QUILT: CONTINUAL CROSS-ENTROPY LEARNING FROM MOMENTS AND CENTROIDS

Anonymous authors
Paper under double-blind review

ABSTRACT

Exact moments support continual least squares, but do not determine the nonlinear updates required by cross-entropy. We introduce QUILT, a readout learner that combines exact historical moments with a bounded, class-wise centroid memory. The moments preserve the supervised label term, the centroids approximate the nonlinear prediction term, and a reusable matrix factorization supports descent on the compressed objective. An exact KL decomposition identifies the loss discarded by compression. An optional covariance term matches its second-order expansion. Under a Gaussian model, the converged corrected classifier is classification-consistent. Experiments cover seven datasets and four frozen encoders, with replay, functional-memory methods, and ACIL/GACL, DS-AL, and AIR under common retained-byte ceilings. The complete learner exceeds a matched RP-ridge head by 0.36 percentage points in a ten-seed replication, with a 95% paired interval of [0.29, 0.42]. At the larger budget, AIR and both QUILT variants have nearly equal mean accuracy. Their temperature-calibrated NLLs are 0.537 for AIR, 0.495 for QUILT-0, and 0.488 for QUILT-C across the 14 settings. The covariance correction itself has a small, conditional effect: changing only that term gives +0.39 points on native features and -0.04 on random features, with losses on individual datasets. Our evidence supports the combined memory design while limiting the accuracy claim for its covariance correction.

1 INTRODUCTION

Pretrained encoders make continual classification feasible with a small trainable readout. The remaining challenge is to retain enough information about earlier examples to update that readout as new classes arrive. Analytic methods retain sufficient statistics for cumulative ridge regression (Zhuang et al., 2022; 2024a; McDonnell et al., 2023). Dual-stream compensation and class reweighting extend this family beyond a single unweighted ridge head (Zhuang et al., 2024b; Fang et al., 2024). Functional memories retain representative inputs to approximate historical predictions or gradients (Khan & Swaroop, 2021; Jung et al., 2025). Moments preserve aggregate geometry, while representative points support nonlinear loss evaluations.

Cross-entropy makes their combination useful. Its gradient depends on the current classifier, so the two ridge statistics are insufficient for exact updates. A weighted centroid memory approximates the nonlinear term while retaining the class masses and means. Together with exact moments, this allows repeated cross-entropy optimization after the original examples are discarded. The practical question is whether this combined summary improves prediction enough to justify its additional storage and optimization.

We propose QUILT, a continual readout with exact moments, a fixed-size class-wise centroid memory, and a majorization-minimization update. The exact Gram matrix supplies a reusable optimization metric; the centroids supply nonlinear loss evaluations. Their feature–label cross-moment equals the original supervised statistic. The retained summary is independent of earlier readouts: it stores the original label term exactly, so optimization can restart without carrying its previous error into the memory. These properties apply to the uncorrected learner, QUILT-0, as well as the corrected variant, QUILT-C.

Compression also discards within-cell variation. We show that the resulting loss gap is exactly a sum of predictive KL divergences. An optional quadratic term, formed from the discarded covariance,

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.

matches this gap near uniform predictions and preserves a global update bound. The approximation has an explicit logit-dependent remainder. It can be loose at trained, confident predictions, so the correction need not improve accuracy. Our descent guarantee concerns the chosen compressed objective; historical loss and test performance remain separate quantities.

A Gaussian model identifies one regime in which correction helps. With two classes and one centroid per class, continued uncorrected optimization can lose the covariance-adjusted direction supplied by its first step. The correction preserves that direction at convergence and gives classification consistency at a fixed centroid count. This is a conditional mechanism, rather than a general explanation of performance on pretrained features.

We evaluate seven datasets using four frozen encoders: five visual datasets with DINOv2 and CLIP, and two text datasets with MiniLM and MPNet. We compare replay, K-priors, Compact Memory, and three analytic families under common retained-byte ceilings. ACIL/GACL, DS-AL, and AIR select their expansion width within the budget; a separate RP-ridge comparison fixes the feature map exactly. Ten additional seeds replicate a 0.36-point gain over that matched head, while the wider analytic comparison shows no overall lead over AIR at the larger budget. The isolated correction adds only 0.39 points on native features and slightly reduces RP accuracy. We therefore assess the complete memory design and the conditional contribution of correction separately.

2 Problem Setting and Motivation

A fixed feature map produces $\phi(x) \in \mathbb{R}^r$, including a constant bias coordinate when used. It may be a pretrained encoder alone or that encoder followed by a fixed nonlinear expansion. Arrival t supplies a labeled batch $\mathcal{D}_t$. A readout $B \in \mathbb{R}^{r\times v}$ predicts from logits $B^\top\phi(x)$ over a fixed, known set of v classes, without an arrival identifier. After n observations, the target empirical objective is

$$F(B) = L(B) + \tfrac{\lambda}{2}\|B\|_F^2, \qquad L(B) = \sum_{i=1}^{n} \ell(B^\top\phi_i, y_i), \qquad \ell(\theta, y) = \log\sum_{c=1}^{v} e^{\theta_c} - \theta_y. \quad (1)$$

Here $\lambda > 0$ regularizes the *sum* of losses, including the bias. A bounded memory of derived features and labels is allowed. We require a fixed number of retained entries for fixed r, v and centroid budget m, independently of the number of observations or arrivals. Temporary storage for the current batch is separate from this retained state.

The moments

$$G = \sum_i \phi_i\phi_i^\top, \qquad H = \sum_i \phi_i e_{y_i}^\top \quad (2)$$

update by addition and determine the joint ridge readout $(G + \lambda_{\mathrm{r}} I)^{-1}H$. For cross-entropy, the gradient is $P(B) - H$, where

$$P(B) = \sum_i \phi_i \,\mathrm{softmax}(B^\top\phi_i)^\top. \quad (3)$$

Its nonlinear dependence on B prevents recovery from a fixed set of polynomial moments for every feature distribution (Proposition 6 in Appendix E). This motivates retaining both moments and representative points. It does not imply an impossibility result for all finite memories, or a universal accuracy advantage for cross-entropy over squared loss.

3 Learning Cross-Entropy from Moments and Centroids

3.1 Retain Moments and Class-Wise Centroids

The moment state (G, H) is exact. The second part of the memory is $\mathcal{Q} = \{(\psi_j, w_j, c_j)\}_{j=1}^{M}$, $M \le m$, where each ψ_j is the weighted mean of one class-pure group of historical features, $w_j > 0$ is its mass, and c_j is its class. At each arrival, we merge old weighted nodes with new features and apply weighted k-means within classes. We assume m is at least the number of observed classes. Centroids are recomputed from their final assignments and are not renormalized after averaging. Repeated merging preserves their interpretation as means of original-example groups, although these groups need not be nearest-neighbor cells of the final nodes.

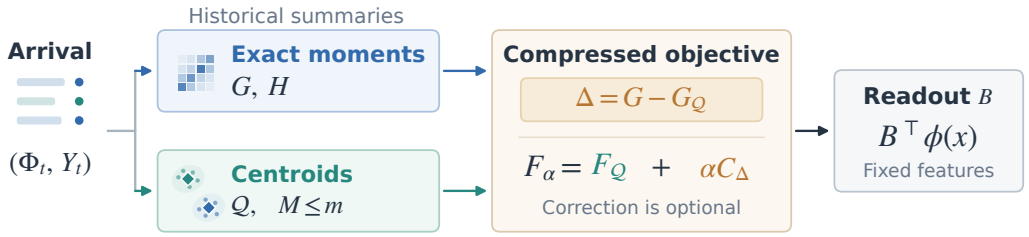


Figure 1: The memory and objective of QUILT. Exact moments retain historical second-order information; centroids support repeated nonlinear loss evaluations. The optional covariance term uses variation discarded by the centroids. Prediction uses the fixed feature map and the readout.

Define the summary moments and residual

$$G_{\mathcal{Q}} = \sum_j w_j \psi_j \psi_j^\top, \qquad H_{\mathcal{Q}} = \sum_j w_j \psi_j e_{c_j}^\top, \qquad \Delta = G - G_{\mathcal{Q}}. \tag{4}$$

For the original-example cells C_j represented by the nodes,

$$H_{\mathcal{Q}} = H, \qquad \Delta = \sum_j \sum_{i \in C_j} (\phi_i - \psi_j)(\phi_i - \psi_j)^\top \succeq 0. \tag{5}$$

The weighted version uses the corresponding sample masses. These identities hold through repeated weighted merges (Appendix E.2).

The nonlinear term is estimated by

$$\widehat{P}(B) = \sum_j w_j \psi_j \,\mathrm{softmax}(B^\top \psi_j)^\top. \tag{6}$$

Because $H_{\mathcal{Q}} = H$, the approximate gradient $\widehat{P}(B) - H$ is exactly the gradient of the weighted centroid loss

$$L_{\mathcal{Q}}(B) = \sum_j w_j \ell(B^\top \psi_j, c_j), \qquad F_{\mathcal{Q}}(B) = L_{\mathcal{Q}}(B) + \tfrac{\lambda}{2}\|B\|_F^2. \tag{7}$$

Both the new arrival and earlier data are represented by $\mathcal{Q}$ when evaluating $\widehat{P}$.

3.2 OPTIONAL CORRECTION FOR DISCARDED COVARIANCE

Let $\Pi_v = I_v - \mathbf{1}\mathbf{1}^\top/v$ and $B_c = B\Pi_v$ remove the common logit shift. We use the corrected objective

$$F_\alpha(B) = F_{\mathcal{Q}}(B) + \alpha C_\Delta(B), \qquad C_\Delta(B) = \frac{1}{2v}\,\mathrm{tr}(B_c^\top \Delta B_c). \tag{8}$$

The uncorrected variant, denoted QUILT-0, has $\alpha = 0$. The corrected variant, QUILT-C, fixes $\alpha = 1$. Section 4 derives this coefficient by matching the original cross-entropy curvature. The correction penalizes readout variation along directions whose variance was discarded by the centroids. Both variants use the same retained summary. The coefficient is a local curvature match, not an accuracy-optimal or universally preferred setting.

The softmax Hessian satisfies the classical bound (Böhning, 1992)

$$\mathrm{diag}(p) - pp^\top \preceq \tfrac{1}{2}\Pi_v \preceq \tfrac{1}{2} I_v. \tag{9}$$

Consequently $A = G/2 + \lambda I_r$ globally majorizes the curvature of F_α whenever $0 \le \alpha \le v/2$, including $\alpha = 1$ for every $v \ge 2$. The update is

$$g_\alpha(B) = \widehat{P}(B) - H + \lambda B + \frac{\alpha}{v}\Delta B \Pi_v, \qquad B^+ = B - A^{-1} g_\alpha(B). \tag{10}$$

One Cholesky factorization of A is reused within each arrival. Classical majorization-minimization (MM) (Böhning & Lindsay, 1988; Krishnapuram et al., 2005) gives

$$F_\alpha(B) - F_\alpha(B^+) \ge \tfrac{1}{2}\,\mathrm{tr}\big(g_\alpha(B)^\top A^{-1} g_\alpha(B)\big). \tag{11}$$

With $\lambda > 0$ and a fixed summary, the iteration converges to the unique minimizer of F_α (Appendix E.4). This does not imply monotonicity of the original historical loss.

Algorithm 1: QUILT with covariance coefficient α; QUILT-C uses $\alpha = 1$.

Retain: $G, H, \mathcal{Q}, B$, initialized to zero or empty.
For each arrival (Φ_t, Y_t), with one-hot labels:

1. Update $G \leftarrow G + \Phi_t^\top \Phi_t$ and $H \leftarrow H + \Phi_t^\top Y_t$.
2. Merge $\mathcal{Q}$ with new labeled features and compress to at most m class-wise weighted centroids.
3. Form $G_{\mathcal{Q}}$, $\Delta = G - G_{\mathcal{Q}}$, and factor $A = G/2 + \lambda I$.
4. Initialize $B \leftarrow [(G + \lambda_{\mathrm{r}} I)^{-1} H]\Pi_v$.
5. Repeat J times: evaluate $g_\alpha(B)$ using Equation (10); solve $AD = -g_\alpha(B)$; set $B \leftarrow B + D$.

Predict: $\arg\max_c [B^\top \phi(x)]_c$.

Resources. Retained tensors occupy $O(r^2 + rv + mr)$ entries. Prediction evaluates the fixed map ϕ and the linear readout B; the latter costs $2rv$ FLOPs. The correction has no inference cost. An optional d-to-D random expansion additionally stores its dD projection entries and costs $2dD$ dense FLOPs per prediction. Updating G costs $O(n_t r^2)$; forming $G_{\mathcal{Q}}$ costs $O(mr^2)$. Factoring A costs $O(r^3)$ once per arrival. Each step costs $O(mrv + r^2 v)$ for the nonlinear term, covariance product, and triangular solves. Centroid construction and batch storage are additional training costs. Exact counters require increasing precision on an indefinitely long stream; the stated bound concerns the number of retained entries.

4 Compression Error and Covariance Correction

4.1 What Compression Removes

The correction follows from the prediction discrepancy inside each cell. Let $p_i = \mathrm{softmax}(B^\top \phi_i)$ and $p_j^{\mathcal{Q}} = \mathrm{softmax}(B^\top \psi_j)$. The following is a Bregman-divergence identity for log-sum-exp (Banerjee et al., 2005), together with a curvature bound.

Proposition 1 (Discarded loss and local correction)**.** *For a class-pure centroid summary,*

$$L(B) - L_{\mathcal{Q}}(B) = \sum_j \sum_{i \in C_j} \mathrm{KL}(p_j^{\mathcal{Q}} \| p_i) \in [0, \tfrac{v}{2} C_\Delta(B)]. \tag{12}$$

If every historical logit vector has class range at most ρ, then

$$e^{-\rho} C_\Delta(B) \le L(B) - L_{\mathcal{Q}}(B) \le e^{\rho} C_\Delta(B), \qquad |L(B) - L_{\mathcal{Q}}(B) - C_\Delta(B)| \le (e^\rho - 1) C_\Delta(B). \tag{13}$$

Thus $L_{\mathcal{Q}} + C_\Delta$ matches L in value, gradient, and Hessian at $B = 0$.

The label term and the first-order Taylor term cancel within each cell because its weighted feature residuals sum to zero. The remaining log-sum-exp divergence is the KL in Equation (12), with the centroid prediction as its first argument. At uniform predictions, the logit Hessian is exactly Π_v / v; replacing the discarded feature covariance by Δ gives C_Δ. This explains both the coefficient and the centering in Equation (8). If $\|\phi_i\|_2 \le R$, we may take $\rho = \sqrt{2} R \|B\Pi_v\|_2$, so the remainder is third order near zero. The bound can be loose at large trained logits; second-order matching is not a uniform claim of accurate approximation. Away from uniform predictions, the softmax curvature changes with the readout. The fixed coefficient can therefore over- or under-weight the missing variation. This leaves room for dataset and optimization interactions even when the corrected objective decreases at every step.

The global interval in Equation (12) is computable from the retained state, but it bounds the loss at a readout, not the improvement between arbitrary updates. Descent on F_α and historical retention are distinct. Appendix E.10 gives a counterexample and a sufficient historical-descent test.

4.2 Covariance in the Update Versus Covariance in the Objective

For $\alpha = 0$, using G instead of $G_{\mathcal{Q}}$ in the MM metric adds

$$\tfrac{1}{4} \mathrm{tr}\big((B^+ - B)^\top \Delta (B^+ - B)\big)$$

to the usual centroid-loss quadratic surrogate. It penalizes the *increment*. It does not change the minimizer of $F_{\mathcal{Q}}$. In contrast, $C_\Delta(B)$ penalizes the *readout* and changes the limiting classifier. A Gaussian model separates these effects explicitly.

Theorem 2 (Covariance geometry at the first step and at convergence)**.** *Consider equal-prior binary classes $X \mid Y = y \sim \mathcal{N}(y\mu, \Sigma)$, where $\mu \neq 0$ and $\Sigma \succ 0$, and retain one population centroid per class. Use no intercept, write $B = (-a/2, a/2)$, and let $\bar{\lambda} > 0$ regularize the* mean *loss, so $\lambda_{\text{sum}} = n\bar{\lambda}$. The centroid objective's minimizer is a positive multiple of μ. Writing $w = (\Sigma + 2\bar{\lambda}I)^{-1}\mu$, an uncorrected exact-Gram MM step from zero and the corrected optimum satisfy*

$$a_G^{(1)} = \frac{2}{1 + \mu^\top w} w, \qquad a_1^\star = cw \quad \text{for some } c > 0. \tag{14}$$

They define the same classifier. As $\bar{\lambda} \downarrow 0$, their direction approaches the Bayes direction $\Sigma^{-1}\mu$. The limiting centroid classifier has strictly larger error than the Bayes classifier unless μ lies in one eigenspace of Σ.

The proof reduces the stationary equations to a scalar times a covariance-adjusted mean and applies the Sherman–Morrison identity. The result connects an early-iteration benefit to a property of the objective: the uncorrected method can lose favorable geometry as optimization continues, whereas the correction retains it at convergence. Ridge and classical discriminant analysis already use covariance information (Fisher, 1936); the theorem identifies how that information enters a centroid-based cross-entropy learner. It concerns classification boundaries in this model, not Bayes-calibrated probabilities or arbitrary encoder distributions. Appendix D tests this mechanism with estimated moments and independent samples; its large synthetic effect is not an estimate of the correction's contribution on pretrained features.

4.3 Statistical and computational consequences

The same model permits a finite-sample comparison. Take $n/2$ independent training examples per class, center them at the midpoint $\widehat{c} = (\bar{X}_+ + \bar{X}_-)/2$, and retain their two empirical centroids. Let $\widehat{\mu} = (\bar{X}_+ - \bar{X}_-)/2$ and let $\widehat{\Sigma}$ be the pooled within-class covariance with divisor n. For a fixed *sum-loss* penalty $\lambda_0 > 0$, the converged corrected classifier has direction

$$\widehat{w} = \left(\widehat{\Sigma} + \frac{2\lambda_0}{n} I\right)^{-1} \widehat{\mu}, \qquad \widehat{h}_C(x) = \operatorname{sign}\left(\widehat{w}^\top (x - \widehat{c})\right). \tag{15}$$

Corollary 3 (Finite-sample classification risk)**.** *Under the Gaussian data model of Theorem 2 in r dimensions, fix $\delta \in (0,1)$ and define $\gamma = \|\Sigma^{-1/2}\mu\|_2 > 0$ and $h = r + \log(4/\delta)$. There are universal constants c_0, c_1 such that, for even $n \geq c_0 h$, with probability at least $1 - \delta$ over the training data,*

$$\mathcal{R}(\widehat{h}_C) - \mathcal{R}^\star \leq c_1 \left[(\gamma + \gamma^{-1})\frac{h}{n} + \frac{\lambda_0^2 \gamma}{n^2 \lambda_{\min}(\Sigma)^2}\right], \tag{16}$$

where $\mathcal{R}$ is population misclassification probability and $\mathcal{R}^\star$ is Bayes risk. The corrected estimator is classification-consistent at fixed centroid count for fixed r, μ, Σ. The converged uncorrected estimator instead approaches a strictly positive excess risk whenever μ is not in one eigenspace of Σ.

Gaussian mean and covariance concentration control the estimated classifier. The risk depends quadratically on its direction error, and equal class priors cancel the first-order error from the estimated midpoint. These facts give the $O(r/n)$ rate for fixed signal strength. Appendix E.6 gives explicit constants and a sample-size experiment. The statement concerns classification under this model; it is not a log-loss consistency result or a distribution-free guarantee for normalized encoder features.

The correction also adds curvature in directions lost by compression. On centered readouts, a valid contraction factor for the objective gap is $1 - \underline{\mu}_\alpha$, where

$$\underline{\mu}_\alpha = \frac{\lambda + (\alpha/v)\lambda_{\min}(\Delta)}{\lambda + \frac{1}{2}\lambda_{\max}(G)}. \tag{17}$$

If Δ/n is uniformly positive definite and $\|G\|_2/n$ stays bounded, the corrected rate bound does not deteriorate with n at fixed $\lambda, \alpha > 0, v$. The uncorrected contraction parameter shrinks as $1/n$. The Gaussian model without an intercept satisfies the required covariance condition. An appended bias has zero residual variance, so this global improvement cannot be inferred for the full parameter space of our encoder experiments. The result identifies a computational benefit under explicit conditions, in addition to the change in the limiting classifier.

Table 1: Final visual test accuracy (%) on native encoder features, with the retained-state budget of QUILT-C at $m = 512$. All QUILT summaries remain unsaturated after the stream. Entries are mean $\pm$ sample standard deviation over seeds $0, 1, 2$; the Mean row first averages settings within each seed. Every method selects its own configuration on validation data. K-priors$^+$ includes the importance-weighted and longer-optimization variants; CM includes its optimizer-sufficiency control. Bold marks the best rounded mean. The corresponding text results and all variant results are in the appendix.

Dataset	Encoder	Ridge	ER	K-priors$^+$	CM	QUILT-0	QUILT-C
CIFAR-100	DINOv2	81.44 $\pm$0.00	77.54 $\pm$0.51	82.52 $\pm$0.22	69.79 $\pm$1.16	83.49 $\pm$0.17	**83.66** $\pm$0.14
	CLIP	73.80 $\pm$0.00	65.90 $\pm$0.63	74.58 $\pm$0.56	59.52 $\pm$1.27	76.78 $\pm$0.04	**77.04** $\pm$0.24
Pets	DINOv2	93.79 $\pm$0.00	84.33 $\pm$2.52	94.55 $\pm$0.24	94.04 $\pm$0.65	94.12 $\pm$0.11	**94.57** $\pm$0.16
	CLIP	87.74 $\pm$0.00	80.75 $\pm$1.88	87.78 $\pm$0.65	87.43 $\pm$0.39	88.37 $\pm$0.26	**89.51** $\pm$0.45
Flowers-102	DINOv2	99.35 $\pm$0.00	85.03 $\pm$2.02	**99.58** $\pm$0.00	99.39 $\pm$0.11	99.54 $\pm$0.02	99.57 $\pm$0.01
	CLIP	92.37 $\pm$0.00	74.07 $\pm$3.23	93.53 $\pm$0.00	91.32 $\pm$0.48	94.29 $\pm$0.15	**94.66** $\pm$0.19
DTD	DINOv2	75.90 $\pm$0.00	71.10 $\pm$1.22	78.05 $\pm$0.12	78.16 $\pm$0.41	77.68 $\pm$0.03	**78.51** $\pm$0.38
	CLIP	70.80 $\pm$0.00	67.32 $\pm$0.43	73.99 $\pm$0.19	73.48 $\pm$0.63	73.32 $\pm$0.17	**74.33** $\pm$0.48
EuroSAT	DINOv2	94.74 $\pm$0.00	92.35 $\pm$0.62	94.13 $\pm$0.21	92.26 $\pm$0.32	**95.38** $\pm$0.14	95.20 $\pm$0.01
	CLIP	94.54 $\pm$0.00	91.15 $\pm$0.92	93.70 $\pm$0.12	91.10 $\pm$0.63	**95.60** $\pm$0.10	95.09 $\pm$0.09
Mean		86.45 $\pm$0.00	78.95 $\pm$0.79	87.24 $\pm$0.11	83.65 $\pm$0.14	87.86 $\pm$0.03	**88.21** $\pm$0.04

5 EXPERIMENTS

Protocol. We use CIFAR-100, Oxford-IIIT Pets, Flowers-102, DTD, and EuroSAT with DINOv2-small and CLIP ViT-B/32, plus Banking77 and 20 Newsgroups with MiniLM and MPNet. The dataset-level main tables cover the ten visual settings; Table 3 also summarizes all 14 settings. Complete setting results remain in the appendix and in the two replication studies below. Each stream has ten class-incremental arrivals and a known output vocabulary. Predictions use every output class, without a task identifier or mask. All features are normalized per example; training, validation, and test pools are separate. Validation selects configurations and then a scalar temperature. The three seeds $0, 1, 2$ define the main comparison; the separate seeds 3–12 test prespecified paired contrasts.

Baselines and memory. The shared-feature comparisons use ridge, ER with AdamW, K-priors (Khan & Swaroop, 2021), and Compact Memory (CM; Jung et al., 2025). K-priors uses real inputs and an anchored functional prior; CM learns synthetic vectors and their weights by matching curvature. We include longer K-prior optimization and a class-mass importance-weighting control, with variants selected on validation data. The CM implementation checks the published update against the authors' code and allows both its minibatch normalization and the corresponding mean-loss normalization. Appendix A.6 details these adaptations. QUILT-0 retains the same moments and centroids as QUILT-C, but fixes $\alpha = 0$ instead of one.

In Tables 1–2, methods share the feature map and the same allowed retained-state bytes. The cap uses the QUILT-C state formula at capacity $\min(m, n_{\text{train}})$, with $m = 512$ or 8192, including any RP matrix. Other methods allocate their own number of memory items within that cap. Teacher aliases are counted once; temporary optimizer workspace is reported separately. Actual allocations and unused bytes appear in Appendix A.5. This comparison measures a storage constraint, not equality of memory cardinality or training time.

5.1 DIRECT COMPARISONS AT COMMON BYTE BUDGETS

Table 1 compares the native-feature learners under a strictly compressed QUILT memory. Corrected QUILT averages 88.21%, exceeding the strengthened K-priors baseline by 0.97 points and CM by 4.57 points. It exceeds CM in all ten settings and K-priors in nine. Both competitors already use CE, and the K-prior comparison includes historical class-mass weighting. These gains concern the complete memory design. The uncorrected QUILT variant is also competitive; the fixed-parameter experiment below isolates the smaller increment from covariance correction.

Table 2: Final visual accuracy with identical RP4096 features, using the retained-state budget of QUILT-C at $m = 8192$. The projection is exactly matched within each seed. All streaming competitors satisfy the common byte cap; their actual memory cardinalities can differ. Means and standard deviations use the same three seeds as Table 1. Both QUILT variants select their own penalty and iteration count on validation. Bold marks the best rounded mean. Small-data saturation and complete results are reported in the appendix.

Dataset	Encoder	RP-ridge	ER	K-priors	CM	QUILT-0	QUILT-C
CIFAR-100	DINOv2	85.39 ±0.07	83.76 ±0.60	84.84 ±0.27	85.32 ±0.18	86.01 ±0.14	**86.02** ±0.12
	CLIP	77.72 ±0.22	75.26 ±0.69	77.98 ±0.06	77.86 ±0.09	79.35 ±0.17	**79.55** ±0.22
Pets	DINOv2	94.37 ±0.31	91.69 ±1.13	**94.99** ±0.32	93.82 ±0.31	94.66 ±0.07	94.66 ±0.07
	CLIP	88.53 ±0.53	73.82 ±3.74	88.33 ±0.10	83.42 ±1.27	**88.63** ±0.07	**88.63** ±0.07
Flowers-102	DINOv2	99.54 ±0.00	97.69 ±0.65	**99.57** ±0.02	98.72 ±0.24	99.56 ±0.00	99.56 ±0.00
	CLIP	93.97 ±0.03	74.42 ±1.18	93.22 ±0.16	77.74 ±0.39	**94.39** ±0.12	**94.39** ±0.12
DTD	DINOv2	78.44 ±0.25	73.72 ±1.61	78.26 ±0.11	75.71 ±0.76	**78.62** ±0.61	**78.62** ±0.61
	CLIP	73.14 ±0.51	62.43 ±2.00	73.79 ±0.20	67.57 ±0.67	**74.15** ±0.52	**74.15** ±0.52
EuroSAT	DINOv2	96.20 ±0.11	94.76 ±0.71	96.05 ±0.11	95.56 ±0.32	**96.64** ±0.17	96.46 ±0.16
	CLIP	95.62 ±0.09	92.13 ±1.25	95.72 ±0.03	93.71 ±0.14	**96.19** ±0.09	95.88 ±0.18
Mean		88.29 ±0.14	81.97 ±0.66	88.28 ±0.06	84.94 ±0.19	**88.82** ±0.08	88.79 ±0.06

Table 3: Analytic families at the two main retained-byte ceilings. Accuracies (%) are averaged over settings within each seed, then reported as mean ± sample standard deviation over seeds $0, 1, 2$. Each analytic family selects its own expansion; Tables 1–2 fix the full feature map. Bold marks the best rounded mean. Scope, complete results and resources appear in Appendix B.

	Native $m = 512$ cap		RP $m = 8192$ cap	
Method	Vision 10	All 14	Vision 10	All 14
Ridge / RP-ridge	86.45 ±0.00	84.58 ±0.00	88.29 ±0.14	86.50 ±0.14
ACIL/GACL	84.81 ±0.08	83.09 ±0.09	88.43 ±0.05	86.74 ±0.07
DS-AL	85.81 ±0.15	84.00 ±0.13	88.25 ±0.09	86.50 ±0.12
AIR	84.80 ±0.12	83.14 ±0.13	88.46 ±0.06	**86.78** ±0.05
QUILT-0	87.86 ±0.03	85.64 ±0.04	**88.82** ±0.08	**86.78** ±0.11
QUILT-C	**88.21** ±0.04	**86.12** ±0.05	88.79 ±0.06	**86.78** ±0.07

Table 2 studies the larger RP representation. Corrected QUILT averages 88.79%, compared with 88.29% for RP-ridge, 88.28% for K-priors, and 84.94% for CM. The added uncorrected control reaches 88.82% at the same actual state size; correction gives no aggregate advantage in this primary RP comparison. K-priors remains higher on Pets/DINOv2 and Flowers/DINOv2. The QUILT-C configuration fixes $m = 8192$, $\alpha = 1$, and the search grid $J \in \{16, 32, 64, 128, 256, 512\}$ before evaluation. RP-ridge uses the same seeded ReLU map and per-example normalization. It is the analytic random-feature head, rather than a reproduction of the complete RanPAC adaptation pipeline. On Pets, Flowers-102, and DTD, the QUILT budget contains the entire fitting pool and Δ vanishes. Those rows evaluate complete-feature-memory CE; they do not establish a benefit from covariance correction. The native-feature offline joint CE reference and every prespecified budget remain in Appendix F.

5.2 A BROADER ANALYTIC COMPARISON

We additionally evaluate ACIL/GACL, DS-AL, and AIR (Zhuang et al., 2022; 2024a;b; Fang et al., 2024). These methods share the frozen encoders, splits, class orders, and byte ceilings, while selecting their own random-feature widths. DS-AL retains its second activation and residual-fitting stream; AIR retains its class reweighting. ACIL and GACL share one row because their ridge solutions coincide in this disjoint-class protocol. The complete 84-case study includes 252 independently refit models, with all choices made on validation data.

At the larger budget, AIR reaches 88.46% on vision, reducing QUILT-C's margin to 0.33 points. Across all 14 settings, their means are nearly identical: 86.783% for AIR and 86.779% for QUILT-

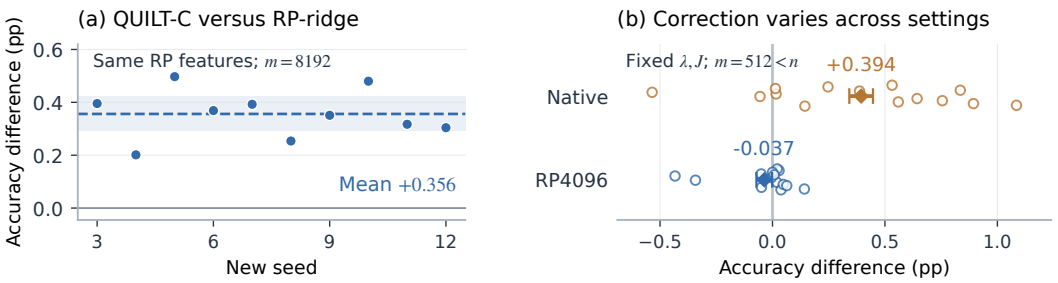


Figure 2: Paired evidence from seeds 3–12, across all 14 fixed settings. (a) Each point is one seed's equal-setting accuracy gain over RP-ridge; the dashed line and shading give the mean and its 95% paired t interval. (b) Only the covariance coefficient changes, at the uncorrected model's validation-selected λ, J and strictly unsaturated $m = 512$. Hollow dots are the 14 setting means; diamonds and bars give the macro mean and its 95% seed-block interval. Setting variation and seed uncertainty differ. The two panels answer different questions.

