WHEN FEATURE RANK MISSES MODEL COLLAPSE: CLASS GEOMETRY IN RECURSIVE CLASSIFIERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Global feature rank can rise even when within- and between-class covariance components both become spectrally more concentrated. We study this behavior in image classifiers recursively retrained on generated images and inherited pseudo-labels. In a three-seed CIFAR-10 study with the original ResNet stem, four replacements increase centered-covariance effective rank by 14.5% and within-to-between-class scatter 3.22-fold, while accuracy falls from 88.4% to 62.2%. Exact entropy accounting shows how variance reweighting conceals concentration of both components. Later generations lose another 9.28 accuracy points while rank rises 3.30%; fixing the initial teacher avoids this additional loss on identical image pools. A standard CIFAR stem raises baseline accuracy to 95.4%, yet retains the full-trajectory mismatch: accuracy reaches 56.2% while rank increases 2.6%. Its later rank changes vary by seed. Alternative spectral definitions also give different directions, and GAN and ImageNet-100 diffusion workflows lose rank as geometry deteriorates. Disjoint subspace and frozen-feature readout analyses corroborate degraded class structure. Low-label audits estimate geometry but do not establish better warning or intervention decisions. These findings identify what increasing covariance rank can conceal, while showing why its interpretation depends on class-conditioned spectra, training stage and predictive performance.

1 INTRODUCTION

Recursive use of model-generated training data can degrade subsequent models, with outcomes that depend on how synthetic and real data are used (Shumailov et al., 2024; Alemohammad et al., 2024; Gerstgrasser et al., 2024; Kazdan et al., 2025). How should a researcher interpret a rising feature rank during that process? Effective rank provides a compact, label-free description of a representation's spectrum (Roy & Vetterli, 2007); singular-value entropy also supports model selection within self-supervised learning methods (Garrido et al., 2023). A rising rank may nevertheless combine changes with different meanings for the task. Accuracy detects predictive degradation, but does not explain whether class-conditioned feature variation becomes more concentrated while the marginal spectrum becomes more dispersed.

We find that *both within- and between-class covariance components can lose effective rank while their sum gains it*. In a three-seed CIFAR-10/DDPM study, four synthetic replacements increase centered-covariance rank by 14.5% and within-to-between-class scatter $3.22\times$, while accuracy falls from 88.4% to 62.2%. Exact entropy accounting identifies the offset: the higher-entropy within-class component gains variance weight, outweighing concentration of the component spectra. Conditional-GAN and held-out ImageNet diffusion workflows exhibit the opposite balance and lose total rank (Figure 1). The empirical question is which internal changes a global rank trend conceals.

The recursive setting requires separating the first real-to-synthetic transition from continued teacher inheritance. In the original-stem CIFAR trajectories, most accuracy loss occurs at the first replacement, yet later generations still lose accuracy and frozen-feature readout performance while rank increases. On identical image pools, fixing the initial labeling teacher avoids this additional loss and changes the later entropy balance. Singular-value entropy scores also increase on these features, whereas participation ratio and stable rank need not. A standard CIFAR input stem retains the full-trajectory covariance-rank mismatch from a stronger real-data baseline, but changes the later rank pattern.

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.

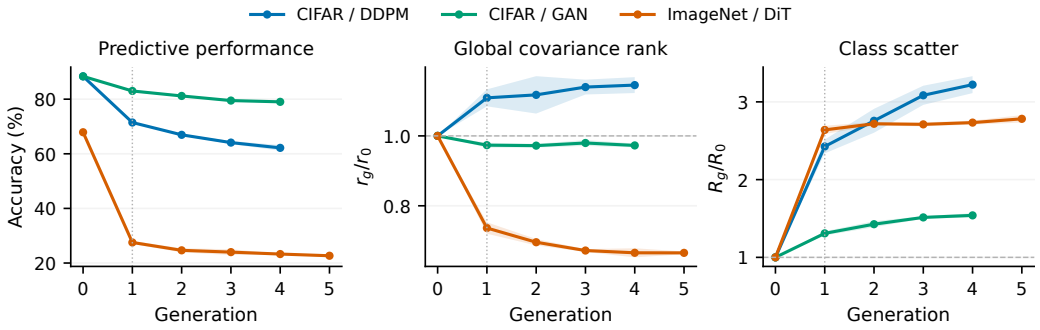

Figure 1: **Predictive degradation, opposite rank directions, worsening class scatter.** Fresh ResNet trajectories on disjoint official-test endpoints. Rank and scatter are normalized within seed to g0; accuracy is unnormalized. Lines and bands show means and one sample SD across three seeds. The dotted line separates the first replacement from later recursion. CIFAR workflows share their initial classifiers; generator, conditioning and sampling differ jointly.

We characterize how component spectra, variance weights and their mixing gap combine during recursive classifier training, and test their dependence on training stage, teacher choice, input stem and rank definition. Disjoint class-subspace and supervised readout analyses test whether the effect extends beyond a scatter ratio. The mathematical identities organize these observations; they do not constitute a new theory of neural learning. Small labeled sets can estimate geometry, but the tested monitoring and retention policies do not establish a consistent advantage over accuracy or unconditional switching.

2 Related Work

Recursive synthetic training. Studies of generative self-training analyze distributional degradation and finite-sample effects (Shumailov et al., 2024; Alemohammad et al., 2024; Suresh et al., 2025). Related regression theory identifies risk floors under synthetic contamination (Dohmatob et al., 2025). Accumulating data and retaining real examples can change the outcome substantially (Gerstgrasser et al., 2024; Kazdan et al., 2025). Other work connects diffusion-model collapse to a transition from generalization to memorization (Shi et al., 2025), or establishes self-verification bounds using a fixed real-trained reference model and confidence-calibration assumptions (Fu et al., 2025). Our fixed-teacher control assigns every generated image the initial classifier's top prediction; it does not implement their candidate-filtering procedure or inherit its guarantees. Our primary workflow instead inherits pseudo-labels under a fixed synthetic input distribution. Confirmation bias in pseudo-labeling is related (Arazo et al., 2020). Generator stability does not isolate label error: input shift, optimization and label imbalance remain possible causes.

Class geometry and dimensionality. Neural collapse describes within-class concentration, simplex organization of class means, alignment with classifier weights and nearest-center decisions near the end of supervised training (Papyan et al., 2020). Unconstrained-feature analyses with weight decay (Zhu et al., 2021) and unit-sphere cross-entropy analyses (Lu & Steinerberger, 2022) clarify conditions for this geometry. Such conditions differ from evaluation on true labels after synthetic pseudo-label training. Experiments that suppress neural collapse can retain generalization (Han et al., 2025). We report a trace scatter ratio and explicitly distinguish it from the canonical pseudoinverse NC1 statistic.

Effective rank summarizes a spectrum (Roy & Vetterli, 2007); feature singular values and covariance eigenvalues yield different summaries. RankMe uses embedding-matrix singular-value entropy for label-free hyperparameter selection within self-supervised learning methods (Garrido et al., 2023). We evaluate this singular-value weighting on supervised penultimate features as a definition-sensitivity check, not as a replication or refutation of its SSL protocol. Dimensional collapse in contrastive learning (Jing et al., 2022) and Hessian spectral collapse in continual learning (Prakash et al.,

2026) motivate other spectral diagnostics. Recursive generator training can reduce input–output Jacobian rank (Batsell et al., 2026); this local derivative spectrum differs from global classifier-feature covariance. Rank and alignment have also been combined for reasoning under distribution shift (Vo & Nguyen, 2026), and REPS combines centered-activation rank, class separation and other components to evaluate compression (Shinde & Sharma, 2026). We therefore locate the contribution in the measured component concentration, stage dependence and teacher comparisons of recursive classifiers, rather than a first-ever combination of rank and geometry.

Diagnostic comparisons. Centered kernel alignment compares paired representations (Kornblith et al., 2019). Because it does not use task labels, we do not interpret a similarity change alone as evidence of preserved or lost class information. Accuracy, negative log likelihood (NLL) and expected calibration error (ECE) measure different predictive properties (Guo et al., 2017). We compare estimation of these quantities at the same label budget. Their different units and targets preclude a claim that the largest relative change is the most sensitive or useful diagnostic.

3 Protocol and measurements

Recursive classifiers. Generation 0 trains a classifier C_0 on labeled real training data. At step t, a fixed pretrained generator produces fresh synthetic images, C_t supplies their hard pseudo-labels, and a newly initialized C_{t+1} is trained on the resulting data. The primary CIFAR studies evaluate four replacements and the ImageNet ResNet study five; Appendices A and F specify the remaining horizons. Section 5.8 separately examines joint generator–classifier retraining.

Fixing the generator holds the image-generating mechanism stable while synthetic input shift, inherited label errors and optimization remain coupled. A separate control fixes the labeling teacher to C_0 on the same generated images (Section 5.3). We use model collapse here to describe finite-horizon degradation on the original real-data task, without claiming an asymptotic limit or collapse of the fixed generator.

Replace uses only the latest synthetic set after generation 0. *Mix-*α uses a fixed-size mixture with real fraction $\alpha \in \{0.25, 0.5, 0.75\}$. *Accumulate* retains the original real set and all previous synthetic sets with the pseudo-labels assigned when each was generated. Accumulation increases dataset size and optimizer steps at fixed epochs, so it is not a compute-matched intervention.

Covariance rank and scatter. For an evaluation feature matrix $F \in \mathbb{R}^{N\times d}$, let x_i be its rows, $\mu = N^{-1}\sum_i x_i$, $F_c = F - \mathbf{1}\mu^\top$, and $\Sigma = N^{-1}F_c^\top F_c$. We apply the effective-rank definition (Roy & Vetterli, 2007) to the *covariance eigenvalues* λ_j:

$$r = \operatorname{erank}(\Sigma) = \exp\left(-\sum_{j:p_j>0} p_j \log p_j\right), \qquad p_j = \frac{\lambda_j}{\sum_l \lambda_l}. \tag{1}$$

This squares the singular values of centered features before normalization. We also compare entropy of singular values of F and F_c, covariance participation ratio $(\operatorname{tr}\Sigma)^2/\operatorname{tr}(\Sigma^2)$, and centered-feature stable rank $\operatorname{tr}\Sigma/\lambda_{\max}(\Sigma)$ (Appendix C). These summaries differ in centering and spectral weighting and need not agree. We use $r = 0$ when total covariance is zero; otherwise $1 \le r \le \min(d, N-1)$.

For labels y_i, class counts n_k, empirical priors $\pi_k = n_k/N$ and means μ_k, define

$$W = \sum_k \pi_k \frac{1}{n_k}\sum_{i:y_i=k}(x_i-\mu_k)(x_i-\mu_k)^\top, \qquad B = \sum_k \pi_k(\mu_k-\mu)(\mu_k-\mu)^\top,$$
$$R = \frac{\operatorname{tr} W}{\operatorname{tr} B}, \qquad \Sigma = W + B. \tag{2}$$

The empirical-prior weights make the decomposition exact even for unequal class sizes. All principal probes are class-balanced. R is a trace-ratio measure inspired by neural-collapse geometry; canonical NC1 instead involves $\operatorname{tr}(WB^\dagger)/K$ (Papyan et al., 2020). The two statistics are generally unequal. We call our measure the *scatter ratio* throughout. Higher R means less separation of class means relative to within-class variation, without determining classification error or every aspect of class geometry. Global translation and nonzero uniform scaling leave both r and R unchanged.

Beyond class centers. We fit a rank-q affine subspace to each class and measure residuals on a disjoint set. The original-stem CIFAR replication splits the 9,800 endpoint images into 490 fitting and 490 evaluation examples per class, excluding the 200 monitoring images; historical analyses use 500/500 per class. For a feature x with true class y, let $e_k(x)$ be the squared residual to class k's fitted affine subspace and $e_{-y}(x) = \min_{k \neq y} e_k(x)$. The projection margin is

$$M_q = \frac{1}{N_{\text{eval}}} \sum_i \frac{e_{-y_i}(x_i) - e_{y_i}(x_i)}{e_{-y_i}(x_i) + e_{y_i}(x_i) + 10^{-12}}. \tag{3}$$

Positive values favor the correct class subspace; smaller values indicate less distinction. We use $q = 10$ and report $q = 5, 20$ as sensitivity checks. Centered NC2/NC3 deviations and network–nearest-center agreement provide additional descriptors (Appendix B).

Limited-label auditing. On historical CIFAR features, a trusted, class-balanced labeled subset is fixed across generations and never used to train the classifiers. Its R_t is normalized by its own R_0. We distinguish an audit against the full 10,000-image evaluation set, which includes the audited examples, from a disjoint audit. The latter reserves 500 examples per class as a reference and draws 20 or 50 examples per class from the remaining 500. Each budget uses 200 subset draws per training seed, with identical examples over generations and identical subsets for all diagnostics. The reference R_t is normalized by its own baseline. We report mean absolute percentage error (MAPE) for normalized scatter and NLL, and mean absolute error in percentage points for accuracy and 15-bin ECE. All errors exclude generation 0. This measures concurrent estimation against a finite reference. Separate fresh studies test monitoring decisions with a fixed 200-image audit (Section 5.7).

4 Interpreting rank through class geometry

Two distinct issues underlie the diagnostic mismatch. First, even an entire covariance matrix cannot identify class discriminability. Second, when covariance changes, its entropy can combine competing changes in within- and between-class structure. We separate the non-identifiability statement from the accounting used to interpret our learned representations.

Identical covariance, different discriminability. Fix Y uniform on $\{-1, +1\}$ and input $U = (2Y + E_1, E_2, E_3)^\top$, where the E_i are independent standard normals, independent of Y. The linear representation

$$X_\rho = \left(\rho U_1 + \sqrt{5(1-\rho^2)}\, U_2,\ U_3\right)^\top, \qquad 0 \le \rho \le 1,$$

has $\mathrm{Cov}(X_\rho) = \mathrm{diag}(5, 1)$ for every ρ, yet $B_\rho = 4\rho^2 e_1 e_1^\top$. Thus

$$R_\rho = \frac{3}{2\rho^2} - 1, \qquad \mathcal{A}_\rho = \Phi\left(\frac{2\rho}{\sqrt{5 - 4\rho^2}}\right), \tag{4}$$

where $\mathcal{A}_\rho$ is Bayes-optimal accuracy and Φ the standard-normal CDF. As ρ decreases to zero, R_ρ diverges and accuracy falls from $\Phi(2)$ to $1/2$ at unchanged covariance and true labels. Appendix C derives the construction and a complementary fixed-feature relabeling identity (Proposition 1). Neither specifies a neural learning update or implies monotone degradation under retraining.

Exact accounting when covariance changes. For nonzero scatter components, let $H_A = \log \mathrm{erank}(A)$ and $p = \mathrm{tr}\, B / \mathrm{tr}\, \Sigma = 1/(1+R)$. Normalizing each matrix by its trace gives $\rho_\Sigma = p\rho_B + (1-p)\rho_W$. Define the entropy-mixing gap

$$\mathcal{J} = H_\Sigma - pH_B - (1-p)H_W, \tag{5}$$

the standard weighted matrix Jensen–Shannon quantity (Majtey et al., 2005). For endpoints $0, t$, write $\Delta q = q_t - q_0$ and $\bar{q} = (q_t + q_0)/2$. Applying $\Delta(ab) = \bar{a}\Delta b + \bar{b}\Delta a$ gives

$$\Delta H_\Sigma = \underbrace{\Delta p(\bar{H}_B - \bar{H}_W)}_{D_p} + \underbrace{\bar{p}\Delta H_B}_{D_B} + \underbrace{(1-\bar{p})\Delta H_W}_{D_W} + \underbrace{\Delta \mathcal{J}}_{D_J}. \tag{6}$$

The first term changes the variance weight assigned to each component; the next two measure changes in the component spectra; the last is the change in mixing gap. All terms use natural-log units. Since B is formed from K centered class means, its rank is at most $K-1$; this bounds its entropy without guaranteeing $H_B < H_W$. If scatter rises, p falls. When the between-class component has lower entropy than the within-class component, that shift makes D_p positive. Total rank can therefore rise even while both component ranks fall, provided reweighting and mixing outweigh spectral concentration. Section 5.4 tests this balance on the stored trajectories.

Equation (6) holds for the measured covariances without a corruption model or alignment of coordinates across generations. $\mathcal{J}$ depends jointly on weights, spectra and relative orientation; D_J is not a pure orientation effect. This is descriptive accounting, not a causal decomposition of training. In particular, actual classifier errors can depend on features within each true class, violating the class-conditional assumption of Proposition 1; Appendix C measures that discrepancy.

5 Experiments

5.1 Settings and evidence scope

The primary comparisons use ResNet-18 on CIFAR-10 and ImageNet-100; historical scope checks add CIFAR-100, ConvNeXt-Tiny and ViT-S/16 (Krizhevsky, 2009; Deng et al., 2009; Russakovsky et al., 2015; He et al., 2016; Liu et al., 2022; Dosovitskiy et al., 2021). Feature dimensions are 512, 768 and 384, respectively. Classifiers start from random initialization at each generation. The original CIFAR ResNet retains the ImageNet-style stem; Section 5.5 examines a standard CIFAR stem. CIFAR uses DDPM with a separate conditional StyleGAN2-ADA comparison; ImageNet uses DiT-XL/2 (Ho et al., 2020; Karras et al., 2020; Peebles & Xie, 2023). Synthetic sets contain 50,000 images. Each trajectory has three seeds, except historical retention with two and the selected policy replication with six. Unless stated otherwise, $\pm$ denotes sample SD in fresh studies and population SD in historical studies; neither is a confidence interval.

Table 1: Evidence roles and label budgets. Rows describe different questions and cohorts; repeated subset draws and shared checkpoints are not independent training replications.

Question	Evaluation and trusted labels	Scope
Rank versus geometry	Fresh CIFAR: 9,800 endpoint images; ImageNet: 5,000 official validation images	Three seeds; DDPM/GAN share g0; new stem refits g0
Subspaces and readout	CIFAR: 4,900 fit + 4,900 disjoint evaluation labels	15 original-stem models; excludes 200 audit images
Small-label estimation	Historical CIFAR: 200 or 500 audit labels; 5,000 disjoint reference labels	200 subset draws per seed, fixed across generations
Monitoring decisions	Fresh CIFAR: 200 audit labels; 9,800 disjoint endpoint labels	Three-seed primary and six-seed selected policy studies

Table 1 separates feature characterization, estimation and decision evaluation. The primary ImageNet results use official validation images, matched preprocessing and a fixed 50,000-row training budget. The historical ImageNet probe uses training-origin images and mismatched pseudo-label preprocessing; it is retained only as descriptive evidence in Appendix E. Historical CIFAR-100 uses a CIFAR-10 generator, confounding dataset and input domain. Full recipes, partitions and post-outcome selections are in Appendices A and F.

5.2 Fresh trajectories show opposite rank directions

The fresh CIFAR-DDPM replication is the primary rank-increase example. After four replacements, accuracy falls from 88.35% to $62.17 \pm 1.04\%$, within-seed scatter rises 3.22 ± 0.11-fold and rank rises $14.54 \pm 2.28\%$ (Table 2). This independently reproduces the historical CIFAR mismatch while separating the 200 monitoring images from the 9,800-image endpoint. Final-epoch accuracy and geometry always refer to the same checkpoint.

With the same baselines and seeds, conditional StyleGAN2-ADA replacements reach $79.02 \pm 0.25\%$ accuracy; scatter grows 1.540 ± 0.028-fold while rank falls $2.73 \pm 0.56\%$ (Tables 28–29). The

Table 2: Primary fresh ResNet trajectories. Accuracy uses disjoint official-test endpoints (9,800 CIFAR or 5,000 ImageNet images). Changes are computed within each seed before averaging; values are means ± sample SD across three seeds. CIFAR GAN reuses DDPM's g0 models. Generator family, conditioning and sampling change jointly.

Workflow (last t)	g0 acc. (%)	gt acc. (%)	R_t/R_0	Δr (%)
CIFAR-10 / DDPM (4)	88.35 ± 0.08	62.17 ± 1.04	3.22 ± 0.11	$+14.54 \pm 2.28$
CIFAR-10 / GAN (4)	88.35 ± 0.08	79.02 ± 0.25	1.54 ± 0.03	-2.73 ± 0.56
ImageNet-100 / DiT (5)	67.89 ± 0.41	22.67 ± 0.59	2.78 ± 0.05	-33.47 ± 0.40

Table 3: First replacement and subsequent CIFAR-DDPM changes. Accuracy uses 9,800 endpoint images and is in percentage points; rank changes are relative percentages. Means ± sample SD of paired changes over three seeds. Both arms share g0/g1; the later comparison holds image pools and recipe fixed.

Interval / teacher	Δ acc. (pp)	R_b/R_a	Δr (%)
Shared first replacement	-16.90 ± 0.31	2.426 ± 0.089	$+10.89 \pm 2.43$
Inherited teacher, g1–g4	-9.28 ± 1.17	1.329 ± 0.075	$+3.30 \pm 0.49$
Fixed C0, g1–g4	$+0.67 \pm 0.10$	0.966 ± 0.027	-3.65 ± 0.07

diagnostic mismatch concerns the interpretation of rank, rather than a universal increase. This post-outcome comparison changes generator family, conditioning and sampling jointly and does not isolate a family effect.

Held-out ImageNet ResNet trajectories start at $67.89 \pm 0.41\%$ accuracy and reach $24.65 \pm 0.75\%$ at generation 2, then 22.67% at generation 5. The prespecified continuation contrast, generation $2 \to 5$, is -1.99 ± 0.32 accuracy points, 1.024 ± 0.025-fold scatter and $-4.37 \pm 1.12\%$ rank (Tables 19–20). Scatter changes little after generation 2 and briefly improves at generation 3 despite further accuracy loss. This post-outcome extension adds depth, not independent seeds or evidence of an asymptotic limit. Matched one-step real-only retraining retains 88.48% accuracy on CIFAR and 68.61% on ImageNet; it changes both inputs and labels and cannot isolate their effects.

The historical configurations broaden the observed directions but also expose a limit of R: a weaker ConvNeXt run loses accuracy from 46.40% to 17.60% while scatter improves from 5.895 to 5.238 (Appendix E). An ImageNet fixed-image conditioning-target control also lowers both accuracy and R; its targets are generator requests, not verified image labels. Thus neither rank nor scatter is a surrogate for predictive performance.

5.3 Separating first replacement from teacher inheritance

The first replacement changes both the input distribution and the source of training labels. Does inheriting later classifiers add to that initial loss? We repeat generations 2–4 with the original real-trained C_0 as the labeling teacher, holding each seed's DDPM images, initialization seeds, 50,000-row budget and 200-epoch recipe fixed. This post-outcome control shares the existing g0/g1 models and trains nine new classifiers; it changes no generated images or evaluation rows (Appendix F).

The first replacement accounts for 64.55% of the mean g0-to-g4 accuracy loss. Inherited-teacher g1–g4 loses a further 9.28 ± 1.17 accuracy points while rank rises $3.30 \pm 0.49\%$ and scatter grows 1.329 ± 0.075-fold (Table 3). With fixed C0, the same later interval instead gains 0.67 ± 0.10 accuracy points and decreases rank and scatter. Its g4 accuracy is $72.12 \pm 0.34\%$, a paired 9.95 ± 1.26 points above inheritance (Table 22).

