WHEN DOES A SAMPLES-PER-CLASS CHOICE TRANSFER IN SUPERVISED CONTRASTIVE LEARNING?

Anonymous authors
Paper under double-blind review

ABSTRACT

Class-balanced contrastive batches trade class coverage for more samples per class. We examine when a preference between $K = 2$ and $K = 4$ remains valid across training budgets, objectives, and evaluation methods. In single-view CIFAR-100, native supervised contrastive learning favors $K = 4$ early and $K = 2$ late, including at a prespecified lower learning rate. On identical $K = 4$ batches, SupCon's existing L_{in} removes a within-positive KL term while retaining the shared denominator and improves late retrieval. Reused-seed extensions give both L_{in} and an averaged per-target objective their own learning-rate selection: they improve over native $K = 4$ by 4.11 and 3.64 Recall@1 points, but differ from $K = 2$ by only $+0.08$ and -0.40 points at the late hold-out head endpoint. The same alternatives differ by 2.44 points at a shorter budget. Paired head–encoder analyses and a post-hoc linear probe further separate objective improvement from the remaining K preference. Pretrained CUB and SOP extend the per-target objective improvement to class-disjoint retrieval, including development-selected stopping on CUB. These results show that the objective gain at a fixed batch composition and the evidence for choosing that composition are distinct, and must be evaluated at the intended budget and readout.

1 INTRODUCTION

Class-balanced sampling allocates a fixed source-image batch $B_{\text{src}} = GK$ to G classes and K samples per class. Increasing K provides more same-class partners while reducing class coverage per update. Class-balanced batches are used in metric learning (Hermans et al., 2017; Roth et al., 2020), where sampling affects performance (Wu et al., 2017); here we study them with supervised contrastive objectives (Khosla et al., 2020). A choice of K is therefore also a choice about the examples and positive relations used to train the representation. The relevant question is whether the resulting preference remains valid after changing how long the model is trained, how positives enter the loss, or which representation is deployed.

We investigate this question through three matched comparisons: native SupCon at $K = 4$ versus $K = 2$, a changed objective versus native SupCon on the same $K = 4$ batches, and the changed procedure versus $K = 2$. Their relationship is simple, but the experiments reveal materially different conclusions. A large improvement over native $K = 4$ can coexist with a small difference from $K = 2$; conversely, under a shorter budget, the improved $K = 4$ procedure can retain a substantial advantage. We use the three comparisons to identify these cases, rather than treating an objective gain as evidence for a batch composition.

Budget and objective jointly determine the observed preference. Independent seed blocks reproduce an early $K = 4$ advantage and a late $K = 2$ advantage in single-view CIFAR-100. An extended learning-rate study preserves the endpoint reversal while locating empirical early maxima inside the grid. In matched, reused-seed extensions, both L_{in} and an averaged per-target objective improve over native $K = 4$ after per-arm LR selection. At 9,600 updates their differences from $K = 2$ are -0.01 and $+2.43$ pp, respectively; at 38,400 they are $+0.08$ and -0.40. Thus the budget range in which a positive difference is observed depends on the objective, rather than following an intrinsic optimal-K trajectory.

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.

A controlled objective change identifies what is removed. We compare SupCon's L_{out} and L_{in} on identical $K = 4$ batches. Rewriting their definitions yields a within-positive KL term, the Jensen gap between the two losses (Khosla et al., 2020, Eqs. (2)–(3)). Removing it preserves each true-negative logit coefficient and the sum of positive coefficients at a fixed state. It changes how that sum is distributed across positives and improves late head retrieval in this recipe. A separate target-estimator comparison and an adverse clipping control delimit this interpretation.

The remaining difference must be evaluated at the intended readout. On the same test queries and checkpoints, L_{in} improves over native $K = 4$ more at the head than at the encoder. A direct paired analysis also finds a readout difference in its comparison against $K = 2$. A post-hoc frozen-feature linear probe shows that the large head-retrieval gains do not translate into equally large classification differences from $K = 2$. Pretrained CUB and SOP reproduce an objective improvement with much smaller differences from $K = 2$; on CUB the improvement remains under development-selected stopping. Stronger-input results include positive $K = 4$ advantages, whereas two-view experiments change the native pattern. Our contribution is the linked empirical evidence distinguishing these outcomes, with complete comparison families and explicit evaluation scope.

2 Setup and comparison design

Batches, objectives, and readouts. Let V be the views per source position. All-view same-label construction gives $B_{\text{enc}} = VB_{\text{src}}$ anchors and $m = VK - 1$ positives per anchor. The main setting has $B_{\text{src}} = 96$ and $V = 1$, so $K = 2/4$ corresponds to $G = 48/24$. Write F for native L_{out} at $K = 4$, K_2 for native L_{out} at $K = 2$, and O for an alternative objective at $K = 4$. The alternatives are L_{in} and the averaged per-target objective $\bar{M}$, defined in Section 4. At $V = 1, K = 2$, all these objectives coincide because there is only one positive. Thus $O - K_2$ is also a same-objective comparison of K in this single-view setting.