C. QUILT-0 reaches 88.82% on vision and 86.784% across all 14 settings at the larger budget, showing that competitiveness does not require correction. Mean calibrated NLL distinguishes their predictive losses: 0.537 for AIR, 0.495 for QUILT-0 and 0.488 for QUILT-C, with each temperature fitted on validation. The accuracy results do not establish superiority over the analytic family.

5.3 Does the RP gain replicate?

The initial three-seed RP gain was 0.276 points, with a paired interval $[-0.233, 0.784]$. We therefore fixed ten additional seeds, all 14 settings, and the same representation, budgets, and selection grids before running a separate replication. Its mean gain is 0.356 points, with a 95% paired interval of $[0.290, 0.422]$ (Figure 2a). All ten seed-level means are positive (two-sided exact sign test, $p = 0.00195$); a paired seed-block bootstrap gives $[0.301, 0.410]$. The visual mean gain is 0.546 points; the text estimate is -0.118, with interval $[-0.279, 0.043]$. These intervals concern algorithmic seed variation on the fixed benchmark splits. The datasets and repeated test examples are not treated as independent replicates. All setting-level effects are reported, including Flowers/DINOv2's near-zero estimate and the unfavorable text results.

5.4 What the covariance correction contributes

A separate study fixes $m = 512 < n$ for both representations and all settings. The primary rule chooses λ, J using the uncorrected model's validation scores, then changes only α from zero to one. Each pair has the same summary, initialization, feature map, and actual retained bytes. On the ten fresh seeds, this intervention improves native-feature accuracy by 0.394 points, with interval $[0.342, 0.447]$; all ten seed-level effects are positive. The visual-only gain is 0.290 points. This average is modest and heterogeneous: the 14 setting means range from -0.535 to $+1.085$ points. EuroSAT/CLIP loses accuracy in all ten seeds; CIFAR-100/DINOv2 and Flowers/DINOv2 gain only 0.016 and 0.013 points.

The RP primary effect is instead -0.037 points, with interval $[-0.071, -0.002]$ (Figure 2b). The prespecified reverse comparison selects parameters for the corrected model and then removes the correction; its RP difference is $+0.391$ points. Independently selecting parameters for both arms gives $+0.098$ points on RP features. These are different estimands: the positive retuned results do not replace the primary fixed-configuration result. The correction changes the useful operating configuration, including both penalty and iteration count. Both primary comparisons improve temperature-calibrated NLL while increasing raw NLL, consistent with a change in geometry and confidence scale. Appendix F.4 reports every setting, selector, and metric. Thus the correction is an optional, configuration-sensitive component, not the main source of the complete learner's accuracy. A retrospective validation choice between $\alpha = 0$ and one is reported in Appendix C; it does not replace the fixed-component intervention or establish an accuracy guarantee.

6 Related work

Analytic continual learning. ACIL solves continual regularized least squares exactly (Zhuang et al., 2022). GACL extends the recursion to reappearing classes (Zhuang et al., 2024a); with fixed features and disjoint classes, its solution equals ACIL's. RanPAC combines pretrained features, random projections, and an analytic head (McDonnell et al., 2023). DS-AL adds a second activation and residual-fitting stream (Zhuang et al., 2024b), while AIR reweights the objective to address class imbalance (Fang et al., 2024). We evaluate these distinct objectives and allow their projection widths to use the available byte budget. GKEAL studies kernel embeddings in few-shot CIL (Zhuang et al., 2023); F-OAL couples recursive least squares to fused pretrained features (Zhuang et al., 2024c), and AnaCP also changes the representation (Momeni et al., 2025). These representation and protocol choices extend beyond our controlled head comparison. Appendix B specifies which components are measured and which complete pipelines are not reproduced.

Functional memories. K-priors and Compact Memory retain inputs or synthetic directions that approximate historical logistic gradients (Khan & Swaroop, 2021; Jung et al., 2025). FROMP and Gaussian-process functional regularization preserve predictions at selected inputs (Pan et al., 2020; Titsias et al., 2020). Our method combines exact moments with class-wise centroids and explicitly characterizes their loss approximation, with an optional covariance correction.

Majorization and covariance penalties. The logistic Hessian bound and constant-curvature MM updates are classical (Böhning, 1992; Böhning & Lindsay, 1988; Krishnapuram et al., 2005). Stochastic and incremental majorization aggregate surrogates (Mairal, 2013; 2015), and PMF-CL uses quadratic task summaries in continual learning (Nair et al., 2026). Noise-based and dropout regularization also connect second-order loss expansions to covariance penalties (Bishop, 1995; Wager et al., 2013). We use this machinery to recover the covariance specifically removed by a continual centroid summary; neither covariance regularization nor MM is claimed as a new general principle.

Coresets and representation scope. Logistic coresets approximate an objective with weighted points (Huggins et al., 2016; Munteanu et al., 2018). Our additive, readout-dependent analysis does not imply a uniform multiplicative coreset guarantee. The experiments isolate readout learning on frozen visual and sentence encoders (Oquab et al., 2024; Radford et al., 2021; Reimers & Gurevych, 2019). They do not compare complete encoder-adaptation pipelines or language-model pretraining.

7 Discussion and limitations

The common-budget comparisons support combining supervised centroids with exact moments for continual CE. The uncorrected learner is already competitive, and the corrected learner's gain over the matched RP-ridge head replicates. That comparator does not represent the full analytic family: AIR matches the larger-budget 14-setting mean. We do not establish general superiority over analytic continual learning. The correction has a small positive native average, near-zero or negative effects elsewhere, and persistent losses on EuroSAT. Its positive macro interval does not imply a reliable improvement on each dataset. It should remain an optional approximation, assessed against $\alpha = 0$ on held-out validation data.

The guarantees establish descent on the corrected objective, local approximation to historical CE, and classification geometry under a Gaussian model. They do not establish per-task retention or distribution-free superiority over SGD. Repeated compression can accumulate error, and the logit-range bound can be loose at confident predictions. All encoders remain frozen and the output universe is known. State grows quadratically with feature dimension and linearly with class count. Changing the encoder, expanding the vocabulary without bound, and learning without pretraining remain open settings. The additional analytic experiments use three development seeds; the ten-seed RP replication does not provide a confirmatory comparison against DS-AL or AIR.

Reproducibility statement

Appendix A specifies preprocessing, configurations, selection rules, and resource accounting. Appendix E gives proofs, and Appendix G describes numerical checks. Tables and empirical figures are regenerated from saved run files by an accompanying rebuild script, whose manifest records every contributing run.

AI use statement

Generative AI assisted with writing and coding for this work. The authors verified all results and take full responsibility for the content.

References

Arindam Banerjee, Srujana Merugu, Inderjit S. Dhillon, and Joydeep Ghosh. Clustering with Bregman divergences. *Journal of Machine Learning Research*, 6(58):1705–1749, 2005. URL https://www.jmlr.org/papers/v6/banerjee05b.html.

Chris M. Bishop. Training with noise is equivalent to Tikhonov regularization. *Neural Computation*, 7(1):108–116, 1995. doi: 10.1162/neco.1995.7.1.108. URL https://doi.org/10.1162/neco.1995.7.1.108.

Dankmar Böhning. Multinomial logistic regression algorithm. *Annals of the Institute of Statistical Mathematics*, 44(1):197–200, 1992. doi: 10.1007/BF00048682.

Dankmar Böhning and Bruce G. Lindsay. Monotonicity of quadratic-approximation algorithms. *Annals of the Institute of Statistical Mathematics*, 40(4):641–663, 1988. doi: 10.1007/BF00049423.

Iñigo Casanueva, Tadas Temčinas, Daniela Gerz, Matthew Henderson, and Ivan Vulić. Efficient intent detection with dual sentence encoders. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pp. 38–45, 2020. doi: 10.18653/v1/2020.nlp4convai-1.5. URL https://aclanthology.org/2020.nlp4convai-1.5/.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3606–3613, 2014. doi: 10.1109/CVPR.2014.461. URL https://www.robots.ox.ac.uk/~vgg/data/dtd/.

Di Fang, Yinan Zhu, Runze Fang, Cen Chen, Ziqian Zeng, and Huiping Zhuang. AIR: Analytic imbalance rectifier for continual learning, 2024. URL https://arxiv.org/abs/2408.10349.

R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2):179–188, 1936. doi: 10.1111/j.1469-1809.1936.tb02137.x. URL https://doi.org/10.1111/j.1469-1809.1936.tb02137.x.

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, 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. EuroSAT: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019. doi: 10.1109/JSTARS.2019.2918242. URL https://doi.org/10.1109/JSTARS.2019.2918242.

Jonathan H. Huggins, Trevor Campbell, and Tamara Broderick. Coresets for scalable bayesian logistic regression. In *Advances in Neural Information Processing Systems*, 2016. arXiv:1605.06423.

Yohan Jung, Hyungi Lee, Wenlong Chen, Thomas Möllenhoff, Yingzhen Li, Juho Lee, and Mohammad Emtiyaz Khan. Compact memory for continual logistic regression. In *Advances in Neural Information Processing Systems*, 2025. arXiv:2511.09167.

Mohammad Emtiyaz Khan and Siddharth Swaroop. Knowledge-adaptation priors. In *Advances in Neural Information Processing Systems*, 2021. arXiv:2106.08769.

Balaji Krishnapuram, Lawrence Carin, Mário A. T. Figueiredo, and Alexander J. Hartemink. Sparse multinomial logistic regression: fast algorithms and generalization bounds. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(6):957–968, 2005. doi: 10.1109/TPAMI.2005.127.

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.

Ken Lang. NewsWeeder: Learning to filter netnews. In *Machine Learning Proceedings 1995*, pp. 331–339, 1995. doi: 10.1016/B978-1-55860-377-6.50048-7. URL https://doi.org/10.1016/B978-1-55860-377-6.50048-7.

Julien Mairal. Stochastic majorization-minimization algorithms for large-scale optimization. In *Advances in Neural Information Processing Systems*, 2013. arXiv:1306.4650.

Julien Mairal. Incremental majorization-minimization optimization with application to large-scale machine learning. *SIAM Journal on Optimization*, 25(2):829–855, 2015. arXiv:1402.4419.

Mark D. McDonnell, Dong Gong, Amin Parvaneh, Ehsan Abbasnejad, and Anton van den Hengel. RanPAC: Random projections and pre-trained models for continual learning. In *Advances in Neural Information Processing Systems*, pp. 12022–12053, 2023. arXiv:2307.02251.

Saleh Momeni, Changnan Xiao, and Bing Liu. AnaCP: Toward upper-bound continual learning via analytic contrastive projection. In *Advances in Neural Information Processing Systems*, 2025. arXiv:2511.13880.

Alexander Munteanu, Chris Schwiegelshohn, Christian Sohler, and David P. Woodruff. On coresets for logistic regression. In *Advances in Neural Information Processing Systems*, 2018. arXiv:1805.08571.

Srijith Nair, Atilla Eryilmaz, and Jia Liu. PMF-CL: Pareto-minimal-forgetting continual learner for conflicting tasks. *arXiv preprint arXiv:2605.19145*, 2026.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Proceedings of the Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pp. 722–729, 2008. doi: 10.1109/ICVGIP.2008.47. URL https://doi.org/10.1109/ICVGIP.2008.47.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=a68SUt6zFt.

Pingbo Pan, Siddharth Swaroop, Alexander Immer, Runa Eschenhagen, Richard E. Turner, and Mohammad Emtiyaz Khan. Continual deep learning by functional regularisation of memorable past. In *Advances in Neural Information Processing Systems*, 2020. arXiv:2004.14070.

O. M. Parkhi, A. Vedaldi, A. Zisserman, and C. V. Jawahar. Cats and dogs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3498–3505, 2012. doi: 10.1109/CVPR.2012.6248092. URL https://doi.org/10.1109/CVPR.2012.6248092.

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

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992, 2019. doi: 10.18653/v1/D19-1410. URL https://aclanthology.org/D19-1410/.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. MPNet: Masked and permuted pre-training for language understanding. In *Advances in Neural Information Processing Systems*, volume 33, pp. 16857–16867, 2020. URL https://papers.nips.cc/paper_files/paper/2020/hash/c3a690be93aa602ee2dc0ccab5b7b67e-Abstract.html.

Michalis K. Titsias, Jonathan Schwarz, Alexander G. de G. Matthews, Razvan Pascanu, and Yee Whye Teh. Functional regularisation for continual learning with gaussian processes. In *International Conference on Learning Representations (ICLR)*, 2020. arXiv:1901.11356.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2018. doi: 10.1017/9781108231596.

Stefan Wager, Sida Wang, and Percy S. Liang. Dropout training as adaptive regularization. In *Advances in Neural Information Processing Systems*, volume 26, 2013. URL https://papers.nips.cc/paper_files/paper/2013/hash/38db3aed920cf82ab059bfccbd02be6a-Abstract.html.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Advances in Neural Information Processing Systems*, volume 33, pp. 5776–5788, 2020. URL https://papers.nips.cc/paper_files/paper/2020/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Huiping Zhuang, Zhenyu Weng, Hongxin Wei, Renchunzi Xie, Kar-Ann Toh, and Zhiping Lin. ACIL: Analytic class-incremental learning with absolute memorization and privacy protection. In *Advances in Neural Information Processing Systems*, 2022. arXiv:2205.14922.

Huiping Zhuang, Zhenyu Weng, Run He, Zhiping Lin, and Ziqian Zeng. GKEAL: Gaussian kernel embedded analytic learning for few-shot class incremental task. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7746–7755, 2023. doi: 10.1109/CVPR52729.2023.00748.

Huiping Zhuang, Yizhu Chen, Di Fang, Run He, Kai Tong, Hongxin Wei, Ziqian Zeng, and Cen Chen. GACL: Exemplar-free generalized analytic continual learning. In *Advances in Neural Information Processing Systems*, 2024a. arXiv:2403.15706.

Huiping Zhuang, Run He, Kai Tong, Ziqian Zeng, Cen Chen, and Zhiping Lin. DS-AL: A dual-stream analytic learning for exemplar-free class-incremental learning. In *AAAI Conference on Artificial Intelligence*, 2024b. arXiv:2403.17503.

Huiping Zhuang, Yuchen Liu, Run He, Kai Tong, Ziqian Zeng, Cen Chen, Yang Wang, and Lap-Pui Chau. F-OAL: Forward-only online analytic learning with fast training and low memory footprint in class incremental learning. In *Advances in Neural Information Processing Systems*, 2024c. arXiv:2403.15751.

A Experimental details and reproducibility

A.1 Datasets and frozen representations

The visual datasets are CIFAR-100 (Krizhevsky, 2009), Oxford-IIIT Pets (Parkhi et al., 2012), Flowers-102 (Nilsback & Zisserman, 2008), DTD (Cimpoi et al., 2014), and EuroSAT (Helber et al., 2019). Text classification uses Banking77 (Casanueva et al., 2020) and 20 Newsgroups (Lang, 1995).

Table 4 lists the seven classification benchmarks. The split seed is 20260911, independently of the continual-learning seeds. CIFAR-100 and Pets hold out 20% of their official training pool for validation; Banking77 and 20 Newsgroups hold out 10% using stratification. Flowers-102 and DTD use their published validation splits. EuroSAT uses an explicitly constructed, stratified 60/20/20 split because it has no universally prescribed train/test partition. Validation examples are never merged back into training.

Table 4: Dataset sizes after splitting. Each vision dataset is evaluated with DINOv2-small and CLIP ViT-B/32; each text dataset with MiniLM and MPNet.

Dataset	Classes	Training	Validation	Test
CIFAR-100	100	40,000	10,000	10,000
Oxford-IIIT Pets	37	2,944	736	3,669
Flowers-102	102	1,020	1,020	6,149
DTD	47	1,880	1,880	1,880
EuroSAT RGB	10	16,200	5,400	5,400
Banking77	77	9,000	1,003	3,080
20 Newsgroups	20	10,184	1,130	7,532

Data provenance. Vision uses the pinned Hugging Face mirrors `uoft-cs/cifar100`, `timm/oxford-iiit-pet`, `dpdl-benchmark/oxford_flowers102`, and `dpdl-benchmark/dtd`; EuroSAT uses its original archive through Torchvision. The DTD mirror supplies the three published splits but does not identify its partition number, which we leave unspecified. Banking77 uses PolyAI's original CSV files and category order. 20 Newsgroups uses the verified by-date archive with the standard removal of headers, footers, and quoted text. Source-row splits are disjoint. The latter preprocessing leaves 270/30/215 empty strings in training/validation/test and nine distinct nonempty texts shared between training and test. These original benchmark rows are retained; no test filtering or class rebalancing is performed. The extraction manifests record source revisions, sample identities, split hashes, and content-overlap checks.

Encoders. We use DINOv2 (Oquab et al., 2024), CLIP (Radford et al., 2021), and released sentence encoders based on MiniLM (Wang et al., 2020) and MPNet (Song et al., 2020) with the Sentence-BERT framework (Reimers & Gurevych, 2019). The vision representations are the 384-dimensional final CLS token of `facebook/dinov2-small` and the 512-dimensional projected image embedding of `openai/clip-vit-base-patch32`, with their pretrained image processors. Text uses `sentence-transformers/all-MiniLM-L6-v2` (384 dimensions, maximum length 256 tokens) and `sentence-transformers/all-mpnet-base-v2` (768 dimensions, maximum length 384), with released pooling rules. The resolved model revisions begin with `ed25f3a3`, `3d74acf9`, `1110a243`, and `e8c3b32e`, respectively; manifests contain the full commit and weight-file hashes.

All encoders remain frozen in evaluation mode. Features are extracted in float32, then independently normalized to unit ℓ_2 norm. No mean, variance, whitening, or other population statistic is fitted. Precomputing these features consequently does not change which training examples are available to a readout at an arrival. The experiments concern classification on pretrained representations; they do not adapt the encoder or measure generative language-model training.

A.2 STREAMS, SELECTION, AND METRICS

For each setting, seeds $0, 1, 2$ independently determine a NumPy class permutation, which is divided by `array_split` into ten arrivals. A sample enters training only in its class's arrival. The same seed controls centroid initialization and, where applicable, the random feature map and replay sampling. All readouts know the complete output class set, receive no task identifier at prediction, and apply no task-specific mask. A constant bias coordinate is appended to every representation and receives the same regularization as other coefficients.

Validation protocol. Every method selects its configuration by final validation accuracy, breaking ties with raw validation NLL. The compact centroid study searches $\lambda \in \{10^{-3}, 10^{-2}, 10^{-1}, 1, 10\}$ and $J \in \{16, 32, 64\}$ updates per arrival. The correction coefficient is fixed at 0 or 1 for the two reported methods; it is not selected from test results. Ridge and all baseline penalty grids additionally include 100. Larger native-feature runs use the same penalty grid and separately select from $J \in \{16, 32, 64\}$ and $\{16, 32, 64, 128, 256\}$. The RP QUILT-C study separately selects from the original grid and $J \in \{16, 32, 64, 128, 256, 512\}$, again with the same five penalties. The large-budget native reference uses $m = 8192$ with the expanded grid; the RP performance reference uses $m = 8192$ with the expanded grid including 512. The new main native comparison instead uses $m = 512$ with $J \leq 256$, from the development cohort of the strict-correction study. These configurations were fixed before their respective test evaluations. All requested budgets and both iteration grids are reported, including cases where a smaller grid has higher test accuracy. The validation set contains all classes and is available to the evaluation procedure for selecting a configuration for the complete stream. No validation or test row contributes to a training moment, centroid, optimizer step, or replay buffer. Thus the results use standard offline model selection for a continual learner, rather than hyperparameter adaptation from test outcomes.

After model selection, each method fits one positive temperature τ (Guo et al., 2017) by minimizing validation NLL over $\log \tau \in [-4, 6]$ using bounded scalar optimization. Both the hyperparameters and temperature are fixed before computing test metrics. We retain raw and calibrated NLL separately. Temperature scaling leaves predicted classes unchanged; its inclusion gives squared-loss readouts the same opportunity to correct their logit scale as CE readouts.

Final accuracy is the proportion of correctly classified examples in the complete test set, not an unweighted average of arrival accuracies. NLL is in nats per example. Main-table entries report the mean and sample standard deviation across the three seeds. Aggregate gains first average seeds within a dataset–encoder setting and then give each of the 14 settings equal weight. Such aggregates summarize these benchmarks; they are not estimates over a sampled population of datasets. Ridge's final solution is order independent, so its seed variation is zero up to numerical precision.

A.3 CENTROID METHODS AND BASELINE IMPLEMENTATIONS

QUILT and its ablations. The streaming summary stores exact G, H and at most $m \in \{128, 512, 2048, 4096, 8192, 16384\}$ class-stratified centroids on native features, and $m \in \{2048, 8192\}$ on RP4096 features. The compact three-budget study also evaluates the uncorrected and centroid-metric variants. Weighted k-means uses 25 Lloyd iterations, a final barycenter update, and class allocations initially proportional to the square root of class mass, adjusted to respect the budget. After each arrival, the implementation checks the budget and the relative cross-moment error $\|H_{\mathcal{Q}} - H\|_F / \|H\|_F \leq 10^{-9}$. All moment arithmetic, readouts, and centroid optimization use float64. The measured number of nodes M is reported separately from the requested m. If m covers the fitting pool, every fitting feature can be retained and Δ vanishes up to floating-point error.

Each arrival restarts from ridge with penalty 1. The update matrix is $G/2 + \lambda I$, whose Cholesky factor is computed once and reused for the J updates. Readouts are centered across output classes after each step. Corrected QUILT adds the covariance-residual gradient $(G - G_{\mathcal{Q}})B/v$ for the centered readout; $\alpha = 0$ omits it. The centroid-metric ablation keeps $\alpha = 0$ and replaces the update matrix by $G_{\mathcal{Q}}/2 + \lambda I$. It retains the same ridge initialization and summary, changing the update metric while retaining the use of G for initialization. Each variant selects λ, J independently from the same grid, so the reported comparison is not a control with identical selected hyperparameters.

Centroid construction and ridge restarts do not depend on earlier readouts. The evaluator therefore builds each streaming summary once, searches final readouts on the last summary, then reconstructs

the selected trajectory from arrival checkpoints. A numerical check requires its final readout to match the selected candidate within 10^{-8}. Checkpoints and the feature archive belong to the evaluator; they never supply extra examples to a learner update. Historical-loss audits read training examples only after configuration selection and do not influence model choice.

The RP evaluator shares the exact moments and matrix factorizations between the two memory budgets and batches the five independent penalty paths. It evaluates the same update as Equation (10). A numerical comparison with the frozen unbatched implementation checks all 30 penalty/iteration checkpoints, with identical validation accuracies and NLL differences below 5×10^{-16} on the verification problem. The selected RP configurations are independently reconstructed at all ten arrivals for the continual metrics.

Ridge and random-feature ridge. Both accumulate a Gram matrix and target cross-moment and solve a ridge problem at each arrival. RP-ridge maps the d input features through a seeded $d \times 4096$ Gaussian matrix with variance $1/d$, applies ReLU and per-example ℓ_2 normalization, and appends a bias. This is a random-feature analytic baseline, not a reproduction of a complete representation-adaptation pipeline. Selected ridge configurations are independently refitted after the sweep. For every RP QUILT-C case, the projection is generated by the same CPU float64 random-number sequence as the corresponding archived RP-ridge run. Projection hashes agree in all 84 results. Recomputed RP-ridge validation and test accuracies agree exactly; NLL differences are below 3×10^{-14}. Thus both RP columns use the same map $\phi_{\mathrm{RP}}(x) = [\mathrm{normalize}(\mathrm{ReLU}(R^\top z(x))); 1]$, where $z(x)$ is the normalized native encoder output.

Experience replay. ER maintains a uniform Algorithm R reservoir of feature/label pairs. The cardinality study reports capacities $512, 4096, 8192, 16384$ under the same optimizer and validation protocol. For the new main comparisons, ER capacity is instead derived from the common byte cap, accounting for its optimizer moments as described below. It trains for ten epochs per arrival with new minibatches of at most 128 examples, each augmented by a uniformly sampled historical minibatch of at most equal size. Each unique new example enters the reservoir exactly once after the arrival's epochs. AdamW uses learning rates $\{10^{-3}, 10^{-2}, 10^{-1}\}$, $(\beta_1, \beta_2) = (0.9, 0.999)$, and $\epsilon = 10^{-8}$. The decoupled weight decay is λ/n_t, where n_t is the cumulative number of unique training examples. The 18 validation configurations use vectorized, independent optimizer states. After selection, a separate run with standard `torch.optim.AdamW` supplies the final predictions and timing; its coefficients are checked against the vectorized run. The maximum absolute coefficient difference is 2.49×10^{-14} in the original 42 trials and 6.04×10^{-14} in the 126 larger-buffer trials.

Offline joint CE. This native-feature reference retains the complete training history and fits a linear readout by full-batch L-BFGS with strong-Wolfe line search, history size 20, and at most 300 iterations per penalty. Validation selects the penalty and checkpoints spaced 25 iterations apart. The objective is summed CE plus $\lambda\|B\|_F^2/2$, internally divided by the training-set size for optimization. The saved gradient diagnostics quantify optimization progress; the selected checkpoint is not assumed to be a certified minimizer or a performance ceiling. Its representation differs from the RP configurations; it is not an offline reference in their expanded feature space.

A.4 STATE, TIMING, AND ARCHIVED EVIDENCE

State counts describe numerical storage at an arrival boundary, with $1\,\mathrm{MB} = 10^6$ bytes. QUILT stores G, H, B, centroid coordinates and masses, and integer labels; the covariance residual requires no extra persistent matrix. Ridge stores G, H, B. Both RP-ridge and RP QUILT-C additionally store the fixed projection matrix. ER includes the classifier, both Adam moments, reservoir features and labels, and counters. The offline reference includes all retained training features and labels; its L-BFGS workspace is reported separately. The shared encoder, feature cache used by the evaluator, validation/test arrays, checkpoint archive, temporary factorizations, and Python-object overhead are excluded. These are tensor-storage requirements, not peak device-memory measurements. Scalar calibration and control metadata do not affect the reported MB precision.

Timing runs use a single NVIDIA H20. For QUILT, summary construction and the selected readout updates are recorded separately; the former also includes summary checks and checkpoint copying. Its total evaluator time includes the validation search and historical/test evaluation and is not pre-

sented as selected-model training time. Baselines likewise record selected-run and complete-sweep times separately. The selected ER timing comes from its independent standard-optimizer refit. Encoder extraction is excluded from readout timing. For RP QUILT-C, shared projection, moment formation, factorization, summary construction, and complete validation-search times are stored separately. Shared operations are not counted as separately executed work for both budgets.

The core result field `head_flops` is per-example inference work, $2rv$. The baseline field with that name instead estimates total training arithmetic; these two fields are not directly compared. Training estimates use $2mnk$ for matrix multiplication, $r^3/3$ for Cholesky, and $2r^2v$ for the associated solves, with projection work included for RP-ridge. They exclude softmax, elementwise optimizer operations, and the encoder. FLOP counts are not latency measurements. The primary resource table computes inference work directly: $2(d+1)v$ for a native readout and $2d \cdot 4096 + 2 \cdot 4097v$ for the shared-map RP heads. The additional analytic families use their own widths, as detailed in Appendix B.

There are 126 distinct core trials (14 settings, three seeds, three budgets) and 42 baseline trials in the compact study. The expansion adds 126 requested native-budget results, 126 larger-buffer ER trials, and 84 RP QUILT-C results. Of the native requests, 90 are distinct effective runs; the other 36 reuse an identical saturated-budget run with the same feature cache, seed, and protocol. The reuse is explicitly recorded and is not an independent replicate. There are also 42 fixed-configuration RP QUILT-C retention reconstructions. Run JSONs contain selected configurations, validation grids, calibration parameters, test metrics, source/cache SHA256 hashes, and software metadata. Core trials also save task-performance matrices and historical-loss audits. Baseline selection decisions are persisted with a timestamp before test evaluation. An independent audit verifies the 42 baseline files, reproduces their selection and state arithmetic, and matches their ridge metrics to all 126 corresponding core trials. Duplicate file copies are not counted as replicates.

Readout runs use Python 3.12.11, PyTorch 2.7.1, and NumPy 1.26.0; baseline calibration uses SciPy 1.16.0. Feature manifests additionally pin encoder revisions, weights, processors, source-row splits, and tensor hashes. Complete source hashes and launch arguments accompany the archived results for the main extension runs and their baselines, and the extraction and evaluation entry points are documented alongside the released code. The additional large-budget, large-replay, and random-projection results are archived in the same way, with manifests and independent audits.

A.5 COMMON RETAINED-STATE BUDGETS

The direct comparison uses a single frozen list of 168 cases: 14 settings, three seeds, two feature maps (native and RP4096), and two reference budgets ($m = 512, 8192$). For feature dimension r including the bias, v output classes, and $M_{\max} = \min(m, n_{\text{train}})$, the allowed retained bytes are

$$S_m = S_\phi + 8\big[r^2 + 2rv + M_{\max}(r+2)\big] + 16. \tag{18}$$

Here $S_\phi = 0$ for native features and $S_\phi = 8d \cdot 4096$ for the fixed RP matrix applied to d native coordinates. The last 16 bytes are two int64 counters. All numerical learning tensors use float64. The budget depends only on dimensions, training count, and requested memory, not on accuracy.

Every method receives the same cap for the same setting, seed, and feature map. Actual usage can be lower because of integer allocations, small fitting pools, or unused capacity. No tensor padding is introduced to make byte counts equal. The field named `reference_actual_m` in the original cap file is the capacity bound $M_{\max}$; measured centroid cardinalities are reported separately. The new tables standardize counters to 16 bytes; some older result files omitted counters or counted only one. This small accounting difference does not change any prediction.

The encoder and evaluator's feature caches are common inputs. Validation/test arrays, archived trajectories, the current input batch, and temporary optimizer workspaces are excluded from retained state. All matrices or buffers needed at an arrival boundary are included. When the current readout also serves as the next teacher, the shared tensor is counted once; a trainable copy used during an update is temporary workspace. These are retained-state constraints, not claims about equal peak GPU allocation.

A.6 DIRECT FUNCTIONAL-MEMORY BASELINES

K-priors. Our implementation follows the anchored prior of Khan & Swaroop (2021), Equations 8 and 13, with class-balanced random real-input memory as in the Compact Memory authors' multi-class experiment. Let B_- be the previous readout, u_j a retained feature, and $p_j^- = \mathrm{softmax}(B_-^\top u_j)$. At each arrival it minimizes

$$\sum_{(\phi_i,y_i)\in\mathcal{D}_t} \ell(B^\top\phi_i, y_i) + \sum_{j\in\mathcal{M}_{t-1}} a_j \ell(B^\top u_j, p_j^-) + \frac{\delta}{2}\|B - B_-\|_F^2, \tag{19}$$

where soft-label CE is $\ell(\theta, p) = -p^\top \log\mathrm{softmax}(\theta)$. The first arrival uses the ordinary zero-centered weight penalty. The unit-weight version has $a_j = 1$, without an implicit N/M multiplier. It retains no Gram matrix or target cross-moment.

A seeded class-balanced policy divides total capacity among observed classes, randomly reduces earlier groups when necessary, and samples the new group. Discarded rows are never reread. This changes the original growing-memory experiment to a fixed total budget while preserving the prior objective. The implementation stores one readout, retained features and integer labels, class counts, two global counters, and any RP matrix. Its capacity is

$$K = \left\lfloor \frac{S_m - S_\phi - 8rv - 8v - 16}{8(r+1)} \right\rfloor,$$

with actual occupancy limited by available examples. Teacher probabilities are recomputed temporarily from the aliased readout, not retained as an uncounted second memory.

Each of six penalties $\delta \in \{.001, .01, .1, 1, 10, 100\}$ runs the entire stream independently. Full-batch L-BFGS uses strong-Wolfe line search, history size 20, at most 600 iterations per arrival, and mean-objective gradient tolerance 10^{-9}. After validation selects the penalty, an independent refit checks its coefficients. A separate sufficiency study retains that penalty and repeats the entire stream with a 2000-iteration cap. Validation chooses between the two trajectories before joining their test results. Gradient norms and cap exits are recorded; a finite cap is not described as certified convergence.