Fixing the teacher avoids the additional loss in these trajectories, but leaves the first-transition gap: fixed-C0 g4 still has 2.343 ± 0.023-fold scatter and $6.84 \pm 2.27\%$ higher rank than g0. The mismatch therefore occurs even without continuing teacher inheritance. Teacher choice also changes target-label frequencies; unknown true generated-image labels prevent separating marginal shifts from conditional prediction errors. We next ask which covariance changes accompany the distinct later trajectories.

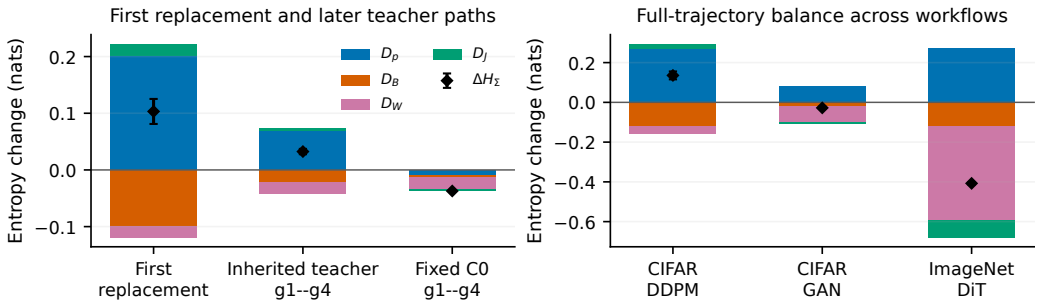

Figure 2: **Variance reweighting can mask component concentration.** Signed bars show mean entropy-accounting terms; black diamonds show total ΔH_Σ with one sample SD over three seeds. Left: the first CIFAR replacement and each teacher's later interval. Right: complete original synthetic trajectories. Positive and negative terms are stacked separately. These are descriptive endpoint identities, not causal effects of training; numerical terms and all stages are in Tables 8 and 10.

Table 4: Rank-definition sensitivity on the same inherited-teacher CIFAR features. Paired relative changes (%), mean ± sample SD across three seeds. The uncentered SVD score uses the zero-epsilon limit of RankMe's weighting on supervised features; no SSL model-selection claim is tested. Definitions and all workflows are in Appendix C.

Global summary	g0–g4	g1–g4
Covariance entropy	$+14.54 \pm 2.28$	$+3.30 \pm 0.49$
Uncentered SVD entropy	$+16.53 \pm 1.81$	$+6.01 \pm 3.38$
Centered SVD entropy	$+18.05 \pm 2.40$	$+3.87 \pm 3.03$
Participation ratio	-0.61 ± 0.74	$+2.89 \pm 1.65$
Feature stable rank	-19.02 ± 1.38	$+2.74 \pm 7.33$

5.4 Accounting for opposite rank directions

The original full-trajectory accounting covers 51 stored ResNet models in CIFAR-DDPM, CIFAR-GAN, ImageNet-DiT and their one-step real-only controls. Stage and definition analyses add the nine fixed-C0 models, retaining the same true-label evaluation rows. These analyses were selected after observing the trajectories. Equation (6) separates the measured change in log rank into variance reweighting, component spectral changes and the entropy-mixing gap (Appendices C and C).

In all nine original synthetic g0-to-final seed comparisons, R rises while *both* H_B and H_W decrease. CIFAR-DDPM's total rank nonetheless increases: mean reweighting, $+0.267$ nats, exceeds the combined component contribution, -0.157, with $+0.025$ from mixing. In CIFAR-GAN and ImageNet-DiT, component concentration instead dominates (Figure 2, right). The same direction of class-scatter deterioration is therefore compatible with either total-rank direction.

The balance also differs after the first replacement (Figure 2, left). In inherited-teacher g1–g4, mean reweighting of $+0.0694$ nats outweighs the -0.0412 component contribution; mixing adds $+0.0042$, leaving $\Delta H_\Sigma = +0.0324$. Fixed C0 instead gives mean reweighting -0.0085 and $\Delta H_\Sigma = -0.0372$. Later component entropies decrease on average, with a slight within-class increase for inherited seed 102. Each interval is decomposed directly: differences between g0-relative attribution columns would generally give different terms.

5.5 Sensitivity to rank definition and input stem

Is the increasing-rank observation specific to squaring singular values or centering features? On the same original-stem CIFAR g0–g4 features, uncentered and centered singular-value entropy ranks increase by $16.53 \pm 1.81\%$ and $18.05 \pm 2.40\%$, respectively; both increase in all three seeds. Covariance participation ratio changes by $-0.61 \pm 0.74\%$, while centered-feature stable rank decreases $19.02 \pm 1.38\%$ (Table 4). Their later-stage changes differ again. The same features therefore give

Table 5: CIFAR input-stem sensitivity. Means ± sample SD across three seeds; scatter folds and rank changes are computed within seed. Both students and inherited teachers use their respective stem; image pools and partitions are shared, but all five CIFAR-stem models per seed are newly trained.

Input stem	Acc. g0 (%)	Acc. g4 (%)	R_4/R_0	Δr (%)
Original ImageNet stem	88.35 ± 0.08	62.17 ± 1.04	3.222 ± 0.109	$+14.54 \pm 2.28$
CIFAR stem	95.41 ± 0.18	56.18 ± 1.77	10.498 ± 1.142	$+2.60 \pm 1.08$

opposite trend signs under different spectral weightings. Full definitions, numerical checks and GAN/ImageNet/control results are in Appendix C.

The original 7×7, stride-2 input stem gives a weaker CIFAR baseline than a 3×3, stride-1 stem without max-pooling. To test whether the mismatch depends on that choice, we train all 15 g0–g4 models again with the latter stem in both students and teachers, retaining the same DDPM image pools, seeds, 200-epoch recipe and evaluation partitions. Baseline accuracy rises to $95.41 \pm 0.18\%$, but reaches $56.18 \pm 1.77\%$ at g4 as scatter grows 10.498 ± 1.142-fold and covariance rank rises $2.60 \pm 1.08\%$ (Table 5). Both component entropies fall and total rank rises in every seed. Here reweighting alone does not offset concentration: mean contributions are $+0.304$ nats from reweighting, -0.322 from components and $+0.044$ from mixing, yielding $+0.026$ total.

This replication is stage- and definition-dependent. During g1–g4, accuracy loses another 13.97 ± 0.78 points, but covariance rank decreases in two seeds and increases in one. Full-trajectory centered singular-value entropy increases on average, whereas its uncentered counterpart decreases on average. Thus the covariance-rank mismatch survives a stronger baseline, without establishing a universal later-stage or alternative-rank direction. This post-outcome architecture check reuses images and seed values; it changes teacher predictions as well as student architecture. Appendix F retains every generation, rank definition, direct entropy contrast and independent replay check.

5.6 Corroboration beyond a trace ratio

The scatter ratio summarizes only trace allocation. To check whether its deterioration accompanies a broader loss of class distinction, we use disjoint subspace fits on all 15 CIFAR-DDPM models with the original ImageNet-style stem. The rank-10 projection margin falls from 0.515 ± 0.004 to 0.134 ± 0.007 by generation 4; ranks 5 and 20 also decrease (Figure 9). Historical margins show the same direction, with a much smaller change in orientation-only subspace overlap (Appendix B). The margin depends on class centers and residual scales as well as subspace orientation.

On the same split, ridge readouts recover 3.60 ± 1.17 accuracy points over the original head at g4, but remain 22.20 ± 1.24 points below their own g0 performance. The decline separates into 15.12 ± 0.73 points at the first replacement and a further 7.08 ± 1.04 points during g1–g4 (Table 12). Later degradation therefore persists under this alternative decoder. This analysis uses 4,900 fitting labels and internal cross-validation; it proves neither irreversible information loss nor optimal linear separability. Its 4,900 evaluation rows differ from the 9,800-row geometry endpoint.

5.7 Estimation and monitoring decisions

On historical CIFAR features, 20 labels per class estimate normalized R against a disjoint reference with $11.6 \pm 1.0\%$ MAPE; 50 per class give $7.6 \pm 1.2\%$ (Table 13). The same 20-label subsets estimate accuracy with 2.50-percentage-point mean absolute error. These different targets and units establish neither a superior diagnostic nor earlier warning. The subset study measures concurrent estimation; the following decision studies use fixed 200-image audit sets on newly trained models (Appendix D).

Under full replacement, scatter, accuracy, ECE and NLL all alarm at g1; geometry- and accuracy-triggered replay follow identical paths. A selected six-seed replication initially gives geometry-triggered switching 1.27 ± 1.74 accuracy points over accuracy-triggered switching, but consumes more real rows and equals unconditional generation-2 switching in every seed (Table 34). Common calibration reduces the gain to 0.14 ± 0.35 points: one win, five ties, still with unequal real-data ex-

posure and fixed-switch equivalence (Tables 37 and 41). On 60 held-out real-retraining pairs, scatter alarms in 2/60 and accuracy in 14/60; the realized operating points remain unmatched (Table 39). All exploratory outcomes, including favorable retrospective ROC rankings (Table 43), remain in Appendix F. These results support measuring geometry for interpretation; they do not establish that it improves decisions.

5.8 Real-data retention and joint retraining

At generation 6, Mix-25%, Mix-50% and Mix-75% reach $R = 0.840$, 0.660 and 0.527, with accuracies 77.74%, 83.11% and 86.31%. Accumulate reaches $R = 0.587$ and 85.67% accuracy; all are less degraded than replacement (Appendix D, Figure 5). Mix-75% scatter rises 24.9% and accuracy falls 2.25 points from its own baseline. These different units do not establish comparative sensitivity. Two retained-data seeds and accumulation's larger training budget limit intervention comparisons.

In joint DDPM–classifier retraining, mean R rises from 0.501 to 7.700 and rank falls from 12.91 to 3.74 by generation 6; accuracy reaches 15.78%. The generation-1 rank increase precedes classifier training on a fine-tuned generator's outputs. Geometric and dimensional degradation coexist, but unmeasured generated-image distributions prevent attribution to generator mode collapse.

6 Discussion and conclusion

Within- and between-class component entropies can both fall while total covariance entropy rises. In the CIFAR-DDPM trajectories, changes in component variance weights and the entropy-mixing gap offset component concentration; concentration dominates in the GAN and ImageNet workflows. This accounting describes the covariance balance of learned representations; the learning mechanism that produces it remains unidentified.

With the original stem, later inherited-teacher models lose accuracy and ridge-readout performance as entropy rank rises; fixing C_0 avoids the additional loss on the same image pools. A standard CIFAR stem retains the full-trajectory mismatch from a 95.4% baseline, but later rank decreases in two of three seeds. Teacher choice also changes target marginals, leaving its effect on label-error rates unidentified. Participation ratio and stable rank can have different signs, and scatter itself has accuracy counterexamples. Evidence remains image-only, with three-seed primary comparisons and finite horizons.

Small labeled sets estimate concurrent geometry, but the tested geometry-triggered policies do not establish better decisions than accuracy or unconditional retention. The results support class-conditioned geometry as a tool for interpreting global rank alongside predictive performance. Prospective comparisons at comparable false-alarm rates and real-data budgets remain necessary to establish its value for choosing interventions.

AI USE STATEMENT

The authors used AI assistants, including ChatGPT and Codex, to support literature searches, hypothesis and claim development, proof drafting and checking, experimental design, data preparation, method implementation, plotting code, result interpretation, and manuscript drafting and editing. All authors reviewed and verified the mathematical claims and proofs. The diffusion and GAN models in Appendices A and F generated the synthetic training datasets. Verification procedures are documented in Appendix G. The authors take responsibility for the final text, claims, references, code and experimental artifacts.

ETHICS STATEMENT

This work studies existing image benchmarks and synthetic training data. Its aggregate rank and scatter measures do not certify class-wise fairness, privacy, image fidelity or deployment safety. A stable aggregate score can obscure failures affecting particular classes, and the trusted labels used for auditing need to reflect the intended evaluation population. We report adverse monitoring and intervention results alongside favorable geometric observations; the evidence does not justify autonomous intervention on the basis of geometry alone. Dataset and model use remains subject to the respective providers' terms; the accompanying source artifact documents data preparation without redistributing benchmark images or pretrained generator weights.

REPRODUCIBILITY STATEMENT

Appendices A–G specify training configurations, data partitions, metric definitions, covariance proofs, sampling procedures and artifact lineage. The accompanying anonymous source artifact contains executable analyses, frozen training-source snapshots, verification records, and scripts that regenerate the reported tables and figures. Appendices C and F detail the stage, rank-definition and input-stem checks. Historical measurements, fresh runs, post-outcome analyses and protocol corrections are explicitly distinguished; artifact checks document what can be reproduced from the included records and what requires the original data and checkpoints.

REFERENCES

Sina Alemohammad, Josue Casco-Rodriguez, Lorenzo Luzi, Ahmed Imtiaz Humayun, Hossein Babaei, Daniel LeJeune, Ali Siahkoohi, and Richard G. Baraniuk. Self-Consuming Generative Models Go MAD. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/ebc042e767de551803ccfcc45e2454f5-Abstract-Conference.html.

Eric Arazo, Diego Ortego, Paul Albert, Noel E. O'Connor, and Kevin McGuinness. Pseudo-Labeling and Confirmation Bias in Deep Semi-Supervised Learning. In *Proceedings of the International Joint Conference on Neural Networks*, pp. 1–8, 2020. doi: 10.1109/IJCNN48605.2020.9207304. URL https://doi.org/10.1109/IJCNN48605.2020.9207304.

Patrick Batsell, Thomas Walker, and Richard Baraniuk. A Geometric Perspective on Recursive Synthetic Training. In *2nd DeLTa Workshop at ICLR*, 2026. URL https://raw.githubusercontent.com/PatrickBats/Synthetic-Training-Geometry/9dd881ef3ca48fe242cf27a429dfc77bf5fbe87b/paper/A_Geometric_Perspective_on_Recursive_Synthetic_Training.pdf. Non-archival workshop paper.

C. J. Clopper and E. S. Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934. doi: 10.1093/biomet/26.4.404. URL https://doi.org/10.1093/biomet/26.4.404.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848. URL https://www.image-net.org/static_files/papers/imagenet_cvpr09.pdf.

Elvis Dohmatob, Yunzhen Feng, Arjun Subramonian, and Julia Kempe. Strong Model Collapse. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/284afdc2309f9667d2d4fb9290235b0c-Abstract-Conference.html.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Shi Fu, Yingjie Wang, Yuzhu Chen, Li Shen, and Dacheng Tao. Self-Verification Provably Prevents Model Collapse in Recursive Synthetic Training. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-1213. URL https://papers.neurips.cc/paper_files/paper/2025/hash/3380e8116452e0efbf36f35d95e88c94-Abstract-Conference.html.

Quentin Garrido, Randall Balestriero, Laurent Najman, and Yann LeCun. RankMe: Assessing the Downstream Performance of Pretrained Self-Supervised Representations by Their Rank. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 10929–10974. PMLR, 2023. URL https://proceedings.mlr.press/v202/garrido23a.html.

Matthias Gerstgrasser, Rylan Schaeffer, Apratim Dey, Rafael Rafailov, Tomasz Korbak, Henry Sleight, Rajashree Agrawal, John Hughes, Dhruv Bhandarkar Pai, Andrey Gromov, Dan Roberts, Diyi Yang, David L. Donoho, and Sanmi Koyejo. Is Model Collapse Inevitable? Breaking the Curse of Recursion by Accumulating Real and Synthetic Data. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=5B2K4LRgmz.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On Calibration of Modern Neural Networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330. PMLR, 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Ting Han, Linara Adilova, Henning Petzka, Jens Kleesiek, and Michael Kamp. Flatness is Necessary, Neural Collapse is Not: Rethinking Generalization via Grokking. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-3561. URL https://proceedings.nips.cc/paper_files/paper/2025/hash/99c85e0fac93cff6b4d40ce7e67eda84-Abstract-Conference.html.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016. URL https://openaccess.thecvf.com/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html.

Jesse Hemerik and Jelle Goeman. Exact testing with random permutations. *TEST*, 27(4):811–825, 2018. doi: 10.1007/s11749-017-0571-1. URL https://link.springer.com/article/10.1007/s11749-017-0571-1.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html.

Li Jing, Pascal Vincent, Yann LeCun, and Yuandong Tian. Understanding Dimensional Collapse in Contrastive Self-supervised Learning. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=YevsQ05DEN7.

Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training Generative Adversarial Networks with Limited Data. In *Advances in Neural Information Processing Systems*, volume 33, pp. 12104–12114, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/8d30aa96e72440759f74bd2306c1fa3d-Abstract.html.

Joshua Kazdan, Rylan Schaeffer, Apratim Dey, Matthias Gerstgrasser, Rafael Rafailov, David L. Donoho, and Sanmi Koyejo. Collapse or Thrive: Perils and Promises of Synthetic Data in a Self-Generating World. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 29469–29494. PMLR, 2025. URL https://proceedings.mlr.press/v267/kazdan25a.html.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of Neural Network Representations Revisited. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3519–3529. PMLR, 2019. URL https://proceedings.mlr.press/v97/kornblith19a.html.

Alex Krizhevsky. Learning Multiple Layers of Features from Tiny Images. Technical report, University of Toronto, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A ConvNet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11976–11986, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Liu_A_ConvNet_for_the_2020s_CVPR_2022_paper.html.

Jianfeng Lu and Stefan Steinerberger. Neural collapse under cross-entropy loss. *Applied and Computational Harmonic Analysis*, 59:224–241, 2022. doi: 10.1016/j.acha.2021.12.011. URL https://doi.org/10.1016/j.acha.2021.12.011.

A. P. Majtey, P. W. Lamberti, and D. P. Prato. Jensen-Shannon divergence as a measure of distinguishability between mixed quantum states. *Physical Review A*, 72(5):052310, 2005. doi: 10.1103/PhysRevA.72.052310. URL https://doi.org/10.1103/PhysRevA.72.052310.

Vardan Papyan, X. 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. doi: 10.1073/pnas.2015509117. URL https://doi.org/10.1073/pnas.2015509117.

William Peebles and Saining Xie. Scalable Diffusion Models with Transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4195–4205, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Peebles_Scalable_Diffusion_Models_with_Transformers_ICCV_2023_paper.html.

Arjun Prakash, Naicheng He, Kaicheng Guo, Saket Tiwari, Tyrone Serapio, Ruo Yu Tao, Amy Greenwald, and George Konidaris. Spectral Collapse Drives Loss of Plasticity in Deep Continual Learning. In *International Conference on Machine Learning*, 2026. URL https://icml.cc/virtual/2026/poster/64388.

Olivier Roy and Martin Vetterli. The Effective Rank: A Measure of Effective Dimensionality. In *Proceedings of the 15th European Signal Processing Conference*, pp. 606–610, 2007. URL https://zenodo.org/records/40328.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115 (3):211–252, 2015. doi: 10.1007/s11263-015-0816-y. URL https://doi.org/10.1007/s11263-015-0816-y.

Lianghe Shi, Meng Wu, Huijie Zhang, Zekai Zhang, Molei Tao, and Qing Qu. A Closer Look at Model Collapse: From a Generalization-to-Memorization Perspective. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-1357. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/39d3b980155d3d52613e412b018004a4-Abstract-Conference.html.

Tushar Shinde and Avinash Kumar Sharma. Do Coresets, Pruning, and Quantization Preserve Neural Network Representations? In *Proceedings of GRaM: the Second Edition of the Workshop on Geometry-grounded Representation Learning and Generative Modeling*, volume 326 of *Proceedings of Machine Learning Research*, pp. 489–496. PMLR, 2026. URL https://proceedings.mlr.press/v326/shinde26a.html.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. AI models collapse when trained on recursively generated data. *Nature*, 631(8022): 755–759, 2024. doi: 10.1038/s41586-024-07566-y. URL https://doi.org/10.1038/s41586-024-07566-y.

Ananda Theertha Suresh, Andrew Thangaraj, and Aditya Nanda Kishore Khandavally. Rate of Model Collapse in Recursive Training. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 1396–1404. PMLR, 2025. URL https://proceedings.mlr.press/v258/suresh25a.html.

Quynh Vo and Cong-Duy T Nguyen. Collapse-Aware Regularization for Reliable Reasoning Under Distribution Shift. In *Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence*, volume 337 of *Proceedings of Machine Learning Research*, pp. 6999–7022. PMLR, 2026. URL https://proceedings.mlr.press/v337/vo26a.html.

Zhihui Zhu, Tianyu Ding, Jinxin Zhou, Xiao Li, Chong You, Jeremias Sulam, and Qing Qu. A Geometric Analysis of Neural Collapse with Unconstrained Features. In *Advances in Neural Information Processing Systems*, volume 34, pp. 29820–29834, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/f92586a25bb3145facd64ab20fd554ff-Abstract.html.

A Training and evaluation protocols

CIFAR data and ResNet. CIFAR-10 and CIFAR-100 each provide 50,000 real training and 10,000 test images. Their test sets contain 1,000 and 100 examples per class, respectively. Every geometric measurement uses the same ordered real evaluation set within a run, with trusted ground-truth labels. Test labels are not used to create the classifier's synthetic training targets. Training uses a random 32×32 crop with padding 4, horizontal flip, and normalization with mean $(0.4914, 0.4822, 0.4465)$ and standard deviation $(0.2470, 0.2435, 0.2616)$. The same normalization is used for CIFAR-100. Test and pseudo-label inputs use tensor conversion and normalization without random augmentation.

The ResNet-18 implementation retains the torchvision stem (a 7×7, stride-2 convolution and max-pooling), rather than replacing it with a CIFAR-specific stem. Its feature is the 512-dimensional pooled vector before the linear classifier. Replacement and retained-real-data experiments use 200 epochs, batch size 256, SGD with learning rate 0.1, momentum 0.9, weight decay 5×10^{-4}, cosine learning-rate decay and automatic mixed precision. Each generation is initialized independently with seed $s + t$, for base seed $s \in \{42, 43, 44\}$ in replacement and $s \in \{42, 43\}$ in retention. The checkpoint and features come from the final epoch. Historical `test_accuracy` fields denoted the best epoch; all present comparisons use `final_accuracy`. Predictive diagnostics for local feature reanalysis are reconstructed from the saved final linear head and feature matrix.

ConvNeXt. The main ConvNeXt-Tiny configuration uses a 3×3, stride-1 input convolution, random initialization, AdamW with learning rate 2×10^{-4} and weight decay 0.05, batch size 128, 200 epochs and cosine decay. The 768-dimensional features are taken after global average pooling and the classifier's LayerNorm, before the linear head. The separately reported weak-initialization configuration uses 100 epochs and learning rate 10^{-3} with the same batch size and optimizer. These are distinct run families, not additional seeds for one configuration. Changing both learning rate and epochs prevents attribution of their difference to either factor alone.

Synthetic CIFAR images. Diffusion-based CIFAR studies use the pretrained unconditional checkpoint `google/ddpm-cifar10-32`; 50 denoising steps generate 50,000 images per generation.

