoint generalisation remain open.

The audit applies to any classifier aggregating fixed class losses with shared nonnegative weights. It can assess whether alternative supports or input-dependent weighting families expand the attainable decisions. Whether a fitted rule realises these possibilities on new inputs remains a separate empirical question.

REPRODUCIBILITY STATEMENT

Deterministic scripts verify the algebra and small-scale enumerations. Analysis plans, file hashes, prespecified criteria, and all failed predictions are included in the supplement, and the preregistration and amendment texts accompany the code release. For every exact row the supplement also contains the exported LP model, the VIPR infeasibility certificate, the exact rational pointwise witnesses, a standard-library verifier that checks the LP/proof correspondence and runs the VIPR checker, and a second script that recomputes the pointwise and shared witnesses from the LP alone and verifies them in rational arithmetic. The bundle covers the four audited SD 1.5 tables and the four PIL ℓ_1 noise tables; raw loss caches are not distributed; Appendix Q distinguishes preregistered, amended, and post-hoc analyses. The Jacobian scale-up kept its estimator and threshold but pooled away two preregistered diagnostics, disclosed in Appendix L.

AI USE STATEMENT

The authors defined the research question, designed the study, approved every registered analysis before execution, made all decisions on scope, interpretation and release, and reviewed each result with its verification record. Generative AI tools assisted under this direction with writing, code implementation and review, mathematical checking, and the drafting of controls and proof steps that the authors then examined. All reported numbers come from executed code and were checked against recorded artifacts or deterministic verification scripts. No AI system is an author; the authors take full responsibility for the text, claims and artifacts.

ETHICS STATEMENT

This work analyses existing public generative models and standard benchmarks; it introduces no new model, data collection, or deployment and involves no human subjects. It inherits the biases, licensing constraints, and compute footprint of the underlying models and datasets. We present the label-aware ceiling as a diagnostic, not an attainable system, to avoid overstating downstream capability.

REFERENCES

Jan Adem and Willy Gochet. Aggregating classifiers with mathematical programming. *Computational statistics & data analysis*, 47(4):791–807, 2004.

Edoardo Amaldi and Viggo Kann. The complexity and approximability of finding maximum feasible subsystems of linear relations. *Theoretical computer science*, 147(1-2):181–210, 1995.

Franz Aurenhammer. Power diagrams: properties, algorithms and applications. *SIAM journal on computing*, 16(1):78–96, 1987.

Alexander Barvinok. *A course in convexity*, volume 54. American Mathematical Society, 2002.

Cristina Bazgan, Stefan Ruzika, Clemens Thielen, and Daniel Vanderpooten. The power of the weighted sum scalarization for approximating multiobjective optimization problems. *Theory of Computing Systems*, 66(1):395–415, 2022.

Roi Benita, Miki Elad, and Joseph Keshet. Spectral analysis of diffusion models with application to schedule design. *Advances in Neural Information Processing Systems*, 38:2073–2127, 2025.

Andreas Brieden, Peter Gritzmann, and Fabian Klemm. Constrained clustering via diagrams: A unified theory and its application to electoral district design. *European Journal of Operational Research*, 263(1):18–34, 2017.

Joël Bun, Jean-Philippe Bouchaud, and Marc Potters. Cleaning large correlation matrices: tools from random matrix theory. *Physics Reports*, 666:1–109, 2017.

Huanran Chen, Yinpeng Dong, Shitong Shao, Zhongkai Hao, Xiao Yang, Hang Su, and Jun Zhu. Diffusion models are certifiably robust classifiers. *Advances in Neural Information Processing Systems*, 37:50062–50097, 2024a.

Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. PixArt-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. In *International conference on learning representations*, volume 2024, pp. 57611–57640, 2024b.

Kevin KH Cheung, Ambros Gleixner, and Daniel E Steffy. Verifying integer programming results. In *International Conference on Integer Programming and Combinatorial Optimization*, pp. 148–160. Springer, 2017.

Kevin Clark and Priyank Jaini. Text-to-image diffusion models are zero shot classifiers. *Advances in Neural Information Processing Systems*, 36:58921–58937, 2023.

Pedro Domingos and Michael Pazzani. On the optimality of the simple Bayesian classifier under zero-one loss. *Machine learning*, 29(2):103–130, 1997.

Leon Eifler and Ambros Gleixner. Safe and verified Gomory mixed-integer cuts in a rational mixed-integer program framework. *SIAM Journal on Optimization*, 34(1):742–763, 2024.

Robert Erdahl and Konstantin Rybnikov. On Voronoi's two tilings of the cone of metrical forms. *Rend. Circ. Mat. Palermo (2) Suppl*, 70:279–296, 2002.

Bernhard N Flury. Common principal components in k groups. *Journal of the American Statistical Association*, 79(388):892–898, 1984.

Bernhard W Flury and Martin J Schmid. Quadratic discriminant functions with constraints on the covariance matrices: Some asymptotic results. *Journal of multivariate analysis*, 40(2):244–261, 1992.

John Gleeson and Jennifer Ryan. Identifying minimally infeasible subsystems of inequalities. *ORSA Journal on Computing*, 2(1):61–63, 1990.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer, 2 edition, 2009.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Priyank Jaini, Kevin Clark, and Robert Geirhos. Intriguing properties of generative classifiers. In *International Conference on Learning Representations*, volume 2024, pp. 13898–13923, 2024.

Yujin Jeong, Arnas Uselis, Seong Joon Oh, and Anna Rohrbach. Diffusion classifiers understand compositionality, but conditions apply. In *Advances in Neural Information Processing Systems*, volume 38, 2025. Datasets and Benchmarks Track.

Samuel Karlin and William J Studden. *Tchebycheff systems: With applications in analysis and statistics*. Interscience Publishers, 1966.

Diederik Kingma and Ruiqi Gao. Understanding diffusion objectives as the ELBO with simple data augmentation. *Advances in Neural Information Processing Systems*, 36:65484–65516, 2023.

M. G. Krein and A. A. Nudelman. *The Markov moment problem and extremal problems*, volume 50 of *Translations of Mathematical Monographs*. American Mathematical Society, 1977.

Olivier Ledoit and Sandrine Péché. Eigenvectors of some large sample covariance matrix ensembles. *Probability Theory and Related Fields*, 151(1):233–264, 2011.

Alexander C Li, Mihir Prabhudesai, Shivam Duggal, Ellis Brown, and Deepak Pathak. Your diffusion model is secretly a zero-shot classifier. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2206–2217. IEEE, 2023.

Xiang Li, Yixiang Dai, and Qing Qu. Understanding generalizability of diffusion models requires rethinking the hidden Gaussian structure. *Advances in neural information processing systems*, 37: 57499–57538, 2024.

Xiang Li, Rongrong Wang, and Qing Qu. Towards understanding the mechanisms of classifier-free guidance. *Advances in Neural Information Processing Systems*, 38:62835–62889, 2025.

Quinn McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, 1947.

Andrew Ng and Michael Jordan. On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes. *Advances in neural information processing systems*, 14, 2001.

Abhishek Panigrahi, Bingbin Liu, Sadhika Malladi, Andrej Risteski, and Surbhi Goel. Progressive distillation induces an implicit curriculum. In *International Conference on Learning Representations*, volume 2025, pp. 23694–23742, 2025.

Vardan Papyan, Xiao Y Han, and David L Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40): 24652–24663, 2020.

Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pp. 3498–3505. IEEE, 2012.

Gábor Pataki. On the rank of extreme matrices in semidefinite programs and the multiplicity of optimal eigenvalues. *Mathematics of operations research*, 23(2):339–358, 1998.

Friedrich Pukelsheim. *Optimal design of experiments*. SIAM, 2006.

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 *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Harish G Ramaswamy and Shivani Agarwal. Convex calibration dimension for multiclass loss matrices. *Journal of Machine Learning Research*, 17(14):1–45, 2016.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 10674–10685. ieee, 2022.

Achill Schürmann. *Computational geometry of positive definite quadratic forms: Polyhedral reduction theories, algorithms, and applications*, volume 48. American Mathematical Society, 2009.

Alan D Sokal. Real-variables characterization of generalized Stieltjes functions. *Expositiones Mathematicae*, 28(2):179–185, 2010.

Binxu Wang and John J Vastola. The unreasonable effectiveness of Gaussian score approximation for diffusion models and its applications. *arXiv preprint arXiv:2412.09726*, 2024.

Yanghao Wang and Long Chen. Noise matters: Optimizing matching noise for diffusion classifiers. *Advances in Neural Information Processing Systems*, 38:131926–131951, 2025.

Kilian Q Weinberger, John Blitzer, and Lawrence Saul. Distance metric learning for large margin nearest neighbor classification. *Advances in neural information processing systems*, 18, 2005.

David V Widder. The Stieltjes transform. *Transactions of the American Mathematical Society*, 43 (1):7–60, 1938.

Jee Seok Yoon, Junghyo Sohn, Wootaek Jeong, and Heung-Il Suk. ExpertDiff: Head-less model reprogramming with diffusion classifiers for out-of-distribution generalization. In *IJCAI*, pp. 6866–6874, 2025.

A BOUNDARY WITH ADJACENT THEORY

Table 2 is a boundary map over the audited works cited here, not evidence that no other related result exists. In particular, the rotationally-invariant-estimator row concedes the eigenbasis obstruction, and the calibration-dimension row concedes the closest dimension-counting precedent. The distinction we use is the object being constrained together with the equivalence relation imposed on it.

Table 2: **Adjacent results and the novelty boundary.** "No implication" means that the cited theorem does not yield our result; it is not a claim that the literature contains no closer theorem.

Family	Prior problem and result	Relation to this work
Convex calibration dimension (Ramaswamy & Agarwal, 2016)	*Object:* Surrogate prediction-space dimension. *Equivalence:* Calibration uniformly over $p \in \Delta_n$. *Assumptions:* Convex surrogate; loss matrix. *Conclusion:* Lower bounds exclude every lower-dimensional calibrated surrogate.	*Overlap:* Dimension count excludes a constrained family. *Distinction:* Different space, equivalence relation, and normal-cone mechanism.
Voronoi reduction; L-type domains (Schürmann, 2009; Erdahl & Rybnikov, 2002)	*Object:* Positive quadratic form. *Equivalence:* Same lattice Delaunay tiling. *Assumptions:* Lattice; infinitely many sites. *Conclusion:* Triangulations have full-dimensional domains; nonsimplicial tilings lie on lower-dimensional faces.	*Overlap:* Partitions metric space by induced diagrams. *Distinction:* We fix labelled cells a.e., rather than the lattice Delaunay type.
Rotationally invariant estimators (Ledoit & Péché, 2011) (Bun et al., 2017)	*Object:* Eigenvalues, with eigenvectors fixed. *Equivalence:* Restriction to estimator family. *Assumptions:* Spectral shrinkage. *Conclusion:* Eigenvalue transforms cannot repair eigenvector mismatch.	*Overlap:* Supplies the eigenvector rigidity behind Corollary 3. *Distinction:* Our general block target can commute with Σ_q; dimension can still obstruct it.
Kiefer invariance (Pukelsheim, 2006)	*Object:* Design measure. *Equivalence:* Invariance under group action. *Assumptions:* Compact group; feasible designs closed under group averaging. *Conclusion:* Averaging produces an invariant feasible design.	*Overlap:* Intersects a family with an invariant set. *Distinction:* Our spectral family need not be closed under that averaging.
Classifier aggregation by MILP (Adem & Gochet, 2004)	*Object:* Weights over fixed base classifiers. *Equivalence:* Training-sample classification. *Assumptions:* Fixed classifiers; design-set error objective. *Conclusion:* A MILP approximates the aggregation problem to select weights.	*Overlap:* Optimises shared weights against classification error. *Distinction:* We compare timestep pooling with a pointwise envelope and certify bounds on frozen loss tables.

Continued on next page

Table 2 (continued)

Family	Prior problem and result	Relation to this work
Metric learning (Weinberger et al., 2005)	*Object:* Freely chosen Mahalanobis metric. *Equivalence:* Optimisation objective, not partition equality. *Assumptions:* Learned metric. *Conclusion:* Selects a metric for large-margin neighbours.	*Overlap:* Treats the decision metric as the object. *Distinction:* We ask what is reachable when the metric family is fixed.
Convex/SDP dimension arguments (Barvinok, 2002; Pataki, 1998)	*Object:* Convex sets and semidefinite feasible sets. *Equivalence:* No partition equivalence imposed. *Assumptions:* Convexity; affine constraints in SDP. *Conclusion:* Face dimensions yield rank bounds for extreme feasible matrices.	*Overlap:* Uses dimensions to constrain matrix families. *Distinction:* We identify $\mathcal{E}$ and prove transversality for our structured family in Theorem 14.
Power diagrams and realisability (Aurenhammer, 1987; Brieden et al., 2017)	*Object:* Additive diagram weights. *Equivalence:* Diagram feasibility or support for a clustering. *Assumptions:* Fixed sites and costs; cluster-mass constraints; additive weights. *Conclusion:* LP duality characterises feasibility and support.	*Overlap:* Relates labelled partitions to diagram parameters. *Distinction:* It neither restricts metrics to our spectral family nor yields our negative result.

B THE TWO OBSTRUCTIONS, SEPARATED

The unreachable set is instance-dependent, so we name the instances swept below. For $\Sigma_q = \mathrm{diag}(0.7, 2.1)$ and, in the same coordinates, $\Sigma_p = \left(\begin{smallmatrix} 1.6 & 0.55 \\ 0.55 & 0.9 \end{smallmatrix}\right)$, sweeping the axis angle over $[0, \pi/2)$ gives a reachable measure of 17.33%; the instance of Figure 4 gives 42.16%. For both instances the linear count passes at *every* angle, so none of the shortfall comes from the linear count. Sweeping the anisotropy budget on the latter, $\mathrm{cond}(\Sigma_q)$ from 1.05 to 20 lifts the reachable fraction from 0.019 to 0.441, and an unbounded budget reaches only 0.442 — the residual is the Euclidean endpoint together with the angles admitting no positive-definite solution, neither of which a larger anisotropy budget removes. The discretised local search behind these sweeps selects and evaluates on one Monte Carlo draw, so its grid result does not support a population-cost conclusion.

C THE DEPTH OF THE IMPOSSIBILITY

For fixed (k, m) and large d, the preserving span has dimension fraction $\dim \mathfrak{E} / \dim \mathrm{Sym}(d) = 1 - \frac{2k}{d+1} + O(d^{-2})$: most metric degrees of freedom lie in the unconstrained $S^{\perp}$ block. Figure 5 plots Δ against d at fixed k with the most favourable $m = k$. The irrelevant fraction and the depth $-\Delta$ move in opposite directions as k changes. When $C \leq d + 1$, the choice $k = C - 1$ is the largest value the class contrasts admit: along these $m = k$ envelopes any smaller k weakly raises the irrelevant fraction (strictly unless $k = d$), so this choice is conservative for the harmless-directions reading. In contrast, any smaller k weakly raises Δ (by $d - k$ per unit step along the envelope) and therefore reduces the depth $-\Delta$, so the reported depths are the largest, most favourable end for the impossibility claim rather than conservative ones. Neural collapse (Papyan et al., 2020) reduces within-class variability and arranges centred class means as a simplex equiangular tight frame, whose contrast span has dimension $C - 1$. It therefore does not justify a smaller k. Any further rank reduction would require separate evidence; the dependence on k above follows from our dimension formula.

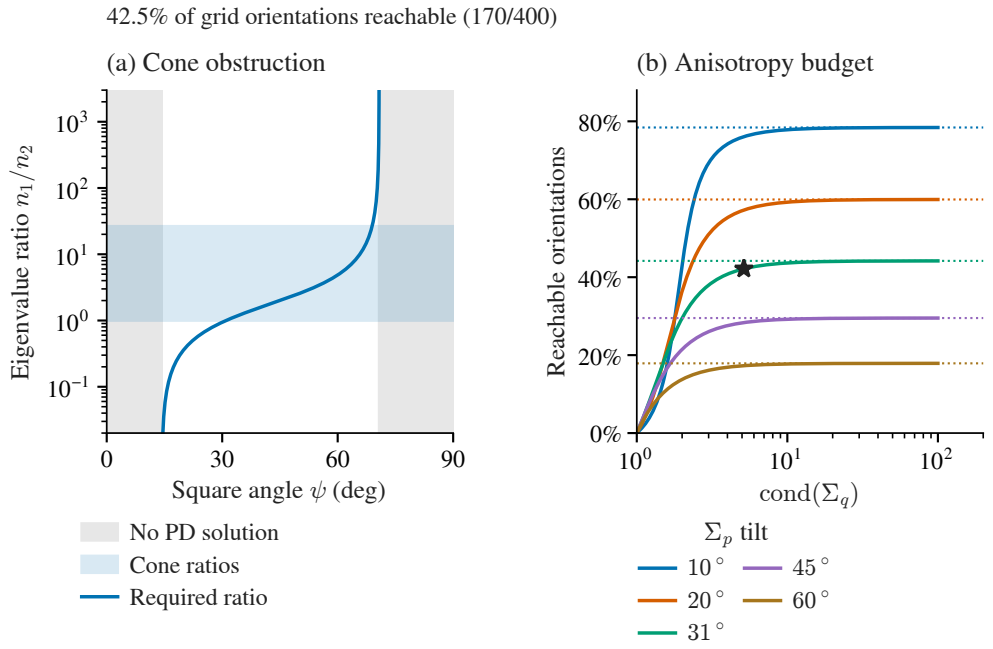

Figure 4: **The linear and the spectral obstruction separate inside one configuration family.** The plotted instance uses Σ_q eigenvalues $(0.35, 1.8)$ and $\Sigma_p = R(31°)\,\mathrm{diag}(0.6, 2.4)R(31°)^\top$ with $R(\theta)$ the rotation by θ; it is distinct from the 17.33% control reported immediately above. **(a)** For 2-D square configurations whose axes are rotated by ψ, the linear obstruction is passed for *every* ψ ($\Delta = 1$, so a solution ray always exists), but a Bayes-preserving metric additionally requires the eigenvalue ratio n_1/n_2 drawn as the curve, and only the shaded band $(1, \mathrm{cond}(\Sigma_q)^2) = (1, 26.449)$ is reachable by a nonnegative weighting (endpoints only in the closure). Grey columns mark angles for which the required n is not even positive definite. On the recorded grid $\psi = \pi i/400$, $i = 0, \ldots, 399$, the reachable fraction is $170/400 = 42.5\%$; the numerically computed Lebesgue fraction over $\psi \in [0, \pi/2)$ is 42.156%. The remainder splits into 37.24% with no positive-definite solution, 18.58% positive definite but below the cone, and 2.03% above it. **(b)** The reachable fraction as a function of the anisotropy budget $\mathrm{cond}(\Sigma_q)$, for several tilts of Σ_p against the Σ_q eigenbasis. It saturates: for the instance of panel (a) an *infinite* budget would still only reach 44.18%. Unbounded anisotropy therefore removes only 2.03 percentage points of failure. The remaining angles either admit no positive-definite solution or require $n_1/n_2 < 1$, outside the cone at every anisotropy budget.

At the SD 1.5 latent dimensionality ($d = 16{,}384$, $C = 10$, $k = 9$, $m = 1$) the decision-relevant codimension is $147{,}419$ out of $134{,}225{,}920$ degrees of freedom, so $\dim \mathfrak{E}/\dim \mathrm{Sym}(d) = 99.890\%$. The actual harmless set is the intersection of $\mathfrak{E}$ with the symmetric positive-definite (SPD) cone, $\mathcal{E} = \mathfrak{E} \cap \mathrm{SPD}$, not every additive perturbation in $\mathfrak{E}$; the ratio is an ambient dimension count, neither the probability that a random perturbation is harmless nor invariance around an arbitrary baseline metric. It says which degrees of freedom an exactly preserving metric may use once positivity and the complementary 0.110% constraints are satisfied.

D Predicting timestep allocation at fixed compute

We test whether noise-stability diagnostics can predict how to allocate a fixed compute budget between timesteps and noise draws. Here $R^\star$ is the first tested noise-trial budget whose 95% Wilson lower bound on independent-block decision agreement reaches 0.90, and R_{sat} is the first after which

Table 3: **Dimensions of the Bayes-preserving set and the spectral family.** $\dim \mathrm{Sym}(d) = d(d+1)/2$, $\dim \mathfrak{E} = m + \binom{r_0+1}{2}$, $\operatorname{codim} \mathfrak{E} = kd - \binom{k}{2} - m$, and $\Delta = d(1-k) + \binom{k}{2} + m$. Here $k = \dim S$ is the dimension of the contrast span, m the number of its non-orthogonal blocks, $r_0 = d - k$, $\mathfrak{E}$ the linear span of the Bayes-preserving set (Section 5), and L the d-dimensional spectral family (Section 5.2). The last column reports only the *linear* test. In general position, $\dim(L \cap \mathfrak{E}) = \max(0, \Delta)$; $\Delta > 0$ is necessary, not sufficient, for reachability. Genericity concerns covariance pairs at fixed whitened means (Theorem 14). Proposition 4 gives failure on a positive-measure set of angles within the planar square family; this is not a full-dimensional measure claim for arbitrary four-centre configurations. Appendix T gives an open set of planar two-class ($k = 1$) configurations with a cone-only obstruction. Figure 8 supplies a separately certified cone-only control. The displayed SD 1.5 row uses m=1 ($\Delta = -131{,}035$); the most favourable admissible split, m=9, still gives $\Delta = -131{,}027$. Rows are generated by the verification script.

Configuration	d	k	m	$\operatorname{codim} \mathfrak{E}$	irrelevant d.o.f.	Δ	$\Delta > 0$
Collinear means (any C)	2	1	1	1	66.667%	1	yes
Square, two orthogonal blocks	2	2	2	1	66.667%	1	yes
Triangle, connected full span	2	2	1	2	33.333%	0	no
Four points in general position	3	3	1	5	16.667%	−2	no
CLIP-scale embedding, $C = 10$	512	9	1	4,571	96.519%	−4,059	no
SD 1.5 latent, $C = 10$, m=1	16,384	9	1	147,419	99.890%	−131,035	no

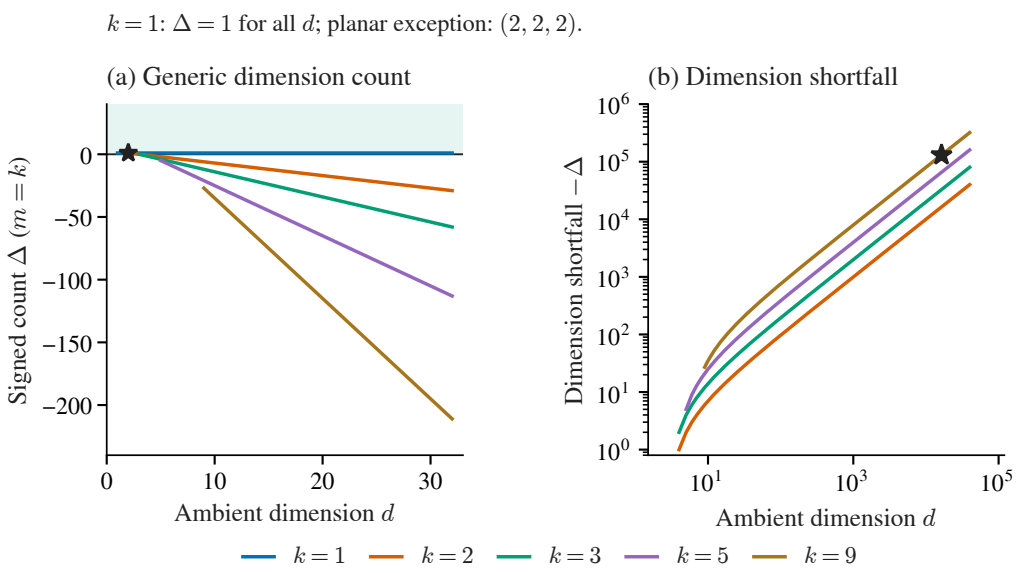

Figure 5: **The signed generic dimension count Δ becomes increasingly negative as the contrast span grows.** Each curve fixes k and takes the most favourable $m = k$, so these are upper envelopes; an exhaustive enumeration of all 37,820 admissible triples with $1 \le m \le k \le d \le 60$ confirms that $k = 1$ and $(d, k, m) = (2, 2, 2)$ are the only configurations with $\Delta > 0$. At the SD 1.5 latent scale ($d = 4 \cdot 64 \cdot 64 = 16{,}384$, $C = 10$, $k = 9$) the best case over admissible m is $\Delta = -131{,}027$. Δ is a signed dimension deficit, not a distance; the metric question is $\delta_\star$ (Theorem 5). When $\Delta \le 0$, the well-specified locus $\Sigma_q = \Sigma_p$ lies inside the exceptional set.

accuracy remains within two points of its maximum-budget value. Figure 6 records both rather than treating accuracy saturation as per-image stability.

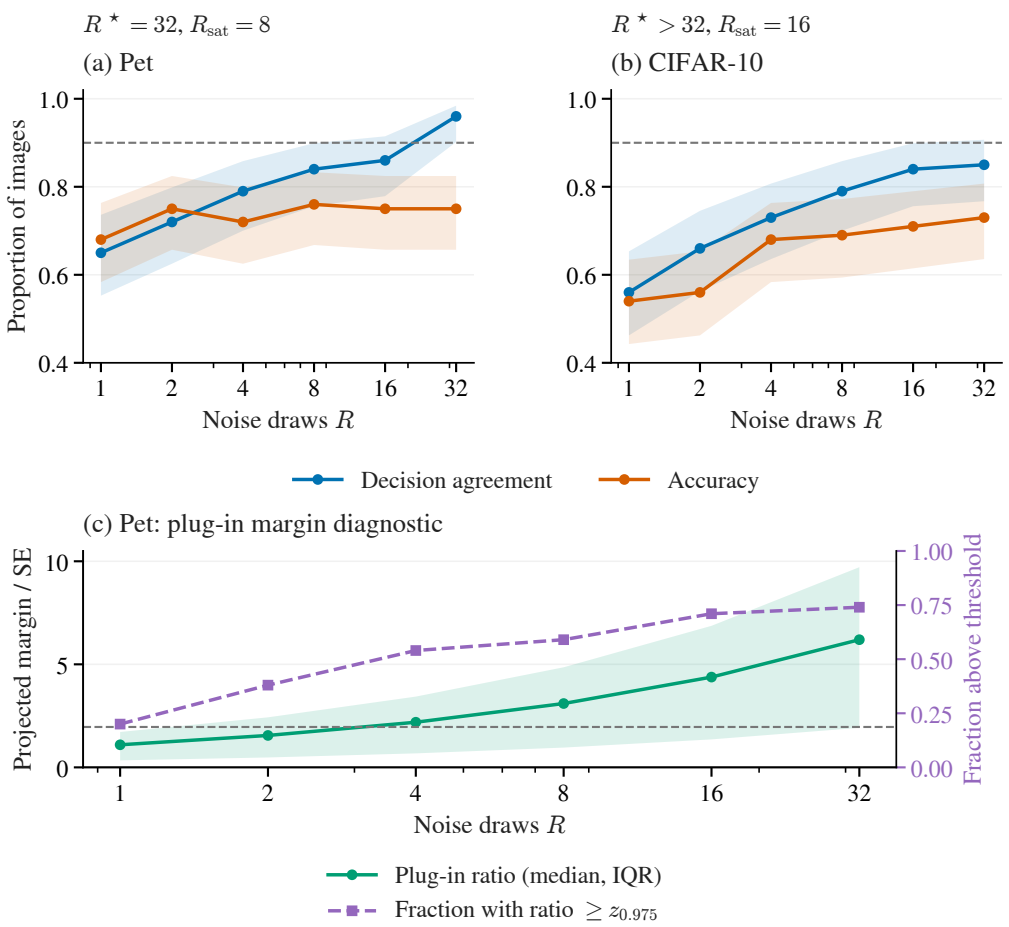

Figure 6: **Accuracy saturation does not ensure stable per-image decisions. (a)** Pet: the 95% Wilson lower bound on decision agreement between independent noise blocks first reaches the preregistered 0.90 at $R^\star = 32$, whereas aggregate accuracy saturates at $R = 8$. **(b)** CIFAR-10 replication: this criterion is not met through $R = 32$ ($R^\star > 32$), while accuracy saturates at $R = 16$; only the within-dataset ordering $R^\star > R_{\text{sat}}$ is compared. Agreement uses two R-draw blocks; accuracy uses one R-draw prefix, with saturation assessed through $R = 64$ (not shown). **(c)** Pet: a plug-in margin diagnostic. The winning class, runner-up, mean loss contrast $\widehat{m}$, and contrast standard deviation $\widehat{s}$ are estimated once using all 64 draws. We plot $\widehat{m}\sqrt{R}/\widehat{s}$ and the fraction of images with this ratio at least $z_{0.975} \simeq 1.96$, the standard-normal quantile. Its increase with R follows from the imposed $R^{-1/2}$ standard-error scaling; it is not an independent measurement of decision stability or a coverage guarantee. Bands in (a) and (b) are Wilson intervals over 100 images; the band in (c) is the interquartile range of the plug-in ratios.

We specified a predicted allocation $\hat{T}^\star(N)$ for $N \in \{4, 12, 24\}$ before running the grid and evaluated it on a disjoint split of 185 images (Table 4, upper block). The prediction matched the observed best allocation in 2 of 3 budgets against a chance level of 0.1875 (the budgets offer 3, 4, and 4 candidate allocations, so independent uniformly random forecasts match with probability 1/3, 1/4, 1/4), with no budget failing a 5-point non-inferiority test; one match, at $N = 12$, was decided by the preregistered tie-break from an exact tie at 145/185, and the miss at $N = 4$ was inconclusive at that budget's 11.7-point minimum detectable difference.

A power extension specified before its execution evaluates the same frozen forecast on 740 images, a nested superset of the 185 evaluation images disjoint from the forecast-fitting set. It matches 0 of 3 allocations (Table 4, lower block): at $N = 12$, $T = 3$ scores 0.754 versus 0.739 for $T = 2$; at

Table 4: **The preregistered allocation test and its power extension.** $\hat{T}^\star$ was fixed before evaluation on images disjoint from the forecast-fitting set and retained unchanged for the extension. "Match" compares $\hat{T}^\star$ with the observed best allocation, tie-broken towards smaller T by the registered rule. Chance level is 0.1875 for at least 2 of 3 under independent uniformly random forecasts over 3, 4, and 4 candidates. The 185-image block is the registered result; the 740-image block is a nested power extension declared in advance, on which the forecast matches 0 of 3 — the $N = 12$ tie breaks the other way and $N = 24$ moves one allocation. Non-inferiority is unchanged at both sizes. The two blocks are not independent: the smaller image set is contained in the larger. CI denotes the 95% Clopper–Pearson interval for the paired difference conditional on the observed discordant pairs; it is not an unconditional interval for a population accuracy difference.

N	$\hat{T}^\star$	observed best	accuracy	match	non-inferiority (5 pt margin)
registered, $n = 185$					
4	2	T=4, R=1	0.697	no	inconclusive; gap +5.95 pt, CI $[-2.14, +13.32]$
12	2	T=2, R=6	0.784	tie	pass
24	24	T=24, R=1	0.811	yes	pass
power extension, $n = 740$ *(nested superset)*					
4	2	T=4, R=1	0.670	no	inconclusive; gap +3.51 pt, CI $[-0.38, +7.31]$
12	2	T=3, R=4	0.754	no	pass
24	24	T=6, R=4	0.788	no	pass

$N = 24$, $T = 6$ scores 0.788 versus 0.778 for $T = 24$. Non-inferiority is unchanged (2 pass, 1 inconclusive, 0 fail). The 185-image analysis remains the original registered test.