Class-mass importance weighting. At the native S_{512} budget we also run all six penalties with weights $a_j = N_c/K_c$ for a retained point from class c. Here N_c is its historical count and K_c its resident count. This preserves each historical class's mass using the existing labels and counters, with no additional persistent state. The full stream uses L-BFGS with a 2000-iteration cap. Class-wise sums of the weights equal the historical counts; resident examples and storage are checked against the unit-weight version. For a fixed teacher and candidate, this is a stratified importance estimator. Because sequentially fitted teachers depend on past memories, we do not claim unconditional unbiasedness of the adaptive training gradient. The native main-table K-priors$^+$ entry selects among unit600, unit2000, and importance-weighted2000 using validation only. All three results remain in the appendix.

Compact Memory. We independently implement Jung et al. (2025)'s weighted teacher prior and PPCA memory updates, with numerical comparisons against the authors' public implementation. The prior has the form of Equation (19), using learned unit vectors and nonnegative weights. Memory fitting matches the sum of the class-diagonal Hessian blocks, using $\sum_c p_c(1-p_c)$. The published per-entry probability clipping is 10^{-4}; clipped rows are not renormalized.

For cap S_m, CM may retain $K_{\max} = \lfloor (S_m - S_\phi - 8rv - 16)/(8(r+1)) \rfloor$ vectors. It follows the authors' growing-memory construction within this fixed cap: $K_t = \min(K_{\max}, \lceil tK_{\max}/10 \rceil)$. Each arrival retains earlier vectors and initializes additional directions from its current features. If more directions than current examples are needed, distinct Gaussian combinations of those features initialize the remainder. No future input or duplicated-vector padding is used. The count is not truncated at feature rank or sample size. Updating memory after the last arrival prepares the retained state for another arrival; it does not alter that arrival's fitted readout.

CM uses 100 full Adam epochs per arrival, learning rate 0.1, batch size 1024, standard Adam moments, and ten PPCA EM iterations with noise 10^{-4} and denominator jitter 10^{-6}. Adam is restarted at each arrival. The authors' code divides the new-example CE by 1024 even for a short batch, and divides the prior by the current arrival's size. We also evaluate the mean-loss variant using actual

batch size for the new data. All six penalties are evaluated under both normalizations; validation selects among all 12 configurations. Each normalization's selected model is retained separately.

PPCA is evaluated with algebraically equivalent latent- or feature-space solves, including the original jitter, without truncating the learned memory. The feature-space form avoids an unnecessarily large latent inverse when $K > r$. Independent tests compare the update and prior gradients with the authors' functions. Additional GPU tests cover low rank, the introduction of new feature directions, and repeated EM steps. These numerical checks support the implementation; they are not accuracy results.

The native S_{512} comparison additionally tests optimization sufficiency. It fixes the original validation-selected penalty and normalization and replaces the Adam head update with full-batch L-BFGS, at most 2000 iterations, history size 50, and gradient tolerance 10^{-9}. The growing memory and ten EM steps are unchanged. For the mean-loss variant the summed objective is Equation (19). For the fixed-1024 variant, its expected per-epoch objective multiplies the new-example sum by $\kappa = N_t/(1024\lceil N_t/1024\rceil)$, while leaving the prior unchanged. Subtracting the teacher entropy replaces soft CE by KL plus a constant and improves line-search numerics without changing the gradient. Validation selects the head variant before its test evaluation; both results and all optimization diagnostics are preserved. Of 420 head-control arrivals, 120 miss the nominal 10^{-9} stopping flag; the largest final mean-objective gradient infinity norm is 2.95×10^{-8}. We report these residuals rather than treating a finite iteration cap as a convergence certificate.

ER under the same byte caps. ER uses the original 18-configuration AdamW grid and ten epochs per arrival. Its persistent classifier, two Adam moment tensors, reservoir features/labels, projection, and counters give capacity

$$K_{\mathrm{ER}} = \min\left(n_{\mathrm{train}}, \left\lfloor \frac{S_m - S_\phi - 24rv - 16}{8(r+1)} \right\rfloor\right).$$

A standard-AdamW refit checks the selected vectorized candidate. Three RP/CIFAR cases exceeded the coefficient-equality tolerance before test access. For those cases, all 18 validation candidates were independently refitted with standard AdamW and selection repeated on that grid. The selected model then passed an additional refit without relaxing the tolerance. Both grids and the recovery reason are archived.

A.7 Paired replications and strictly unsaturated interventions

The RP replication fixes ten new seeds, 3–12, all 14 settings, $m = 8192$, $\alpha = 1$, the RP4096 construction, and the original selection grids before computing its results. Each seed determines the projection and class order for both methods. Every seed produces one equal-setting mean difference. The primary estimate averages these ten blocks. We report a two-sided 95% Student t interval, a paired t test, a 20,000-draw seed-block percentile bootstrap with analysis seed 20260911, and an exact two-sided sign test. The sign test concerns sign balance, not a distribution-free test of the mean. Setting-wise tests additionally report Holm adjustments.

The correction study has 364 cases: 280 new-seed cases across the two feature maps, and 84 development cases at seeds $0, 1, 2$. Every case has exactly 512 nodes, fewer than its fitting examples. Both arms share all retained tensors, initialization, class order, and actual bytes; only α changes. The primary selector uses the uncorrected model's validation accuracy and raw NLL to choose λ, J. A prespecified reverse selector uses the corrected model instead. Independently tuned comparisons and the legacy $J \leq 64$ grid are also retained. These contrasts are kept distinct, including the unfavorable primary RP result.

The intervals quantify randomness in the algorithm on these fixed benchmark splits. They do not treat repeated test rows, different encoders of one dataset, or the 14 settings as independent draws from a population of unseen datasets. Three-seed development tables and ten-seed replication intervals are never pooled without being identified.

B ANALYTIC CONTINUAL-LEARNING COMPARISONS

The shared-feature ridge control does not cover all mechanisms in analytic continual learning. We add ACIL/GACL, DS-AL, and AIR under both main retained-byte ceilings, using all 14 settings and seeds $0, 1, 2$. The resulting 84 cases contain 252 selected models. The comparison uses the original frozen encoder caches, class orders, training/validation/test splits, and ten disjoint-class arrivals. Every output class remains eligible at prediction time, including classes not yet observed in the stream. These are adaptations of the analytic heads to this common protocol. The scope of each comparison is explicit below.

Table 5: Coverage of analytic continual learning. Equality of the ridge objective does not imply equality of complete representation-learning pipelines.

Method	Distinct component	Coverage in this paper
ACIL/GACL	Exact recursive ridge; GACL also supports recurring classes	Measured with a validation-selected buffer width. The two heads coincide on our disjoint-class streams; they are not counted as separate competitors.
RanPAC	Pretrained features, random expansion, analytic classifier	The RP-ridge head is measured with exactly the same map as RP-QUILT. The complete first-session adaptation pipeline is not reproduced.
DS-AL	Main ReLU stream, tanh compensation, previous-label cleansing	Both streams are measured, with independently selected penalties and a selected compensation ratio, including zero.
AIR	Class-balanced analytic objective	Measured using its exact specialization to completed, disjoint classes. Reappearing-class AIR requires a different retained state.
GKEAL	Gaussian kernel embedding for few-shot CIL	Discussed as a distinct kernel method; its embedding and few-shot protocol are not evaluated here.
F-OAL	Feature fusion and forward-only online RLS	The ridge objective is covered; the fused representation and online arrival protocol are not reproduced.
AnaCP	Analytic contrastive projection	Discussed as a representation-changing approach; its learned projection is not evaluated.

ACIL's exactness follows from the joint ridge normal equations (Zhuang et al., 2022). GACL's exposed/unexposed-class decomposition generalizes the recursion to recurring classes; when classes never recur, its extra exposed-class contribution vanishes (Zhuang et al., 2024a). This equality is conditional on an identical feature map, penalty and label encoding. It is not a claim that every analytic method reduces to our original RP-ridge implementation. In particular, the DS-AL compensation targets and AIR reweighting change the fitted objective. GKEAL, F-OAL, and AnaCP also change the representation or evaluation protocol (Zhuang et al., 2023; 2024c; Momeni et al., 2025). Our performance claims are restricted to the measured heads.

B.1 FEATURE EXPANSIONS, SELECTION, AND INDEPENDENT REFITS

The added methods follow the authors' shared implementation of the no-bias `RandomBuffer`. For native feature dimension d, its $q \times d$ matrix has independent entries uniform on $[-1/\sqrt{d}, 1/\sqrt{d}]$. The input features retain the benchmark's per-example normalization. The main activation is ReLU; DS-AL uses tanh for its compensation stream. There is no extra output-feature normalization or appended head bias. This differs from the Gaussian, normalized RP4096 map in the strictly shared-feature experiment; the two comparisons answer different questions.

Each method selects from the feasible widths in $\{128, 256, 512, 1024, 2048, 4096, 8192\}$, together with the largest feasible multiple of 64, capped at 8192. Feasibility uses only dimensions and the allowed byte ceiling. A single seeded random matrix supplies nested prefixes across widths. The selected tensor is copied at its actual width, so unselected projection columns are not retained. The width grid was fixed before these experiments.

The ridge penalties are $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 100\}$. DS-AL independently searches both penalties and the compensation ratio $\{0, 0.1, 0.2, 0.4, 0.6, 0.8, 1, 1.2, 1.4, 1.6, 1.8, 2\}$. Final validation accuracy selects configurations, breaking ties by raw validation NLL. Each selected model is then independently refit through all ten arrivals, and its coefficients and validation predictions are checked against the sweep result. One positive temperature is fitted on validation with log-temperature bounds $[-4, 6]$. The configuration and calibration are saved before evaluating any test prediction. Four native-cap DS-AL fits reach the lower temperature bound; all other added analytic fits have interior temperatures. Their NLLs reflect this stated, bounded calibration procedure. The original data splits and test pools are never reselected.

B.2 Algorithms and retained state

Write Z_t for an arrival's expanded features and Y_t for its one-hot labels. An ordinary analytic learner with Gram A and exact ridge readout W can recover its cross-moment as $(A+\gamma I)W$. Consequently the update

$$A^+ = A + Z_t^\top Z_t, \qquad W^+ = (A^+ + \gamma I)^{-1}\big[(A + \gamma I)W + Z_t^\top Y_t\big] \tag{20}$$

needs only A and W between arrivals. We implement the solve by Cholesky factorization. It is algebraically equivalent to recursive least squares, without retaining a separate cross-moment. Independent full-moment accumulators verify its normal equations but are not inputs to the learner. They are audit workspace.

DS-AL. Both activations use the same frozen projection. After the main stream's arrival update, the compensation target is

$$R_t = (Y_t - Z_{\mathrm{M},t} W_{\mathrm{M},t}) D_t, \tag{21}$$

where D_t is diagonal and keeps only the current arrival's class columns. This is the published previous-label-cleansing rule. The compensation stream fits these targets using its own Gram, readout, and penalty; old residual targets are not regenerated using a later main readout. The prediction is $Z_{\mathrm{M}} W_{\mathrm{M}} + c_{\mathrm{comp}} Z_{\mathrm{C}} W_{\mathrm{C}}$. We use one analytic update per complete arrival, corresponding to the phase-level formulation in DS-AL, Equations 13–18 (Zhuang et al., 2024b). During the sweep, compensation solves can be postponed because their previous readouts do not enter future targets. The independent selected refit updates both streams at every arrival.

AIR under disjoint-class arrivals. Let n_c be the observed count for class c and let v_t classes have arrived. The authors' class weight is s_t/n_c, where $s_t = v_t/(\sum_{c:n_c>0} 1/n_c)$. Define

$$\bar{A}_t = \sum_{c:n_c>0} \frac{1}{n_c} \sum_{i:y_i=c} z_i z_i^\top, \qquad \bar{H}_t = \sum_{c:n_c>0} \frac{1}{n_c} \sum_{i:y_i=c} z_i e_c^\top. \tag{22}$$

AIR's exact optimum is $(\bar{A}_t + (\gamma/s_t)I)^{-1}\bar{H}_t$ (Fang et al., 2024). When every class is completed in one arrival, its count never changes afterward. Its normalized moments can therefore be added once to the aggregate. The selected learner retains $\bar{A}_t$, W_t, and class counts, reconstructing $\bar{H}_t$ from its previous normal equation before the next update. This gives the same solution as the authors' generalized implementation while avoiding per-class Gram storage in this specific protocol. The implementation rejects class reappearance; this storage reduction is not claimed for generalized AIR streams.

Bytes and computational scope. All learning tensors use float64. At width q, ACIL/GACL retains $8(dq + q^2 + qv) + 16$ bytes; AIR additionally retains $8v$ bytes of counts. DS-AL retains $8(dq + 2q^2 + 2qv) + 16$ bytes. These totals include the random projection, current readout(s), Gram(s), and two int64 counters. Factorizations, current-arrival tensors, independent audit oracles and evaluator snapshots are temporary, separately identified storage. The saved artifact contains earlier readouts for evaluation; they never enter a later update. Timings of the selected refit include the numerical audit and are not used to rank method speed. Different feature widths also have different inference costs, so equal retained bytes do not imply equal inference compute. The dominant dense inference arithmetic is $2dq + 2qv$ for ACIL/GACL or AIR, and $2dq + 4qv$ for DS-AL, excluding their common encoder and elementwise activations.

B.3 VERIFICATION AND COMPLETE RESULTS

The implementation was checked against the authors' public recursive-linear and AIR modules, pinned to commit `0afd6bf87622e808ce7f730f8556e78229b9e758`. Small independent tests compare batch ridge, recursive ridge with recurring classes, class-balanced ridge, and the two DS-AL streams. The maximum absolute difference is 7.21×10^{-11}. The source files, license, equations inspected, and hashes are archived. For every benchmark case, an independent audit reconstructs validation selection, verifies all retained bytes, and regenerates logits and every arrival prediction from the saved projection and readouts. All 84 audits pass; reconstructed logits agree exactly with the saved logits. The aggregate has 972,300 checks. These checks establish implementation and artifact consistency, not statistical independence of configurations.

Table 6: Native-cap analytic comparison: final accuracy (%) over seeds 0, 1, 2. Each method uses the common frozen encoder and retained-byte ceiling, with its own stated expansion. QUILT-0 and QUILT-C have independently validation-selected penalties and iteration counts.

Dataset	Encoder	Ridge	ACIL/GACL	DS-AL	AIR	QUILT-0	QUILT-C
CIFAR-100	DINOv2	81.44 ±0.00	79.65 ±0.15	80.81 ±0.28	79.65 ±0.15	83.49 ±0.17	**83.66** ±0.14
	CLIP	73.80 ±0.00	71.16 ±0.37	72.92 ±0.16	71.16 ±0.37	76.78 ±0.04	**77.04** ±0.24
Pets	DINOv2	93.79 ±0.00	93.36 ±0.39	93.05 ±0.41	93.39 ±0.36	94.12 ±0.11	**94.57** ±0.16
	CLIP	87.74 ±0.00	84.85 ±0.52	87.46 ±0.19	84.72 ±0.44	88.37 ±0.26	**89.51** ±0.45
Flowers-102	DINOv2	99.35 ±0.00	99.20 ±0.02	99.13 ±0.32	99.20 ±0.02	99.54 ±0.02	**99.57** ±0.01
	CLIP	92.37 ±0.00	90.50 ±0.37	91.69 ±0.27	90.50 ±0.37	94.29 ±0.15	**94.66** ±0.19
DTD	DINOv2	75.90 ±0.00	74.57 ±0.61	76.03 ±0.55	74.57 ±0.61	77.68 ±0.03	**78.51** ±0.38
	CLIP	70.80 ±0.00	68.62 ±0.19	70.37 ±0.61	68.62 ±0.19	73.32 ±0.17	**74.33** ±0.48
EuroSAT	DINOv2	94.74 ±0.00	93.11 ±0.25	93.62 ±0.27	93.06 ±0.29	**95.38** ±0.14	95.20 ±0.01
	CLIP	94.54 ±0.00	93.04 ±0.25	92.99 ±0.20	93.12 ±0.17	**95.60** ±0.10	95.09 ±0.09
Banking77	MiniLM	88.31 ±0.00	88.68 ±0.30	88.26 ±0.34	89.39 ±0.27	90.58 ±0.06	**91.67** ±0.12
	MPNet	92.01 ±0.00	92.32 ±0.21	91.75 ±0.03	92.58 ±0.32	92.79 ±0.16	**93.28** ±0.25
20NG	MiniLM	**67.72** ±0.00	64.58 ±0.31	67.07 ±0.12	64.64 ±0.23	66.63 ±0.09	67.26 ±0.36
	MPNet	**71.57** ±0.00	69.57 ±0.11	70.78 ±0.24	69.38 ±0.21	70.33 ±0.32	71.41 ±0.21

Table 7: RP-cap analytic comparison: final accuracy (%) over seeds 0, 1, 2. Each method uses the common frozen encoder and retained-byte ceiling, with its own stated expansion. QUILT-0 and QUILT-C have independently validation-selected penalties and iteration counts.

Dataset	Encoder	Ridge	ACIL/GACL	DS-AL	AIR	QUILT-0	QUILT-C
CIFAR-100	DINOv2	85.39 ±0.07	**86.14** ±0.07	85.71 ±0.16	**86.14** ±0.07	86.01 ±0.14	86.02 ±0.12
	CLIP	77.72 ±0.22	78.30 ±0.07	77.75 ±0.02	78.30 ±0.07	79.35 ±0.17	**79.55** ±0.22
Pets	DINOv2	94.37 ±0.31	94.61 ±0.22	94.59 ±0.18	94.61 ±0.22	**94.66** ±0.07	**94.66** ±0.07
	CLIP	88.53 ±0.53	88.23 ±0.34	88.11 ±0.69	88.21 ±0.33	**88.63** ±0.07	**88.63** ±0.07
Flowers-102	DINOv2	99.54 ±0.00	99.50 ±0.16	**99.57** ±0.05	99.50 ±0.16	99.56 ±0.00	99.56 ±0.00
	CLIP	93.97 ±0.03	94.02 ±0.14	93.74 ±0.52	94.02 ±0.14	**94.39** ±0.12	**94.39** ±0.12
DTD	DINOv2	78.44 ±0.25	78.33 ±0.36	78.24 ±0.27	78.33 ±0.36	**78.62** ±0.61	**78.62** ±0.61
	CLIP	73.14 ±0.51	73.12 ±0.21	72.64 ±0.25	73.12 ±0.21	**74.15** ±0.52	**74.15** ±0.52
EuroSAT	DINOv2	96.20 ±0.11	96.52 ±0.12	96.40 ±0.07	96.51 ±0.13	**96.64** ±0.17	96.46 ±0.16
	CLIP	95.62 ±0.09	95.50 ±0.22	95.78 ±0.21	95.86 ±0.08	**96.19** ±0.09	95.88 ±0.18
Banking77	MiniLM	93.26 ±0.22	93.31 ±0.37	93.33 ±0.38	**93.61** ±0.18	93.21 ±0.19	93.21 ±0.14
	MPNet	94.09 ±0.18	**94.22** ±0.12	94.06 ±0.26	94.21 ±0.10	93.84 ±0.22	93.86 ±0.17
20NG	MiniLM	68.58 ±0.52	**69.76** ±0.05	68.82 ±0.99	69.51 ±0.03	68.14 ±0.90	68.39 ±0.42
	MPNet	72.19 ±0.56	72.81 ±0.93	72.22 ±0.34	**73.03** ±0.25	71.58 ±0.14	71.51 ±0.18

Table 8: Native-cap allocations. Each method cell gives its selected width range across three seeds, followed by mean actual retained MiB. The cap includes the projection, learner matrices and counters. Unused capacity is the cap minus actual usage; no padding tensors are added.

Dataset	Encoder	Cap MiB	ACIL/GACL	DS-AL	AIR
CIFAR-100	DINOv2	3.23	448 / 3.19	320 / 2.99	448 / 3.19
	CLIP	4.80	512 / 4.39	384 / 4.34	512 / 4.39
Pets	DINOv2	2.86	384 / 2.36	256–320 / 2.42	384 / 2.36
	CLIP	4.31	512 / 4.14	384 / 3.97	512 / 4.14
Flowers-102	DINOv2	3.24	448 / 3.19	320 / 3.00	448 / 3.19
	CLIP	4.82	512 / 4.40	384 / 4.35	512 / 4.40
DTD	DINOv2	2.92	384 / 2.39	320 / 2.73	384 / 2.39
	CLIP	4.39	512 / 4.18	384 / 4.03	512 / 4.18
EuroSAT	DINOv2	2.70	384 / 2.28	320 / 2.55	384 / 2.28
	CLIP	4.10	512 / 4.04	384 / 3.81	512 / 4.04
Banking77	MiniLM	3.09	384 / 2.48	320 / 2.88	384 / 2.48
	MPNet	8.43	704 / 8.32	512 / 7.60	704 / 8.32
20NG	MiniLM	2.76	384 / 2.31	320 / 2.60	384 / 2.31
	MPNet	7.76	512–640 / 6.34	512 / 7.16	640 / 6.97

Table 9: RP-cap allocations. Each method cell gives its selected width range across three seeds, followed by mean actual retained MiB. The cap includes the projection, learner matrices and counters. Unused capacity is the cap minus actual usage; no padding tensors are added.

Dataset	Encoder	Cap MiB	ACIL/GACL	DS-AL	AIR
CIFAR-100	DINOv2	402.50	6976 / 397.04	4992 / 402.49	6976 / 397.04
	CLIP	406.50	6976 / 403.85	4928 / 397.33	6976 / 403.85
Pets	DINOv2	234.44	2048–5312 / 103.17	2048–3776 / 124.36	2048–5312 / 103.17
	CLIP	238.44	5312 / 237.53	3776 / 234.44	5312 / 237.53
Flowers-102	DINOv2	178.34	1024–4544 / 109.79	3264 / 177.21	1024–4544 / 109.79
	CLIP	182.34	4096–4544 / 168.27	3264 / 180.39	4096–4544 / 168.27
DTD	DINOv2	201.79	4928 / 201.49	3456 / 194.85	4928 / 201.49
	CLIP	205.79	4864 / 201.24	3520 / 205.34	4864 / 201.24
EuroSAT	DINOv2	396.88	6976 / 392.25	4096–4992 / 353.30	6976 / 392.25
	CLIP	400.88	4096–6976 / 314.15	4992 / 400.51	6976 / 399.06
Banking77	MiniLM	401.06	4096–6976 / 311.35	4096–4992 / 358.10	6976 / 395.82
	MPNet	413.06	6912 / 409.06	4928 / 405.23	6912 / 409.06
20NG	MiniLM	397.50	6976 / 392.78	4096–4992 / 354.02	6976 / 392.78
	MPNet	409.50	6912 / 406.05	4096–4928 / 361.04	6912 / 406.05

Table 10: Native-cap predictive loss: raw / temperature-calibrated NLL in nats, averaged over the three seeds. One scalar temperature is selected on validation separately for each method.

Dataset	Encoder	ACIL/GACL	DS-AL	AIR	QUILT-C
CIFAR-100	DINOv2	4.238/0.809	3.973/0.734	4.238/0.809	0.620/0.567
	CLIP	4.280/1.182	3.985/1.072	4.280/1.182	0.895/0.837
Pets	DINOv2	2.995/0.262	3.048/0.242	2.995/0.261	0.277/0.184
	CLIP	3.204/0.500	3.559/1.322	3.178/0.517	0.578/0.311
Flowers-102	DINOv2	3.989/0.054	3.979/0.084	3.989/0.054	0.139/0.038
	CLIP	4.259/0.359	4.116/0.322	4.259/0.359	0.571/0.195
DTD	DINOv2	3.442/1.004	3.303/0.918	3.442/1.004	0.834/0.762
	CLIP	3.502/1.232	3.363/1.138	3.502/1.232	0.990/0.920
EuroSAT	DINOv2	1.688/0.221	1.443/0.199	1.687/0.218	0.256/0.145
	CLIP	1.667/0.221	1.650/0.214	1.668/0.216	0.238/0.146
Banking77	MiniLM	3.840/0.475	3.823/0.476	3.834/0.456	0.377/0.322
	MPNet	3.729/0.352	3.433/0.358	3.725/0.347	0.309/0.268
20NG	MiniLM	2.659/1.296	2.436/1.195	2.665/1.282	1.115/1.116
	MPNet	2.602/1.144	2.491/1.081	2.596/1.141	1.026/1.019

Table 11: RP-cap predictive loss: raw / temperature-calibrated NLL in nats, averaged over the three seeds. One scalar temperature is selected on validation separately for each method.

Dataset	Encoder	ACIL/GACL	DS-AL	AIR	QUILT-C
CIFAR-100	DINOv2	3.966/0.608	3.975/0.617	3.966/0.608	0.644/0.530
	CLIP	4.072/0.952	4.069/0.954	4.072/0.952	0.756/0.720
Pets	DINOv2	2.867/0.232	2.844/0.225	2.867/0.231	0.290/0.193
	CLIP	3.035/0.434	3.032/0.434	3.035/0.434	0.441/0.363
Flowers-102	DINOv2	3.960/0.044	3.801/0.041	3.960/0.044	0.092/0.038
	CLIP	4.054/0.238	4.066/0.241	4.054/0.238	0.604/0.205
DTD	DINOv2	3.303/0.903	3.269/0.903	3.303/0.903	0.936/0.862
	CLIP	3.400/1.067	3.348/1.106	3.400/1.067	1.049/0.997
EuroSAT	DINOv2	1.579/0.112	1.563/0.115	1.580/0.112	0.180/0.110
	CLIP	1.600/0.140	1.573/0.133	1.588/0.132	0.159/0.119
Banking77	MiniLM	3.567/0.341	3.548/0.331	3.567/0.327	0.300/0.289
	MPNet	3.531/0.304	3.534/0.304	3.531/0.301	0.263/0.248
20NG	MiniLM	2.557/1.144	2.536/1.172	2.560/1.142	1.131/1.131
	MPNet	2.523/1.050	2.487/1.059	2.537/1.028	1.028/1.032

Table 12: Native-cap analytic retention. AIA pools all seen-class test examples at each arrival; all output classes remain eligible. Signed forgetting uses each earlier task's best accuracy before the last arrival minus its final accuracy. Entries are means and sample standard deviations of three setting-averaged seed blocks.

Group	Method	AIA (%)	Forgetting (pp)	Worst task (%)
Vision 10	ACIL/GACL	90.09 $\pm$0.96	5.72 $\pm$0.12	74.99 $\pm$0.83
Vision 10	DS-AL	90.19 $\pm$1.12	6.13 $\pm$0.30	75.50 $\pm$0.88
Vision 10	AIR	90.08 $\pm$1.00	5.72 $\pm$0.12	75.08 $\pm$0.80
Text 4	ACIL/GACL	85.18 $\pm$0.60	7.34 $\pm$1.15	64.95 $\pm$2.53
Text 4	DS-AL	85.58 $\pm$0.47	8.23 $\pm$1.01	64.17 $\pm$4.29
Text 4	AIR	85.23 $\pm$0.83	7.00 $\pm$0.68	67.10 $\pm$3.15
All 14	ACIL/GACL	88.68 $\pm$0.86	6.18 $\pm$0.41	72.12 $\pm$0.55
All 14	DS-AL	88.88 $\pm$0.94	6.73 $\pm$0.49	72.27 $\pm$0.75
All 14	AIR	88.70 $\pm$0.95	6.08 $\pm$0.28	72.80 $\pm$0.60

Table 13: RP-cap analytic retention. AIA pools all seen-class test examples at each arrival; all output classes remain eligible. Signed forgetting uses each earlier task's best accuracy before the last arrival minus its final accuracy. Entries are means and sample standard deviations of three setting-averaged seed blocks.

Group	Method	AIA (%)	Forgetting (pp)	Worst task (%)
Vision 10	ACIL/GACL	92.12 ± 0.84	4.15 ± 0.23	80.48 ± 0.63
Vision 10	DS-AL	91.99 ± 0.99	4.76 ± 0.23	79.96 ± 0.49
Vision 10	AIR	92.14 ± 0.84	4.14 ± 0.24	80.58 ± 0.66
Text 4	ACIL/GACL	87.57 ± 0.59	5.69 ± 1.01	70.83 ± 2.87
Text 4	DS-AL	87.29 ± 0.67	6.86 ± 0.36	70.04 ± 3.35
Text 4	AIR	87.63 ± 0.56	5.60 ± 0.59	72.17 ± 2.94
All 14	ACIL/GACL	90.82 ± 0.76	4.59 ± 0.34	77.72 ± 0.69
All 14	DS-AL	90.65 ± 0.85	5.36 ± 0.22	77.13 ± 0.99
All 14	AIR	90.85 ± 0.75	4.56 ± 0.26	78.18 ± 0.78

B.4 The uncorrected learner at the larger RP budget

We add the missing $m = 8192$ QUILT-0 ablation for all 14 settings and development seeds $0, 1, 2$. Its projection and streaming centroid construction exactly match the original QUILT-C configuration. The same validation grid searches $\lambda \in \{0.001, 0.01, 0.1, 1, 10\}$ and $J \in \{16, 32, 64, 128, 256, 512\}$, with α fixed to zero. The summary is rebuilt, and the archived corrected readout is independently reproduced before the new test evaluation. A second, single-path solver checks the selected uncorrected readout. Actual retained bytes are equal between the two variants, including two standardized counters. These are independently tuned method results, not the strict fixed-parameter intervention at $m = 512$. They are additional development ablations and do not extend the ten-seed confirmatory cohort.

C Correction heterogeneity and optional selection

The positive native seed-block interval summarizes the average over the fixed benchmark. It does not show that correction reliably helps a new setting. At fixed λ, J, native setting means span -0.535 to $+1.085$ points: 12 are positive and two negative. Across the 140 native cases, 106 improve, six tie, and 28 decline. RP has eight positive, one tied and five negative setting means; its 140 cases contain 64 improvements, 14 ties and 62 declines. Both EuroSAT RP settings lose accuracy in all ten seeds. These counts describe paired outcomes; they are not treated as independent replicates. Near-zero effects are not interpreted as meaningful practical gains.

A practical option is to retain $\alpha = 0$ as a candidate and select whether to enable correction on validation. We examine that rule retrospectively using the already saved $m = 512$ candidates. Within each case, we compare the two independently selected (λ, J) configurations by validation accuracy, then raw validation NLL, and prefer $\alpha = 0$ on an exact tie. The chosen model and its previously validation-fitted temperature supply the test result. This uses no test score in the selection calculation, but the decision to study this rule was made after the original experiments; it is exploratory. It must not replace the prespecified intervention, and it is not an independent replication or a guarantee against test regression. Any gain combines correction with model selection, including changes in penalty and iteration count. No new training runs are counted for this analysis.

Table 14: Retrospective validation choice between independently tuned QUILT-0 and QUILT-C at $m = 512$, using seeds 3–12. The gain is relative to independently tuned QUILT-0 and includes changes in λ, J; it is not the isolated correction effect. Uncertainty shown is sample standard deviation across ten seed-block means. Case counts are descriptive, not independent replicates.

Features	$\alpha = 1$ selected	Gain vs. QUILT-0	Setting means $+/0/-$	Case losses
Native	105/140	0.52 ± 0.07	13 / 1 / 0	2/140
RP4096	92/140	0.13 ± 0.06	11 / 2 / 1	24/140

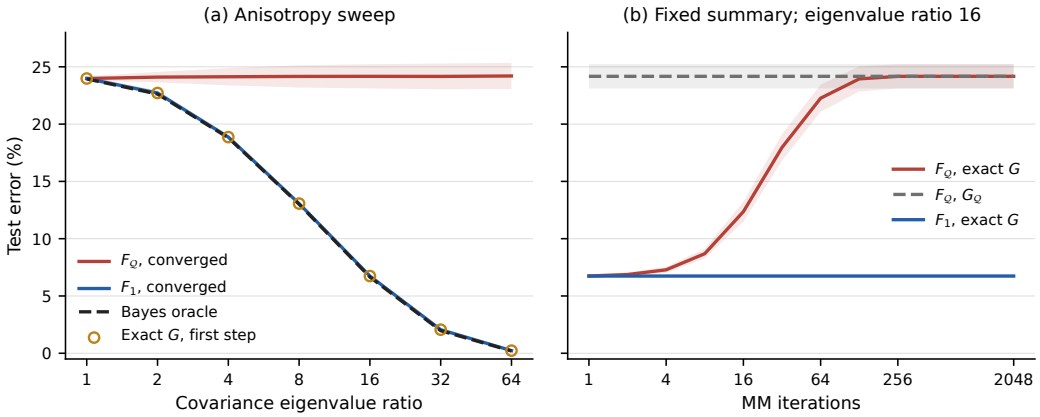


Figure 3: Covariance in the update and in the objective. Balanced Gaussian classes, ten seeds, 1024 training and 40,000 independent test examples per seed. Left: covariance trace is fixed while anisotropy varies. Right: at covariance ratio 16, continued optimization of the uncorrected centroid objective loses the first step's favorable geometry. The corrected objective retains it. Shading is one sample standard deviation. This experiment uses the raw Gaussian features and zero initialization assumed by Theorem 2.