At a matched budget, protocol, readout, and seed population,

$$\underbrace{R_O - R_{K_2}}_{\text{difference from } K=2} = \underbrace{R_F - R_{K_2}}_{\text{native difference}} + \underbrace{R_O - R_F}_{\text{within-}K\text{ improvement}} . \tag{1}$$

This relation is a comparison convention, not an assumed learning law. With one view and the same optimizer recipe at both K values, the equality of objectives at $K = 2$ also makes $R_O - R_F$ the objective-by-K interaction: $[R_{O,4} - R_{O,2}] - [R_{F,4} - R_{F,2}]$. It does not separate the positive-count and class-coverage changes in a native K comparison. After LR selection, Eq. (1) applies to consistently selected procedures; that comparison changes the selected recipe as well as the objective.

Table 1: Experiment map. Fixed-recipe interventions and LR-selected procedure comparisons answer different questions. Appendix G maps the evidence types and seed blocks.

Study	Representation and evaluation	Main comparison / evidence type
CIFAR reference	Small conv; seen-class retrieval	Native K, objective controls; independent seed blocks
Native LR study	Same conv; hold-out head	Four rates, 12 new seeds; fixed and selected comparisons
Objective LR extensions	Same conv; head and encoder	Four rates each for L_{in} and $\bar{M}$; reused seeds
Paired readout analysis	Same checkpoint, same split	L_{in} block; post-hoc paired differences
Pretrained CUB / SOP	ImageNet ResNet-50; disjoint classes	Per-procedure development selection; stopping analysis
Input / view studies	Strong conv or two-view ResNet-18	Scope checks; separate view-count interaction

Training and evaluation. The CIFAR reference uses an eight-layer convolutional encoder with batch normalization, a 256-d pooled feature, and a normalized 128-d projection head. AdamW uses constant LR 10^{-3}, weight decay 10^{-4}, and $\tau = .05$. Weak input resizes to 64 pixels and flips horizontally. Training and hold-out splits contain 45,000 and 5,000 CIFAR-100 images; the latter has 50 images per class. Retrieval uses the same split for queries and gallery, excludes self-matches,

and reports Recall@1, Recall@5, and MRR@10. This seen-class task differs from class-disjoint CUB/SOP retrieval. Head and encoder are evaluated from the same checkpoint within each procedure. A separate frozen-feature linear probe measures top-1 classification accuracy. The official CIFAR test set was already examined during the research program. Appendix A gives architecture, evaluation, and pretrained specifications.

Evidence and uncertainty. Differences are Recall@1 percentage points (pp), unless stated otherwise. We report 95% paired-t intervals across training seeds, conditional on the evaluation set. Original two-sided sign-test decisions, excluding exact ties and correcting within each prespecified family, are retained separately: they assess directional balance, not the mean. Equivalence uses a dedicated margin-based design. LR-selected intervals and sign-formula values are nominal summaries because leave-one-seed-out folds overlap and share evaluation queries. Reanalyses and reused-seed extensions are identified explicitly; none converts non-rejection into equality or supplies study-wide multiplicity control.

3 Native preference changes with training budget

At the reference rate, independent CIFAR seed blocks give $F - K_2 = +3.14/-2.92$ pp at 9,600/38,400 updates (eight seeds), $+3.67/-3.08$ on a second eight-seed block, and $+2.78/-2.77$ on twelve further seeds. A predicted early $K3 - K4$ ordering changes from $+1.38$ to an unresolved -0.01 on replication. We therefore focus on the pairwise reversal, not a unique optimal-K path.

A reversal at a prespecified lower LR. A separate block trains K_2 and F at $\{.125, .25, .5, 1\} \times 10^{-3}$ on twelve new seeds, giving 96 trajectories to 38,400 updates. Within an arm and seed, all rates receive the same batch sequence. At fixed 2.5×10^{-4}, $K = 4$ leads by 3.75 pp at 4,800 and trails by 3.70 at 38,400 (Table 2). Each endpoint has a unanimous direction across the twelve seeds. A post-hoc paired change between these endpoints is -7.45 pp $[-8.49, -6.41]$. The early endpoint was moved from 9,600 to 4,800 using earlier discovery data before this block began; the original 9,600 read remains reported.

Table 2: All four primary native LR comparisons. The fixed-rate rows provide the primary inferential evidence (two-sided sign $p = .00049$ each, family threshold .0125). Selected rows have nominal intervals; their recorded sign-formula values are .00635 and .00049. Appendix B.2 preserves the original IDs and seed values.

LR rule	Updates	$F - K_2$	$n_+/12$	95% interval
Fixed 2.5×10^{-4}	4,800	$+3.75$	12	$[3.16, 4.33]$
Fixed 2.5×10^{-4}	38,400	-3.70	0	$[-4.53, -2.87]$
Per-arm LOSO	4,800	$+3.17$	11	$[2.41, 3.92]$
Per-arm LOSO	38,400	-4.04	0	$[-4.89, -3.18]$

Lower-rate extension and selection sensitivity. At each budget, leave-one-seed-out (LOSO) selection chooses the rate maximizing the other eleven seeds' hold-out hit count, with ties favoring the higher rate. An extension adds $.0625\times$ and $.03125\times$ for both native arms through 9,600 updates. On this six-rate grid, both empirical maxima at 4,800 are interior, at $.125\times$. Neither added rate is selected, so the $+3.17$ result is unchanged seed by seed: the extension checks the search boundary, rather than providing another independent replication. Late maxima on the original four-rate grid are also interior.