The forecast therefore fails to identify the best observed allocations on the larger sample. At $n = 185$, no pairwise two-sided exact McNemar test (McNemar, 1947) reaches $p < 0.05$. At $n = 740$, the within-budget accuracy ranges are 1.6–5.8 points, and four of the 15 pairwise comparisons have unadjusted $p < 0.05$: $T = 1$ versus $T = 4$ at $N = 4$ ($p = 0.0113$), and $T = 1$ versus $T = 2, 3, 6$ at $N = 12$ ($p = 0.0086, 0.00027, 0.0028$, respectively). None of the frozen forecast's comparisons with the observed best allocation reaches $p < 0.05$. These comparisons do not establish a population-optimal allocation or equivalence between policies. The two results are not independent evidence — the smaller set is contained in the larger — so we report them as one nested sequence, not as a replication.

E Compute-matched policy comparison

We test whether distributing a fixed budget across more timesteps improves accuracy. Holding compute at 12 forward passes per class on 4,000 held-out CIFAR-10 images, the six-timestep grid and a two-timestep pair $\{550, 750\}$ *selected on a disjoint validation split* have a paired difference of +0.30 points (exact McNemar $p = 0.70$; interval $[-1.12, +1.72]$, conditional on the observed discordant images). This does not establish equivalence or adequate power for smaller effects. The same design separates the grid from every *fixed* two-timestep policy registered, by 3.40 ($p = 7 \times 10^{-7}$), 3.45 and 21.20 points. But both live inside a pre-narrowed interior $[250, 750]$. On a grid spanning $[50, 950]$ at the same compute and on the same 1,600 images the comparison reverses: the six-timestep spread scores 0.401 against 0.714 for the same spread confined to the interior (paired gap 31.25 points, $p = 2 \times 10^{-96}$), and a two-timestep interior pair beats the six-timestep full-range grid by 30.75; the pair $\{50, 230\}$ scores 0.164 against chance 0.100. Within the interior, the grid outperforms the tested fixed pairs but has no resolved advantage over the validation-selected pair at this budget. Extending the grid to the extremes substantially reduces accuracy on these 32×32 inputs. Thus timestep placement matters even when both the number of timesteps and the evaluation budget are held fixed.

Resolution dependence of timestep allocation. We preregistered a resolution ladder on Oxford-IIIT Pet and bicubically downsampled each image to a square before the same fixed preprocessing returned it to 512^2. The grid $[50, 230, 410, 590, 770, 950]$, budget $N = 12$, and policy compar-

ison are fixed at every rung. Five policies share the budget: the broad six-point grid, the low pair $\{50, 230\}$, the high pair $\{770, 950\}$, the fixed mid pair $\{410, 590\}$, and the "chamber", the two-timestep pair selected on the validation split at each rung (for example $\{410, 770\}$ at 64^2 and $\{410, 590\}$ at 128^2 and native). Table 5 reports the test-set results.

Table 5: Resolution dependence of fixed-budget Pet policies. Accuracies are fractions; paired differences are percentage points (pp). The chamber pair is validation-selected at each resolution.

resolution	test n	low $[50, 230]$	high $[770, 950]$	broad − chamber (pp)
native	259	0.575	0.479	$+0.39$ ($p = 1.00$)
128^2	148	0.486	0.277	$+1.35$ ($p = 0.87$)
64^2	148	0.122	0.304	-27.03 ($p = 7.2 \times 10^{-7}$)
32^2	259	0.042	0.266	-16.22 ($p = 3.7 \times 10^{-6}$)

Fixed-mid sensitivity. A post-hoc reanalysis removes the rung-specific chamber selection and compares the same two policies at both middle resolutions. Both use 12 forwards per class and the same 148 held-out images (Table 6).

Table 6: Fixed-mid sensitivity on matched Pet images. Both policies use 12 forward passes per class.

pixelation	broad full	fixed mid $[410, 590]$	broad − mid (pp)
64^2	$38/148$ (25.68%)	$84/148$ (56.76%)	-31.08
128^2	$97/148$ (65.54%)	$95/148$ (64.19%)	$+1.35$

The resolution interaction is $+32.43$ points. A 2,000-replicate class-stratified image bootstrap gives the descriptive percentile interval $[23.65, 41.89]$ points; this is a fixed-cache post-hoc sensitivity analysis, not a confirmatory population interval.

The low-noise accuracy collapses monotonically towards chance $1/37 = 0.027$ as resolution falls; the high-noise band does not show the same monotone collapse towards chance. The broad-versus-chamber contrast reverses sign between 128^2 and 64^2. At both 128^2 and native resolution, the broad grid ranks first, followed by the validation-selected and fixed-mid policies (which both use $[410, 590]$ and are tied), then the low and high pairs. The $64^2 \to 128^2$ comparison uses the same 185 images, labels, seed, and split, changing only pixelation side. It therefore identifies the transition without relying on a correlation across separately sampled runs. Cross-rung comparisons are not all equally controlled: 64^2 and 128^2 use five images per class ($n = 185$, test 148), whereas 32^2 and native use ten ($n = 370$, test 259). This changes precision and the sampled images, but not the within-rung policy contrasts or the controlled $64^2 \to 128^2$ result.

A simple input-level explanation is that at $t = 50$ the image is nearly clean, so class evidence can live in fine detail that downsampling destroys. By $t = 950$ the forward noise has already destroyed that detail, so reducing input resolution has much less left to remove. This is consistent with the observed shift away from broad coverage and towards policies concentrated in the middle of the schedule. The Pet intervention shows that resolution alone is sufficient to reproduce the sign reversal; it does not establish that resolution is the only difference relevant to CIFAR-10. The 32^2 CIFAR-10 result is therefore resolution-confounded and cannot be extrapolated to native resolution. The experiment does not isolate how the UNet denoiser internally implements this dependence.

Floating-point precision versus noise variability. At fixed draws, 16-bit (fp16) versus 32-bit (fp32) floating-point predictions disagree on 0.75% of per-cell $\arg\min$s, versus 73.3% across draws: a $98\times$ reduction ($125\times$ on CIFAR-10); pooled decisions agree $100/100$. Under these settings, changing numerical precision affects decisions much less than changing noise draws; this comparison does not establish precision invariance for other protocols.

Budget dependence. The policy comparison uses a single seed; its ranking is stable across three compute levels ($N = 6, 12, 24$ forward passes per class) but the interior/two-point tie is not: at $N = 24$ the six-timestep grid leads by 1.55 points ($p = 0.026$, which does not survive correction for the three comparisons), so we report an ordering at these budgets rather than a placement rule that transfers.

F SPECTRAL RECONSTRUCTION OF TIMESTEP-LOSS PROFILES

This appendix tests a spectral implication of the Gaussian model against alternative smooth bases (Section 5.3). In that model every class contrast $L_{tc} - L_{tc'}$ has timestep dependence confined to the family $a_\lambda(t) = \sigma_t^2/(\sigma_t^2 + \lambda)^2$ over the eigenvalues λ of Σ_q, with σ_t fixed by the model noise schedule, and with the *same* λ's for every image.

Basis selection and evaluation. The amended, preregistered strict test selects $K = 2$ values of λ from a fixed 25-point log-grid by least squares on half the images (even noise trials), and evaluates the out-of-image, out-of-noise relative residual $\|D_{\text{odd}} - B\widehat{C}_{\text{even}}\|_F / \|D_{\text{odd}}\|_F$ on the other half (odd trials). Here B is the basis selected on the training images; for each evaluation image, $\widehat{C}_{\text{even}}$ is fitted to that image's even-trial contrasts and evaluated on its odd-trial contrasts. Thus image holdout applies to basis selection, while noise holdout applies to reconstruction. We give two competing families — K real-exponent monomials in a normalised $\log \sigma_t$, and K power laws $(\sigma_t^2)^\alpha$, with 25 equally spaced exponents in $[-2, 2]$ — the same rule: choose K of 25 distinct functions on the training images, so all three families select among $\binom{25}{2} = 300$ distinct two-dimensional spans. Over 20 random image splits the theory family's median residual is 0.322 on Pet and 0.436 on CIFAR-10 against 0.379/0.483 (monomials) and 0.370/0.480 (power law), beating them on $20/20$, $20/20$, $20/20$ and $19/20$ resplits — a stability count, not a test, since resplits share images and noise. A *frozen* theory basis with $\lambda \in \{0.5, 2\}$ — straddling unit latent variance, fixed before any fit and given no selection at all — beats the same 300-way-selected comparators $20/20$ on both datasets (0.330 and 0.431), while the single frozen $\lambda = 1$ wins $0/20$ — consistent with, though not uniquely diagnostic of, a multi-component spectral model, since a one-function basis also has less to fit with.

Independent-panel replication. Because those resplits reuse 100 images, we registered a replication on $4{,}957$ CIFAR-10 images from a separate run never analysed before, with all three bases frozen (the comparators keeping the bases they had selected on the first 100) and one independent sign per image: the theory basis has the smaller held-out residual on 55.5% and 56.5% of images against the two comparators (one-sided binomial $p = 8 \times 10^{-15}$ and 3×10^{-20}), and $\lambda = 1$ loses on 72% and 70%. The image-level test detects a small directional advantage, but median residuals remain near 0.77 for all three methods on this panel with three noise draws per half. The per-image, best-λ version of the test, registered first, gives the same ordering with paired Wilcoxon $p \leq 1.2 \times 10^{-15}$. The initial comparison favours the spectral family under matched basis selection; on the independent panel, the fixed spectral basis has smaller errors on a majority of images. This does not by itself establish lower mean error. The large residuals limit the absolute predictive fit, and six timesteps cannot distinguish the spectral model from every smooth family.

Decision reconstruction. Reconstructing losses from the per-image fit and recomputing the reachability LP reproduces the measured true-class reachability on 88% of Pet images against a 98% noise baseline (partial transfer under the prespecified rule) and 94% of CIFAR-10 against 96% (meeting the prespecified transfer criterion).

A post-hoc shared-gap check of the per-image reconstruction. We further analyse the original per-image $K = 2$ reconstruction above, after exposure to its alignment results. Each image selects its own two kernel parameters from the fixed 25-point grid and fits centred class losses on even noise trials; this is not a fit of one covariance shared across images. The comparison uses the same 100 images and $T = 6$ timesteps, with measured even and odd $R = 32$ halves as matched references, rather than the $T = 21$, $R = 64$ headline tables. Each image's scale is its maximum absolute even-half true-class contrast; the same scale and $\tau = 10^{-6}$ apply to all three tables (Table 7).

The reconstructed tables express positive shared gaps, but reduce O from the odd-half values $97/90$ to $87/84$ (Pet/CIFAR-10); their certified true-reachability masks agree with the odd halves on

Table 7: **Shared gaps on matched six-timestep measured and reconstructed tables.** All counts are out of 100. A counts strict uniform-margin hits (equal to uniform argmin accuracy in these six cells). O and L have rational witnesses on the materialised binary64 contrasts; U is a floating MILP bound. Thus all gap brackets remain *solver-grade*, despite coincident endpoints. Each cell has one 600-second MILP cap and at most one fixed-subset margin-LP repair, recounted at the original strict threshold. The analysis is descriptive and has no registered success criterion.

Dataset	Table	A	O	$[L, U]$	Gap bracket
Pet	Measured even	76	97	$[84, 84]$	$[13, 13]$
Pet	Measured odd	78	97	$[84, 84]$	$[13, 13]$
Pet	Spectral, fit on even	75	87	$[76, 76]$	$[11, 11]$
CIFAR-10	Measured even	72	94	$[79, 79]$	$[15, 15]$
CIFAR-10	Measured odd	72	90	$[78, 78]$	$[12, 12]$
CIFAR-10	Spectral, fit on even	73	84	$[75, 75]$	$[9, 9]$

$88/94$ images, versus $98/96$ for measured even versus odd. All 100 images in each comparison have resolved pointwise certificates. When applied to measured odd losses without further fitting, the adopted reconstructed-table weights correctly classify $75/72$ images, versus $80/72$ for the measured-even incumbents and $78/72$ for uniform weights. These transfers depend on the selected incumbents. For the selected incumbents, intersecting the odd table's set of reachable-but-missed images with the corresponding even and reconstructed sets gives $9/13$ versus $5/13$ on Pet, and $8/12$ versus $5/12$ on CIFAR-10, using the odd set size as denominator. However, six and four of those odd-table misses are unreachable in the reconstructed Pet and CIFAR-10 tables, respectively. Restricting each comparison to images reachable in both tables gives overlap fractions $9/12$ versus $5/7$ on Pet and $8/12$ versus $5/8$ on CIFAR-10. The compared sets differ in size and reachable domain, and optimal incumbents can be non-unique; these descriptive overlaps do not identify canonical conflict sets. Positive or similar gap sizes do not establish the same conflict sets, spectral specificity, or a common Gaussian mechanism. This is a same-image representation check with an independent noise half, and does not validate the Gaussian bridge.

Numerically unresolved 16-dimensional covariance fits. A source-only protocol on the fresh 500-image PIL mean-squared-error (MSE) panel uses a class-balanced $300/200$ source/target split and 500 disjoint calibration images to fix a 16-dimensional projection and class means. Neither its Cholesky fit nor its Jacobi-reparameterised continuation meets the fixed optimisation eligibility criterion; neither produces a target prediction or accesses target losses for evaluation. This is numerical non-resolution, not evidence against the Gaussian model family. The completed 32-dimensional study in Appendix Z is a separate protocol.

G Shared-weight optimisation and held-out evaluation

Table 8 collects the gap brackets discussed in this appendix.

Relation to published SD 1.5 accuracy. Table 10 of Li et al. (2023) reports 87.5% Oxford-IIIT Pet and 85.5% CIFAR-10 accuracy on SD 1.5, whereas uniform pooling on our two audited $T = 21$ panels gives $76/100$ and $72/100$. These are not like-for-like reproductions: Li et al. sample over the full training-time range with an adaptive class-pruning procedure and treat ℓ_1 versus squared ℓ_2 as a dataset-level choice; our reachability audit fixes squared ℓ_2, a shared-noise cache, and the interior support $[250, 750]$, without adaptive pruning. We did not fully decompose the protocol differences; a later paired preprocessing control is reported in Appendix N.4. Accuracy and reachability both depend on the scoring protocol and noise draws; this comparison does not isolate their relationship. The $R = 32$ halves and the $T = 101$ dense-grid check retain the qualitative gap. A preregistered reference-protocol re-measurement (row F of Table 29) moves panel C towards the published operating point: a wider timestep range on a 21-point grid $\{50, 95, \ldots, 950\}$, $R = 64$, the same fixed panel and scorer, and the ℓ_1 loss that Li et al. (2023) select for CIFAR-10, recorded from the same forward passes as the squared ℓ_2 loss; the analysis criteria and a 0.80 reference-accuracy threshold

Table 8: Rowwise decomposition on four of the six fixed tables in Table 1. A is uniform accuracy under the row's strict threshold, O the label-aware pointwise ceiling, and $G^\star$ the shared-weight optimum. Every row reports $O - G^\star = (O - A) - (G^\star - A)$. On the SD 1.5 rows machine-checked exact certificates establish O and $G^\star$, so those entries are single values; PixArt-α remains solver-grade and keeps intervals that retain the dependence on the same unknown $G^\star$. All rows have $n = 100$; SD uses $\tau = 10^{-6}$ and PixArt $\tau = 10^{-12}$ on per-image max-normalised contrasts.

Dataset	Model	(T, R)	A	O	shared $G^\star$	$O - A$	$G^\star - A$	$O - G^\star$
Pet	SD 1.5	$(21, 64)$	76	97	86	21	10	11
CIFAR-10	SD 1.5	$(21, 64)$	72	91	81	19	9	10
Pet	PixArt-α	$(6, 32)$	44	78	$[48, 52]$	34	$[4, 8]$	$[26, 30]$
CIFAR-10	PixArt-α	$(6, 32)$	46	78	$[55, 57]$	32	$[9, 11]$	$[21, 23]$

were specified before the first forward pass. On the primary ℓ_1 arm, uniform accuracy is $79/100$ against 78 on the interior grid (a reference operating point below the prespecified 80% threshold and the published 85.5%), $O_\tau = 98$, and the shared bracket is $[L, U] = [87, 92]$ with L postchecked and U a solver bound, so $O_\tau - G^\star_\tau \in [6, 11]$: the registered persistence criterion $O_\tau - U \geq 5$ holds and the closure criterion $O_\tau - L \leq 2$ fails. Both $R = 32$ halves persist ($O_\tau - U = 5$ and 7); at $\tau \in \{0, 10^{-8}, 10^{-4}\}$ the persistence margin $O_\tau - U$ stays at 6 and only the solver incumbent L moves ($O_\tau - L$ between 7 and 11); and the supporting squared-ℓ_2 arm on the same forward passes (uniform accuracy 65, bracket $[90, 93]$ against $O_\tau = 99$) persists as well. The two arms separate the scorer from the coupling: writing $O_\tau - G^\star_\tau = (O_\tau - A) - (G^\star_\tau - A)$, the ℓ_1 arm has $O_\tau - A = 19$ and $G^\star_\tau - A \in [8, 13]$ while the squared-ℓ_2 arm has $O_\tau - A = 34$ and $G^\star_\tau - A \in [25, 28]$. Thus the lower-accuracy arm has more recoverable headroom under shared weighting. Both gap lower bounds are 6, but the overlapping brackets do not establish equal sharing costs. This is a protocol intervention on one fixed panel and checkpoint, solver-grade and without an exact certificate; changing the support and the loss alone left uniform accuracy below the published 85.5%. The later PIL control reaches $89/100$ on this panel, but the external-validity question remains open.

The shared-weight programs. Write $D_i \in \mathbb{R}^{T \times (C-1)}$ for image i's normalised contrast matrix, $(D_i)_{tj} = (L_{tj} - L_{ty_i})/s_i$ with the registered per-image scale s_i, so that $(D_i^\top w)_j$ is the margin of the true label over competitor j under w and image i is strictly correct when $\min_j (D_i^\top w)_j > \tau$; the pointwise LP of Equation 1 maximises that margin per image. The shared program is a grouped maximum-feasible-subsystem formulation (Amaldi & Kann, 1995), written as the mixed-integer linear program

$$\max_{w \in \Delta_T,\ z \in \{0,1\}^n} \sum_{i=1}^{n} z_i \quad \text{s.t.} \quad (D_i^\top w)_j \ \geq \ \tau - M_i(1 - z_i) \ \text{ for every } i \text{ and every } j \neq y_i,$$
$$M_i = \max\bigl\{0,\ \tau - \min_{t,j}(D_i)_{tj}\bigr\}, \tag{5}$$

where the big-M choice makes an unselected image's constraints vacuous ($z_i = 0$ asks only for $(D_i^\top w)_j \geq \min\{\tau, \min_{t,j}(D_i)_{tj}\}$, which every simplex w satisfies). The constraints are closed ($\geq \tau$), so the optimum of equation 5 upper-bounds the strict count $G^\star_\tau$ and need not be attained at strict margin by the weighting it returns: U is the solver's bound on this closed optimum, L is the strict recount of the returned w, and the certification instance replaces the objective by the feasibility constraint $\sum_i z_i \geq K$, whose infeasibility proves $G^\star_\tau \leq K - 1$. The bracket closes only when a separately verified strict witness reaches U.

This separates the uniform policy's errors into

$$n - A = (n - O_\tau) + (O_\tau - G^\star_\tau) + (G^\star_\tau - A).$$

The terms count individual non-reachability, the cost of sharing, and remaining improvement within the shared family on this fixed table. The last term is neither a forecast of held-out improvement nor a cross-entropy optimiser gap. A bracket for $G^\star_\tau$ couples the latter two terms; their interval endpoints cannot be combined independently.

Exact certification of the upper bounds. We use rational MILP certification and the VIPR verification framework (Cheung et al., 2017; Eifler & Gleixner, 2024). Every coefficient of the frozen per-image max-normalised binary64 contrasts, including $\tau = 10^{-6}$, is exported exactly through `float.as_integer_ratio()`; a round-trip check verifies the exported model against these dyadic rationals. The SCIP 10 solver (Solving Constraint Integer Programs) in exact rational mode, with presolve, separation and conflict analysis disabled, proves the closed $K = 87$ and $K = 82$ problems infeasible for Pet and CIFAR-10. VIPR verifies both proofs. Exact pointwise primal witnesses and dual mixtures over rival classes establish $O_\tau = 97/91$: the dual bound $\max_t(D_i p)_t \le \tau$ certifies the remaining $3/9$ images unreachable. Shared rational witnesses attain $86/81$ strict successes, matching the upper bounds and establishing exact gaps of $11/10$ on the frozen tables. The supplement contains the exported LP models, certificate files, rational pointwise witnesses, hashes, checker records and a verifier. These results concern the fixed tables and do not establish population optima.

A post-hoc class-stratified image bootstrap resamples with replacement within true class and refits the shared weighting in every one of $1{,}000$ replicates. All replicates return usable brackets. For Pet, $O - L$ has median 13 and percentile interval $[8, 19]$, while the conservative $O - U$ has median 10 and interval $[6, 14]$; for CIFAR-10 the corresponding values are 13 $[7, 19]$ and 8 $[4, 13]$. The conservative intervals exclude zero on both datasets. This is stability to image-panel resampling on tables already averaged at $R = 64$, not a fresh sample or a treatment of finite-noise, model, checkpoint, or timestep-grid uncertainty.

Held-out evaluation of a fixed incumbent. The validation-selected shared weighting is a postchecked feasible incumbent, with global optimality unresolved at the solver time limit. The held-out experiment evaluates this frozen weighting, not the best possible target weighting.

Both policies are evaluated on the same $4{,}500$ previously unused CIFAR-10 images, exactly 450 per class. This paired evaluation gives $3{,}468$ uniform-correct and $3{,}517$ incumbent-correct images, with discordances $(b, c) = (190, 141)$ and a balanced paired difference of $+1.0889$ points. The preregistered class-stratified Wald standard error is 0.3962 points, giving $t = 2.7486$, a one-sided $p = 0.0030$, a one-sided 95% lower bound of $+0.4373$ points, and a descriptive two-sided interval $[+0.3124, +1.8654]$. These inferential quantities have an image-superpopulation interpretation only under within-class exchangeability; the fixed-panel counts are descriptive without it. The mandatory centred, studentised within-class bootstrap sensitivity gives $p = 0.0032$ and interval $[+0.3209, +1.8683]$ points. Effects are not class-uniform: six class differences are positive, one is zero, and three are negative, ranging from -10.00 to $+11.56$ points. Neither this heterogeneity nor any secondary oracle analysis changes the registered primary estimand, but both constrain its interpretation. On the same complete cache, exact dyadic witnesses certify that the true class is per-input reachable on $4{,}331/4{,}500 = 96.24\%$ of images. This label-aware, image-specific lower bound is 18.09 points above the frozen incumbent's 78.16% accuracy, but it permits a different weighting per image and is therefore a pointwise-ceiling gap, not an attainable improvement by one shared rule.

H From the denoising loss to Equation 2

Equation 2 is the only bridge between the diffusion classifier and the metric geometry of Sections 5.1–5.2, so we give it in full, together with the assumptions it consumes. Nothing here is new; it is written out because everything downstream rests on it.

Parameterisation. Stable Diffusion (Rombach et al., 2022) is trained with ϵ-prediction on a variance-preserving (VP) forward process (Ho et al., 2020), $x_t = \bar\alpha_t^{1/2} x_0 + (1 - \bar\alpha_t)^{1/2}\epsilon$ with $\epsilon \sim \mathcal{N}(0, I)$ and $\bar\alpha_t = \prod_{s \le t}(1 - \beta_s)$ fixed by the model schedule (`scaled_linear`, $\beta \in [0.00085, 0.012]$, 1000 steps). Dividing by $\bar\alpha_t^{1/2}$ puts it in variance-exploding (VE) form,

$$\tilde x_t := \bar\alpha_t^{-1/2} x_t = x_0 + \sigma_t \epsilon, \qquad \sigma_t^2 := \frac{1 - \bar\alpha_t}{\bar\alpha_t}, \tag{6}$$

with the *same* ϵ. This defines the σ of equation 2: it is not a free parameter but the log signal-to-noise ratio (log-SNR) reparameterisation of the model schedule, and every σ_t quoted in Section 5 and Appendix F is computed from that $\bar\alpha_t$. Here x_0 is the variational autoencoder (VAE) latent after the scaling factor, which is why λ of order one corresponds to unit latent variance.

The one modelling step. Assume the model's class-conditional law for the clean latent is Gaussian with a covariance shared across classes,

$$q(x_0 \mid c) = \mathcal{N}(\nu_c, \Sigma_q), \tag{7}$$

and that the network realises the *exact* minimiser of its own denoising objective under equation 7. Then $\tilde{x}_t \mid c \sim \mathcal{N}(\nu_c, \Sigma_q + \sigma_t^2 I)$ and the optimal ϵ-predictor is the affine map

$$\hat{\epsilon}(\tilde{x}_t, c, \sigma) = -\sigma \nabla_{\tilde{x}} \log q_\sigma(\tilde{x}_t \mid c) = \sigma\,(\Sigma_q + \sigma^2 I)^{-1}(\tilde{x}_t - \nu_c) =: D_\sigma(\tilde{x}_t - \nu_c). \tag{8}$$

This is the step that is idealised, and no theorem in this paper implies a real UNet satisfies it. The predictor is linear in the centred input $\tilde{x}_t - \nu_c$, with D_σ the same for every c; both are testable, and Appendix L probes their local necessary conditions without establishing that a real UNet satisfies the idealisation.

What survives the $\arg\min$**.** Substituting equation 6 into equation 8, $\hat{\epsilon} - \epsilon = D_\sigma(x_0 - \nu_c) + (\sigma D_\sigma - I)\epsilon$, and since $\mathbb{E}[\epsilon] = 0$ the cross term vanishes:

$$L(x_0, c, \sigma) = \mathbb{E}_\epsilon \big\|\hat{\epsilon} - \epsilon\big\|^2 = \underbrace{(x_0 - \nu_c)^\top D_\sigma^\top D_\sigma (x_0 - \nu_c)}_{\text{class-dependent}} + \underbrace{\operatorname{tr}\big[(\sigma D_\sigma - I)(\sigma D_\sigma - I)^\top\big]}_{\text{class-independent}}. \tag{9}$$

The second term depends on σ and Σ_q but not on c, so it is common to every candidate label and cancels in the $\arg\min$ for any weighting. The shared covariance also makes $D_\sigma^\top D_\sigma$ common to all classes. With $\Sigma_q = Q\,\mathrm{diag}(\lambda)Q^\top$,

$$D_\sigma^\top D_\sigma = \sigma^2(\Sigma_q + \sigma^2 I)^{-2} = Q\,\mathrm{diag}\Big(\tfrac{\sigma^2}{(\sigma^2+\lambda_k)^2}\Big)Q^\top, \tag{10}$$

which is $v(\sigma)$ of equation 2. Weighting the timesteps by any $w \geq 0$ and dropping the common term,

$$\sum_t w_t\, L(x_0, c, \sigma_t) \;\overset{c}{=}\; (x_0 - \nu_c)^\top M_w (x_0 - \nu_c), \quad M_w = Q\,\mathrm{diag}\Big(\textstyle\sum_t w_t \tfrac{\sigma_t^2}{(\sigma_t^2+\lambda_k)^2}\Big)Q^\top = f_w(\Sigma_q), \tag{11}$$

where $\overset{c}{=}$ is equality up to a constant in c. This is equation 2, and the rule $\arg\min_c (x - \nu_c)^\top M_w (x - \nu_c)$ is the nearest-centroid form used throughout.

The full assumption list. Collected in one place, the assumptions of Sections 5.1 and 5.2 are: (i) the true law is $x \mid c \sim \mathcal{N}(\nu_c, \Sigma_p)$ with equal priors and distinct means; (ii) the model's law is equation 7, Gaussian with covariance Σ_q shared across classes; (iii) the denoiser is the exact optimum equation 8 for that law; and (iv) **the model's class means are the true ones**, the same ν_c in equation 7 as in the true law. Assumption (iv) is what confines the analysis to *covariance* misspecification: if the model used $\tilde{\nu}_c \neq \nu_c$ the rule would compare distances to different centroids, so preserving the true Bayes cells would also require matching their boundary offsets. Theorem 1 would not apply as stated. Mean misspecification is outside this paper.

I Checkpoint dependence of reachability and spectral reconstruction

Every SD 1.5 number in Section 4 comes from one set of model weights. We therefore re-ran the CIFAR-10 protocol unchanged — same images, seed, timesteps, trials, precision and scorer, changing only the checkpoint — on two further checkpoints: `CompVis/stable-diffusion-v1-1`, an earlier point of the *same training run* — 515,000 further steps at 512^2 reach the shared v1-2 checkpoint and 595,000 more reach v1-5, so 1.11M steps separate the two — and `Lykon/dreamshaper-8`, a finetuned descendant of it. All three share the architecture, the ϵ-parameterisation and the noise schedule, so σ_t is identical. Under the idealisation, however, each checkpoint can have its own Σ_q, including both its eigenvalues and eigenvectors. This comparison does not include SD 2.x. The evaluation criteria were fixed before measurement.

The preregistered questions concern transfer of the invariant/unreachable-truth reachability profile, the reconstruction advantage of a fixed spectral basis with $\lambda \in \{0.5, 2\}$ over two matched controls, and the ordering $R^\star > R_{\mathrm{sat}}$ (Table 9).

Table 9: CIFAR-10 reachability across checkpoints. Reachable counts are mean class-set sizes; invariant and unreachable-truth entries are image percentages.

CIFAR-10, $T = 6$, $R = 64$	uniform acc.	reachable	invariant	unreach. truth	R_{sat} / $R^\star$
SD 1.5 (Appendix M)	0.73	3.09	15%	9%	16 / > 32
SD 1.1	0.70	2.52	29%	14%	4 / 32
DreamShaper-8	0.15	3.87	0%	56%	4 / > 32

The budget ordering transfers, but the reachability profile does not. The budget separation $R^\star > R_{\text{sat}}$ holds on both new checkpoints. The reachability profile does not: the registered criterion required invariant $\geq 10\%$ and unreachable-truth > 0 on all three, and DreamShaper-8 gives 0% invariant. This fails the prespecified profile-transfer criterion. In the same comparison, DreamShaper-8 scores 0.15 against a chance level of 0.10, whereas SD 1.1 scores 0.70 and has more invariant images than SD 1.5 (29% versus 15%). These descriptive differences do not isolate accuracy as the cause of the failed profile transfer; above-chance accuracy alone does not establish the registered reachability profile.

The fixed spectral basis does not transfer. We registered that the frozen theory basis $\lambda \in \{0.5, 2\}$ would beat the two null families on both new checkpoints, with the nulls keeping the bases selected on SD 1.5, without selection on either target checkpoint. This criterion fails. On SD 1.1 the theory basis has larger residuals, with median held-out residual 0.516 against 0.468 and 0.500, winning on only 38% and 45% of images; on DreamShaper-8 it wins on 55% and 63%, short of the registered $p < 10^{-3}$ on both. Neither checkpoint meets the prespecified criterion. Under the model, the λ are eigenvalues of a checkpoint's own Σ_q, so the theory does not require them to transfer as fixed numbers. Appendix L probes local within- and cross-timestep necessary conditions at one checkpoint, not whether eigenvalues transfer between checkpoints, and does not establish transfer of this fixed basis. Allowing checkpoint-specific eigenvalues is a possible model-based interpretation, but was not the registered prediction.

Post-hoc interpretation. First, neither new checkpoint is a different model family: one is the same run earlier and the other a finetune of it, so these runs test transfer across *training time* and finetuning. The later PixArt-α experiment in Appendix K crosses to a text-conditioned Transformer family, but tests finite-table jointness and support range rather than this frozen-basis transfer. Second, a structure carried by an intermediate checkpoint need not survive to a later one; that a teacher's implicit curriculum lives in the intermediate checkpoints and not the converged one is reported, in a quite different setting, by Panigrahi et al. (2025). Both readings postdate the result and neither is offered as a mechanism.