D A CONDITIONAL GAUSSIAN MECHANISM

The following experiment illustrates the assumptions of Theorem 2. Its large two-class effect is distinct from the small and heterogeneous component effects on normalized encoder features.

D.1 A CONTROLLED TEST OF COVARIANCE GEOMETRY

We use two-dimensional classes with means $\pm(0.5, 0.5)$ and covariance $\Sigma = \mathrm{diag}(2/(1+k), 2k/(1+k))$, retaining one centroid per class. All centering and moments use training data; $\bar{\lambda} = 0.02$ is fixed. At $k = 16$, the converged uncorrected objective has $24.17 \pm 1.06\%$ test error, compared with $6.74 \pm 0.15\%$ for both the exact-Gram first step and the corrected optimum. The Bayes oracle gives $6.65 \pm 0.10\%$. Figure 3 shows that the uncorrected path approaches its poorer limiting classifier despite decreasing its training objective. The corrected boundary is close to Bayes, but its uncalibrated cross-entropy remains higher (0.292 versus 0.166); the theorem concerns classification geometry. In the population Gaussian model, as the ridge penalty vanishes, a scalar temperature recovers the Bayes posterior (Appendix E.7). This is a model-specific explanation of direction and confidence.

E PROOFS AND SUPPLEMENTARY ANALYSIS

E.1 WEIGHTED NOTATION AND NORMALIZATION

Fix one observed history and one feature map. Assign each example a mass $a_i > 0$; ordinary empirical data have $a_i = 1$. Write $n = \sum_i a_i$, $G = \sum_i a_i \phi_i \phi_i^\top$, and $H = \sum_i a_i \phi_i e_{y_i}^\top$. A class-pure partition $\{C_j\}$ defines

$$w_j = \sum_{i \in C_j} a_i, \qquad \psi_j = \frac{1}{w_j} \sum_{i \in C_j} a_i \phi_i, \qquad c_j = y_i \quad (i \in C_j).$$

All identities below use exact arithmetic. The ridge coefficient λ multiplies the regularizer of the loss *sum*; a mean-loss coefficient $\bar{\lambda}$ corresponds to $\lambda = n\bar{\lambda}$. The Gaussian theorem uses the mean convention explicitly. The implementation uses float64 for moments and optimization; encoder features are cached in float32.

E.2 Centroid identities and repeated merging

For each cell, $\sum_{i \in C_j} a_i(\phi_i - \psi_j) = 0$. Expanding outer products gives

$$\sum_{i \in C_j} a_i \phi_i \phi_i^\top = w_j \psi_j \psi_j^\top + \sum_{i \in C_j} a_i (\phi_i - \psi_j)(\phi_i - \psi_j)^\top,$$
$$\sum_{i \in C_j} a_i \phi_i e_{y_i}^\top = w_j \psi_j e_{c_j}^\top.$$

Summing proves Equation (5), including $H_{\mathcal{Q}} = H$ and $\Delta \succeq 0$. A repeated weighted merge preserves the hypothesis. If node k is the exact centroid of original cell C_k, merging a same-class collection J produces

$$\psi' = \frac{\sum_{k \in J} w_k \psi_k}{\sum_{k \in J} w_k} = \frac{\sum_{i \in \cup_{k \in J} C_k} a_i \phi_i}{\sum_{i \in \cup_{k \in J} C_k} a_i}.$$

New examples enter as singleton cells, so induction covers the entire stream. The proof uses cells inherited through the merge tree, which need not be nearest-node cells in the final point set. The learner need not store that tree. Renormalizing a centroid after averaging would generally invalidate the identities, which is why only the original encoder features are normalized.

E.3 Exact discarded loss and the local remainder

Let $f(\theta) = \log \sum_c e^{\theta_c}$, $\theta_i = B^\top \phi_i$, $\theta_j^{\mathcal{Q}} = B^\top \psi_j$, and $d_i = \theta_i - \theta_j^{\mathcal{Q}}$. The label term is linear and cancels when summed within a cell. So does $(p_j^{\mathcal{Q}})^\top d_i$. Therefore

$$L(B) - L_{\mathcal{Q}}(B) = \sum_j \sum_{i \in C_j} a_i \left[f(\theta_i) - f(\theta_j^{\mathcal{Q}}) - (p_j^{\mathcal{Q}})^\top d_i \right] \tag{23}$$
$$= \sum_j \sum_{i \in C_j} a_i \, \mathrm{KL}(p_j^{\mathcal{Q}} \| p_i). \tag{24}$$

The last equality follows by substituting $\log p_c = \theta_c - f(\theta)$. The order of the KL arguments is thus fixed by the Bregman divergence.

For any $u \in \mathbb{R}^v$ and probability vector p, write $J(p) = \mathrm{diag}(p) - pp^\top$. Its variance representation is

$$u^\top J(p) u = \min_b \sum_c p_c (u_c - b)^2.$$

Using $p_{\min} \le p_c \le p_{\max}$ gives

$$p_{\min} \|\Pi_v u\|^2 \le u^\top J(p) u \le p_{\max} \|\Pi_v u\|^2. \tag{25}$$

For the upper bound, evaluate the minimum at the arithmetic mean of u; for the lower bound, minimize the unweighted sum of squared deviations. If $\mathrm{osc}(\theta) = \max_c \theta_c - \min_c \theta_c \le \rho$, then $p_{\min} \ge e^{-\rho}/v$ and $p_{\max} \le e^{\rho}/v$. The oscillation is convex, so the same range bound holds for every segment from a centroid logit to an original logit. Taylor's integral remainder therefore implies

$$\frac{e^{-\rho}}{2v} \|\Pi_v d_i\|^2 \le f(\theta_i) - f(\theta_j^{\mathcal{Q}}) - (p_j^{\mathcal{Q}})^\top d_i \le \frac{e^{\rho}}{2v} \|\Pi_v d_i\|^2.$$

Sum and use the residual-covariance identity to obtain Equation (13). The absolute-error form follows because $1 - e^{-\rho} \le e^{\rho} - 1$.

At $B = 0$, the original and centroid losses both equal $n \log v$. Their gradients agree by the first-moment identity. Their Hessians differ by $\Pi_v \otimes \Delta / v$, exactly the Hessian of C_Δ. For bounded features,

$$\mathrm{osc}(B^\top \phi_i) \le \sqrt{2} \|B \Pi_v\|_2 \|\phi_i\|_2 \le \sqrt{2} R \|B \Pi_v\|_2.$$

Since $C_\Delta(B) = O(\|B\|^2)$, the error after correction is $O(\|B\|^3)$ near zero. For normalized encoder features with a unit bias coordinate, $R \le \sqrt{2}$, so the computable envelope $\rho = 2\|B\Pi_v\|_2$ is valid. It need not be numerically useful at a trained readout.

For completeness, Böhning's bound $J(p) \preceq \Pi_v/2$ follows from

$$\mathrm{Var}_p(u_c) \le \frac{(\max_c u_c - \min_c u_c)^2}{4} \le \frac{1}{2}\|\Pi_v u\|^2.$$

It gives the global upper endpoint $L - L_{\mathcal{Q}} \le \mathrm{tr}(B_c^\top \Delta B_c)/4 = (v/2)C_\Delta$, while convexity gives the lower endpoint zero. This completes Proposition 1.

A local Fisher remainder. A refinement can be used when the original cell residuals are available for analysis. Set $s_i = \mathrm{osc}(d_i)$ and $t_i = d_i^\top J(p_j^{\mathcal{Q}})d_i$. Then

$$h_-(s_i)t_i \le \mathrm{KL}(p_j^{\mathcal{Q}}\|p_i) \le h_+(s_i)t_i, \qquad h_\pm(s) = \frac{e^{\pm s} \mp s - 1}{s^2}, \quad h_\pm(0) = \tfrac{1}{2}. \tag{26}$$

Along $\theta + td$, the second derivative of f is the variance of the categorical values d_c, and its derivative is their third central moment. The absolute third central moment is at most $\mathrm{osc}(d)$ times that variance. Integrating the resulting differential inequality bounds the second derivative between $e^{-s_i t}t_i$ and $e^{s_i t}t_i$. Integrating once more against $1 - t$ proves Equation (26). The implemented correction uses the uniform Fisher matrix Π_v/v; it does not store a covariance matrix per centroid or assume that trained probabilities remain uniform.

E.4 Corrected MM and convergence

The covariance term has gradient $\Delta B \Pi_v/v$. For $0 \le \alpha \le v/2$,

$$\begin{aligned}\nabla^2 F_\alpha(B) &\preceq I_v \otimes (\tfrac{1}{2}G_{\mathcal{Q}} + \lambda I) + \tfrac{\alpha}{v}\Pi_v \otimes \Delta \\ &\preceq I_v \otimes (\tfrac{1}{2}G + \lambda I) = I_v \otimes A.\end{aligned}$$

The last step uses $\Pi_v \preceq I_v$ and $\Delta \succeq 0$. For $g = g_\alpha(B)$, the tangent upper quadratic is

$$F_\alpha(B + D) \le F_\alpha(B) + \langle g, D\rangle + \tfrac{1}{2}\,\mathrm{tr}(D^\top A D).$$

Its minimizer is $D = -A^{-1}g$, giving Equation (11). The objective is coercive and λ-strongly convex, so its minimizer exists and is unique. Let $E_k = F_\alpha(B_k) - \min F_\alpha$. Strong convexity gives $\|g_\alpha(B_k)\|_F^2 \ge 2\lambda E_k$, hence

$$E_{k+1} \le \left(1 - \frac{\lambda}{\lambda_{\max}(A)}\right) E_k.$$

The iterates converge by strong convexity. This is an iteration-count statement on a fixed summary. The bound can have an unfavorable condition number; it is not a claim of faster optimization than a stochastic method.

For $\alpha = 0$, replacing the centroid curvature $G_{\mathcal{Q}}/2$ by $G/2$ adds $\mathrm{tr}(D^\top \Delta D)/4$ to the surrogate. This changes the penalty on each increment, but leaves the stationary equation $\widehat{P}(B) - H + \lambda B = 0$ unchanged. Adding $C_\Delta(B)$ to the objective changes that stationary equation. This is the distinction tested by the matched-parameter experiments.

E.5 Proof of the Gaussian mechanism

Use the mean-loss convention of Theorem 2. For equal class priors and population centroids $\pm\mu$,

$$\mathcal{F}_0(a) = \log(1 + e^{-a^\top \mu}) + \tfrac{\bar\lambda}{4}\|a\|^2, \qquad \mathcal{F}_1(a) = \mathcal{F}_0(a) + \tfrac{1}{8}a^\top \Sigma a.$$

The normalized original and centroid Gram matrices are $G = \Sigma + \mu\mu^\top$ and $G_{\mathcal{Q}} = \mu\mu^\top$. Write $s(t) = 1/(1 + e^t)$. The uncorrected stationary equation is

$$-s(a^\top \mu)\mu + \tfrac{\bar\lambda}{2}a = 0.$$

It has a unique solution by strong convexity, and the solution is a positive multiple of μ. For the scalar-margin parameter a, the matrix-readout MM step is $a^+ = a - 2A^{-1}\nabla_a \mathcal{F}_0(a)$, where $A = G/2 + \bar\lambda I$. At zero it gives

$$a_G^{(1)} = 2(\Sigma + \mu\mu^\top + 2\bar\lambda I)^{-1}\mu = \frac{2}{1 + \mu^\top w}w, \qquad w = (\Sigma + 2\bar\lambda I)^{-1}\mu.$$

The second equality is Sherman–Morrison. Using $G_{\mathcal{Q}}$ instead gives a positive multiple of μ.

The corrected stationary equation is

$$-s(a^\top\mu)\mu + \tfrac{1}{4}\Sigma a + \tfrac{\bar{\lambda}}{2}a = 0, \qquad a = 4s(a^\top\mu)w.$$

Again strong convexity gives a unique minimizer, and the scalar factor is positive. Its classifier therefore agrees with the exact-Gram first step.

For any a with $a^\top\mu > 0$, Gaussian test error is

$$\Pr(Ya^\top X < 0) = \Phi_{\mathrm{N}}\left(-\frac{a^\top\mu}{\sqrt{a^\top\Sigma a}}\right),$$

where Φ_{N} is the standard normal CDF. Cauchy–Schwarz in the Σ inner product bounds the squared signal-to-noise ratio by $\mu^\top\Sigma^{-1}\mu$, with equality exactly when $a \parallel \Sigma^{-1}\mu$. As $\bar{\lambda} \downarrow 0$, $w \to \Sigma^{-1}\mu$. The centroid direction μ attains the same error only when $\Sigma\mu$ is proportional to μ. This proves the classification statement. The Bayes log-odds have coefficient $2\Sigma^{-1}\mu$; the corrected coefficient generally has a different scale. The theorem therefore does not establish Bayes probability calibration.

The algebra for equal first-step and corrected-limit directions also holds for balanced finite samples after centering at the midpoint of the two training class means, using their empirical residual covariance. The Gaussian assumption supplies the population error formula and Bayes comparison, rather than the matrix identity itself.

E.6 Finite-sample risk and the effect of sample size

We prove Corollary 3 for a balanced training sample, with $n/2$ independent observations from each Gaussian class. Let

$$\widehat{\mu} = \frac{\bar{X}_+ - \bar{X}_-}{2}, \qquad \widehat{c} = \frac{\bar{X}_+ + \bar{X}_-}{2}, \qquad \widehat{\Sigma} = \frac{1}{n}\sum_{y=\pm1}\sum_{i:Y_i=y}(X_i - \bar{X}_y)(X_i - \bar{X}_y)^\top.$$

The covariance divisor is n, although its Wishart degrees of freedom are $n-2$. After centering at $\widehat{c}$, the normalized moments are $G/n = \widehat{\Sigma} + \widehat{\mu}\widehat{\mu}^\top$ and $\Delta/n = \widehat{\Sigma}$. For the fixed sum penalty λ_0, the corrected mean objective is

$$\widehat{\mathcal{F}}_1(a) = \log(1 + e^{-a^\top\widehat{\mu}}) + \frac{1}{8}a^\top\widehat{\Sigma}a + \frac{\lambda_0}{4n}\|a\|^2.$$

Its stationary equation gives the direction in Equation (15). There is no freely optimized intercept here: the decision threshold is obtained from the training midpoint.

Deterministic perturbation. Set $\zeta = \Sigma^{-1/2}\mu$, $\gamma = \|\zeta\| > 0$, and

$$E = \Sigma^{-1/2}\widehat{\Sigma}\Sigma^{-1/2} - I, \quad e_\mu = \Sigma^{-1/2}(\widehat{\mu} - \mu), \quad b = \Sigma^{-1/2}\widehat{c}, \quad \beta = \frac{2\lambda_0}{n\lambda_{\min}(\Sigma)}.$$

Suppose $\|E\|_2 \le \epsilon \le 1/2$ and $\|e_\mu\|, \|b\| \le \eta$. The whitened classifier vector is

$$z = \Sigma^{1/2}\widehat{w} = \left(I + E + \frac{2\lambda_0}{n}\Sigma^{-1}\right)^{-1}(\zeta + e_\mu).$$

The ridge term is positive semidefinite, so

$$q := \|z - \zeta\| \le \frac{\eta + (\epsilon+\beta)\gamma}{1-\epsilon} \le 2[\eta + (\epsilon+\beta)\gamma]. \tag{27}$$

Almost surely $z \neq 0$. Write $u = z/\|z\|$, $s = u^\top\zeta$, and $t = u^\top b$. The exact conditional population risk is

$$\mathcal{R}(\widehat{h}_C) = \tfrac{1}{2}\Phi_{\mathrm{N}}(-s+t) + \tfrac{1}{2}\Phi_{\mathrm{N}}(-s-t), \qquad \mathcal{R}^\star = \Phi_{\mathrm{N}}(-\gamma).$$

The normalization inequality $\|z/\|z\| - \zeta/\gamma\| \le 2q/\gamma$ does not require $q < \gamma$. Since both normalized vectors have unit norm,

$$\gamma - s = \frac{\gamma}{2}\|u - \zeta/\gamma\|^2 \le \frac{2q^2}{\gamma}.$$

Also $\sup_x |\Phi'_{\mathrm{N}}(x)| = 1/\sqrt{2\pi}$ and $\sup_x |\Phi''_{\mathrm{N}}(x)| = 1/\sqrt{2\pi e}$. The equal-prior average cancels the term linear in t, giving

$$\mathcal{R}(\widehat{h}_C) - \mathcal{R}^\star \le \frac{2q^2}{\gamma\sqrt{2\pi}} + \frac{\eta^2}{2\sqrt{2\pi e}} \tag{28}$$

$$\le \frac{8[\eta + (\epsilon+\beta)\gamma]^2}{\gamma\sqrt{2\pi}} + \frac{\eta^2}{2\sqrt{2\pi e}}. \tag{29}$$

A weaker first-order bound, $[5\eta + 4(\epsilon+\beta)\gamma]/\sqrt{2\pi}$, also follows from Lipschitz continuity and can be preferable for small γ. The second-order bound is the source of the fixed-signal fast rate.

Gaussian concentration. The whitened mean and midpoint errors each have distribution $\mathcal{N}(0, I_r/n)$. The pooled within-class covariance satisfies

$$\Sigma^{-1/2}\widehat{\Sigma}\Sigma^{-1/2} = \frac{1}{n}Z^\top Z, \qquad Z \in \mathbb{R}^{(n-2)\times r}, \quad Z_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0,1).$$

Standard Gaussian norm and singular-value concentration (Vershynin, 2018), with

$$t_0 = \sqrt{2\log(4/\delta)}, \qquad \eta = \frac{\sqrt{r}+t_0}{\sqrt{n}}, \qquad u_0 = \frac{\sqrt{r}+t_0}{\sqrt{n-2}}, \qquad \epsilon = \frac{2}{n} + 2u_0 + u_0^2,$$

bound both mean-error events and both singular-value events jointly with probability at least $1-\delta$. For $u_0 < 1$, the covariance eigenvalues lie between $\frac{n-2}{n}(1-u_0)^2$ and $\frac{n-2}{n}(1+u_0)^2$. The $2/n$ term retains the divisor-n bias explicitly.

To obtain absolute constants, set $h = r + \log(4/\delta)$ and $s_0 = \sqrt{h/n}$. If $n \ge 256h$, then $\eta \le 2s_0$ and $\epsilon < 6s_0 \le 3/8$. Furthermore,

$$\frac{(\eta+\epsilon\gamma)^2}{\gamma} \le (4/\gamma + 24 + 36\gamma)s_0^2 \le 48(\gamma+\gamma^{-1})s_0^2.$$

Apply $(x+y)^2 \le 2x^2 + 2y^2$ in Equation (29). The coefficient of $(\gamma+\gamma^{-1})h/n$ is less than 307, and that of $\lambda_0^2\gamma/[n^2\lambda_{\min}(\Sigma)^2]$ is less than 26. Thus $c_0 = 256$ and $c_1 = 320$ are sufficient in Equation (16). These conservative constants are not intended as tight numerical intervals at the benchmark sample sizes.

The corrected classifier consequently converges in classification risk to Bayes at fixed centroid count. The uncorrected classifier has direction $\widehat{\mu}$ and midpoint $\widehat{c}$, so consistency of empirical means instead gives

$$\mathcal{R}_{Q,\infty} = \Phi_{\mathrm{N}}\left(-\frac{\|\mu\|^2}{\sqrt{\mu^\top\Sigma\mu}}\right) > \Phi_{\mathrm{N}}(-\gamma)$$

unless μ lies in one eigenspace of Σ. This proves the claimed statistical distinction. It does not imply log-loss consistency of the corrected objective.

The stronger MM bound. On the centered readout subspace, the Hessian lower bound is $I_v \otimes (\lambda I + \alpha\Delta/v)$. Its smallest eigenvalue is at least $\lambda + \alpha\lambda_{\min}(\Delta)/v$. Combining this with the descent bound and strong convexity proves Equation (17). When Δ/n is bounded below by a positive-definite matrix and $\|G\|_2/n$ is bounded above, the corrected contraction parameter remains bounded away from zero for fixed $\alpha > 0, v, \lambda$. An appended constant feature gives a zero row and column in Δ. Thus the positive-definite residual condition does not apply to the full parameter space of the normalized encoder experiments.

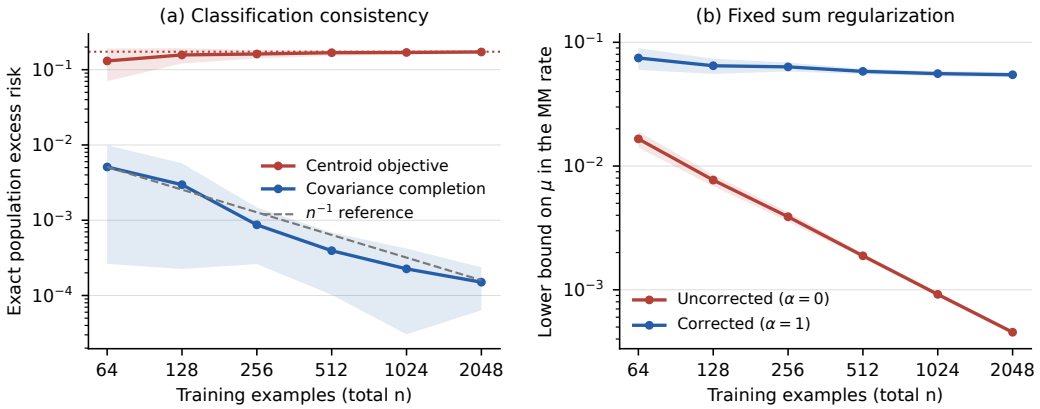


Figure 4: Sample-size study under the Gaussian model, ten training seeds, fixed sum penalty $\lambda_0 = 1$, and covariance ratio 16. Left: exact conditional Gaussian population excess risk of the fitted classifiers, evaluated using the known generating distribution. Right: the conditional MM contraction-parameter lower bounds. The corrected excess decreases with sample size while the centroid-only classifier approaches a positive floor. These are model-based risk and curvature calculations, separate from the pretrained-encoder experiments.

The study uses $n \in \{64, 128, 256, 512, 1024, 2048\}$ with the same μ and covariance ratio as Figure 3. Training data alone determine the fitted classifier; 40,000 independent test samples per seed are also retained. The corrected mean population excess decreases from 0.005099 to 0.0001503, with descriptive log–log slope -1.077. The centroid-only excess approaches its theoretical floor 0.173279 (0.171902 at $n = 2048$). The plotted population risks use the analytic Gaussian formula to avoid test-sampling noise at small excess values.

E.7 What scalar calibration can recover in the Gaussian model

Let $Y \in \{-1, +1\}$ have equal class probabilities and $X \mid Y = y \sim \mathcal{N}(y\mu, \Sigma)$, where $\mu \neq 0$ and $\Sigma \succ 0$. There is no intercept, and the population class centroids are $\pm\mu$. Write $\sigma(t) = (1 + e^{-t})^{-1}$ and $\gamma^2 = \mu^\top \Sigma^{-1} \mu$. For a positive temperature τ, the readout $B = (-a/2, a/2)$ predicts $q_{a,\tau}(x) = \sigma(a^\top x/\tau)$. The Bayes probability is $p_\star(x) = \sigma(2\mu^\top \Sigma^{-1} x)$. Define population binary cross-entropy in nats by

$$\mathcal{L}(q) = \mathbb{E}\big[-\mathbf{1}_{\{Y=+1\}} \log q(X) - \mathbf{1}_{\{Y=-1\}} \log(1 - q(X))\big],$$

and let $\mathcal{R}(h) = \Pr(h(X) \neq Y)$.

Corollary 4 (Direction determines what temperature can recover)**.** *At zero mean-loss ridge penalty, the corrected centroid objective*

$$\mathcal{F}_{1,0}(a) = \log(1 + e^{-a^\top \mu}) + \tfrac{1}{8} a^\top \Sigma a$$

has the unique minimizer

$$a_{\mathrm{C}} = c_\gamma \Sigma^{-1} \mu, \qquad c_\gamma = 4\sigma(-c_\gamma \gamma^2), \qquad 0 < c_\gamma < 2.$$

The temperature $\tau_\star = c_\gamma/2$ gives $q_{a_{\mathrm{C}},\tau_\star} = p_\star$ exactly, and hence $\mathcal{L}(q_{a_{\mathrm{C}},\tau_\star}) = \mathcal{L}(p_\star)$.

In contrast, fix any uncorrected centroid direction $a_0 = t\mu$, $t > 0$, as obtained by minimizing the uncorrected population objective with a positive isotropic penalty. Its best positive temperature produces the coefficient

$$\frac{a_0}{\tau_0} = \frac{2\|\mu\|^2}{\mu^\top \Sigma \mu}\, \mu, \qquad \tau_0 = \frac{t\, \mu^\top \Sigma \mu}{2\|\mu\|^2}.$$

If $\Sigma\mu$ is not proportional to μ, no temperature recovers $p_\star$. More quantitatively, with $\rho_\mu = \|\mu\|^2/\sqrt{\mu^\top \Sigma \mu}$,

$$\inf_{\tau > 0} \big\{\mathcal{L}(q_{a_0,\tau}) - \mathcal{L}(p_\star)\big\} \geq \frac{1}{2}\big[\Phi_{\mathrm{N}}(-\rho_\mu) - \Phi_{\mathrm{N}}(-\gamma)\big]^2 > 0. \tag{30}$$

If $\Sigma\mu$ is proportional to μ, the best calibrated mean direction also recovers $p_\star$.

Proof. Since $\Sigma \succ 0$, $\mathcal{F}_{1,0}$ is strongly convex and coercive. Its stationary equation is

$$-\sigma(-a^\top \mu)\mu + \tfrac{1}{4}\Sigma a = 0.$$

Consequently $a = c\Sigma^{-1}\mu$ and $c = 4\sigma(-c\gamma^2)$. The function $c - 4\sigma(-c\gamma^2)$ is strictly increasing, is negative at zero, and is positive at two because $\gamma > 0$. This proves existence and uniqueness of $c_\gamma \in (0, 2)$. The stated temperature gives $a_{\mathrm{C}}/\tau_\star = 2\Sigma^{-1}\mu$, the Gaussian Bayes coefficient.

For the mean direction, put $Z = \mu^\top X$. Conditionally on $Y = y$, Z is Gaussian with mean $y\|\mu\|^2$ and variance $\mu^\top\Sigma\mu$. Thus

$$\Pr(Y = +1 \mid Z = z) = \sigma\left(\frac{2\|\mu\|^2}{\mu^\top\Sigma\mu} z\right).$$

Binary cross-entropy is strictly proper. This conditional probability minimizes risk over all functions of Z, and it belongs to the positive temperature family, proving the formula for τ_0. Because the Gaussian mixture has a positive density on all of $\mathbb{R}^r$ and σ is injective, equality with $p_\star$ almost surely requires $\mu \parallel \Sigma^{-1}\mu$, equivalently $\Sigma\mu \parallel \mu$.

For completeness, cross-entropy excess has the exact representation

$$\mathcal{L}(q) - \mathcal{L}(p_\star) = \mathbb{E}\,\mathrm{KL}\big(\mathrm{Ber}(p_\star(X))\|\mathrm{Ber}(q(X))\big).$$

On the event that the classifier induced by q disagrees with the Bayes classifier, $|2p_\star - 1| \le 2|p_\star - q|$. Pinsker's inequality and Cauchy–Schwarz therefore give

$$\mathcal{R}(h_q) - \mathcal{R}^\star \le 2\mathbb{E}|p_\star(X) - q(X)| \le \sqrt{2\{\mathcal{L}(q) - \mathcal{L}(p_\star)\}}.$$

Every positive temperature preserves $h_\mu(x) = \mathrm{sign}(\mu^\top x)$. Its error is $\Phi_{\mathrm{N}}(-\rho_\mu)$, whereas $\mathcal{R}^\star = \Phi_{\mathrm{N}}(-\gamma)$. Cauchy–Schwarz gives $\rho_\mu \le \gamma$, with equality precisely when $\Sigma\mu \parallel \mu$. Squaring the preceding risk inequality proves Equation (30), including strictness. □

A consistency consequence with independent calibration data. Fix the dimension and Gaussian distribution above. From balanced training samples of total size n, let $\widehat{\mu}_n$ be half the difference of the two class means, $\widehat{b}_n$ their midpoint, and $\widehat{\Sigma}_n$ the pooled within-class residual covariance with divisor n. Let a_n minimize

$$\log(1 + e^{-a^\top \widehat{\mu}_n}) + \tfrac{1}{8}a^\top\widehat{\Sigma}_n a + \frac{\lambda_0}{4n}\|a\|^2, \qquad \lambda_0 > 0 \text{ fixed.}$$

Let $\widehat{\tau}_{n,k}$ minimize empirical cross-entropy on k independent validation examples over a fixed compact interval $[\tau_{\min}, \tau_{\max}] \subset (0, \infty)$ containing $\tau_\star$, using logits $a_n^\top(X - \widehat{b}_n)/\tau$. Write $q_{n,k}(x) = \sigma(a_n^\top(x - \widehat{b}_n)/\widehat{\tau}_{n,k})$. As $n, k \to \infty$,

$$a_n \to a_{\mathrm{C}}, \qquad \widehat{\tau}_{n,k} \to \tau_\star, \qquad \mathcal{L}(q_{n,k}) \to \mathcal{L}(p_\star) \quad \text{in probability,}$$

and the induced classification risk tends to $\mathcal{R}^\star$. Indeed, the training moments converge and $w_n = (\widehat{\Sigma}_n + 2\lambda_0 I/n)^{-1}\widehat{\mu}_n$ converges to $\Sigma^{-1}\mu$. The stationary equation $a_n = c_n w_n$ has the continuous unique root $c_n = 4\sigma(-c_n\widehat{\mu}_n^\top w_n)$, proving readout convergence; also $\widehat{b}_n \to 0$. On bounded sets of readouts and offsets and the stated temperature interval, logistic loss is bounded by an integrable multiple of $1 + \|X\|$ and has an integrable Lipschitz envelope. A uniform law of large numbers and continuity of population risk therefore reduce validation selection to its limiting population problem. That problem has the unique minimizer $\tau_\star$ by strict propriety and the nondegenerate Gaussian logit. The argmin theorem gives temperature and log-loss convergence; the excess-risk inequality above gives classification consistency. No finite-sample rate is claimed.

Scope. The exact posterior identity concerns the population limit with $\bar{\lambda} = 0$ and a well-specified balanced Gaussian mixture. The existing anisotropy experiment instead fixes $\bar{\lambda} = 0.02$; its finite regularization can alter the direction, so the identity is not an assertion about that experiment's calibrated output. This corollary explains why direction and logit scale must be assessed separately for the corrected and uncorrected centroid objectives. It provides neither a distribution-free calibration guarantee nor a separation from ridge regression.