Each image receives the preceding classifier's argmax label. The use of this CIFAR-10 generator for CIFAR-100 creates an image-domain mismatch that is inseparable from differences in the classification task in the present comparison. A fixed generator fixes its sampling distribution, but neither fixes the sample realization nor establishes that labels are the only degrading component of the learning pipeline. The archived checkpoint and configuration records, rather than an unpinned download of a model with the same name, define the historical experiment.

Data retention. Mix-α randomly selects $50{,}000\alpha$ original real examples and $50{,}000(1-\alpha)$ examples from the latest synthetic set without replacement within each set. It does not balance the training classes. Accumulate uses all 50,000 original real examples and all synthetic sets generated so far, preserving the labels originally assigned to each synthetic example. Its training size is $50{,}000(t+1)$ at generation t. At 200 epochs this increases total optimizer steps; reported improvements cannot be interpreted as compute-matched comparisons. The historical random choices may depend on interruption/resume history; current code isolates the mixture-sampling RNG for future runs.

Historical ImageNet probe. The subset consists of the first 100 ImageNet synsets in the explicit sorted list in `code/data/imagenet.py`. Generation 0 uses 129,395 real training images. The probe is the last 50 training images per class in the loader's filename order, totaling 5,000 images; those examples were *not excluded* from generation-0 training in the historical code. Probe features are stored in reverse class order because the selection loop traverses the dataset backward. The same ordering is preserved over generations. Accuracy on this probe is a training-origin measurement, not official ImageNet validation accuracy. Later training sets have 50,000 synthetic images, introducing a sample-count change as well.

The classifier is a randomly initialized timm ViT-S/16 with 384-dimensional features, trained for 90 epochs with batch size 64, SGD at learning rate 10^{-3}, momentum 0.9, weight decay 5×10^{-4} and cosine decay. The classifier training path uses random resized crops to 224 pixels, horizontal flips and ImageNet normalization; probe extraction uses resize to 256, center crop to 224 and the same normalization. These runs did not use AdamW, RandAugment or Mixup.

DiT-XL/2, `facebook/DiT-XL-2-256`, samples 500 images for each of conditioning indices 0–99 using 50 steps. The recorded generator revision is `eab87f77abd5aef071a632f08807fbaab0b704d0`. All synthetic training labels, including those for the first synthetic classifier, are overwritten by classifier predictions. In the historical pseudo-label path, CIFAR normalization and a direct resize were used instead of the ImageNet evaluation transforms. This mismatch is another limitation of that probe experiment. The separate fresh experiment below corrects preprocessing and uses official validation; it does not retroactively correct the historical probe.

Joint retraining. The CIFAR-10 self-consuming loop uses three base seeds and seven classifier generations. The classifier uses the ResNet settings above with batch size 128. Beginning after the first synthetic set, the DDPM is fine-tuned for 50 epochs on that synthetic image set, starting from the preceding generator checkpoint when available. Fine-tuning uses batch size 64, AdamW with learning rate 10^{-5}, weight decay 10^{-6}, cosine decay, noise-prediction mean-squared error and gradient clipping at 1. Generation 1's classifier still trains on images from the original pretrained DDPM. The first classifier exposed to outputs of a fine-tuned generator is generation 2. We report the whole trajectory rather than treating the first transition as evidence about a retrained generator.

B METRIC DEFINITIONS AND NUMERICAL CONVENTIONS

All revised covariance computations use double precision for saved-feature analysis. Negative eigenvalues from round-off are clipped to zero, and eigenvalues below machine epsilon times the largest eigenvalue are omitted from entropy. Population covariance normalization $1/N$ is used; a common rescaling of covariance leaves rank and the trace ratio invariant. The total-covariance identity uses sample class priors. Singleton classes contribute zero within-class scatter and their actual mass to between-class scatter. The diagnostic ratio is reported as infinite when its between-class denominator vanishes. On nondegenerate features, the numerical implementation is invariant to translation, common scale and feature-coordinate rotations; it is not invariant to arbitrary nonspherical rescaling of coordinates.

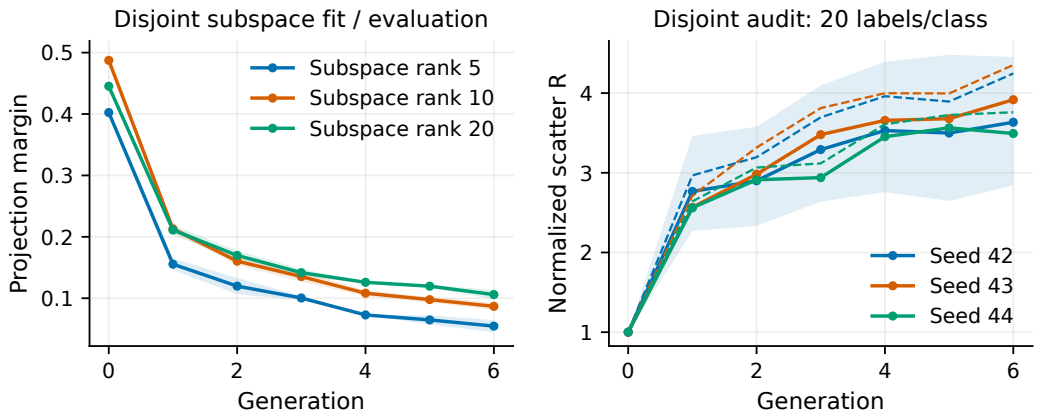

Figure 3: **Geometry beyond a global spectrum.** Left: projection margins on examples disjoint from those used to fit class subspaces; bands indicate one standard deviation across three training seeds. Right: the 20-label-per-class audit against a disjoint reference, shown separately for each seed. Solid lines are subset means; dashed lines are reference trajectories. The shaded region for seed 42 is its 2.5th–97.5th percentile across subset draws, not uncertainty across training seeds.

NC2 and NC3. Let $A \in \mathbb{R}^{K\times d}$ have rows $\mu_k - \mu$ and let H contain the corresponding classifier weight rows. We use

$$\mathrm{NC2} = \left\| \frac{AA^\top}{\|AA^\top\|_F} - \frac{I_K - \mathbf{1}\mathbf{1}^\top/K}{\sqrt{K-1}} \right\|_F, \tag{7}$$

$$\mathrm{NC3} = \left\| \frac{H}{\|H\|_F} - \frac{A}{\|A\|_F} \right\|_F. \tag{8}$$

NC2 combines class-mean equinorm and equiangular structure through the normalized Gram matrix. NC3 uses one Frobenius normalization per matrix, rather than normalizing each class separately. Classifier biases are not part of this alignment measure. All experiments in the reported NC2/NC3 comparison use balanced evaluation classes. Zero-norm configurations are explicitly flagged in the code instead of interpreted as ideal neural collapse.

Network–nearest-center agreement is measured by fitting class centers on alternating examples within each class and evaluating the classifier's predictions and Euclidean nearest-center decisions on the other half. Agreement is not accuracy: two decision rules can agree on wrong labels. Its value declines from 0.993 to 0.849 across the CIFAR-10 replacement trajectory.

Subspace fitting. For each class, alternating test examples in the saved ordering form a 500/500 fit/evaluation split. A randomized SVD fits the centered class-specific subspace, with four power iterations and random state $123 + k$ for class k. The fit mean is the affine offset. With basis U_k, residuals are $e_k(x) = \max\{\|x - \mu_k\|^2 - \|U_k^\top(x - \mu_k)\|^2, 0\}$; the maximum guards against numerical round-off. Pairwise overlap is $\|U_i^\top U_j\|_F^2/q$, averaged over class pairs. This orientation-only quantity differs from the affine projection margin. The exact same fit/evaluation rule applies to all generations, and no evaluated example fits its own class subspace.

Predictive quantities. For saved features x_i and final classifier (H, b), logits are $Hx_i + b$ and probabilities are their softmax. NLL is $-N^{-1}\sum_i \log p_i(y_i)$ in natural-log units. ECE uses 15 equal-width confidence bins and the weighted absolute difference between mean confidence and accuracy per nonempty bin. Neither temperature fitting nor model selection uses the audit labels. Limited-sample ECE has sampling and binning effects and is not treated as an unbiased population estimate. CKA, when used by the pipeline, centers both feature matrices and compares the same ordered examples, following Kornblith et al. (2019).

Historical subspace and audit trajectories. The rank-10 projection margin decreases from 0.487 ± 0.009 to 0.087 ± 0.006, an 82.2% reduction (Figure 3). Rank-5 and rank-20 fits give

Table 6: CIFAR-10 replacement diagnostics, means over three seeds. Accuracy and ECE are percentages. These are concurrent descriptors, not standardized effect sizes. Per-seed measurements and variances are retained in the analysis JSON.

t	R	r	NC2	NC3	Acc. (%)	ECE (%)	NLL
0	0.424	12.279	0.304	0.247	88.483	7.576	0.504
1	1.159	14.777	0.580	0.493	70.877	20.489	1.481
2	1.318	14.771	0.601	0.486	66.823	24.664	1.854
3	1.464	14.993	0.594	0.477	63.743	28.197	2.195
4	1.601	15.543	0.627	0.496	60.750	31.078	2.529
5	1.610	14.706	0.629	0.495	58.613	33.942	2.839
6	1.686	14.415	0.648	0.509	56.267	36.661	3.160

reductions of 86.3% and 76.2%. Average pairwise rank-10 subspace overlap changes from 0.738 to 0.768, a much smaller shift. The loss in margin therefore should not be described as wholesale convergence of all class subspaces to one orientation: affine centers, residual scales and the nearest competing class also enter the margin. Centered class-mean and classifier-alignment deviations worsen, and split-sample network–nearest-center agreement decreases from 99.3% to 84.9% (Appendix B). These measurements characterize distinct aspects of class structure; they do not independently establish its cause.

C Covariance decomposition and diagnostic counterexamples

Relabeling fixed features. Let $X \in \mathbb{R}^d$ be square-integrable with true class Y and corrupted class Z. Assume a class-conditional label channel: $X \perp Z \mid Y$. Write $W_Y = \mathbb{E}[\mathrm{Cov}(X \mid Y)]$, $B_Y = \mathrm{Cov}(\mathbb{E}[X \mid Y])$, and define W_Z, B_Z analogously. These are population quantities for the *same* representation X.

Proposition 1 (Transfer between scatter components)**.** *Under the stated conditional independence, set* $\Delta = \mathbb{E}[\mathrm{Cov}(\mathbb{E}[X \mid Y] \mid Z)] \succeq 0$*. Then*

$$W_Z = W_Y + \Delta, \qquad B_Z = B_Y - \Delta, \qquad W_Z + B_Z = W_Y + B_Y = \mathrm{Cov}(X). \tag{9}$$

Consequently, every statistic depending only on $\mathrm{Cov}(X)$*, including its effective rank, is invariant under this relabeling.*

Proof. Put $m(Y) = \mathbb{E}[X \mid Y]$. Conditional independence gives $\mathbb{E}[X \mid Y, Z] = m(Y)$ and $\mathrm{Cov}(X \mid Y, Z) = \mathrm{Cov}(X \mid Y)$. Applying total covariance conditionally on Z and then averaging yields

$$\mathbb{E}[\mathrm{Cov}(X \mid Z)] = \mathbb{E}[\mathrm{Cov}(X \mid Y)] + \mathbb{E}[\mathrm{Cov}(m(Y) \mid Z)].$$

Thus $W_Z = W_Y + \Delta$. Also $\mathbb{E}[X \mid Z] = \mathbb{E}[m(Y) \mid Z]$, so total covariance applied to $m(Y)$ gives $B_Y = B_Z + \Delta$. The marginal law of X is unchanged. □

Let $w = \mathrm{tr}\, W_Y$, $b = \mathrm{tr}\, B_Y$ and $\delta = \mathrm{tr}\, \Delta$. When $b > 0$ and $0 \le \delta < b$,

$$R_Z - R_Y = \frac{w + \delta}{b - \delta} - \frac{w}{b} = \frac{\delta(b + w)}{b(b - \delta)} \ge 0. \tag{10}$$

Strict increase requires $\delta > 0$, not merely a nonzero label-error rate. A permutation of class names can have many label errors with $\Delta = 0$. If $\delta = b > 0$, between-class scatter vanishes and the ratio diverges. For balanced classes and symmetric corruption probability ε, the special case is $B_Z = a^2 B_Y$ with $a = 1 - K\varepsilon/(K-1)$, and $R_Z = (R_Y + 1 - a^2)/a^2$ when $a \neq 0$ (Appendix C). This formula does not assert monotonicity over arbitrary channels or over the full noise-probability range.

Proposition 1 requires only a finite second moment and class-conditional corruption, not Gaussianity, equal class priors or isotropic class covariances. Equation (10) follows by bringing $(w + \delta)/(b - \delta) - w/b$ over the common denominator $b(b - \delta)$. Because $B_Z \succeq 0$, the trace transfer satisfies $0 \le \delta \le b$. If $b = 0$, the initial ratio itself is undefined or infinite and the finite-ratio comparison does not apply.

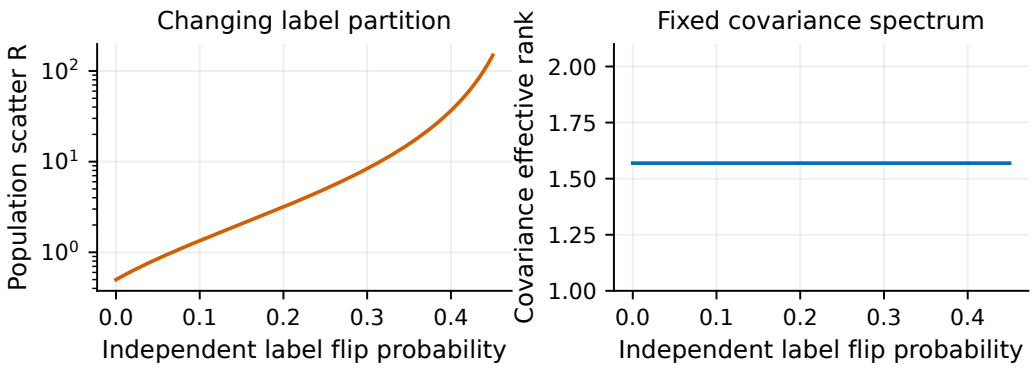

Figure 4: Exact binary population example. Label-conditioned scatter deteriorates while the marginal feature covariance, its eigenvalues and its effective rank stay fixed. The scatter axis is logarithmic. The assumptions concern relabeling of fixed features and do not describe a neural-network update.

For the balanced symmetric special case, let $P(Y = k) = 1/K$ and

$$P(Z = j \mid Y = k) = \begin{cases} 1 - \varepsilon, & j = k, \\ \varepsilon/(K-1), & j \neq k. \end{cases}$$

Then $P(Z = j) = 1/K$, and Bayes' rule gives the same weights for $P(Y = k \mid Z = j)$. If $m_k = \mathbb{E}[X \mid Y = k]$ and $\mu = K^{-1} \sum_k m_k$, conditional independence yields

$$\begin{aligned} \mathbb{E}[X \mid Z = j] - \mu &= (1 - \varepsilon) m_j + \frac{\varepsilon}{K-1} \sum_{k \neq j} m_k - \mu \\ &= \left(1 - \frac{K\varepsilon}{K-1}\right)(m_j - \mu) = a(m_j - \mu). \end{aligned}$$

Thus $B_Z = a^2 B_Y$ and $W_Z = W_Y + (1 - a^2) B_Y$. Taking traces proves the stated ratio. At $\varepsilon = (K-1)/K$, $a = 0$ and Z is independent of Y; all class-conditional means coincide. Beyond that point, a^2 can increase again. Binary label reversal at $\varepsilon = 1$ is a permutation and recovers the original scatter ratio. A positive classification-label error rate therefore cannot by itself imply strictly increasing scatter.

For the plotted binary example, $W_Y = I_2$, $B_Y = 4 e_1 e_1^\top$, and $\Sigma = \mathrm{diag}(5, 1)$. Hence $R_Y = 2/4 = 1/2$, $R_Z = (1/2 + 1 - a^2)/a^2 = 1.5/(1 - 2\varepsilon)^2 - 1$, and $\mathrm{erank}(\Sigma) = \exp[-(5/6)\log(5/6) - (1/6)\log(1/6)]$ for every flip probability. These are exact population formulas, not estimates from a trained classifier.

Changing representations with fixed labels. The blind spot also admits an explicit construction with unchanged input data and labels. Let Y be uniform on $\{-1, +1\}$ and let E_1, E_2, E_3 be independent standard normals, independent of Y. Fix the observed input $U = (2Y + E_1, E_2, E_3)^\top$. For $\rho \in [0, 1]$, consider the two-dimensional linear representation

$$X_\rho = f_\rho(U) = \left(\rho U_1 + \sqrt{5(1 - \rho^2)}\, U_2,\ U_3\right)^\top. \tag{11}$$

The map uses only U; the joint law of (U, Y) is the same for every ρ. Conditional on Y, the mean is $2\rho Y e_1$ and the common covariance is $\mathrm{diag}(5 - 4\rho^2, 1)$, because the noise variance in the first coordinate is $\rho^2 + 5(1 - \rho^2)$. Consequently,

$$W_\rho = \mathrm{diag}(5 - 4\rho^2, 1), \qquad B_\rho = 4\rho^2 e_1 e_1^\top, \qquad \Sigma_\rho = W_\rho + B_\rho = \mathrm{diag}(5, 1). \tag{12}$$

The full covariance and its effective rank are constant. For $\rho > 0$, the equal-prior Gaussian classes have a Bayes decision boundary at $X_{\rho,1} = 0$, giving

$$R_\rho = \frac{6 - 4\rho^2}{4\rho^2} = \frac{3}{2\rho^2} - 1, \qquad \mathcal{A}_\rho = \Phi\left(\frac{2\rho}{\sqrt{5 - 4\rho^2}}\right), \tag{13}$$

where Φ is the standard-normal CDF and $\mathcal{A}_\rho$ is Bayes-optimal accuracy. The argument of Φ has derivative $10/(5-4\rho^2)^{3/2} > 0$ on $[0,1]$. Thus, as ρ decreases from 1 to 0, accuracy decreases from $\Phi(2)$ to $1/2$, while the scatter ratio grows from $1/2$ without bound. At $\rho = 0$, X_ρ is independent of Y and between-class scatter is zero. This is an existence construction with an externally chosen representation parameter. It keeps the covariance fixed while allowing higher moments to change and specifies no recursive learning rule, alarm lead time or trajectory followed by the trained classifiers.

Without conditional independence. The correction for feature-dependent labels can also be written exactly. Let $m(Y) = \mathbb{E}[X \mid Y]$, $\mu = \mathbb{E}[X]$, $a(Z) = \mathbb{E}[m(Y) \mid Z] - \mu$ and $r(Z) = \mathbb{E}[X - m(Y) \mid Z]$. Both a and r have zero mean, and $\mathbb{E}[X \mid Z] - \mu = a(Z) + r(Z)$. Define $B_{\mathrm{cc}} = \mathbb{E}[aa^\top]$, $H = \mathbb{E}[rr^\top]$ and $J = \mathbb{E}[ar^\top]$. Then

$$B_Z = B_{\mathrm{cc}} + H + J + J^\top, \qquad W_Z = W_Y + \Delta - H - J - J^\top, \tag{14}$$

where $\Delta = B_Y - B_{\mathrm{cc}} = \mathbb{E}[\mathrm{Cov}(m(Y) \mid Z)] \succeq 0$. Expanding $(a+r)(a+r)^\top$ gives the first identity, and subtracting it from $\mathrm{Cov}(X) = W_Y + B_Y$ gives the second. Conditional independence sets $r = 0$, recovering Proposition 1; in general the correction $H + J + J^\top$ need not have a fixed sign. This is an identity for fixed features, not a model of representation learning.

Empirical assumption check. After the preceding outcomes, we evaluated all 15 fresh CIFAR replacement models, nine original SGD ImageNet models and six matched real-only controls. Here Y is the official real-image label and Z the same checkpoint's prediction on that image, rather than a synthetic training target from its predecessor. We compare the observed partition with a hypothetical class-conditional channel having identical true/predicted confusion counts. Its class-conditional feature mixtures give B_{cc} exactly. A finite-sample control independently permutes the observed Z assignments within each true class 100 times per model, preserving the full confusion matrix and error rate. If n_{kj} is the confusion count, n_k the true-class size and n_j the predicted-class size, the exact expected permutation between trace is $\mathrm{tr}\, B_{\mathrm{cc}} + v$, where

$$v = \frac{1}{N} \sum_{j:n_j>0} \frac{1}{n_j} \sum_k \frac{n_{kj}(n_k - n_{kj})}{n_k - 1} \mathrm{tr}\, \widehat{\Sigma}_k. \tag{15}$$

Here $\widehat{\Sigma}_k$ uses divisor n_k and all evaluated $n_k > 1$. The expression follows from the variance of a size-n_{kj} sum sampled without replacement within true class k. Exhaustive enumeration of 12 assignments in an unequal-class toy example verifies the correction.

In all 30 fixed models, prediction-labeled scatter is below true-label scatter, and the actual between trace exceeds all 100 within-class permutation values. This also occurs in the real-only controls, so it is not specific to recursive degradation. On the ImageNet replacement trajectory, mean true-label scatter rises from 5.731 to 9.811 while prediction-labeled scatter remains 3.626 to 3.620; the class-conditional constructions instead give 16.030 and 51.183. The finite-population correction is substantial for ImageNet, and is reported separately (Table 7). Thus replacing actual prediction errors by a class-conditional channel does not reproduce their measured geometry, and self-labeling can make geometry appear more favorable. All 30 checkpoint heads, ordered arrays, full covariance identities and 3,000 permuted partitions were independently verified. These are conditional, post-outcome fixed-model comparisons, not independent training replicates, population significance tests, a usable label-free correction or an explanation of retraining dynamics. They do not establish the true labels of generated images. The original trusted-label results are unchanged.

Entropy accounting across learned representations. Section 4 defines the entropy terms and the exact finite-change decomposition. For $\rho_A = A/\,\mathrm{tr}\, A$, the identity $\rho_\Sigma = p\rho_B + (1-p)\rho_W$ makes $\mathcal{J}$ the weighted matrix Jensen–Shannon quantity. Apply the product identity to pH_B and $(1-p)H_W$ in Equation (5); collecting the two Δp terms proves Equation (6). No Gaussianity, class-conditional corruption, isotropy or neural learning model is assumed. Each component is invariant to a common orthogonal coordinate change and global feature scaling. No alignment of feature coordinates across generations is required. $\mathcal{J}$ depends on relative orientation, spectra and weights together; D_J is not a pure orientation effect. A zero-trace B or W makes its normalized entropy undefined and is rejected by the implementation. None occurs in this cohort.

The post-outcome cohort fixes all seeds 101/102/103 and all generations of the completed ResNet workflows: CIFAR-DDPM g0–g4, CIFAR-GAN g0–g4, ImageNet-DiT g0–g5, and both one-step

Table 7: Post-outcome fixed-feature label audit, mean ± sample SD across three seeds. R_Y uses official labels; R_Z uses the same classifier's predictions. R_{cc} is the class-conditional mixture with the identical confusion counts. R^*_{perm} evaluates the ratio at the exact expected between scatter under within-true-class label permutations; it is not the mean of the ratios. All 30 models and 100 permutations per model are recorded; the table shows trajectory endpoints and real-only controls. Real-only denotes an independently initialized generation-1 retrain.