J SHARED GAPS ON NEW IMAGES AND DENSER TIMESTEP GRIDS

The fresh-panel protocol (C) was specified before measurement, apart from a timing-only rehearsal whose losses were not used. Its 100 CIFAR-10 indices have zero overlap with the reconstructed 9,707-index union recorded in prior project artifacts. They are a deterministic sample from the 293 indices left after that exclusion, not a representative population sample. Endpoint E reuses the panel and common $R = 32$ cells while nesting grids over $[250, 750]$. The endpoints are dependent and address different questions: sensitivity to image selection and to grid density. Their criteria were registered separately.

Persistence required $O-U \geq 5$, fixed before results. The $T = 101$ even and odd $R = 16$ halves both retain a gap lower bound of 6; the registered τ sensitivity and true-label nesting checks also satisfy their prespecified criteria without numerical instability. Within E, the positive per-image reference scale computed from the full $T = 101, R = 32$ table is reused at every rung, so zero-extension preserves the normalised margin. C and E use different registered endpoint-reference normalisations, so their common-cell $T = 21$ bracket values are not interchangeable. Nor is monotone convergence of the gap assumed: the ladder is a finite-grid description, not a limit argument.

Table 10: Fresh-panel (C) and dependent nested-grid (E) results at $\tau = 10^{-6}$; every row meets the prespecified persistence criterion $O - U \geq 5$. A, O, and $G^\star$ use the same rowwise strict-threshold estimand, and the last three columns display $O - G^\star = (O - A) - (G^\star - A)$. L is independently postchecked and U is solver-derived. The dagger marks the two full-table rows where exact certificates establish O and $G^\star$, making those entries single values; all other rows keep solver-grade intervals.

endpoint	noise block	T	A	O	shared $G^\star$	$O - A$	$G^\star - A$	$O - G^\star$
C	full $R = 64$	21	78	97	89	19	11	$8^\dagger$
	even $R = 32$	21	77	98	$[86, 89]$	21	$[9, 12]$	$[9, 12]$
	odd $R = 32$	21	79	99	$[87, 90]$	20	$[8, 11]$	$[9, 12]$
E	full $R = 32$	6	76	95	$[87, 88]$	19	$[11, 12]$	$[7, 8]$
	full $R = 32$	11	77	97	$[86, 88]$	20	$[9, 11]$	$[9, 11]$
	full $R = 32$	21	77	98	$[85, 90]$	21	$[8, 13]$	$[8, 13]$
	full $R = 32$	101	80	99	92	19	12	$7^\dagger$

For the two daggered full-table instances only, SCIP exact-mode infeasibility certificates checked by `viprchk` establish $G^{>}_\tau \leq 89$ at C/$T = 21$ and $G^{>}_\tau \leq 92$ at E/$T = 101$, and an exactly verified shared weighting attains strict margin on 89 and 92 images, establishing $G^\star_\tau = 89$ and 92. Exact rational pointwise primal witnesses at the same binary64 τ establish $O_\tau \geq 97$ and ≥ 99, and exact dual witnesses certify the remaining 3 and 1 images unreachable, establishing $O_\tau = 97$ and 99: the daggered gaps are exactly 8 and 7 on those dyadic table readings. This does not certify either L endpoint of the undaggered rows; all split halves, sensitivity rows, and E rungs $T \in \{6, 11, 21\}$ remain solver-grade. Thus C addresses specificity to the original 100 images, and E addresses a coarse finite $T = 21$ grid within $[250, 750]$; neither is a population bound, a continuum-support theorem, or an independent pair of replications.

K Cross-family finite-table replication on PixArt-α

PixArt-α is a text-conditioned latent Transformer with ϵ-prediction; its eight-channel output stores the predicted noise in the first four channels. The replication fixes one checkpoint revision and reuses the ordered 100 CIFAR-10 indices, class-prompt strings, and cell-seed rule of the SD 1.5 run. Twelve PixArt indices are matched by noise scale to the SD 1.5 coordinates $\{50, 230, 250, 350, 410, 450, 550, 590, 650, 750, 770, 950\}$. The seed is derived from the SD coordinate, not the mapped PixArt index. This is common-random-number coupling rather than bit-identical noise: the source run drew fp16 noise and PixArt drew fp32. The VAE weights, T5 (text-to-text transformer) rather than CLIP text conditioning, denoiser family and training, scheduler, and precision are not experimentally held fixed (the VAE is loaded from the PixArt repository), so the experiment is a cross-model-family replication rather than a causal architecture ablation.

The CIFAR-10 measurement contains 384,000 class-conditioned evaluations and 38,400 complete, finite image–timestep–trial cells. Its measurement design was fixed before measurement; the analysis plan was fixed during data collection, before examining the results. Thus analysis planning followed the start of measurement, rather than preceding it. Reproduction uses the frozen executed runner, whose hash and verification records accompany the release.

Table 11: PixArt-α CIFAR-10 shared-weight bounds on the full noise block and its halves. Counts are out of 100; upper bounds remain solver-grade.

noise block	per-input O	verified L	solver U	$O - U$	$O - L$
full $R = 32$	78	55	57	21	23
even $R = 16$	80	54	56	24	26
odd $R = 16$	79	53	55	24	26

Here L is recomputed by applying the returned weighting at the registered strict margin; U is the floating-point MILP dual bound. Table 11 therefore gives solver-grade brackets, not exact-rational strict-margin optima. The full block and both halves meet the prespecified criterion $O - U \geq 5$ at cutoffs $0, 10^{-12}, 10^{-10}, 10^{-8}$. The positive gap therefore persists across the tested noise blocks and cutoffs on this panel.

Table 12: PixArt-α CIFAR-10 support intervention. Each added pair contributes two timesteps; all arms use the same 100 images and 32 noise draws per timestep.

full-$R = 32$ arm	invariant / 100	mean reachable	uniform accuracy
base six	10	3.38	0.46
base $+\{410, 590\}$ (interior)	10	3.57	0.47
base $+\{230, 770\}$ (near)	5	3.80	0.44
base $+\{770, 950\}$ (high-only)	3	4.40	0.48
base $+\{50, 950\}$ (far)	0	5.48	0.33

The secondary range rule reaches its threshold exactly (Table 12): the interior-minus-far contrast is $10/100$, with the directional ordering $10 \geq 5 \geq 0$. Its even and odd $R = 16$ contrasts are only 8 and 7, although both retain the same sign, and the pair-arm accuracy range is 15 points, above the prespecified 10-point accuracy-range threshold. The full-block contrast meets the support-effect criterion, but the noise halves do not; nor is the effect isolated from classification strength. The SD 1.5 resolution experiment also identifies 32×32 inputs as sensitive to support location. The support result is therefore secondary and is assessed separately from the shared-weight gap, whose criterion does not imply it.

Native-resolution Oxford-IIIT Pet replication. The Pet arm repeats the fixed CIFAR-10 design with only the dataset (and hence 37 classes) changed: the same 100-image order as the SD 1.5 Pet cache, the same twelve coordinates, the same seed rule, 1,420,800 class-conditioned forward-equivalents. Its analysis plan was independently checked and fixed during measurement, before examining the results; it was not a pre-measurement registration. At the primary cutoff the base-support pointwise ceiling is $O = 78$; the independently postchecked shared weighting realises $L = 48$ and the floating-point solver supplies $U = 52$, hence $26 \leq O - G^\star \leq 30$. Both $R = 16$ halves (24–26 and 25–27) and all four LP cutoffs retain the positive gap and meet the prespecified persistence criterion. The support-effect co-primary does not meet its criterion: the interior-minus-far invariant contrast is $5 - 0 = 5$ against the registered $10/100$ threshold with the directional ordering $5 \geq 2 \geq 0$ preserved, base-support uniform accuracy 0.44 above the 0.40 diagnostic threshold, and a 2-point pair-arm accuracy span within the 10-point accuracy-range threshold. The shared gap and support effect thus have different replication outcomes; neither establishes population, checkpoint, or continuous-support behaviour.

L Local necessary conditions in the real UNet

Appendix H isolates the single idealisation the theory rests on: that the network realises the exact optimum equation 8 for a Gaussian law with a covariance *shared* across classes. Under equation 8 the Jacobian $J = \partial\hat{\epsilon}/\partial\tilde{x}$ equals $D_\sigma = \sigma(\Sigma_q + \sigma^2 I)^{-1}$, hence is (S) symmetric, (C) identical across classes, and (L) independent of $\tilde{x}$. The first probe below tests those three necessary conditions locally and at each timestep separately. A second probe tests a necessary cross-timestep resolvent identity. Neither is sufficient for the idealisation. The VP-to-VE rescaling of equation 6 cancels from the within-timestep ratios; the cross-timestep identity instead retains its scale factors explicitly.

Estimator. For $u, v \sim \mathcal{N}(0, I_d)$ independent and any matrix M, $\mathbb{E}[\langle v, Mu\rangle^2] = \|M\|_F^2$. Writing $a = \langle v, Ju\rangle$ and $b = \langle u, Jv\rangle$,

$$r_S = \sqrt{\frac{\mathbb{E}[(a-b)^2]}{\mathbb{E}[a^2]}} = \frac{\|J - J^\top\|_F}{\|J\|_F}, \qquad r_C = \frac{\|J^{(c)} - J^{(c')}\|_F}{\|J\|_F}, \qquad r_L = \frac{\|J_x - J_{x'}\|_F}{\|J\|_F}, \quad (12)$$

each estimated from Jacobian-vector products (JVPs). The scale is absolute: 0 means the property holds exactly, $\sqrt{2}$ is the large-d root-mean-square ensemble value for a mean-zero i.i.d. unstructured matrix (the ratio of expected squared norms is exactly $2(1-1/d)$), and 2 for a purely antisymmetric one. Jacobian-vector products are taken by central finite differences with the same attention kernel the measurement scorer runs; forward-mode automatic differentiation has no rule for the fused attention kernel, and forcing the reference kernel made one timestep exceed fifty minutes. The two finite-difference steps below provide an empirical check on sensitivity to the step scale; no upper-bound interpretation is used.

Estimator validation. To check sensitivity to asymmetry, we first apply the same code to linear maps with known answers: a symmetric M gives $r_S = 0.0000$, a random M gives 1.389 (against $\sqrt{2} = 1.414$), an antisymmetric M gives 2.0000 exactly, and adding a bias changes nothing. On $M = S + \varepsilon(A - A^\top)/2$ the measured value tracks $\|M - M^\top\|_F / \|M\|_F$ to two decimals down to $\varepsilon = 0.01$ (true 0.0199, measured 0.0185), so the detection floor is about 0.02.

The main estimate uses 100 images. At each (t, h) condition we pool 3,200 projections for r_S and 800 sampled class-pair or input-pair projections for r_C and r_L (Table 13).

Table 13: Pooled local Jacobian residuals at two relative finite-difference steps. These are necessary-condition diagnostics, not Gaussian-model validation.

	$t = 250$		$t = 450$		$t = 750$	
step h_{rel}	10^{-3}	10^{-2}	10^{-3}	10^{-2}	10^{-3}	10^{-2}
r_S symmetry	0.0670	0.0665	0.0532	0.0531	0.0321	0.0311
r_C class-invariance	0.0340	0.0305	0.0424	0.0405	0.0370	0.0416
r_L constancy in x	0.1186	0.1190	0.0948	0.0888	0.0601	0.0654

Local residuals. On a frozen Stable Diffusion 1.5, at the preregistered primary step $h_{\mathrm{rel}} = 10^{-3}$ the timestep medians are $r_S = 0.0532$ and $r_C = 0.0370$ — 3.8% and 2.6% of the unstructured scale $\sqrt{2}$. The registered decision statistic is therefore $\max(0.0532, 0.0370) = 0.0532$, below the prespecified 0.30 threshold for this local necessary-condition diagnostic. At $h_{\mathrm{rel}} = 10^{-2}$ the corresponding medians are 0.0531 and 0.0405. The pooled estimates remain similar under this tenfold step change, arguing against finite-difference scale as their dominant source. In the original $n = 4$ run the primary-step medians were 0.0491 and 0.0359; thus the same local necessary-condition conclusion survives the 25× increase to 100 images. None of these within-timestep values tests the cross-timestep shared-covariance structure.

Cross-timestep resolvent check. Write $a_t = \sqrt{\bar{\alpha}_t}$ and $K_t = a_t J_t^{\mathrm{VP}}$. One covariance would imply, for every pair, $\sigma_t K_s - \sigma_s K_t = (\sigma_t^2 - \sigma_s^2) K_s K_t$. The registered matrix-free test uses 50 CIFAR-10 images, 32 random directions, and pairs among $t \in \{250, 450, 750\}$; its five controls retain every measured Jacobian action while permuting the three σ values. Residuals pool $\|Lu - Hu\|^2$ over the registered directions and normalise by the pooled $\|Lu\|^2 + \|Hu\|^2$, where L and H denote the two sides of the identity. At the primary step, the pooled actual residual is 0.0269 versus 0.8706 for the median wrong-σ control. For the image-level median-pair log contrast D, the equal-class mean is -3.4921, or $\rho = \exp(\overline{D}) = 0.0304$ (one-sided 99% upper bound 0.0336, bootstrap-t $p = 1.0 \times 10^{-5}$). All three pair medians are negative, and the two finite-difference arms give mean $D = -3.4921$ and -3.4925. The comparison meets the prespecified criterion. Omitting the VP factors instead gives residual 0.7621, an internal check on the scaling algebra. The raw commutator is secondary and receives no absolute cutoff.

Native-resolution replication. An unconditional replication was fixed before the CIFAR-10 result was inspected and then run on 50 previously unused native-resolution Oxford-IIIT Pet images, five from each of ten preregistered breeds. Its estimand is the equal-weight mean over those breeds, not the full 37-breed population. With the same timesteps, directions, controls, and decision rule, the primary-step residual is 0.0401 under the actual σ assignment versus 0.8503 under the median

wrong-σ control ($21.2\times$ smaller). The equal-breed mean is $D = -3.0022$, or $\rho = 0.0497$ (one-sided 99% upper bound 0.0541, bootstrap-t $p = 1/(100{,}000 + 1)$). Pair medians for $(250, 450)$, $(450, 750)$, and $(250, 750)$ are -1.9719, -3.0132, and -3.4309, respectively; the two step-size arms give mean $D = -3.0022$ and -3.0029. This comparison also meets the prespecified criterion. This replicates the local necessary-condition comparison beyond upsampled 32×32 CIFAR, but it is not a causal resolution comparison because the dataset, classes, images, and prompts also differ.

Held-out operator-action prediction. A stronger registered probe asks whether the source action at $s = 250$ predicts target actions at $t \in \{450, 750\}$. Using the VE Jacobian $K_s = \sqrt{\bar{\alpha}_s} J_s^{\mathrm{VP}}$ defined above, put $R_s = K_s/\sigma_s$ and $\delta = \sigma_t^2 - \sigma_s^2$. The test solves $(I + \delta R_s)x = R_s v$ by unrestarted generalised minimal residual iteration (GMRES) and predicts $K_t v = \sigma_t x$ before opening the direct target JVP. The experiment uses 400 image–step–direction cells and $42{,}400$ UNet forwards. Predictive performance was examined only after all measurements were complete.

The prediction test is numerically inconclusive because its GMRES residuals and finite-difference step sensitivity exceed the prespecified tolerances. Across the two finite-difference steps, the shared/naive error ratios are 0.0300 and 0.0466, whereas shared/isotropic is 2.0534 and 3.1860. Thus the resolvent transform is 21–33$\times$ better than naive transfer but 2.05–3.19$\times$ worse than v/σ_t. The maximum full and projected GMRES residuals are 1.2245 and 0.1580, both above their 10^{-3} tolerances, although the Arnoldi orthogonality residual is only 1.55×10^{-15}; the two registered step-size ratios are both 1.552, outside $[0.8, 1.25]$. A separately prespecified wrong-σ sensitivity analysis has no direction reversal: its median log shared/wrong ratios are $(-0.826, -3.917)$ at $h = 10^{-3}$ and $(-0.638, -3.519)$ at $h = 10^{-2}$, all negative. This directional advantage does not resolve the numerical convergence and step-sensitivity failures.

A post-hoc technical convergence probe then replaced the finite-difference source action with exact JVP (which also required the MATH attention backend) on one already exposed Pet image from the cross-timestep probe. For the $(250, 450)$ pair, the residual falls from the original finite-difference value 1.224528 to 0.042134, a $29.06\times$ change, but remains above 10^{-3} after all 60 iterations. The probe opens no target JVP and emits no efficacy statistic. Because the JVP method and the attention backend change together, this comparison does not identify finite-difference error alone as the source of the reduction.

The isotropic baseline was necessary to distinguish the fitted predictor from a nondiscriminative alternative. Before GPU execution, the synthetic test exhibited wrong-law relative error 0.0022 in a high-noise regime, nearly matched by the isotropic error 0.0020; an absolute-error rule could therefore have passed a nondiscriminative predictor. Requiring the shared arm to beat v/σ_t prevented that spurious conclusion here. Consequently this experiment does not establish one shared Σ_q or a theory–UNet bridge; the cross-timestep probe's sampled necessary-condition residual comparison remains the extent of the positive cross-timestep evidence.

Scope of the local evidence. The estimated r_L is nonzero and is the largest of the three residuals in all six conditions, ranging from 0.0601 to 0.1190. This indicates input dependence of the measured Jacobian actions; it does not establish an approximation bound for the Gaussian denoiser. At each tested noise level separately, the Jacobian is nearly symmetric and the pooled differences across sampled class pairs are small. These are local necessary conditions for the shared form, but three unrelated covariances, one per timestep, could satisfy all of them. This within-timestep probe does not test whether a single Σ_q generates the maps across timesteps, including their shared eigenbasis and resolvent law, and therefore does not establish that the score is that of a Gaussian. The scale-up covers 100 images from one dataset and one checkpoint, but its pooled artifact did not retain the preregistered per-image dispersion, so it does not quantify image-to-image heterogeneity. The same omission prevents the preregistered diagnostic that re-runs the first four images: the estimates remain below the fixed threshold, but the numerical shift from the original run cannot be decomposed as planned. The two cross-timestep dataset probes are stronger than these within-timestep checks, but their results concern only the sampled local actions. They neither reconstruct a single Σ_q nor prove a shared eigenbasis, Gaussian score, or global affinity; they cover two datasets at one checkpoint.

M CALIBRATION NULLS FOR THE REACHABILITY COUNTS

Section 4 reports small reachable sets. The count alone does not identify their cause: a mean of 2.55 reachable classes could reflect the frozen model's timestep structure, or it could be what any six-dimensional nonnegative mixture over 37 items would give. The controls below distinguish these explanations. All act on the same trial-averaged table L_{tc} the linear program consumes, and all reuse that program unmodified, so every difference below is a difference in the input table.

Table 14: **Reachable-set sizes under marginal-preserving and rank-preserving controls.** Null A permutes the class axis independently at each (image, timestep): the exact multiset of class losses at every (image, timestep) survives, only the cross-timestep coupling is destroyed. Null B redraws each row i.i.d. Gaussian with that row's mean and standard deviation. Null C keeps the class *ordering* at every (image, timestep) — so every per-timestep $\arg\min$ is unchanged — and replaces the values by an evenly spaced grid over the row's range, destroying only the margins. Stochastic nulls are 10 repeats, mean $\pm$ s.d. (their invariant fraction is 0 in every repeat); null C is deterministic.

	Oxford-IIIT Pet ($C = 37$)			CIFAR-10 ($C = 10$)		
Table	reachable	invariant	unr. truth	reachable	invariant	unr. truth
Measured	2.55	30%	3%	3.09	15%	9%
A class permutation	15.76 ± 0.28	0%	$60 \pm 5\%$	7.05 ± 0.12	0%	$31 \pm 5\%$
B moment matched	15.72 ± 0.45	0%	$58 \pm 4\%$	7.34 ± 0.09	0%	$29 \pm 4\%$
C rank preserving	2.74	30%	3%	3.50	15%	10%

Two comparisons are informative. Null A yields $6.2\times$ and $2.3\times$ as many reachable classes as the measured tables, and reduces the invariant fraction to zero: destroying the cross-timestep coupling while keeping every marginal leaves *no* image whose label a weighting cannot move, against 30% and 15% measured. These differences implicate cross-timestep loss structure rather than the number of weighting coordinates alone. Under null C, equally spacing losses while preserving their order and range changes the counts by 0.19 and 0.41 and the invariant fractions not at all. Together, these controls identify cross-timestep class-order agreement as a proximate explanation of reachable-set size, without attributing this agreement to a denoiser mechanism.

Densifying the support. This nested-grid comparison uses raw trial-mean tables without per-rung normalisation and the same absolute cutoff 10^{-12}. Under this fixed estimand, $\mathcal{T}_1 \subseteq \mathcal{T}_2$ implies $\mathcal{R}_{\mathcal{T}_1}(x) \subseteq \mathcal{R}_{\mathcal{T}_2}(x)$, since any w on $\mathcal{T}_1$ extends by zeros. Reachable sets can therefore only expand; the question is by how much. We refined $[250, 750]$ from the 6 evaluated timesteps to 11 and then 21, *at fixed* $R = 64$ noise trials per timestep — holding the compute budget $N = T \cdot R$ fixed instead would lower R as T rises, and the split-half analysis below reports that lower R inflates the reachable count, so the two effects would be inseparable. The three grids are nested, so one $T = 21$ table supplies all three rungs. The evaluation criteria were fixed before measurement (Table 15).

Table 15: Reachability on nested timestep grids at fixed $R = 64$. Reachable and null-A entries are mean set sizes; ρ is their null/measured ratio and the ceiling is a fraction of images.

dataset	grid	reachable	null A	ρ	invariant	unreach. truth	ceiling
Pet	$T = 6$	2.55	16.06	6.30	30%	3%	0.97
	$T = 11$	2.72	29.59	10.88	27%	3%	0.97
	$T = 21$	2.92	36.64	12.55	26%	3%	0.97
CIFAR-10	$T = 6$	3.09	7.06	2.29	15%	9%	0.91
	$T = 11$	3.43	9.44	2.75	15%	9%	0.91
	$T = 21$	3.81	9.99	2.62	12%	9%	0.91

Unreachable-truth and the ceiling do not move within either dataset; their endpoints also separate across datasets (3% and 0.97 on Pet versus 9% and 0.91 on CIFAR-10). The invariant fraction falls by four Pet points and three CIFAR-10 points across a $3.5\times$ denser grid. The registered primary

endpoint was the null-relative ratio ρ_T, because the raw fractions are diluted mechanically by the monotonicity above while the null count grows too. Pet continues to separate from the null, $6.30 \to 10.88 \to 12.55$, whereas CIFAR-10 saturates by $T = 11$, $2.29 \to 2.75 \to 2.62$; both exceed the registered support-robust thresholds 4.12 and 1.52 at the final rung. Both null counts approach the class count at $T = 21$ (36.64 of 37 and 9.99 of 10), so the later rungs of ρ_T are ceiling-limited on both datasets: the Pet ratio keeps rising because its measured count moves little while the null count reaches its ceiling, so that rise is not further evidence of structure. The $T = 6$ null A means here (16.06) and in Table 14 (15.76 ± 0.28) come from two scripts with different random-number keys; the older script salted its key with Python's process-dependent string hash without recording the hash seed, so 15.76 identifies the released draw but is not regenerable from the recorded seed alone. The Pet–CIFAR-10 divergence is consistent with the separate resolution intervention but, because it compares datasets rather than resolutions within one image set, does not identify resolution as its cause. The preregistered exchangeability-based statistic also meets its fixed thresholds ($10.2, 9.5, 9.4$ on Pet and $2.1, 1.9, 1.7$ on CIFAR-10), but it uses the exchangeable baseline that the matched top-k analysis below shows is not the right structural comparator. We preserve the registered result without using it as evidence that the reachable set beats a label-blind alternative. These ladders test density inside $[250, 750]$, not a wider support range.

Noise-block stability at $T = 21$. A post-hoc recomputation uses the even and odd $R = 32$ trials from the same completed tensors. On Pet the two halves have mean reachable counts $3.35/3.51$, invariant fractions $19\%/20\%$, and true-unreachable fractions $1\%/1\%$; on CIFAR-10 they give $4.14/4.16$, $13\%/12\%$, and $4\%/6\%$. The even/odd agreement rates for the true-reachable indicator, the invariant indicator, and every image–class reachability cell are $100\%/95\%/97.35\%$ on Pet and $98\%/95\%/89.8\%$ on CIFAR-10. Full-R versus half-block true-reachable agreement is 98% for either Pet half and $95\%/97\%$ for the CIFAR halves. This supports qualitative per-image stability but not noise-independence of the rates. Exact post-hoc primal and explicit-dual checks classify a cell as reachable when its lower bound exceeds 10^{-8} and unreachable when its upper bound is below -10^{-8}; all other cells remain unresolved. These checks resolve all 3,700 Pet and 1,000 CIFAR-10 full-block image–class games and both noise halves. Pet invariance changes markedly from 26% at full R to $19\%/20\%$ in the halves, whereas CIFAR-10 is nearly flat at 12% versus $13\%/12\%$. This is a post-hoc certification with zero new GPU forwards, not fresh confirmation: it removes numerical-tolerance doubt on the realised serialised tables, not Monte Carlo sampling uncertainty, continuous-support reachability, or a population rate.

Support-range follow-up. An exposed-cache, low-R diagnostic first changed Pet invariance from 22.4% to 1.9% when an interior $T = 6$ grid was united with the disjoint grid $\{50, 230, 410, 590, 770, 950\}$. It motivated a fixed-design follow-up (the analysis plan was finalised during measurement) using the same, order-identical 100 images as the dense $T = 21$ table and $R = 64$ throughout. The four pair-addition arms are count-matched, each adding exactly two timesteps — interior $\{410, 590\}$, near-outside $\{230, 770\}$, high-only $\{770, 950\}$, far-symmetric $\{50, 950\}$; the base itself has (invariant count, mean reachable, accuracy) $(26, 2.92, 0.76)$. Interior, near-outside, high-only, and far-symmetric additions yield invariant counts $26/20/10/2$, mean reachable counts $2.94/3.26/4.68/5.55$, and accuracies $0.75/0.77/0.76/0.77$. Thus their roughly equal classification strength does not explain the twenty-four-point invariance range. Adding four outside points gives $(2, 5.81, 0.78)$ and all six gives $(2, 5.85, 0.79)$ for (invariant count, mean reachable, accuracy), with no image having unreachable truth in either arm. The analysis-plan all-six point prediction was $4/100$; the observed $2/100$ has absolute error 0.02, reported without a post-hoc success tolerance. The high-only arm is diagnostic rather than part of the primary comparison. Its exact historical accuracy match to the far-symmetric arm was selected from a separate, independent and previously exposed validation cache of the compute-matched policy comparison (Appendix E); in this experiment their measured accuracies differ by one point, so we do not call them matched.

The support-effect criterion is met at all four LP cutoffs, without numerical instability. After excluding the eight images overlapping the exploratory cache, base/interior/near/high/far counts are $24/24/19/10/2$ out of 92; even and odd $R = 32$ blocks give respectively $19/15/5/2$ and $20/16/7/3$ across the four pair-addition arms, preserving their ordering. The directional secondary gradient $26 \geq 20 \geq 2$ also holds. The measurement design was fixed before data collection; the analysis plan was amended during collection and finalised before examining intervention results. The main predictive test was the four-outside-point arm (invariant count $\leq 9/100$), observed at

$2/100$: the interior-arm prediction ($\geq 20/100$) was largely determined by the exposed base value, and an exhaustive enumeration confirms that the two jointly imply the contrast prediction (≥ 10) because $20 - 9 = 11 \geq 10$. The three criteria therefore do not provide three independent predictive successes. The experiment reproduces the exploratory low-R support-range effect on the same images at $R = 64$; it establishes neither continuous-support nor another-checkpoint behaviour, global shared weighting, or the Gaussian theorem.

The pointwise ceiling. If the uniform weighting clears the chosen strict margin, its predicted label is reachable because that weighting is feasible. Thus $O_\tau \geq A$ when A counts uniform decisions under the same strict-margin criterion; a unique $\arg\min$ alone does not guarantee a margin above $\tau > 0$. Uniform predictions have no ties on the two tables compared below. The ceiling is an upper bound computed with label access, not an attainable gain, and we do not report it as one. The exchangeable calculation $(n_{\mathrm{reach}} - 1)/(C - 1)$ gives baselines of 9.3% and 40.9% on the uniform-wrong Pet and CIFAR-10 images, but Null C shows that exchangeability is the wrong structural comparator: the measured sets largely inherit a shared class ranking. We therefore use a matched, label-blind set-valued comparison. For each image, take the $|\mathcal{R}_\tau(x)|$ classes with smallest uniformly pooled loss, where $|\mathcal{R}_\tau(x)|$ is computed from the label-free raw-margin mask. That mask agrees cellwise with the registered normalised mask on these tables. This top-k set has mean Jaccard 0.815 and 0.744 with $\mathcal{R}_\tau(x)$ and contains the truth on 99 and 90 images, versus 97 and 91 for the reachable set. The paired discordances are only 2 and 5 (two-sided exact McNemar $p = 0.5$ and 1.0); these tests have little or no power to distinguish the endpoints, and non-rejection is not equivalence. Over the uniform-wrong sets specifically, the counts are $23/24$ versus $21/24$ on Pet and $18/28$ versus $19/28$ on CIFAR-10 (top-k versus reachable). The comparison covers all $4{,}700$ image–class cells and supports the rank-agreement mechanism, while O_τ remains a valid upper bound for the weighting family by definition.

N Sensitivity analyses and transfer on fixed loss tables

These analyses examine noise variability, loss ordering, input-dependent weighting, evaluation protocols, and the distinction between feasible decisions and fitted accuracy. The six-table follow-ups use the original SD 1.5 Pet and CIFAR-10 tables, panel C and its reference-protocol re-measurement F, and the two PixArt-α tables; they do not include the dense-grid endpoint E of Table 1.

Scope and conventions. Repeated-noise comparisons condition on their image panels; nested or pooled blocks are dependent. Optimisation brackets are not confidence intervals, and fixed-weight transfer counts are not target optima. These conventions apply throughout this appendix; experiment-specific exceptions and qualifications are stated locally.

The analyses below followed inspection of the loss tables and results in Table 29. The PixArt and PIL fresh-noise, dense-allocation and class-pruning re-measurements have separate preregistrations. The PIL preprocessing control freezes its protocol before new measurement but remains descriptive; the other follow-ups have no preregistered decision criterion. Unless explicitly marked exact, reported shared-weighting intervals are floating-point solver brackets with postchecked feasible incumbents, not exact certificates. All but the Gaussian illustration and fresh-image study use $n = 100$ images per table. For margin-based counts and brackets, cross-fitting, controls, conditional-family and dense-allocation analyses use the strict margin $\tau = 10^{-6}$ on per-image max-normalised contrasts. The reachable-set analysis instead uses the manuscript's raw-table LP cutoff 10^{-12}. Cross-fitting, controls, reachable-set and conditional-family analyses restrict PixArt-α to its six SD-matched timesteps; cross-fit incumbents use HiGHS for 120 s on one thread and control brackets use it for 60 s. The operating-point table retains each source table's reported τ.

N.1 Noise variability and decision transfer

Noise cross-fitting. We first ask whether weights selected on one noise block remain effective on another, without allowing the target block to influence the weights. Each table's R noise trials are split into interleaved halves (even and odd trial indices; $32/32$ for SD 1.5, $16/16$ for PixArt-α), each half is averaged and normalised with its own per-image maximum absolute contrast, and the per-image maximum-margin witnesses and a shared incumbent are selected on the source half

only. Neither is re-optimised on the target half. Table 16 counts the source-reachable images whose own witness is still strictly correct on the target half, and the strict count of the transferred shared incumbent on each half.

Table 16: **Source-half witnesses and incumbents evaluated on the other noise half.** Survivors are counted over source-reachable images; the shared column gives strict counts on the source and then on the target half for the same weighting. Both directions reuse the same data and are not independent replicates.