E.8 A SMALLER FIXED MAJORANT

Proposition 5 (A smaller majorizing matrix for the corrected objective)**.** *Fix a class-pure centroid summary, with $\Delta = G - G_{\mathcal{Q}} \succeq 0$, $v \geq 2$, and $\lambda > 0$. For the same corrected objective F_1 used in Equation (8), define*

$$A_{\text{tight}} = \tfrac{1}{2} G_{\mathcal{Q}} + \tfrac{1}{v}\Delta + \lambda I_r = \tfrac{1}{v} G + (\tfrac{1}{2} - \tfrac{1}{v}) G_{\mathcal{Q}} + \lambda I_r. \tag{31}$$

Then $\nabla^2 F_1(B) \preceq I_v \otimes A_{\text{tight}}$ for every B, and $0 \prec A_{\text{tight}} \preceq A = G/2 + \lambda I_r$. For $g = \nabla F_1(B)$ and $D_{\text{tight}} = -A_{\text{tight}}^{-1} g$,

$$F_1(B) - F_1(B + D_{\text{tight}}) \geq \tfrac{1}{2} \operatorname{tr}(g^\top A_{\text{tight}}^{-1} g) \geq \tfrac{1}{2} \operatorname{tr}(g^\top A^{-1} g). \tag{32}$$

Thus the certified quadratic decrement is at least as large at the same readout and for the same gradient.

Proof. The centroid loss has Hessian bounded by $I_v \otimes G_{\mathcal{Q}}/2$. The covariance term has exact Hessian $\Pi_v \otimes \Delta/v$. Since $\Pi_v \preceq I_v$ and $\Delta \succeq 0$,

$$\nabla^2 F_1(B) \preceq I_v \otimes (\tfrac{1}{2} G_{\mathcal{Q}} + \lambda I_r) + \tfrac{1}{v} \Pi_v \otimes \Delta \preceq I_v \otimes A_{\text{tight}}.$$

The matrix is positive definite because $\lambda > 0$, and

$$A - A_{\text{tight}} = (\tfrac{1}{2} - \tfrac{1}{v})\Delta \succeq 0.$$

Minimizing its tangent upper quadratic gives the first inequality in Equation (32). Inversion reverses the order of positive-definite matrices, so $A_{\text{tight}}^{-1} \succeq A^{-1}$, proving the second inequality. No commutativity assumption on $G_{\mathcal{Q}}$ and Δ is needed. □

This is an algebraic refinement for a fixed summary. The two metrics coincide when $v = 2$ or $\Delta = 0$, and their decrement certificates can also agree for particular gradients. The proposition compares guaranteed lower bounds; it makes no ordering claim for the actual objectives attained by the two candidate steps. It establishes neither lower wall-clock time nor higher historical or test accuracy. All reported results use A, with the same matrix in their paired correction comparisons. The smaller matrix is an unmeasured optimization option for the same F_1, not a change to those experiments.

E.9 POLYNOMIAL MOMENTS DO NOT DETERMINE THE ANCHOR TERM

Proposition 6 (Moment insufficiency)**.** *For every finite degree k, two finitely supported probability measures on $[-1, 1]$ can agree on all moments of degree at most k while giving different values of $P(B)$ for a common binary linear readout B. The examples can be chosen to give equal-sized empirical histories.*

Proof. Increase k to at least two if necessary, which only strengthens the claim. For scalar features and $B = (1, 0)$, one entry of $P(B)$ integrates $h(x) = x/(1 + e^{-x})$. This analytic function is not a polynomial on $[-1, 1]$. Choose $k + 1$ distinct rational nodes and interpolate h there with a polynomial p of degree at most k. Since $h \neq p$, there is another rational node where they differ. For these $k + 2$ nodes x_i, the vector $(h(x_i))_i$ is not in the row space of the degree-k Vandermonde matrix V. There is consequently a vector $a \in \ker V$ with $\sum_i a_i h(x_i) \neq 0$. The nodes are rational and the nullspace is one-dimensional, so a may be chosen rational.

The degree-zero constraint gives $\sum_i a_i = 0$. Let $a_i^+ = \max(a_i, 0)$ and $a_i^- = \max(-a_i, 0)$, with common positive mass $M = \sum_i a_i^+ = \sum_i a_i^-$. Then $\mu_+ = M^{-1} \sum_i a_i^+ \delta_{x_i}$ and $\mu_- = M^{-1} \sum_i a_i^- \delta_{x_i}$ have equal moments through degree k but different integrals of h. Scaling rational masses to integers realizes them as equal-sized empirical histories. Assigning the same label to every example also matches H, since the first feature moment agrees. Matching the second moment matches G. □

This proposition does not say that a centroid memory is the only possible solution. It rules out exact recovery from any fixed finite polynomial moment collection. Other functional summaries, and approximation with explicit distributional assumptions, remain possible.

E.10 HISTORICAL LOSS, PREDICTION CHANGES, AND SCOPE

Let $q(B) = (v/2)C_\Delta(B)$. For any readout, the exact centroid identities give

$$F_{\mathcal{Q}}(B) \le F(B) \le F_{\mathcal{Q}}(B) + q(B).$$

Thus the sufficient condition

$$F_{\mathcal{Q}}(B^+) + q(B^+) \le F_{\mathcal{Q}}(B)$$

guarantees historical descent for any candidate update, including one obtained from F_α. It is not a necessary condition. The benchmark updates are not filtered by this test.

Descent on an approximate objective alone is insufficient. For scalar features $x = (-3, -1, 1, 3)$, labels $(0, 0, 1, 1)$, no intercept, and $\lambda = 0.2$, compress each class to a centroid -2 or 2 of mass two. With $B = (-a/2, a/2)$,

$$F(a) = 2\log(1 + e^{-3a}) + 2\log(1 + e^{-a}) + 0.05a^2, \qquad F_{\mathcal{Q}}(a) = 4\log(1 + e^{-2a}) + 0.05a^2.$$

At the original minimizer $a_\star \simeq 2.159967$, an uncorrected exact-Gram MM step gives $a^+ \simeq 2.138205$. It lowers $F_{\mathcal{Q}}$ from 0.28612520 to 0.28378333 while raising F from 0.45463717 to 0.45471168. The historical sufficient test correctly rejects the step.

The retained Gram also gives an exact empirical logit-change identity: for an unweighted history with feature matrix Φ, $\|\Phi D\|_F^2 = \operatorname{tr}(D^\top G D)$. Since the cross-entropy gradient in logits has norm at most $\sqrt{2}$,

$$\frac{|L(B + D) - L(B)|}{n} \le \sqrt{\frac{2}{n} \operatorname{tr}((D\Pi_v)^\top G (D\Pi_v))}.$$

It may be loose at useful update sizes. Neither this bound nor the corrected-MM result establishes per-task retention, test-risk monotonicity, or stability of data-dependent clustering. The loss-sum intervals and Gaussian theorem have the explicit scopes given above.

F ADDITIONAL EXPERIMENTAL RESULTS

F.1 COMPLETE DIRECT COMPARISONS

The following tables retain all 14 settings, both feature maps, and both byte budgets. Text settings are included even when they favor a baseline. Each budget is an allowed retained-state ceiling; the resource tables report actual allocations. The native S_{512} main comparator includes the validation-selected K-prior and CM sufficiency controls.

Table 15: Final test accuracy (%). All 14 settings on native encoder features, under the common persistent-state cap S_{512}. K-priors selects among unit-weight 600/2000-iteration and importance-weighted 2000-iteration trajectories on validation data. CM jointly selects its penalty and loss normalization on validation data. At this native small-cap operating point, validation then selects the original Adam-100 or the prespecified L-BFGS-2000 stream at that fixed penalty and normalization; both are reported separately in the CM sensitivity table. QUILT-0 and QUILT-C each select their own penalty and iteration count; these columns are method comparisons, not the strictly matched correction contrast. Entries are mean ± sample standard deviation over seeds 0–2. Aggregate rows use seed-level equal-setting means. Bold marks the best displayed mean, including ties; it does not denote statistical significance.

Dataset	Encoder	Ridge	ER	K-priors$^+$	CM	QUILT-0	QUILT-C
CIFAR-100	DINOv2	81.44 ±0.00	77.54 ±0.51	82.52 ±0.22	69.79 ±1.16	83.49 ±0.17	**83.66** ±0.14
	CLIP	73.80 ±0.00	65.90 ±0.63	74.58 ±0.56	59.52 ±1.27	76.78 ±0.04	**77.04** ±0.24
Pets	DINOv2	93.79 ±0.00	84.33 ±2.52	94.55 ±0.24	94.04 ±0.65	94.12 ±0.11	**94.57** ±0.16
	CLIP	87.74 ±0.00	80.75 ±1.88	87.78 ±0.65	87.43 ±0.39	88.37 ±0.26	**89.51** ±0.45
Flowers-102	DINOv2	99.35 ±0.00	85.03 ±2.02	**99.58** ±0.00	99.39 ±0.11	99.54 ±0.02	99.57 ±0.01
	CLIP	92.37 ±0.00	74.07 ±3.23	93.53 ±0.00	91.32 ±0.48	94.29 ±0.15	**94.66** ±0.19
DTD	DINOv2	75.90 ±0.00	71.10 ±1.22	78.05 ±0.12	78.16 ±0.41	77.68 ±0.03	**78.51** ±0.38
	CLIP	70.80 ±0.00	67.32 ±0.43	73.99 ±0.19	73.48 ±0.63	73.32 ±0.17	**74.33** ±0.48
EuroSAT	DINOv2	94.74 ±0.00	92.35 ±0.62	94.13 ±0.21	92.26 ±0.32	**95.38** ±0.14	95.20 ±0.01
	CLIP	94.54 ±0.00	91.15 ±0.92	93.70 ±0.12	91.10 ±0.63	**95.60** ±0.10	95.09 ±0.09
Banking77	MiniLM	88.31 ±0.00	86.26 ±0.13	91.46 ±0.44	88.38 ±0.28	90.58 ±0.06	**91.67** ±0.12
	MPNet	92.01 ±0.00	89.05 ±0.62	92.66 ±0.34	91.23 ±1.04	92.79 ±0.16	**93.28** ±0.25
20NG	MiniLM	**67.72** ±0.00	60.84 ±0.64	66.33 ±0.31	59.43 ±1.91	66.63 ±0.09	67.26 ±0.36
	MPNet	**71.57** ±0.00	67.42 ±0.56	70.63 ±0.20	66.57 ±0.91	70.33 ±0.32	71.41 ±0.21
Vision mean		86.45 ±0.00	78.95 ±0.79	87.24 ±0.11	83.65 ±0.14	87.86 ±0.03	**88.21** ±0.04
Text mean		79.91 ±0.00	75.89 ±0.21	80.27 ±0.05	76.40 ±0.85	80.08 ±0.07	**80.90** ±0.09
All 14 mean		84.58 ±0.00	78.08 ±0.62	85.25 ±0.07	81.58 ±0.34	85.64 ±0.04	**86.12** ±0.05

Table 16: Final test accuracy (%). All 14 settings on native encoder features, under the common persistent-state cap S_{8192}. K-priors selects between unit-weight 600/2000-iteration trajectories on validation data at the original validation-selected penalty. CM jointly selects its penalty and loss normalization on validation data. QUILT-C uses the archived expanded validation grid at requested $m = 8192$. Entries are mean ± sample standard deviation over seeds 0–2. Aggregate rows use seed-level equal-setting means. Bold marks the best displayed mean, including ties; it does not denote statistical significance.

Dataset	Encoder	Ridge	ER	K-priors	CM	QUILT-C
CIFAR-100	DINOv2	81.44 ±0.00	82.75 ±0.48	83.75 ±0.36	83.92 ±0.16	**85.03** ±0.12
	CLIP	73.80 ±0.00	74.52 ±0.46	76.54 ±0.42	76.44 ±0.31	**78.90** ±0.04
Pets	DINOv2	93.79 ±0.00	85.18 ±3.34	**94.99** ±0.00	93.83 ±0.37	94.90 ±0.00
	CLIP	87.74 ±0.00	80.70 ±1.87	**88.91** ±0.00	87.14 ±0.45	87.98 ±0.00
Flowers-102	DINOv2	99.35 ±0.00	85.44 ±2.38	**99.58** ±0.00	99.28 ±0.07	99.54 ±0.00
	CLIP	92.37 ±0.00	74.07 ±3.23	93.53 ±0.00	89.99 ±0.64	**94.36** ±0.00
DTD	DINOv2	75.90 ±0.00	71.26 ±0.63	**78.99** ±0.00	78.10 ±0.39	78.88 ±0.00
	CLIP	70.80 ±0.00	67.64 ±0.67	**74.41** ±0.00	72.29 ±0.30	74.20 ±0.00
EuroSAT	DINOv2	94.74 ±0.00	93.98 ±0.90	95.77 ±0.13	95.18 ±0.03	**95.94** ±0.07
	CLIP	94.54 ±0.00	92.27 ±1.25	95.60 ±0.10	93.71 ±0.03	**95.67** ±0.06
Banking77	MiniLM	88.31 ±0.00	90.67 ±0.52	**92.99** ±0.00	91.62 ±0.13	92.49 ±0.07
	MPNet	92.01 ±0.00	91.48 ±0.44	**93.51** ±0.00	92.53 ±0.06	93.37 ±0.07
20NG	MiniLM	67.72 ±0.00	62.75 ±0.64	**67.96** ±0.02	67.32 ±0.54	67.71 ±0.23
	MPNet	71.57 ±0.00	66.95 ±0.66	**71.79** ±0.00	70.52 ±0.80	71.19 ±0.08
Vision mean		86.45 ±0.00	80.78 ±1.01	88.21 ±0.09	86.99 ±0.14	**88.54** ±0.02
Text mean		79.91 ±0.00	77.96 ±0.49	**81.56** ±0.01	80.50 ±0.31	81.19 ±0.08
All 14 mean		84.58 ±0.00	79.98 ±0.85	86.31 ±0.06	85.13 ±0.16	**86.44** ±0.02

Table 17: Final test accuracy (%). All 14 settings on the shared 4096-dimensional random projection, under the common persistent-state cap S_{512}. K-priors selects between unit-weight 600/2000-iteration trajectories on validation data at the original validation-selected penalty. CM jointly selects its penalty and loss normalization on validation data. QUILT-0 and QUILT-C each select their own penalty and iteration count; these columns are method comparisons, not the strictly matched correction contrast. Entries are mean ± sample standard deviation over seeds 0–2. Aggregate rows use seed-level equal-setting means. Bold marks the best displayed mean, including ties; it does not denote statistical significance.

Dataset	Encoder	RP-ridge	ER	K-priors	CM	QUILT-0	QUILT-C
CIFAR-100	DINOv2	85.39 ±0.07	81.63 ±0.75	81.40 ±0.54	82.22 ±0.34	85.52 ±0.21	**85.63** ±0.17
	CLIP	77.72 ±0.22	72.40 ±0.91	73.04 ±0.23	72.74 ±0.53	**78.36** ±0.16	**78.36** ±0.03
Pets	DINOv2	94.37 ±0.31	91.69 ±1.13	**94.99** ±0.32	94.30 ±0.17	94.69 ±0.07	94.76 ±0.34
	CLIP	88.53 ±0.53	73.82 ±3.74	88.33 ±0.10	85.73 ±0.84	88.53 ±0.35	**89.48** ±0.34
Flowers-102	DINOv2	99.54 ±0.00	97.69 ±0.65	99.57 ±0.02	98.74 ±0.13	99.53 ±0.03	**99.58** ±0.03
	CLIP	93.97 ±0.03	74.42 ±1.18	93.22 ±0.16	79.50 ±0.23	94.25 ±0.05	**94.45** ±0.11
DTD	DINOv2	78.44 ±0.25	73.72 ±1.61	78.26 ±0.11	76.29 ±1.08	78.39 ±0.48	**79.13** ±0.45
	CLIP	73.14 ±0.51	62.43 ±2.00	73.79 ±0.20	69.18 ±0.39	74.10 ±0.35	**74.11** ±0.59
EuroSAT	DINOv2	96.20 ±0.11	94.28 ±0.48	95.40 ±0.19	95.28 ±0.35	**96.46** ±0.15	96.17 ±0.14
	CLIP	95.62 ±0.09	92.48 ±0.74	94.89 ±0.29	93.65 ±0.11	**95.97** ±0.09	95.57 ±0.13
Banking77	MiniLM	93.26 ±0.22	90.57 ±0.50	92.80 ±0.15	92.29 ±0.35	93.43 ±0.07	**93.46** ±0.22
	MPNet	94.09 ±0.18	91.90 ±0.27	93.55 ±0.22	93.10 ±0.16	94.11 ±0.15	**94.19** ±0.17
20NG	MiniLM	**68.58** ±0.52	63.94 ±0.61	66.76 ±0.29	66.43 ±0.69	68.41 ±0.15	68.56 ±0.14
	MPNet	**72.19** ±0.56	68.18 ±0.67	71.00 ±0.36	70.51 ±0.80	71.64 ±0.39	72.03 ±0.14
Vision mean		88.29 ±0.14	81.46 ±0.71	87.29 ±0.12	84.76 ±0.16	88.58 ±0.08	**88.72** ±0.07
Text mean		82.03 ±0.27	78.65 ±0.35	81.03 ±0.15	80.58 ±0.02	81.90 ±0.10	**82.06** ±0.08
All 14 mean		86.50 ±0.14	80.65 ±0.59	85.50 ±0.11	83.57 ±0.12	86.67 ±0.05	**86.82** ±0.06

Table 18: Final test accuracy (%). All 14 settings on the shared 4096-dimensional random projection, under the common persistent-state cap S_{8192}. K-priors selects between unit-weight 600/2000-iteration trajectories on validation data at the original validation-selected penalty. CM jointly selects its penalty and loss normalization on validation data. QUILT-C uses the archived expanded validation grid at requested $m = 8192$. Entries are mean ± sample standard deviation over seeds 0–2. Aggregate rows use seed-level equal-setting means. Bold marks the best displayed mean, including ties; it does not denote statistical significance.

Dataset	Encoder	RP-ridge	ER	K-priors	CM	QUILT-C
CIFAR-100	DINOv2	85.39 ±0.07	83.76 ±0.60	84.84 ±0.27	85.32 ±0.18	**86.02** ±0.12
	CLIP	77.72 ±0.22	75.26 ±0.69	77.98 ±0.06	77.86 ±0.09	**79.55** ±0.22
Pets	DINOv2	94.37 ±0.31	91.69 ±1.13	**94.99** ±0.32	93.82 ±0.31	94.66 ±0.07
	CLIP	88.53 ±0.53	73.82 ±3.74	88.33 ±0.10	83.42 ±1.27	**88.63** ±0.07
Flowers-102	DINOv2	99.54 ±0.00	97.69 ±0.65	**99.57** ±0.02	98.72 ±0.24	99.56 ±0.00
	CLIP	93.97 ±0.03	74.42 ±1.18	93.22 ±0.16	77.74 ±0.39	**94.39** ±0.12
DTD	DINOv2	78.44 ±0.25	73.72 ±1.61	78.26 ±0.11	75.71 ±0.76	**78.62** ±0.61
	CLIP	73.14 ±0.51	62.43 ±2.00	73.79 ±0.20	67.57 ±0.67	**74.15** ±0.52
EuroSAT	DINOv2	96.20 ±0.11	94.76 ±0.71	96.05 ±0.11	95.56 ±0.32	**96.46** ±0.16
	CLIP	95.62 ±0.09	92.13 ±1.25	95.72 ±0.03	93.71 ±0.14	**95.88** ±0.18
Banking77	MiniLM	**93.26** ±0.22	91.27 ±0.09	92.89 ±0.20	91.58 ±0.34	93.21 ±0.14
	MPNet	**94.09** ±0.18	92.34 ±0.28	93.56 ±0.12	92.87 ±0.25	93.86 ±0.17
20NG	MiniLM	**68.58** ±0.52	64.52 ±0.88	67.63 ±0.05	66.90 ±0.94	68.39 ±0.42
	MPNet	**72.19** ±0.56	68.39 ±0.87	71.73 ±0.11	70.28 ±1.06	71.51 ±0.18
Vision mean		88.29 ±0.14	81.97 ±0.66	88.28 ±0.06	84.94 ±0.19	**88.79** ±0.06
Text mean		**82.03** ±0.27	79.13 ±0.48	81.45 ±0.07	80.41 ±0.52	81.74 ±0.09
All 14 mean		86.50 ±0.14	81.16 ±0.38	86.33 ±0.05	83.65 ±0.18	**86.78** ±0.07

Table 19: Temperature-calibrated test NLL in nats per example; lower is better. Each temperature is fitted only on validation data. All 14 settings on native encoder features, under the common persistent-state cap S_{512}. K-priors selects among unit-weight 600/2000-iteration and importance-weighted 2000-iteration trajectories on validation data. CM jointly selects its penalty and loss normalization on validation data. At this native small-cap operating point, validation then selects the original Adam-100 or the prespecified L-BFGS-2000 stream at that fixed penalty and normalization; both are reported separately in the CM sensitivity table. QUILT-0 and QUILT-C each select their own penalty and iteration count; these columns are method comparisons, not the strictly matched correction contrast. Entries are mean ± sample standard deviation over seeds 0–2. Aggregate rows use seed-level equal-setting means. Bold marks the best displayed mean, including ties; it does not denote statistical significance.

Dataset	Encoder	Ridge	ER	K-priors^{+}	CM	QUILT-0	QUILT-C
CIFAR-100	DINOv2	0.741 ±0.000	0.779 ±0.031	0.574 ±0.002	1.076 ±0.025	0.604 ±0.001	**0.567** ±0.001
	CLIP	1.094 ±0.000	1.219 ±0.020	0.861 ±0.018	1.456 ±0.030	0.877 ±0.007	**0.837** ±0.021
Pets	DINOv2	0.247 ±0.000	0.567 ±0.145	0.191 ±0.012	0.207 ±0.030	0.206 ±0.016	**0.184** ±0.003
	CLIP	0.466 ±0.000	0.672 ±0.147	0.387 ±0.017	0.377 ±0.018	0.386 ±0.041	**0.311** ±0.017
Flowers-102	DINOv2	0.050 ±0.000	0.546 ±0.108	**0.037** ±0.000	0.045 ±0.008	0.038 ±0.001	0.038 ±0.000
	CLIP	0.294 ±0.000	1.525 ±0.171	0.235 ±0.000	0.323 ±0.025	0.210 ±0.002	**0.195** ±0.011
DTD	DINOv2	0.949 ±0.000	1.015 ±0.058	**0.741** ±0.004	0.748 ±0.004	0.768 ±0.001	0.762 ±0.031
	CLIP	1.176 ±0.000	1.227 ±0.011	0.927 ±0.010	0.928 ±0.003	1.004 ±0.027	**0.920** ±0.012
EuroSAT	DINOv2	0.169 ±0.000	0.231 ±0.014	0.172 ±0.002	0.229 ±0.007	0.150 ±0.002	**0.145** ±0.002
	CLIP	0.167 ±0.000	0.255 ±0.025	0.182 ±0.001	0.262 ±0.020	**0.140** ±0.001	0.146 ±0.001
Banking77	MiniLM	0.484 ±0.000	0.501 ±0.022	**0.304** ±0.004	0.404 ±0.004	0.369 ±0.012	0.322 ±0.005
	MPNet	0.345 ±0.000	0.420 ±0.051	**0.262** ±0.012	0.313 ±0.024	0.313 ±0.005	0.268 ±0.004
20NG	MiniLM	1.184 ±0.000	1.345 ±0.052	1.131 ±0.002	1.381 ±0.046	1.199 ±0.013	**1.116** ±0.007
	MPNet	1.062 ±0.000	1.134 ±0.017	**1.001** ±0.003	1.147 ±0.011	1.080 ±0.031	1.019 ±0.013
Vision mean		0.535 ±0.000	0.804 ±0.043	0.431 ±0.001	0.565 ±0.003	0.438 ±0.002	**0.410** ±0.002
Text mean		0.769 ±0.000	0.850 ±0.003	**0.675** ±0.003	0.812 ±0.019	0.740 ±0.010	0.681 ±0.004
All 14 mean		0.602 ±0.000	0.817 ±0.030	0.500 ±0.001	0.635 ±0.007	0.525 ±0.002	**0.488** ±0.000

Table 20: Temperature-calibrated test NLL in nats per example; lower is better. Each temperature is fitted only on validation data. All 14 settings on native encoder features, under the common persistent-state cap S_{8192}. K-priors selects between unit-weight 600/2000-iteration trajectories on validation data at the original validation-selected penalty. CM jointly selects its penalty and loss normalization on validation data. QUILT-C uses the archived expanded validation grid at requested $m = 8192$. Entries are mean ± sample standard deviation over seeds 0–2. Aggregate rows use seed-level equal-setting means. Bold marks the best displayed mean, including ties; it does not denote statistical significance.

Dataset	Encoder	Ridge	ER	K-priors	CM	QUILT-C
CIFAR-100	DINOv2	0.741 ±0.000	0.586 ±0.013	0.547 ±0.009	0.539 ±0.004	**0.502** ±0.002
	CLIP	1.094 ±0.000	0.881 ±0.020	0.805 ±0.006	0.801 ±0.004	**0.737** ±0.001
Pets	DINOv2	0.247 ±0.000	0.514 ±0.141	0.181 ±0.000	0.212 ±0.012	**0.176** ±0.000
	CLIP	0.466 ±0.000	0.641 ±0.135	**0.343** ±0.000	0.393 ±0.009	0.393 ±0.000
Flowers-102	DINOv2	0.050 ±0.000	0.574 ±0.072	**0.037** ±0.000	0.050 ±0.007	**0.037** ±0.000
	CLIP	0.294 ±0.000	1.525 ±0.171	0.235 ±0.000	0.366 ±0.029	**0.203** ±0.000
DTD	DINOv2	0.949 ±0.000	1.007 ±0.051	**0.726** ±0.000	0.771 ±0.005	0.729 ±0.000
	CLIP	1.176 ±0.000	1.212 ±0.040	**0.917** ±0.000	0.984 ±0.002	0.929 ±0.000
EuroSAT	DINOv2	0.169 ±0.000	0.173 ±0.015	0.123 ±0.000	0.142 ±0.004	**0.118** ±0.001
	CLIP	0.167 ±0.000	0.225 ±0.027	**0.126** ±0.001	0.183 ±0.002	0.128 ±0.002
Banking77	MiniLM	0.484 ±0.000	0.357 ±0.007	**0.265** ±0.000	0.304 ±0.003	0.285 ±0.007
	MPNet	0.345 ±0.000	0.321 ±0.005	**0.230** ±0.000	0.266 ±0.007	0.262 ±0.002
20NG	MiniLM	1.184 ±0.000	1.267 ±0.022	**1.086** ±0.000	1.120 ±0.017	1.148 ±0.010
	MPNet	1.062 ±0.000	1.151 ±0.004	**0.973** ±0.000	1.035 ±0.034	1.019 ±0.005
Vision mean		0.535 ±0.000	0.734 ±0.045	0.404 ±0.001	0.444 ±0.005	**0.395** ±0.000
Text mean		0.769 ±0.000	0.774 ±0.007	**0.638** ±0.000	0.681 ±0.013	0.678 ±0.001
All 14 mean		0.602 ±0.000	0.745 ±0.034	**0.471** ±0.001	0.512 ±0.006	0.476 ±0.000

Table 21: Temperature-calibrated test NLL in nats per example; lower is better. Each temperature is fitted only on validation data. All 14 settings on the shared 4096-dimensional random projection, under the common persistent-state cap S_{512}. K-priors selects between unit-weight 600/2000-iteration trajectories on validation data at the original validation-selected penalty. CM jointly selects its penalty and loss normalization on validation data. QUILT-0 and QUILT-C each select their own penalty and iteration count; these columns are method comparisons, not the strictly matched correction contrast. Entries are mean ± sample standard deviation over seeds 0–2. Aggregate rows use seed-level equal-setting means. Bold marks the best displayed mean, including ties; it does not denote statistical significance.

Dataset	Encoder	RP-ridge	ER	K-priors	CM	QUILT-0	QUILT-C
CIFAR-100	DINOv2	0.631 ±0.005	0.634 ±0.025	0.656 ±0.016	0.611 ±0.007	0.589 ±0.001	**0.580** ±0.006
	CLIP	0.976 ±0.001	0.988 ±0.018	0.938 ±0.005	0.941 ±0.005	0.862 ±0.004	**0.853** ±0.003
Pets	DINOv2	0.245 ±0.005	0.287 ±0.048	**0.173** ±0.001	0.208 ±0.007	0.221 ±0.009	0.209 ±0.020
	CLIP	0.394 ±0.037	1.006 ±0.189	0.358 ±0.007	0.438 ±0.021	0.389 ±0.013	**0.320** ±0.013
Flowers-102	DINOv2	0.043 ±0.001	0.101 ±0.011	**0.037** ±0.001	0.068 ±0.011	0.038 ±0.001	0.038 ±0.000
	CLIP	0.242 ±0.003	1.054 ±0.110	0.243 ±0.003	0.741 ±0.020	0.207 ±0.002	**0.194** ±0.002
DTD	DINOv2	0.889 ±0.006	0.946 ±0.090	**0.732** ±0.009	0.873 ±0.007	0.863 ±0.028	0.756 ±0.003
	CLIP	1.123 ±0.040	1.456 ±0.061	**0.922** ±0.012	1.142 ±0.014	1.032 ±0.024	1.029 ±0.052
EuroSAT	DINOv2	0.125 ±0.004	0.168 ±0.016	0.131 ±0.005	0.136 ±0.005	**0.113** ±0.003	0.118 ±0.003
	CLIP	0.139 ±0.001	0.219 ±0.014	0.145 ±0.008	0.182 ±0.008	**0.126** ±0.003	0.130 ±0.001
Banking77	MiniLM	0.346 ±0.008	0.344 ±0.009	**0.268** ±0.003	0.289 ±0.010	0.329 ±0.002	0.305 ±0.023
	MPNet	0.304 ±0.009	0.293 ±0.009	**0.231** ±0.001	0.251 ±0.006	0.290 ±0.003	0.284 ±0.003
20NG	MiniLM	1.197 ±0.043	1.198 ±0.009	1.129 ±0.007	1.140 ±0.010	1.187 ±0.009	**1.114** ±0.004
	MPNet	1.071 ±0.033	1.079 ±0.029	**1.004** ±0.010	1.023 ±0.017	1.091 ±0.031	1.008 ±0.024
Vision mean		0.481 ±0.001	0.686 ±0.030	0.434 ±0.005	0.534 ±0.005	0.444 ±0.005	**0.423** ±0.004
Text mean		0.729 ±0.021	0.729 ±0.010	**0.658** ±0.004	0.676 ±0.004	0.724 ±0.006	0.678 ±0.010
All 14 mean		0.552 ±0.005	0.698 ±0.023	0.498 ±0.005	0.574 ±0.003	0.524 ±0.005	**0.496** ±0.005

Table 22: Temperature-calibrated test NLL in nats per example; lower is better. Each temperature is fitted only on validation data. All 14 settings on the shared 4096-dimensional random projection, under the common persistent-state cap S_{8192}. K-priors selects between unit-weight 600/2000-iteration trajectories on validation data at the original validation-selected penalty. CM jointly selects its penalty and loss normalization on validation data. QUILT-C uses the archived expanded validation grid at requested $m = 8192$. Entries are mean ± sample standard deviation over seeds 0–2. Aggregate rows use seed-level equal-setting means. Bold marks the best displayed mean, including ties; it does not denote statistical significance.