Dataset	Model	Acc. (%)	R_Y	R_Z	R_{cc}	R^*_{perm}
CIFAR-10	$g = 0$	88.35 ± 0.08	0.503 ± 0.002	0.366 ± 0.002	0.896 ± 0.003	0.896 ± 0.003
CIFAR-10	$g = 4$	62.17 ± 1.04	1.620 ± 0.048	0.724 ± 0.030	4.135 ± 0.122	4.125 ± 0.122
CIFAR-10	Real-only	88.48 ± 0.18	0.492 ± 0.005	0.359 ± 0.004	0.879 ± 0.010	0.879 ± 0.010
ImageNet-100	$g = 0$	34.96 ± 0.67	5.731 ± 0.092	3.626 ± 0.019	16.030 ± 0.494	12.645 ± 0.298
ImageNet-100	$g = 2$	15.67 ± 0.12	9.811 ± 0.132	3.620 ± 0.047	51.183 ± 1.260	26.493 ± 0.328
ImageNet-100	Real-only	34.91 ± 0.26	5.756 ± 0.007	3.626 ± 0.005	16.134 ± 0.104	12.708 ± 0.063

Table 8: Accounting for opposite rank changes on stored ResNet features: g0–g4 on CIFAR and g0–g5 on ImageNet. Values are paired g0-to-endpoint changes, means ± sample SD over three reused seeds, in nats. $D_p + D_B + D_W + D_J = \Delta \log \mathrm{erank}(\Sigma)$: variance-weight, between-spectrum, within-spectrum and entropy-mixing terms (Appendix C). This identity describes the measured covariances, not causal effects.

Workflow	$\Delta \log \mathrm{erank}$	D_p	D_B	D_W	D_J
CIFAR DDPM	$+0.136 \pm 0.020$	$+0.267 \pm 0.007$	-0.119 ± 0.002	-0.038 ± 0.011	$+0.025 \pm 0.002$
CIFAR GAN	-0.028 ± 0.006	$+0.080 \pm 0.003$	-0.021 ± 0.003	-0.080 ± 0.002	-0.007 ± 0.003
ImageNet DiT	-0.408 ± 0.006	$+0.270 \pm 0.003$	-0.123 ± 0.004	-0.470 ± 0.003	-0.086 ± 0.001

real-only controls. Reused g0 checkpoints give 51 distinct models and 60 workflow/generation rows. The cohort, input hashes and analysis/verification sources were frozen before computing component entropies, after the original rank, scatter and accuracy outcomes were known. We retain every generation and all 45 within-seed g0-to-later contrasts. CIFAR uses the 9,800 endpoint rows disjoint from the 200 audit rows; ImageNet uses all 5,000 official validation rows, retaining the disclosed near-duplicates. Generator comparisons and baselines retain their original confounding and reuse limitations.

The analyzer forms W from class-centered residuals and B from empirical-prior-weighted centered class means, using float64 and divisor n. It checks $W + B$ against directly centered total covariance. An independent verifier computes rectangular-factor SVDs for the total features, class residuals and weighted class means, avoiding the analyzer's covariance eigendecompositions. All 51 spectra, entropies, parent rank/scatter values and 45 contrasts agree; maximum entropy disagreement is below 6×10^{-14} nats. Previously verified checkpoints and ordered arrays identify these inputs; this analysis performs no new training or raw-image replay. Analytic tests cover unequal priors, singular components, noncommuting changes, orthogonal transforms, scaling, label permutation and undefined zero traces.

Table 8 reports the three synthetic primary contrasts; Table 9 retains both real-only controls. In all nine synthetic g0-to-final comparisons, both component entropies decrease, but decreasing between-class variance weight contributes positively to total entropy. Whether that reweighting offsets component concentration differs across workflows. This is descriptive covariance accounting, not evidence for isotropic noise injection, a causal generator-family effect, a universal trajectory or a new learning theorem. The artifact provides full endpoint and contrast CSVs and all component spectra; means and sample SD use the three existing seeds.

Rank definitions and direct stage comparisons. The subsequent analysis adds all nine verified fixed-C0 endpoints to the original 51-model cohort: 60 distinct models and 75 workflow/generation rows, with shared g0/g1 checkpoints explicitly reused. It retains every original generation, both real-only controls and all three seeds. These analyses were prompted by review after the original trajectories and an initial rank-sensitivity feasibility calculation were observed; the portable source

Table 9: Post-outcome entropy accounting for matched real-only g1 retrains. These are one-step controls, not real-data trajectories matched to four or five synthetic steps. Values are paired g0-to-endpoint changes, means ± sample SD over three reused seeds, in nats. $D_p + D_B + D_W + D_J = \Delta \log \mathrm{erank}(\Sigma)$: variance-weight, between-spectrum, within-spectrum and entropy-mixing terms (Appendix C). This identity describes the measured covariances, not causal effects.

Workflow	$\Delta \log \mathrm{erank}$	D_p	D_B	D_W	D_J
CIFAR real	-0.012 ± 0.012	-0.004 ± 0.002	-0.001 ± 0.002	-0.004 ± 0.008	-0.002 ± 0.001
ImageNet real	-0.002 ± 0.006	-0.006 ± 0.007	$+0.004 \pm 0.004$	$+0.001 \pm 0.003$	-0.000 ± 0.001

release is not a preregistration. It performs no new neural training or raw-image inference. The additional standard-stem training study is reported separately in Appendix F.

For singular values $s_j(A)$, write $r_s(A) = \exp[-\sum_j q_j \log q_j]$, where $q_j = s_j(A)/\sum_l s_l(A)$ and zero terms contribute zero. We compare $r_s(F)$, $r_s(F_c)$, the covariance rank in Equation (1), and

$$r_{\mathrm{PR}} = \frac{(\mathrm{tr}\,\Sigma)^2}{\mathrm{tr}(\Sigma^2)} = \frac{1}{\sum_j p_j^2}, \qquad r_{\mathrm{SR}} = \frac{\|F_c\|_F^2}{\|F_c\|_2^2} = \frac{1}{\max_j p_j}.$$

Thus singular-value entropy uses square roots of covariance eigenvalues after centering, whereas participation ratio and stable rank emphasize the largest variance directions differently. The reported stable rank belongs to the centered feature matrix, not to Σ as a matrix. The uncentered score follows RankMe's singular-value entropy definition (Garrido et al., 2023), applied here to supervised classifier penultimate features rather than self-supervised projector embeddings. Its primary value uses the continuous $\epsilon = 0$ limit. A separate sensitivity uses $q_j = s_j/\sum_l s_l + 10^{-7}$ without renormalization; this explicitly chosen stabilization is not claimed to reproduce an author implementation. It gives a mean CIFAR g0–g4 increase of 16.527%, with unchanged signs. Centered and uncentered entropy scores increase in all three CIFAR seeds; participation ratio decreases in two and stable rank in all three over that interval. Table 11 retains the other workflows and controls.

Every feature and label file is checked against its prior hash (120 files). Float64 eigendecompositions of total, within and between covariance are compared with direct rectangular-factor SVDs on all 60 models; the original 51 full spectra and all parent rank/scatter values are reproduced. Uncentered singular-value entropy is independently checked against the uncentered Gram matrix. These checks concern numerical definitions and retained inputs, not independent training replication. The artifact retains full component spectra, per-model definition scores, source identities and all intermediate generations.

For each workflow, stages are g0–g1, g1–final and g0–final; one-step real-only controls have only g0–g1. This gives 42 paired seed contrasts. Each uses the two endpoints' own averages in Equation (6). Total entropy differences add across consecutive intervals, but the separate midpoint attribution terms generally do not. Table 10 reports the direct calculations, including the fixed teacher. The inherited later-stage within-class entropy rises slightly for seed 102, from 2.98341 to 2.98583 nats, despite its decline on average. This exception is retained rather than generalized away. Table 12 similarly recomputes differences of the existing independently checked readout endpoints, without refitting or changing their cross-validation procedure.

D Audit sampling and additional conditions

Full-set audit. The earlier low-label analysis uses 200 stratified draws at each budget from the full CIFAR-10 test set, seeded with `20260724`. A subset is reused across generations for each draw. Subset and full-set trajectories are each normalized by their own generation-0 scatter ratio. At 20 examples per class, the mean post-baseline percentage error is 9.671%, with post-baseline Pearson correlation 0.725. Including the shared normalized baseline would raise the correlation to 0.925; that is not the reported correlation. Percentiles of seed-averaged subset trajectories quantify subsampling variation, not a confidence interval over independently trained models. Because the reference contains the subset, this is explicitly an overlapping-reference analysis.

Table 10: Direct endpoint entropy accounting for every retained workflow and stage. All values are nats, means ± sample SD across three seeds. Shared first-replacement rows are repeated for context, not independent replications. Each interval uses its own endpoint averages.

Workflow / interval	ΔH_Σ	D_p	D_B	D_W	D_J
CIFAR DDPM 0–1	$+0.103 \pm 0.022$	$+0.201 \pm 0.008$	-0.099 ± 0.014	-0.019 ± 0.013	$+0.021 \pm 0.003$
CIFAR DDPM 1–4	$+0.032 \pm 0.005$	$+0.069 \pm 0.014$	-0.021 ± 0.008	-0.020 ± 0.021	$+0.004 \pm 0.004$
CIFAR DDPM 0–4	$+0.136 \pm 0.020$	$+0.267 \pm 0.007$	-0.119 ± 0.002	-0.038 ± 0.011	$+0.025 \pm 0.002$
CIFAR GAN 0–1	-0.027 ± 0.007	$+0.050 \pm 0.004$	-0.014 ± 0.002	-0.053 ± 0.003	-0.010 ± 0.004
CIFAR GAN 1–4	-0.001 ± 0.009	$+0.028 \pm 0.007$	-0.007 ± 0.003	-0.025 ± 0.002	$+0.003 \pm 0.001$
CIFAR GAN 0–4	-0.028 ± 0.006	$+0.080 \pm 0.003$	-0.021 ± 0.003	-0.080 ± 0.002	-0.007 ± 0.003
ImageNet DiT 0–1	-0.306 ± 0.022	$+0.264 \pm 0.005$	-0.107 ± 0.002	-0.385 ± 0.016	-0.078 ± 0.002
ImageNet DiT 1–5	-0.101 ± 0.023	$+0.010 \pm 0.004$	-0.012 ± 0.003	-0.093 ± 0.019	-0.007 ± 0.000
ImageNet DiT 0–5	-0.408 ± 0.006	$+0.270 \pm 0.003$	-0.123 ± 0.004	-0.470 ± 0.003	-0.086 ± 0.001
CIFAR real 0–1	-0.012 ± 0.012	-0.004 ± 0.002	-0.001 ± 0.002	-0.004 ± 0.008	-0.002 ± 0.001
ImageNet real 0–1	-0.002 ± 0.006	-0.006 ± 0.007	$+0.004 \pm 0.004$	$+0.001 \pm 0.003$	-0.000 ± 0.001
CIFAR fixed C0 0–1	$+0.103 \pm 0.022$	$+0.201 \pm 0.008$	-0.099 ± 0.014	-0.019 ± 0.013	$+0.021 \pm 0.003$
CIFAR fixed C0 1–4	-0.037 ± 0.001	-0.008 ± 0.007	-0.005 ± 0.015	-0.021 ± 0.013	-0.003 ± 0.009
CIFAR fixed C0 0–4	$+0.066 \pm 0.021$	$+0.190 \pm 0.006$	-0.106 ± 0.010	-0.036 ± 0.011	$+0.018 \pm 0.005$

Table 11: Global-rank changes from g0 to each workflow's last generation (%), means ± sample SD across three seeds. SVD denotes singular-value entropy; PR denotes covariance participation ratio; SR denotes centered-feature stable rank. CIFAR horizons are g4, ImageNet DiT g5, real-only controls g1.

Workflow	Cov. entropy	Raw SVD	Centered SVD	PR	SR
CIFAR DDPM	$+14.54 \pm 2.28$	$+16.53 \pm 1.81$	$+18.05 \pm 2.40$	-0.61 ± 0.74	-19.02 ± 1.38
CIFAR GAN	-2.73 ± 0.56	-8.48 ± 0.56	-8.51 ± 0.49	-1.54 ± 0.48	-3.00 ± 2.97
ImageNet DiT	-33.47 ± 0.40	-16.79 ± 0.33	-12.47 ± 0.30	-32.87 ± 0.89	-14.00 ± 3.90
CIFAR real	-1.17 ± 1.20	-0.85 ± 3.22	-0.85 ± 2.96	-0.84 ± 0.74	$+1.13 \pm 1.88$
ImageNet real	-0.17 ± 0.60	-0.15 ± 0.25	-0.19 ± 0.31	$+0.30 \pm 0.65$	$+0.79 \pm 1.58$
CIFAR fixed C0	$+6.84 \pm 2.27$	$+8.86 \pm 3.64$	$+13.52 \pm 4.71$	-6.36 ± 1.93	-27.71 ± 3.85

Disjoint audit. The new analysis uses random seed `20260922` to permute indices within each class. The first half becomes the reference (500/class); draws without replacement select 20 or 50 examples per class from the other half. The split and selected indices are fixed across generations and shared across diagnostics. For a normalized scatter reference z_{st} and a subset estimate $\hat{z}_{sjt}$ for seed s, draw j and post-baseline generation t, the error reported for seed s is

$$\mathrm{MAPE}_s = \frac{100}{200T} \sum_{j=1}^{200} \sum_{t=1}^{T} \left| \frac{\hat{z}_{sjt} - z_{st}}{z_{st}} \right|.$$

NLL uses the same formula without baseline normalization. Accuracy and ECE use $100(200T)^{-1} \sum_{j,t} |\hat{z}_{sjt} - z_{st}|$. Means and standard deviations in Table 13 are computed across these per-training-seed averages. Correlated generations and repeated overlapping subset draws are not counted as new training seeds. The reference is itself finite, so these errors combine reference uncertainty and estimation mismatch. This subsampling analysis does not select alert thresholds or evaluate intervention policies; those distinct questions are examined in Section 5.7 and Appendix F.

The retention conditions give smaller normalized-scatter estimation errors than replacement in this analysis, while NLL percentage error is larger for some better-performing conditions. Relative error depends on both sampling variability and the size of the reference denominator. It is not evidence that a particular diagnostic is intrinsically worse when real data are retained. Practical cost also differs: accuracy needs predicted classes and labels; NLL/ECE need probabilities; covariance and geometry additionally require access to penultimate representations.

E VARIABILITY, BOUNDARY CASES AND EVIDENCE LINEAGE

Table 12: Stage-specific readout accuracy changes on the same 4,900 evaluation examples per seed, excluding 4,900 fitting and 200 audit images. Means ± sample SD in percentage points; fitting and hyperparameter selection are unchanged from the verified frozen-feature analysis.

Interval	Original head	Ridge	Nearest center
g0–g1	-16.52 ± 0.47	-15.12 ± 0.73	-16.46 ± 0.69
g1–g4	-9.21 ± 1.17	-7.08 ± 1.04	-7.45 ± 0.75
g0–g4	-25.73 ± 0.74	-22.20 ± 1.24	-23.91 ± 0.54

Table 13: Matched-budget estimation against a disjoint reference on CIFAR-10 replacement. Mean ± training-seed standard deviation of errors averaged over generations 1–6 and 200 subset draws. Each column estimates its own reference quantity; percentage errors and percentage-point errors are not directly comparable.

Labels/class	R MAPE (%)	Acc. MAE (pp)	ECE MAE (pp)	NLL MAPE (%)
20	11.56 ± 1.00	2.50 ± 0.04	2.36 ± 0.09	8.14 ± 0.11
50	7.60 ± 1.23	1.55 ± 0.05	1.52 ± 0.05	5.20 ± 0.17

Historical rank and scatter comparisons. On CIFAR-10 with ResNet, R increases $3.97\times$, while the mean covariance rank increases by 17.4% (Table 15, Figure 6). Final accuracy decreases from 88.48% to 56.27%. Features and accuracy refer to the same final-epoch checkpoint; selecting the highest accuracy during training would change the endpoint without changing the saved features. The result supports a blind spot of rank as a stand-alone class-geometry monitor.

Rank decreases by 57.2% on CIFAR-100 with ResNet and by 25.9% on CIFAR-10 with the 200-epoch ConvNeXt configuration, while R increases in both. Thus we do not infer a common sign of rank change or mutually exclusive forms of collapse. Nor is R universally monotone: the ConvNeXt mean rises to 1.736 at generation 4 and then decreases to 1.693. In a separate, weaker 100-epoch ConvNeXt configuration, accuracy falls from 46.40% to 17.60% while R improves from 5.895 to 5.238 (Appendix E). This is a useful boundary: a trace ratio can also miss predictive degradation.

Distinct ConvNeXt families. The 200-epoch family is identified by the prefix `cifar10_ddpm_convnext_pilot_s` with seeds 42, 43 and 44; the 100-epoch family by the prefix `cifar10_ddpm_convnext_replace_s` with the same seeds. An earlier centered-rank audit used the latter family, whereas the main scatter/accuracy comparison used the former. The present manuscript recomputes rank, scatter and centered geometry from the same stored features within each family. For the 200-epoch configuration, the corrected rank change is $19.11 \rightarrow 14.17$ (-25.9%), rather than the approximately -62.3% change of the separate weak configuration. No cross-family metric pairing is used in the figures or tables.

Which historical claims are not used. Uncentered singular-value entropy is not used as covariance effective rank. An earlier noncentered class-mean deviation is not used as NC2. Stored best-epoch accuracy is not paired with final-epoch geometry. Historical regularizer comparisons do not support mitigation claims because versions used detached features, inconsistent rank definitions or evaluation-derived targets. Corrected experimental code provides live gradients and training-only calibration; it does not turn old runs into new evidence. The labeled real-image noise control repeatedly perturbs original images while retaining true labels, changing both input and labeling processes relative to DDPM replacement. It is therefore not a factorial isolation of label noise and is not used to claim that label errors are the primary cause. Incomplete ImageNet accumulation attempts are also not included.

Remaining limitations. Three seeds (and two for historical retention) are insufficient for narrow population-level confidence claims. The same saved ordering is used within historical runs, but historical feature files do not store unique image identifiers; deterministic loader order and equal label sequences support alignment without independently proving image identity from the feature arrays alone. The fresh ImageNet studies record ordered validation IDs and a dataset manifest, covering two ViT replacement steps and a five-step post-outcome ResNet continuation on 100 selected

Table 14: Disjoint-reference audit across CIFAR-10 data-retention conditions. All generations after baseline and all draws are averaged within each seed, then across seeds; replacement has three seeds, the other conditions two. Error units match Table 13.

Condition	Labels/class	R MAPE (%)	Acc. MAE (pp)	ECE MAE (pp)	NLL MAPE (%)
replace	20	11.56	2.50	2.36	8.14
replace	50	7.60	1.55	1.52	5.20
mix25	20	8.28	2.40	2.28	12.57
mix25	50	5.48	1.45	1.35	8.07
mix50	20	7.06	2.28	2.27	15.08
mix50	50	4.94	1.45	1.41	9.38
mix75	20	6.50	2.04	1.99	16.63
mix75	50	4.47	1.22	1.17	10.35
accumulate	20	6.33	2.01	2.00	16.15
accumulate	50	4.35	1.24	1.19	9.83

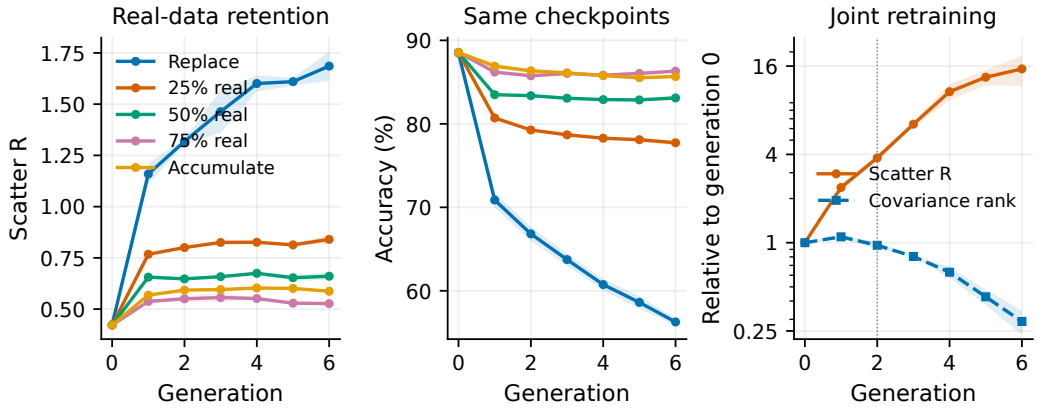

Figure 5: **Workflow changes the outcome.** Left and middle: CIFAR-10 real-data retention, with two seeds per retained-data condition and three for replacement. Right: jointly retrained generator/classifier trajectories, normalized per seed, on a logarithmic vertical axis. Bands are one standard deviation. The first classifier trained on outputs of a fine-tuned generator is generation 2; generation 1 still uses the pretrained generator.

classes with one pretrained generator family. These finite trajectories do not establish an asymptotic outcome. We do not measure generator FID, diversity or class-conditional coverage in the joint loop and cannot identify generator mode collapse from classifier accuracy alone. A fixed real probe, balanced class quotas, access to intermediate features and stable class definitions are operational assumptions. No experiments here establish performance under changing class priors, open-world classes or deployment distribution shift.

F Fresh replication and matched intervention methods

Numerical replay scopes, source bindings and execution amendments for these studies are collected in Appendix G.1.

ImageNet recipe and horizon summary. The original SGD ViT trajectory loses 19.29 accuracy points by generation 2; within-seed scatter rises 1.712 ± 0.035-fold and rank falls $11.00 \pm 0.22\%$ (Table 17). All three ImageNet recipes show concurrent geometric and dimensional degradation over two replacements, without establishing an ImageNet rank-increase blind spot.

The post-outcome ResNet-18 recipe starts at $67.89 \pm 0.41\%$ accuracy and reaches $24.65 \pm 0.75\%$ by generation 2; paired scatter grows 2.718 ± 0.023-fold and rank falls $30.43 \pm 0.91\%$. Matched real-only generation-1 retraining reaches 68.61%, versus 27.53% with pseudo-synthetic inputs; SGD ViT

Table 15: Replacement endpoints (means over three seeds). $T = 6$ for CIFAR and $T = 4$ for ImageNet. †ImageNet is a training-origin probe with historical preprocessing limitations, not a held-out accuracy benchmark. Variability is reported in Appendix E and the artifact.