Table	Witnesses A→B	Witnesses B→A	Shared A→B	Shared B→A
SD 1.5 Pet, T=21	$93/99$	$90/99$	$83 \to 80$	$84 \to 82$
SD 1.5 CIFAR-10, T=21	$86/96$	$83/94$	$80 \to 73$	$78 \to 74$
SD 1.5 CIFAR-10 panel C, T=21	$90/98$	$90/99$	$86 \to 84$	$88 \to 85$
SD 1.5 panel C, reference ℓ_1	$97/99$	$97/98$	$90 \to 83$	$88 \to 86$
PixArt-α CIFAR-10, T=6	$60/80$	$64/79$	$53 \to 49$	$54 \to 48$
PixArt-α Pet, T=6	$59/76$	$54/78$	$49 \to 46$	$52 \to 41$

Witness survival is 88–99% on the four SD 1.5 follow-up tables (Table 16) and 69–81% on the two PixArt-α tables, and the target counts are unchanged at every fixed-weight threshold in $\{0, 10^{-12}, 10^{-8}, 10^{-6}, 10^{-4}\}$. This measures fixed-weight stability on the exposed noise halves; the per-image witnesses use labels. A preregistered PixArt-α CIFAR-10 re-measurement uses a fresh noise seed on the same 100 images, six timesteps and $R = 32$. With the count, denominator, τ and comparator fixed before measurement, 126 of the 159 source-half witnesses survive on the fresh halves (registered pooled Wilson interval 72.3–84.8%). The 159 witnesses come from 87 images: 72 contribute twice, 15 once, and 13 contribute no eligible witness. Because images contribute repeated witnesses with differing eligibility, the registered binomial interval has no established population coverage. Keeping both half-records and eligibility together, a descriptive bootstrap over all 100 images (100,000 resamples, seed 20260911) gives 71.6–86.2% as an image-cluster sensitivity check. The exposed comparator $124/159$ lies inside the registered interval; this inclusion alone does not establish equal survival probabilities. Per half the source-reachable witnesses are 80 and 79, of which 61 and 65 survive the fresh draw (60 and 64 on the exposed split half), and the frozen shared incumbents score 49 and 47 on the fresh halves against 53 and 54 on their source halves (exploratory). A replay of 360 forward passes of the original cache through the re-measurement path reproduced it exactly.

Descriptive trial-budget ladder on the original MSE tables. To assess sensitivity to noise averaging, we re-optimise both endpoints using the ordered prefixes $r = 0, \ldots, R-1$ of the two already exposed SD 1.5 caches, for $R \in \{1, 2, 4, 8, 16, 32, 64\}$, keeping the original 100 images and 21 timesteps in $[250, 750]$. Each prefix has its own per-image max-absolute contrast normalisation. These are nested prefixes, not the interleaved halves in Table 16. The primary cutoff is $\tau = 10^{-6}$, with $0, 10^{-8}, 10^{-4}$ sensitivity cutoffs at $R = 32, 64$; each of the 26 MILPs has a single-thread 1,800 s cap. All cells completed and an independent recount matched the saved pointwise, uniform and feasible-incumbent counts (Table 17).

The primary gap lower bounds are $9, 11, 10, 12, 13, 11, 11$ for Pet and $20, 17, 16, 16, 15, 11, 10$ for CIFAR-10. The Pet sequence is nonmonotone; the CIFAR-10 sequence declines with plateaus, alongside changes in O. The $R = 64$ uniform and pointwise counts match the original tables, and the new floating brackets contain the separately certified exact optima $86/81$. A weaker new incumbent does not revise those exact optima. The sequence describes the jointly changing noise budget and prefix-specific normalisation on these tables.

Fresh-noise re-optimisation on the original CIFAR-10 table. A follow-up specified before its new measurements retains the original 100 CIFAR-10 images, SD 1.5 checkpoint, tensor preprocessing, squared loss and 21 timesteps $\{250, 275, \ldots, 750\}$. It measures new global trials 64–127 under the original master seed 20260813; the original cache used trials 0–63. Before measurement, a replay reproduces 400 original class-conditioned loss scalars bitwise. The new block completes all 1,344,000 class-conditioned evaluations. Both new tables use the original $R64$ per-image

Table 17: **All** 26 **descriptive trial-budget cells.** Each dataset contributes 13 cells. A is the strict uniform count, O the pointwise LP count, and $[L, U]$ the floating shared-weight bracket with a strict incumbent recount. The gap column is $[O - U, O - L]$, not an exact certificate or a confidence interval. All closed MILPs returned optimal solver status; this does not establish a strict-count optimum. Sensitivities below the rule use the same prefixes as the primary rows.

		Pet				CIFAR-10			
R	τ	A	O	$[L, U]$	gap	A	O	$[L, U]$	gap
1	10^{-6}	77	100	$[83, 91]$	$[9, 17]$	63	99	$[72, 79]$	$[20, 27]$
2	10^{-6}	73	100	$[79, 89]$	$[11, 21]$	66	98	$[73, 81]$	$[17, 25]$
4	10^{-6}	79	100	$[84, 90]$	$[10, 16]$	68	97	$[74, 81]$	$[16, 23]$
8	10^{-6}	78	99	$[82, 87]$	$[12, 17]$	72	98	$[78, 82]$	$[16, 20]$
16	10^{-6}	79	99	$[81, 86]$	$[13, 18]$	71	96	$[77, 81]$	$[15, 19]$
32	10^{-6}	76	97	$[84, 86]$	$[11, 13]$	71	93	$[79, 82]$	$[11, 14]$
64	10^{-6}	76	97	$[82, 86]$	$[11, 15]$	72	91	$[77, 81]$	$[10, 14]$
32	0	76	97	$[84, 86]$	$[11, 13]$	71	93	$[79, 82]$	$[11, 14]$
32	10^{-8}	76	97	$[83, 86]$	$[11, 14]$	71	93	$[80, 82]$	$[11, 13]$
32	10^{-4}	76	97	$[84, 86]$	$[11, 13]$	71	93	$[79, 82]$	$[11, 14]$
64	0	76	97	$[84, 86]$	$[11, 13]$	72	91	$[77, 81]$	$[10, 14]$
64	10^{-8}	76	97	$[83, 86]$	$[11, 14]$	72	91	$[78, 81]$	$[10, 13]$
64	10^{-4}	75	97	$[83, 86]$	$[11, 14]$	72	91	$[76, 81]$	$[10, 15]$

maximum-absolute contrast scale and strict binary64 $\tau = 10^{-6}$. Thus normalisation is held fixed here, unlike the descriptive prefix ladder above.

Table 18: **Original-headline noise re-optimisation.** Counts are out of 100. The original row uses its previously certified optimum. For the fresh and pooled rows, pointwise counts and shared lower bounds have exact rational witnesses, while shared upper bounds are floating MILP bounds. Their gap intervals are optimisation brackets, not confidence intervals. Pooled $R128$ reuses both blocks and is not an independent replication.

Noise table	A	O_τ	$[L, U]$	Gap	Shared grade
Original $R64$	72	91	$[81, 81]$	10	exact
Fresh $R64$ (primary)	71	94	$[78, 81]$	$[13, 16]$	solver
Pooled $R128$	71	93	$[79, 80]$	$[13, 14]$	solver

Each new table is re-optimised once with the prespecified $1{,}800$ s MILP cap. Both solvers return optimal status for their closed floating formulations; exact strict recounts give the lower bounds in Table 18. The fresh block meets the primary positive-gap criterion, since $94 - 81 = 13 > 0$; the pooled table does not enter that rule. Independent checking reproduces the contrast hashes, every pointwise witness and both shared strict counts. The original-table, PIL and fresh-panel results were already known when this fixed-image noise follow-up was specified. The gap therefore persists on this fresh noise block of the original image panel.

N.2 Class ordering and loss spacing

Rank and coordinate-alignment controls. Table 19 recomputes the gap bracket $[O_\tau - U,\ O_\tau - L]$ on each table after replacing every loss row by its legacy rank-plus-range transform, by its canonical midranks with the range discarded, and after permuting the timestep coordinates independently per image under two fixed seeds. The canonical transform leaves a positive lower bound on all six tables, and on five of the six its lower bound exceeds the real-score upper bound. Rank structure alone therefore produces comparable or larger sharing gaps: the size of $O_\tau - G^\star_\tau$ is not evidence specific to the calibration of the diffusion loss. The imagewise permutations preserve pointwise reachability by an explicit witness map while changing the coordinates a single shared vector sees;

their brackets move heterogeneously and mostly overlap the real ones. With unique winners and no strict-margin boundary cases, preservation of the invariant fractions in Appendix M follows from row-order preservation.

Table 19: **Gap brackets $[O_\tau - U,\ O_\tau - L]$ under rank and permutation controls** (floating solver brackets, 60 s, one thread; not the exact closures of Table 29).

Table	real	rank+range	canonical	perm. 1	perm. 2
SD 1.5 Pet	$[11, 14]$	$[13, 17]$	$[12, 15]$	$[10, 13]$	$[10, 11]$
SD 1.5 CIFAR-10	$[10, 13]$	$[12, 17]$	$[14, 18]$	$[9, 13]$	$[8, 11]$
SD 1.5 CIFAR-10 panel C	$[8, 9]$	$[11, 15]$	$[16, 18]$	$[9, 12]$	$[9, 13]$
SD 1.5 panel C, ref. ℓ_1	$[6, 9]$	$[8, 12]$	$[14, 17]$	$[7, 12]$	$[8, 11]$
PixArt-α CIFAR-10	$[21, 21]$	$[24, 25]$	$[27, 30]$	$[26, 28]$	$[25, 25]$
PixArt-α Pet	$[26, 26]$	$[31, 33]$	$[32, 34]$	$[29, 30]$	$[28, 29]$

Reachable sets under the rank control. Recomputing each image's reachable class set with the manuscript's per-class linear program on the canonical-midrank tables shows what the control changes. The true class is never reachable only under the real table (zero of 600 images), although 23 other real-reachable image–class memberships are absent under the control. The invariant counts are identical on every table, and the control has 0.2–0.8 more reachable classes per image on average; the per-image Jaccard overlap with the real sets averages 0.87–0.93 across the six tables, with 43–78 of 100 sets identical. The uniform strict count falls under the control (for example $79 \to 55$ on the reference ℓ_1 table), since averaging midranks discards loss magnitudes. The legacy rank-plus-range transform gives exactly the real-scale sets of the canonical one. Thus the reachable sets are comparatively insensitive to these changes in loss spacing, whereas uniform-pooling accuracy is not; set preservation and accuracy preservation are distinct.

Preserving winners while permuting the remaining losses. On the two original six-timestep SD diagnostic tables ($n = 100$, $R = 64$), we keep every per-timestep winner, including ties, and permute the remaining class losses independently within each row. This label-blind transformation preserves the complete row multiset. We apply 200 fixed-seed transformations per table and recompute reachable sets, truth inclusion, and uniform strict accuracy (Table 20). Keeping the winners alone does not reproduce the real tables' thin reachable sets: the assignment of the nonwinning losses also matters. The invariant counts remain 30 on Pet and 15 on CIFAR-10 throughout, as expected from fixing the winner structure. This isolates the role of nonwinning losses in pointwise set size.

Table 20: Winner-preserving controls on the original $T = 6$ diagnostic tables. Reachability uses the raw-margin LP cutoff 10^{-12}; uniform strict accuracy uses a unique minimum of the mean loss. Control entries are means (min–max) over 200 transformations of the same exposed table, not independent panels or confidence intervals. Truth and accuracy entries are counts out of 100.

Dataset	Loss table	Mean reachable	Truth included	Uniform strict
Pet	Real	2.55	97	76
	Canonical midrank	2.74	97	74
	Winners preserved	7.84 (7.01–8.51)	94.7 (92–99)	69.4 (63–76)
CIFAR-10	Real	3.09	91	73
	Canonical midrank	3.50	90	60
	Winners preserved	4.70 (4.43–5.02)	93.0 (89–97)	71.1 (66–76)

N.3 Conditional weighting and compute allocation

A small label-blind conditional family. We test whether simple input dependence reduces the cost of sharing weights. We partition images into $K \in \{1, 2, 4\}$ nested clusters by deterministic

two-means bisection of their per-timestep runner-up-margin profiles, without using labels. Each cluster receives its own shared weighting. Summing per-cluster solver upper bounds (120 s per cluster) bounds the attainable strict count. For noise transfer, both clustering and weights are fitted on the source half and applied unchanged to the target half. Table 21 reports every cluster count and direction.

Table 21: Conditional weighting on six fixed tables, with counts out of 100. Triples list $K = 1, 2, 4$ in that order. O_τ is the full-table pointwise ceiling; the upper bounds concern the full-table clustered family and remain solver-grade. Transfer columns give target strict counts with source-fitted clusters and weights; the two directions reuse the same images and noise.

Table	O_τ	Full-table upper $K = 1, 2, 4$	Even $\to$ odd $K = 1, 2, 4$	Odd $\to$ even $K = 1, 2, 4$
SD Pet	97	86/86/86	78/78/78	81/82/81
SD CIFAR-10	91	81/81/83	72/67/63	73/74/73
SD panel C	97	89/89/91	85/81/83	86/81/80
SD reference ℓ_1	98	92/92/92	82/80/83	85/79/80
PixArt-α CIFAR-10	78	57/57/58	50/47/49	49/47/46
PixArt-α Pet	78	52/52/54	45/44/39	40/39/38

The $K = 4$ upper bounds are only 0–2 counts above those for $K = 1$ and remain below the pointwise ceilings. The change in upper bounds does not determine the change in optimal strict counts. In transfer, $K = 4$ gives no higher count in 11 of 12 directions; the exception is $82 \to 83$ on the reference ℓ_1 even-to-odd comparison. This clustering rule therefore provides no consistent noise-transfer improvement despite enlarging the weighting family.

Dense timestep allocation on panel C. A preregistered re-measurement of the same 100 CIFAR-10 images evaluates a 500-point grid over $[2, 999]$ with $R = 2$ draws per point, i.e. 1,000 class-conditioned evaluations per class against $21 \times 64 = 1{,}344$ for the reference $T = 21$ table, so density, repeat allocation and total budget change together and no single factor is isolated.[1] Under the primary ℓ_1 reduction, fixed before the run, full-grid uniform averaging followed by loss argmin gives $78/100$ correct (Wilson 95% interval 68.9–85.0, the registration's single nominal interval) against $79/100$ on the reference table; the exploratory ℓ_2 reduction gives $60/100$. Under this joint protocol change, uniform accuracy is no closer to the published operating point. The comparison does not isolate the effect of grid density.

To examine shared-weight limits at this finer resolution, we average the 500 points into $B = 21$ bins. This defines a weighting family with one weight per bin ($w_t = v_b/|b|$); within it the pointwise ceiling is 98 under both registered bin rules and the shared bracket is $[87, 93]$ (equal-width bins) and $[88, 92]$ (bins nearest the reference grid, whose unequal sizes give a binned uniform count of 80), giving respective gap brackets $[5, 11]$ and $[6, 10]$, both meeting the registered persistence threshold $O_\tau - U \geq 5$; the ℓ_2 reduction gives $[5, 7]$ under both rules. These brackets apply only to the binned family, not the 500-dimensional free family. Uniform bin masses $v_b = 1/21$ differ from full-grid uniform weighting, which requires $v_b = |b|/500$, and changing the family can change both endpoints O_τ and $G^\star_\tau$, so the gap itself has no corresponding monotonicity guarantee.

The registered prefix-pruning emulator obtained $27/100$ under primary ℓ_1 and $25/100$ under secondary squared ℓ_2. Its first stage sampled only the lowest 5% of the grid, so these counts do not evaluate a published adaptive sampler and carry no shared-gap claim.

Budget-matched two-stage class pruning. To test whether reallocating compute from unlikely classes improves accuracy, a separate preregistered comparison on panel C fixes the SD 1.5 scorer, ℓ_1 loss and 13,440 class-conditioned evaluations per image. Arm R averages 1,344 draws for each of all ten classes, with timesteps sampled uniformly from $\{0, \ldots, 999\}$. Arm A first evaluates 224 draws for each class, retains the five classes with lowest mean loss, and spends another 2,240 draws on each retained class; its final loss averages both stages. Overlapping draw indices use

[1] Forty-nine of the 500 grid points lie outside the $[50, 950]$ range of the $T = 21$ reference grid.

common timesteps and noise, with the seed and hashed tie-breaking fixed before measurement. Both arms give $79/100$ argmin-correct decisions, equal to the reference-grid comparator P1. The registered primary is Arm A's accuracy (Wilson 95% interval 70.02–85.83%, the single nominal interval). Relative to R, A corrects one image and loses one; relative to P1, R corrects four and loses four. These paired contrasts are reported as counts and point differences only, without intervals or equivalence claims. The original-path replay reproduces 400 class-conditioned evaluations exactly before each arm. Both measurements complete within their fixed caps. R versus P1 changes timestep sampling and noise; A versus R also changes the retained class set. The comparison evaluates these custom procedures jointly. Image-dependent class removal lies outside the fixed-table weighting family, so it supplies no bounds on O_τ or $G^\star_\tau$.

N.4 Preprocessing, fresh noise, and new images

PIL versus tensor preprocessing. To assess sensitivity to preprocessing, a protocol fixed before measurement changes only the image preprocessing path on the same exposed 100-image CIFAR-10 panel C: the original tensor bicubic resize is replaced by PIL bicubic resize, centre crop, RGB conversion, tensor conversion and normalisation, following the upstream `get_transform` implementation of Li et al. (2023) (commit `e9f772d`). The SD 1.5 checkpoint, prompts, grid $\{50, 95, \ldots, 950\}$, $R = 64$, image-indexed noise seeds, fp16 inference and fp32 residual accumulation are fixed. Before new measurement, an original-path replay reproduces 400 class-conditioned evaluations exactly. The new arm completes all $1{,}344{,}000$ evaluations. This contrast includes PIL quantisation, clipping and resize-backend differences; it is not a pure operation-order intervention or a replication of the full published pipeline.

Table 22: Paired preprocessing comparison on 100 panel-C images. Only the PIL ℓ_1 row has an exact shared optimum; the other rows retain solver-grade bounds.

Loss	Preprocessing	A	O_τ	$[L, U]$	Gap
ℓ_1	Tensor	79	98	$[87, 92]$	$[6, 11]$
ℓ_1	PIL	89	100	$[97, 97]$	3 (exact)
Squared ℓ_2	Tensor	65	99	$[90, 93]$	$[6, 9]$
Squared ℓ_2	PIL	81	100	$[90, 95]$	$[5, 10]$

Each endpoint and loss in Table 22 uses its own full-table per-image maximum absolute contrast scale and $\tau = 10^{-6}$. Strict and argmin uniform counts agree. PIL corrects ten previously incorrect images under ℓ_1 and sixteen under squared ℓ_2, losing none under either loss. MILPs use a $1{,}800\,\mathrm{s}$ cap on one thread. The PIL ℓ_1 row additionally has exact certification on the same dyadic table and binary64 τ: all 100 pointwise strict witnesses verify, a shared rational vector verifies 97 strict decisions, and a SCIP/VIPR certificate proves the closed 98-correct problem infeasible. The shared witness is generated at 2×10^{-6} but accepted only at the original 10^{-6} threshold. Independent rational recounts, a full LP–proof input comparison and a proof-checker rerun verify the result. Thus $O_\tau = 100$, $G^\star_\tau = 97$, and sharing costs exactly three images at $89/100$ observed accuracy. The tensor ℓ_1 row retains a solver-grade gap lower bound of six; the same-panel comparison places the PIL gap below that bound. The other rows remain solver-grade. The 89% accuracy on this exposed panel is not directly comparable to the published 85.5% benchmark result.

Fresh-noise re-optimisation at the PIL operating point. The remaining noise question is whether a shared gap persists when both endpoints are re-optimised on new draws, rather than transferring existing witnesses. A follow-up specified before its new measurements keeps the same 100 panel-C images, PIL preprocessing, SD 1.5 checkpoint, prompts, ℓ_1 loss and 21-point grid $\{50, 95, \ldots, 950\}$. It measures 192 additional draws per timestep, using global trial indices 64–255 under the original image/timestep/trial keyed seed rule (master seed 20260901). The fresh blocks A (64–159) and B (160–255) each have $R = 96$ and share no draw index with each other or the original $R = 64$ table. They use disjoint keyed pseudorandom draws on the same images. Original tensor and PIL replays reproduce 400 class-conditioned evaluations each exactly before all $4{,}032{,}000$ new evaluations complete within the fixed measurement cap.

Every new table uses the *original PIL* $R = 64$ per-image maximum absolute contrast scale and binary64 $\tau = 10^{-6}$; scales are not refitted on the fresh blocks. The prespecified analysis re-optimises O_τ and the shared MILP separately on A, B and the pooled table. Its single primary requires a strictly positive gap lower bound in *both* fresh blocks, with no five-image effect threshold. The exact gaps of two and three images satisfy this rule; pooled $R = 256$, which combines the original 64 and fresh 192 draws, is secondary.

Table 23: **Exact sharing gaps after fresh-noise re-optimisation on the same PIL panel.** All counts are out of 100 at the fixed source-$R64$ scale and strict margin $\tau = 10^{-6}$. Both endpoints are certified on each normalised dyadic table. The original row is a reference; only the two fresh rows enter the primary comparison. Pooled reuses all three draw sets and is not a third independent replication or a new image panel.

Noise table	R	Uniform A	O_τ	$G^\star_\tau$	Gap
Original PIL	64	89	100	97	3
Fresh A	96	90	99	97	2
Fresh B	96	89	100	97	3
Original + A + B	256	90	99	96	3

All three new gaps have exact certificates (Table 23). On the same dyadic contrasts and original τ, 298 pointwise primal and two dual witnesses certify $O_\tau = 99/100/99$. Shared rational witnesses certify $97/97/96$ strict decisions, and SCIP/VIPR proofs establish infeasibility of the respective closed $K = 98/98/97$ problems. Candidates are generated at 2×10^{-6} and accepted only at the original 10^{-6} margin. An independent implementation verifies every pointwise and shared witness, the full correspondence between each proof's root model and the intended dyadic constraints, and a rerun of the pinned proof checker. Appendix Q describes the released evidence and reproduction scope.

The high-accuracy PIL separation survives fresh-noise re-optimisation, while the conflicting images can change: panel position 0 is pointwise unreachable in A and pooled, although it belongs to the original PIL three-image conflict.

PIL measurements on a fresh 500-image panel. A protocol specified before measurement uses 500 analysis-unused CIFAR-10 train-split images, fifty per class, disjoint from the previous analysis panels and the reserved reference-protocol panel. The checkpoint, prompts, PIL preprocessing, ℓ_1 loss and grid $\{50, 95, \ldots, 950\}$ are fixed, with $R = 64$ and master seed 20260913. This is a new finite-table panel, not a benchmark-test replication or a claim that these images were absent from model pretraining. Each image uses its full-$R64$ maximum absolute contrast scale and strict binary64 $\tau = 10^{-6}$.

Uniform strict and argmin counts both equal $439/500$ (87.8%). Exact rational checks verify 497 pointwise primal witnesses and three dual witnesses, closing $O_\tau = 497$. A rational shared witness attains 462 strict successes and the floating MILP upper bound is 462, giving a solver-grade gap bracket $[35, 35]$ (7.0 percentage points). The coincident bounds do not establish an exact optimum. This meets the prespecified primary criterion of a strictly positive gap lower bound. The lower witness uses the post-hoc subset-margin refinement specified below.

All six solvers return optimal status for the *closed* MILP relaxation; strict postchecks remain separate. The full solve takes 3598 s. The combined upper is $\min(497, 462, 94{+}96{+}96{+}98{+}94) = 462$, and the largest full-panel rational recount among these six MILP incumbents is 458.

A post-hoc refinement, specified after inspecting the original result and before refitting, fixes the 462 images with largest exact margins under its saved incumbent (ties by image row) and solves one max-min-margin LP with a 600 s cap, using the same contrasts, full timestep space and τ. This is a reconstructed candidate set; the original binary assignment was not saved. The resulting rational simplex weighting has minimum margin exceeding 10^{-3} on the selected set and exactly 462 strict successes on the full panel. An independent rational recount certifies the improved global lower bound; the floating upper bound remains $U = 462$.

Table 24: **Shared-weight analysis on the fresh PIL 500-image panel.** All six planned MILPs and the post-hoc subset LP are reported. Each lower bound is an exact rational strict recount; upper bounds are floating MILP bounds. The LP supplies only a lower bound. Block rows use fixed, class-balanced groups of 100 images and permit separate weights. Their lower bounds must not be summed as one global policy. "Full recount" evaluates each returned vector on all 500 images.

Analysis	Images	Local L	Local U	Full recount	Cap (s)
Full panel	500	458	462	458	3600
Block 1	100	90	94	442	1200
Block 2	100	91	96	441	1200
Block 3	100	94	96	424	1200
Block 4	100	95	98	439	1200
Block 5	100	90	94	427	1200
Subset LP	462	462	–	462	600

Exact infeasibility of the closed $K = 463$ problem remains unproved after a six-hour SCIP attempt with a 512 MB certificate-output cap. The upper bound and hence the gap bracket remain solver-grade.

A weighting fixed on the original PIL $R64$ table before measurement also gives $439/500$, with sixteen policy-only and sixteen uniform-only correct images. Its paired improvement is 0.0 points; a descriptive class-stratified image bootstrap (10,000 resamples of fixed predictions) gives the percentile interval $[-2.2, 2.2]$ points. Hence the diagnostic shared gap persists on this new panel, while the frozen policy shows no accuracy improvement.

Partial ℓ_1 summaries were exposed during technical diagnosis before the registered analysis completed; the measured data, analysis code, six budgets and primary rule were unchanged.

N.5 Incompatible decisions and the weighting domain

Real-table incompatibility witnesses. To make the shared gap concrete, we identify small groups whose members are individually reachable but cannot all be classified correctly by one weighting. For each individually reachable image define the convex set $S_i = \{w \in \Delta_T : D_{ij}^\top w > \tau \; \forall j\}$. We search the complete timestep space for conflicting collections with empty intersection, then verify both a joint dual upper bound and a strict feasible witness after every single-image removal. Such a collection is inclusion-minimal, not necessarily smallest in cardinality; it is an image-level analogue of an irreducible infeasible subsystem (Gleeson & Ryan, 1990). If q verified collections are disjoint and all their images are counted in O_τ, every shared weighting must fail on at least one image per collection, giving $G_\tau^\star \le O_\tau - q$.

A deterministic greedy search finds $q = 8/6/6/5$ on the four exact audited SD tables (Pet, CIFAR-10, C and dense E), explaining lower bounds of that size within their exact gaps $11/10/8/7$. The PIL ℓ_1 table gives two disjoint conflicts of sizes six and three, accounting for a lower bound of two within its exact gap of three. These searches are descriptive; the reported dual and deletion witnesses are checked with exact rational arithmetic on the same normalised dyadic tables and τ as their parent certificates (Figure 7).

Allowing signed weights on two fixed tables. A post-exposure descriptive control expands the simplex to $\mathcal{W}_\pm = \{w : \|w\|_1 \le 1\}$ on the original CIFAR-10 MSE and PIL ℓ_1 tables, each with $n = 100$, $T = 21$ and $R = 64$. It keeps each table's original per-image normalisation and binary64 $\tau = 10^{-6}$. Writing $w = u - v$, $u, v \ge 0$ and $\sum_t (u_t + v_t) \le 1$ gives pointwise LPs and a shared MILP with $M_i = \tau + \max_{t,j} |D_{itj}|$. The floating implementation rounds M_i upward and $\tau - M_i$ downward, so it relaxes the closed constraints; every lower bound is recounted at the original strict threshold using rational arithmetic. The original exact simplex witness remains a feasible lower-bound candidate (Table 25).

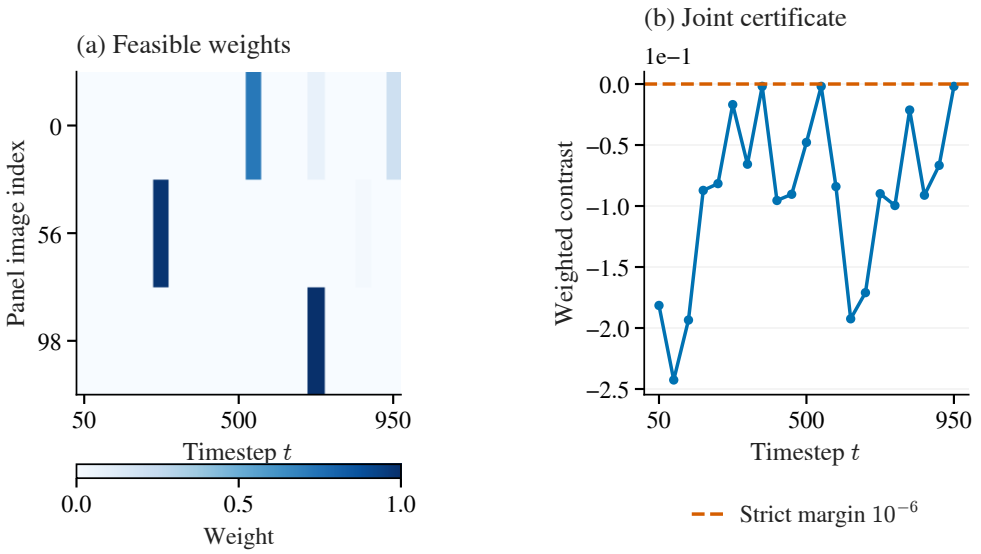

Figure 7: **Three individually reachable images have incompatible shared requirements.** Original PIL ℓ_1 table ($R = 64$), panel indices $0, 56, 98$. (a) Each row is a feasible per-image weighting. (b) Nonnegative dual weights p_{ij} sum to one over the selected images and rival constraints. All 21 coefficients $a_t = \sum_{i,j} p_{ij} D_{itj}$ lie below τ; their exact maximum is approximately -0.001918. Thus $\sum_t a_t w_t \le \max_t a_t < \tau$ for every simplex weighting, whereas satisfying every selected strict constraint would make this weighted average exceed τ. The contradiction applies in the full 21-timestep space.

Table 25: Signed-weight scope control on two exposed tables. All counts are out of 100. Simplex O and $G^\star$ are exact; signed $O = 100$ is certified by per-image rational witnesses. Signed shared bounds have exact feasible lower bounds and floating upper bounds, so both signed gap brackets remain solver-grade. This is a descriptive comparison with no prespecified decision criterion or likelihood-weighting interpretation.

Table	Uniform	Simplex O	Simplex $G^\star$	Signed O	Signed $[L, U]$	Signed gap
Original CIFAR-10 MSE	72	91	81	100	$[81, 92]$	$[8, 19]$
PIL CIFAR-10 ℓ_1	89	100	97	100	$[97, 100]$	$[0, 3]$

The MSE shared solve reaches its 1,800 s cap. The PIL solve reports an optimal closed-relaxation value of 100; that value does not certify 100 strict successes. The independently checked lower bounds use the original simplex witnesses (81/97), both with zero negative mass and unit ℓ_1 norm. The signed gap is positive on MSE and unresolved on PIL. Expanding the family cannot decrease O or $G^\star$, but their difference need not be monotone. These bounds concern the unit ℓ_1 ball; the signed gradient fit below uses unrestricted weights.

N.6 Surrogate fitting and image transfer

Cross-entropy fitting versus attainable strict counts. We compare a fitted weighting with the best strict count attainable in its family, to distinguish the fitted rule's shortfall from the cost of sharing. The fit uses softmax cross-entropy over cached losses and full-batch Adam for 2,000 steps, with no early stopping; it does not reproduce a published method. The simplex fit reaches 66/66/72/75 strict correct in sample on the four SD 1.5 follow-up tables when constrained to the simplex, below uniform (76/72/78/79) and 12–20 below the certified or solver-grade shared optimum (86/81/89/[87, 92]); the signed variant, outside the family the certificate covers, reaches 76/72/84/88. On PixArt-α, shared brackets recomputed at this control's $\tau = 10^{-6}$ are $[54, 57]$

(CIFAR-10) and $[50, 52]$ (Pet) against $39/51$ and $44/44$ (simplex/signed); Table 1 reports the registered $\tau = 10^{-12}$ brackets $[55, 57]$ and $[48, 52]$. Cross-entropy is a surrogate for the strict count, so this shortfall mixes surrogate mismatch with finite-step optimisation and is not an optimiser gap. Its magnitude also depends on loss normalisation, as the scale control below shows.