Dataset	Encoder	RP-ridge	ER	K-priors	CM	QUILT-C
CIFAR-100	DINOv2	0.631 ±0.005	0.555 ±0.018	0.508 ±0.003	**0.502** ±0.002	0.530 ±0.035
	CLIP	0.976 ±0.001	0.886 ±0.018	0.751 ±0.005	0.754 ±0.007	**0.720** ±0.002
Pets	DINOv2	0.245 ±0.005	0.287 ±0.048	**0.173** ±0.001	0.225 ±0.014	0.193 ±0.028
	CLIP	0.394 ±0.037	1.006 ±0.189	**0.358** ±0.007	0.509 ±0.035	0.363 ±0.010
Flowers-102	DINOv2	0.043 ±0.001	0.101 ±0.011	**0.037** ±0.001	0.068 ±0.013	0.038 ±0.001
	CLIP	0.242 ±0.003	1.054 ±0.110	0.243 ±0.003	0.810 ±0.022	**0.205** ±0.004
DTD	DINOv2	0.889 ±0.006	0.946 ±0.090	**0.732** ±0.009	0.910 ±0.015	0.862 ±0.033
	CLIP	1.123 ±0.040	1.456 ±0.061	**0.922** ±0.012	1.204 ±0.022	0.997 ±0.055
EuroSAT	DINOv2	0.125 ±0.004	0.158 ±0.016	**0.110** ±0.000	0.129 ±0.005	**0.110** ±0.003
	CLIP	0.139 ±0.001	0.225 ±0.025	0.122 ±0.001	0.180 ±0.005	**0.119** ±0.000
Banking77	MiniLM	0.346 ±0.008	0.322 ±0.006	**0.262** ±0.001	0.314 ±0.005	0.289 ±0.009
	MPNet	0.304 ±0.009	0.292 ±0.012	**0.228** ±0.006	0.269 ±0.007	0.248 ±0.016
20NG	MiniLM	1.197 ±0.043	1.187 ±0.024	**1.104** ±0.001	1.147 ±0.031	1.131 ±0.018
	MPNet	1.071 ±0.033	1.072 ±0.018	**0.983** ±0.002	1.040 ±0.037	1.032 ±0.019
Vision mean		0.481 ±0.001	0.667 ±0.027	**0.396** ±0.003	0.529 ±0.008	0.414 ±0.008
Text mean		0.729 ±0.021	0.718 ±0.011	**0.644** ±0.001	0.693 ±0.020	0.675 ±0.006
All 14 mean		0.552 ±0.005	0.682 ±0.016	**0.467** ±0.002	0.576 ±0.008	0.488 ±0.007

Table 23: Actual retained state under S_{512}, native features. Method columns are mean actual MB (10^6 bytes) over three seeds. $M_{\mathcal{Q}}$ is the measured QUILT node count or seed range. The common cap is an allowance; unused bytes are retained in the run records. Projection storage and two int64 counters are included. Common encoders, evaluator snapshots, and temporary optimizer workspace are excluded.

Dataset	Encoder	Cap (MB)	Ridge	ER	K-priors	CM	QUILT-0/C	$M_{\mathcal{Q}}$
CIFAR-100	DINOv2	3.39	1.80	3.39	3.38	3.39	3.39	512
	CLIP	5.04	2.93	5.03	5.03	5.03	5.04	512
Pets	DINOv2	3.00	1.41	3.00	3.00	3.00	3.00	512
	CLIP	4.52	2.41	4.52	4.51	4.51	4.52	512
Flowers-102	DINOv2	3.40	1.81	3.40	3.40	3.40	3.40	512
	CLIP	5.05	2.94	5.05	4.61	5.05	5.05	512
DTD	DINOv2	3.06	1.48	3.06	3.06	3.06	3.06	512
	CLIP	4.60	2.49	4.60	4.60	4.60	4.60	512
EuroSAT	DINOv2	2.83	1.25	2.83	2.83	2.83	2.83	512
	CLIP	4.30	2.19	4.30	4.29	4.29	4.30	512
Banking77	MiniLM	3.25	1.66	3.24	3.24	3.24	3.25	512
	MPNet	8.84	5.68	8.83	8.83	8.83	8.84	512
20NG	MiniLM	2.89	1.31	2.89	2.89	2.89	2.89	512
	MPNet	8.13	4.98	8.13	8.13	8.13	8.13	512

Table 24: Actual retained state under S_{8192}, native features. Method columns are mean actual MB (10^6 bytes) over three seeds. $M_{\mathcal{Q}}$ is the measured QUILT node count or seed range. The common cap is an allowance; unused bytes are retained in the run records. Projection storage and two int64 counters are included. Common encoders, evaluator snapshots, and temporary optimizer workspace are excluded.

Dataset	Encoder	Cap (MB)	Ridge	ER	K-priors	CM	QUILT-C	$M_{\mathcal{Q}}$
CIFAR-100	DINOv2	27.16	1.80	27.16	27.16	27.16	27.16	8192
	CLIP	36.68	2.93	36.68	36.67	36.67	36.68	8192
Pets	DINOv2	10.53	1.41	9.43	9.21	10.53	10.53	2944
	CLIP	14.54	2.41	12.56	12.26	14.54	14.54	2944
Flowers-102	DINOv2	4.97	1.81	4.09	3.46	4.97	4.97	1020
	CLIP	7.14	2.94	5.45	4.61	7.14	7.14	1020
DTD	DINOv2	7.30	1.48	6.24	5.95	7.29	7.30	1880
	CLIP	10.24	2.49	8.31	7.92	10.23	10.24	1880
EuroSAT	DINOv2	26.61	1.25	26.61	26.61	26.61	26.61	8192
	CLIP	35.94	2.19	35.93	35.93	35.93	35.94	8192
Banking77	MiniLM	27.02	1.66	27.02	27.02	27.02	27.02	8192
	MPNet	56.21	5.68	56.20	55.91	56.20	56.21	8192
20NG	MiniLM	26.67	1.31	26.67	26.67	26.67	26.63	8180
	MPNet	55.51	4.98	55.50	55.50	55.50	55.44	8182

Table 25: Actual retained state under S_{512}, shared RP features. Method columns are mean actual MB (10^6 bytes) over three seeds. $M_{\mathcal{Q}}$ is the measured QUILT node count or seed range. The common cap is an allowance; unused bytes are retained in the run records. Projection storage and two int64 counters are included. Common encoders, evaluator snapshots, and temporary optimizer workspace are excluded.

Dataset	Encoder	Cap (MB)	RP-ridge	ER	K-priors	CM	QUILT-0/C	$M_{\mathcal{Q}}$
CIFAR-100	DINOv2	170.21	153.42	170.21	170.21	170.21	170.21	512
	CLIP	174.41	157.62	174.40	174.40	174.40	174.41	512
Pets	DINOv2	166.08	149.29	112.74	110.31	166.08	166.08	512
	CLIP	170.28	153.49	116.93	114.51	170.27	170.28	512
Flowers-102	DINOv2	170.34	153.55	56.05	49.37	170.34	170.34	512
	CLIP	174.54	157.75	60.25	53.56	174.53	174.54	512
DTD	DINOv2	166.74	149.95	78.84	75.76	166.73	166.74	512
	CLIP	170.93	154.14	83.03	79.95	170.93	170.93	512
EuroSAT	DINOv2	164.31	147.52	164.31	164.31	164.31	164.31	512
	CLIP	168.51	151.72	168.50	168.50	168.50	168.51	512
Banking77	MiniLM	168.70	151.91	168.70	168.70	168.70	168.70	512
	MPNet	181.29	164.50	181.28	181.28	181.28	181.29	512
20NG	MiniLM	164.97	148.18	164.96	164.96	164.96	164.97	512
	MPNet	177.55	160.76	177.55	177.55	177.55	177.55	512

Table 26: Actual retained state under S_{8192}, shared RP features. Method columns are mean actual MB (10^6 bytes) over three seeds. $M_{\mathcal{Q}}$ is the measured QUILT node count or seed range. The common cap is an allowance; unused bytes are retained in the run records. Projection storage and two int64 counters are included. Common encoders, evaluator snapshots, and temporary optimizer workspace are excluded.

Dataset	Encoder	Cap (MB)	RP-ridge	ER	K-priors	CM	QUILT-C	$M_{\mathcal{Q}}$
CIFAR-100	DINOv2	422.05	153.42	422.05	422.02	422.02	422.05	8192
	CLIP	426.25	157.62	426.25	426.22	426.22	426.25	8192
Pets	DINOv2	245.83	149.29	112.74	110.31	245.81	245.83	2944
	CLIP	250.03	153.49	116.93	114.51	250.00	250.03	2944
Flowers-102	DINOv2	187.00	153.55	56.05	49.37	186.99	187.00	1020
	CLIP	191.19	157.75	60.25	53.56	191.19	191.19	1020
DTD	DINOv2	211.60	149.95	78.84	75.76	211.58	211.60	1880
	CLIP	215.79	154.14	83.03	79.95	215.78	215.79	1880
EuroSAT	DINOv2	416.15	147.52	416.15	416.12	416.12	416.15	8192
	CLIP	420.35	151.72	420.35	420.32	420.32	420.35	8192
Banking77	MiniLM	420.55	151.91	315.21	310.16	420.51	420.55	8192
	MPNet	433.13	164.50	327.79	322.75	433.10	433.13	8192
20NG	MiniLM	416.81	148.18	348.42	347.11	416.78	416.43	8180–8181
	MPNet	429.39	160.76	361.00	359.69	429.36	429.06	8182

F.2 DIRECT-BASELINE SENSITIVITY

The K-prior tables keep the original unit-weight implementation, longer optimization, and native small-budget importance weighting visible. The CM tables retain both head optimizers at the original selected penalty and normalization. The selected variants are determined from validation scores, not from the test columns below.

Table 27: K-priors accuracy (%) on native features under S_{512}. Unit 600 and Unit 2000 use the same anchored objective, selected penalty, and resident examples; Unit val. selects their iteration budget on validation accuracy, then raw NLL. IS 2000 uses the prespecified importance weighting and its own validation-selected penalty. Family val. selects among Unit 600, Unit 2000, and IS 2000 by the same validation rule. Every branch is retained. Entries are mean ± seed standard deviation; aggregate rows use equal-setting seed blocks. Bold refers only to displayed means.

Dataset	Encoder	Unit 600	Unit 2000	Unit val.	IS 2000	Family val.
CIFAR-100	DINOv2	69.82 ±1.37	69.82 ±1.37	69.82 ±1.37	**82.52** ±0.22	**82.52** ±0.22
	CLIP	60.61 ±1.03	60.61 ±1.03	60.61 ±1.03	**74.58** ±0.56	**74.58** ±0.56
Pets	DINOv2	94.29 ±0.06	94.29 ±0.06	94.29 ±0.06	**94.71** ±0.05	94.55 ±0.24
	CLIP	87.40 ±0.33	87.40 ±0.33	87.40 ±0.33	**88.05** ±0.18	87.78 ±0.65
Flowers-102	DINOv2	**99.58** ±0.00	**99.58** ±0.00	**99.58** ±0.00	**99.58** ±0.00	**99.58** ±0.00
	CLIP	**93.53** ±0.00	**93.53** ±0.00	**93.53** ±0.00	**93.53** ±0.00	**93.53** ±0.00
DTD	DINOv2	77.84 ±0.25	77.84 ±0.25	77.84 ±0.25	**78.17** ±0.08	78.05 ±0.12
	CLIP	**73.99** ±0.19	**73.99** ±0.19	**73.99** ±0.19	73.56 ±0.96	**73.99** ±0.19
EuroSAT	DINOv2	90.69 ±0.28	90.69 ±0.28	90.69 ±0.28	**94.13** ±0.21	**94.13** ±0.21
	CLIP	90.66 ±0.70	90.66 ±0.70	90.66 ±0.70	**93.70** ±0.12	**93.70** ±0.12
Banking77	MiniLM	88.26 ±0.33	88.26 ±0.33	88.26 ±0.33	**91.46** ±0.44	**91.46** ±0.44
	MPNet	90.77 ±0.41	90.77 ±0.41	90.77 ±0.41	**92.66** ±0.34	**92.66** ±0.34
20NG	MiniLM	56.49 ±1.86	56.49 ±1.86	56.49 ±1.86	**66.33** ±0.31	**66.33** ±0.31
	MPNet	63.49 ±1.43	63.49 ±1.43	63.49 ±1.43	**70.63** ±0.20	**70.63** ±0.20
Vision mean		83.84 ±0.35	83.84 ±0.35	83.84 ±0.35	**87.25** ±0.15	87.24 ±0.11
Text mean		74.75 ±0.76	74.75 ±0.76	74.75 ±0.76	**80.27** ±0.05	**80.27** ±0.05
All 14 mean		81.24 ±0.46	81.24 ±0.46	81.24 ±0.46	**85.26** ±0.10	85.25 ±0.07

Table 28: K-priors accuracy (%) on native features under S_{8192}. Unit 600 and Unit 2000 use the same anchored objective, selected penalty, and resident examples; Unit val. selects their iteration budget on validation accuracy, then raw NLL. The IS sensitivity study was prespecified only for native features at the smaller cap. Every branch is retained. Entries are mean ± seed standard deviation; aggregate rows use equal-setting seed blocks. Bold refers only to displayed means.

Dataset	Encoder	Unit 600	Unit 2000	Unit val.
CIFAR-100	DINOv2	**83.75** ±0.36	**83.75** ±0.36	**83.75** ±0.36
	CLIP	76.54 ±0.42	**76.55** ±0.42	76.54 ±0.42
Pets	DINOv2	**94.99** ±0.00	**94.99** ±0.00	**94.99** ±0.00
	CLIP	**88.91** ±0.00	**88.91** ±0.00	**88.91** ±0.00
Flowers-102	DINOv2	**99.58** ±0.00	**99.58** ±0.00	**99.58** ±0.00
	CLIP	**93.53** ±0.00	**93.53** ±0.00	**93.53** ±0.00
DTD	DINOv2	**78.99** ±0.00	**78.99** ±0.00	**78.99** ±0.00
	CLIP	**74.41** ±0.00	**74.41** ±0.00	**74.41** ±0.00
EuroSAT	DINOv2	**95.77** ±0.13	**95.77** ±0.13	**95.77** ±0.13
	CLIP	**95.60** ±0.10	**95.60** ±0.10	**95.60** ±0.10
Banking77	MiniLM	**92.99** ±0.00	**92.99** ±0.00	**92.99** ±0.00
	MPNet	**93.51** ±0.00	**93.51** ±0.00	**93.51** ±0.00
20NG	MiniLM	**67.96** ±0.02	**67.96** ±0.02	**67.96** ±0.02
	MPNet	**71.79** ±0.00	**71.79** ±0.00	**71.79** ±0.00
Vision mean		**88.21** ±0.09	**88.21** ±0.09	**88.21** ±0.09
Text mean		**81.56** ±0.01	**81.56** ±0.01	**81.56** ±0.01
All 14 mean		**86.31** ±0.06	**86.31** ±0.06	**86.31** ±0.06

Table 29: K-priors accuracy (%) on shared RP features under S_{512}. Unit 600 and Unit 2000 use the same anchored objective, selected penalty, and resident examples; Unit val. selects their iteration budget on validation accuracy, then raw NLL. The IS sensitivity study was prespecified only for native features at the smaller cap. Every branch is retained. Entries are mean ± seed standard deviation; aggregate rows use equal-setting seed blocks. Bold refers only to displayed means.

Dataset	Encoder	Unit 600	Unit 2000	Unit val.
CIFAR-100	DINOv2	**81.40** ±0.54	**81.40** ±0.54	**81.40** ±0.54
	CLIP	**73.04** ±0.23	**73.04** ±0.23	**73.04** ±0.23
Pets	DINOv2	**94.99** ±0.32	**94.99** ±0.32	**94.99** ±0.32
	CLIP	**88.33** ±0.10	**88.33** ±0.10	**88.33** ±0.10
Flowers-102	DINOv2	**99.57** ±0.02	**99.57** ±0.02	**99.57** ±0.02
	CLIP	**93.22** ±0.16	**93.22** ±0.16	**93.22** ±0.16
DTD	DINOv2	**78.26** ±0.11	**78.26** ±0.11	**78.26** ±0.11
	CLIP	**73.79** ±0.20	**73.79** ±0.20	**73.79** ±0.20
EuroSAT	DINOv2	**95.40** ±0.19	**95.40** ±0.19	**95.40** ±0.19
	CLIP	**94.89** ±0.29	94.88 ±0.29	**94.89** ±0.29
Banking77	MiniLM	**92.80** ±0.15	**92.80** ±0.15	**92.80** ±0.15
	MPNet	**93.55** ±0.22	**93.55** ±0.22	**93.55** ±0.22
20NG	MiniLM	**66.76** ±0.29	**66.76** ±0.29	**66.76** ±0.29
	MPNet	**71.00** ±0.36	**71.00** ±0.36	**71.00** ±0.36
Vision mean		**87.29** ±0.12	**87.29** ±0.12	**87.29** ±0.12
Text mean		**81.03** ±0.15	**81.03** ±0.15	**81.03** ±0.15
All 14 mean		**85.50** ±0.11	**85.50** ±0.11	**85.50** ±0.11

Table 30: K-priors accuracy (%) on shared RP features under S_{8192}. Unit 600 and Unit 2000 use the same anchored objective, selected penalty, and resident examples; Unit val. selects their iteration budget on validation accuracy, then raw NLL. The IS sensitivity study was prespecified only for native features at the smaller cap. Every branch is retained. Entries are mean ± seed standard deviation; aggregate rows use equal-setting seed blocks. Bold refers only to displayed means.

Dataset	Encoder	Unit 600	Unit 2000	Unit val.
CIFAR-100	DINOv2	**84.84** ±0.27	**84.84** ±0.27	**84.84** ±0.27
	CLIP	**77.98** ±0.06	**77.98** ±0.06	**77.98** ±0.06
Pets	DINOv2	**94.99** ±0.32	**94.99** ±0.32	**94.99** ±0.32
	CLIP	**88.33** ±0.10	**88.33** ±0.10	**88.33** ±0.10
Flowers-102	DINOv2	**99.57** ±0.02	**99.57** ±0.02	**99.57** ±0.02
	CLIP	**93.22** ±0.16	**93.22** ±0.16	**93.22** ±0.16
DTD	DINOv2	**78.26** ±0.11	**78.26** ±0.11	**78.26** ±0.11
	CLIP	**73.79** ±0.20	**73.79** ±0.20	**73.79** ±0.20
EuroSAT	DINOv2	**96.05** ±0.11	**96.05** ±0.11	**96.05** ±0.11
	CLIP	**95.72** ±0.04	**95.72** ±0.03	**95.72** ±0.03
Banking77	MiniLM	**92.89** ±0.20	**92.89** ±0.20	**92.89** ±0.20
	MPNet	**93.56** ±0.12	**93.56** ±0.12	**93.56** ±0.12
20NG	MiniLM	**67.63** ±0.05	**67.63** ±0.05	**67.63** ±0.05
	MPNet	**71.73** ±0.11	**71.73** ±0.11	**71.73** ±0.11
Vision mean		88.27 ±0.06	**88.28** ±0.06	**88.28** ±0.06
Text mean		**81.45** ±0.07	**81.45** ±0.07	**81.45** ±0.07
All 14 mean		**86.33** ±0.05	**86.33** ±0.05	**86.33** ±0.05

Table 31: K-priors calibrated NLL (nats) on native features under S_{512}. Unit 600 and Unit 2000 use the same anchored objective, selected penalty, and resident examples; Unit val. selects their iteration budget on validation accuracy, then raw NLL. IS 2000 uses the prespecified importance weighting and its own validation-selected penalty. Family val. selects among Unit 600, Unit 2000, and IS 2000 by the same validation rule. Every branch is retained. Entries are mean ± seed standard deviation; aggregate rows use equal-setting seed blocks. Bold refers only to displayed means.

Dataset	Encoder	Unit 600	Unit 2000	Unit val.	IS 2000	Family val.
CIFAR-100	DINOv2	1.106 ±0.026	1.106 ±0.026	1.106 ±0.026	**0.574** ±0.002	**0.574** ±0.002
	CLIP	1.437 ±0.023	1.437 ±0.023	1.437 ±0.023	**0.861** ±0.018	**0.861** ±0.018
Pets	DINOv2	0.202 ±0.001	0.202 ±0.001	0.202 ±0.001	**0.183** ±0.006	0.191 ±0.012
	CLIP	0.406 ±0.005	0.406 ±0.005	0.406 ±0.005	**0.375** ±0.011	0.387 ±0.017
Flowers-102	DINOv2	**0.037** ±0.000	**0.037** ±0.000	**0.037** ±0.000	**0.037** ±0.000	**0.037** ±0.000
	CLIP	**0.235** ±0.000	**0.235** ±0.000	**0.235** ±0.000	**0.235** ±0.000	**0.235** ±0.000
DTD	DINOv2	0.746 ±0.010	0.746 ±0.010	0.746 ±0.010	**0.733** ±0.007	0.741 ±0.004
	CLIP	**0.927** ±0.010	**0.927** ±0.010	**0.927** ±0.010	0.929 ±0.013	**0.927** ±0.010
EuroSAT	DINOv2	0.272 ±0.009	0.272 ±0.009	0.272 ±0.009	**0.172** ±0.002	**0.172** ±0.002
	CLIP	0.281 ±0.017	0.281 ±0.017	0.281 ±0.017	**0.182** ±0.001	**0.182** ±0.001
Banking77	MiniLM	0.420 ±0.008	0.420 ±0.008	0.420 ±0.008	**0.304** ±0.004	**0.304** ±0.004
	MPNet	0.330 ±0.007	0.330 ±0.007	0.330 ±0.007	**0.262** ±0.012	**0.262** ±0.012
20NG	MiniLM	1.497 ±0.037	1.497 ±0.037	1.497 ±0.037	**1.131** ±0.002	**1.131** ±0.002
	MPNet	1.259 ±0.033	1.259 ±0.033	1.259 ±0.033	**1.001** ±0.003	**1.001** ±0.003
Vision mean		0.565 ±0.005	0.565 ±0.005	0.565 ±0.005	**0.428** ±0.003	0.431 ±0.001
Text mean		0.877 ±0.014	0.877 ±0.014	0.877 ±0.014	**0.675** ±0.003	**0.675** ±0.003
All 14 mean		0.654 ±0.008	0.654 ±0.008	0.654 ±0.008	**0.498** ±0.003	0.500 ±0.001

Table 32: K-priors calibrated NLL (nats) on native features under S_{8192}. Unit 600 and Unit 2000 use the same anchored objective, selected penalty, and resident examples; Unit val. selects their iteration budget on validation accuracy, then raw NLL. The IS sensitivity study was prespecified only for native features at the smaller cap. Every branch is retained. Entries are mean ± seed standard deviation; aggregate rows use equal-setting seed blocks. Bold refers only to displayed means.

Dataset	Encoder	Unit 600	Unit 2000	Unit val.
CIFAR-100	DINOv2	**0.547** ±0.009	**0.547** ±0.009	**0.547** ±0.009
	CLIP	**0.805** ±0.006	**0.805** ±0.006	**0.805** ±0.006
Pets	DINOv2	**0.181** ±0.000	**0.181** ±0.000	**0.181** ±0.000
	CLIP	**0.343** ±0.000	**0.343** ±0.000	**0.343** ±0.000
Flowers-102	DINOv2	**0.037** ±0.000	**0.037** ±0.000	**0.037** ±0.000
	CLIP	**0.235** ±0.000	**0.235** ±0.000	**0.235** ±0.000
DTD	DINOv2	**0.726** ±0.000	**0.726** ±0.000	**0.726** ±0.000
	CLIP	**0.917** ±0.000	**0.917** ±0.000	**0.917** ±0.000
EuroSAT	DINOv2	**0.123** ±0.000	**0.123** ±0.000	**0.123** ±0.000
	CLIP	**0.126** ±0.001	**0.126** ±0.001	**0.126** ±0.001
Banking77	MiniLM	**0.265** ±0.000	**0.265** ±0.000	**0.265** ±0.000
	MPNet	**0.230** ±0.000	**0.230** ±0.000	**0.230** ±0.000
20NG	MiniLM	**1.086** ±0.000	**1.086** ±0.000	**1.086** ±0.000
	MPNet	**0.973** ±0.000	**0.973** ±0.000	**0.973** ±0.000
Vision mean		**0.404** ±0.001	**0.404** ±0.001	**0.404** ±0.001
Text mean		**0.638** ±0.000	**0.638** ±0.000	**0.638** ±0.000
All 14 mean		**0.471** ±0.001	**0.471** ±0.001	**0.471** ±0.001

Table 33: K-priors calibrated NLL (nats) on shared RP features under S_{512}. Unit 600 and Unit 2000 use the same anchored objective, selected penalty, and resident examples; Unit val. selects their iteration budget on validation accuracy, then raw NLL. The IS sensitivity study was prespecified only for native features at the smaller cap. Every branch is retained. Entries are mean ± seed standard deviation; aggregate rows use equal-setting seed blocks. Bold refers only to displayed means.

Dataset	Encoder	Unit 600	Unit 2000	Unit val.
CIFAR-100	DINOv2	**0.656** ±0.016	**0.656** ±0.016	**0.656** ±0.016
	CLIP	**0.938** ±0.005	**0.938** ±0.005	**0.938** ±0.005
Pets	DINOv2	**0.173** ±0.001	**0.173** ±0.001	**0.173** ±0.001
	CLIP	**0.358** ±0.007	**0.358** ±0.007	**0.358** ±0.007
Flowers-102	DINOv2	**0.037** ±0.001	**0.037** ±0.001	**0.037** ±0.001
	CLIP	**0.243** ±0.003	**0.243** ±0.003	**0.243** ±0.003
DTD	DINOv2	**0.732** ±0.009	**0.732** ±0.009	**0.732** ±0.009
	CLIP	**0.922** ±0.012	**0.922** ±0.012	**0.922** ±0.012
EuroSAT	DINOv2	**0.131** ±0.005	**0.131** ±0.005	**0.131** ±0.005
	CLIP	**0.145** ±0.008	**0.145** ±0.008	**0.145** ±0.008
Banking77	MiniLM	**0.268** ±0.003	**0.268** ±0.003	**0.268** ±0.003
	MPNet	**0.231** ±0.001	**0.231** ±0.001	**0.231** ±0.001
20NG	MiniLM	**1.129** ±0.007	**1.129** ±0.007	**1.129** ±0.007
	MPNet	**1.004** ±0.010	**1.004** ±0.010	**1.004** ±0.010
Vision mean		**0.434** ±0.005	**0.434** ±0.005	**0.434** ±0.005
Text mean		**0.658** ±0.004	**0.658** ±0.004	**0.658** ±0.004
All 14 mean		**0.498** ±0.005	**0.498** ±0.005	**0.498** ±0.005

Table 34: K-priors calibrated NLL (nats) on shared RP features under S_{8192}. Unit 600 and Unit 2000 use the same anchored objective, selected penalty, and resident examples; Unit val. selects their iteration budget on validation accuracy, then raw NLL. The IS sensitivity study was prespecified only for native features at the smaller cap. Every branch is retained. Entries are mean ± seed standard deviation; aggregate rows use equal-setting seed blocks. Bold refers only to displayed means.

Dataset	Encoder	Unit 600	Unit 2000	Unit val.
CIFAR-100	DINOv2	**0.508** ±0.003	**0.508** ±0.003	**0.508** ±0.003
	CLIP	**0.751** ±0.005	**0.751** ±0.005	**0.751** ±0.005
Pets	DINOv2	**0.173** ±0.001	**0.173** ±0.001	**0.173** ±0.001
	CLIP	**0.358** ±0.007	**0.358** ±0.007	**0.358** ±0.007
Flowers-102	DINOv2	**0.037** ±0.001	**0.037** ±0.001	**0.037** ±0.001
	CLIP	**0.243** ±0.003	**0.243** ±0.003	**0.243** ±0.003
DTD	DINOv2	**0.732** ±0.009	**0.732** ±0.009	**0.732** ±0.009
	CLIP	**0.922** ±0.012	**0.922** ±0.012	**0.922** ±0.012
EuroSAT	DINOv2	**0.110** ±0.000	**0.110** ±0.000	**0.110** ±0.000
	CLIP	**0.122** ±0.001	**0.122** ±0.001	**0.122** ±0.001
Banking77	MiniLM	**0.262** ±0.001	**0.262** ±0.001	**0.262** ±0.001
	MPNet	**0.228** ±0.006	**0.228** ±0.006	**0.228** ±0.006
20NG	MiniLM	**1.104** ±0.001	**1.104** ±0.001	**1.104** ±0.001
	MPNet	**0.983** ±0.002	**0.983** ±0.002	**0.983** ±0.002
Vision mean		**0.396** ±0.003	**0.396** ±0.003	**0.396** ±0.003
Text mean		**0.644** ±0.001	**0.644** ±0.001	**0.644** ±0.001
All 14 mean		**0.467** ±0.002	**0.467** ±0.002	**0.467** ±0.002

Table 35: CM head-optimization sensitivity: accuracy (%). Native features under S_{512}, all 14 settings and three seeds. Adam 100 is the unchanged production result. L-BFGS 2000 reruns the full ten-arrival stream at the original validation-selected penalty and normalization, preserving the PPCA update count, memory-growth rule, seed, and retained-state cap. Validation selects between the two completed streams by accuracy, then raw NLL. The selected column is the CM comparator in the native small-cap table. Both branches remain visible; L-BFGS is an explicit optimization variant, not the published Adam implementation. Entries show mean ± seed standard deviation; aggregate rows use equal-setting seed blocks.

Dataset	Encoder	Adam 100	L-BFGS 2000	Validation-selected
CIFAR-100	DINOv2	69.77 ±1.17	69.75 ±1.18	**69.79** ±1.16
	CLIP	58.38 ±1.14	**59.52** ±1.27	**59.52** ±1.27
Pets	DINOv2	94.04 ±0.65	**94.20** ±0.46	94.04 ±0.65
	CLIP	**87.43** ±0.39	87.35 ±0.67	**87.43** ±0.39
Flowers-102	DINOv2	99.39 ±0.11	**99.40** ±0.04	99.39 ±0.11
	CLIP	**91.32** ±0.48	90.63 ±0.41	**91.32** ±0.48
DTD	DINOv2	**78.21** ±0.35	**78.21** ±0.50	78.16 ±0.41
	CLIP	**73.48** ±0.63	**73.48** ±0.03	**73.48** ±0.63
EuroSAT	DINOv2	91.03 ±0.45	**92.26** ±0.32	**92.26** ±0.32
	CLIP	87.91 ±0.12	**91.10** ±0.63	**91.10** ±0.63
Banking77	MiniLM	87.72 ±0.75	**88.51** ±0.12	88.38 ±0.28
	MPNet	90.85 ±0.63	**91.29** ±0.96	91.23 ±1.04
20NG	MiniLM	59.39 ±1.92	**59.43** ±1.91	**59.43** ±1.91
	MPNet	66.57 ±0.91	**66.81** ±1.01	66.57 ±0.91
Vision mean		83.10 ±0.18	83.59 ±0.18	**83.65** ±0.14
Text mean		76.13 ±1.02	**76.51** ±0.90	76.40 ±0.85
All 14 mean		81.11 ±0.40	81.57 ±0.38	**81.58** ±0.34

Table 36: CM head-optimization sensitivity: calibrated NLL (nats). Native features under S_{512}, all 14 settings and three seeds. Adam 100 is the unchanged production result. L-BFGS 2000 reruns the full ten-arrival stream at the original validation-selected penalty and normalization, preserving the PPCA update count, memory-growth rule, seed, and retained-state cap. Validation selects between the two completed streams by accuracy, then raw NLL. The selected column is the CM comparator in the native small-cap table. Both branches remain visible; L-BFGS is an explicit optimization variant, not the published Adam implementation. Entries show mean ± seed standard deviation; aggregate rows use equal-setting seed blocks.

Dataset	Encoder	Adam 100	L-BFGS 2000	Validation-selected
CIFAR-100	DINOv2	1.077 ± 0.025	1.078 ± 0.024	$\mathbf{1.076} \pm 0.025$
	CLIP	1.516 ± 0.029	$\mathbf{1.456} \pm 0.030$	$\mathbf{1.456} \pm 0.030$
Pets	DINOv2	0.207 ± 0.030	$\mathbf{0.201} \pm 0.022$	0.207 ± 0.030
	CLIP	$\mathbf{0.377} \pm 0.018$	0.387 ± 0.021	$\mathbf{0.377} \pm 0.018$
Flowers-102	DINOv2	$\mathbf{0.045} \pm 0.008$	$\mathbf{0.045} \pm 0.006$	$\mathbf{0.045} \pm 0.008$
	CLIP	$\mathbf{0.323} \pm 0.025$	0.344 ± 0.027	$\mathbf{0.323} \pm 0.025$
DTD	DINOv2	$\mathbf{0.748} \pm 0.004$	$\mathbf{0.748} \pm 0.003$	$\mathbf{0.748} \pm 0.004$
	CLIP	$\mathbf{0.928} \pm 0.003$	0.952 ± 0.009	$\mathbf{0.928} \pm 0.003$
EuroSAT	DINOv2	0.269 ± 0.007	$\mathbf{0.229} \pm 0.007$	$\mathbf{0.229} \pm 0.007$
	CLIP	0.365 ± 0.012	$\mathbf{0.262} \pm 0.020$	$\mathbf{0.262} \pm 0.020$
Banking77	MiniLM	0.431 ± 0.034	$\mathbf{0.402} \pm 0.006$	0.404 ± 0.004
	MPNet	0.328 ± 0.006	$\mathbf{0.311} \pm 0.021$	0.313 ± 0.024
20NG	MiniLM	1.384 ± 0.047	1.383 ± 0.045	$\mathbf{1.381} \pm 0.046$
	MPNet	$\mathbf{1.147} \pm 0.011$	$\mathbf{1.147} \pm 0.012$	$\mathbf{1.147} \pm 0.011$
Vision mean		0.585 ± 0.004	0.570 ± 0.003	$\mathbf{0.565} \pm 0.003$
Text mean		0.823 ± 0.024	$\mathbf{0.811} \pm 0.019$	0.812 ± 0.019
All 14 mean		0.653 ± 0.010	0.639 ± 0.007	$\mathbf{0.635} \pm 0.007$

F.3 Independent paired replication

The primary replication includes all 14 original dataset–encoder settings, including the text settings summarized in Table 3. Its ten new seeds were fixed before their results were computed. Each seed is a complete block: the two learners share their data, class order, and exact RP matrix, and their accuracy differences are averaged equally over the 14 settings. The paired interval therefore uses ten observations, not 140 independent observations.