Setting	R_0	R_T	r_0	r_T	Acc.$_0$ (%)	Acc.$_T$ (%)
C10 / ResNet	0.424	1.686	12.28	14.41	88.48	56.27
C100 / ResNet	1.756	6.993	134.46	57.54	59.42	4.80
C10 / ConvNeXt	0.739	1.693	19.11	14.17	89.47	57.02
IN100 / ViT†	2.558	10.893	116.02	120.05	74.91	9.10

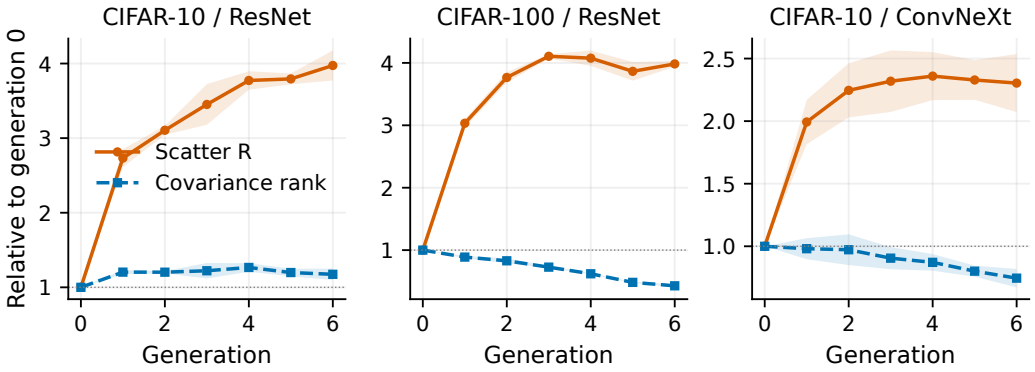

Figure 6: Historical held-out replacement trajectories, normalized within each seed. Rank and scatter summarize the same final checkpoint. Lines and bands show means and one population standard deviation over three seeds. CIFAR-100 uses the CIFAR-10 generator and therefore changes the input domain as well as the task.

controls similarly retain 34.91%, versus 18.81%. Architecture and recipe change jointly on reused data and seeds (Appendix F).

Extending the same ResNet trajectories through generation 5 yields 22.67% accuracy. The prespecified continuation contrast, generation $2 \to 5$, is -1.99 ± 0.32 accuracy points, 1.024 ± 0.025-fold scatter and $-4.37 \pm 1.12\%$ rank (Tables 19–20). Scatter changes little and its mean briefly improves at generation 3 despite further accuracy loss. This post-outcome continuation adds depth, not independent seeds, and establishes neither monotone geometric degradation nor an asymptotic limit.

A fixed-image control using conditioning IDs instead of pseudo-labels also lowers both accuracy and R, reinforcing that lower scatter need not improve prediction. Conditioning IDs are not verified image labels. Excluding two screened cross-synset training/validation near-duplicates leaves the ImageNet trajectory materially unchanged; other overlap and DiT pretraining remain untested (Appendix F).

Three questions require different comparisons. The replication first asks whether global rank can increase while trusted-label class geometry deteriorates. This requires features, labels and final-checkpoint measurements from the same models. It does not require the scatter ratio to outperform accuracy: accuracy measures predictive success, whereas the geometric analysis asks what a global spectral summary leaves unspecified. Agreement between geometry and accuracy is compatible with the diagnostic mismatch between geometry and rank. Conversely, the weak ConvNeXt and conditioning-ID controls show why scatter itself cannot replace accuracy.

The monitoring question is whether a limited labeled audit can identify a change and support a useful alarm. Estimation error against a disjoint reference quantifies the first part. Alarm timing additionally depends on the score, baseline, threshold and frequency of inspection. A score can change substantially and still fail to alarm early under a calibrated rule. A nominal level shared by two rules also need not give the same realized false-alarm rate, as the held-out real-retraining pairs

Table 16: Endpoint variability across the three training seeds for each configuration. Standard deviations use divisor n, matching all plotted bands.

Configuration	R_T	r_T	Acc.$_T$ (%)
C10 / ResNet	1.69 ± 0.07	14.41 ± 0.85	56.27 ± 0.41
C100 / ResNet	6.99 ± 0.03	57.54 ± 2.79	4.80 ± 0.22
C10 / ConvNeXt	1.69 ± 0.07	14.17 ± 0.58	57.02 ± 0.92
IN100 / ViT†	10.89 ± 0.15	120.05 ± 1.37	9.10 ± 0.33
C10 / ConvNeXt weak	5.24 ± 0.37	8.50 ± 0.25	17.60 ± 0.71
C10 / joint loop	7.70 ± 1.91	3.74 ± 0.69	15.78 ± 0.21

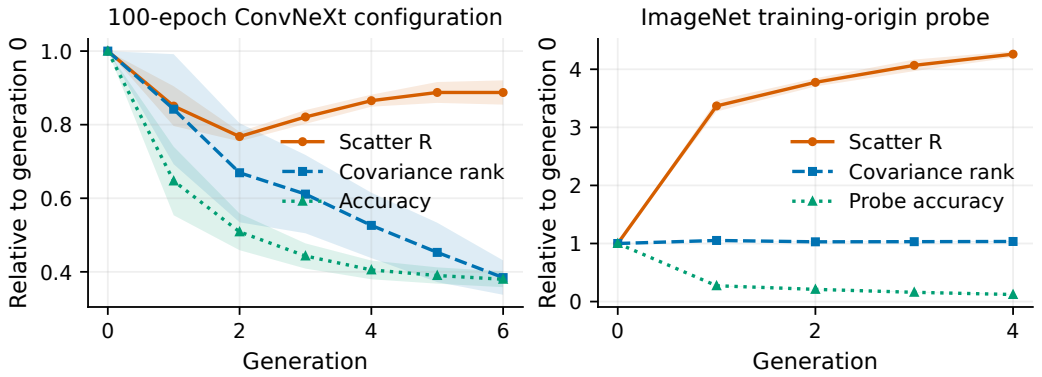

Figure 7: Additional scope checks. Left: the weak 100-epoch ConvNeXt configuration loses accuracy despite an improved endpoint scatter ratio. Right: descriptive ImageNet training-origin trajectories; the probe and historical pseudo-label preprocessing limit their interpretation. Quantities are normalized within each seed; bands are one standard deviation over three seeds.

demonstrate. The ROC comparison evaluates empirical ranking on a selected cohort; it does not turn the earlier threshold comparisons into a prospective or sequential guarantee.

The intervention question concerns what happens after an alarm. Our replay and retention comparisons use the next generation as the first actionable step. A policy can improve its endpoint because it consumes more real training data, because it changes the time of the switch, or both. Comparison with an unconditional switch is therefore necessary: the six-seed geometry policy matches the fixed generation-2 switch throughout. The common-calibration analysis replays available stored branches and preserves their limitations; it cannot supply an untrained counterfactual branch. Neither the original policy difference nor that replay establishes a unique benefit from observing geometry.

Labels have separate roles. In the fresh CIFAR replacement study, the same 200 official-test images provide all four audit scores, and the other 9,800 images provide endpoint evaluation. The audit labels are never classifier-training targets. Real examples retained by an intervention come from the training set and constitute an additional resource beyond the audit budget; the number of real training rows must therefore accompany any policy comparison. The frozen-feature readout is a separate diagnostic: it splits the 9,800 endpoint images into 4,900 fitting and 4,900 evaluation examples, excluding the audit set. Its supervised fitting labels are not part of a 200-label monitoring claim. Internal fitting folds choose the ridge penalty without using readout evaluation labels. A better fitted readout demonstrates recoverable structure for that decoder; its residual degradation does not prove that every possible decoder would fail.

Official ImageNet validation. The new study uses seeds 101, 102 and 103, each with generation 0 and two synthetic replacements. The first 100 ImageNet-1K synsets in sorted WNID order map to local labels and DiT conditioning indices 0–99. We choose 500 official training images per class before outcomes (50,000 total) and use all 50 official validation images per class (5,000 total). The initial selection contained 49 training files byte-identical to validation files. Before any classifier

Table 17: ImageNet-100 official-validation trajectories (mean ± sample SD, three seeds): 90 epochs, 50,000 training rows and 5,000 validation images. Post-outcome AdamW/ResNet studies reuse original real data and generated image pools with new pseudo-labels. Settings change jointly; rank is compared within trajectories because feature dimensions differ.

Classifier recipe	g	Accuracy (%)	Scatter R	Rank r
ViT-S/16, SGD	0	34.96 ± 0.67	5.731 ± 0.092	137.409 ± 0.261
	1	18.81 ± 0.34	9.051 ± 0.138	126.463 ± 0.865
	2	15.67 ± 0.12	9.811 ± 0.132	122.289 ± 0.315
ViT-S/16, AdamW	0	43.13 ± 1.12	4.764 ± 0.126	132.285 ± 0.674
	1	19.11 ± 0.47	8.818 ± 0.040	108.882 ± 0.569
	2	16.64 ± 0.08	9.013 ± 0.257	103.779 ± 1.176
ResNet-18, SGD	0	67.89 ± 0.41	1.703 ± 0.011	101.985 ± 0.477
	1	27.53 ± 0.58	4.494 ± 0.074	75.101 ± 1.390
	2	24.65 ± 0.75	4.628 ± 0.061	70.952 ± 0.604

training, we retained validation unchanged and replaced exactly those 49 training files with deterministic same-class reserves. The final manifest records the original and replacement IDs, ordering, class mapping and SHA-256 hashes; no selected training image is byte-identical to a validation image. This excludes byte-identical classifier-training overlap, without establishing that the external pretrained DiT corpus is disjoint from validation or free of near duplicates.

After observing the original outcomes, a perceptual-hash screen of selected training and validation images, including flips, returned 247 candidate pairs at Hamming distance at most eight. We visually reviewed all 247 pairs in three batches: 25 selected by low hash distance or pixel difference, 40 additional pairs at resized RGB correlation at least 0.5, and the remaining 182 regardless of correlation. Two byte-different validation photographs matched selected training photographs under different official synsets; no further same-source pair was found. Table 17 retains the original endpoints. Excluding only the two confirmed matches changes mean generation-0/2 accuracies from 34.960/15.673% to 34.974/15.680%, the ratio of mean scatter values from 1.7120 to 1.7111, and mean rank change from -11.003% to -11.010%. An overinclusive stress test excludes all 111 validation rows with nearest training-image pHash distance at most eight, including nonduplicates: mean generation-0/2 accuracies are 34.942/15.675%, mean within-seed scatter fold 1.7117 and rank change -11.134%, versus full-set values of 1.7123 and -11.003% for the latter two measures. This 4,889-image post-outcome subset is not an independent benchmark; overlap outside the pHash candidate set and external DiT-pretraining overlap remain untested. Contact sheets and ImageNet images are not redistributed.

The classifier is the same randomly initialized ViT-S/16 family as above, trained for 90 epochs per generation, batch size 64, SGD at learning rate 10^{-3}, momentum 0.9, weight decay 5×10^{-4} and cosine decay. Generation 0 uses 50,000 real images. For each subsequent generation the pinned DiT-XL/2 snapshot draws 50,000 new class-conditional 256×256 images with 50 denoising steps; the preceding final classifier assigns hard pseudo-labels using resize to 256, center crop to 224 and ImageNet normalization. Training applies random resized crops, flips and the same normalization. All three classifiers within a seed use 50,000 training rows, independently initialized weights and fresh synthetic image pools. The official validation set is used only for endpoint measurements, with final-checkpoint accuracy, features and logits evaluated in manifest order. There is no ImageNet audit-triggered policy in this study.

Exploratory classifier-recipe sensitivity. After seeing the SGD results, we specified a paired three-seed sensitivity using the same ImageNet manifest, validation order, ViT-S/16 initialization seeds, 90 epochs per classifier and exact generated image pools. The replacement recipe jointly uses AdamW at learning rate 5×10^{-4}, weight decay 0.05, five linear warmup epochs from 10% of target rate, then cosine decay. Each preceding classifier assigns new pseudo-labels to the shared images. Generation-0/1/2 accuracy is 43.13 ± 1.12/19.11 ± 0.47/16.64 ± 0.08%, scatter ratio is 4.764 ± 0.126/8.818 ± 0.040/9.013 ± 0.257, and effective rank is 132.285 ± 0.674/108.882 ± 0.569/103.779 ± 1.176 (mean ± sample SD across training seeds). Paired accuracy gains over SGD are 8.17 ± 1.78, 0.30 ± 0.51 and 0.97 ± 0.20 percentage points at generations 0/1/2. Within-recipe

generation-0-to-2 accuracy falls 26.49 ± 1.19 points, scatter rises 1.892 ± 0.049-fold and rank falls $21.55 \pm 0.85\%$. Excluding the two visually confirmed validation duplicates changes mean generation-2 accuracy from 16.640% to 16.647%. All nine saved models were independently checked against their training logs, checkpoint heads, ordered validation arrays, source and image-pool hashes, and pseudo-label lineage. This post-outcome study changes multiple classifier settings together and does not identify an individual optimizer or label-noise effect.

Exploratory ResNet architecture and recipe sensitivity. After the original SGD ViT, AdamW ViT and CIFAR follow-ups, we specified a further three-seed sensitivity using an ImageNet-stem ResNet-18 with 512-dimensional features. The corrected real split, validation order, initialization seed schedule and exact generated image pools are unchanged; each ResNet predecessor supplies its own pseudo-labels. Every classifier starts from random weights and trains for 90 epochs with batch 256, SGD learning rate 0.1, momentum 0.9, weight decay 10^{-4} and a factor-0.1 learning-rate reduction after epochs 30 and 60. The same crop, flip and normalization transforms are used. Architecture and training settings change jointly, so this is neither an isolated architecture/optimizer effect nor a new generator replicate. Table 17 reports every generation; Table 18 gives every seed's accuracy, scatter, rank, ECE and NLL.

Paired generation-0 accuracy gains over the original SGD and AdamW ViT recipes are 32.93 ± 0.98 and 24.75 ± 1.00 percentage points, respectively. From generation 0 to 2, accuracy falls 43.23 ± 0.36 points, scatter rises 2.718 ± 0.023-fold and rank falls $30.43 \pm 0.91\%$ (mean $\pm$ sample SD). Thus the geometric and dimensional degradation also occurs from a substantially higher initial accuracy. Excluding the two confirmed near-duplicate validation rows changes mean generation-2 accuracy from 24.653% to 24.663%; the full-set endpoint remains primary. The study still has three reused seeds, two synthetic replacements and no ImageNet audit-triggered intervention.

Table 18: ImageNet ResNet-18 sensitivity and matched real-only endpoints. 1R denotes generation-1 retraining on real data with its original labels. All metrics use final checkpoints and the complete official validation set. ECE uses 15 equal-width bins; accuracy and ECE are percentages, NLL uses natural logarithms.

Seed	g	Acc. (%)	R	r	ECE (%)	NLL
101	0	67.60	1.710	101.492	10.45	1.332
101	1	26.86	4.537	76.651	31.94	4.243
101	2	24.30	4.694	71.628	36.33	4.761
101	1R	68.70	1.664	100.822	9.59	1.311
102	0	67.70	1.708	102.445	10.46	1.307
102	1	27.90	4.408	74.686	31.92	4.214
102	2	24.14	4.615	70.465	37.07	4.859
102	1R	69.02	1.651	102.092	9.73	1.300
103	0	68.36	1.690	102.018	10.23	1.326
103	1	27.84	4.535	73.967	32.22	4.310
103	2	25.52	4.575	70.762	34.46	4.570
103	1R	68.10	1.699	102.534	10.41	1.341

ImageNet continuation through generation 5. After observing the ResNet generation-0–2 results, we fixed a continuation through generations 3, 4 and 5 for the same seeds 101–103. Each added generation draws 50,000 fresh images from the same pinned DiT sampler, uses the preceding final classifier as its pseudo-label teacher, and trains a newly initialized ResNet with the unchanged 90-epoch recipe. It does not fine-tune the preceding classifier's weights. The real split, class mapping, official validation order, sampling settings and generation-indexed seed rules are unchanged. Before continuation outcomes, we specified generation $2 \to 5$ as the primary contrast and $0 \to 5$ as secondary. Nine added models extend three existing trajectories; generations do not provide additional independent seeds.

Tables 19–21 retain every generation, paired contrast and seed. From generation 2 to 5, accuracy falls 1.99 ± 0.32 points, scatter changes 1.024 ± 0.025-fold and rank falls $4.37 \pm 1.12\%$. The mean scatter ratio first falls slightly, from 4.628 at generation 2 to 4.615 at generation 3, while mean

Table 19: ImageNet-100 ResNet-18 continuation through five synthetic-training generations (mean $\pm$ sample SD, three reused seeds). Generations 0–2 are retained verified endpoints; 3–5 were specified after observing them. The classifier recipe, fixed DiT generator and 5,000-image official validation are unchanged. ECE uses 15 equal-width bins; NLL uses natural logarithms.

g	Acc. (%)	Scatter R	Rank r	ECE (%)	NLL
0	67.89 ± 0.41	1.703 ± 0.011	101.985 ± 0.477	10.38 ± 0.13	1.321 ± 0.013
1	27.53 ± 0.58	4.494 ± 0.074	75.101 ± 1.390	32.03 ± 0.17	4.255 ± 0.049
2	24.65 ± 0.75	4.628 ± 0.061	70.952 ± 0.604	35.95 ± 1.34	4.730 ± 0.147
3	23.99 ± 1.06	4.615 ± 0.063	68.528 ± 0.215	37.11 ± 0.90	4.896 ± 0.143
4	23.29 ± 0.13	4.654 ± 0.072	67.870 ± 1.358	38.15 ± 0.60	5.039 ± 0.099
5	22.67 ± 0.59	4.736 ± 0.054	67.850 ± 0.355	39.16 ± 1.44	5.139 ± 0.118

Table 20: Paired changes for the ImageNet horizon extension (mean $\pm$ sample SD across three reused seeds). The prespecified primary contrast is $g2 \to g5$; $g0 \to g5$ is secondary. Ratios and percentage changes are computed within each seed before averaging. "2 excluded" removes the two previously identified validation near-duplicates. This post-outcome continuation does not test diagnostic intervention value.

Contrast	Validation	ΔAcc. (pp)	R_b/R_a	Δr (%)
$2 \to 5$	Full	-1.99 ± 0.32	1.024 ± 0.025	-4.37 ± 1.12
$2 \to 5$	2 excluded	-1.99 ± 0.32	1.024 ± 0.025	-4.37 ± 1.13
$0 \to 5$	Full	-45.22 ± 0.36	2.782 ± 0.048	-33.47 ± 0.40
$0 \to 5$	2 excluded	-45.24 ± 0.36	2.781 ± 0.048	-33.47 ± 0.40

accuracy also falls; seed 101 has a generation-2-to-5 scatter fold below one. Over generation $0 \to 5$, accuracy falls 45.22 ± 0.36 points, scatter grows 2.782 ± 0.048-fold and rank falls $33.47 \pm 0.40\%$. Thus most observed degradation precedes the continuation, and later predictive losses need not produce a comparably large or monotone scatter change. These finite, reused-seed trajectories do not establish a steady state, an ImageNet rank-increase blind spot or a diagnostic intervention benefit.

The complete study independently verifies all nine new checkpoints, their 90-epoch schedules, all generated-pool hashes and all 450,000 predecessor-model pseudo-label decisions. It additionally replays 45,000 raw official-validation images and checks 1,800 CPU versus TF32-disabled CUDA predictions/features under the parent's unchanged tolerances. The retained generation-0–2 records remain separately identified. Excluding the two previously confirmed near-duplicate validation rows changes generation-5 mean accuracy from 22.667% to 22.676%; Table 20 reports both scopes. The artifact includes 36 endpoint rows, 12 paired-change rows and all 900 assigned-class counts for the nine new pools. This continuation changes neither the parent results nor the limited scope of the overlap audit.

Matched ResNet real-only retraining control. After observing the ResNet generation-0/1 outcomes, we specified a matched generation-1 control using the same corrected 50,000 real images and their original labels. Initialization seeds, the 90-epoch SGD/StepLR recipe, batch size, augmentations and validation set match pseudo-synthetic generation 1. Table 18 reports all three controls as 1R. Mean accuracy is $68.61 \pm 0.47\%$, a paired 41.07 ± 0.79 percentage points above pseudo-synthetic generation 1 and $+0.72 \pm 0.86$ points relative to real-data generation 0. Within-seed scatter fold versus generation 0 is 0.982 ± 0.021 and rank change is $-0.17 \pm 0.60\%$ (mean $\pm$ sample SD). Fresh initialization and retraining on unchanged real data do not reproduce the large synthetic-data drop. Images and assigned labels change together in this contrast, so their separate effects remain unidentified. The same three seeds and selected configuration are reused, and no diagnostic intervention is tested. Independent verification rehashes all 50,000 real training images and checks frozen sources, matched recipes, complete epoch logs, checkpoint heads and metrics. All 15,000 raw validation-image replays match saved features/logits exactly, and the additional 600 CPU versus TF32-disabled CUDA comparisons pass. Excluding the two confirmed near-duplicate validation images changes mean accuracy from 68.607% to 68.634%; the full-set endpoint remains primary.

Table 21: All ImageNet ResNet-18 horizon endpoints on the complete official validation. Generations 0–2 are historical verified models and 3–5 are the added continuation; generations do not add independent training seeds. Accuracy and ECE are percentages.

Seed	g	Acc. (%)	R	r	ECE (%)	NLL
101	0	67.60	1.710	101.492	10.45	1.332
101	1	26.86	4.537	76.651	31.94	4.243
101	2	24.30	4.694	71.628	36.33	4.761
101	3	23.84	4.611	68.679	37.15	4.910
101	4	23.36	4.678	68.276	37.56	4.974
101	5	22.66	4.678	67.582	39.43	5.144
102	0	67.70	1.708	102.445	10.46	1.307
102	1	27.90	4.408	74.686	31.92	4.214
102	2	24.14	4.615	70.465	37.07	4.859
102	3	23.02	4.680	68.624	38.00	5.033
102	4	23.36	4.711	68.978	38.76	5.153
102	5	22.08	4.746	67.715	40.45	5.254
103	0	68.36	1.690	102.018	10.23	1.326
103	1	27.84	4.535	73.967	32.22	4.310
103	2	25.52	4.575	70.762	34.46	4.570
103	3	25.12	4.554	68.282	36.19	4.747
103	4	23.14	4.572	66.354	38.15	4.990
103	5	23.26	4.784	68.252	37.60	5.019

Fixed-image conditioning-label control. After the original SGD outcomes, we specified a separate three-seed generation-1 control on each seed's *exact same* 50,000 DiT images. It keeps the original ViT-S/16 initialization seed, 90-epoch SGD recipe, augmentations and official validation set, but replaces the preceding classifier's pseudo-labels with the generator's conditioning IDs. The latter are balanced at 500 assigned targets per class; original pseudo-label counts span 101–1,715 across the three seeds. Conditioning IDs agree with the original pseudo-labels on $67.39 \pm 0.67\%$ of images. Conditioning-target validation accuracy is $16.38 \pm 0.47\%$, versus $18.81 \pm 0.34\%$ under pseudo-labels, a within-seed paired difference of -2.43 ± 0.24 percentage points (mean $\pm$ sample SD). The conditioned models have scatter ratio 8.472 ± 0.042 versus 9.051 ± 0.138 and centered effective rank 108.577 ± 0.135 versus 126.463 ± 0.865. All three saved models were independently checked against fixed-image identities, frozen source and checkpoint hashes, training logs, validation order, classifier heads and recomputed metrics. This post-outcome single-generation control does not support accuracy recovery from assigning conditioning IDs. Those IDs are generator requests rather than human-verified image classes; the intervention also changes the target-label marginal, so it cannot isolate true-label error or establish multi-generation behavior.