Fitted on fifty images and scored on the other fifty, no weighting exceeds uniform by more than six images in any direction (four on the SD 1.5 tables), and the in-half MILP incumbent's 1–6 image advantage over uniform does not consistently carry over (held-out difference -2 to $+3$). On the four CIFAR-10 tables, this even/odd image split separates ground-truth classes $(0, 2, 4, 6, 8$ versus $1, 3, 5, 7, 9)$; the evaluation classes occur as negative candidates but not as positive labels during fitting. Both Pet halves contain all 37 classes.

For reproducibility, this control uses a float64 NumPy implementation of Adam with learning rate 0.05, $(\beta_1, \beta_2) = (0.9, 0.999)$, $\epsilon = 10^{-8}$, zero initial moments, and bias correction. The simplex arm sets $w = \mathrm{softmax}(\theta)$ with $\theta = 0$ initially; the signed arm directly optimises $w \in \mathbb{R}^T$, initialised at $1/T$. Each update uses every fit image and candidate class, with no regularisation, minibatching, or learning-rate schedule. Before fitting, losses are averaged over noise draws and divided by their standard deviation across all fit-image, timestep, and class entries (population convention, `ddof=0`); the scale is frozen for evaluation. The objective is mean cross-entropy with logits $-\sum_t w_t L_{itc}/s$. The final $2{,}000$-step iterate is evaluated without checkpoint selection. The original transfer control uses even/odd manifest positions in both directions; the separate class-stratified scale control below uses seed 20260909 and reports every scale.

Balanced-split transfer of a shared incumbent. Using exactly the class-stratified halves below, with five images per CIFAR-10 class on each side, we fit a shared MILP incumbent on one half (120 s, one thread) and evaluate the unchanged vector on the other half. For the three SD $T = 21$ tables (original CIFAR-10, fresh C and reference F) and the PixArt CIFAR-10 $T = 6$ table, the held-out correct counts are respectively $32/38$, $43/41$, $43/39$ and $21/27$, against uniform $35/37$, $41/37$, $40/39$ and $18/28$ (two directions, each out of 50). The differences range from -3 to $+4$ images. Each direction evaluates a different fitted vector; their counts cannot be pooled into one shared-vector result. Even with class-balanced halves, these fitted shared weights do not consistently outperform uniform weighting on the held-out images.

Scale sensitivity of the surrogate fit. A separate descriptive control fixes its design and code before refitting the exposed caches. Write $X_{itc} = L_{itc} - C^{-1}\sum_j L_{itj}$ and compute both $s_0 = \mathrm{std}(L)$ and $s_c = \mathrm{std}(X)$ on the fit images only. For any fixed scale s, centring preserves cross-entropy and class decisions; changing s changes the objective with logits $-\sum_t w_t X_{itc}/s$. We compare s_0 with s_c and report all three additional sensitivity scales $0.1s_c, 10s_c, 100s_c$, without selecting a scale by evaluation accuracy. Each column uses the same convex simplex solver (sequential least-squares programming, SLSQP; uniform initialisation, analytic gradient, tolerance 10^{-12}, at most $2{,}000$ iterations). Table 26 reproduces the original count at s_0 but shows substantial scale sensitivity. In particular, the $0.1s_c$ arm leaves SD deficits of $8/9/3/[1, 6]$ images, whereas the designated s_c comparison remains below uniform on all four SD 1.5 follow-up tables. Thus the original 12–20 image shortfall does not characterise CE fitting in general. With the original $2{,}000$-step Adam update and only the scale changed to s_c, the six counts are $67/71/76/77/48/43$ in table order, versus $66/66/72/75/39/44$ at s_0.

For a separate image-transfer check we fix a class-stratified $50/50$ split (seed 20260909): each CIFAR-10 half has five images per class, and each Pet half contains all 37 classes. Scale estimation and fitting use only the fit half. Table 27 reports both directions for every fixed scale. The two directions fit different vectors, so their summed count is not the performance of one shared vector to compare with $G^\star_\tau$. No temperature is selected by evaluation accuracy. Even the $0.1s_c$ setting, which improves all six in-sample counts relative to s_0, does not outperform uniform in every held-out direction. Thus improved surrogate fitting on the source table does not ensure an accuracy gain on other images.

For every saved fit we independently recompute the objective, gradient and strict count. Convexity bounds CE objective suboptimality on the simplex by $\nabla f(w)^\top w - \min_t \nabla f(w)_t$. Across the 90 SLSQP fits this expression is at most 8.13×10^{-7} in float64; 14 exceed the declared 10^{-7} diagnostic threshold and remain included in the results. These are numerical objective diagnostics, not

Table 26: CE scale control, strict correct out of 100 on the fit table. All scales were fixed before refitting; s_c is the designated comparison and the other multiples are sensitivity arms, not a selected-temperature method. Shared optima use this control's $\tau = 10^{-6}$, so the PixArt rows differ from the $\tau = 10^{-12}$ brackets of Table 1; brackets retain their solver-grade status.

Table	Uniform	s_0	$0.1s_c$	s_c	$10s_c$	$100s_c$	$G^\star_\tau$
SD Pet	76	66	78	67	66	66	86
SD CIFAR-10	72	66	72	71	66	66	81
SD fresh C	78	72	86	76	72	72	89
SD reference F	79	75	86	78	75	75	$[87, 92]$
PixArt CIFAR-10	46	39	51	48	39	39	$[54, 57]$
PixArt Pet	44	44	46	43	44	44	$[50, 52]$

Table 27: Class-stratified image transfer under the CE scale control. Each entry is strict correct out of 50, with fit direction $a \to b$ / $b \to a$. All scales are calculated on the corresponding fit half; the complete set of sensitivity arms is reported.

Table	Uniform	s_0	$0.1s_c$	s_c	$10s_c$	$100s_c$
SD Pet	35/41	34/34	37/39	34/33	34/34	34/34
SD CIFAR-10	35/37	29/33	33/40	30/39	29/33	29/33
SD fresh C	41/37	32/38	43/41	35/37	32/38	32/38
SD reference F	40/39	37/34	43/40	38/38	37/34	37/34
PixArt CIFAR-10	18/28	18/26	24/28	19/26	18/26	18/26
PixArt Pet	22/22	22/22	21/24	21/22	22/22	22/22

exact certificates or bounds on strict-count suboptimality. The control changes neither the weighting family nor its previously reported O_τ and $G^\star_\tau$.

N.7 Interpreting gap and risk magnitudes

Accuracy–gap operating points. Table 28 places the eight tables in the original audited gap map on the two normalisations of the shared gap that the structural bound $O_\tau - G^\star_\tau \le 100 - A$ suggests, where A is the uniform strict count. All counts are reported elsewhere in this paper; this comparison adds only the two ratios. The points jointly vary model, grid (T, R), τ, loss reduction and image panel. The ℓ_2 and reference ℓ_1 rows use the same cache; exact rows are marked. Within SD 1.5 the accuracy range 65–80 does not order the gap monotonically; a lower uniform accuracy alone does not mechanically produce a large shared gap. The normalisations place the families differently. Later PIL panels appear in Appendix N.4, including the fresh 500-image panel in Table 24; the R-ladder reuses existing panels (Appendix N.1).

Gaussian effect-size illustration. To distinguish metric distance from classification error, we vary mean separation while holding the Bayes-preserving span fixed. On the noncommuting instance of Appendix O with $\Sigma_p(s) = I + s(AA^\top - I)$, $s \in \{0, 0.25, 1\}$, whitened means $r\{0, e_1, e_2\}$, $r \in \{0.5, 1.5, 3\}$, six noise scales, two seeds, $n = 90$ per table and nested $R \in \{32, 64, 256\}$ tables built from analytic expected losses with the same noise coupling, the six-scale distance at $s = 1$ is $\delta_{T,\star} \approx 1.718$; the separately certified continuum statement is the lower bound $\delta_\star \ge 12\sqrt{2}/11 \approx 1.543$, not equality. Scaling the means preserves the Bayes-preserving span and $\delta_{T,\star}$ but changes both the finite-table gaps and the excess risk of the distance-minimising fixed-grid candidate, estimated on 30,000 independent samples with paired Monte Carlo error as 1.20–1.40, 3.42–3.80 and 2.52–2.77 percentage points at $r = 0.5, 1.5, 3$ (ranges over seeds). These are candidate risks, not the family-optimal excess risk. Expected-loss sharing gaps stay positive on these instances; the isotropic $s = 0$ series is retained as a control. On the same instances a simplex weighting chosen by directly minimising the 0–1 risk on a separate 30,000-sample fit set (a finite deterministic search, not a proven optimum) has the same observed test excess risk as the distance minimiser at $s = 1$ in four of

Table 28: **Shared gap against uniform accuracy on the original eight-table gap map**, with the gap divided by the structural ceiling $100 - A$ and by the total headroom $O_\tau - A$. Brackets follow $G^\star_\tau \in [L, U]$.

Table	A	O_τ	$G^\star_\tau$	$O_\tau - G^\star_\tau$	$/(100 - A)$	$/(O_\tau - A)$
SD 1.5 Pet, T=21 (exact)	76	97	86	11	0.46	0.52
SD 1.5 CIFAR-10, T=21 (exact)	72	91	81	10	0.36	0.53
SD 1.5 panel C, T=21 (exact)	78	97	89	8	0.36	0.42
SD 1.5 panel E, T=101 (exact)	80	99	92	7	0.35	0.37
SD 1.5 panel C, reference ℓ_1	79	98	$[87, 92]$	$[6, 11]$	$[0.29, 0.52]$	$[0.32, 0.58]$
SD 1.5 panel C, same cache ℓ_2	65	99	$[90, 93]$	$[6, 9]$	$[0.17, 0.26]$	$[0.18, 0.26]$
PixArt-α CIFAR-10, T=6	46	78	$[55, 57]$	$[21, 23]$	$[0.39, 0.43]$	$[0.66, 0.72]$
PixArt-α Pet, T=6	44	78	$[48, 52]$	$[26, 30]$	$[0.46, 0.54]$	$[0.76, 0.88]$

six instances and an observed difference smaller than each candidate's paired-to-Bayes Monte Carlo standard error in the other two, and improves on it by 0.03–0.33 percentage points at $s = 0.25$; this finite search finds no lower observed test excess risk at $s = 1$. At fixed covariances and noise scales, mean separation changes the evaluated candidates' excess risks and finite-table gaps without changing the metric distance on the fixed grid.

O THE EXACT DISTANCE AND ITS DUAL CERTIFICATE

Retain the equal-prior shared-covariance Gaussian model, fixed distinct whitened means, $\Sigma_p, \Sigma_q \succ 0$, and nonnegative finite weightings. For $\sigma > 0$ set $G_\sigma := \Sigma_p^{1/2} Q \operatorname{diag}(v(\sigma)) Q^\top \Sigma_p^{1/2}$, let $A_\sigma = dG_\sigma / \operatorname{tr} G_\sigma$ and define the closed normalised reachable hull and target slice

$$\mathcal{R}_d = \operatorname{conv} \overline{\{A_\sigma : \sigma > 0\}}, \qquad \mathcal{E}_d = \{B \in \mathcal{E} : \operatorname{tr} B = d\}.$$

Theorem 5 (Exact distance and dual certificate)**.** *With $\Pi_{\mathfrak{E}^\perp}$ the Frobenius projection onto $\mathfrak{E}^\perp$,*

$$\begin{aligned} \delta_\star &:= \min_{B \in \mathcal{R}_d} \operatorname{dist}_F(B, \mathcal{E}_d) = \min_{B \in \mathcal{R}_d} \|\Pi_{\mathfrak{E}^\perp} B\|_F \\ &= \max_{\substack{Y \in \mathfrak{E}^\perp \\ \|Y\|_F \le 1}} \min_{\sigma \in [0,\infty]} \langle Y, A_\sigma \rangle. \end{aligned} \tag{13}$$

Hence $\delta_\star > 0$ exactly when the closed reachable hull misses $\mathcal{E}_d$, and under the stated Gaussian model $\delta_\star > 0$ implies a strictly positive infimum of excess 0–1 *risk over $\mathcal{R}_d$.*

Proof outline. The projection $\Pi_{\mathfrak{E}} = I - \Pi_{\mathfrak{E}^\perp}$ preserves positive definiteness and trace, giving the nearest point in $\mathcal{E}_d$, and convex separation of the projected hull gives the last equality. The complete argument follows the exact controls below.

Any positive dual value certifies the separation simultaneously at every noise scale. The risk infimum ranges over arbitrary finite-support weightings, not one prescribed grid, and the theorem does not calibrate it as a numerical function of $\delta_\star$. For a prescribed grid, $\mathcal{R}_{T,d} := \operatorname{conv}\{A_{\sigma_t} : 1 \le t \le T\} \subseteq \mathcal{R}_d$ is the trace-normalised hull of the metrics reachable on that grid, so its distance $\delta_{T,\star}$ to $\mathcal{E}_d$ satisfies $\delta_{T,\star} \ge \delta_\star$.

Under the same model, excess population 0–1 risk on the compact trace-normalised hull is bounded above and below by constants times squared distance to the preserving set; the hull-optimal excess risk is of order $\delta_\star^2$ with unevaluated instance-dependent constants (Theorem 6, Corollary 7).

These plotted instances are exact certificate regressions, not demonstrations of the genericity condition in Corollary 2. All three controls, and the ϵ-path below, take the full-span single-block target $\mathfrak{E} = \operatorname{span}\{I\}$ ($k = d$, $m = 1$, requiring $C \ge d + 1$), so $\mathcal{E}_d = \{I\}$ and $\delta_\star$ is the distance from the normalised hull to I; another mean configuration would give other distances for the same covariances. The covariances are $\Sigma_q = I$, $\Sigma_p = \operatorname{diag}(1, 2, 4)$ ($d = 3$) for the linear-space obstruction, $\Sigma_q = \operatorname{diag}(1, 2)$, $\Sigma_p = \operatorname{diag}(4, 1)$ ($d = 2$) for the cone-only control, and $\Sigma_p = \Sigma_q = I$

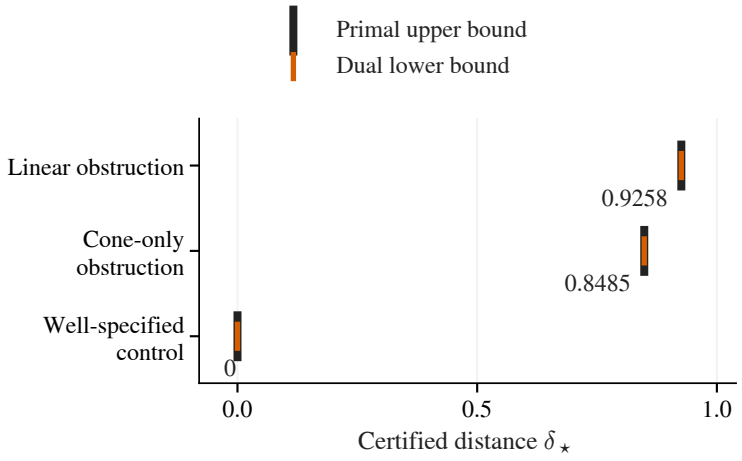

Figure 8: **Exact primal–dual distance controls.** Primal mixtures and continuum dual witnesses coincide for a linear-space obstruction ($\delta_\star = \sqrt{42}/7$), a cone-only obstruction ($\delta_\star = 3\sqrt{2}/5$, whose primal uses the compactification endpoint $\sigma = \infty$), and a reachable zero control. All inequalities are certified by exact rational polynomial algebra; these are deterministic theorem regressions, not UNet measurements.

($d = 2$) for the reachable control. The linear-obstruction and reachable controls use $\Sigma_q = I$, so their trace-normalised reachable hull collapses to one point; the former is obstructed by the direct span computation $L_{\mathrm{eff}} \cap \mathfrak{E} = \{0\}$ for the effective spectral span $L_{\mathrm{eff}} = \mathrm{span}\{\Sigma_p\}$ of Appendix R.3, not by the sign of Δ alone; the cone-only control exercises a nontrivial Stieltjes interval but lies in the planar $\Delta = 0$ exceptional case. Thus the figure itself contains no simple-spectrum, noncommuting generic-regime instance.

In that cone-only control, $M = \Sigma_p^{-1} = \mathrm{diag}(1/4, 1)$ gives the SPD metric $B = \Sigma_p^{1/2} M \Sigma_p^{1/2} = I \in L \cap \mathcal{E}$: the spectral linear span admits a Bayes-preserving solution. However, its required spectral coefficient ratio $n_1/n_2 = 1/4$ lies outside the nonnegative cone's interval $(1, 4)$, since each positive-scale atom has ratio $((\sigma^2 + 2)/(\sigma^2 + 1))^2$ and every nonzero nonnegative mixture has a ratio in the same interval. Equivalently, the whitened atom ratio is in $(4, 16)$ whereas I requires one. The closest compactified atom is $\mathrm{diag}(8/5, 2/5)$, recovering the certified $\delta_\star = 3\sqrt{2}/5$. This isolates the same conic restriction as Proposition 4, with a different mean configuration and without an angle-measure claim.

A nondegenerate exact demo. A separate $d = 3$, $k = 2$, $m = 1$ construction fills that scope gap without claiming an exact value of $\delta_\star$. For whitened means $0, e_1, e_2$, set

$$A = \begin{pmatrix} 3 & 3 & 3 \\ 1 & 0 & -1 \\ 0 & 1 & -1 \end{pmatrix}, \qquad \Sigma_p = AA^\top, \quad Q = \Sigma_p^{-1/2} A, \quad \Sigma_q = Q\,\mathrm{diag}(1, 2, 3) Q^\top.$$

Here Q is orthogonal, Σ_q has simple spectrum, and $[\Sigma_p, \Sigma_q] \neq 0$ because $[A^\top A, \mathrm{diag}(1, 2, 3)] \neq 0$. The three spectral generators together with the two-dimensional preserving span have exact rank 5 in $\mathrm{Sym}(3)$ (a verifier-computed minor is 225), hence $L \cap \mathfrak{E} = \{0\}$ in the $\Delta = -1$ regime. Moreover $Y_0 = \mathrm{diag}(1, -1, 0) \in \mathfrak{E}^\perp$ pairs with the three generator columns as $(8, 9, 8)$, whose trace weights are $(10, 10, 11)$. After trace normalisation, writing $t = \sigma^2$ and $B_t = A_\sigma$ for the trace-normalised atom, clearing the continuum denominators in $\langle Y_0, B_t \rangle \geq 24/11$ gives $3(t+3)^2(27t^2+70t+51)/11$, coefficientwise positive for $t \geq 0$. Therefore the unit witness $Y_0/\sqrt{2}$ certifies $\delta_\star \geq 12\sqrt{2}/11$ at every compactified scale. This is an exact positive lower bound for one idealised instance, not equality for $\delta_\star$ and not a UNet measurement.

A separately certified path with $\Sigma_q = I$, $\Sigma_p = \mathrm{diag}(1+\epsilon, 1)$, and the same target $\mathfrak{E} = \mathrm{span}\{I\}$ ($d = 2$) has $\delta_\star = \sqrt{2}\,\epsilon/(2+\epsilon)$ for $\epsilon \in \{1, 1/2, 1/4, 1/8, 1/16\}$ and tends to the reachable $\epsilon = 0$ case; it illustrates continuity towards the well-specified case, not generic transversality.

Let P_a be the orthogonal projector onto S_a, let $r_a = \dim S_a$, and let P_0 project onto $S^\perp$. The Frobenius projection onto $\mathfrak{E}$, the linear span of the Bayes-preserving set, is

$$\Pi_{\mathfrak{E}}(B) = \bigoplus_{a=1}^{m} \frac{\mathrm{tr}(P_a B P_a)}{r_a} I_{S_a} \;\oplus\; P_0 B P_0|_{S^\perp}\,. \tag{14}$$

Indeed, the block-scalar matrices and the unrestricted symmetric $S^\perp$ block are mutually orthogonal subspaces, which gives equation 14. If $B \succ 0$, every displayed scalar is positive and the final compression is positive definite; moreover the traces of the blocks sum to $\mathrm{tr}\, B$. Hence for every $B \succ 0$ with trace d, $\Pi_{\mathfrak{E}}(B) \in \mathcal{E}_d$, and Pythagoras gives

$$\mathrm{dist}_F(B, \mathcal{E}_d) = \|\Pi_{\mathfrak{E}^\perp} B\|_F. \tag{15}$$

For completeness, the noise scale is compactified by its two limits. Since $\Sigma_q \succ 0$, A_σ extends continuously to $\sigma = 0$ and $\sigma = \infty$ after trace normalisation. Its image is therefore compact and uniformly positive definite, as is its convex hull $\mathcal{R}_d$. In finite dimension this convex hull equals the closure of the finite convex hull of the interior atoms. Normalising any nonzero finite conic combination of the generators gives a convex combination of the A_σ; conversely, a finite convex combination is obtained by rescaling its coefficients by $\mathrm{tr}\, G_\sigma$. Thus $\mathcal{R}_d$ is exactly the closure of trace-normalised metrics from arbitrary finite-support schedules at positive noise scales; a prescribed finite grid gives a smaller polytope, and an endpoint is only a limit when no finite noise scale attains it.

Set $K = \Pi_{\mathfrak{E}^\perp}\mathcal{R}_d$ and let $z_\star$ be its closest point to zero. If $z_\star \neq 0$, the projection variational inequality with $Y_\star = z_\star/\|z_\star\|_F$ gives $\langle Y_\star, z\rangle \geq \|z_\star\|_F$ for every $z \in K$; if $z_\star = 0$, take $Y_\star = 0$. Conversely, for every $\|Y\|_F \leq 1$, $\min_{z\in K}\langle Y, z\rangle \leq \langle Y, z_\star\rangle \leq \|z_\star\|_F$. Because $Y \in \mathfrak{E}^\perp$ removes the projection and a linear functional has the same minimum over a convex hull as over its compact generating curve, these inequalities prove equation 13. They also prove the stated zero-versus-positive alternative.

Finally let $R(B)$ denote population 0–1 risk under the Gaussian model. Decision boundaries have probability zero, so dominated convergence makes R continuous on the compact set $\mathcal{R}_d$. The Bayes posterior has a unique maximiser away from those boundaries; hence any rule that disagrees with it on a positive-measure set has strictly larger risk. By Theorem 1, $R(B) = R_{\mathrm{Bayes}}$ can therefore occur in $\mathcal{R}_d$ only when $B \in \mathcal{E}_d$. If $\delta_\star > 0$, compactness now yields

$$\min_{B\in\mathcal{R}_d} \{R(B) - R_{\mathrm{Bayes}}\} > 0.$$

This argument is qualitative: it neither identifies the value of that minimum nor calibrates it as a function of $\delta_\star$. Finite positive-scale normalised weightings are dense in $\mathcal{R}_d$, so continuity makes the same positive value their infimum; an endpoint need not be attained.

The released deterministic verifier exercises all equalities above. It supplies exact primal and dual witnesses for the three controls in Figure 8; after clearing positive denominators, each continuum inequality is an identically zero or coefficientwise nonnegative rational polynomial on $\sigma^2 \geq 0$. It also checks a positive exceptional path converging to a reachable instance and regression-tests trace and positive-definiteness preservation, self-adjointness, idempotence, and the Pythagorean identity of equation 14. Thus no sampled scale grid or Monte Carlo estimate is used as a certificate. Appendix S extends these deterministic theorem regressions to finite labelled tables: exact rational examples separate partition distance from the shared gap. Its positive-distance, zero-gap example uses the $\epsilon = 1$ member of the path above.

P A UNIFORM QUADRATIC RISK LAW ON COMPACT FAMILIES

This appendix strengthens the qualitative risk floor stated after Theorem 5: on any fixed compact trace-normalised SPD family, excess 0–1 risk under the Gaussian model is *two-sidedly quadratic*

in the Frobenius distance to the exact Bayes-preserving set $\mathcal{E}_d$. The constants exist but are hull-dependent and are not evaluated numerically, so nothing here upgrades the statement in Appendix O that the risk floor is not a calibrated numerical function of $\delta_\star$. Throughout we work in whitened coordinates $z = \Sigma_p^{-1/2}x$, so $z \mid c \sim \mathcal{N}(\mu_c, I_d)$ with equal priors and the distinct whitened means $\mu_1, \dots, \mu_C$ of Section 5; $R(B)$ is the population 0–1 risk of the metric rule $h_B(z) = \min \arg\min_i (z-\mu_i)^\top B(z-\mu_i)$ (ties broken by smallest index; ties are null); in these coordinates $B = I$ realises the Bayes rule, so $R_{\mathrm{Bayes}} = R(I)$. We write $\varphi_i(z) = (2\pi)^{-d/2} e^{-\|z-\mu_i\|^2/2}$, $\delta_{ij} = \mu_i - \mu_j$, $s_{ij} = \frac{1}{2}(\mu_i + \mu_j)$, and $g_{ij}^B(z) = \delta_{ij}^\top B(z - s_{ij})$, so that $g_{ij}^I \geq 0$ cuts out the Euclidean cell $V_i = \bigcap_{j \neq i} \{g_{ij}^I \geq 0\}$ with pair faces $F_{ij} = V_i \cap V_j$; a pair is active (written $e = \{i, j\} \in E$) when F_{ij} has a $(d-1)$-dimensional facet, matching Section 5. For each active e we fix once and for all an ordered representative (i, j) and write $\delta_e = \delta_{ij}$, $s_e = s_{ij}$, $n_e = \delta_e / \|\delta_e\|$, $F_e = F_{ij}$, and $H_e = H_{ij}$ for the $\{i, j\}$ bisector hyperplane $\{g_{ij}^I = 0\}$. Swapping the representative flips the signs of δ_e, n_e, q_e^H, and u_e, but leaves every risk-level object unchanged: the displaced point $u_e n_e$, the segment $\{y + v n_e : v \in I_e(y)\}$, the pair loss J_e, and the quadratic form Q_A (where q_e^H enters squared) are orientation-independent, and the two winner cases below exchange accordingly. The Bayes-preserving span is $\mathfrak{E} = \{A = A^\top : A\delta_e \parallel \delta_e \ \forall e \in E\}$ with $\mathcal{E} = \mathfrak{E} \cap \mathrm{SPD}$ and $\mathcal{E}_d = \{A \in \mathcal{E} : \operatorname{tr} A = d\}$, as in Theorem 1 and Appendix O; $\mathfrak{E}$ is unrelated to the covariance spectral span L of Section 5.2, and the two never mix below.

Theorem 6 (Compact-family uniform quadratic risk law)**.** *Fix $d \geq 2$, $C \geq 2$, distinct means, and any nonempty compact family $\mathcal{H} \subset \{B \succ 0 : \operatorname{tr} B = d\}$. There are constants $0 < c_{\mathcal{H}} \leq C_{\mathcal{H}} < \infty$, depending only on the mean configuration and $\mathcal{H}$, such that for every $B \in \mathcal{H}$,*

$$c_{\mathcal{H}} \operatorname{dist}_F(B, \mathcal{E}_d)^2 \ \leq \ R(B) - R_{\mathrm{Bayes}} \ \leq \ C_{\mathcal{H}} \operatorname{dist}_F(B, \mathcal{E}_d)^2. \tag{16}$$

(If $\mathcal{H} \subset \mathcal{E}_d$ both sides vanish and the claim is trivial.)

Corollary 7 (Schedule risk floor at the hull)**.** *Take $\mathcal{H} = \mathcal{R}_d$, the compact normalised reachable hull, and $\delta_\star$ from Theorem 5. Then*

$$c_{\mathcal{R}_d} \delta_\star^2 \ \leq \ \inf_w \big[R(B_w) - R_{\mathrm{Bayes}}\big] = \min_{B \in \mathcal{R}_d} \big[R(B) - R_{\mathrm{Bayes}}\big] \ \leq \ C_{\mathcal{R}_d} \delta_\star^2, \tag{17}$$

where the infimum runs over normalised finite positive-scale weightings. The infimum need not be attained by a finite schedule: an optimiser over the closed hull may require a compactification-endpoint atom, which no finite positive-scale schedule realises.

The constants in equation 16–equation 17 may depend on the means, the active Voronoi geometry, the uniform SPD margin of the family, and the far-set risk minimum; they are not evaluated, are not universal across covariance instances, quantify no real UNet, and assert no rate as $\delta_\star \to 0$ across model families. Compactness is essential for the uniform bound: at the end of the appendix we give a two-class example on the *unnormalised* SPD cone where the distance stays $\sqrt{2}$ while the excess risk tends to 0.