The mean gain is 0.356 points, with paired t interval $[0.290, 0.422]$, paired t-test $p = 6.79 \times 10^{-7}$, and exact two-sided seed-sign $p = 0.00195$. The bootstrap interval is $[0.301, 0.410]$. All ten seed-block differences are positive. The visual mean is positive in nine of ten individual settings; Flowers-102 with DINOv2 is -0.008 points and its interval spans zero. The text aggregate is -0.118 points with interval $[-0.279, 0.043]$. These estimates concern this fixed benchmark suite under algorithmic randomness. They do not establish a universal gain over future datasets.

Table 37: Complete new-seed RP replication, seeds 3–12, fixed primary $m = 8192$. Accuracies are percentages. Each setting's gain and 95% paired t interval use its ten paired seeds. Holm-adjusted two-sided paired-t p values are secondary setting-wise comparisons; the all-setting macro difference is the prespecified primary endpoint.

Dataset	Encoder	RP-ridge	QUILT-C	Gain [95% CI]	Holm p
CIFAR-100	DINOv2	85.42	86.20	$+0.78$ $[+0.64, +0.93]$	$7.37e-06$
	CLIP	77.53	79.37	$+1.84$ $[+1.66, +2.02]$	$3.51e-08$
Pets	DINOv2	94.60	95.01	$+0.40$ $[+0.20, +0.60]$	0.0119
	CLIP	88.44	88.48	$+0.04$ $[-0.41, +0.48]$	1
Flowers-102	DINOv2	99.56	99.55	-0.01 $[-0.03, +0.01]$	1
	CLIP	93.96	94.47	$+0.51$ $[+0.34, +0.67]$	0.00068
DTD	DINOv2	78.52	78.65	$+0.13$ $[-0.10, +0.36]$	0.973
	CLIP	73.08	74.28	$+1.20$ $[+0.87, +1.53]$	0.000214
EuroSAT	DINOv2	96.31	96.51	$+0.19$ $[+0.08, +0.31]$	0.039
	CLIP	95.52	95.89	$+0.37$ $[+0.23, +0.52]$	0.00276
Banking77	MiniLM	93.13	93.36	$+0.23$ $[+0.06, +0.40]$	0.0913
	MPNet	93.99	93.97	-0.02 $[-0.23, +0.18]$	1
20NG	MiniLM	68.80	68.49	-0.30 $[-0.57, -0.03]$	0.157
	MPNet	72.23	71.85	-0.38 $[-0.69, -0.07]$	0.129

Table 38: The replication first averages paired gains within each seed over the indicated fixed settings. All intervals and tests then use ten seed blocks. The exact sign test concerns sign balance; the bootstrap resamples complete seed blocks.

Fixed setting group	Gain [95% CI]	Bootstrap 95% CI	Sign p
All 14 (primary)	$+0.356$ $[+0.290, +0.422]$	$[+0.301, +0.410]$	0.00195
Vision 10	$+0.546$ $[+0.477, +0.615]$	$[+0.484, +0.595]$	0.00195
Text 4	-0.118 $[-0.279, +0.043]$	$[-0.247, +0.015]$	0.34375

F.4 Strictly unsaturated correction study

All 364 cases are complete, with 140 new-seed cases per feature map and 84 separate development cases. Every case actually retains 512 nodes, fewer than its fitting examples. The smallest residual-trace fraction across the study is 0.01985, so none of these effects comes from a saturated summary. Both arms share the same moment tensors, nodes, initialization, and storage. The primary intervention uses the uncorrected model's validation-selected penalty and iteration count. The reverse selector and the independently tuned comparison were specified separately and do not replace that intervention.

The native primary effect is $+0.394$ points, with ten positive seed blocks. The RP primary effect is -0.037 points, with only two positive blocks. Its paired t interval lies just below zero, whereas the exact seed-sign p value is 0.1094. These procedures address different summaries of the paired distribution; neither provides evidence for a positive RP primary effect.

Selection behavior explains why the reverse and independently tuned comparisons must remain separate. On RP features, the uncorrected arm selects $J \leq 32$ in 97 of 140 cases, whereas the corrected arm does so in 66. The corrected arm selects $J = 512$ in 32 cases, compared with 11 for the uncorrected arm. Both penalty and iteration count may change, so these observations do not identify iteration count as the sole source of the interaction. The legacy $J \leq 64$ primary effects are $+0.309$ points on native features and -0.038 on RP features. Restricting the iteration grid does not remove the missing positive RP component effect.

Calibrated NLL decreases by 0.02490 and 0.00896 in the native and RP primary comparisons, respectively, while raw NLL increases by 0.02305 and 0.01671. Each arm's temperature is fitted on validation data. Classification geometry, raw confidence, and calibrated predictive loss therefore remain distinct outcomes.

Table 39: All-setting fresh correction effects at strictly unsaturated $m = 512$. Every pair keeps its summary, initialization, penalty, iteration count, and actual retained bytes fixed. The selector rules define distinct conditional contrasts. Accuracy is in percentage points and NLL in nats.

Features	Configuration rule	Acc. effect [95% CI]	Raw NLL Δ	Cal. NLL Δ
Native	$\alpha = 0$ selector (primary)	$+0.394$ $[+0.342, +0.447]$	0.023	-0.025
Native	$\alpha = 1$ selector (reverse)	$+0.917$ $[+0.837, +0.996]$	-0.015	-0.050
RP4096	$\alpha = 0$ selector (primary)	-0.037 $[-0.071, -0.002]$	0.017	-0.009
RP4096	$\alpha = 1$ selector (reverse)	$+0.391$ $[+0.312, +0.469]$	0.000	-0.026

Table 40: Native correction effects by setting, averaged over ten fresh seeds. All cases have 512 nodes and more than 512 fitting examples. The independently selected column allows separate validation-selected penalties and iteration counts; it is not the fixed-parameter intervention.

Dataset	Encoder	Primary effect	Reverse effect	Independently selected
CIFAR-100	DINOv2	$+0.016$	$+0.503$	$+0.274$
	CLIP	$+0.144$	$+0.559$	$+0.289$
Pets	DINOv2	$+0.387$	$+0.316$	$+0.327$
	CLIP	$+1.085$	$+2.614$	$+1.300$
Flowers-102	DINOv2	$+0.013$	$+0.024$	$+0.039$
	CLIP	$+0.559$	$+0.748$	$+0.738$
DTD	DINOv2	$+0.755$	$+0.883$	$+0.697$
	CLIP	$+0.532$	$+0.957$	$+0.957$
EuroSAT	DINOv2	-0.056	$+0.000$	-0.022
	CLIP	-0.535	-0.328	-0.663
Banking77	MiniLM	$+0.643$	$+1.334$	$+1.195$
	MPNet	$+0.247$	$+0.617$	$+0.451$
20NG	MiniLM	$+0.894$	$+2.586$	$+0.753$
	MPNet	$+0.835$	$+2.017$	$+0.690$

Table 41: RP4096 correction effects by setting, averaged over ten fresh seeds. All cases have 512 nodes and more than 512 fitting examples. The independently selected column allows separate validation-selected penalties and iteration counts; it is not the fixed-parameter intervention.

Dataset	Encoder	Primary effect	Reverse effect	Independently selected
CIFAR-100	DINOv2	-0.005	$+0.068$	$+0.029$
	CLIP	$+0.038$	$+0.058$	$+0.027$
Pets	DINOv2	$+0.008$	$+0.147$	$+0.071$
	CLIP	$+0.142$	$+1.259$	$+0.807$
Flowers-102	DINOv2	-0.000	-0.003	$+0.010$
	CLIP	$+0.063$	$+0.231$	$+0.046$
DTD	DINOv2	$+0.048$	$+0.755$	$+0.388$
	CLIP	$+0.021$	$+0.846$	$+0.245$
EuroSAT	DINOv2	-0.343	-0.030	-0.191
	CLIP	-0.433	-0.261	-0.439
Banking77	MiniLM	$+0.019$	$+0.231$	-0.003
	MPNet	$+0.029$	$+0.127$	$+0.049$
20NG	MiniLM	-0.049	$+1.348$	$+0.139$
	MPNet	-0.050	$+0.697$	$+0.200$

F.5 The earlier cardinality-based comparison

Tables 42 and 43 report calibrated and raw NLL for the earlier $m = 8192$ configurations. The two RP methods use identical projections and data; each method selects its own penalty and, where applicable, iteration count on validation data.

Table 42: Temperature-calibrated test NLL in nats per example for the earlier cardinality-based comparison; lower is better. Native and RP QUILT-C use $m = 8192$ with validation-selected $J \leq 256$ and $J \leq 512$, respectively. ER uses 8192 stored examples. Entries give mean ± sample standard deviation over seeds; the last row uses seed-level equal-setting means. Calibration, when applied, is fitted separately on validation data.

		Native encoder features			4096 RP features	
Dataset	Encoder	Ridge	ER	QUILT-C	Ridge	QUILT-C
CIFAR-100	DINOv2	0.741 ±0.000	0.592 ±0.005	0.502 ±0.002	0.631 ±0.005	0.530 ±0.035
	CLIP	1.094 ±0.000	0.886 ±0.007	0.737 ±0.001	0.976 ±0.001	0.720 ±0.002
Pets	DINOv2	0.247 ±0.000	0.514 ±0.141	0.176 ±0.000	0.245 ±0.005	0.193 ±0.028
	CLIP	0.466 ±0.000	0.641 ±0.135	0.393 ±0.000	0.394 ±0.037	0.363 ±0.010
Flowers-102	DINOv2	0.050 ±0.000	0.574 ±0.072	0.037 ±0.000	0.043 ±0.001	0.038 ±0.001
	CLIP	0.294 ±0.000	1.525 ±0.171	0.203 ±0.000	0.242 ±0.003	0.205 ±0.004
DTD	DINOv2	0.949 ±0.000	1.007 ±0.051	0.729 ±0.000	0.889 ±0.006	0.862 ±0.033
	CLIP	1.176 ±0.000	1.212 ±0.040	0.929 ±0.000	1.123 ±0.040	0.997 ±0.055
EuroSAT	DINOv2	0.169 ±0.000	0.164 ±0.016	0.118 ±0.001	0.125 ±0.004	0.110 ±0.003
	CLIP	0.167 ±0.000	0.220 ±0.023	0.128 ±0.002	0.139 ±0.001	0.119 ±0.000
Banking77	MiniLM	0.484 ±0.000	0.354 ±0.013	0.285 ±0.007	0.346 ±0.008	0.289 ±0.009
	MPNet	0.345 ±0.000	0.319 ±0.004	0.262 ±0.002	0.304 ±0.009	0.248 ±0.016
20NG	MiniLM	1.184 ±0.000	1.287 ±0.017	1.148 ±0.010	1.197 ±0.043	1.131 ±0.018
	MPNet	1.062 ±0.000	1.160 ±0.020	1.019 ±0.005	1.071 ±0.033	1.032 ±0.019
Mean		0.602 ±0.000	0.747 ±0.034	0.476 ±0.000	0.552 ±0.005	0.488 ±0.007

Table 43: Raw test NLL in nats per example for the earlier cardinality-based comparison; lower is better. Native and RP QUILT-C use $m = 8192$ with validation-selected $J \leq 256$ and $J \leq 512$, respectively. ER uses 8192 stored examples. Entries give mean ± sample standard deviation over seeds; the last row uses seed-level equal-setting means. Calibration, when applied, is fitted separately on validation data.

		Native encoder features			4096 RP features	
Dataset	Encoder	Ridge	ER	QUILT-C	Ridge	QUILT-C
CIFAR-100	DINOv2	4.185 ±0.000	0.593 ±0.004	0.554 ±0.007	4.006 ±0.000	0.644 ±0.095
	CLIP	4.219 ±0.000	0.994 ±0.052	0.770 ±0.001	4.099 ±0.000	0.756 ±0.008
Pets	DINOv2	2.991 ±0.000	0.621 ±0.094	0.225 ±0.000	2.837 ±0.003	0.290 ±0.046
	CLIP	3.105 ±0.000	0.656 ±0.118	0.402 ±0.000	3.079 ±0.048	0.441 ±0.068
Flowers-102	DINOv2	4.145 ±0.000	1.278 ±0.583	0.113 ±0.000	3.798 ±0.001	0.092 ±0.026
	CLIP	4.268 ±0.000	3.253 ±0.060	0.558 ±0.000	4.068 ±0.002	0.604 ±0.107
DTD	DINOv2	3.418 ±0.000	1.284 ±0.025	0.786 ±0.000	3.295 ±0.000	0.936 ±0.044
	CLIP	3.408 ±0.000	1.212 ±0.039	0.945 ±0.000	3.373 ±0.041	1.049 ±0.137
EuroSAT	DINOv2	1.658 ±0.000	0.164 ±0.016	0.163 ±0.001	1.588 ±0.006	0.180 ±0.002
	CLIP	1.638 ±0.000	0.221 ±0.023	0.187 ±0.018	1.596 ±0.006	0.159 ±0.011
Banking77	MiniLM	3.836 ±0.000	0.386 ±0.035	0.313 ±0.024	3.562 ±0.006	0.300 ±0.023
	MPNet	3.737 ±0.000	0.433 ±0.060	0.272 ±0.014	3.551 ±0.005	0.263 ±0.012
20NG	MiniLM	2.609 ±0.000	1.575 ±0.108	1.147 ±0.006	2.562 ±0.022	1.131 ±0.010
	MPNet	2.555 ±0.000	1.466 ±0.133	1.006 ±0.004	2.524 ±0.021	1.028 ±0.013
Mean		3.269 ±0.000	1.010 ±0.066	0.532 ±0.003	3.139 ±0.004	0.562 ±0.008

F.6 Larger memories and iteration grids

Tables 44–46 report every requested larger-memory and iteration-grid comparison. Different grids define separate validation-selection procedures, so their test-accuracy differences are not estimates of the effect of changing J at a fixed penalty. Native $m = 16384$ retains all fitting features in 12 of the 14 settings; CIFAR-100 remains compressed. At $m = 8192$, both native and RP summaries retain all fitting features on Pets, Flowers-102, and DTD. These saturated cases are complete-feature-memory CE readouts rather than evidence for the covariance correction under compression.

Table 44: Native-feature QUILT-C accuracy (%) for every larger budget, with original ($J \leq 64$) and expanded ($J \leq 256$) validation grids. Both grids select their own regularization and iteration count. Actual node counts may saturate at the training-set size; identical saturated runs are reused explicitly. All entries show mean ± sample standard deviation over three seeds.

		$m = 4096$		$m = 8192$		$m = 16384$	
Dataset	Encoder	$J \leq 64$	$J \leq 256$	$J \leq 64$	$J \leq 256$	$J \leq 64$	$J \leq 256$
CIFAR-100	DINOv2	84.49 ±0.09	84.67 ±0.13	84.78 ±0.24	85.03 ±0.12	85.27 ±0.06	85.49 ±0.17
	CLIP	78.19 ±0.07	78.60 ±0.03	78.39 ±0.08	78.90 ±0.04	78.87 ±0.09	79.42 ±0.05
Pets	DINOv2	94.39 ±0.00	94.90 ±0.00	94.39 ±0.00	94.90 ±0.00	94.39 ±0.00	94.90 ±0.00
	CLIP	88.69 ±0.00	87.98 ±0.00	88.69 ±0.00	87.98 ±0.00	88.69 ±0.00	87.98 ±0.00
Flowers-102	DINOv2	99.53 ±0.00	99.54 ±0.00	99.53 ±0.00	99.54 ±0.00	99.53 ±0.00	99.54 ±0.00
	CLIP	94.36 ±0.00	94.36 ±0.00	94.36 ±0.00	94.36 ±0.00	94.36 ±0.00	94.36 ±0.00
DTD	DINOv2	78.67 ±0.00	78.88 ±0.00	78.67 ±0.00	78.88 ±0.00	78.67 ±0.00	78.88 ±0.00
	CLIP	74.36 ±0.00	74.20 ±0.00	74.36 ±0.00	74.20 ±0.00	74.36 ±0.00	74.20 ±0.00
EuroSAT	DINOv2	95.63 ±0.10	95.64 ±0.10	95.90 ±0.09	95.94 ±0.07	95.93 ±0.00	95.93 ±0.00
	CLIP	95.48 ±0.11	95.51 ±0.10	95.67 ±0.06	95.67 ±0.06	95.81 ±0.00	95.81 ±0.00
Banking77	MiniLM	92.41 ±0.24	92.69 ±0.20	92.34 ±0.12	92.49 ±0.07	92.40 ±0.00	92.53 ±0.00
	MPNet	93.32 ±0.11	93.55 ±0.21	93.37 ±0.08	93.37 ±0.07	93.44 ±0.00	93.44 ±0.00
20NG	MiniLM	67.67 ±0.27	67.69 ±0.23	67.71 ±0.23	67.71 ±0.23	67.64 ±0.00	67.64 ±0.00
	MPNet	71.42 ±0.18	71.44 ±0.21	71.19 ±0.08	71.19 ±0.08	71.81 ±0.00	71.81 ±0.00
Mean		86.33 ±0.03	86.40 ±0.02	86.38 ±0.04	86.44 ±0.02	86.51 ±0.00	86.57 ±0.01

Table 45: Accuracy (%) of QUILT-C on the archived RP-ridge feature maps. All budgets and both validation grids are reported; the RP performance configuration is $m = 8192$ with $J \leq 512$. Entries give three-seed means and sample standard deviations.

		$m = 2048$		$m = 8192$	
Dataset	Encoder	$J \leq 64$	$J \leq 512$	$J \leq 64$	$J \leq 512$
CIFAR-100	DINOv2	85.90 ±0.02	85.90 ±0.02	86.10 ±0.04	86.02 ±0.12
	CLIP	78.80 ±0.13	78.80 ±0.13	79.26 ±0.08	79.55 ±0.22
Pets	DINOv2	94.72 ±0.13	94.67 ±0.11	94.69 ±0.12	94.66 ±0.07
	CLIP	88.61 ±0.37	88.60 ±0.44	88.55 ±0.11	88.63 ±0.07
Flowers-102	DINOv2	99.58 ±0.04	99.56 ±0.00	99.58 ±0.04	99.56 ±0.00
	CLIP	94.39 ±0.12	94.39 ±0.12	94.39 ±0.12	94.39 ±0.12
DTD	DINOv2	78.62 ±0.61	78.62 ±0.61	78.62 ±0.61	78.62 ±0.61
	CLIP	74.31 ±0.65	74.15 ±0.52	74.31 ±0.65	74.15 ±0.52
EuroSAT	DINOv2	96.22 ±0.16	96.22 ±0.16	96.46 ±0.16	96.46 ±0.16
	CLIP	95.83 ±0.02	95.81 ±0.00	95.89 ±0.17	95.88 ±0.18
Banking77	MiniLM	93.33 ±0.18	93.32 ±0.20	93.27 ±0.21	93.21 ±0.14
	MPNet	94.03 ±0.17	93.92 ±0.10	93.96 ±0.06	93.86 ±0.17
20NG	MiniLM	68.31 ±0.15	68.26 ±0.14	68.40 ±0.41	68.39 ±0.42
	MPNet	71.92 ±0.38	71.90 ±0.34	71.96 ±0.45	71.51 ±0.18
Mean		86.75 ±0.05	86.72 ±0.07	86.82 ±0.06	86.78 ±0.07

Table 46: Reservoir ER accuracy (%) as the memory budget grows. All other optimizer and validation-grid rules are held fixed. The 512-example column is the archived baseline; larger budgets are separately executed runs. Entries give three-seed means and sample standard deviations.

Dataset	Encoder	$m = 512$	$m = 4096$	$m = 8192$	$m = 16384$
CIFAR-100	DINOv2	76.26 ±0.29	81.29 ±0.12	82.42 ±0.21	83.41 ±0.52
	CLIP	60.28 ±1.41	72.73 ±0.45	74.33 ±0.17	75.41 ±0.20
Pets	DINOv2	84.45 ±3.39	85.18 ±3.34	85.18 ±3.34	85.18 ±3.34
	CLIP	78.66 ±1.30	80.70 ±1.87	80.70 ±1.87	80.70 ±1.87
Flowers-102	DINOv2	85.61 ±2.87	85.44 ±2.38	85.44 ±2.38	85.44 ±2.38
	CLIP	72.60 ±3.15	74.07 ±3.23	74.07 ±3.23	74.07 ±3.23
DTD	DINOv2	70.87 ±0.85	71.26 ±0.63	71.26 ±0.63	71.26 ±0.63
	CLIP	65.43 ±0.23	67.64 ±0.67	67.64 ±0.67	67.64 ±0.67
EuroSAT	DINOv2	91.09 ±0.90	93.64 ±0.50	94.30 ±0.61	93.99 ±0.62
	CLIP	89.65 ±1.38	92.29 ±0.64	92.46 ±0.75	92.29 ±1.26
Banking77	MiniLM	83.38 ±1.45	90.22 ±0.29	90.71 ±0.59	90.67 ±0.52
	MPNet	86.21 ±0.26	91.02 ±0.25	91.53 ±0.46	91.48 ±0.44
20NG	MiniLM	59.59 ±1.69	61.94 ±1.34	62.38 ±0.59	62.75 ±0.91
	MPNet	67.00 ±0.19	67.88 ±0.83	66.94 ±0.67	67.02 ±0.75
Mean		76.51 ±0.74	79.66 ±0.75	79.95 ±0.80	80.09 ±0.85

Table 47: Retained state for the earlier cardinality-based comparison in MB (10^6 bytes). Native and RP QUILT-C use $m = 8192$; ER uses an 8192-example reservoir. Projection storage is included for both RP methods. Shared encoders, evaluator caches and snapshots, and temporary training workspaces are excluded. Scalar counter conventions do not affect the displayed precision.

		Native encoder features			4096 RP features	
Dataset	Encoder	Ridge	ER	QUILT-C	Ridge	QUILT-C
CIFAR-100	DINOv2	1.80 ±0.00	26.22 ±0.00	27.16 ±0.00	153.42 ±0.00	422.05 ±0.00
	CLIP	2.93 ±0.00	34.92 ±0.00	36.68 ±0.00	157.62 ±0.00	426.25 ±0.00
Pets	DINOv2	1.41 ±0.00	9.43 ±0.00	10.53 ±0.00	149.29 ±0.00	245.83 ±0.00
	CLIP	2.41 ±0.00	12.56 ±0.00	14.54 ±0.00	153.49 ±0.00	250.03 ±0.00
Flowers-102	DINOv2	1.81 ±0.00	4.09 ±0.00	4.97 ±0.00	153.55 ±0.00	187.00 ±0.00
	CLIP	2.94 ±0.00	5.45 ±0.00	7.14 ±0.00	157.75 ±0.00	191.19 ±0.00
DTD	DINOv2	1.48 ±0.00	6.24 ±0.00	7.30 ±0.00	149.95 ±0.00	211.60 ±0.00
	CLIP	2.49 ±0.00	8.31 ±0.00	10.24 ±0.00	154.14 ±0.00	215.79 ±0.00
EuroSAT	DINOv2	1.25 ±0.00	25.39 ±0.00	26.61 ±0.00	147.52 ±0.00	416.15 ±0.00
	CLIP	2.19 ±0.00	33.81 ±0.00	35.94 ±0.00	151.72 ±0.00	420.35 ±0.00
Banking77	MiniLM	1.66 ±0.00	26.01 ±0.00	27.02 ±0.00	151.91 ±0.00	420.55 ±0.00
	MPNet	5.68 ±0.00	51.88 ±0.00	56.21 ±0.00	164.50 ±0.00	433.13 ±0.00
20NG	MiniLM	1.31 ±0.00	25.48 ±0.00	26.63 ±0.00	148.18 ±0.00	416.43 ±0.02
	MPNet	4.98 ±0.00	50.83 ±0.00	55.44 ±0.00	160.76 ±0.00	429.06 ±0.00
Mean		2.45 ±0.00	22.90 ±0.00	24.74 ±0.00	153.84 ±0.00	334.67 ±0.00

Table 48: Actual node counts at requested $m = 8192$, full-feature-history flags, and dense head inference work per example. Y means every training feature is retained. Ranges, if present, span the three seeds. Native methods use $2(d+1)v$ FLOPs; both RP methods use $2d \cdot 4096 + 2 \cdot 4097v$, including the fixed projection. ReLU, normalization, and common encoder work are excluded; these are arithmetic estimates, not latency measurements.

			Actual nodes		Full history	Head MFLOPs	
Dataset	Encoder	n	Native	RP	Native/RP	Native	RP
CIFAR-100	DINOv2	40000	8192	8192	N/N	0.077	3.965
	CLIP	40000	8192	8192	N/N	0.103	5.014
Pets	DINOv2	2944	2944	2944	Y/Y	0.028	3.449
	CLIP	2944	2944	2944	Y/Y	0.038	4.497
Flowers-102	DINOv2	1020	1020	1020	Y/Y	0.079	3.982
	CLIP	1020	1020	1020	Y/Y	0.105	5.030
DTD	DINOv2	1880	1880	1880	Y/Y	0.036	3.531
	CLIP	1880	1880	1880	Y/Y	0.048	4.579
EuroSAT	DINOv2	16200	8192	8192	N/N	0.008	3.228
	CLIP	16200	8192	8192	N/N	0.010	4.276
Banking77	MiniLM	9000	8192	8192	N/N	0.059	3.777
	MPNet	9000	8192	8192	N/N	0.118	6.922
20NG	MiniLM	10184	8180	8180–8181	N/N	0.015	3.310
	MPNet	10184	8182	8182	N/N	0.031	6.455

F.7 Compact native-feature study and offline reference

The original compact study uses the smaller $J \in \{16, 32, 64\}$ grid and fixed cardinalities rather than common byte caps. Its offline joint CE reference uses native features, so it does not describe the optimum in the RP feature space.

Table 49: Native-feature offline reference, accuracy (%), three-seed means. † uses the complete fitting history and is outside the retained-memory comparison; it is not a certified optimum. No RP offline optimum is inferred from this table.

Dataset	Encoder	Ridge	Q-C, $m = 512$	Q-C, $m = 8192$	Joint CE†
CIFAR-100	DINOv2	81.44	83.66	85.03	86.06
	CLIP	73.80	77.04	78.90	78.93
Pets	DINOv2	93.79	94.57	94.90	94.74
	CLIP	87.74	89.51	87.98	88.69
Flowers-102	DINOv2	99.35	99.57	99.54	99.58
	CLIP	92.37	94.66	94.36	93.20
DTD	DINOv2	75.90	78.51	78.88	78.94
	CLIP	70.80	74.33	74.20	74.41
EuroSAT	DINOv2	94.74	95.20	95.94	95.87
	CLIP	94.54	95.09	95.67	95.70
Banking77	MiniLM	88.31	91.67	92.49	92.95
	MPNet	92.01	93.28	93.37	93.51
20NG	MiniLM	67.72	67.26	67.71	67.96
	MPNet	71.57	71.41	71.19	70.90

Probabilities and memory budgets. Tables 50 and 51 distinguish validation-temperature calibration from raw classifier confidence. All temperatures are fitted after the accuracy-based configuration choice and before test evaluation. Tables 52–54 report the compact budget study, including the centroid-metric variant. Its hyperparameters are independently selected; the matched experiment below removes that difference. At $m = 2048$, both Flowers-102 and DTD fit entirely in memory.

Table 50: Temperature-calibrated test cross-entropy (nats per example); lower is better. Each temperature is fitted only on validation data. Three seeds, $m = 512$ for centroid methods.

Dataset	Encoder	Ridge	RP ridge	ER	QUILT-0	QUILT-C	Joint CE†
CIFAR-100	DINOv2	0.741 ±0.000	0.631 ±0.005	0.819 ±0.024	0.604 ±0.001	0.586 ±0.008	0.469 ±0.000
	CLIP	1.094 ±0.000	0.976 ±0.001	1.446 ±0.021	0.877 ±0.007	0.845 ±0.008	0.702 ±0.000
Pets	DINOv2	0.247 ±0.000	0.245 ±0.005	0.514 ±0.134	0.211 ±0.009	0.197 ±0.006	0.183 ±0.000
	CLIP	0.466 ±0.000	0.394 ±0.037	0.744 ±0.094	0.386 ±0.041	0.324 ±0.002	0.364 ±0.000
Flowers-102	DINOv2	0.050 ±0.000	0.043 ±0.001	0.579 ±0.072	0.038 ±0.000	0.038 ±0.000	0.037 ±0.000
	CLIP	0.294 ±0.000	0.242 ±0.003	1.555 ±0.151	0.210 ±0.002	0.202 ±0.005	0.254 ±0.000
DTD	DINOv2	0.949 ±0.000	0.889 ±0.006	1.036 ±0.044	0.781 ±0.009	0.777 ±0.034	0.726 ±0.000
	CLIP	1.176 ±0.000	1.123 ±0.040	1.300 ±0.037	0.990 ±0.020	0.933 ±0.005	0.917 ±0.000
EuroSAT	DINOv2	0.169 ±0.000	0.125 ±0.004	0.268 ±0.017	0.150 ±0.002	0.145 ±0.002	0.120 ±0.000
	CLIP	0.167 ±0.000	0.139 ±0.001	0.304 ±0.034	0.140 ±0.001	0.146 ±0.001	0.124 ±0.000
Banking77	MiniLM	0.484 ±0.000	0.346 ±0.008	0.600 ±0.063	0.370 ±0.012	0.350 ±0.014	0.265 ±0.000
	MPNet	0.345 ±0.000	0.304 ±0.009	0.488 ±0.007	0.314 ±0.004	0.291 ±0.004	0.230 ±0.000
20NG	MiniLM	1.184 ±0.000	1.197 ±0.043	1.378 ±0.045	1.199 ±0.013	1.118 ±0.006	1.085 ±0.000
	MPNet	1.062 ±0.000	1.071 ±0.033	1.153 ±0.020	1.080 ±0.031	1.028 ±0.008	1.009 ±0.000

Table 51: Raw test cross-entropy before temperature calibration. These retain the different confidence scales of ridge and logistic objectives.