Matched ViT real-only retraining control. After the original SGD outcomes, we repeated generation-1 training on the same corrected 50,000 real ImageNet-100 images and official validation set, with the original ViT-S/16 initialization seed, augmentations, SGD settings and 90-epoch budget. Across seeds 101/102/103, real-only accuracy is 35.20/34.84/34.70%, scatter ratio is 5.756/5.749/5.764 and centered effective rank is 136.481/137.141/137.521. Mean accuracy is $34.91 \pm 0.26\%$, a paired 16.11 ± 0.59 percentage points above the original pseudo-synthetic generation-1 models and -0.05 ± 0.71 points relative to the original real-data generation-0 models (mean $\pm$ sample SD). Mean within-seed scatter fold relative to generation 0 is 1.005 ± 0.015 and rank change is $-0.26 \pm 0.56\%$. All three controls were independently checked against the manifest, frozen source, 90-epoch logs, saved checkpoints, classifier heads and ordered official-validation arrays. Fresh initialization and retraining alone do not explain the large pseudo-synthetic generation-1 drop; the real-only contrast changes both inputs and labels and therefore does not isolate either cause.

Matched CIFAR-10 policies. The new CIFAR experiment also uses seeds 101–103, but generation 0 plus four synthetic replacements. It trains ResNet-18 for 200 epochs, batch size 128, SGD at learning rate 0.1 with momentum 0.9, weight decay 5×10^{-4} and cosine decay; the historical study above used batch size 256 and a longer trajectory. The pretrained `google/ddpm-cifar10-32` snapshot supplies 50,000 fresh images with 50 steps at each generation, and the predecessor clas-

sifier assigns pseudo-labels. At generation 0, before model outcomes, we reserve 20 official-test examples per class (200 total) for auditing; the other 9,800 form the endpoint. Separately, 50 originally labeled training images per class (500 total) are retained for replay. Each replay model has 12,500 balanced repeated real rows and 37,500 synthetic rows; replacement models have 50,000 synthetic rows. These are retained original training labels, not additional audit annotations.

The four policies are replacement, replay at every synthetic generation, scatter-triggered replay and accuracy-triggered replay. The latter two act one generation after a first alarm and stay on thereafter. The scatter, accuracy, 15-bin ECE and NLL audits all use the same 200 labels, fixed across generations. For each seed, each alarm threshold is the 95th percentile of 2,000 corresponding differences between independent class-stratified bootstrap samples of the generation-0 audit, fixed before later outcomes. This estimates variation from resampling the fixed baseline audit, not a population false-alarm rate over independent retrainings. An independently initialized real-only generation-1 retrain in each seed probes this additional source of false alarms. Generated images and training seeds are matched across policies; when their actual training arrays and seeds coincide, policies reuse the same fitted model, so policy aliases are not additional training replicates.

At generation 4, replacement reaches $62.17 \pm 1.04\%$, always replay $62.56 \pm 1.28\%$, and both alarm-triggered policies $62.72 \pm 1.35\%$ (mean $\pm$ sample SD). Their paired gains over replacement are 0.39 ± 0.36 and 0.55 ± 0.48 percentage points, respectively. All four scores first alarm at generation 1 in all replacement seeds, so geometry and accuracy choose identical fitted paths. Each score has zero alarms on the three real-only retrains, which is too few to estimate a reliable false-alarm rate. These results show a small replay benefit at the tested retained-label budget without geometric lead time or a unique geometry-triggered policy.

Fixed-initial-teacher comparison. After inspecting the original rank, scatter and accuracy outcomes, we specified a separate control on all three fresh CIFAR seeds. Each generation uses the exact original 50,000-image DDPM pool but assigns labels with the seed's fixed real-trained C_0, instead of the preceding classifier. Generation 1 already uses C_0, so g0/g1 are shared and only g2–g4 are newly trained. The initialization seed at generation g remains the base seed plus $1000g$, and the unchanged trainer uses the 200-epoch recipe above. Evaluation retains the original 9,800 endpoint and 200 audit rows. No audit score selects the control or its training data. The 51-model entropy-accounting cohort is unchanged; these nine models form a separate control.

The plan binds the original 27-file trainer, new runner, 15 retained models, nine image pools and evaluation manifests before producing any new targets or checkpoints. The prespecified primary comparison is the within-seed g4 fixed-minus-inherited difference in accuracy, scatter and rank. Every g2–g4 endpoint and its paired comparison is retained, along with all ten assigned-class counts per pool and teacher. The initial independent numerical verifier was frozen before the first completed new model; a disclosed device-selection correction after training preserves those numerical checks. Means and sample SD summarize three reused seeds, not independent new baselines or a population significance test.

Fixed-teacher accuracy at g2/g3/g4 is 71.97/71.89/72.12%, compared with inherited-teacher accuracy 66.94/64.10/62.17%. The g4 paired advantage is 9.95 ± 1.26 points; scatter is 0.442 ± 0.052 lower and rank 1.000 ± 0.063 lower. Fixed and inherited teachers disagree on $28.74 \pm 0.39\%$ of g4 training labels. Their class-frequency vectors also change. Thus the comparison holds generated inputs fixed and tests the teacher choice, without isolating a change in true-label error from a change in target marginals. It neither identifies true generated-image classes nor predicts behavior under other generators, horizons or training recipes. Table 23 gives all 24 distinct shared, inherited and fixed-teacher endpoints; the artifact includes every target count.

Standard CIFAR input-stem sensitivity. This post-outcome architecture study reuses the original DDPM image pools, seed values 101–103 and official evaluation partitions, while refitting all five classifiers per seed from random initialization. The initial convolution changes from 7×7, stride 2 to 3×3, stride 1, padding 1; max-pooling becomes the identity. The residual stages, 512-dimensional feature output and ten-class head retain their original definitions. The new convolution uses Kaiming-normal initialization with fan-out/ReLU gain. All other frozen training modules are unchanged. The model has 11,173,962 trainable parameters. The revised architecture is used by

Table 22: Fixed-initial-teacher control on the same CIFAR DDPM pools. Initial g0 and first-replacement g1 models are shared. Each later classifier is trained from scratch with its original generation seed and 200-epoch recipe. Values are means ± sample SD over three seeds; the final row is a within-seed difference (accuracy in percentage points). This post-outcome control changes the labeling teacher, including its target marginals.

Checkpoint / label teacher	Accuracy (%)	Scatter R	Rank r
g0, real-trained (shared)	88.35 ± 0.08	0.503 ± 0.002	12.99 ± 0.09
g1, teacher C0 (shared)	71.45 ± 0.24	1.220 ± 0.048	14.41 ± 0.31
g4, inherited teacher	62.17 ± 1.04	1.620 ± 0.048	14.88 ± 0.31
g4, fixed C0 teacher	72.12 ± 0.34	1.178 ± 0.013	13.88 ± 0.29
g4, fixed minus inherited	$+9.95 \pm 1.26$	-0.442 ± 0.052	-1.00 ± 0.06

Table 23: All retained and new fixed-teacher comparison endpoints. Shared g0/g1 records are the original verified models; only the nine fixed-C0 g2–g4 models are newly trained. Accuracy uses each seed's unchanged 9,800-image endpoint.

Seed	Generation	Teacher	Accuracy (%)	R	r
101	0	shared	88.33	0.5004	12.8902
101	1	shared	71.66	1.1953	14.4316
101	2	inherited	67.06	1.4684	15.1997
101	2	fixed C0	72.39	1.1563	14.2212
101	3	inherited	64.60	1.5737	14.9247
101	3	fixed C0	71.81	1.1452	13.8530
101	4	inherited	62.45	1.6752	14.8271
101	4	fixed C0	72.39	1.1738	13.8966
102	0	shared	88.44	0.5047	13.0549
102	1	shared	71.18	1.2756	14.6997
102	2	inherited	67.10	1.3477	14.1528
102	2	fixed C0	71.77	1.1647	14.3454
102	3	inherited	64.23	1.5980	14.9480
102	3	fixed C0	72.00	1.2002	14.6470
102	4	inherited	63.04	1.5967	15.2131
102	4	fixed C0	71.73	1.1929	14.1603
103	0	shared	88.30	0.5037	13.0346
103	1	shared	71.51	1.1895	14.0906
103	2	inherited	66.66	1.3422	14.1933
103	2	fixed C0	71.76	1.1393	13.7988
103	3	inherited	63.47	1.4821	14.5501
103	3	fixed C0	71.86	1.2289	14.4705
103	4	inherited	61.02	1.5882	14.6052
103	4	fixed C0	72.23	1.1677	13.5877

both the student trainer and the pseudo-label producer; the teacher at generation g is that seed's new-stem model at $g - 1$.

Generation 0 uses the same 50,000 official training examples. Every replacement uses the corresponding original 50,000-image generated pool, with new inherited pseudo-labels. All models use 200 epochs, batch size 128, SGD with learning rate 0.1, momentum 0.9 and weight decay 5×10^{-4}, cosine scheduling, random crop/flip and mixed-precision training. Model seeds remain $s + 1000g$; epoch seeds add $10007e$. The original 200 audit images stay disjoint from the 9,800 endpoint images. The study, source hashes, complete input-pool identities, primary g0–g4 contrast and all intermediate generations were frozen before obtaining new scientific outcomes. Its selection after the original findings remains an explicit limitation; it adds an architecture check on reused data, rather than a new generator or dataset replication.

Table 24 reports all fifteen endpoints. Full g0–g4 covariance rank increases by 2.41%, 3.76% and 1.62% for seeds 101–103, respectively, despite accuracy losses of 39.86, 37.06 and 40.79 percentage points and scatter folds 9.183, 11.250 and 11.059. Both within- and between-class entropies

Table 24: All fifteen newly trained CIFAR-stem models, evaluated on the retained 9,800 true-labeled endpoint images per seed. Entropies use natural logarithms; accuracy is in percent. Three trajectories, not fifteen independent replications.

Seed	Gen.	Acc.	R	r	H_B	H_W
101	0	95.42	0.1418	10.381	2.1552	3.0224
101	1	70.22	0.9359	11.188	1.9515	2.6809
101	2	66.27	1.0329	11.120	1.9878	2.5859
101	3	58.82	1.1258	10.364	1.9002	2.4918
101	4	55.56	1.3025	10.632	1.9612	2.4750
102	0	95.23	0.1314	10.134	2.1550	2.9643
102	1	72.29	0.9140	10.160	1.9430	2.5179
102	2	66.33	1.0406	9.746	1.8521	2.4677
102	3	61.13	1.1931	10.679	1.9351	2.5071
102	4	58.17	1.4788	10.515	1.9508	2.4224
103	0	95.59	0.1274	10.096	2.1562	2.9635
103	1	67.94	1.0183	11.807	1.9912	2.7022
103	2	58.05	1.1150	10.510	1.8636	2.5631
103	3	57.40	1.2606	10.756	1.9076	2.5463
103	4	54.81	1.4092	10.259	1.9488	2.4121

Table 25: Direct stage contrasts for the CIFAR stem, means $\pm$ sample SD across three seeds. Accuracy changes are percentage points, rank changes relative percentages; all changes are computed within seed.

Interval	Δ acc. (pp)	R_b/R_a	Δr (%)
g0–g1	-25.27 ± 2.35	7.181 ± 0.724	$+8.33 \pm 8.36$
g1–g4	-13.97 ± 0.78	1.464 ± 0.133	-4.86 ± 8.31
g0–g4	-39.23 ± 1.94	10.498 ± 1.142	$+2.60 \pm 1.08$

decrease in every full-trajectory comparison. Variance reweighting contributes $+0.304$ nats on average, the component terms sum to -0.322, and mixing contributes $+0.044$. The positive mixing change is needed for the positive net balance here; reweighting alone does not outweigh the component terms.

The first replacement accounts for 64.40% of the mean total accuracy loss. During g1–g4, all three seeds lose further accuracy, while covariance rank changes by -4.97%, $+3.50\%$ and -13.11%. Within-class entropy decreases in all three later comparisons, but between-class entropy slightly increases in seeds 101 and 102. The later-stage mismatch is therefore not consistent across seeds for this stem. Full-trajectory centered singular-value entropy changes by $+6.30 \pm 3.97\%$ on average, uncentered singular-value entropy by $-2.79 \pm 5.26\%$, participation ratio by $-5.62 \pm 3.56\%$ and feature stable rank by $-6.61 \pm 9.82\%$. Tables 25–27 retain complete stage and definition comparisons. The original-stem subspace/readout results below were not repeated for the new stem.

Conditional GAN generator comparison. After the diffusion, retention and diagnostic analyses, we specified a four-step CIFAR-10 replacement study using the official conditional CIFAR-10 StyleGAN2-ADA checkpoint (Karras et al., 2020). The checkpoint remains fixed throughout. We reuse the three verified real-data generation-0 classifiers and seeds 101–103; each new classifier starts from its prescribed independent initialization and uses the same 200-epoch ResNet-18 recipe as the fresh DDPM study. These are matched extensions of the same baselines, not three additional independent baseline replications. The 200 reserved audit images remain excluded from the 9,800-image endpoint; no audit alarm, model selection or intervention is performed.

Each generation samples 50,000 images, with 5,000 requested from each of ten conditioning classes, from the pinned `G_ema` checkpoint. The preceding classifier, rather than the conditioning ID, supplies each training label. Balanced generation requests therefore do not guarantee balanced or correct pseudo-labels. All 120 assigned-class counts across the twelve pools are retained. Sampling uses 512-dimensional Gaussian latents, truncation $\psi = 1$, constant synthesis noise, batch size 64 and float32 reference operators. The latent seed is the base seed plus $30{,}000 + 1{,}000(g-1)$; the

Table 26: Rank-definition changes for the CIFAR stem (%), means ± sample SD of paired relative changes across three seeds. Definitions are identical to Table 4.

Global summary	g0–g1	g1–g4	g0–g4
Covariance entropy	$+8.33 \pm 8.36$	-4.86 ± 8.31	$+2.60 \pm 1.08$
Uncentered SVD entropy	$+19.27 \pm 7.29$	-18.39 ± 4.50	-2.79 ± 5.26
Centered SVD entropy	$+37.40 \pm 9.28$	-22.45 ± 4.86	$+6.30 \pm 3.97$
Participation ratio	-9.97 ± 8.72	$+5.64 \pm 12.57$	-5.62 ± 3.56
Feature stable rank	-22.30 ± 11.04	$+22.70 \pm 26.94$	-6.61 ± 9.82

Table 27: Direct endpoint entropy accounting for the CIFAR stem (nats), means ± sample SD across three seeds. Each interval is decomposed independently. The four signed contributions sum to total log-rank change.

Interval	ΔH_Σ	D_p	D_B	D_W	D_J
g0–g1	$+0.078 \pm 0.077$	$+0.278 \pm 0.025$	-0.135 ± 0.019	-0.106 ± 0.026	$+0.041 \pm 0.012$
g1–g4	-0.052 ± 0.087	$+0.054 \pm 0.008$	-0.004 ± 0.014	-0.106 ± 0.053	$+0.004 \pm 0.015$
g0–g4	$+0.026 \pm 0.011$	$+0.304 \pm 0.004$	-0.131 ± 0.004	-0.191 ± 0.002	$+0.044 \pm 0.009$

archive identifies the checkpoint, latent streams, generated pools and pinned source revision. This protocol changes generator family, conditional versus unconditional generation, sampling and the resulting input distribution together, so it does not isolate a causal effect of generator family.

The prespecified primary comparison is each seed's generation-0-to-4 accuracy change, scatter fold and effective-rank percentage change. Tables 28–29 retain every generation and the paired contrasts, including the retained DDPM generation 4 and real-only generation 1. The latter is a single retraining control, not a four-generation real-data trajectory. We report means and sample standard deviations across the three reused seeds; generations do not increase the independent sample count.

Generation 4 reaches $79.02 \pm 0.25\%$ accuracy, versus $88.35 \pm 0.08\%$ at generation 0. The primary within-seed changes are -9.33 ± 0.24 accuracy points, 1.540 ± 0.028-fold scatter and $-2.73 \pm 0.56\%$ rank. For seeds 101/102/103, accuracy changes are $-9.59/-9.29/-9.11$ points, scatter folds are $1.538/1.513/1.569$, and rank changes are $-2.71/-3.29/-2.18\%$. Mean accuracy decreases at each replacement, but rank fluctuates; seed 102 also shows a small accuracy increase and scatter decrease from generation 3 to 4. Relative to retained DDPM generation 4, the paired GAN accuracy gain is 16.85 ± 1.14 points, scatter fold is 0.478 ± 0.019, and rank change is $-15.05 \pm 2.18\%$. All assigned-class counts lie between 4,614 and 5,384 per 50,000-image pool. The GAN endpoint combines worse class scatter with a lower rank than its real-data baseline, qualifying the DDPM rank-increase finding; the workflow comparison does not attribute the difference to generator family alone.

The checkpoint's training membership and absence of test overlap are not independently established, and conditioning IDs are not verified true image labels. These finite CIFAR trajectories test a second generator workflow, without establishing asymptotic collapse, deployment performance or a unique geometric monitoring benefit.

Exploratory partial-real monitoring. After the original replacement and replay outcomes, we specified a separate three-seed 50%-real CIFAR-10 run to examine whether the severe replacement setting masked any difference in alarm timing. For generations 1–4, each 200-epoch model trains on a fixed class-balanced selection of 25,000 original real examples plus 25,000 examples selected from the matched original unconditional DDPM pool and pseudo-labeled by its new predecessor. The generation-0 classifiers, 200 labeled test-audit examples, disjoint 9,800-example test endpoints and four bootstrap thresholds are reused unchanged. The thresholds measure fixed-baseline audit resampling variation, not retraining-population error rates; alarms remain observational and never choose a training policy. Table 30 reports every endpoint and audit alarm. All 12 new models were independently checked against their frozen Python sources, real and synthetic selection identities, pseudo-label lineage, 200-epoch logs, final checkpoints, classifier heads, ordered evaluation arrays and frozen cutoffs. Scatter first alarms at generation 1 in all three seeds; accuracy alarms then only

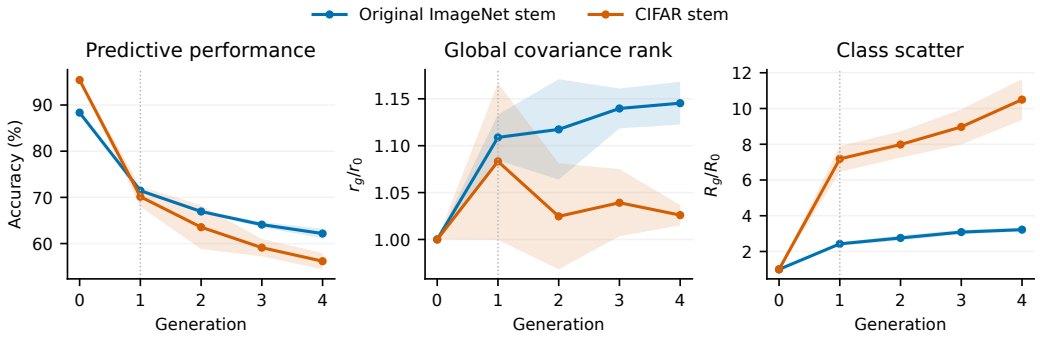

Figure 8: **Input-stem sensitivity on the same CIFAR-DDPM image pools.** All g0–g4 models for the CIFAR stem are newly trained. Lines and bands show means and one sample SD across three seed trajectories; rank and scatter normalize each trajectory to its own baseline. The full-trajectory covariance-rank mismatch remains, while its magnitude and later evolution differ.

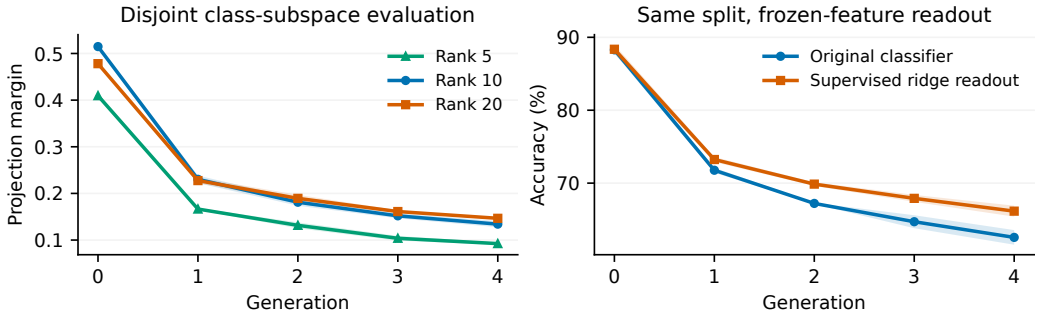

Figure 9: **Class structure and decoder recovery with the original ImageNet-style stem.** Subspaces and supervised readouts use the same 4,900 fitting and 4,900 disjoint evaluation images, excluding the 200 monitoring images. Left: projection margins for three subspace ranks. Right: original classifier and ridge readout accuracy on identical evaluation rows. Bands show one sample SD over three seeds. The fitting label budget is separate from small-label auditing.

in seed 103 and never through generation 4 in seeds 101 and 102. ECE first alarms at generation 3 only in seed 103, and NLL only in seed 102. The generation-1 scatter changes exceed their cutoffs by 0.0174, 0.1781 and 0.0376, respectively, so one margin is small. Generation-4 accuracy is $83.44 \pm 0.26\%$ versus $62.17 \pm 1.04\%$ for paired replacement, a within-seed gain of 21.27 ± 0.84 points; within the mixed-data trajectory scatter rises 1.442 ± 0.038-fold and rank rises $10.19 \pm 2.21\%$ (mean $\pm$ sample SD). These observations come from a setting chosen after the original outcomes, with only three seeds and reused generator pools. They do not establish an intervention advantage, calibrated false-alarm probability or general lead time over accuracy. A separate post-outcome audit counted the assigned training targets in every seed and generation, verifying each pseudo-label file and the selected generation-1 labels against the matched original full pool. For the 25,000 synthetic rows, total variation of the class marginal from uniform ranges from 0.317 to 0.379 and the entropy-effective class count from 6.31 to 7.31 of 10. The balanced 25,000 real rows halve that total variation, but the full training-target marginal still has total variation 0.159–0.189 and effective count 8.79–9.18; mean full-target total variation rises from 0.164 at generation 1 to 0.182 at generation 4. The 200-label test audit itself remains class-balanced. This descriptive analysis cannot identify how much of the representation change is caused by assigned-label imbalance versus other changes in the synthetic data or classifier trajectory. All 12 count vectors accompany the artifact.