P.1 Projection, fixed cells, and the exact zero set

Lemma 8 (Positivity- and trace-preserving projection)**.** *Decompose $\mathbb{R}^d = S_1 \oplus \cdots \oplus S_m \oplus S_0$ into the active blocks and their complement, with orthogonal projectors P_a, P_0 and $r_a = \dim S_a$. The Frobenius projection onto $\mathfrak{E}$ is $\Pi_{\mathfrak{E}} B = \bigoplus_a \frac{\operatorname{tr}(P_a B P_a)}{r_a} I_{S_a} \oplus P_0 B P_0$. Consequently $B \succeq \mu I$ implies $\Pi_{\mathfrak{E}} B \succeq \mu I$, and $\operatorname{tr} \Pi_{\mathfrak{E}} B = \operatorname{tr} B$ since $I \in \mathfrak{E}$. Hence for $B \succ 0$ with $\operatorname{tr} B = d$, the point $A := \Pi_{\mathfrak{E}} B$ lies in $\mathcal{E}_d$, $N := \Pi_{\mathfrak{E}^\perp} B$ satisfies $\operatorname{dist}_F(B, \mathcal{E}) = \operatorname{dist}_F(B, \mathcal{E}_d) = \|N\|_F$, and $\mathcal{K} := \Pi_{\mathfrak{E}} \mathcal{H}$ is a compact subset of $\mathcal{E}_d$ with the same uniform lower spectral bound as $\mathcal{H}$.*

Proof. The block form of $\Pi_{\mathfrak{E}}$ is the orthogonality relation $\langle B - \Pi_{\mathfrak{E}} B, A' \rangle = 0$ for all $A' \in \mathfrak{E}$, checked blockwise. Each scalar block average of a matrix $\succeq \mu I$ is $\geq \mu$ and each compression $P_0 B P_0$ is $\succeq \mu P_0$, giving positivity; the trace identity is $I \in \mathfrak{E}$. Pythagoras and $A \in \mathcal{E}_d$ give the distance identities, i.e. equation 15. □

Lemma 9 (Cells are fixed on $\mathcal{E}$; exact zero set)**.** *For every $A \in \mathcal{E}$ the closed cells of the metric rule h_A coincide with the Euclidean cells, $V_i^A = V_i$ for all i; in particular $g_{ij}^A \geq 0$ on V_i with equality exactly on F_{ij}. Moreover $R(B) = R_{\mathrm{Bayes}}$ iff $B \in \mathcal{E}$, and R is continuous on compact subsets of the SPD cone.*

Proof. On an active pair $A\delta_e = \alpha_e(A)\delta_e$ with $\alpha_e(A) > 0$, so each facet-defining inequality is rescaled by a positive factor and $V_i^A \subseteq V_i$; since both cell systems cover $\mathbb{R}^d$ and Euclidean cell interiors are disjoint, interiors satisfy the reverse inclusion and closures agree. The equivalence $R(B) = R_{\text{Bayes}} \iff B \in \mathcal{E}$ is Theorem 1 together with the a.e. uniqueness of the posterior winner; continuity is dominated convergence off the null tie set, as in Appendix O. □

P.2 A COMPACT WEAK-SHARPNESS LEMMA

Lemma 10 (Compact fixed-face weak sharpness)**.** *Let $P \subset \mathbb{R}^d$ be a nonempty polyhedron and Θ a compact parameter set, and let $f_\theta(x) = c_\theta^\top x + b_\theta$ depend continuously on θ with $f_\theta \geq 0$ on P. (i) If the zero set $\{x \in P : f_\theta(x) = 0\}$ equals one common nonempty face F for every θ, then there is $\gamma_F > 0$ with $f_\theta(x) \geq \gamma_F \operatorname{dist}(x, F)$ for all $x \in P$, $\theta \in \Theta$. (ii) If the zero set is empty for every θ, then $\inf_{P\times\Theta} f_\theta > 0$.*

Proof. Boundedness below forces $c_\theta \perp \operatorname{lin}(P)$, so pass to the pointed quotient and fix a Minkowski–Weyl representation $P = \operatorname{conv}\{v_1, \dots, v_p\} + \operatorname{cone}\{r_1, \dots, r_q\} + \operatorname{lin}(P)$. For (i) with $F \neq P$ pick a vertex $v_0 \in F$; given $x \in P$, move the coefficients of all vertices outside F onto v_0 and drop the rays outside $\operatorname{rec} F$ to produce $y \in F$ with $x - y = \sum_{v_a \notin F} \lambda_a (v_a - v_0) + \sum_{r_b \notin \operatorname{rec} F} \mu_b r_b$. Every outside vertex has $f_\theta(v_a) > 0$ and every outside ray has $c_\theta^\top r_b > 0$ (a zero slope would generate a zero ray leaving the common face, a contradiction), so by finiteness and compactness $\eta := \min_\theta \min\{f_\theta(v_a), c_\theta^\top r_b\} > 0$ over the outside generators. With C_P a bound on the generator norms, $\operatorname{dist}(x, F) \leq \|x - y\| \leq C_P(\sum \lambda_a + \sum \mu_b) \leq (C_P/\eta) f_\theta(x)$. For (ii) the minimum over P is attained at the finitely many quotient vertices and is a positive continuous function of θ on a compact set. □

Applying Lemma 10 with $P = V_i$, $f = g_{ij}^A$, $\Theta = \mathcal{K}$ (finitely many ordered pairs) yields constants $\gamma_\mathcal{K}, m_\mathcal{K} > 0$: nonempty F_{ij} obey $g_{ij}^A(z) \geq \gamma_\mathcal{K} \operatorname{dist}(z, F_{ij})$ on V_i, empty F_{ij} obey $g_{ij}^A \geq m_\mathcal{K}$ on V_i. Applying it inside the affine hull of each active facet F_e to the third-class slacks $a_{e,k}^A = g_{ik}^A|_{F_e}$ (zero face $G_{e,k} = F_e \cap V_k$) yields facet constants $\gamma_\mathcal{K}^{\text{fac}}, m_\mathcal{K}^{\text{fac}}$, with the conventions $1, +\infty$ when the index family is empty.

P.3 UNIFORM TRANSVERSE EXPANSION

Fix $A \in \mathcal{K}$ and $N = \varepsilon H \in \mathfrak{E}^\perp$ with $\|H\|_F = 1$, and let $B = A + \varepsilon H$ (SPD for small ε by the uniform margin). Put $\Delta_{\min} = \min_{i\neq j} \|\delta_{ij}\|$, $\Delta_{\max} = \max_{i\neq j} \|\delta_{ij}\|$, $S_{\max} = \max \|s_{ij}\|$, $K_R = \Delta_{\max}(R + 1 + S_{\max})$ and the moving-boundary scale $w_R = \varepsilon K_R$.

Proposition 11 (Uniform local expansion)**.** *Define the quadratic form*

$$Q_A(H) = \frac{1}{2C} \sum_{e=\{i,j\}\in E} \frac{1}{\alpha_e(A)^2 \|\delta_e\|} \int_{F_e} \varphi_i(y) \left[\delta_e^\top H(y - s_e)\right]^2 d\mathcal{H}^{d-1}(y). \tag{18}$$

Then

$$\sup_{\substack{A\in\mathcal{K},\ N\in\mathfrak{E}^\perp \\ 0<\|N\|_F\leq\eta}} \frac{\left|R(A+N) - R_{\text{Bayes}} - Q_A(N)\right|}{\|N\|_F^2} \longrightarrow 0 \qquad (\eta \downarrow 0). \tag{19}$$

Proof. Off the null tie set, split the disagreement set $\{h_B \neq h_I\} \cap \mathbb{B}_{R-2}$ (the Euclidean ball of radius $R - 2$; write $\mathcal{D}_{ij}(R - 2)$ for the piece with true winner i and perturbed winner j, and H_{ij} for the $\{i, j\}$ bisector hyperplane) by the ordered pair (true winner i, perturbed winner j); on such a point $0 \leq g_{ij}^A(z) \leq \varepsilon|\delta_{ij}^\top H(z - s_{ij})| \leq w_R$, so once $w_R < m_\mathcal{K}$ empty pairs cannot disagree and nonempty pairs satisfy $\operatorname{dist}(z, F_{ij}) \leq w_R/\gamma_\mathcal{K}$.

Singular part. Call a disagreement point singular if its pair is inactive, or if its orthogonal projection $\pi_e(z)$ onto the active bisector leaves F_e or comes within $b_\mathcal{K} w_R$ of the relative boundary $\partial_{\text{rel}} F_e$ (a fixed $b_\mathcal{K} \geq 2c_\mathcal{K}^{\text{slack}}/\gamma_\mathcal{K}^{\text{fac}}$ suffices, with $c_\mathcal{K}^{\text{slack}} = 1 + (\beta + 1)\Delta_{\max} U_\mathcal{K}$, $\beta = \max_\mathcal{K} \|A\|_{\text{op}}$, $U_\mathcal{K} = 2/(a_-\Delta_{\min})$, and $0 < a_- \leq \alpha_e(A) \leq a_+$ the compact eigenfactor bounds). Two polyhedral identities control this set: a nonempty *inactive* face is covered by the triple intersections $G_{ijk} =$

$V_i \cap V_j \cap V_k$, and the relative boundary of an *active* facet is the union of its G_{ijk}. Whenever $G_{ijk} \neq \varnothing$ the normals δ_{ij}, δ_{ik} are linearly independent (collinear normals with a common point would force $\mu_j = \mu_k$), so each $L_{ijk} = H_{ij} \cap H_{ik}$ is an affine flat of codimension two. Hence the singular set is covered by the pair-labelled tubes $\{z \in \mathcal{D}_{ij}(R-2) : \mathrm{dist}(z, G_{ijk}) \le D_{\mathcal{K}} w_R\}$, where one may take the explicit compact constant $D_{\mathcal{K}} = c_{n,\mathcal{K}} + \max\{\gamma_{\mathcal{K}}^{-1}, b_{\mathcal{K}}\}$ with $c_{n,\mathcal{K}} = 1/(a_- \Delta_{\min})$: an active disagreement point lies within $c_{n,\mathcal{K}} w_R$ of its bisector, and its distance to the relevant face or facet boundary is bounded by $\gamma_{\mathcal{K}}^{-1} w_R$ or $b_{\mathcal{K}} w_R$ respectively. On such a tube the point is within $D_{\mathcal{K}} w_R$ of *its own* bisector H_{ij}, where $\varphi_i = \varphi_j$; bounded Gaussian gradients give a density gap $O(w_R)$, and the codimension-two tube volume in $\mathbb{B}_R$ is $O(w_R^2(1+R)^{d-2})$. Summing the finitely many ordered triples (overlaps cost only a finite factor),

$$0 \;\le\; R_{\mathrm{sing}} \;\le\; C_{\mathrm{sing},\mathcal{K}}\, w_R^3\, (1+R)^{d-2}. \tag{20}$$

Regular part. On the regular base $F_{e,\mathrm{reg}}^{\varepsilon,R} = \{y \in F_e : \|y\| \le R-1,\ \mathrm{dist}_{H_e}(y, \partial_{\mathrm{rel}} F_e) > b_{\mathcal{K}} w_R\}$ write points as $y + v n_e$ with $n_e = \delta_e/\|\delta_e\|$. Exactly, $g_{ij}^{A+\varepsilon H}(y+vn_e) = [\alpha_e(A)\|\delta_e\| + \varepsilon\, \delta_e^\top H n_e]\, v + \varepsilon\, \delta_e^\top H(y - s_e)$, whose bracket is uniformly positive for small ε; the perturbed boundary displacement is therefore

$$u_e(y) = -\,\frac{\varepsilon\, \delta_e^\top H(y - s_e)}{\alpha_e(A)\|\delta_e\| + \varepsilon\, \delta_e^\top H n_e}, \qquad |u_e(y)| \le U_{\mathcal{K}} w_R < 1. \tag{21}$$

The choice of $b_{\mathcal{K}}$ makes every third-class slack on the base dominate the boundary movement, so with $I_e(y) := [\min\{0, u_e(y)\}, \max\{0, u_e(y)\}]$ only classes i, j compete under both rules on the segment $\{y + v n_e : v \in I_e(y)\}$, and the positive bracket in equation 21 identifies the ordered winners: for $u_e(y) > 0$ and $0 < v < u_e(y)$ the true winner is i and the perturbed winner j; for $u_e(y) < 0$ and $u_e(y) < v < 0$ the roles swap. Define the strip $\mathcal{D}_{e,\mathrm{reg}}^{\varepsilon,R} := \{y + v n_e : y \in F_{e,\mathrm{reg}}^{\varepsilon,R},\ v \in \mathrm{int}\, I_e(y)\}$, the exact nonnegative pair loss

$$\begin{aligned} J_e\big(y, u_e(y)\big) &= \frac{1}{C} \int_{I_e(y)} \big|\varphi_i(y + v n_e) - \varphi_j(y + v n_e)\big|\, dv, \\ R_{\mathrm{reg}}^{\varepsilon,R} &:= \sum_{e \in E} \int_{F_{e,\mathrm{reg}}^{\varepsilon,R}} J_e\big(y, u_e(y)\big)\, d\mathcal{H}^{d-1}(y). \end{aligned} \tag{22}$$

The affine map $(y, v) \mapsto y + v n_e$ is bijective with unit Jacobian, so equation 22 is exactly the excess risk carried by $\bigcup_e \mathcal{D}_{e,\mathrm{reg}}^{\varepsilon,R}$: interval endpoints are ties (null), open strips of distinct active edges are disjoint (a common point would carry two distinct ordered winner changes), and a coverage check shows every non-singular disagreement point of $\mathbb{B}_{R-2}$ lies in exactly one strip — a strip may stick out of $\mathbb{B}_{R-2}$, but only into genuine disagreement inside $\mathbb{B}_R$, so equation 22 undercounts nothing in the core and overcounts no non-disagreement. The density difference vanishes on H_e with normal derivative $\|\delta_e\|\varphi_i(y)$, and a second-order Taylor bound with cubic remainder controlled by the surface-integrable weight $(1 + \|y\|)^2 \sum_{c \in \{i,j\}} e^{-\|y - \mu_c\|^2/4}$ gives

$$\Big| R_{\mathrm{reg}}^{\varepsilon,R} - \varepsilon^2 Q_{A,\mathrm{reg}}^{\varepsilon,R}(H) \Big| \le C_{\mathrm{reg},\mathcal{K}}\, \varepsilon^3, \quad Q_{A,\mathrm{reg}}^{\varepsilon,R}(H) = \frac{1}{2C} \sum_e \frac{\int_{F_{e,\mathrm{reg}}^{\varepsilon,R}} \varphi_i\, [\delta_e^\top H(\cdot - s_e)]^2\, d\mathcal{H}^{d-1}}{\alpha_e(A)^2 \|\delta_e\|}, \tag{23}$$

after replacing u_e by its first-order term (the replacement error is $O(\varepsilon^2(1 + \|y\|))$ pointwise, hence $O(\varepsilon^3)$ after integration).

Collar and tails. The base omitted from equation 18 splits into the surface tail $\|y\| > R - 1$, whose contribution is $\varepsilon^2 \tau(R-1)$ with $\tau(r) = \sum_e \int_{F_e \cap \mathbb{B}_r^c} (1 + \|y\|)^2 \varphi_i\, d\mathcal{H}^{d-1} \downarrow 0$, and a boundary collar of width $O(w_R)$ inside finitely many codimension-one tubes of the flats containing the $G_{e,k}$, of surface measure $O(w_R(1+R)^{d-2})$; with $q_e^H(y) := \delta_e^\top H(y - s_e)$ bounded by $|q_e^H| \le K_R$ there, the collar costs $O(w_R^3(1+R)^{d-2})$. Choosing $R(\varepsilon) = M + \sqrt{d} + \sqrt{8 \log(1/\varepsilon)} + 3$ (with $M = \max_i \|\mu_i\|$) makes $\Pr\{\|Z\| > R - 2\} \le \varepsilon^4$ (for Z the whitened observation) and $w_R = O(\varepsilon \sqrt{\log(1/\varepsilon)})$, so equation 20, equation 23, the collar, and the tail are all $o(\varepsilon^2)$, uniformly over $\mathcal{K}$ and the unit sphere of $\mathfrak{E}^\perp$. Combining the four pieces proves equation 19. □

P.4 COERCIVITY, THE LOCAL LAW, AND GLOBALISATION

Lemma 12 (Kernel and coercivity)**.** *$Q_A(N) \geq 0$, and for $N \in \mathfrak{E}^\perp$, $Q_A(N) = 0$ iff $N = 0$. Moreover with $\kappa = \min\{Q_I(N) : N \in \mathfrak{E}^\perp, \|N\|_F = 1\} > 0$ and $\Lambda = \max\{Q_I(N) : N \in \mathfrak{E}^\perp, \|N\|_F = 1\} < \infty$,*

$$\frac{\kappa}{a_+^2}\|N\|_F^2 \;\leq\; Q_A(N) \;\leq\; \frac{\Lambda}{a_-^2}\|N\|_F^2 \qquad (A \in \mathcal{K},\ N \in \mathfrak{E}^\perp). \tag{24}$$

Proof. If $Q_A(N) = 0$ then $\delta_e^\top N(y - s_e) = 0$ on a relatively open subset of every active facet, which forces $N\delta_e \parallel \delta_e$ for every active e, i.e. $N \in \mathfrak{E}$; intersect with $\mathfrak{E}^\perp$. The unit sphere of $\mathfrak{E}^\perp$ is compact and Q_I is continuous and positive there, giving κ, Λ; Q_A differs from Q_I only through the per-edge factors $\alpha_e(A)^{-2} \in [a_+^{-2}, a_-^{-2}]$. □

Proof of Theorem 6. Write $B = A + N$ with $A = \Pi_{\mathfrak{E}} B \in \mathcal{K}$ and $N \in \mathfrak{E}^\perp$, $\|N\|_F = \mathrm{dist}_F(B, \mathcal{E}_d)$ (Lemma 8). Choose $\eta_{\mathcal{K}} > 0$ so small that the remainder in equation 19 is below $\kappa/(2a_+^2)$ and $\eta_{\mathcal{K}} < \frac{1}{2}\min_{\mathcal{K}} \lambda_{\min}$. On the near set $\{\|N\|_F \leq \eta_{\mathcal{K}}\}$, Proposition 11 and equation 24 sandwich $R(B) - R_{\mathrm{Bayes}}$ between $\frac{\kappa}{2a_+^2}\|N\|_F^2$ and $(\frac{\Lambda}{a_-^2} + \frac{\kappa}{2a_+^2})\|N\|_F^2$. The far set $\{B \in \mathcal{H} : \|N\|_F \geq \eta_{\mathcal{K}}\}$, if nonempty, is compact and disjoint from $\mathcal{E}_d$, so by Lemma 9 its excess-risk minimum m_{far} is positive, while $d_E := \mathrm{dist}_F(\cdot, \mathcal{E}_d)$ is bounded on $\mathcal{H}$ by some $D_{\mathcal{H}}$ and the excess risk by some $M_{\mathcal{H}}$; hence $\frac{m_{\mathrm{far}}}{D_{\mathcal{H}}^2} d_E^2 \leq R - R_{\mathrm{Bayes}} \leq \frac{M_{\mathcal{H}}}{\eta_{\mathcal{K}}^2} d_E^2$ there. Taking the worse of the near and far constants gives equation 16. □

Proof of Corollary 7. Take minima in equation 16 over the compact hull $\mathcal{R}_d$. For the lower bound, every $B \in \mathcal{R}_d$ has $d_E(B) \geq \delta_\star$, so $R(B) - R_{\mathrm{Bayes}} \geq c_{\mathcal{R}_d}\delta_\star^2$ and the same holds for the risk minimum. For the upper bound, evaluate the right inequality of equation 16 at a *distance* minimiser $B_\star$ (which exists by compactness and attains $d_E(B_\star) = \delta_\star$): the risk minimum is at most $R(B_\star) - R_{\mathrm{Bayes}} \leq C_{\mathcal{R}_d}\delta_\star^2$. Finite positive-scale normalised weightings are dense in $\mathcal{R}_d$ (Appendix O), and R is continuous, so their infimum equals the hull minimum; we claim only the infimum for finite schedules, since a hull optimiser may require a compactification-endpoint atom. □

Necessity of scale control. For $d = 2$, $C = 2$, $\mu_{1,2} = \pm(a/2, 0)$ one has $\mathcal{E} = \{\mathrm{diag}(r, s) : r, s > 0\}$. On the unnormalised cone take $B_n = nI + \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$: the distance $\mathrm{dist}_F(B_n, \mathcal{E}) = \sqrt{2}$ is constant while $R(B_n) - R_{\mathrm{Bayes}} = \Phi(-\frac{an}{2\sqrt{n^2+1}}) - \Phi(-\frac{a}{2}) = \Theta(n^{-2}) \to 0$, so no uniform lower constant exists without the trace slice. After normalising, $\widetilde{B}_n = I + N/n$ with $N = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$ has distance $\Theta(n^{-1})$ and excess risk $\Theta(n^{-2})$, matching equation 16. An analytic two-class identity ($R(B) = \Phi(-\frac{ab}{2\sqrt{b^2+e^2}})$ for $B = \left(\begin{smallmatrix} b & e \\ e & 2-b \end{smallmatrix}\right)$, with $\mathrm{dist}_F(B, \mathcal{E}_d)^2 = 2e^2$) reproduces the quadratic ratio $a\, e^{-a^2/8}/(8\sqrt{2\pi}\, b^2)$ in the small-e limit; this served as an independent consistency check of the constants' existence, not as part of the proof.

Scope. Theorem 6 is a statement about the idealised Gaussian model on one fixed compact family, and the proof uses the following assumptions: distinct *fixed* means ($\Delta_{\min} > 0$ drives the tie hyperplanes and the codimension-two triple geometry), $d \geq 2$ and $C \geq 2$, equal priors with a common *nonsingular* covariance (whitening fixes the mid-bisectors and the strictly positive normal density derivative), SPD metrics with the trace slice (projection positivity and removal of the scale degeneracy), and compactness of the family (uniform $a_\pm$, the tail/local radius, and the far-set positive minimum; see the counterexample above). It licenses none of the following: a universal or covariance-independent constant; a rate in $\delta_\star$ across model families or as the mean configuration varies; a calibrated numerical risk value; any statement about the real UNet's risk; or an extension to unequal priors, class-dependent covariances, or non-Gaussian boundary contact.

Q PREREGISTRATION AND REPRODUCTION

Audit identity and margin tolerance. Throughout the paper, "independently audited" means reverified by a separate, mechanically distinct implementation run against the released artifacts; it is

Table 29: Global accounting of the six audited reachability tables, plus the reference-protocol remeasurement of panel C (row F). All rows use $n = 100$ images and strict margin $> \tau$ on per-image max-normalised contrasts (C, E and F reuse one registered endpoint-reference scale per endpoint); all rows use the squared-ℓ_2 denoising loss except F, whose primary arm uses the ℓ_1 loss of the published CIFAR-10 protocol. A is uniform accuracy under the row's strict threshold, O the label-aware pointwise ceiling (a diagnostic, not attainable performance), and $G^\star$ the shared-weight optimum. "Exact" means machine-checked exact rational certificates establish both O and $G^\star$ on the row's dyadic table reading; "solver" (solver-grade) means L is independently postchecked feasible and U is a floating-point solver-derived upper bound. Registration status is per package; C and E are dependent endpoints of one preregistration, not independent replications.

Measurement protocols

Table	Support	(T, R)	τ	Status
SD 1.5, Pet original	$[250, 750]$	$(21, 64)$	10^{-6}	post-hoc
SD 1.5, CIFAR-10 original	$[250, 750]$	$(21, 64)$	10^{-6}	post-hoc
SD 1.5, CIFAR-10 fresh zero-overlap (C)	$[250, 750]$	$(21, 64)$	10^{-6}	preregistered
SD 1.5, CIFAR-10 C panel, dense grid (E)	$[250, 750]$	$(101, 32)$	10^{-6}	preregistered
SD 1.5, CIFAR-10 C panel, reference protocol (F)	$[50, 950]$	$(21, 64)$	10^{-6}	preregistered
PixArt-α, Pet original	6-pt base	$(6, 32)$	10^{-12}	plan during meas.
PixArt-α, CIFAR-10 original	6-pt base	$(6, 32)$	10^{-12}	plan during meas.

Reachability and shared-weight bounds

Table	A	O	$G^\star$	$O - G^\star$	Grade
SD 1.5, Pet original	76	97	86	11	exact
SD 1.5, CIFAR-10 original	72	91	81	10	exact
SD 1.5, CIFAR-10 fresh zero-overlap (C)	78	97	89	8	exact
SD 1.5, CIFAR-10 C panel, dense grid (E)	80	99	92	7	exact
SD 1.5, CIFAR-10 C panel, reference protocol (F)	79	98	$[87, 92]$	$[6, 11]$	solver
PixArt-α, Pet original	44	78	$[48, 52]$	$[26, 30]$	solver
PixArt-α, CIFAR-10 original	46	78	$[55, 57]$	$[21, 23]$	solver

not an author-independent external review. The SD strict-margin tolerance $\tau = 10^{-6}$ on per-image max-normalised contrasts and the PixArt-α tolerance $\tau = 10^{-12}$ were fixed in their respective analysis plans before analysis. For the original two tables, the losses had already been examined, so the shared-weight analysis is post-hoc. Each study reports strict-margin sensitivity across its prespecified cutoffs, without qualitative change in the reported separation; the C/E and PixArt-α studies also report their prespecified criteria.

Original MSE scoring protocol. Oxford-IIIT Pet uses the literal template "a photo of a {breed}" with underscores replaced by spaces and no article adjustment; CIFAR-10 uses the scorer's ten fixed prompt strings, such as "a photo of an airplane" and "a photo of a car". Images are re-

sized bicubically to a short edge of 512, centre-cropped to 512×512, and normalised to $[-1, 1]$. Each cell loss is the mean squared ϵ-prediction error over latent coordinates (the squared-ℓ_2 loss up to the constant factor $1/d$) of the conditional branch alone; neither the SD probe nor the PixArt runners build an unconditional embedding or apply classifier-free guidance. The six-point grid is $\{250, 350, 450, 550, 650, 750\}$ and the 21-point grid runs from 250 to 750 in steps of 25; noise draws are seeded per cell (below). The PixArt-α runners reuse the same prompt lists (Appendix K).

Model and data versions. The original SD 1.5 and PixArt measurements use the following Hugging Face repositories and immutable revisions; dataset revisions refer to the test parquet files. These identifiers specify the source weights and data, while the ordered image manifests specify the actual experimental panels.

- SD 1.5: `stable-diffusion-v1-5/stable-diffusion-v1-5`, revision `451f4fe16113bff5a5d2269ed5ad43b0592e9a14`.
- PixArt-α: `PixArt-alpha/PixArt-XL-2-512x512`, revision `50f702106901db6d0f8b67eb88e814c56ded2692`.
- CIFAR-10: `uoft-cs/cifar10`, revision `0b2714987fa478483af9968de7c934580d0bb9a2`.
- Oxford-IIIT Pet: `timm/oxford-iiit-pet`, revision `089695c834a7deb60505b7cc506672db1c31a6aa`.

The original CIFAR-10 panel has ten images per class. The Pet panel has two or three images per breed, totalling 100; it is not an exactly balanced three-per-breed sample. PixArt reuses the corresponding ordered SD panels. Fresh panels and their disjointness checks are specified in their respective experimental appendices.

Numerical and noise settings. In these original measurements, the image is converted to a float tensor before antialiased bicubic resizing; the later PIL-resize controls change this order explicitly. Latents use the VAE posterior mean multiplied by the checkpoint's scaling factor, with no posterior sampling. SD runs in fp16 with TensorFloat-32 (TF32) enabled; PixArt's denoiser, VAE, text encoder, and noise use fp32 with TF32 disabled. Residuals are reduced in fp32 and the scalar cell losses are stored in float64. Each class batch contains all 10 CIFAR-10 or 37 Pet classes; the SD trial chunk is 8. All classes in an image–timestep–trial cell share the same noisy latent. The original master seed is 20260813. Its UTF-8 key is the decimal string `seed|dataset_index|timestep|trial`; an eight-byte BLAKE2b digest, read as an unsigned big-endian integer and masked to 63 bits, seeds an NVIDIA CUDA generator for that cell. The image identifier is the dataset index, not its position in the selected panel. PixArt uses the SD timestep in this key and the noise-level-matched PixArt timestep in its scheduler. Its six-point SD grid maps to $\{194, 256, 320, 387, 458, 531\}$; the other six measured points are the support intervention in Appendix K. Both schedulers are loaded from their pinned model repositories. The SD scheduler has 1,000 training timesteps and a scaled-linear beta schedule with endpoints 0.00085 and 0.012. Table 29 gives each panel's T, R, and strict-margin threshold. Later noise blocks, preprocessing controls, and the 500-image panel retain their separately stated seeds and reference scales; the original seed is not a default for those experiments.

Software and optimisation settings. The original SD metadata record Python 3.12.3, PyTorch 2.13.0+cu130, CUDA 13.0, Diffusers 0.39.0, and Transformers 5.15.0 on the NVIDIA GB10; these are the executed versions, rather than the earlier probe's unfulfilled x86 software pins. The gradient-fitting control records NumPy 2.5.2, and the scale control additionally records SciPy 1.18.0. For the original $T = 21$ shared-weight analysis, pointwise LPs use SciPy's HiGHS interface; shared MILPs use a 1,800-second cap and zero requested relative MILP gap. Unspecified solver tolerances retain that run's library defaults. A time-limited solve therefore reports a bracket, even when zero gap was requested. Later controls use their stated caps (including 120 seconds for fit-half incumbents); these are not interchangeable with the original cap. The separate rational verification uses SCIP 10 and VIPR as detailed in Appendix G. The gradient and convex surrogate fitting settings are given with their results in Appendix N.

Reproduction scope. Reanalysing a recorded table requires its cell losses, ordered image and class manifests, reference normalisation, and the analysis version identified by the package's

hashes. Regenerating those losses additionally requires the executed scorer, model and data revisions, preprocessing order, cell seeds, precision, and batching; matching a master seed alone does not establish bitwise replay across software or hardware changes. The release includes analysis sources, configuration and verification records and, for every exact row, the exported LP model, the VIPR certificate, the exact rational pointwise witnesses, a verifier (`verify_certificates.py`) that checks the certificate's constraint set against the LP and runs `viprchk`, and a script (`recompute_witnesses.py`) that recovers O_τ and a shared lower witness with $L = K - 1$ from the LP alone, verified in exact rational arithmetic; Table 30 lists the recomputed endpoints and run times for all eight released tables. Measured loss caches are not distributed; the LP coefficients, finite decimals of the stored binary64 contrasts, are the released form of each certified table. Public evaluation of other rows recomputes floating-point bounds. The clean SD scorer evaluates one noise cell at a time, whereas the original harness groups trials in chunks of eight. The original batch settings above describe the recorded measurements, not a promise of bitwise identity with the public evaluator.

Table 30: **Independent recomputation of every released exact table from its LP model.** Pointwise witnesses close O_τ exactly; a floating MILP incumbent, refined by floating max-min-margin LPs on its selected set and recounted in rational arithmetic, attains $L = K - 1$, matching the VIPR certificate $G^\star_\tau \le K - 1$. Times are elapsed seconds for the witness recomputation in the recorded runs; `viprchk` verifies each certificate in under two seconds.

Table	T	K	O_τ	L	$G^\star_\tau$	s
SD 1.5 Pet, original	21	87	97	86	86	2.5
SD 1.5 CIFAR-10, original	21	82	91	81	81	2.4
SD 1.5 CIFAR-10, panel C	21	90	97	89	89	1.1
SD 1.5 CIFAR-10, panel C, dense (E)	101	93	99	92	92	5.4
SD 1.5 CIFAR-10, PIL ℓ_1	21	98	100	97	97	0.8
SD 1.5 CIFAR-10, PIL ℓ_1, fresh A	21	98	99	97	97	0.4
SD 1.5 CIFAR-10, PIL ℓ_1, fresh B	21	98	100	97	97	0.7
SD 1.5 CIFAR-10, PIL ℓ_1, pooled	21	97	99	96	96	0.9

Registration status. Experiments registered before measurement, studies whose analysis plans were fixed during measurement, and post-hoc analyses are identified separately in Table 29 and the relevant experiments. The original registration fixes the data, image selection, timestep grid, trial counts, split-half estimator, agreement and saturation thresholds, stopping rule and artifact schema before measurement. Dated amendments and original records accompany the code. For PixArt, measurement began before the analysis plan was fixed, but results were not examined until afterwards; it is not a pre-measurement registration.

Two thresholds in these appendices are close to their boundaries and we report the margin rather than only whether the criterion is met: $R^\star = 32$ rests on $96/100$, whose Wilson lower bound is 0.9016 against a threshold of 0.90; and the $N = 12$ allocation match is decided by the preregistered tie-break rule from an exact tie at $145/185$, so a tie-break in the other direction would have made the count 1 of 3.

Scale of the empirical runs. All runs are on a single NVIDIA GB10 system. The Pet noise-stability run totals $1{,}420{,}800$ UNet forward passes over two equal $R{=}32$ and $R{=}64$-extension stages (each $710{,}400$; 4.50 h at a measured 43.8/s per stage) and its CIFAR-10 replication $384{,}000$ (2.71 h at 39.4/s); the allocation grid is $568{,}135$ (3.60 h); the compute-matched policy comparison on $5{,}000$ CIFAR-10 images is $1{,}800{,}000$ at $N = 12$ with a further $3{,}600{,}000$ at $N = 24$ and $720{,}000$ for the full-range grid; the precision comparison and figure material are small. The reachability and single-weighting measurements use the cached losses and no GPU. The runs enumerated in this paragraph total roughly 8.5 million forward passes; the later resolution ladder and Jacobian probes are reported separately. The PixArt-α replication adds $384{,}000$ class-conditioned Transformer evaluations over 27.7 h.

Scoring implementation. All audited fp16 SD table runs import and SHA-check `sd15_diffusion_classifier_probe.py`. The production harness reuses its prompt construction, preprocessing, model loading, scheduler, and cell-loss expression, while changing (i) GPU discovery, (ii) the outer scorer from the trial-averaged `score_image` to a cellwise cache that retains every (t, r, c) loss, and (iii) noise generation to per-cell `blake2b(seed | image | timestep | trial)` seeds. For Pet it additionally batches all 37 classes at once rather than the original scorer default of 10; the artifacts state that this can change low-order bits. The E1 and nested-support artifacts record these deviations under `adapter_deviations`; later SD packages record hashes of the fixed scorer and shared harness instead of repeating that field. The per-cell seeds are what make the caches nested across R; without them a run at $R = 4$ would not contain the run at $R = 2$ and no split-half curve would be comparable across R. PixArt requires a separate pinned runner because the denoiser, text encoder, scheduler, and output layout differ; its analyser instead hash-checks that runner and all registered estimator sources before opening the completed cache.