The observed transition remains LR-dependent. At 9,600, the selected native difference is -0.80 pp, compared with $+3.41$ at the reference rate. A post-hoc bootstrap that reruns selection gives $[-1.44, +0.19]$ at 9,600 and $[-4.59, -2.73]$ at 38,400. The late full-sample best-vs-best value is -3.61, versus LOSO's -4.04, illustrating sensitivity to selection among near-tied rates. These procedures and the bootstrap's duplicate-position convention are specified in Appendix B.3.

Replication scope. A scratch ResNet gives $+4.64/-2.76$ pp at 9,600/38,400; a terminal-budget linear-decay control gives $+4.61/-2.93$. Fresh class-disjoint CUB gives $+2.35/-3.21$, although

its early absolute R@1 is only 9.61/11.96% for K2/K4. Seen-class CUB's late contrast misses its sign criterion. These results, their complete families, and pass-matched external failures are in Appendix B; they support specified comparisons, not a universal crossover rule.

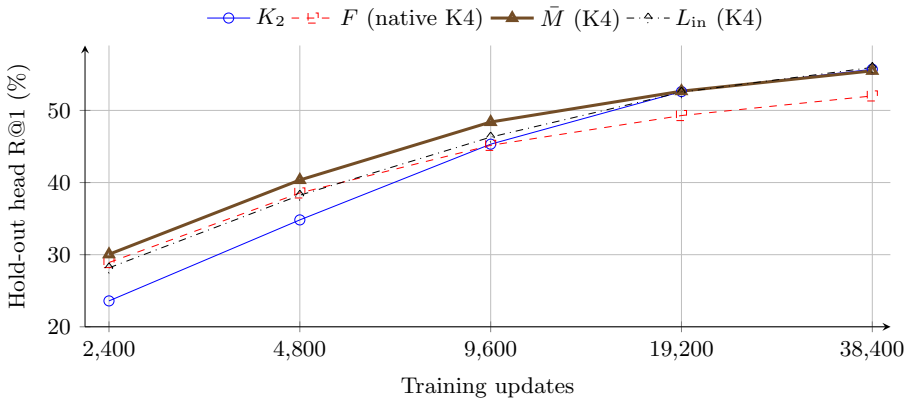

Figure 1: Absolute hold-out head R@1 at the common rate 2.5×10^{-4}, twelve matched seeds. These are descriptive mean levels from the native and both objective LR blocks; no per-arm variance bands are inferred from contrast intervals. All procedures improve over these reads, so the native reversal need not imply a decline in K4's absolute retrieval.

4 Objective Interventions and LR-Selected Comparisons

4.1 A Shared-Denominator Intervention

For an anchor, let P contain its m positives and let z_j be temperature-scaled cosine similarity. Write $A = \sum_{p \in P} e^{z_p}$, $Z = \sum_{n \in N} e^{z_n} > 0$ for true negatives, and $S = Z + A$. The three deterministic objectives are

$$L_F = \log S - \frac{1}{m} \sum_{p \in P} z_p, \qquad L_{\text{in}} = \log S - \log A + \log m, \tag{2}$$

$$L_{\bar{M}} = \frac{1}{m} \sum_{p \in P} \left[\log(Z + e^{z_p}) - z_p \right]. \tag{3}$$

L_F and L_{in} are SupCon's outside- and inside-log alternatives (Khosla et al., 2020, Eqs. (2)–(3)). Equation (3) matches UniMoCo's alternative averaged form, not its main UniCon loss (Dai et al., 2021, Appendix 6.2, Eq. (15)). It also matches SINCERE after reindexing and averaging ordered positive pairs (Feeney & Hughes, 2023, Eq. (12)), and ϵ-SupInfoNCE at $\epsilon = 0$ after dividing its positive sum by m, without the additional FairKL regularizer (Barbano et al., 2023, Eq. (4)). It is distinct from deleting all positives from the denominator.

With $u_p = 1/m$ and $\widetilde{q}_p = e^{z_p}/A$, the SupCon Jensen gap is

$$L_F = L_{\text{in}} + \mathrm{KL}(u \| \widetilde{q}). \tag{4}$$

For $p_j = e^{z_j}/S$, differentiation gives

$$\partial_{z_n} L_F = \partial_{z_n} L_{\text{in}} = p_n, \qquad \sum_{p \in P} \partial_{z_p} L_F = \sum_{p \in P} \partial_{z_p} L_{\text{in}} = -Z/S. \tag{5}$$

Individual positive coefficients change from $p_p - 1/m$ to $p_p - \widetilde{q}_p < 0$. Thus the intervention retains the shared denominator, true-negative coefficients, and total positive coefficient at a fixed state, while reallocating coefficients among positives. The KL term is identically zero when $m = 1$, connecting this change directly to the single-view $K = 2$ reference. At $K = 4$ the input batches and class coverage are held fixed. Both L_{in} and $\bar{M}$ attract every positive, but their allocations differ: within

a fixed state, L_{in} gives larger coefficient magnitudes to higher-scoring positives, whereas $\bar{M}$ gives larger magnitudes to lower-scoring ones (Appendix C.1). This distinction motivates comparing their budget-dependent behavior rather than treating them as interchangeable repulsion-free objectives.

On twelve new seeds at the reference recipe, removing the KL term improves late hold-out head R@1 by $+3.56$ pp $[3.05, 4.06]$; all twelve differences are positive. $\bar{M}$ improves by $+3.31$ on the same seeds, and $\bar{M} - L_{\text{in}} = -0.25$ $[-0.77, 0.28]$ is unresolved under the original sign criterion. A per-target denominator is therefore unnecessary for an improvement in this recipe. Early $\bar{M} - L_{\text{in}}$ is $+1.72$, so the objectives are not interchangeable across budgets. Appendix C.3 retains all six prespecified comparisons, including both unresolved ones.