Generation-1 class-weighting control. After the mixed-data outcomes and label-marginal audit, we specified a three-seed generation-1 control that reuses exactly the same 25,000 real and 25,000 synthetic images, their assigned labels, predecessor checkpoints, initialization seeds, evalu-

Table 28: CIFAR-10 generator-workflow sensitivity (mean ± sample SD across three reused seeds). All GAN replacement generations use the same retained real-data $g0$ models and disjoint 9,800-image endpoints. DDPM $g4$ and real-only $g1$ are retained comparisons; the control is not a four-generation real-only trajectory. The GAN generator is fixed. ECE uses 15 equal-width bins; NLL uses natural logarithms.

Condition	Acc. (%)	Scatter R	Rank r	ECE (%)	NLL
Real $g0$	88.35 ± 0.08	0.503 ± 0.002	12.993 ± 0.090	7.18 ± 0.11	0.474 ± 0.006
GAN $g1$	83.00 ± 0.44	0.658 ± 0.014	12.648 ± 0.015	11.17 ± 0.34	0.759 ± 0.015
GAN $g2$	81.18 ± 0.23	0.717 ± 0.022	12.631 ± 0.134	13.05 ± 0.20	0.903 ± 0.022
GAN $g3$	79.49 ± 0.36	0.761 ± 0.011	12.729 ± 0.099	14.60 ± 0.14	0.999 ± 0.014
GAN $g4$	79.02 ± 0.25	0.774 ± 0.014	12.638 ± 0.105	15.00 ± 0.39	1.042 ± 0.029
DDPM $g4$	62.17 ± 1.04	1.620 ± 0.048	14.882 ± 0.308	29.27 ± 1.42	2.267 ± 0.132
Real-only $g1$	88.48 ± 0.18	0.492 ± 0.005	12.841 ± 0.114	7.09 ± 0.25	0.465 ± 0.008

Table 29: Paired generator-workflow changes (mean ± sample SD). Ratios and changes are computed within each seed before averaging; all signs are retained. GAN $g0 \rightarrow g4$ is the primary comparison. This post-outcome setting changes generator family, conditioning and sampling jointly, reuses earlier seeds, and does not test monitoring or intervention advantage. Conditioning IDs are not verified image labels.

Contrast	ΔAcc. (pp)	R_b/R_a	Δr (%)
Real $g0 \rightarrow$ GAN $g4$	-9.33 ± 0.24	1.540 ± 0.028	-2.73 ± 0.56
Real $g0 \rightarrow$ DDPM $g4$	-26.18 ± 0.98	3.222 ± 0.109	14.54 ± 2.28
Real $g0 \rightarrow$ Real-only $g1$	0.12 ± 0.15	0.979 ± 0.010	-1.17 ± 1.20
DDPM $g4 \rightarrow$ GAN $g4$	16.85 ± 1.14	0.478 ± 0.019	-15.05 ± 2.18

ation examples and frozen audit cutoffs. The only training change is cross-entropy target weighting by $50{,}000/(10n_c)$, where n_c is the assigned count for class c among the 50,000 training rows; the weights range from 0.407 to 1.618 across seeds. All three 200-epoch models passed independent input, source, checkpoint, classifier-head, ordered-array and metric verification. For seeds 101/102/103, unweighted versus weighted generation-1 endpoint accuracies are 84.44/84.39%, 84.87/84.99% and 84.61/85.01%. The frozen scatter-alarm margins (observed log-change minus cutoff) change from $+0.0174$ to -0.0131, $+0.1781$ to $+0.1182$ and $+0.0376$ to $+0.0731$, respectively. Thus the scatter-only alarm disappears in seed 101, persists in seed 102, and remains tied to the accuracy alarm in seed 103. The assigned-label marginal itself is unchanged. This post-outcome, single-generation control shows that one near-cutoff alarm is objective-sensitive; it does not isolate true-class imbalance, establish general lead time or test an intervention policy.

Sensitivity to trusted-set selection. After the monitoring, class-weighting and intervention outcomes, we drew 200 alternative audits per seed from the 9,800 original endpoint images, excluding the original 200 audit examples. Each draw samples 20 images per class without replacement and fixes their identities across all model generations and controls. Its four cutoffs use only the selected generation-0 outputs and the same 2,000-pair stratified-bootstrap rule. Table 31 reports conditional alarm frequencies: generation-1 scatter-only alarms occur in 67.5–69.5% of draws for unweighted models and 67.0–76.5% for the matched weighted models. Thus weighting does not eliminate scatter-only alarms across alternative audits. The three fixed real-only controls have scatter alarms in 0/200 draws each and accuracy alarms in 2/200, 2/200 and 0/200. All 600 selected sets and 4,200 baseline/control/model score sets were checked against the original scalar audit implementation; full scalar bootstrap calibration was additionally replayed for the first draw of each seed. These overlapping draws reuse three fixed model trajectories and do not increase the number of independent training replicates, calibrate future false-alarm probability or establish counterfactual policy performance. Retrospective class-balanced selection uses the known test-pool labels and does not establish an acquisition cost of only 200 annotations. Original audit and endpoint measurements are unchanged.

Table 30: Exploratory CIFAR-10 monitoring with 25,000 real and 25,000 synthetic training rows at each generation. Accuracy, scatter ratio R and effective rank r use 9,800 official-test endpoint images; alarms use a disjoint fixed 200-image audit and the matched generation-0 thresholds. S/A/E/N denote scatter/accuracy/ECE/NLL alarms; a dash means none. All 12 post-baseline models were independently verified.

Seed	Gen.	Accuracy (%)	R	r	Alarms
101	0	88.33	0.500	12.890	–
	1	84.44	0.713	14.491	S
	2	83.54	0.743	14.604	S
	3	83.64	0.735	14.552	S
	4	83.65	0.734	14.521	S
102	0	88.44	0.505	13.055	–
	1	84.87	0.704	14.164	S
	2	83.53	0.733	14.507	S
	3	83.00	0.723	14.103	S,N
	4	83.50	0.738	14.304	S,N
103	0	88.30	0.504	13.035	–
	1	84.61	0.706	14.276	S,A
	2	83.51	0.709	13.919	S,A
	3	83.44	0.715	14.106	S,A,E
	4	83.15	0.705	14.125	S,A

Table 31: Post-outcome trusted-set selection sensitivity on fixed CIFAR models. Entries are percentages of 200 class-balanced audit draws per seed. Generation-1 columns compare unweighted and class-weighted training; generation-4 timing uses the unweighted 50%-real trajectory. S/A denote scatter/accuracy. In the generation-4 group, S only means no accuracy alarm by generation 4; S first means both eventually alarm but scatter does so earlier. The generation-4 accuracy-only category is zero in all seeds. These are conditional sampling frequencies, not new training replicates or population false-alarm estimates.

	Generation-1 S only		First alarms by generation 4				
Seed	Unweighted	Weighted	S only	S first	Tied	A first	Neither
101	68.0	67.0	38.5	33.5	27.5	0.0	0.5
102	69.5	68.0	24.0	48.5	26.5	0.5	0.5
103	67.5	76.5	30.5	42.5	25.5	0.5	1.0

Paired-accuracy comparator. We additionally checked the original audits and all 600 alternative audits with a post-outcome paired correctness comparator. For each later model versus generation 0 on the same 200 images, let L count correct-to-wrong transitions and G wrong-to-correct transitions. The one-sided binomial tail is $p = \Pr\{\mathrm{Binomial}(L+G, 1/2) \geq L\}$, with $p = 1$ if there are no discordant predictions. We report $p \leq 0.05$ and the more stringent $p \leq 0.0125$ rule, which allocates a nominal 0.05 budget across four generation comparisons. The binomial null assumes independent discordant pairs with equally probable directions; class-balanced sampling does not establish that null for heterogeneous future data. All per-class counts are retained. On the original audits, both rules preserve the previous timing: scatter alone alarms in seeds 101/102 through generation 4, while both monitors first alarm at generation 1 in seed 103. Across alternative audits, the unweighted generation-1 scatter-only frequencies decrease from 67.5–69.5% with the original accuracy cutoff to 54–56% with the paired 0.05 rule; weighted-model frequencies are 52–69.5% (Table 32). The fixed real-only controls have paired-accuracy alarms in 8/200, 5/200 and 1/200 draws at 0.05, and none at 0.0125. Thus the pattern persists but its frequency depends on the comparator. Every one of the 3,618 model-pair comparisons was independently recounted from logits; binomial tails were checked using exact integer sums. Geometry thresholds are unchanged and the operating points are not matched. Neither nominal rule establishes a future false-alarm rate, independent replication or a counterfactual intervention outcome.

Table 32: Post-outcome paired-accuracy sensitivity. L/G counts correct-to-wrong/wrong-to-correct transitions from generation 0 to the unweighted generation-1 model on the original audit; p is the one-sided binomial tail. Remaining entries are percentages of 200 alternative audits with a generation-1 scatter alarm but no paired-accuracy alarm, for unweighted (U) and weighted (W) training. Geometry cutoffs remain unchanged; the monitors do not have established equal false-alarm operating points.

	Original audit, U		$\alpha = 0.05$		$\alpha = 0.0125$	
Seed	L/G	p	U	W	U	W
101	10/11	0.668188	54.5	52.0	76.0	74.5
102	9/7	0.401810	56.0	54.0	81.0	79.0
103	20/3	0.000244	54.0	69.5	76.5	79.0

Alarm-triggered real-data intervention. After observing the 50%-real trajectories and class-weighting outcomes, we specified a separate three-seed policy pilot with the same generation-0/1 models, 200-label audit, 9,800-example endpoint, frozen thresholds, generator pools, SGD recipe and 50,000-row budget. A first scatter or accuracy alarm switches permanently, one generation later, from 25,000 real/25,000 pseudo-synthetic rows to 37,500 distinct class-balanced real/12,500 pseudo-synthetic rows. The additional real rows are drawn from the original CIFAR-10 training set, with original labels; 12,500 synthetic indices are nested within the corresponding 50%-real pool. All three scatter alarms occur at generation 1, while only seed 103 has an accuracy alarm then. Consequently, geometry switches all seeds from generation 2, accuracy switches only seed 103, and an unconditional generation-2 switch is exactly identical to geometry in fitted model identity. Table 33 gives generation-4 official-test accuracy. Geometry minus accuracy is $+2.94$, $+2.89$ and $+0.00$ points in seeds 101/102/103, or $+1.94 \pm 1.68$ points (mean $\pm$ sample SD). Geometry minus staying 50% real is $+3.09 \pm 0.30$ points. The new generation-2 full training-target total variation from uniform is 0.081/0.088/0.088, versus 0.163/0.174/0.177 in the matched 50%-real trajectories. All nine new generation-2–4 models passed independent checks of source and input hashes, pseudo-label lineage, 200-epoch logs, checkpoints, classifier heads, ordered endpoint and audit arrays, and frozen alarms. This selected post-outcome comparison changes both the amount of correctly labeled real training data and the training-target marginal. It cannot establish a unique value of the geometry alarm, a calibrated population false-alarm rate or a general intervention benefit; the same generator pools and three seeds are reused. The plan, verifier and per-seed trajectories accompany the artifact.

Table 33: Post-outcome CIFAR-10 intervention at generation 4. All policies share generation 0 and the 50%-real generation 1. A frozen audit alarm switches the subsequent generations to 75% real training. Entries are official-test accuracy (%); the final column is geometry minus accuracy policy in percentage points. The unconditional switch at generation 2 uses exactly the geometry-policy models in all seeds.

Seed	Stay 50%	Geometry alarm	Accuracy alarm	Difference
101	83.65	86.59	83.65	+2.94
102	83.50	86.39	83.50	+2.89
103	83.15	86.58	86.58	+0.00
Mean $\pm$ SD	83.44$\pm$0.26	86.52$\pm$0.11	84.58$\pm$1.74	+1.94$\pm$1.68

Six-seed replication in the selected retention setting. After the preceding outcomes, we froze a new protocol and independent verification contract before fitting or inspecting seeds 201–206. Each seed uses a new generation-0 classifier, a new class-balanced 200-image audit, new retained-real subsets and four newly sampled 50,000-image DDPM pools. The other 9,800 official-test images form a disjoint endpoint. All models use 200 epochs, batch size 128, SGD with learning rate 0.1, momentum 0.9, weight decay 5×10^{-4} and cosine decay. Generation-0 outputs alone calibrate 2,000-pair bootstrap cutoffs. The six policies are staying 50% real, scatter-triggered switching, accuracy-triggered switching, paired-accuracy switching at nominal $p \leq 0.05$ and $p \leq 0.0125$,

Table 34: CIFAR-10 policy replication in the previously selected 50%-to-75% real-data setting. Entries are means ± sample SD across all six new seeds. Δ is geometry minus the row policy at g4, in percentage points. Real rows sum g1–g4 training exposure (thousands), not newly acquired annotations. The final column counts identical g4 checkpoints with geometry; such ties provide no unique diagnostic advantage. Monitor false-alarm operating points are not matched.

Policy	g4 acc. (%)	Δ (pp)	Real rows (10^3)	Shared g4
Stay 50%	83.28 ± 0.42	$+3.09 \pm 0.41$	100.00 ± 0.00	0/6
Geometry	86.37 ± 0.30	$+0.00 \pm 0.00$	137.50 ± 0.00	6/6
Accuracy	85.09 ± 1.63	$+1.27 \pm 1.74$	118.75 ± 17.23	2/6
Paired .05	85.56 ± 1.47	$+0.81 \pm 1.44$	125.00 ± 15.81	3/6
Paired .0125	85.09 ± 1.63	$+1.27 \pm 1.74$	118.75 ± 17.23	2/6
Fixed g2	86.37 ± 0.30	$+0.00 \pm 0.00$	137.50 ± 0.00	6/6

Table 35: All six new-seed g4 accuracies (%) and first intervention generations in parentheses. A dash means no intervention within g1–g4. Policies use their own previous checkpoint and the same 200-label audit; the disjoint 9,800 endpoint images never choose an intervention. Fixed g2 switches independently of diagnostic outputs.

Seed	Stay 50%	Geometry	Accuracy	Paired .05	Paired .0125	Fixed g2
201	83.84 (–)	86.33 (2)	86.03 (4)	86.33 (2)	86.03 (4)	86.33 (2)
202	83.49 (–)	86.84 (2)	83.49 (–)	85.98 (4)	83.49 (–)	86.84 (2)
203	82.58 (–)	86.26 (2)	82.58 (–)	82.58 (–)	82.58 (–)	86.26 (2)
204	83.13 (–)	85.93 (2)	85.93 (2)	85.93 (2)	85.93 (2)	85.93 (2)
205	83.30 (–)	86.47 (2)	86.15 (3)	86.15 (3)	86.15 (3)	86.47 (2)
206	83.32 (–)	86.38 (2)	86.38 (2)	86.38 (2)	86.38 (2)	86.38 (2)

and unconditional switching at generation 2. Every switch permanently raises the real fraction to 75% with a one-generation lag. Each policy uses its own preceding classifier; fitted models are shared only for identical inputs and lineage. All six seeds and complete generation-0–4 trajectories are retained, including absent, late and tied alarms.

Table 34 reports paired outcomes and training exposure. Geometry reaches $86.37 \pm 0.30\%$ generation-4 accuracy versus $85.09 \pm 1.63\%$ for accuracy-triggered switching, $85.56 \pm 1.47\%$ for paired accuracy at 0.05, and $83.28 \pm 0.42\%$ for staying 50% real. Paired geometry gains are respectively 1.27 ± 1.74, 0.81 ± 1.44 and 3.09 ± 0.41 percentage points (mean ± sample SD, six seeds). Geometry and unconditional generation-2 switching share the same fitted checkpoints in all six seeds. Geometry uses 137,500 real training rows summed over generations 1–4, versus an average 118,750 for accuracy-triggered switching; this counts repeated exposure to already labeled training data, not annotation acquisition. Table 35 retains every per-seed endpoint and first intervention time. Table 36 reports all uninterrupted alarm times and all six matched real-only controls, which have zero alarms under every rule. Six controls give little precision about retraining-population false alarms, and the monitors' operating points remain unmatched.

This adds stochastic replication within a previously selected CIFAR setting; it does not remove setting selection, disentangle increased real-data exposure from changed target marginals, or establish a unique geometric intervention benefit.

Common real-retraining null calibration. To compare diagnostic scores under one calibration rule, we froze a further study before fitting 19 new baseline/retraining pairs (seeds 501–519 and 1501–1519). Each of the 38 models uses the unchanged 200-epoch ResNet CIFAR recipe and all 50,000 original real training examples with their labels. A new class-balanced 200-image audit is shared within each pair; its remaining 9,800 official-test images form the disjoint endpoint. The six earlier real-only pairs and all 24 uninterrupted 50%-real generation-1–4 comparisons are held out from calibration. All five scores use the same 19 pairs: log scatter change, accuracy loss in integer audit-image counts, ECE change, NLL change, and negative exact one-sided paired-correctness tail probability. Scores retain their signs without clipping.

Table 36: First alarms on each uninterrupted 50%-real trajectory (top; dash means none through g4) and alarms on the matched real-only g1 controls (bottom). A g4 alarm has no intervention effect within this horizon. Six controls do not establish a precise retraining-population false-alarm rate; paired thresholds are nominal.

Seed	Scatter	Accuracy	ECE	NLL	Paired .05	Paired .0125
201	1	3	3	1	1	3
202	1	–	–	–	3	–
203	1	–	–	1	–	–
204	1	1	1	1	1	1
205	1	2	3	2	2	2
206	1	1	1	1	1	1
Matched real-only g1 control alarms (0/1)						
201	0	0	0	0	0	0
202	0	0	0	0	0	0
203	0	0	0	0	0	0
204	0	0	0	0	0	0
205	0	0	0	0	0	0
206	0	0	0	0	0	0

For each score, let s_i be calibration values and s a target value, with larger values indicating stronger evidence against the real-retraining null. A uniformly randomized ordering of all 49 calibration/target keys was fixed before the new training and independently of its outcomes. With priorities u_i, u, the upper rank is $r = 1 + \sum_{i=1}^{19} \mathbf{1}\{(s_i, u_i) > (s, u)\}$, using lexicographic order; we alarm at $r/20 \leq 0.05$. We also report the conservative rank $1 + \sum_i \mathbf{1}\{s_i \geq s\}$, retaining exact rational paired tails. The usual Monte Carlo rank construction provides background (Hemerik & Goeman, 2018). Under exchangeable null pairs and independent uniform tie ordering, symmetry makes the augmented rank uniform on 1–20. This is a marginal 5% statement over calibration and tie randomness, not a guarantee conditional on the realized calibration or proof of retraining exchangeability. The fixed CIFAR data and recipe define the null; it does not represent arbitrary benign shifts. Reusing this calibration across scores and four generations provides no familywise or sequential error control.

Table 37 reports every primary alarm count. At generation 1, scatter/accuracy/ECE/NLL/paired-accuracy alarm in 6/5/2/5/5 of six uninterrupted trajectories, respectively; real-only control counts are 1/1/0/0/1. All three control alarms occur in seed 204. Only seed 202 retains an earlier scatter alarm than accuracy: generation 1 versus generation 3. ECE never alarms through generation 4 in seeds 202 and 203; Table 38 preserves every first-alarm time. Tie-broken and conservative alarms agree in all 150 target-score comparisons. Thus common retraining-null calibration reduces the initial scatter-only contrast to one of six trajectories and reveals alarms absent under the earlier bootstrap cutoffs. Six controls give imprecise error estimates and do not establish equal operating points. No model is retrained under the new alarm rules; the preceding original-cutoff policy comparison is unchanged. The separate stored-path replay below examines these ranks conditional on existing fitted trajectories.

These experiments reuse the selected CIFAR setting and underlying data; they add training realizations, not independent datasets.

Held-out real-retraining error evaluation. Before fitting either calibration or evaluation models, we froze a separate evaluation of 60 new real-only baseline/retraining pairs (seeds 601–660 and 1601–1660). All 120 models use the unchanged 200-epoch CIFAR trainer and all 50,000 real training images and labels. Each pair has a new balanced 200-image audit and disjoint 9,800-image endpoint. The 19 calibration pairs and all preceding target scores remain fixed; the six earlier real-only controls are reported separately. No evaluation outcomes select these frozen rank thresholds, seeds or stopping times.

The primary rule embeds the old 49-key priority order into the sorted first 49 of 109 independent uniform draws and assigns the remaining 60 draws to the new evaluation keys. This retains every old score-rank comparison while supplying fresh evaluation priorities; it does not redraw the

Table 37: Common real-retraining null calibration. Each cell gives alarm counts with the prespecified tie ordering, then conservative inclusive ranks. All five scores use the same 19 new real-only retraining pairs and nominal per-comparison level 5%. Controls are six held-out real-only pairs; g1 cases are all six uninterrupted 50%-real trajectories. This is a diagnostic sensitivity, not a new intervention policy. Nominal validity requires exchangeable null pairs; six controls do not establish equal realized false-alarm rates or a precise population rate.

Score	Real controls (T, C)	Stay 50% g1 (T, C)
Scatter	1/6, 1/6	6/6, 6/6
Accuracy	1/6, 1/6	5/6, 5/6
ECE	0/6, 0/6	2/6, 2/6
NLL	0/6, 0/6	5/6, 5/6
Paired accuracy	1/6, 1/6	5/6, 5/6

Table 38: First common-null-calibrated alarms on all uninterrupted trajectories. Cells give tie-broken/conservative first generations; a dash means none through g4. A g4 alarm has no one-generation-lag intervention within this horizon. The same frozen calibration set is shared by scores and cases. This analysis does not control repeated-look error or measure performance under these new alarm rules. Scores are not clipped: a tied zero or negative signed change can trigger the tie-broken rule. The accompanying CSV retains every signed score, tie priority and both ranks.

Seed	Scatter	Accuracy	ECE	NLL	Paired accuracy
201	1/1	1/1	3/3	1/1	1/1
202	1/1	3/3	–/–	3/3	3/3
203	1/1	1/1	–/–	1/1	1/1
204	1/1	1/1	1/1	1/1	1/1
205	1/1	1/1	3/3	1/1	1/1
206	1/1	1/1	1/1	1/1	1/1

old ordering. We evaluate the same five signed scores with both this tie-broken rule (T) and conservative inclusive ranks (C). Table 39 reports every alarm count and equal-tailed two-sided 95% Clopper–Pearson interval (Clopper & Pearson, 1934). Coverage assumes independent identically distributed evaluation training, audit and tie draws conditional on the fixed data, recipe, calibration and embedded old priorities. It averages over new evaluation tie draws rather than conditioning on all their realized values. The conservative rule uses no tie randomness. Coverage is per score/rule, not simultaneous across ten intervals; the frozen pseudorandom implementation does not prove the sampling assumptions.

Primary scatter/accuracy/ECE/NLL/paired-accuracy counts are 2/14/2/4/10 of 60; conservative counts are 2/10/2/4/10. Scatter's observed rate is 3.33% with interval $[0.41, 11.53]\%$, versus accuracy's 23.33% with $[13.38, 36.04]\%$; conservative accuracy gives 16.67% with $[8.29, 28.52]\%$. These conditional estimates differ from the nominal level averaged over calibration randomness, and this realized calibration does not match the monitors' operating points. The scatter interval also does not establish a rate at most 5%. All twenty paired score/rule comparisons are retained in Table 40. For scatter versus accuracy, T yields 0 scatter-only, 12 accuracy-only, 2 joint and 46 absent alarms; C yields 1/9/1/49. These are descriptive discordances, not equivalence or intervention tests. The study adds stochastic realizations of one dataset and recipe, with no evidence for deployment-shift calibration, sequential control or unique diagnostic utility.