Theory. Deterministic scripts regression-test small instances and the algebraic identities; the proofs establish the general statements. The algebraic checks use no GPU and no random seed that affects their conclusions.

R PROOFS

R.1 CONNECTIVITY OF THE ACTIVE GRAPH

Lemma 13 (Connectivity)**.** *For distinct $\mu_1, \dots, \mu_C$ the graph whose edges are the active pairs is connected. Hence the active contrasts span S, all means lie in a common coset $\mu_1 + S$, and a metric block-diagonal with respect to $S \oplus S^\perp$ has its $S^\perp$ block contribute the same constant to every class.*

Proof. The cells $V_1, \dots, V_C$ are closed convex sets covering $\mathbb{R}^d$ with pairwise disjoint interiors. Put

$$N = \bigcup_{\{i,j\} \notin E} (V_i \cap V_j) \;\cup\; \bigcup_{i<j<l} (V_i \cap V_j \cap V_l),$$

where E is the set of active pairs. If $\{i, j\} \notin E$ then $V_i \cap V_j$ is a convex subset of the bisecting hyperplane H_{ij} that is not $(d-1)$-dimensional, hence has affine dimension at most $d-2$. A triple intersection lies in $H_{ij} \cap H_{il}$; for $j \neq l$ these hyperplanes are distinct, so the intersection is either empty or an affine subspace of dimension $d-2$. Thus N is contained in a finite union of affine subspaces of codimension at least 2.

Each μ_i is strictly closer to itself than to any other centre, so $\mu_i \in \operatorname{int} V_i$ and every cell has nonempty interior. Fix $i \neq j$ and choose $p \in \operatorname{int} V_i$ outside the finite union of affine subspaces covering N; such a point exists because that union is Lebesgue-null. For each of these subspaces A, we have $p \notin A$, and the set of points q for which the segment $[p, q]$ meets A is contained in the cone over A with apex p, which has dimension at most $(d-2)+1 = d-1$ and is therefore Lebesgue-null; the same holds for the finite union, and for the finitely many q making $[p, q]$ lie inside one of the bisecting hyperplanes. Since $\operatorname{int} V_j$ is nonempty and open, we may choose $q \in \operatorname{int} V_j$ with $[p, q] \cap N = \varnothing$ and with $[p, q]$ contained in no bisecting hyperplane.

The Voronoi diagram is the polyhedral complex cut out by the finitely many hyperplanes H_{ab}, and the segment meets each in at most one point, so $[p, q]$ passes through finitely many cells. Every transition point avoids N, hence belongs to exactly two cells and to no triple intersection, and the corresponding pair is active. Reading the transitions in order gives a path $i = k_0, k_1, \dots, k_r = j$ in the active graph. As i, j were arbitrary the graph is connected, and $\mu_i - \mu_j = \sum_s (\mu_{k_s} - \mu_{k_{s+1}})$ exhibits every contrast as a sum of active contrasts, so the active contrasts span S.

For the last clause, write $\mu_c = \mu_1 + \sigma_c$ with $\sigma_c \in S$ and split $x = x_S + x_0$ along $S \oplus S^\perp$. If B is block-diagonal the cross terms vanish and $(x - \mu_c)^\top B (x - \mu_c)$ equals a term depending on c only through x_S, plus $(x_0 - \mu_{1,0})^\top B_0 (x_0 - \mu_{1,0})$, which is the same for every c. □

R.2 THEOREM 1 (THE BAYES-PRESERVING SET)

Proof. All means have the same $S^\perp$ component $u = P_{S^\perp}\mu_i$. Translating both x and every μ_i by $-u$ leaves all differences $x - \mu_i$ unchanged, so assume without loss of generality that $\mu_i \in S$. Let $B \succ 0$ be the metric under test (for a weighting, $B = B_w$) and $m_{ij} = \frac{1}{2}(\mu_i + \mu_j)$.

Necessity. Cells are convex polyhedra with nonempty interior, so almost-everywhere agreement forces exact equality of every cell. If (i, j) is active, the common facet is $(d-1)$-dimensional and lies in both full-space bisecting hyperplanes, which therefore coincide; their normals give $B\delta_{ij} = \alpha_{ij}\delta_{ij}$ with $\alpha_{ij} = \delta_{ij}^\top B\delta_{ij}/\|\delta_{ij}\|^2 > 0$. The offsets match because $\mu_i^\top B\mu_i - \mu_j^\top B\mu_j = 2m_{ij}^\top B\delta_{ij}$. The active contrasts span S by Lemma 13, hence $BS \subseteq S$; symmetry of B then also makes $S^\perp$ invariant, so the S–$S^\perp$ cross block vanishes. We may now set $T = B|_S$. Symmetry of T gives $\alpha_e\langle\delta_e, \delta_f\rangle = \alpha_f\langle\delta_e, \delta_f\rangle$, so non-orthogonal active contrasts share an eigenvalue and T is the scalar α_a on each S_a; the remaining $S^\perp$ block is an arbitrary positive-definite B_0.

Sufficiency. For $B = \bigoplus_a \alpha_a I_{S_a} \oplus B_0 \in \mathcal{E}$, set $T = B|_S$. For active (i, j) we have $B\delta_{ij} = \alpha_a\delta_{ij}$ with $\alpha_a > 0$, and scaling a half-space normal by a positive number leaves the half-space unchanged, so the constraint is *identical* to the Euclidean one. A non-redundant Voronoi constraint is exactly one that defines a facet, i.e. exactly an active pair, so the Euclidean cell is cut out by the active constraints alone. Intersecting all constraints in B-form and discarding the inactive ones gives $V_i^B \subseteq V_i$ for every i. Both families cover $\mathbb{R}^d$ with pairwise disjoint interiors, so no cell can shrink, and $V_i^B = V_i$. □

Inactive contrasts. For $B \in \mathcal{E}$, write an inactive contrast as $\delta_{ij} = \sum_a \delta^{(a)}$ with $\delta^{(a)} \in S_a$. Then $B\delta_{ij} = \sum_a \alpha_a\delta^{(a)}$ need not be parallel to δ_{ij}. Nevertheless, the metric preserves every facet-defining constraint and moves only redundant ones; no margin condition on inactive pairs is required.

R.3 GENERICITY IS ONE POLYNOMIAL: THE THEOREM BEHIND "GENERIC"

Fix the whitened means, hence the orthogonal decomposition $\mathbb{R}^d = S_1 \oplus \cdots \oplus S_m \oplus S_0$ and the subspace $\mathfrak{E}$; write $D = \dim \mathrm{Sym}(d)$ and $q = \operatorname{codim} \mathfrak{E} = D - \dim \mathfrak{E}$. Fix an eigenvalue-multiplicity pattern $\mathbf{d} = (d_1, \ldots, d_p)$ for Σ_q, with p distinct eigenvalues, and let $I_1, \ldots, I_p$ be the coordinate projectors onto consecutive blocks of sizes d_s. A pair (Σ_p, Σ_q) with $\Sigma_q = Q\,\mathrm{diag}_{\mathbf{d}}(\lambda)Q^\top$ determines $H := \Sigma_p^{1/2}Q \in GL(d)$ up to the choice of Q, which is unique only up to right multiplication by $O(\mathbf{d}) := \prod_s O(d_s)$; but $HI_sH^\top$ is invariant under $H \mapsto HR$ for $R \in O(\mathbf{d})$, so $L_{\text{eff}} = \{\sum_s n_s HI_sH^\top : n \in \mathbb{R}^p\}$, W, and $\mathcal{G}$ below are well-defined functions of (Σ_p, Σ_q) (for distinct eigenvalues $p = d$ and $L_{\text{eff}} = L$ of Section 5.2). Conversely every class $HO(\mathbf{d})$ arises: writing $H = PU$ (polar), $(\Sigma_p, \Sigma_q) = (P^2,\ U\,\mathrm{diag}_{\mathbf{d}}(\lambda)U^\top)$ for any admissible λ. Genericity is stated for $H \in GL(d)$, where $\mathcal{G}$ is a polynomial, and transferred as follows: the parameter space $\mathrm{SPD} \times \mathrm{SPD}_{\mathbf{d}}$ is a smooth manifold, the map $\Phi : (\Sigma_p, \Sigma_q) \mapsto \{HI_sH^\top\}_s$ into $\mathrm{Sym}(d)^p$ is smooth and $O(\mathbf{d})$-invariant, and $\mathcal{G}$ factors through it: since the rows of W are linear in the tuple $\{HI_sH^\top\}_s$, $\mathcal{G} = \widehat{\mathcal{G}}(\{HI_sH^\top\}_s)$ for a polynomial $\widehat{\mathcal{G}}$ on $\mathrm{Sym}(d)^p$; the exceptional set in $\mathrm{SPD} \times \mathrm{SPD}_{\mathbf{d}}$ is Φ^{-1} of the exceptional set of a polynomial that is not identically zero on $GL(d)$, hence not identically zero on the image of Φ (which contains $\{HI_sH^\top\}$ for every $H \in GL(d)$), hence a proper analytic subset of a connected manifold and therefore Lebesgue-null and nowhere dense there. Algebraicity holds in H-space; only nullity is transferred. Let $\pi : \mathrm{Sym}(d) \to \mathrm{Sym}(d)/\mathfrak{E} \cong \mathbb{R}^q$ be the quotient map and let $W(H) \in \mathbb{R}^{p\times q}$ have s-th row $\pi(HI_sH^\top)$ — concretely, the components of $HI_sH^\top$ on the off-block entries between distinct blocks and on the traceless part of each diagonal block S_a, $a \geq 1$; its entries are quadratic polynomials in H. Put $t = \min(p, q)$.

Theorem 14 (Explicit genericity)**.** *(a) For $H \in GL(d)$, $\dim(L_{\text{eff}} \cap \mathfrak{E}) = p - \operatorname{rank} W(H)$; hence $\dim(L_{\text{eff}} \cap \mathfrak{E}) = \max(0, \Delta_{\text{eff}})$ with $\Delta_{\text{eff}} = p - q$ if and only if $\operatorname{rank} W(H) = t$. (b) The set $\{H : \operatorname{rank} W(H) < t\}$ is a proper Zariski-closed subset of $\mathbb{R}^{d\times d}$; equivalently the single polynomial $\mathcal{G}(H) = \det(WW^\top)$ (for $p \leq q$; $\det(W^\top W)$ for $p > q$) is not identically zero. (c) Hence $\{\mathcal{G} = 0\}$ has Lebesgue measure zero and empty interior in $GL(d)$, and its $O(\mathbf{d})$-invariant preimage in (Σ_p, Σ_q) is Lebesgue-null and nowhere dense there.*

Proof. (a) The $HI_sH^\top$ are linearly independent: $\sum_s n_s HI_sH^\top = 0$ gives, after multiplying by H^{-1} and $H^{-\top}$, $\sum_s n_s I_s = 0$, so $n = 0$. Thus $T_H : n \mapsto \sum_s n_s HI_sH^\top$ is injective, $L_{\text{eff}} \cap \mathfrak{E} = T_H(\ker(\pi \circ T_H))$, and $\dim \ker(\pi \circ T_H) = p - \operatorname{rank} W(H)$. Since $\operatorname{rank} W \le t$ always, $p - \operatorname{rank} W \ge p - t = \max(0, p - q)$, with equality iff $\operatorname{rank} W = t$.

(b) By polarisation $\{hh^\top : h \in \mathbb{R}^d\}$ spans $\mathrm{Sym}(d)$, and π is onto, so $\{\pi(hh^\top)\}$ spans the q-dimensional quotient; choose $h_1, \dots, h_t$ with $\pi(h_1h_1^\top), \dots, \pi(h_th_t^\top)$ independent. Let H^0 have first column of block s equal to h_s for $s \le t$ and all other entries zero, so $H^0 I_s (H^0)^\top = h_s h_s^\top$ for $s \le t$ and $\operatorname{rank} W(H^0) = t$. H^0 need not be invertible, but $U_1 = \{H : \operatorname{rank} W(H) \ge t\}$ and $U_2 = \{H : \det H \ne 0\}$ are nonempty Zariski-open sets ($H^0 \in U_1$, $I \in U_2$) in the irreducible space $\mathbb{R}^{d\times d}$, hence dense, hence they meet; any $H^* \in U_1 \cap U_2$ is an invertible point with $\operatorname{rank} W(H^*) = t$ — its existence is what the argument delivers; we do not construct it in closed form. So some $t \times t$ minor of W is a nonzero polynomial in H, and $\{\operatorname{rank} W < t\}$ lies in its zero set. By Cauchy–Binet $\mathcal{G}$ is the sum of squares of the maximal minors, so $\mathcal{G} \ne 0$ iff some maximal minor is nonzero.

(c) The real zero set of a nonzero polynomial has Lebesgue measure zero and empty interior. For the transfer, $\widehat{\mathcal{G}} \circ \Phi$ is a real-analytic function on the connected manifold $\mathrm{SPD} \times \mathrm{SPD}_{\mathbf{d}}$ that is not identically zero (its value at the point corresponding to any H^* with $\mathcal{G}(H^*) \ne 0$ is nonzero), and the zero set of a non-identically-zero real-analytic function on a connected manifold is Lebesgue-null with empty interior. □

Two remarks. The statement is in *whitened* coordinates: the blocks come from the active facets of $\mu_c = \Sigma_p^{-1/2}\nu_c$, so with raw means fixed $\mathfrak{E}$ would move with Σ_p and "generic in (Σ_p, Σ_q)" would not be well-posed; fixing whitened means removes the circularity, and, for each fixed Σ_p, whitening is a diffeomorphism in the raw means; a Fubini argument therefore transfers the conclusion to almost every raw configuration. And the exceptional set is null but not empty, and not uniformly harmless: it contains points where $\Sigma_p \in \mathcal{E}$ already, and also points with $\Sigma_p \notin \mathfrak{E}$ where a finite nonnegative weighting is exactly Bayes-preserving; we claim nothing about it beyond its measure. Multiplicity only helps the impossibility: $\Delta_{\text{eff}} = \Delta - (d - p) \le \Delta$.

Finite schedules and the well-specified case. With G_σ as in Appendix O, an actual schedule reaches only $\mathrm{cone}\{G_{\sigma_t} : 1 \le t \le T\}$, whose span lies in L_{eff} and has dimension at most $\min(T, p)$. The continuous spectral space and cone are therefore favourable relaxations: failure of either applicable relaxation to intersect the preserving set rules out every finite schedule, which may face additional obstructions. The commutator obstruction of Corollary 3 vanishes in the well-specified case $\Sigma_q = \Sigma_p$, without implying that every prescribed finite schedule reaches the Bayes metric. In the regime $k \ge 2$, $d \ge 3$ of Corollary 2, that case lies inside the exceptional set: the continuous ELBO weighting measure $d\sigma/\sigma$ gives $f_w(\lambda) = 1/(2\lambda)$, so $M_w = \frac{1}{2}\Sigma_q^{-1}$ has eigenvalue vector in $\overline{\mathcal{C}}$, and $\Sigma_q = \Sigma_p$ gives $B_w = \frac{1}{2}I \in \mathcal{E}$, forcing $\mathcal{G} = 0$.

R.4 Corollary 2 (which triples admit $\Delta > 0$)

Proof. The blocks $S_1, \dots, S_m$ are mutually orthogonal and each has dimension at least one, so $m \le k$, and $k \le d$; admissible triples are exactly $1 \le m \le k \le d$. Write $\Delta = -d(k-1) + \binom{k}{2} + m$.

If $k = 1$ then $\Delta = m \ge 1 > 0$: this is the collinear family, for every d. Let $k \ge 2$. Using $m \le k$ and $d \ge k$,

$$\Delta \;\le\; -k(k-1) + \tfrac{k(k-1)}{2} + k \;=\; \tfrac{k(3-k)}{2}.$$

For $k \ge 4$ the right-hand side is negative, and for $k = 3$ it is 0, so $\Delta > 0$ forces $k = 2$. With $k = 2$ we have $\Delta = 1 + m - d$ and $m \le 2$, hence $\Delta \le 3 - d$, so $\Delta > 0$ forces $d = 2$ and then $\Delta = m - 1 > 0$ forces $m = 2$. The only triple with $k \ge 2$ and $\Delta > 0$ is therefore $(d, k, m) = (2, 2, 2)$.

The consequence is immediate: for $k \ge 2$ and $d \ge 3$ the count gives $\Delta \le 0$, and under $\mathcal{G} \ne 0$ the transversality estimate is an equality, so $\dim(L \cap \mathfrak{E}) = 0$; since $\mathfrak{E}$ is a linear space and $\mathcal{E}$ is contained in it minus the boundary, no nonzero element of L is Bayes-preserving. □

R.5 Proposition 4 (nonnegative-cone obstruction)

We first record the two-dimensional description of $\mathcal{C}$ that the proposition uses.

Lemma 15. *Let $d = 2$ and let Σ_q have eigenvalues $\lambda_1 < \lambda_2$ in the basis Q. The set of ratios n_1/n_2 realised by $n \in \mathcal{C} \setminus \{0\}$ is exactly the* open *interval $(1, \mathrm{cond}(\Sigma_q)^2)$; its closure $[1, \mathrm{cond}(\Sigma_q)^2]$ is realised by $\overline{\mathcal{C}}$, the endpoints corresponding to the limits $\sigma \to \infty$ and $\sigma \to 0$.*

Proof. For a single generator, $v(\sigma)_1/v(\sigma)_2 = \big((\sigma^2 + \lambda_2)/(\sigma^2 + \lambda_1)\big)^2$, which is continuous and strictly decreasing in σ^2, with limits $\mathrm{cond}(\Sigma_q)^2$ as $\sigma \to 0$ and 1 as $\sigma \to \infty$, neither attained at finite $\sigma > 0$. So the generators realise exactly the open interval. A conic combination $n = \sum_t c_t v(\sigma_t)$ with $c_t \geq 0$, not all zero, has n_1/n_2 equal to a weighted average of the individual ratios, with weights proportional to $c_t v(\sigma_t)_2$. Each positively weighted ratio lies in $(1, \mathrm{cond}(\Sigma_q)^2)$, so their average does too. The same argument applies to any nonzero integral combination with finite components. Hence $\mathcal{C}$ realises exactly $(1, \mathrm{cond}(\Sigma_q)^2)$, and taking closure gives the closed interval. Where the closed interval $[1, \mathrm{cond}(\Sigma_q)^2]$ is quoted, it refers to $\overline{\mathcal{C}}$; the reachable-fraction computations are unaffected because the endpoint set has measure zero in ψ. □

Proof of Proposition 4. Take the square family: the four whitened centres are the corners of a unit square whose axes $u_1(\psi), u_2(\psi)$ are rotated by ψ, so $k = 2$ and the two orthogonal axes give $m = 2$; hence $\Delta = 2(1-2) + 1 + 2 = 1 > 0$ for every ψ and the linear count never obstructs.

By Theorem 1, $B_w = \Sigma_p^{1/2} M_w \Sigma_p^{1/2}$ is Bayes-preserving exactly when it acts as a positive scalar on each axis, i.e. when u_1, u_2 are eigenvectors of B_w, i.e. when $u_1^\top B_w u_2 = 0$. Writing $A = \Sigma_p^{1/2}$ and $M_w = Q\,\mathrm{diag}(n)Q^\top$, this is the single linear equation

$$\sum_k n_k \,(Q^\top A u_1)_k \,(Q^\top A u_2)_k = 0 \tag{25}$$

in n — one equation in $d = 2$ unknowns, so wherever its coefficient vector is nonzero its solution set is a line through the origin (at most one ray of it lies in the positive quadrant), which is exactly $\dim(L \cap \mathfrak{E}) = \Delta = 1$. The two coefficients are $(Q^\top A u_1)_k (Q^\top A u_2)_k$, $k = 1, 2$; each is a trigonometric polynomial of degree 2 in ψ that is not identically zero for the instance below ($Q = I$ and $A_{k1} A_{k2} \neq 0$ at $\psi = 0$), so each has finitely many zeros on $[0, \pi/2)$, and we exclude that finite set. Solving equation 25 gives the required ratio

$$r(\psi) = \frac{n_1}{n_2} = -\,\frac{(Q^\top A u_1)_2 \,(Q^\top A u_2)_2}{(Q^\top A u_1)_1 \,(Q^\top A u_2)_1},$$

a continuous function of ψ off its poles. A Bayes-preserving metric is reachable at ψ if and only if $r(\psi) > 0$ — otherwise the ray carries no positive-definite metric at all — and $r(\psi)$ lies in the interval of Lemma 15. In particular, $0 < r(\psi) < 1$ gives a positive-definite Bayes-preserving metric in L, but none reachable by nonnegative weighting.

It remains to exhibit an instance where that set has positive measure, which needs only one angle and continuity. Take $\Sigma_q = \mathrm{diag}(0.7, 2.1)$, so $\mathrm{cond}(\Sigma_q)^2 = 9$, and, in the same coordinates, $\Sigma_p = \left(\begin{smallmatrix} 1.6 & 0.55 \\ 0.55 & 0.9 \end{smallmatrix}\right)$. At $\psi_0 = 0.7$, interval arithmetic (the 2×2 SPD square root has the closed form $(\Sigma_p + \sqrt{\det \Sigma_p}\, I)/\sqrt{\mathrm{tr}\,\Sigma_p + 2\sqrt{\det \Sigma_p}}$, so every step is a validated $+, -, \times, /, \sqrt{\cdot}, \cos, \sin$) encloses the denominator in $[-0.67022, -0.67021]$, bounded away from 0. It encloses $r(\psi_0)$ in an interval of width below 10^{-49} wholly inside $(0, 1)$; the value rounds to 0.6248587 to seven decimal places. The certificate script prints the interval endpoints at fifty-digit precision. With this covariance pair fixed, continuity gives $0 < r(\psi) < 1$ on a nonempty open interval $I \subset (0, \pi/2)$ around ψ_0. Thus L admits a positive-definite preserving metric throughout I, but the reachable cone admits none. This proves the claimed positive angle measure. The percentages we quote for this instance — reachable 17.33%; 30.31% with no positive-definite solution, 50.68% with $r < 1$, 1.68% with $r > \mathrm{cond}(\Sigma_q)^2$ — are numerical root-finding results (the reachable fraction agrees with a 4000-point grid to 6×10^{-5}) and are reported as measurements, not as part of the proof. □

Remark 16. The split above is what makes the statement quantitative: of the 82.67% that fails, only 1.68 percentage points sit above the anisotropy ceiling and could be bought with a larger-anisotropy

model, while the 50.68 points below $r = 1$ are blocked by the $\sigma \to \infty$ Euclidean end of the curve, which does not move with $\mathrm{cond}(\Sigma_q)$. Figure 4(b) shows the same shape at a different instance, where the budget sweep saturates strictly below one. Figure 8 gives a separate exact cone-only control whose required coefficient ratio lies below this same interval (Appendix O).

S SEPARATING PARTITION DISTANCE FROM FINITE-TABLE COUPLING

The distance $\delta_{T,\star}$ concerns preservation of a population Bayes partition. The gap $O_\tau - G^\star_\tau$ concerns simultaneous strict decisions on a labelled finite table. Neither quantity's positivity determines the other's, even for expected Gaussian losses. The following examples concern noise expectations, not finite-R denoiser measurements.

Lemma 17 (A common successful atom has no shared cost)**.** *On a finite contrast table at threshold $\tau \geq 0$, suppose one timestep $t_\star$ satisfies $\min_j D_{it_\star j} > \tau$ for every image counted by O_τ. Then the simplex vertex $e_{t_\star}$ attains $G^\star_\tau = O_\tau$.*

Proof. The common vertex attains all O_τ successes, and $G^\star_\tau \leq O_\tau$ by definition. □

One sufficient condition is $D_{itj} = a_t b_{ij}$ with common $a_t > 0$: an image is reachable only if $\min_j b_{ij} > 0$, and the atom maximising a_t simultaneously maximises all such margins. At $\tau > 0$, other weightings can still have different strict counts.

Positive partition distance with zero shared cost. Take $d = 2$, $\Sigma_p = \mathrm{diag}(2, 1)$, $\Sigma_q = I$, and whitened means $0, e_1, e_2$. Their active-contrast graph is connected, and their mean differences span $\mathbb{R}^2$, so $\mathcal{E}_2 = \{I\}$. Every expected-loss metric in whitened coordinates has the form $G_\sigma = [\sigma^2/(\sigma^2 + 1)^2]\Sigma_p$, hence $A_\sigma = \mathrm{diag}(4/3, 2/3)$ and $\delta_{T,\star} = \delta_\star = \sqrt{2}/3 > 0$ for every nonempty finite grid. Expected rival-minus-truth losses lie on a common positive ray, so Lemma 17 gives zero shared cost on every finite labelled panel. Per-image positive normalisation preserves the factorisation. Partition misspecification need not cause conflicts between weighting choices.

Zero partition distance with positive shared cost. Conversely, take $\Sigma_p = \Sigma_q = \mathrm{diag}(1, 4)$, whitened means $\pm(1, 1)$, and $\sigma_1^2 = 1$, $\sigma_2^2 = 2$. Then the whitened metrics are $G_{\sigma_1} = \mathrm{diag}(1/4, 4/25)$ and $G_{\sigma_2} = (2/9)I$, so $\delta_{T,\star} = 0$. Consider the labelled whitened points $x_+ = (1, -5/4)$ with label $+$ and $x_- = (1, -6/5)$ with label $-$. The two expected squared-loss contrasts, divided by $d = 2$, are respectively $(1/10, -1/9)$ and $(-29/250, 4/45)$ across the two timesteps. After the per-image maximum-absolute normalisation of Section 4, these are $(9/10, -1)$ and $(-1, 200/261)$. Writing $\theta = w_1$, strict correctness at $\tau = 0$ requires $\theta > 10/19$ and $\theta < 200/461$, respectively. The intervals are disjoint and shrink for $\tau > 0$, while each point remains individually reachable for $0 \leq \tau < 200/261$. Thus $O_\tau = 2$ and $G^\star_\tau = 1$ throughout that range. A Bayes-preserving atom need not fit every realised label in an overlapping Gaussian classification problem.

These examples establish logical independence of the two positive-obstruction statements; they do not validate the bridge to a real denoiser.

T A CONE-ONLY OBSTRUCTION ON AN OPEN TWO-CLASS FAMILY

Proposition 4 measures angles inside a prescribed square family. The following separate construction shows that a cone-only obstruction also persists on an open set when $d = C = 2$; it makes no full-dimensional claim for the four-centre square stratum.

Proposition 18. *There is a nonempty open set of equal-prior, two-class planar Gaussian configurations for which the continuous spectral linear space contains a positive-definite Bayes-preserving metric, but the nonnegative spectral cone contains none. Here $k = m = 1$ and $\Delta = 1$.*

Proof. Start with whitened means $\pm\delta/2$, $\delta = (1, 1)$, $\Sigma_p = \mathrm{diag}(4, 1/4)$ and $\Sigma_q = \mathrm{diag}(1, 2)$, so $A = \Sigma_p^{1/2} = \mathrm{diag}(2, 1/2)$ and $Q = I$. The spectral linear space contains $M = \Sigma_p^{-1} = \mathrm{diag}(1/4, 4)$, giving the Bayes-preserving whitened metric $B = AMA = I$. For a spectral metric

$M = \mathrm{diag}(n_1, n_2)$, Theorem 1 requires $B\delta = \alpha\delta$ with $\alpha > 0$, hence $(n_1, n_2) = (\alpha/4, 4\alpha)$ and $n_1/n_2 = 1/16$. But the positive atoms have ratio $((\sigma^2+2)/(\sigma^2+1))^2 \in (1, 4)$. Every nonzero nonnegative combination has ratio greater than one, so no such combination is Bayes-preserving, even with unrestricted positive noise scales.

To check openness rather than only one example, order the distinct eigenvalues of Σ_q as $\lambda_1 < \lambda_2$ and choose its eigenbasis Q continuously near this instance. Define $a = Q^\top A\delta$ and $b = Q^\top A^{-1}\delta$. The condition $B\delta = \alpha\delta$ is $n_k a_k = \alpha b_k$. At the displayed instance both ratios b_k/a_k are positive and $(b_1/a_1)/(b_2/a_2) = 1/16 < 1$. Nonzero denominators, positivity, the strict ratio inequality, and distinct ordered eigenvalues persist on a neighbourhood of $(\Sigma_p, \Sigma_q, \delta)$. There $n_k = b_k/a_k > 0$ defines a positive-definite Bayes-preserving metric in the spectral linear space, whereas every nonnegative spectral combination still has $n_1/n_2 > 1$. For two distinct eigenvalues the continuous atoms span the full diagonal spectral space (two distinct positive noise scales already suffice). Thus the obstruction persists on an open set of the fixed $d = C = 2$ parameter space, also after arbitrary small changes of the common centre or the corresponding raw means. It has positive Lebesgue measure in that space, not merely positive angle measure in the square family. □

U TWO SELECTION REGIMES FOR SMOOTH RECONSTRUCTION CONTROLS

This post-exposure descriptive follow-up asks whether the finite-table behaviour in Appendix F distinguishes spectral reconstruction from two smooth comparators under matched selection freedom. It uses the same 100 images per dataset, six timesteps and even/odd $R = 32$ MSE noise halves. The original alignment and subsequent reconstruction results were already known; this is not a new preregistered test.

Each family chooses two functions from a fixed 25-function grid, giving 300 candidate bases. The spectral functions are $s/(s+\lambda)^2$ with $s = \sigma_t^2$ and 25 logarithmically spaced λ values from 10^{-2} to 10^2. The monomial functions are $x(s)^k$, where $x(s)$ rescales $\log s$ over the six timesteps to $[0.05, 1.05]$ and k ranges over 25 equally spaced values in $[0, 3]$. The power-law functions are s^a with 25 equally spaced exponents in $[-2, 2]$. These are the function families of the strict alignment control, but the present fitting and decision analysis are distinct.

In the *per-image* regime, each image selects its basis from even-noise all-pair class-contrast residuals. In the *shared-basis* regime, one basis per dataset minimises the sum of squared contrast residuals across all 100 even-noise images. Both regimes then fit centred class-loss coefficients separately for each image. Thus the second regime shares basis parameters, including λ for the spectral family, but does not impose a common Gaussian covariance or the full Gaussian score constraints. Neither regime holds out images; only the noise half is held out.

The comparison contains six measured-even, measured-odd and per-image spectral cells and ten additional comparator or shared-basis cells. All cells use the same even-reference scale, six-timestep simplex and strict $\tau = 10^{-6}$ as Appendix F. Each new shared MILP has a 600-second cap, followed by the same rational strict recount and conditional subset-margin repair rule. All new MILPs report floating optimality; the shared and gap brackets retain solver grade. Pointwise decisions resolve on all 100 images in all 16 cells. There is no fitted-model selection based on these gaps.

All three families produce positive gaps in both regimes on both datasets (Table 31). Spectral gaps are closer to the measured-odd gap on Pet, whereas the monomial comparator is closer on CIFAR-10. The selected shared-basis spectral and monomial incumbents have the same conditional miss-overlap counts, $5/6$ on Pet and $6/9$ on CIFAR-10, without implying equality of their missed sets. Their odd-noise hit counts are $71/64$ and $71/72$, respectively (Pet/CIFAR-10). The selected spectral incumbents thus score below uniform weighting ($78/72$) on both odd-noise tables. These observations give no consistent spectral advantage across the reported descriptive criteria. In these tables, the occurrence of a positive gap alone does not distinguish the spectral family from these two controls; this is not an equivalence test or evidence that every smooth family behaves alike. The Gaussian bridge remains unvalidated.

Released 100-row pointwise-status and selected-incumbent hit masks, together with a CPU utility, reproduce these overlaps without private inputs or witness payloads. Unresolved pointwise statuses would be excluded with coverage reported, never coded as unreachable.