Dataset	Encoder	Ridge	RP ridge	ER	QUILT-0	QUILT-C	Joint CE†
CIFAR-100	DINOv2	4.185 ±0.000	4.006 ±0.000	1.177 ±0.008	0.715 ±0.104	0.675 ±0.022	0.471 ±0.000
	CLIP	4.219 ±0.000	4.099 ±0.000	1.460 ±0.025	0.908 ±0.002	0.906 ±0.033	0.755 ±0.000
Pets	DINOv2	2.991 ±0.000	2.837 ±0.003	0.615 ±0.077	0.271 ±0.014	0.275 ±0.000	0.180 ±0.000
	CLIP	3.105 ±0.000	3.079 ±0.048	0.749 ±0.090	0.507 ±0.070	0.678 ±0.023	0.364 ±0.000
Flowers-102	DINOv2	4.145 ±0.000	3.798 ±0.001	1.284 ±0.560	0.335 ±0.001	0.348 ±0.001	0.045 ±0.000
	CLIP	4.268 ±0.000	4.068 ±0.002	3.247 ±0.064	0.608 ±0.065	0.642 ±0.060	0.319 ±0.000
DTD	DINOv2	3.418 ±0.000	3.295 ±0.000	1.303 ±0.020	0.843 ±0.021	0.857 ±0.004	0.779 ±0.000
	CLIP	3.408 ±0.000	3.373 ±0.041	1.303 ±0.040	1.057 ±0.050	1.011 ±0.004	0.918 ±0.000
EuroSAT	DINOv2	1.658 ±0.000	1.588 ±0.006	0.375 ±0.010	0.181 ±0.020	0.256 ±0.009	0.126 ±0.000
	CLIP	1.638 ±0.000	1.596 ±0.006	0.312 ±0.030	0.194 ±0.001	0.238 ±0.011	0.127 ±0.000
Banking77	MiniLM	3.836 ±0.000	3.562 ±0.006	0.682 ±0.094	0.424 ±0.025	0.424 ±0.022	0.310 ±0.000
	MPNet	3.737 ±0.000	3.551 ±0.005	0.515 ±0.006	0.382 ±0.043	0.358 ±0.013	0.277 ±0.000
20NG	MiniLM	2.609 ±0.000	2.562 ±0.022	1.398 ±0.055	1.193 ±0.016	1.118 ±0.004	1.097 ±0.000
	MPNet	2.555 ±0.000	2.524 ±0.021	1.588 ±0.005	1.080 ±0.036	1.039 ±0.013	1.013 ±0.000

Table 52: Centroid-budget comparison at $m = 128$. Each method selects its own penalty and iteration count on validation data; these are not matched-parameter ablations.

Dataset	Encoder	Ridge	$G_{\mathcal{Q}}$ metric	QUILT-0	QUILT-C
CIFAR-100	DINOv2	81.44 ±0.00	80.03 ±0.18	82.03 ±0.02	82.03 ±0.02
	CLIP	73.80 ±0.00	69.19 ±0.43	75.32 ±0.03	75.37 ±0.02
Pets	DINOv2	93.79 ±0.00	93.45 ±0.28	93.68 ±0.24	94.18 ±0.46
	CLIP	87.74 ±0.00	82.78 ±0.56	88.13 ±0.16	89.36 ±0.79
Flowers-102	DINOv2	99.35 ±0.00	99.35 ±0.01	99.37 ±0.00	99.44 ±0.01
	CLIP	92.37 ±0.00	90.29 ±0.07	93.62 ±0.00	94.04 ±0.02
DTD	DINOv2	75.90 ±0.00	75.23 ±0.78	76.38 ±0.37	76.84 ±0.49
	CLIP	70.80 ±0.00	68.12 ±0.08	71.61 ±0.20	71.93 ±0.64
EuroSAT	DINOv2	94.74 ±0.00	92.43 ±0.16	95.09 ±0.10	95.08 ±0.02
	CLIP	94.54 ±0.00	91.09 ±0.29	95.24 ±0.08	94.96 ±0.07
Banking77	MiniLM	88.31 ±0.00	85.27 ±0.26	89.58 ±0.06	90.08 ±0.46
	MPNet	92.01 ±0.00	87.65 ±0.10	92.37 ±0.06	92.80 ±0.05
20NG	MiniLM	67.72 ±0.00	65.52 ±0.52	66.44 ±0.42	67.27 ±0.33
	MPNet	71.57 ±0.00	69.29 ±0.71	70.24 ±0.11	70.96 ±0.06

Table 53: Centroid-budget comparison at $m = 512$. Each method selects its own penalty and iteration count on validation data; these are not matched-parameter ablations.

Dataset	Encoder	Ridge	$G_{\mathcal{Q}}$ metric	QUILT-0	QUILT-C
CIFAR-100	DINOv2	81.44 ±0.00	82.55 ±0.15	83.49 ±0.17	83.75 ±0.10
	CLIP	73.80 ±0.00	73.91 ±0.04	76.78 ±0.04	76.92 ±0.09
Pets	DINOv2	93.79 ±0.00	93.82 ±0.10	94.06 ±0.22	94.37 ±0.06
	CLIP	87.74 ±0.00	86.44 ±0.10	88.37 ±0.26	88.97 ±0.03
Flowers-102	DINOv2	99.35 ±0.00	99.50 ±0.02	99.50 ±0.02	99.51 ±0.02
	CLIP	92.37 ±0.00	93.54 ±0.14	94.29 ±0.15	94.47 ±0.02
DTD	DINOv2	75.90 ±0.00	77.75 ±0.03	78.21 ±0.31	78.53 ±0.25
	CLIP	70.80 ±0.00	71.45 ±0.81	73.39 ±0.29	74.06 ±0.21
EuroSAT	DINOv2	94.74 ±0.00	93.86 ±0.18	95.38 ±0.14	95.20 ±0.01
	CLIP	94.54 ±0.00	93.22 ±0.39	95.60 ±0.10	95.09 ±0.09
Banking77	MiniLM	88.31 ±0.00	89.39 ±0.36	90.58 ±0.06	91.22 ±0.28
	MPNet	92.01 ±0.00	91.29 ±0.37	92.91 ±0.10	92.86 ±0.14
20NG	MiniLM	67.72 ±0.00	66.48 ±0.05	66.63 ±0.09	67.32 ±0.37
	MPNet	71.57 ±0.00	69.99 ±0.02	70.33 ±0.32	71.34 ±0.11

Table 54: Centroid-budget comparison at $m = 2048$. Each method selects its own penalty and iteration count on validation data; these are not matched-parameter ablations.

Dataset	Encoder	Ridge	$G_{\mathcal{Q}}$ metric	QUILT-0	QUILT-C
CIFAR-100	DINOv2	81.44 ±0.00	83.48 ±0.03	84.09 ±0.24	84.23 ±0.12
	CLIP	73.80 ±0.00	76.20 ±0.09	77.49 ±0.18	77.88 ±0.02
Pets	DINOv2	93.79 ±0.00	94.43 ±0.15	94.45 ±0.14	94.59 ±0.12
	CLIP	87.74 ±0.00	88.51 ±0.11	88.56 ±0.19	88.78 ±0.10
Flowers-102	DINOv2	99.35 ±0.00	99.53 ±0.00	99.53 ±0.00	99.53 ±0.00
	CLIP	92.37 ±0.00	94.36 ±0.00	94.36 ±0.00	94.36 ±0.00
DTD	DINOv2	75.90 ±0.00	78.67 ±0.00	78.67 ±0.00	78.67 ±0.00
	CLIP	70.80 ±0.00	74.36 ±0.00	74.36 ±0.00	74.36 ±0.00
EuroSAT	DINOv2	94.74 ±0.00	94.80 ±0.08	95.43 ±0.08	95.44 ±0.09
	CLIP	94.54 ±0.00	95.10 ±0.04	95.76 ±0.06	95.28 ±0.10
Banking77	MiniLM	88.31 ±0.00	91.52 ±0.36	91.77 ±0.14	91.85 ±0.09
	MPNet	92.01 ±0.00	92.64 ±0.07	92.99 ±0.06	93.10 ±0.10
20NG	MiniLM	67.72 ±0.00	67.29 ±0.32	67.07 ±0.29	67.49 ±0.30
	MPNet	71.57 ±0.00	70.90 ±0.11	71.06 ±0.19	71.39 ±0.08

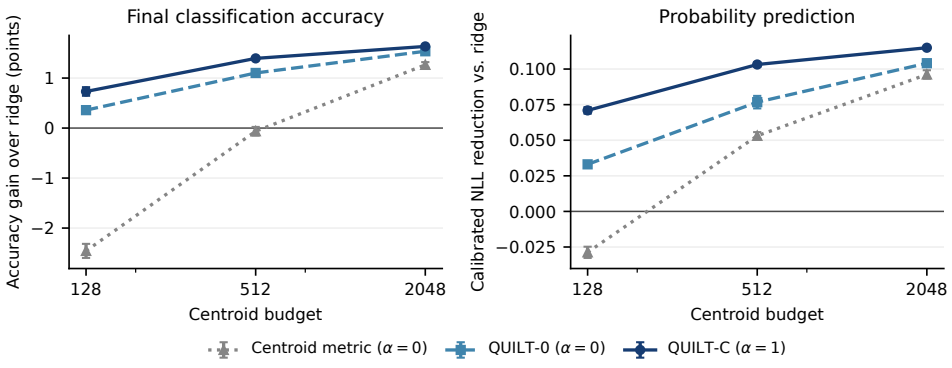


Figure 5: The original three-budget study on native features. Each variant selects its own regularization and iteration count; the paired parameter interventions below address that distinction. Points average the 14 settings, and error bars show the sample standard deviation of three seed-level averages.

F.8 MATCHED-PARAMETER EXPERIMENTS

For each of the 126 core runs, we freeze the uncorrected original-Gram model's validation-selected λ, J, ridge initialization, feature cache, and centroid summary. We evaluate the archived control readout, a readout changing only the metric to $G_{\mathcal{Q}}$, and a readout changing only α to one. Each alternative fits its own scalar temperature on validation examples. No configuration is selected from test performance. The reference readouts are bitwise identical to the archived controls; an independent recomputation agrees to relative error at most 1.61×10^{-12}.

Across all budgets, using $G_{\mathcal{Q}}$ reduces normalized F_0 by 0.03356 but increases the original regularized historical objective by 0.15924. In this original native-feature study, correction improves mean accuracy and calibrated NLL while raw NLL increases. The later RP interventions in Appendix F.4 do not show the same accuracy benefit. The distinction between classification and confidence is consistent with the Gaussian theorem: recovering useful classification geometry does not fix confidence scale.

Table 55: Matched interventions at the original QUILT-0 configuration. Positive accuracy changes are better; negative NLL changes are better. Values are paired means over 14 settings and three seeds. The covariance correction improves calibrated NLL while increasing raw NLL.

	Only $G \to G_{\mathcal{Q}}$			Only $\alpha : 0 \to 1$		
m	Acc. Δ	Raw NLL Δ	Cal. NLL Δ	Acc. Δ	Raw NLL Δ	Cal. NLL Δ
128	-7.44	$+0.336$	$+0.348$	$+0.41$	$+0.026$	-0.028
512	-3.20	$+0.157$	$+0.155$	$+0.32$	$+0.022$	-0.022
2048	-0.88	$+0.047$	$+0.050$	$+0.14$	$+0.015$	-0.011

F.9 CONTINUAL EVALUATION

For each arrival t, seen-pooled accuracy evaluates the union of the test examples belonging to classes observed by t, weighted by their example counts. Prediction still uses the full output space, without a task-specific mask. Average incremental accuracy (AIA) averages these values over the ten arrivals. Forgetting averages $\max_{j \le i < T} R_{ij} - R_{Tj}$ over old tasks $j < T$. It can be negative; low forgetting alone can reflect poor initial learning. Final worst-task accuracy is $\min_j R_{Tj}$.

The 42 fixed-configuration baseline refits recover all 126 original endpoint accuracies and raw NLLs exactly. Core AIA is recomputed from its saved task matrices using the same test-task sample counts. Its stored all-test accuracy trajectory includes future classes and is not used as AIA. The new native S_{512} configuration has 42 additional fixed-parameter trajectory reconstructions. Their final accuracies agree exactly with the selected main-table models, with readout differences below 10^{-8}. The following aggregates compare the primary configurations with the same validation-selected K-prior and CM variants as the main accuracy tables.

Table 56: Primary continual metrics, native features under S_{512}, averaged equally over all 14 setting means. Native QUILT-C trajectories reconstruct the exact new main-table configurations; endpoints are checked against their archived readouts. K-priors and CM use the same validation-selected variants as the corresponding accuracy comparison.

Method	AIA ↑	Forgetting ↓	Final worst task ↑
Ridge	89.56	5.49	74.05
K-priors	90.16	6.94	75.50
CM	87.54	12.53	66.90
QUILT-C	90.50	4.90	78.27

Table 57: Primary continual metrics, RP4096 features under S_{8192}, averaged equally over all 14 setting means. Native QUILT-C trajectories reconstruct the exact new main-table configurations; endpoints are checked against their archived readouts. K-priors and CM use the same validation-selected variants as the corresponding accuracy comparison.

Method	AIA ↑	Forgetting ↓	Final worst task ↑
Ridge	90.68	4.68	77.64
K-priors	90.69	5.30	78.10
CM	88.77	3.14	71.54
QUILT-C	90.87	4.58	78.67

For the primary RP QUILT-C model, 42 independent trajectory reconstructions copy the archived validation-selected configurations and temperatures without reselection. Their final readout coefficients agree with the corresponding primary models within 10^{-8}, and endpoint accuracies agree exactly. The following tables retain the earlier $m = 8192$ native and RP trajectories. They are distinct from the new native S_{512} main configurations.

Table 58: Average incremental accuracy for the earlier large-budget configurations: native QUILT-C $m = 8192, J \leq 256$ and RP QUILT-C $m = 8192, J \leq 512$. AIA and worst-task accuracy are percentages; forgetting is in percentage points. Entries show three-seed means and sample standard deviations; aggregate uncertainty uses seed-level equal-setting means. Ridge trajectories reuse their unchanged validation-selected configurations.

		Native encoder features		4096 RP features	
Dataset	Encoder	Ridge	QUILT-C	Ridge	QUILT-C
CIFAR-100	DINOv2	88.10 ±0.42	90.07 ±0.70	89.78 ±0.81	90.47 ±0.66
	CLIP	82.84 ±0.60	85.42 ±0.65	84.91 ±0.47	85.84 ±0.56
Pets	DINOv2	95.52 ±1.74	96.57 ±1.19	96.17 ±1.55	96.43 ±1.32
	CLIP	90.68 ±2.05	92.38 ±1.71	92.56 ±2.10	92.52 ±1.50
Flowers-102	DINOv2	99.61 ±0.12	99.71 ±0.11	99.71 ±0.11	99.70 ±0.09
	CLIP	95.34 ±0.86	96.29 ±0.58	96.06 ±0.71	96.22 ±0.62
DTD	DINOv2	83.80 ±2.67	85.20 ±2.56	84.96 ±2.71	84.88 ±2.48
	CLIP	78.69 ±3.67	81.54 ±3.26	80.76 ±3.19	81.63 ±3.41
EuroSAT	DINOv2	97.52 ±0.34	97.96 ±0.34	98.05 ±0.22	98.21 ±0.23
	CLIP	97.48 ±0.37	97.89 ±0.39	97.88 ±0.38	97.98 ±0.35
Banking77	MiniLM	92.55 ±0.29	94.74 ±0.12	94.98 ±0.29	95.25 ±0.16
	MPNet	94.71 ±0.14	95.41 ±0.30	95.85 ±0.18	95.90 ±0.01
20NG	MiniLM	77.08 ±0.99	76.43 ±1.27	77.69 ±1.25	77.40 ±0.55
	MPNet	79.95 ±0.93	79.36 ±1.08	80.15 ±1.40	79.77 ±0.94
Mean		89.56 ±0.82	90.64 ±0.70	90.68 ±0.82	90.87 ±0.66

Table 59: Forgetting for the earlier large-budget configurations: native QUILT-C $m = 8192, J \leq 256$ and RP QUILT-C $m = 8192, J \leq 512$. AIA and worst-task accuracy are percentages; forgetting is in percentage points. Entries show three-seed means and sample standard deviations; aggregate uncertainty uses seed-level equal-setting means. Ridge trajectories reuse their unchanged validation-selected configurations.

		Native encoder features		4096 RP features	
Dataset	Encoder	Ridge	QUILT-C	Ridge	QUILT-C
CIFAR-100	DINOv2	7.35 ±0.25	5.67 ±0.50	5.17 ±0.30	5.22 ±0.27
	CLIP	9.90 ±0.12	7.18 ±0.07	7.85 ±0.21	7.01 ±0.32
Pets	DINOv2	2.41 ±0.34	2.36 ±0.12	2.76 ±0.43	2.58 ±0.23
	CLIP	3.74 ±0.84	5.45 ±0.52	4.77 ±0.34	4.59 ±0.55
Flowers-102	DINOv2	0.20 ±0.18	0.14 ±0.14	0.18 ±0.10	0.13 ±0.13
	CLIP	3.10 ±0.24	2.01 ±0.34	2.09 ±0.18	1.88 ±0.42
DTD	DINOv2	8.14 ±0.38	7.13 ±1.36	7.01 ±1.13	6.88 ±1.39
	CLIP	8.31 ±0.40	7.54 ±0.45	8.07 ±0.44	7.83 ±0.56
EuroSAT	DINOv2	2.80 ±0.44	2.20 ±0.07	1.86 ±0.28	1.71 ±0.19
	CLIP	2.81 ±0.26	2.15 ±0.16	2.17 ±0.20	2.15 ±0.27
Banking77	MiniLM	4.82 ±0.36	2.67 ±0.10	2.28 ±0.43	2.35 ±0.25
	MPNet	3.42 ±0.43	2.39 ±0.11	2.16 ±0.30	2.18 ±0.16
20NG	MiniLM	10.47 ±2.41	9.89 ±1.63	10.22 ±1.89	10.30 ±1.44
	MPNet	9.42 ±2.10	8.96 ±1.11	8.99 ±1.39	9.30 ±1.21
Mean		5.49 ±0.34	4.70 ±0.30	4.68 ±0.30	4.58 ±0.34

Table 60: Final worst-task accuracy for the earlier large-budget configurations: native QUILT-C $m = 8192, J \leq 256$ and RP QUILT-C $m = 8192, J \leq 512$. AIA and worst-task accuracy are percentages; forgetting is in percentage points. Entries show three-seed means and sample standard deviations; aggregate uncertainty uses seed-level equal-setting means. Ridge trajectories reuse their unchanged validation-selected configurations.

		Native encoder features		4096 RP features	
Dataset	Encoder	Ridge	QUILT-C	Ridge	QUILT-C
CIFAR-100	DINOv2	75.23 ±1.83	80.57 ±1.62	80.50 ±1.21	81.57 ±1.31
	CLIP	66.53 ±3.56	74.80 ±1.47	72.03 ±3.09	75.70 ±1.22
Pets	DINOv2	84.62 ±2.52	88.38 ±0.70	86.70 ±0.91	86.79 ±0.70
	CLIP	77.35 ±2.54	78.94 ±1.90	78.19 ±2.60	78.68 ±0.97
Flowers-102	DINOv2	97.85 ±0.20	98.29 ±0.54	98.17 ±0.66	98.29 ±0.54
	CLIP	81.51 ±3.73	87.41 ±1.27	84.20 ±2.11	86.07 ±1.57
DTD	DINOv2	63.71 ±5.69	67.79 ±1.46	67.21 ±3.61	67.33 ±4.25
	CLIP	53.21 ±6.64	61.12 ±4.08	56.54 ±7.40	58.92 ±2.63
EuroSAT	DINOv2	87.60 ±0.00	91.27 ±0.42	91.20 ±0.35	91.53 ±0.23
	CLIP	86.60 ±0.00	90.47 ±0.12	90.00 ±0.20	91.00 ±0.80
Banking77	MiniLM	81.52 ±1.25	89.27 ±1.48	90.28 ±0.88	90.16 ±1.53
	MPNet	89.06 ±0.83	90.21 ±1.18	91.12 ±1.93	91.15 ±1.18
20NG	MiniLM	43.15 ±4.14	49.81 ±3.35	48.21 ±5.01	49.63 ±5.02
	MPNet	48.75 ±5.33	55.58 ±3.94	52.55 ±6.31	54.55 ±3.93
Mean		74.05 ±0.59	78.85 ±0.48	77.64 ±1.06	78.67 ±0.80

The original compact-memory study is retained below with explicit $m = 512$ labels; it uses different QUILT configurations from the main tables.

Table 61: Continual metrics, averaged equally over the 14 setting means. Three seeds per setting, $m = 512$. AIA and worst-task accuracy are percentages; forgetting is in percentage points.

Method	AIA ↑	Forgetting ↓	Final worst task ↑
Ridge	89.56	5.49	74.05
RP ridge	90.68	4.68	77.64
ER	82.61	13.25	51.57
Centroid metric	89.70	5.84	75.87
QUILT-0	90.29	5.20	77.76
QUILT-C	90.42	5.01	78.20

Table 62: Average incremental accuracy by setting, mean ± sample standard deviation across three seeds.

Dataset	Encoder	Ridge	RP ridge	ER	Centroid metric	QUILT-0	QUILT-C
CIFAR-100	DINOv2	88.10 ±0.42	89.78 ±0.81	84.53 ±0.70	88.42 ±0.66	89.07 ±0.70	89.20 ±0.64
	CLIP	82.84 ±0.60	84.91 ±0.47	75.64 ±0.30	82.30 ±0.69	84.38 ±0.76	84.51 ±0.78
Pets	DINOv2	95.52 ±1.74	96.17 ±1.55	90.94 ±0.84	95.88 ±1.53	96.08 ±1.28	96.06 ±1.61
	CLIP	90.68 ±2.05	92.56 ±2.10	86.64 ±2.29	91.55 ±1.83	92.43 ±1.59	92.62 ±1.86
Flowers-102	DINOv2	99.61 ±0.12	99.71 ±0.11	83.92 ±7.98	99.69 ±0.10	99.69 ±0.10	99.69 ±0.10
	CLIP	95.34 ±0.86	96.06 ±0.71	62.14 ±1.29	95.81 ±0.65	96.31 ±0.56	96.36 ±0.55
DTD	DINOv2	83.80 ±2.67	84.96 ±2.71	76.70 ±2.85	84.76 ±2.59	84.98 ±2.53	84.89 ±2.93
	CLIP	78.69 ±3.67	80.76 ±3.19	75.71 ±3.07	80.13 ±2.81	81.49 ±3.23	81.72 ±3.30
EuroSAT	DINOv2	97.52 ±0.34	98.05 ±0.22	95.38 ±0.35	97.14 ±0.45	97.82 ±0.31	97.70 ±0.29
	CLIP	97.48 ±0.37	97.88 ±0.38	94.82 ±0.54	96.69 ±0.51	97.88 ±0.41	97.58 ±0.30
Banking77	MiniLM	92.55 ±0.29	94.98 ±0.29	89.57 ±0.14	93.28 ±0.15	93.66 ±0.23	93.95 ±0.49
	MPNet	94.71 ±0.14	95.85 ±0.18	91.69 ±0.21	94.29 ±0.57	95.11 ±0.04	95.17 ±0.10
20NG	MiniLM	77.08 ±0.99	77.69 ±1.25	73.44 ±1.13	76.76 ±1.20	76.15 ±1.20	76.76 ±1.32
	MPNet	79.95 ±0.93	80.15 ±1.40	75.37 ±1.89	79.16 ±0.97	79.07 ±0.75	79.63 ±0.91

Table 63: Forgetting by setting, mean ± sample standard deviation across three seeds.

Dataset	Encoder	Ridge	RP ridge	ER	Centroid metric	QUILT-0	QUILT-C
CIFAR-100	DINOv2	7.35 ±0.25	5.17 ±0.30	16.02 ±0.67	6.96 ±0.37	6.47 ±0.28	6.40 ±0.17
	CLIP	9.90 ±0.12	7.85 ±0.21	35.57 ±2.08	9.35 ±0.40	8.46 ±0.05	8.48 ±0.19
Pets	DINOv2	2.41 ±0.34	2.76 ±0.43	14.17 ±4.31	3.09 ±0.25	2.87 ±0.20	2.67 ±0.22
	CLIP	3.74 ±0.84	4.77 ±0.34	15.99 ±1.59	5.96 ±0.50	4.64 ±0.24	4.13 ±0.53
Flowers-102	DINOv2	0.20 ±0.18	0.18 ±0.10	−1.67 ±15.16	0.20 ±0.13	0.20 ±0.12	0.20 ±0.11
	CLIP	3.10 ±0.24	2.09 ±0.18	−4.60 ±1.55	2.16 ±0.32	2.13 ±0.48	1.98 ±0.28
DTD	DINOv2	8.14 ±0.38	7.01 ±1.13	8.22 ±1.46	7.77 ±1.27	7.29 ±1.26	6.81 ±1.37
	CLIP	8.31 ±0.40	8.07 ±0.44	21.46 ±1.62	8.92 ±0.34	8.67 ±0.65	8.32 ±0.30
EuroSAT	DINOv2	2.80 ±0.44	1.86 ±0.28	9.05 ±0.99	3.57 ±0.32	2.46 ±0.10	2.54 ±0.27
	CLIP	2.81 ±0.26	2.17 ±0.20	10.44 ±1.87	3.65 ±0.45	2.33 ±0.18	2.44 ±0.05
Banking77	MiniLM	4.82 ±0.36	2.28 ±0.43	14.51 ±2.10	4.54 ±0.71	3.81 ±0.41	3.25 ±0.30
	MPNet	3.42 ±0.43	2.16 ±0.30	12.32 ±0.47	3.64 ±0.36	2.71 ±0.21	2.81 ±0.32
20NG	MiniLM	10.47 ±2.41	10.22 ±1.89	29.84 ±0.19	11.90 ±1.30	10.95 ±1.05	10.79 ±0.97
	MPNet	9.42 ±2.10	8.99 ±1.39	4.16 ±2.13	10.10 ±1.48	9.82 ±1.28	9.33 ±1.30

Table 64: Final worst-task accuracy by setting, mean ± sample standard deviation across three seeds.

Dataset	Encoder	Ridge	RP ridge	ER	Centroid metric	QUILT-0	QUILT-C
CIFAR-100	DINOv2	75.23 ±1.83	80.50 ±1.21	62.13 ±3.93	77.37 ±2.67	78.67 ±2.42	79.30 ±2.33
	CLIP	66.53 ±3.56	72.03 ±3.09	47.87 ±4.19	69.47 ±0.42	72.33 ±1.25	73.13 ±0.97
Pets	DINOv2	84.62 ±2.52	86.70 ±0.91	66.29 ±7.17	85.14 ±1.18	85.37 ±1.20	86.62 ±1.48
	CLIP	77.35 ±2.54	78.19 ±2.60	51.81 ±16.51	77.30 ±2.43	78.85 ±1.69	80.19 ±0.83
Flowers-102	DINOv2	97.85 ±0.20	98.17 ±0.66	23.89 ±35.47	98.29 ±0.54	98.29 ±0.54	98.29 ±0.54
	CLIP	81.51 ±3.73	84.20 ±2.11	0.00 ±0.00	84.01 ±1.88	86.83 ±1.52	87.05 ±1.82
DTD	DINOv2	63.71 ±5.69	67.21 ±3.61	44.54 ±14.01	68.17 ±2.08	68.50 ±2.18	69.79 ±1.54
	CLIP	53.21 ±6.64	56.54 ±7.40	52.54 ±7.61	57.67 ±3.40	59.96 ±3.06	60.54 ±2.11
EuroSAT	DINOv2	87.60 ±0.00	91.20 ±0.35	75.80 ±6.16	86.33 ±0.31	90.33 ±0.50	89.27 ±0.31
	CLIP	86.60 ±0.00	90.00 ±0.20	76.40 ±6.13	86.60 ±0.87	90.73 ±0.31	89.87 ±0.31
Banking77	MiniLM	81.52 ±1.25	90.28 ±0.88	73.54 ±4.33	84.90 ±2.35	86.46 ±2.43	87.29 ±2.39
	MPNet	89.06 ±0.83	91.12 ±1.93	78.23 ±2.35	87.71 ±1.26	90.31 ±1.43	90.00 ±0.83
20NG	MiniLM	43.15 ±4.14	48.21 ±5.01	34.22 ±8.25	46.88 ±4.97	48.79 ±4.34	48.95 ±3.98
	MPNet	48.75 ±5.33	52.55 ±6.31	34.78 ±8.49	52.40 ±4.09	53.16 ±3.74	54.53 ±4.18

F.10 Gaussian protocol and confidence scale

The synthetic study fixes $\mu = (0.5, 0.5)$, $\Sigma = \mathrm{diag}(2/(1+k), 2k/(1+k))$, $k \in \{1, 2, 4, 8, 16, 32, 64\}$, and ten seeds, 0–9. Each seed has 512 training and 20,000 independent test examples per class. Base normal draws are paired across the anisotropy grid within a seed. Training data alone determine centroids, their midpoint, and the residual covariance. There is no feature normalization. The mean-loss penalty is fixed at $\bar{\lambda} = 0.02$, equivalently $\lambda_{\mathrm{sum}} = 20.48$. Neither it nor $\alpha = 1$ is selected on these test outcomes. The Bayes reference uses the known generating distribution.

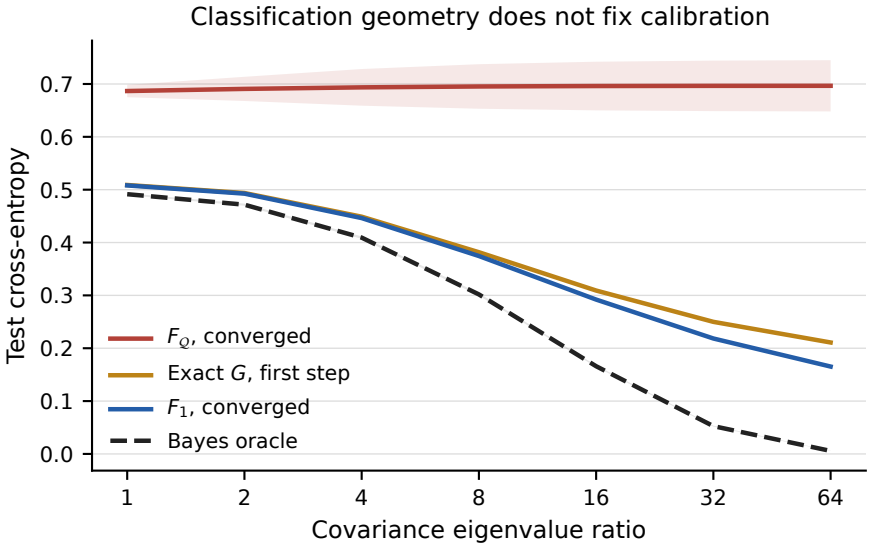


Figure 6: Uncalibrated confidence in the controlled Gaussian experiment. The corrected objective improves classification geometry but generally has a different logit scale from the Bayes predictor. These are raw test cross-entropies, unlike the temperature-calibrated primary NLL comparison.

G Verification and reproducible implementation

The expanded records contain 126 centroid-summary runs, 42 baseline trials, 126 matched-parameter comparisons, and 42 fixed-configuration retention refits. The accuracy-oriented exten-

sion adds 126 native-budget requests (90 distinct effective runs), 126 larger-buffer ER trials, 84 RP QUILT-C results, and 42 fixed-configuration RP retention reconstructions. The three seeds per setting are the experimental replicates; hyperparameter candidates and method variants are not counted as extra independent seeds. Feature extraction, model selection, and test evaluation are recorded separately, with source, model, cache, and result provenance.

The baseline audit checks all selected configurations against their validation grids, matches every ordinary-ridge result to the corresponding core run, and validates the reservoir implementation by independently refitting standard `torch.optim.AdamW`. The maximum coefficient discrepancy is 2.49×10^{-14}. All retention-refit endpoint accuracies and raw NLLs match the frozen records. The matched experiment additionally checks archived readout hashes and fixes the exact same moments and centroids for each comparison.

Independent NumPy/SciPy checks test the KL identity, local and global bounds, corrected gradient, and corrected-MM decrement on 168 bounded-feature probes and 2875 updates. No violation exceeds the stated tolerance. On a nonsymmetric three-class example, local error orders are 2.003 before correction and 2.982 after correction. These finite checks supplement the proofs. The earlier independent centroid checks cover 720 loss-interval probes and 8100 uncorrected-MM updates, and the original two executable ledger suites pass 26 and 21 checks.

The reusable library implementation retains only $G, H, \mathcal{Q}, B$ and fixed-size bookkeeping. Factors, residual matrices, and the current arrival are temporary. Evaluation checkpoints retained by the experimental harness are not learner state and never supply discarded examples to an update. The library uses the same update as the frozen experimental runner; its integration checks compare readouts and objective identities. All eight library integration tests pass, including fixed state size over 30 arrivals, independent autograd checks, agreement with the frozen benchmark solver, and checkpoint recovery.