Separating the objective from target estimation. A two-by-two control crosses shared versus per-target denominators with one sampled positive versus an average over all positives. Averaging improves over the corresponding sampled-target procedure under both objectives, despite conditional gradient unbiasedness. The complete grid and its unresolved late per-target averaging comparison are in Appendix C. A narrower control, Fclip, clips positive coefficients at zero and performs worse than F by 1.73 pp. Unlike Eq. (5), clipping changes the total positive coefficient and can produce a non-integrable gradient field. Its failure concerns that intervention; the successful L_{in} change removes the entire KL term, not repulsion alone.

4.2 Both objective improvements persist after LR selection

Two extensions train $\bar{M}$ and L_{in} at the same four learning rates as the native arms. Each contributes 48 trajectories on the same twelve seeds, paired with the completed K2/F controls. Selected K2 runs are replayed to save checkpoints. The native lower-rate extension contributes another 48 short runs. The new analyses were fixed before their added runs, but deliberately reuse previously observed control seeds: they are registered extensions, not independent confirmations. Temperature and batch size remain fixed.

Table 3: Late comparisons after per-arm LR selection, 38,400 updates, twelve reused seeds. Entries are R@1 differences with nominal 95% paired-t intervals. Both gains over F are positive on all twelve seeds. The final column evaluates encoder features using rates selected on hold-out head R@1. Complete C1–C3 and A1–A3 decisions appear in Appendices B.5 and B.4.

K4 objective	Gain over F Hold-out head	Difference from K2 Hold-out head	Difference from K2 Test encoder
L_{in}	$+4.11$ $[3.53, 4.69]$	$+0.08$ $[-0.49, 0.64]$	-0.49 $[-1.06, 0.08]$
$\bar{M}$	$+3.64$ $[2.96, 4.32]$	-0.40 $[-1.07, 0.27]$	-0.92 $[-1.75, -0.10]$

Table 3 shows that both improvements persist beyond the reference-rate setting. The late native difference is $F - K_2 = -4.04$ pp. Adding each objective gain gives the much smaller difference from K2: $-4.04 + 4.11 \simeq +0.08$ for L_{in} and $-4.04 + 3.64 \simeq -0.40$ for $\bar{M}$. Neither comparison against K2 meets its original sign criterion; this is not an equivalence result. Unlike the fixed-recipe intervention, LR selection compares entire training procedures. As a complementary fixed-rate check, the late $L_{\text{in}} - F$ gain ranges from 3.42 to 4.38 pp across all four common rates, with twelve positive seeds at each rate. The corresponding $\bar{M} - F$ mean ranges from approximately 3.08 to 3.79 pp (Appendix B.4). Thus the gains are not created solely by the selection rule.

The shorter-budget results distinguish the alternatives. At 9,600, $\bar{M} - K_2 = +2.43$ $[1.83, 3.04]$, whereas $L_{\text{in}} - K_2 = -0.01$ $[-0.55, 0.54]$. Their direct paired difference is $L_{\text{in}} - \bar{M} = -2.44$ $[-3.09, -1.78]$, negative on all twelve seeds. By 38,400 that difference is $+0.48$ $[-0.08, 1.04]$. These descriptive comparisons show that objectives with similar late differences from K2 can have different short-budget behavior. At 4,800, both alternatives have positive mean differences from K2, but their empirical LR maxima remain at the lower grid boundary.

The test-encoder comparisons use *head-selected* checkpoints. The mean interval for $\bar{M} - K_2$ is negative, yet its original sign criterion is unmet (four positives, $p = .3877$); $L_{\text{in}} - K_2$ is also unresolved (three positives, $p = .1460$). Mean intervals and sign decisions answer different questions, and both are retained. These separate-split head/encoder estimates are not a paired readout interaction; the next section makes that comparison on the same test queries.

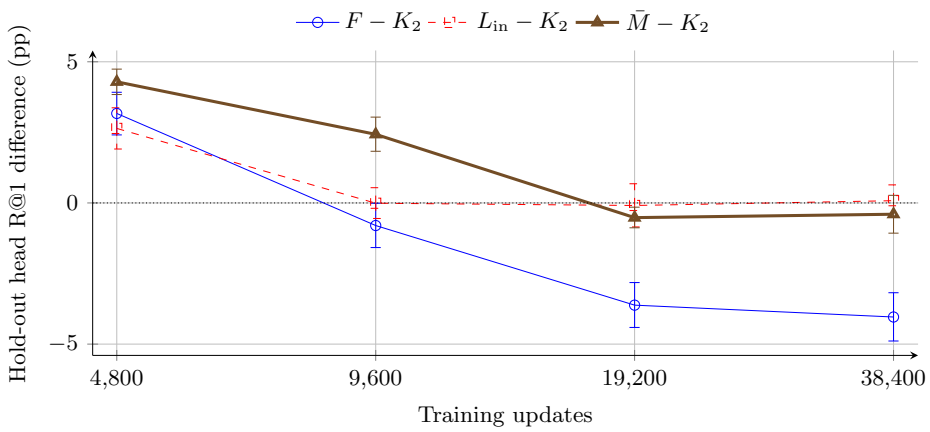

Figure 2: Differences from K2 under each objective after per-arm LR selection on hold-out head R@1, twelve reused seeds. Selection is repeated at each budget on the common four-rate grid. Error bars are nominal 95% paired-t intervals, not selection-calibrated bands. The objectives differ most clearly at the shorter budgets. Non-primary budgets are descriptive; lines join measured means on a base-2 logarithmic axis.