Table 31: Smooth-family reconstruction controls. Shared and gap brackets are solver-grade, including singleton brackets. The measured-odd reference has $(A, O, L, U) = (78, 97, 84, 84)$ on Pet and $(72, 90, 78, 78)$ on CIFAR-10. "Odd hits" applies the selected reconstructed-table incumbent to measured-odd losses. "Miss overlap" divides the intersection of reachable-but-missed sets by the odd-reference missed set restricted to images reachable in both compared tables. This domain is defined separately for each pair; the fractions are descriptive and depend on potentially nonunique incumbent choices.

Dataset	Selection	Family	A	O	$[L, U]$	Gap	Odd hits	Miss overlap
Pet	Per-image	Spectral	75	87	$[76, 76]$	$[11, 11]$	75	$5/7$
Pet	Per-image	Monomial	75	86	$[76, 76]$	$[10, 10]$	78	$4/5$
Pet	Per-image	Power law	73	84	$[77, 77]$	$[7, 7]$	74	$2/4$
Pet	Shared basis	Spectral	74	86	$[75, 75]$	$[11, 11]$	71	$5/6$
Pet	Shared basis	Monomial	74	86	$[77, 77]$	$[9, 9]$	71	$5/6$
Pet	Shared basis	Power law	71	85	$[76, 76]$	$[9, 9]$	72	$2/5$
CIFAR-10	Per-image	Spectral	73	84	$[75, 75]$	$[9, 9]$	72	$5/8$
CIFAR-10	Per-image	Monomial	70	85	$[73, 73]$	$[12, 12]$	70	$6/9$
CIFAR-10	Per-image	Power law	73	82	$[75, 75]$	$[7, 7]$	71	$3/8$
CIFAR-10	Shared basis	Spectral	69	86	$[71, 71]$	$[15, 15]$	64	$6/9$
CIFAR-10	Shared basis	Monomial	72	86	$[72, 72]$	$[14, 14]$	72	$6/9$
CIFAR-10	Shared basis	Power law	71	83	$[73, 73]$	$[10, 10]$	60	$5/9$

V Error-aware transfer from reconstructed tables

The positive reconstructed-table gaps in Appendix U do not by themselves certify a gap on the measured table. We therefore perform a retrospective error-bound check on the same exposed six-timestep MSE tables, without selecting a new basis or changing the earlier fits. Let D_{itj} be the measured-odd rival-minus-truth contrasts and $\widehat{D}_{itj}$ an approximation, using the same measured-even per-image normalisation, timestep order, and rival order. Set

$$e_i = \max_{t,j} |D_{itj} - \widehat{D}_{itj}|.$$

For every simplex weight vector, each weighted contrast differs by at most e_i; the minimum over rivals has the same bound. Writing $O_{\widehat{D}}(\boldsymbol{\eta})$ and $G^*_{\widehat{D}}(\boldsymbol{\eta})$ for strict counts with image-specific thresholds, the resulting feasible-set inclusions give

$$O_{\widehat{D}}(\tau + \boldsymbol{e}) \le O_D(\tau) \le O_{\widehat{D}}(\tau - \boldsymbol{e}),$$
$$G^*_{\widehat{D}}(\tau + \boldsymbol{e}) \le G^*_D(\tau) \le G^*_{\widehat{D}}(\tau - \boldsymbol{e}).$$

Thus a stringent pointwise lower L^+_O and a lenient shared upper U^-_G yield $O_D - G^*_D \ge \max\{0, L^+_O - U^-_G\}$. This is an error-to-gap statement for these finite tables, not a bound from the Gaussian partition distance to an empirical gap.

We use $\tau = 10^{-6}$ and the stored binary64 contrasts. Errors are computed as exact rational differences and rounded upwards; $\tau + e_i$ is rounded upwards and $\tau - e_i$ downwards, retaining negative thresholds. Each pointwise LP proposal is recounted with exact rational simplex weights on both tables. Each shared MILP imposes the closed lenient constraints with outward-rounded coefficients, and has a 600-second cap. Its floating upper remains solver-grade; unresolved pointwise proposals contribute no lower witness. The three shared-basis tables are replayed byte for byte from the previously selected parameters, and the measured-even table serves as a noise baseline.

All eight transfer lower bounds are zero (Table 32); every lenient threshold is negative. The per-image maximum-entry error bound is too loose to transfer a positive gap, even for the measured-even noise baseline. This does not establish zero gap, isolate a spectral failure, or invalidate the directly measured gaps. It shows why similarity of reconstructed and measured gap magnitudes alone is insufficient for the claimed connection.

Table 32: Error-aware transfer on the exposed T6 tables. L_O^+ is a stringent pointwise lower and U_G^- a floating upper for the lenient shared problem. Every resulting gap lower is zero, which is uninformative about the actual gap. The error scale is the common normalised contrast scale, not an accuracy scale.

Dataset	Approximation	Median e_i	L_O^+	U_G^-	Gap lower
Pet	Shared spectral	0.259	7	100	0
Pet	Shared monomial	0.296	3	100	0
Pet	Shared power law	0.294	6	100	0
Pet	Measured even	0.153	21	100	0
CIFAR-10	Shared spectral	0.392	23	100	0
CIFAR-10	Shared monomial	0.431	18	100	0
CIFAR-10	Shared power law	0.420	14	100	0
CIFAR-10	Measured even	0.307	46	99	0

W SYNTHETIC EVALUATION OF A DIRECTIONAL-RESPONSE DIAGNOSTIC

We considered an additional, full-latent MSE diagnostic that avoids fitting a covariance or inverting a denoiser Jacobian. For a shared-covariance Gaussian score with covariance $\Sigma_q \succ 0$, put $s = \sigma^2$, $A_s = (\Sigma_q + sI)^{-1}$ and $d = \nu_c - \nu_b$ for a class pair. At an anchor s_0, its epsilon-output contrast determines the unit direction $v = A_{s_0} d / \|A_{s_0} d\|_2$, provided the denominator is nonzero. Let $F_s(x, h)$ be the difference of the two coordinate-mean MSE losses at clean input $x + hv$, using the same noise for both classes and both perturbations. For latent dimension p, the Gaussian affine score gives the finite-difference identity

$$q(s) = -\frac{F_s(x,h) - F_s(x,-h)}{4hs} = \frac{d^\top A_s^2 v}{p} = \sum_k \frac{a_k}{(\lambda_k + s)^2}, \qquad a_k \geq 0.$$

Indeed, in an eigenbasis of Σ_q with eigenvalues λ_k, $a_k = d_k^2 / [p \|A_{s_0} d\|_2 (\lambda_k + s_0)]$. The response is independent of input, paired noise, and nonzero h in this idealised model. Positivity, decrease and convexity in s are necessary properties, but do not identify this mechanism; $\exp(-s)$ also has them.

We prespecified a synthetic validation test to decide whether to measure real-denoiser responses (Table 33). It used nine scheduler timesteps, five for fitting and four interleaved held out, with ten synthetic source and thirty target images, four class pairs, and perturbation sizes $h = 0.01\|x\|_2$ and $2h$. A two-pole predictor selected one global pair from a fixed 25-point grid, with two nonnegative coefficients per class pair. An isotropic one-pole predictor and a two-term power-law predictor used fixed 25-point grids; the two-term families each had 325 candidate pairs including duplicates. Selection used source-only leave-one-timestep-out prediction error and source pair root-mean-square (RMS) normalisation. The held-out rule required both root-mean-square error (RMSE) ratios to be at most 0.90 and both paired image-bootstrap 95% interval upper endpoints below one. This criterion determined whether the restricted predictor was suitable for real-denoiser measurement; it was not a test of the full Gaussian family.

The 432 float64 identity checks passed with maximum scaled error 1.94×10^{-13}. However, three of 108 fp16 input/output-rounding checks exceeded the fixed tolerance $0.05 \max\{|q|, \text{source RMS}\}$ for response error or h-versus-$2h$ agreement. Together with the three-pole result, this fails the criterion for proceeding to real-denoiser measurement, so no such measurement was performed. The isotropic and nonlinear passes further illustrate that this finite-grid comparison does not identify anisotropy or a Gaussian mechanism. Failure here concerns the proposed numerical protocol and restricted predictor, not real-denoiser Gaussianity.

The released synthetic code supports recomputation without private image data.

Table 33: Synthetic validation of the directional-response predictor. Entries count instances passing the two-baseline prediction rule, not Gaussian detections or real-denoiser results. The two- and three-pole positive groups each required at least 11/12 passes; the latter failed. The separate fp16 arithmetic check also exceeded its tolerance.

Synthetic generating family	Prediction-rule passes
Shared Gaussian, two poles	12/12
Shared Gaussian, three poles	1/12
Isotropic Gaussian calibration	2/3
Exponential response	0/3
Class-dependent covariance	0/3
Input-nonlinear score	2/3
Timestep-dependent covariance	0/3

X Direct class-output contrasts and shared-covariance necessary conditions

Question and scope. The shared-covariance Gaussian model imposes restrictions on the denoiser's class-conditioned outputs that can be inspected before fitting a covariance. This diagnostic tests those restrictions at a specified operating point; it does not predict the empirical shared-weight gap or identify its cause. It is separate from the unsuccessful synthetic directional-response test in Appendix W. Unlike that proposal, it subtracts class-conditioned outputs directly, without a finite-difference step, a two-pole approximation, or a covariance inverse fit.

Necessary relations. Write $x_t = \sqrt{\alpha_t}x + \sqrt{1-\alpha_t}\epsilon$ and $s_t = (1-\alpha_t)/\alpha_t > 0$. For an exact epsilon predictor of a Gaussian law with common covariance $\Sigma_q \succ 0$ and class means m_c, the difference for classes c, b satisfies

$$u_s := -\frac{\epsilon_\theta(x_t, t, c) - \epsilon_\theta(x_t, t, b)}{\sqrt{s}} = (\Sigma_q + sI)^{-1}(m_c - m_b). \tag{26}$$

Thus u_s is independent of the input and sampled noise at fixed class pair and timestep. Fix an anchor s_0. Form columns $P_{cb,s} = u_s - u_{s_0}$ and $Q_{cb,s} = s_0 u_{s_0} - s u_s$ for every selected pair and non-anchor timestep. Subtracting $(\Sigma_q + sI)u_s = (\Sigma_q + s_0 I)u_{s_0}$ gives $\Sigma_q P = Q$; consequently $H = P^\top Q = P^\top \Sigma_q P$ is symmetric positive semidefinite, and $Pz = 0$ implies $Qz = 0$. These are necessary conditions, not an identification theorem for a real denoiser or a test of mean correctness. In particular, Q need not lie in the spatial span of P.

Fixed numerical summaries. We report the antisymmetry $\|H - H^\top\|_F/(\|P\|_F\|Q\|_F)$, the sum of negative eigenvalue magnitudes of $(H + H^\top)/2$ with the same normaliser, and $\|Q - QV_rV_r^\top\|_F/\max(\|Q\|_F, 10^{-30})$, where V_r contains right singular vectors of P with singular values above 10^{-8} times its largest. All singular values and both raw and source-scaled summaries are retained. No Gaussian acceptance threshold is selected from these results. Rank truncation is a numerical approximation: a nonzero truncated dependency residual alone is not a failure of the model. At full column rank the dependency condition is vacuous. The antisymmetry has a quantitative interpretation. For nonzero P, Q, write $a = \|H - H^\top\|_F/(\|P\|_F\|Q\|_F)$. For any symmetric candidate Σ_q, let $E = Q - \Sigma_q P$. Then $H - H^\top = P^\top E - E^\top P$, so $\|H - H^\top\|_F \le 2\|P\|_F\|E\|_F$ and hence $\|Q - \Sigma_q P\|_F/\|Q\|_F \ge a/2$. This bounds the relative secant residual for the observed matrix pair, not classification accuracy, a shared-weight gap, or generalisation. Its floating numerical evaluation is not an exact rational certificate.

Synthetic sensitivity. A fixed central processing unit (CPU) suite uses 24 shared-SPD instances in dimension 64 with 1, 2, 3, 8, 32, or 64 distinct eigenvalues, rotated and unrotated bases, and condition numbers up to 10^6. Twelve alternatives vary the covariance across classes, vary it across timesteps, or add a nonlinear score term. Each instance is evaluated at four additive observation-noise levels ($0, 0.001, 0.01, 0.1$ times its source anchor RMS), giving 144 reported rows. Noise

levels share a draw and are not independent replicates. For pooled means without pair scaling or added observation noise, shared-SPD antisymmetry is 1.73×10^{-16}–1.15×10^{-10} and negative mass is at most 3.09×10^{-16}, whereas the truncated dependency residual can reach 6.01×10^{-5}. The alternative families' antisymmetry ranges are respectively 0.028–0.175, 0.058–0.397, and 0.023–0.130. Timestep-dependent covariance still gives essentially zero input-constancy error: input constancy alone misses that alternative. These fixed controls demonstrate sensitivity and limitations, not a calibrated error rate or exhaustive separation from non-Gaussian models.

Real-output measurement protocol. We use the first 20 exposed images of the original SD 1.5 CIFAR-10 table, split into ten source and ten target images with one of each class in each half. Four disjoint class pairs, nine measured timesteps $\{250, 300, 350, 400, 500, 600, 650, 700, 750\}$, and eight fresh noise draws per image and timestep are fixed before measurement. The original pinned fp16 checkpoint and text embeddings are upcast to fp32; the UNet and noisy-input arithmetic use fp32 with TF32 disabled. The original tensor preprocessing and fp16 VAE posterior-mean latent are retained. The class-output subtraction and $\sqrt{s}$ normalisation use float64, followed by one fp32 storage rounding and float64 analysis. This is a distinct numerical operating point, not a remeasurement of the headline loss table or a claim about full-precision checkpoint weights.

For each pair, we freeze the coordinate RMS of its source-image/noise mean at $t = 500$, then apply that scalar at every timestep. The source means, scales, and hash are frozen before target UNet outputs are generated. The constancy summary is target-output RMSE about the frozen source mean, divided by this pair scale, averaging coordinates, pairs, timesteps, and draws within each image. Its descriptive 95% bootstrap interval resamples the ten target images while keeping the source reference fixed. Source/target crossed with first/last four noise draws give four mean tables; these and their pooled mean receive the Gram diagnostics. Noise-half differences provide a within-panel reference. All nine timesteps are directly measured: this is neither held-out-timestep prediction nor fresh-image population generalisation.

Observed deviations from the Gaussian relations. The target class-output contrasts deviate substantially from the fixed source means: the target constancy root-mean-square error (RMSE), normalised by the source anchor scale, is 7.5403 (descriptive image-bootstrap 95% interval $[6.4509, 8.4130]$). The measurement comprises 11,520 class-forwards. Replay reproduces 60 historical scalar losses bitwise, and all 24 selected repeat cells (48 additional class-forwards) satisfy the fixed fp32 repeatability rule. Results were independently verified.

Table 34 gives the source-scaled Gram summaries for all four split means and their pooled mean. Raw summaries, singular values, per-image errors, and source-target/noise-half comparisons are retained in the public aggregate artifact. Input/noise constancy and the exact Gram relations are not reproduced by these observed outputs. The pooled antisymmetry gives a floating lower bound 0.0710 on the relative secant residual for any symmetric covariance candidate. This bound concerns the observed mean table, with no population or classification interpretation. The checks concern necessary relations of one idealised score model at this numerical operating point. They neither predict the original loss-table gap nor validate the Gaussian bridge. Numerically unresolved covariance fits are reported in Appendix F; a separate target evaluation appears in Appendix Z.

Y Fresh-noise replication on Oxford-IIIT Pet

This experiment tests whether the shared-weight gap persists under independent noise on the original Pet image panel.

Pet independent noise. On the same 100 Oxford-IIIT Pet images and the original 21-timestep MSE grid, we repeated measurement with 64 new noise draws (global trials 64–127), retaining the original R64 normalisation and the threshold $\tau = 10^{-6}$. The fresh R64 table has uniform-pooling accuracy 77/100, pointwise ceiling 98, and a shared-weight bracket $[83, 86]$, giving a gap bracket $[12, 15]$. Pooling the original and fresh draws gives R128, uniform-pooling accuracy 77/100, pointwise ceiling 97, a shared-weight bracket $[82, 86]$, and a gap bracket $[11, 15]$. Both argmin and strict-threshold uniform counts are 77 in each table. Pointwise ceilings and shared lower counts were independently checked using exact rational witnesses on the stored binary64 contrasts; the upper bounds remain floating-point MILP bounds for the closed relaxation, not exact certificates of

Table 34: Direct class-output necessary-condition diagnostics after fixed source-pair scaling. Each split mean contains ten images and four noise draws; pooled uses twenty images and eight draws. Antisymmetry and negative mass use $\|P\|_F\|Q\|_F$; dependency uses $\|Q\|_F$. The rank threshold is 10^{-8} relative to the largest singular value. P has 32 columns; when its rank is 32, the dependency condition is vacuous. These are descriptive diagnostics, not Gaussian-acceptance or causal tests.

Mean table	Rank of P	Antisymmetry	Negative mass	Dependency
Source, first four	32	9.538×10^{-2}	5.690×10^{-1}	1.332×10^{-15}
Source, last four	32	1.003×10^{-1}	5.773×10^{-1}	1.382×10^{-15}
Target, first four	32	1.042×10^{-1}	5.480×10^{-1}	1.187×10^{-15}
Target, last four	32	1.049×10^{-1}	5.596×10^{-1}	1.144×10^{-15}
Pooled	32	1.420×10^{-1}	4.809×10^{-1}	1.033×10^{-15}

the strict optimum. At this numerical-upper-bound grade, the fresh block meets the preregistered persistence criterion. This is a same-image noise robustness check: the pooled table is dependent, and neither table establishes performance on new images or a deployable pointwise classifier.

Z SOURCE-FITTED COVARIANCE AND FROZEN TARGET CONFLICT PROPOSALS

Question and design. Can a covariance fitted on source denoiser outputs propose conflicts that are certified on different measured images? We test a specified Gaussian surrogate on the PIL/fp16 SD 1.5 MSE channels of Appendix N.4, using a $300/200$ source/target split and 500 disjoint calibration images to fix a 32-dimensional subspace. Source fitting and target evaluation were specified before their respective executions; source losses were already exposed. This protocol is distinct from the unresolved 16-dimensional fits in Appendix F.

Source measurement and convex fit. Let $U \in \mathbb{R}^{16384\times32}$ have orthonormal columns, let m_c denote projected calibration means, and let P be the calibration data covariance in that subspace. Restrict the score covariance to

$$S = UAU^\top + v_0(I - UU^\top), \qquad v_0 = \mathrm{tr}(P)/32, \qquad A \succ 0.$$

The means share the same orthogonal component, which cancels from class-loss contrasts. This block restriction does not cover arbitrary full-space covariances; estimated calibration means are not exact population means. For the first source image in each of the ten classes, measure all ten class outputs at the eleven source timesteps $50, 140, \ldots, 950$, with eight new noise draws. Use the original PIL preprocessing, fp16 VAE posterior mean, fp16 UNet/noisy-input arithmetic and class batch of ten. The master seed is 2026092101, with the original dataset-index/timestep/trial derivation. The $8{,}800$ new class-forwards follow 400 original-seed replay forwards; another 60 repeat forwards test the new path. All 800 replayed MSE/ℓ_1 scalars and the repeated outputs agree bitwise.

Average the float64 projected contrasts $-U^\top(\epsilon_c - \epsilon_0)/\sqrt{s_t}$ over source images and draws to obtain v_{tc}, where $s_t = (1-\alpha_t)/\alpha_t$. The Gaussian relation is $Av_{tc} = d_c - s_t v_{tc}$, $d_c = m_c - m_0$. Writing $A = v_0 B$, minimize the symmetric convex quadratic

$$\frac{\mathrm{mean}_{t,c,k}[(v_0 B v_{tc} - d_c + s_t v_{tc})_k^2]}{\mathrm{mean}_{c,k}(d_{c,k}^2)} + \frac{10^{-4}}{32^2}\|B - P/v_0\|_F^2, \qquad 10^{-4}I \preceq B \preceq 10^4 I.$$

A fixed projected-gradient eligibility check, rather than solver status alone, is required. The fit passes in 218 iterations with relative projected-gradient norm 9.69×10^{-8} at threshold 10^{-7}; B's smallest eigenvalue is at the prescribed floor. Equation RMSE, normalised by the RMS of d, is 0.539; this measures the secant-equation residual, not relative scalar-loss prediction error. Numerical eligibility does not imply predictive adequacy: its source-loss MSE relative to the zero-contrast predictor is $140{,}545.55$, versus 0.99879 for isotropic and 0.69929 for spline controls.

The isotropic variance is selected from a fixed 257-point log grid with one adjacent-interval refinement. The spline control uses four Bernstein time bases and affine projected noisy-input/noise

features, with fixed ridge 10^{-4}. Controls fit the 300 source images' existing $R = 64$ MSE contrasts at the eleven source timesteps; the Gaussian fit instead uses ten images' fresh $R = 8$ direct outputs. Nominal parameter counts are 528 and 2,340 for Gaussian and spline. These are unequal-data and unequal-capacity comparisons, not a controlled model-selection test.

Prediction and separation of source and target data. Let o be the calibration origin used to centre the projected class means. For the VP noisy latent x_{itr} and sampled noise ϵ_{itr}, set $\bar{y}_{it} = U^\top(R^{-1}\sum_r x_{itr}/\sqrt{\alpha_t} - o)$ and $\bar{e}_{it} = U^\top R^{-1}\sum_r \epsilon_{itr}$, using the recorded finite-precision inputs. With $H_t = \sqrt{s_t}(A + s_t I)^{-1}$, these means give the class-dependent MSE term

$$\widehat{L}_{itc} = \frac{\|H_t m_c\|^2 - 2\bar{y}_{it}^\top H_t^2 m_c + 2\bar{e}_{it}^\top H_t m_c}{16384}.$$

Class centring removes the common loss term. We fix the fitted models before reading target features, then fix eleven predicted loss tables: Gaussian/isotropic/spline crossed with full/first-half/last-half noise, plus two nominal expected Gaussian/isotropic tables. We fix all seven main arms' label-aware candidate sets and witnesses before obtaining actual target MSE. The noise baseline subsequently proposes on the first 32 draws, freezes its witnesses, and verifies on the last 32. The loader internally materializes the full per-image chunk, including ℓ_1 channels; the proposal routine receives only the permitted first-half MSE view. This is a program-level data-flow separation, not an assertion that only selected bytes were read.

All contrasts are divided by one frozen source class-centred MSE RMS, $a = 0.000193962506589$, and use binary64 $\tau = 10^{-6}$. Exact checks interpret the resulting binary64 coefficients as dyadic rationals. This normalisation differs from the earlier per-image max-normalised tables; their gap counts cannot be imported here. Each arm first proposes strict pointwise weights using LPs, then finds shared-infeasible groups by deterministic descending-index deletion and removes each accepted group before continuing. Rationally checked witnesses are required throughout. The fixed limits are 2,000 LP calls, 10 seconds per LP, 600 seconds per arm, and 50 disjoint groups. These are deletion-reduced proposals, not minimum infeasible subsystems or optimal disjoint packings. Target inputs and labels participate in candidate selection: this is not label-blind prediction or deployment.

Exact transfer criterion. For one proposed group, let $p_{ij} \geq 0$, $\sum_{i,j} p_{ij} = 1$ be its frozen dual, and let w_i be frozen simplex pointwise witnesses. Set $E = D - \widehat{D}$ and

$$\beta = \tau - \max_t \sum_{i,j} p_{ij}\widehat{D}_{itj} > 0, \quad \varepsilon_{\mathrm{dual}} = \max_t \sum_{i,j} p_{ij}E_{itj}, \quad \gamma_i = \min_j \sum_t w_{it}\widehat{D}_{itj} - \tau > 0.$$

If $\varepsilon_{\mathrm{dual}} < \beta$ and $\min_j \sum_t w_{it}E_{itj} > -\gamma_i$ for every member, each image is individually reachable on D while no common weighting reaches them all. Indeed, the dual-weighted actual margin is less than τ for every simplex weighting, whereas reaching all members would make it greater than τ. The pointwise inequalities give strict reachability directly. Thus q such image-disjoint groups certify $O_\tau - G_\tau^* \geq q$. We also report q_{direct}, checking the same frozen witnesses directly on D: dual maximum at most τ and all pointwise margins above τ. This can succeed when the sufficient error conditions fail, and is never smaller than q. No witness is re-optimized on the actual verification table. A restricted ten-timestep result is not a lower bound for the full 21-timestep family.

Prediction error and conflict transfer. Table 35 reports every arm. The Gaussian primary proposes six groups, but neither transfer test certifies any of them. All 80 frozen duals across the nine arms fail the actual dual inequality; some pointwise witnesses survive. The noise baseline also transfers no group. Five arms reach the LP-call limit, and even the other stop reason does not certify absence of additional proposals. Hence zero here is an uninformative gap lower bound, not proof of zero actual coupling.

On the ten held-out timesteps, Gaussian, isotropic and spline class-centred MSE-contrast RMSEs are respectively 0.0521448, 0.000190136 and 0.000182354 (before dividing by a). Gaussian/isotropic RMSE is 274.25 with descriptive image-bootstrap 95% interval $[250.95, 299.95]$; Gaussian/spline is 285.95 $[261.44, 312.77]$. The 10,000 bootstrap replicates resample the 200 images, keeping source fits fixed. Full-grid RMSEs are respectively 0.0634558, 0.000191256 and 0.000183305. The fixed projected model predicts class-centred loss contrasts poorly and provides no certified transfer of its

Table 35: Frozen conflict proposals on 200 PIL/MSE target images. Model proposals are checked on actual $R64$; measured first-half proposals are checked on the independent last half. $T = 10$ uses the held-out timestep positions. Both counts use exact rational checks on fixed binary64 tables; groups are disjoint within an arm, not pooled across arms. The separate nominal isotropic factor control has $O = G^* = 56$ by the common-atom lemma; that equality is not asserted for finite-noise isotropic predictions.

Proposal table	T	Groups	q	q_{direct}	Search stop
Gaussian finite $R64$	21	6	0	0	No further certificate
Gaussian finite $R64$	10	3	0	0	No further certificate
Gaussian expected	21	3	0	0	No further certificate
Isotropic finite $R64$	21	10	0	0	LP-call cap
Isotropic finite $R64$	10	11	0	0	LP-call cap
Spline finite $R64$	21	14	0	0	LP-call cap
Spline finite $R64$	10	16	0	0	LP-call cap
Measured first 32 draws	21	7	0	0	No further certificate
Measured first 32 draws	10	10	0	0	LP-call cap

frozen conflict proposals. With power uncalibrated and zero transfer also observed for the first-32-to-last-32 noise baseline, zero transfer alone does not distinguish surrogate error from finite-noise instability of the selected witnesses or establish that a perfect expected-loss predictor would fail at $R = 64$. These results provide no positive bridge evidence at this operating point; they do not reject all Gaussian approximations, explain the primary ℓ_1 gap, or identify the cause of the observed coupling. The completed protocol includes no target-driven refit or expanded search. Public aggregates, group masks and portable code accompany the result; real loss/output arrays and witness payloads remain outside the release.

Source-only repeatability follow-up. An exploratory loss-based refit holds the calibration-derived 32-dimensional projection and covariance eigenvectors fixed, fitting the shared covariance eigenvalues on 200 source images at eleven timesteps in each of three balanced folds. Frozen out-of-fold models predict a new $R = 64$ noise block on 30 source images without refitting. On the ten timesteps excluded from fitting, Gaussian/isotropic class-centred MSE-contrast RMSE ratio is 1.0036 (descriptive 10,000-replicate image-bootstrap 95% interval $[0.9926, 1.0148]$), showing no resolved predictive advantage. Across the full grid, the old/new $R = 64$ contrast difference has RMS 4.03×10^{-5}; the two new $R = 32$ halves give 5.81×10^{-5}. This independently audited source diagnostic measures repeatability, not fresh-image generalisation, an irreducible prediction-error floor or calibrated conflict-transfer power.

Bidirectional profile-based conflict-set transfer. A separate post-hoc analysis uses the original and independently remeasured $R = 64$ MSE tables on the same 100 images in each of CIFAR-10 and Pet (SD 1.5, tensor preprocessing, $T = 21$, $\tau = 10^{-6}$). Both blocks had previously been analysed; the target data were therefore previously examined, and the study was not preregistered before data collection. For each image, three families fit class-centred losses with signed coefficients on $K = 2$ basis functions, choosing among all 300 pairs from a fixed 25-point grid. The spectral family is $s/(s+\lambda)^2$ with $\lambda \in \mathrm{logspace}(-2, 2, 25)$; comparators are $z(s)^p$, $p \in \mathrm{linspace}(0, 3, 25)$, and s^p, $p \in \mathrm{linspace}(-2, 2, 25)$, where $z(s) = (\log s - \min \log s)/(\max \log s - \min \log s) + 0.05$. Here s is the scheduler noise variance on the full grid. Pair selection minimises class-centred squared error, equivalently all-pair contrast squared error. Nonnegative weights constrain the subsequent timestep combination, not the fitted contrast coefficients. The original-block per-image contrast normalisation is retained in both directions, including as a historical reference in reverse transfer.

Each arm proposes disjoint jointly infeasible sets from images that are individually reachable in both the measured source and, where applicable, its reconstruction. A fixed descending deletion pass reduces each set; no minimum-cardinality claim is made. All four cells' proposals and matched random sets are frozen before cross-evaluation (Table 36). On the destination table, every member must again be individually strictly reachable and the entire unchanged set jointly infeasible. Primal

Table 36: Certified set transfer, q/N: N source-proposed sets and q destination-certified conflicts with all members individually reachable. Raw measured proposals give a noise baseline, not an upper bound. Random counts pool 200 permutations of the spectral source-eligible pool, each partitioned into the spectral set sizes; their ratio equals the mean replicate rate, not 200 independent datasets. This matches source reachability eligibility and size, not class composition. Directions are dependent and reported separately. All cells resolved.

Dataset	Source → destination	Raw	Spectral	Monomial	Power law	Random
CIFAR-10	original → fresh	$2/7$	$0/9$	$1/7$	$1/9$	$6/1800$
CIFAR-10	fresh → original	$3/10$	$0/8$	$0/9$	$0/8$	$6/1600$
Pet	original → fresh	$3/7$	$1/8$	$1/9$	$0/8$	$7/1600$
Pet	fresh → original	$4/8$	$1/7$	$0/8$	$1/7$	$16/1400$

and dual witnesses may be reoptimised there; member deletion, reselection and profile refitting are forbidden. Both properties are checked by exact rational arithmetic on the fixed binary64 contrasts. Thus q transferred disjoint sets certify a gap of at least q within that arm and table; counts across arms, directions or random repetitions cannot be added.

In table order, eligible source-pool sizes for raw/spectral/monomial/power law are $91/86/85/86$, $94/86/87/87$, $97/90/89/88$ and $98/89/89/88$. Spectral sets contain two or three images in the first row, two in the second and third, and two or four in the fourth. Three raw sets lose pointwise reachability in reverse CIFAR-10 transfer, and one in reverse Pet transfer; no reconstructed-arm set does so. Remaining nontransferred sets are jointly feasible, with no numerical unresolved cases. The random replicate-rate 5th/50th/95th percentiles are $(0, 0, 0)$ in the first three rows and $(0, 0, 1/7)$ in the fourth; these are descriptive Monte Carlo summaries, not confidence intervals.

The fixed positive criterion requires spectral q/N to strictly exceed both fitted comparators and the matched-random mean in every dataset and direction. It is not met: only forward Pet meets all three inequalities, and reverse Pet ties power law. This selection rule therefore establishes no consistent spectral advantage, not equivalence between families. Nonzero raw transfer shows that some measured sets survive this particular test; it does not calibrate its power for a perfect expected-loss predictor. Image-specific basis fits impose no shared covariance, eigendirections or means, so even a positive outcome would not validate the shared Gaussian explanation. Public group indices, statuses and recount code reproduce the table; loss arrays and fitted profiles remain undistributed.