Post-outcome replay of calibrated policies. After inspecting the complete policy and calibration results and checking graph coverage, we reconstructed all five common-null policies under both rank rules on the stored training graph. An alarm at generation g first changes generation $g + 1$, permanently raising the real fraction from 50% to 75%. Until this first switch, the path is the verified uninterrupted trajectory; subsequent alarms cannot alter the permanent action. Each transition must match the generation, preceding checkpoint, real-row count, full input-identity hash and training seed. We additionally require an entire previously fitted policy to match the requested schedule and every checkpoint. Thus the analysis reuses specific verified training realizations, rather than

Table 39: Held-out real-retraining alarms under the unchanged 19-pair calibration. All 60 new pairs are included; the original six controls remain separate. T uses the frozen random tie ordering; C uses conservative inclusive ranks. Cells give count/60 (percent) and two-sided 95% Clopper–Pearson limits in percent. Intervals assume conditionally independent, identically distributed training, audit and evaluation-tie draws at this fixed dataset, recipe and calibration; they are per score/rule, not simultaneous. Even 0/60 alarms has a 5.96% upper limit. This does not establish matched error rates or deployment calibration.

Score	T alarms (%)	T 95% CI	C alarms (%)	C 95% CI
Scatter	2/60 (3.3)	[0.41, 11.53]	2/60 (3.3)	[0.41, 11.53]
Accuracy	14/60 (23.3)	[13.38, 36.04]	10/60 (16.7)	[8.29, 28.52]
ECE	2/60 (3.3)	[0.41, 11.53]	2/60 (3.3)	[0.41, 11.53]
NLL	4/60 (6.7)	[1.85, 16.20]	4/60 (6.7)	[1.85, 16.20]
Paired accuracy	10/60 (16.7)	[8.29, 28.52]	10/60 (16.7)	[8.29, 28.52]

Table 40: Paired alarm differences on the same 60 held-out controls. A-only and B-only count discordant pairs; Both and Neither retain the other pairs. The final column is 100(A-only minus B-only)/60, in percentage points. T and C denote the tie-broken and conservative rules. All ten score pairs and both rules are retained. These descriptive comparisons do not test equivalence, establish equal error rates, or measure intervention benefit.

Scores (A / B)	Rule	A-only	B-only	Both	Neither	A−B (pp)
Scatter / Accuracy	T	0	12	2	46	-20.00
Scatter / Accuracy	C	1	9	1	49	-13.33
Scatter / ECE	T	1	1	1	57	+0.00
Scatter / ECE	C	1	1	1	57	+0.00
Scatter / NLL	T	1	3	1	55	-3.33
Scatter / NLL	C	1	3	1	55	-3.33
Scatter / Paired accuracy	T	1	9	1	49	-13.33
Scatter / Paired accuracy	C	1	9	1	49	-13.33
Accuracy / ECE	T	12	0	2	46	+20.00
Accuracy / ECE	C	8	0	2	50	+13.33
Accuracy / NLL	T	12	2	2	44	+16.67
Accuracy / NLL	C	8	2	2	48	+10.00
Accuracy / Paired accuracy	T	4	0	10	46	+6.67
Accuracy / Paired accuracy	C	0	0	10	50	+0.00
ECE / NLL	T	1	3	1	55	-3.33
ECE / NLL	C	1	3	1	55	-3.33
ECE / Paired accuracy	T	0	8	2	50	-13.33
ECE / Paired accuracy	C	0	8	2	50	-13.33
NLL / Paired accuracy	T	2	8	2	48	-10.00
NLL / Paired accuracy	C	2	8	2	48	-10.00

assuming that a new GPU training run would produce identical weights. A generation-4 alarm has no intervention within this horizon.

Table 41 reports complete six-seed aggregates only. Scatter reaches $86.37 \pm 0.30\%$ generation-4 accuracy and equals fixed generation-2 switching in every seed. Accuracy, NLL and paired accuracy all select the same paths: five equal scatter, while seed 202 follows the original paired-0.05 path, switching at generation 4. Scatter minus each comparator is 0.14 ± 0.35 points, with one win and five ties; the single difference is 0.86 points in seed 202. Mean real-row exposure is 137,500 for scatter versus 133,333 for these comparators, still a resource difference. ECE requires an unavailable generation-4 transition from the uninterrupted generation-3 checkpoint in seed 205. Its six-seed mean is therefore unreported; neither a different predecessor nor a five-seed subset substitutes for the missing case. Both rank rules produce identical paths, not independent replications. The artifact retains all 60 score/rule/seed cases (58 available, two referring to the same missing ECE path), 12 fixed-control cases and full input/checkpoint lineage. This post-outcome comparison adds no models, removes neither setting selection nor exposure confounding, and provides no prospective or unique geometric intervention benefit.

Table 41: Post-outcome replay of common-null-calibrated policies on the verified stored training graph. Both rank rules give identical paths. Values are six-seed means ± sample SD; Δ is scatter-policy minus comparator accuracy in percentage points. Exposure is thousands of reused real training rows across g1–g4. Each available path matches a whole previously fitted trajectory, including every predecessor and input identity; no new model is trained. ECE lacks the required g4 transition for seed 205, so its six-seed aggregates are unavailable; the CSV retains all six cases. These results are conditional on stored training realizations and do not establish prospective benefit or matched realized alarm rates.

Policy	Coverage	g4 accuracy (%)	Δ (pp)	Real rows (10^3)
Stay 50%	6/6	83.28 ± 0.42	3.09 ± 0.41	100.00 ± 0.00
Fixed g2	6/6	86.37 ± 0.30	0.00 ± 0.00	137.50 ± 0.00
Scatter	6/6	86.37 ± 0.30	0.00 ± 0.00	137.50 ± 0.00
Accuracy	6/6	86.22 ± 0.22	0.14 ± 0.35	133.33 ± 10.21
ECE	5/6	–	–	–
NLL	6/6	86.22 ± 0.22	0.14 ± 0.35	133.33 ± 10.21
Paired accuracy	6/6	86.22 ± 0.22	0.14 ± 0.35	133.33 ± 10.21

Post-outcome subspace replication. We applied the historical split-sample residual-margin definition to the saved replacement features of the completed CIFAR-DDPM trajectories with the original ImageNet-style stem. The 200 audit images are excluded. Within each seed, the 980 endpoint images per class are divided by fixed alternating class-index order into 490 fitting and 490 disjoint evaluation images; this split is identical across generations. Class-centered randomized SVD with four power iterations fits 20 axes, and nested ranks 5, 10 and 20 are evaluated without reselecting examples. The rank-10 margin decreases from 0.515 ± 0.004 to 0.134 ± 0.007 (mean ± sample SD across three seeds); the mean within-seed reductions are 77.4%, 73.9% and 69.3% for ranks 5, 10 and 20. This is descriptive corroboration, not a predeclared intervention endpoint or a new independent training replicate. Source code and per-seed split hashes are in `analyze_fresh_cifar_subspace.py` and `fresh_cifar_subspace.json`.

Post-outcome frozen-feature readouts. After observing the original training and diagnostic outcomes, we specified a supervised readout analysis of all 15 original-stem CIFAR-DDPM replacement models (seeds 101–103, generations 0–4). Before fitting these readouts, we froze the method, three executed source files and 102 input-file hashes. The feature extractors and original classifier heads remain unchanged. The class-stratified 490/490 split is exactly the preceding subspace-analysis split, with identical row identities across generations and the 200 monitoring examples excluded. Fitting a new decoder uses 4,900 trusted labels per seed, reused across generations; this is separate from the paper's 200-label monitoring budget and is not an intervention result.

We fit an intercept and a ridge-regularized least-squares map to one-hot class targets. Each fit centers features using its training rows and divides by one global RMS coordinate scale; no coordinate-wise rescaling or whitening is performed. For standardized features Z and centered one-hot targets Y_c, the objective is $\|ZB - Y_c\|_F^2/n + \lambda\|B\|_F^2$; training class priors supply the unpenalized intercept. Within each model's fitting partition, five stratified folds select $\lambda \in \{10^{-4}, 10^{-3}, \ldots, 10^{4}\}$ by total validation accuracy, with exact ties choosing the largest value. Class-local fitting position modulo five assigns folds. Each fold refits preprocessing on its 3,920 training rows and validates on 980 rows; the 4,900 final evaluation rows never fit preprocessing or choose regularization. The selected value is then refit on all fitting rows. Euclidean nearest-center classifiers use the same fitting labels; the original network, ridge readout and nearest-center rule share every evaluation row.

Table 42 retains every generation. Ridge accuracy decreases from $88.36 \pm 0.63\%$ to $66.16 \pm 0.75\%$, a paired change of -22.20 ± 1.24 points, compared with -25.73 ± 0.74 points for the original head. At generation 4, ridge improves on the original head by 3.60 ± 1.17 points, so some performance is recoverable with additional supervised fitting, while this particular decoder does not restore baseline performance. Three selections reach the fixed grid boundary: seed 101 generation 4 and seed 103 generation 1 select 10^{-4}; seed 103 generation 0 selects 10^{4}. The grid was not expanded after seeing outcomes. This is not a measurement of optimal linear separability, a causal head-only attribution or a proof of irreversible information loss.

Table 42: Post-outcome frozen-feature readout diagnostic on the three CIFAR-DDPM trajectories with the original ImageNet-style stem. Each model uses 4,900 labeled examples for fitting and internal five-fold regularization selection, and 4,900 disjoint examples for evaluation; the 200 monitoring examples are excluded. All decision rules use the same evaluation rows. Values are means ± sample SD across three seeds; the last row is the within-seed g0-to-g4 change in percentage points. Ridge minus network is also a paired difference. This uses a larger supervised label budget than monitoring and does not establish optimal linear separability or irreversible information loss.

Generation	Network (%)	Ridge (%)	Nearest center (%)	Ridge–network (pp)
g0	88.29 ± 0.48	88.36 ± 0.63	88.38 ± 0.55	0.07 ± 0.15
g1	71.77 ± 0.35	73.24 ± 0.35	71.92 ± 0.38	1.47 ± 0.17
g2	67.22 ± 0.08	69.86 ± 0.26	68.32 ± 0.54	2.65 ± 0.21
g3	64.71 ± 0.89	67.91 ± 0.48	66.09 ± 0.96	3.20 ± 0.76
g4	62.56 ± 1.00	66.16 ± 0.75	64.47 ± 0.38	3.60 ± 1.17
Δ g0–g4 (pp)	-25.73 ± 0.74	-22.20 ± 1.24	-23.91 ± 0.54	–

Post-outcome score-ranking comparison. After observing the completed policy, calibration and 60-pair error-evaluation results, including their unequal alarm counts, we specified a descriptive ROC analysis. Its method, three new source files, existing reporters and exact input reports/archives were frozen before computing the ROC values. For each generation 1–4, the positive condition consists of the six uninterrupted 50%-real trajectories (seeds 201–206); the negative condition consists of the same 60 new real-only generation-1 retraining pairs (seeds 601–660). The underlying CIFAR data, executed classifier trainer, 200-epoch recipe and 200-label audit budget agree. The 19 calibration pairs and six older real-only controls are excluded from these ROC cohorts. We retain all five signed scores and all four generations, using baseline-relative changes without clipping or choosing their direction from the ROC outcomes.

For a fixed score and generation, let u_i be the six recursive-condition changes and v_j the 60 real-control changes. We report

$$\widehat{\mathrm{AUC}} = \frac{1}{360} \sum_{i=1}^{6} \sum_{j=1}^{60} \left[\mathbf{1}\{u_i > v_j\} + \tfrac{1}{2} \mathbf{1}\{u_i = v_j\} \right].$$

Stored floating scores are compared exactly; paired accuracy retains its exact negative binomial-tail fraction. Sorting complete tied score groups gives the empirical ROC. At empirical false-positive fractions $\alpha \in \{0.05, 0.10\}$, we linearly interpolate a tied group: $h_{t,q}(z) = \mathbf{1}\{z > t\} + q\mathbf{1}\{z = t\}$, with $q \in [0, 1]$ chosen so that $60^{-1} \sum_j h_{t,q}(v_j) = \alpha$. The displayed recursive-condition fraction is $6^{-1} \sum_i h_{t,q}(u_i)$. These are analytical expectations for uniform tie inclusion; no new random draws, deployed cutoffs or changes to the frozen rank rules are made.

Table 43 reports all 20 curves' summaries. Generation-1 AUCs for scatter/accuracy/ECE/NLL/paired accuracy are 1.0000/0.9069/0.8639/0.9750/0.9000. At 5% empirical control alarms, the corresponding recursive-condition fractions are 100/66.7/66.7/83.3/66.7%. Scatter AUC is 1.0000 at every generation; NLL also identifies all six generation-3 recursive cases at the displayed 5% level.

These condition labels are not independently verified collapse labels. The setting was selected after an earlier geometry-only pilot, and this analysis was chosen after seeing its completed results. The same six positive seeds recur across generations and the same controls appear in every curve; neither generations nor the 7,200 pair comparisons supply additional independent training seeds. The contrast can reflect input, pseudo-label or class-prior changes. No confidence interval, significance claim, prospective calibration, sequential guarantee or intervention effect follows. Matching these empirical ROC fractions does not repair the unequal operating points of the earlier fixed policies or establish generalization to another dataset or generator.

Table 43: Post-outcome discrimination of uninterrupted 50%-real recursive training from real-only retraining. Each generation uses six recursive-condition seeds and the same 60 real-only pairs. AUC gives half credit for score ties. The last columns are expected recursive-condition identification fractions (%) at exactly 5% and 10% empirical control alarm rates, using analytical randomization within tied score groups. These are descriptive ROC values, not observed integer alarm counts, prospective error control, collapse ground-truth labels or intervention effects. All five scores and four generations are retained; repeated generations do not add independent seeds.

g	Score	AUC	At 5% FPR	At 10% FPR
1	Scatter	1.0000	100.0	100.0
1	Accuracy	0.9069	66.7	83.3
1	ECE	0.8639	66.7	66.7
1	NLL	0.9750	83.3	83.3
1	Paired accuracy	0.9000	66.7	83.3
2	Scatter	1.0000	100.0	100.0
2	Accuracy	0.9306	66.7	70.0
2	ECE	0.9167	50.0	83.3
2	NLL	0.9750	83.3	83.3
2	Paired accuracy	0.9194	66.7	66.7
3	Scatter	1.0000	100.0	100.0
3	Accuracy	0.9472	83.3	83.3
3	ECE	0.9000	66.7	66.7
3	NLL	0.9972	100.0	100.0
3	Paired accuracy	0.9389	83.3	83.3
4	Scatter	1.0000	100.0	100.0
4	Accuracy	0.8917	66.7	83.3
4	ECE	0.8500	33.3	50.0
4	NLL	0.9472	66.7	83.3
4	Paired accuracy	0.8875	66.7	66.7

G ARTIFACT RECONSTRUCTION

The revised artifact preserves historical arrays and checkpoints and writes new analysis outputs. The numerical definition is versioned as `iclr2027-v1`. The local feature reanalysis is in `rebuttal/analyze_revision.py`; explicit remote-run reanalysis is in `rebuttal/recompute_saved_runs.py`. Their outputs are `iclr_revision_analysis.json` and `remote_probe_audit.json`. Split-sample margins and the earlier full-set audit are in `additional_analysis.json`, generated by `analyze_additional_metrics.py`. Source hashes, metric version and explicit run names accompany the reanalyses. The figure/table manifest binds the historical outputs to their sources. Tables for new studies use separate result and verification records listed in the artifact README.

`regenerate_figures.py` produces five figures and five historical tables, recording their hashes and explicit input files in a manifest. Separate table writers and inputs are documented in the artifact README. The reviewed bibliography's fields, primary-source URLs and full-text hashes are bound in `reference_audit.json`; `reference_online_check.json` records the metadata and full-text recheck. This records verification evidence rather than asserting that metadata checks prove every scientific claim.

The executable pipeline now reserves ImageNet probe images before training, uses the evaluation preprocessing for ImageNet pseudo-labels, seeds mixing and generation independently, and rejects resuming legacy metrics or a changed experimental configuration in an existing output directory. Experimental guard targets use generation-0 training batches in training mode on a copied model, leaving the checkpoint, evaluation data and subsequent random state untouched. A supplied missing classifier checkpoint is an error instead of a fallback labeling policy. The new results use the separately versioned fresh protocol specified in Appendix F; pipeline repairs do not turn old runs into corrected evidence.

Tests cover the covariance spectrum definition, translation/scale invariance, the weighted scatter decomposition including singleton classes, simplex geometry, label-corruption invariance, regularizer

gradients, probe exclusion and preprocessing, training-target calibration and interruption/resume behavior. The archived-data replay test is optional when raw features are unavailable. The supplied summary JSON files suffice to regenerate tables and figures; reproducing feature extraction or training additionally requires the original datasets, model dependencies and saved checkpoints. The source package does not redistribute ImageNet images or large trained checkpoints.

G.1 VERIFICATION RECORDS

The records below specify which saved outputs were independently checked and which computations were replayed. Their plans, source identities and verification reports accompany the artifact. These checks concern numerical consistency and execution provenance; their replay counts do not add independent training replications.

Initial fresh-study records and audits. The executable protocol is version `fresh-20260922-v2`. Each run records its protocol, frozen source hashes, pinned generator revision and file hashes, data identities, per-epoch training log, final checkpoint and feature/logit/label arrays. The independent ImageNet audit recomputes metrics from all nine model arrays and the manifest's ordered validation labels, verifies their saved classifier heads, and hashes the 55,000 selected real images and all six 50,000-image synthetic pools. It checks that pseudo-label inputs identify the preceding classifier checkpoint. The corresponding CIFAR audit independently recomputes all distinct generation-0/generation-4 model endpoints and the three real-only controls. These checks audit saved results and lineage; they do not repeat training or prove that a pretrained generator never saw an evaluation image. The summary `fresh_results.json`, verification records, protocol code and the ImageNet dataset manifest accompany the artifact.

ImageNet ResNet replay and backend matching. All nine models passed separate checks of source, dataset and generated-pool hashes, pseudo-label lineage, the complete epoch schedule, checkpoint heads and independently recomputed metrics. A canonical torchvision model and independently specified preprocessing replay all 5,000 validation images per checkpoint; all saved features and logits match exactly. Original extraction stores float32 tensors with TF32 convolution enabled. An initial direct CPU comparison failed because it did not match that backend; the documented amendment matches the original GPU backend and batch size for full replay, then additionally compares 200 images per model on CPU against TF32-disabled CUDA at the unchanged tolerance. The archive preserves the initial verifier, diagnosis and amendment. Training and saved endpoints were unchanged.

Fixed-initial-teacher verification. Independent verification rehashes all retained inputs, checks all nine complete 200-epoch logs and checkpoint artifacts, and reconstructs each classifier separately in torchvision. All 90,000 official-image predictions match the saved argmax decisions; feature/logit comparisons use the original stricter numerical tolerances, with 1,800 additional CPU/CUDA comparisons. All 450,000 newly assigned pseudo-labels are replayed with the producer's original 512-image float16 batches and must agree exactly. Metric recomputation uses float64 and independent covariance calculations. The first g2 labeling pass uses the initial cuDNN deterministic setting (false); later passes use the true setting installed by the original trainer. A separate post-outcome check enables determinism for every g2 pool and changes none of the 150,000 labels. These are checks of existing targets, not additional training examples.

Fixed-teacher device amendment. CUDA-enumeration differences required two restarts before any saved training artifact and a later correction of seed 101's verification device to match its producer's GPU model; physical training UUIDs were not recorded. The artifact preserves both verification releases and the runtime amendments. An AST comparison confirms unchanged numerical checks and tolerances outside release validation and device selection. All three seeds pass the amended verifier without changing any checkpoint, target or evaluation row.

Standard CIFAR stem verification. The independent verifier constructs the modified torchvision architecture directly and strictly maps the saved wrapper tensors. It rejects the original 7×7 checkpoint shape and checks the 4×4 final residual map for a 32×32 input. All 15 complete epoch logs, 150,000 official-image outputs, 3,000 additional CPU/CUDA audit comparisons and

600,000 inherited pseudo-label decisions passed the frozen verification contract. Pseudo-label replay uses the producer's batch size 512, float16 autocast and deterministic cuDNN setting, with exact argmax agreement. GPU selection uses recorded UUIDs. Complete per-seed trajectories and all rank-definition and entropy terms are retained regardless of their direction; outcomes do not select which seed or generation is reported.

GAN workflow verification. The independent audit checks all twelve new and nine retained checkpoints, their saved heads and metrics, and all 210,000 raw audit/endpoint image replays, with 4,200 CPU/CUDA comparisons. It replays all 600,000 new pseudo-label decisions and checks complete generated-pool hashes and latent identities. Independent generator replay covers 640 fixed, unselected images per pool (7,680 in total), plus ten CPU crosschecks per pool (120 total); it does not regenerate every image. The earlier DDPM pseudo-label chain is not newly regenerated.

Six-seed policy verification. Independent verification covers all 58 distinct fitted models, 580,000 official-image feature/logit replays, 11,600 CPU/full-precision-CUDA crosschecks, 925,000 pseudo-label replays, all six complete bootstrap calibrations and all 144 policy decisions. The artifact preserves all 180 seed/policy/generation rows and recounted target marginals for all 46 distinct post-baseline policy models. The plan, frozen training source and verification contract identify the executed study.

Common-null calibration verification. Independent verification checks all 38 new checkpoints, 380,000 raw official-image replays, 7,600 CPU/full-precision-CUDA comparisons and 150 target ranks, binding the original 30 target pairs to the completed policy study. The source artifact includes the frozen plan, training sources and verifier, both tables and all 245 signed calibration/target score rows, including priorities and both ranks.

Held-out null evaluation verification. Independent verification checks all 120 checkpoints, 1,200,000 raw official-image replays, 24,000 CPU/full-precision-CUDA comparisons, 300 new score-ranks and preservation of all 150 old target ranks. All ten intervals satisfy independent binomial-tail checks. The artifact retains the frozen plan, sources and verification contract, both tables, 300 signed score rows, ten interval rows and twenty paired-comparison rows. No calibration or policy result is refitted or pooled into this evaluation.

Stored-feature readout verification. Float64 eigendecomposition results were independently recomputed using direct normal-equation solves for all 675 fold/regularizer comparisons and 220,500 final predictions across the three decision rules. All 15 checkpoints, feature/logit/label hashes, classifier heads and official-label identities were checked. No new backbone training or raw-image inference replay was performed for this stored-feature analysis.

Retrospective ROC verification. A second implementation enumerates all 7,200 score-pair comparisons, 1,105 ROC points, 40 interpolated operating points, 120 positive-seed placements and 40 paired AUC contrasts. The artifact preserves these quantities and all 420 cohort score entries, binding 150 existing checkpoint identities. This analysis adds no training or raw-image inference.