5 Which comparisons remain at the evaluation endpoint?

5.1 Head and encoder lead to different conclusions

At 38,400 updates in the independent reference-rate L_{in} block, the test-head gain over F is 3.74 pp, compared with 0.97 at the encoder. The remaining difference from K2 is $+0.83$ at the head and -0.37 at the encoder. Table 4 supplies the marginal intervals and distinguishes two direct paired questions: whether the *objective improvement* differs across readouts, and whether the *difference from K2* does. Their interactions are $+2.78$ and $+1.20$ pp, respectively. The latter supports readout dependence even though the encoder's individual interval includes zero.

Table 4: Post-hoc paired analysis on identical official-test queries, twelve seeds (900–911), at the reference LR. Head and encoder use the same checkpoint within each procedure. All interaction differences in the bottom two rows are positive across twelve seeds; these analyses are unadjusted and not independent evidence families.

$L_{\text{in}} - K_2$, late test	Head	Encoder
R@1 (pp)	$+0.83$ $[0.25, 1.41]$	-0.37 $[-0.78, 0.05]$
R@5 (pp)	$+1.14$ $[0.74, 1.53]$	$+0.20$ $[-0.11, 0.51]$
MRR@10 (points)	$+0.96$ $[0.50, 1.42]$	-0.12 $[-0.44, 0.21]$
Paired R@1 interaction: head minus encoder		Estimate [95% interval]
$(L_{\text{in}} - F)_{\text{head}} - (L_{\text{in}} - F)_{\text{enc}}$		$+2.78$ $[2.10, 3.45]$
$(L_{\text{in}} - K_2)_{\text{head}} - (L_{\text{in}} - K_2)_{\text{enc}}$		$+1.20$ $[0.63, 1.76]$

The head result also appears in R@5 and MRR@10, while their encoder intervals include zero. For $\bar{M}$, the late head-minus-encoder interaction is $+2.28$ $[1.78, 2.77]$ against F, and $+0.69$ $[0.12, 1.27]$ against K2. Because $F - K_2$ itself is more negative at the head (interaction -1.58), the larger gain against F cannot be attributed solely to properties of the alternative objective. All six contrasts and four readout combinations are in Appendix C.4. These retrieval results motivate checking a supervised readout of the same frozen representations.

5.2 Frozen-feature linear classification

A post-hoc probe evaluates all four reference-rate procedures on twelve seeds at 9,600 and 38,400 updates. It fits linear classifiers on 45,000 training images using frozen 256-d encoder or 128-d

head features. One regularization value per feature type is chosen on a predefined subset of late checkpoints using the hold-out set, before test evaluation, and shared across all fits. This probes the original 10^{-3} checkpoints, not the LR-selected models. Appendix C.5 gives the full protocol, all 24 comparisons, and numerical-sensitivity checks.

Table 5: Post-hoc encoder linear-probe top-1 differences (pp), twelve reference-rate seeds. Intervals are unadjusted 95% paired-t intervals; the 24 probe contrasts do not replace registered retrieval decisions. Small point estimates against K2 are not an equivalence test.

Contrast	9,600 updates	38,400 updates
$F - K_2$	$+2.72\ [2.39, 3.05]$	$-0.52\ [-1.01, -0.03]$
$\bar{M} - K_2$	$+3.55\ [3.15, 3.95]$	$+0.28\ [-0.16, 0.72]$
$L_{\text{in}} - K_2$	$+2.24\ [1.91, 2.57]$	$-0.09\ [-0.56, 0.38]$
$\bar{M} - F$	$+0.83\ [0.56, 1.10]$	$+0.80\ [0.50, 1.11]$
$L_{\text{in}} - F$	$-0.48\ [-0.69, -0.27]$	$+0.43\ [0.12, 0.74]$

The native encoder-probe mean changes from positive to negative, as in retrieval, while the late alternative-objective differences from K2 are small (Table 5). The apparent change of sign for $\bar{M} - K_2$, from -0.52 in encoder retrieval to $+0.28$ in the encoder probe, is a difference in point estimates, not evidence of opposite population rankings. Moreover, $L_{\text{in}} - F$ is negative at the early encoder probe and positive at the head probe ($+0.69$ pp). Linear classification therefore provides a complementary assessment rather than converting the large head-retrieval gain into a broad encoder advantage.

5.3 Small advantages and endpoint-specific equivalence

At the shared reference rate, four CIFAR blocks totaling 52 seeds yield a post-hoc pooled $\bar{M} - K_2$ head estimate of $+0.53$ pp (modified HKSJ interval $[0.07, 0.98]$; prediction interval $[-0.09, 1.14]$). This supports a small positive effect in that fixed recipe and should not be conflated with the LR-selected -0.40 estimate, which evaluates a different procedure.

A separate prospective design evaluates K2 and $\bar{M}$ on 24 new seeds at 76,800 and 153,600 updates, using the reference rate and a one-point margin fixed before these runs. On hold-out head R@1, $\bar{M} - K_2$ is $+0.15$ $[-0.44, 0.74]$ and $+0.28$ $[-0.24, 0.81]$; both satisfy the prespecified non-inferiority and ± 1 TOST criteria. On the twelve shared control seeds, $F - K_2 = -3.07$ $[-3.87, -2.27]$ at 153,600. This is equivalence for $\bar{M}$, at two head-retrieval endpoints and the stated resolution. It does not apply to L_{in}, selected LRs, or unmeasured budgets. Narrower-margin results and the full families are in Appendix C.6.

5.4 Class-disjoint retrieval and development-selected stopping

We fine-tune ImageNet-pretrained ResNet-50 with a normalized 128-d head, frozen BN statistics, 224-pixel inputs, and $B_{\text{src}} = 96, V = 1$. CUB trains on its first 100 classes and evaluates on the other 100; SOP uses its official item split. Development partitions training classes or items. Each procedure independently selects its LR and temperature from a finite grid; all select the same setting within a dataset: $(3 \times 10^{-6}, .05)$ on CUB and $(3 \times 10^{-5}, .05)$ on SOP. The CUB selected and shared-recipe blocks consequently differ in seeds, not hyperparameters.

The CUB search reaches its lower LR boundary, and all development cells have lower MAP@R at 150 than at 30 passes. We therefore report a post-hoc stopping analysis alongside the fixed endpoint: at the already selected hyperparameters, each procedure chooses its checkpoint using mean development MAP@R, with ties favoring earlier stopping. All select 30 passes. The objective improvement remains ($+1.52$ versus $+2.97$ at 150), while $\bar{M} - K_2 = -0.34$ in R@1 and -0.01 $[-0.12, 0.11]$ in MAP@R. The conclusion is metric-specific, and the effect size depends on stopping. At the fixed endpoints, absolute R@1 levels for K2/F/Mbar are 59.84/56.90/59.87 on selected CUB and 74.00/68.34/73.95 on SOP. SOP improves by 5.64 MAP@R points over F at ten passes. Filtering SOP to items with at least four images preserves the objective improvement, although the changed population and seeds preclude attributing its magnitude uniquely to duplicates. These pretrained results concern $\bar{M}$, not L_{in}.

Table 6: Pretrained head retrieval, R@1 differences in pp. Development-selected stopping is a post-hoc CUB analysis using development MAP@R, not final-test choice. Fixed-endpoint comparisons retain their original prespecified decisions. Both CUB seed blocks are shown. Full metric families, including short budgets, are in Appendix D.

Dataset / endpoint rule	Passes	n	$\bar{M} - F$ [95% interval]	$\bar{M} - K_2$ [95% interval]
CUB / dev stopping	30	12	$+1.52\ [1.18, 1.86]$	$-0.34\ [-0.62, -0.07]$
CUB / fixed endpoint	150	12	$+2.97\ [2.57, 3.37]$	$+0.03\ [-0.29, 0.35]$
CUB shared / dev stopping	30	8	$+1.24\ [0.86, 1.63]$	$-0.07\ [-0.53, 0.39]$
CUB shared / fixed	150	8	$+2.71\ [2.32, 3.10]$	$-0.49\ [-0.83, -0.15]$
SOP / fixed	10	12	$+5.61\ [5.35, 5.86]$	$-0.05\ [-0.34, 0.24]$
SOP no duplicates / fixed	10	12	$+4.69\ [4.52, 4.87]$	$-0.08\ [-0.43, 0.28]$

6 Input regime, related work, and scope

Stronger inputs can retain a positive $K = 4$ difference. With strong augmentation in the single-view CIFAR recipe, $F - K_2 = +8.98/+2.09$ pp at 9,600/38,400, contradicting the original late negative prediction. A subsequently specified 76,800-update extension gives -2.38. The objective contrasts are $\bar{M} - F = +3.06/+3.93$ at the two late budgets; $\bar{M} - K_2 = +5.14/+1.55$, positive on all eight seeds, is descriptive. Thus positive K4 advantages remain part of the evidence rather than exceptions to an assumed no-advantage claim.

Two views change both the empirical pattern and the reference. A two-view CIFAR ResNet-18 at 32 pixels, $B_{\text{src}} = 200$, and $\tau = .1$ gives $F - K_2 = -1.17$ at five passes, opposite its early prediction, and -2.04 at 80. The late $\bar{M} - F$ improvement is $+2.84$; $\bar{M} - K_2 = +0.80$ misses its original sign criterion. At $V = 2$, K2 has three positives, so native K2 is no longer a common-objective reference. This last contrast changes both K and the objective: a same-objective $\bar{M}(K4, V2) - \bar{M}(K2, V2)$ comparison remains unmeasured. A separate paired view-count intervention on the strong conv recipe reduces the native difference by $8.63\ [7.73, 9.53]$ at 9,600 updates. It also doubles encoded work and changes normalization inputs (Appendix E).

Relation to prior work. N-pair learning illustrates structured metric-learning batches (Sohn, 2016); subsequent work studies sampling and evaluation protocols (Wu et al., 2017; Roth et al., 2020; Musgrave et al., 2020). Short-horizon meta-optimization can bias hyperparameter learning (Wu et al., 2018), while budgeted training motivates budget-aware schedules (Li et al., 2020). Neither establishes a rule for K. SupCon already compares L_{out} and L_{in}, with a different empirical ordering in its ImageNet setting (Khosla et al., 2020, Table 1). UniMoCo supplies a multi-positive alternative (Dai et al., 2021); SINCERE analyzes intra-class repulsion (Feeney & Hughes, 2023). SimCLR distinguishes encoder representations from projection outputs (Chen et al., 2020), whereas Graf et al. (2021) study supervised-contrastive loss geometry. Our contribution is matched empirical evidence about objective-dependent K preferences, not these prior objectives or geometrical results.

Scope and remaining evidence. The controlled objective study uses a small single-view CIFAR encoder; the completed four-rate extensions cover both L_{in} and $\bar{M}$ at 38,400 updates. LR selection keeps temperature and batch fixed, reuses evaluation queries, and does not establish global optima. Constant-rate trajectories and one terminal-decay control do not cover all budget-aware schedules. The native grid extends beyond $K = 4$, but the objective interventions do not. The post-hoc probe covers the reference-rate checkpoints only; a probe of LR-selected checkpoints and same-objective two-view K comparisons are outside the measured scope. Test-set reuse, post-hoc analyses, and study-level multiplicity limit extrapolation beyond the reported protocols. The supplement separates evidence histories and preserves failed predictions; the supplied companion supports arithmetic and objective checks, not full historical training reproduction.

7 Conclusion

An objective improvement and a samples-per-class preference are different empirical quantities. In the studied single-view setting, $K = 4$ leads early under native SupCon and loses late; changing

the objective can preserve a positive difference over a longer measured budget range, including after per-arm LR selection. Both objective gains persist under per-arm LR selection, while the remaining differences from K2 are much smaller at the late endpoint. Paired retrieval analyses, a post-hoc linear probe, and development-selected stopping further show why the intended evaluation matters. These results support comparing K under the intended objective, budget, selection procedure, and readout, while retaining both positive and negative outcomes rather than assuming a universal preferred composition.

REPRODUCIBILITY STATEMENT

Appendix A specifies training and evaluation, Appendices B.5 and C.5 specify the completed LR extension and probe, and Appendix G maps evidence histories. The companion contains manuscript sources, all transcribed result tables, available numeric seed differences, arithmetic checks, and an independent PyTorch reference for the objective definitions. This reference was not used for the reported training. The complete historical trainer, executed configurations, split manifests, raw per-run outputs, and feature caches are not included. Summary-based checks audit reported arithmetic and algebra, not historical execution or preregistration timing.

AI USE STATEMENT

Generative AI assisted hypothesis and framework refinement, experimental design, implementation and analysis, mathematical derivations and proof checks, literature review, result interpretation, and manuscript preparation, including translation, figures, and reviewer-style critique. AI-generated numerical cases were used only for code and algebra checks, not benchmark data. The authors are responsible for all content.

REFERENCES

Carlo Alberto Barbano, Benoit Dufumier, Enzo Tartaglione, Marco Grangetto, and Pietro Gori. Unbiased supervised contrastive learning. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=Ph5cJSfD2XN.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning (ICML)*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1597–1607, 2020. URL https://proceedings.mlr.press/v119/chen20j.html.

Zhigang Dai, Bolun Cai, Yugeng Lin, and Junying Chen. UniMoCo: Unsupervised, semi-supervised and full-supervised visual representation learning. *arXiv preprint arXiv:2103.10773*, 2021. URL https://arxiv.org/abs/2103.10773.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 248–255, 2009. URL https://image-net.org/static_files/papers/imagenet_cvpr09.pdf.

Patrick Feeney and Michael C. Hughes. SINCERE: Supervised information noise-contrastive estimation REvisited. *arXiv preprint arXiv:2309.14277*, 2023. URL https://arxiv.org/abs/2309.14277v4.

Florian Graf, Christoph Hofer, Marc Niethammer, and Roland Kwitt. Dissecting supervised contrastive learning. In *International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pp. 3821–3830, 2021. URL https://proceedings.mlr.press/v139/graf21a.html.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016. URL https://www.cv-foundation.org/openaccess/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html.

Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017. URL https://arxiv.org/abs/1703.07737.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 18661–18673, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/d89a66c7c80a29b1bdbab0f2a1a94af8-Abstract.html.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. URL https://www.cs.toronto.edu/~kriz/cifar.html.

Mengtian Li, Ersin Yumer, and Deva Ramanan. Budgeted training: Rethinking deep neural network training under resource constraints. In *International Conference on Learning Representations (ICLR)*, 2020. URL https://arxiv.org/abs/1905.04753.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. URL https://arxiv.org/abs/1306.5151.

Kevin Musgrave, Serge Belongie, and Ser-Nam Lim. A metric learning reality check. In *European Conference on Computer Vision (ECCV)*, pp. 681–699, 2020. URL https://doi.org/10.1007/978-3-030-58595-2_41.

Karsten Roth, Timo Milbich, Samarth Sinha, Prateek Gupta, Björn Ommer, and Joseph Paul Cohen. Revisiting training strategies and generalization performance in deep metric learning. In *International Conference on Machine Learning (ICML)*, volume 119 of *Proceedings of Machine Learning Research*, pp. 8242–8252, 2020. URL https://proceedings.mlr.press/v119/roth20a.html.

Kihyuk Sohn. Improved deep metric learning with multi-class N-pair loss objective. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 29, pp. 1857–1865, 2016. URL https://proceedings.neurips.cc/paper/2016/hash/6b180037abbebea991d8b1232f8a8ca9-Abstract.html.

Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. Deep metric learning via lifted structured feature embedding. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4004–4012, 2016. URL https://cvgl.stanford.edu/projects/lifted_struct/.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The Caltech-UCSD Birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011. URL https://authors.library.caltech.edu/records/cvm3y-5hh21.

Chao-Yuan Wu, R. Manmatha, Alexander J. Smola, and Philipp Krähenbühl. Sampling matters in deep embedding learning. In *IEEE International Conference on Computer Vision (ICCV)*, 2017. URL https://arxiv.org/abs/1706.07567.

Yuhuai Wu, Mengye Ren, Renjie Liao, and Roger B. Grosse. Understanding short-horizon bias in stochastic meta-optimization. In *International Conference on Learning Representations (ICLR)*, 2018. URL https://openreview.net/forum?id=H1MczcgR-.

SUPPLEMENTARY MATERIAL

This supplement retains complete comparison families, protocol details and contrary results. Table captions identify budgets, readouts, and evidence types; the companion maps the newly added summaries to source sections. Post-hoc analyses are identified separately.

A EXPERIMENTAL PROTOCOLS AND INFERENCE

Batches and representations. The main source batch has $B_{\mathrm{src}} = GK = 96$ positions. With V views per position, all-view construction yields VB_{src} anchors, $VK - 1$ positives and $V(B_{\mathrm{src}} - K)$ true-negative candidates per anchor. SOP positions can repeat images. At $V = 1, K = 2$, all objectives coincide; at $V = 2, K = 2$, they do not. No $M(K2, V2)$ result is supplied. Scratch experiments use a normalized 128-d head; the L-in block also evaluates the 256-d pooled encoder. Encoder retrieval is not a linear probe.

Scratch recipe and evaluation. The main encoder has four blocks with two 3-by-3 conv–BN–ReLU layers each, with channels 64/128/256/256, max-pooling after the first three blocks, and global average pooling. Its head is a 256–256–128 MLP. There are 2,425,664 parameters including the head and 2,326,976 before it. AdamW (Loshchilov & Hutter, 2019) uses $\beta = (.9, .999)$, weight decay 10^{-4}, constant LR 10^{-3}, no warmup, and $\tau = .05$. Weak input resizes to 64 pixels and flips horizontally. Strong input adds random resized crops at scale [.2,1], color jitter with probability .8 and grayscale with probability .2. The separate schedule block decays linearly to zero at its own terminal budget. CIFAR-100 (Krizhevsky, 2009) is split into 45,000 training and 5,000 hold-out images, with 50 hold-out queries per class. Query and gallery are the same split; self-matches are excluded. R@1 and R@5 use integer hit counts and MRR truncates at 10. The official 10,000-image test set was used earlier in the research program; it is not an untouched research-level test set. Native class-disjoint and seen-class CUB are distinct evaluation populations.

Pretrained retrieval. CUB-200-2011 (Wah et al., 2011) final fitting uses 5,864 images of the first 100 classes, and evaluation uses 5,924 images of the other 100. SOP (Song et al., 2016) uses 59,551 training images/11,318 items and 60,502 evaluation images/11,316 items. Development partitions the training classes or items 80/20. ImageNet-pretrained ResNet-50 (Deng et al., 2009; He et al., 2016) (`IMAGENET1K_V1`) uses a linear 2048-to-128 head, frozen BN running statistics, 224-pixel train crops/flips and resize-256/center-crop-224 evaluation. Pretrained fitting uses AdamW, weight decay 10^{-4}, constant LR without warmup, and a single parameter group, so the head receives the same selected LR as the backbone. The records available for this revision do not resolve the BN affine policy, all crop defaults, or evaluator tie handling and chunking; reproducing those details requires the executed configurations.

Inferential targets. Mean-t intervals target the expected paired seed difference, conditional on the evaluation set and the usual independent-seed/distributional assumptions. Sign tests target directional balance after excluding ties. Original percentile intervals and blockwise decisions are retained. Eight non-tied pairs have minimum two-sided sign $p = .0078125$, exceeding $.05/9$; the earlier three-rate, eight-seed LR family with nine primary cells cannot resolve a sign test at that sample size. Blockwise control is not study-wide control. Reused checkpoints and readouts are dependent; selected final-seed intervals condition on the development choice. The NI block fixes one-sided .025 tests and a data-informed one-point margin before its new runs, corresponding to 95% intervals. The LR temporal contrast, paired readout interactions, stopping selection, and pooled summaries are post hoc.

Execution evidence. Records describe configuration/hash checks and checkpoint rereads. The research records name per-run metrics, configurations, split manifests, and evaluation code. Only the supplied summary tables and selected seed differences are transcribed in this companion; it does not contain the historical trainer or a complete raw-output export. Its new PyTorch objective reference is derived from the equations, not used for the reported training. Synthetic and arithmetic checks do not independently validate historical execution or registration timing. Exceptions remain in Appendix F.

B NATIVE SAMPLER COMPARISONS AND THE EXTENDED LR BLOCK

B.1 ORIGINAL GRIDS, FRESH ENDPOINTS AND BOUNDARIES

The original grids show a changing native preference, not a unique optimal-K path. The fresh $K3 - K4$ contrast at 9,600 is unresolved even when the $K4 - K2$ reversal repeats.

Table 7: Original native K grid and its paired differences; CIFAR reference recipe. Counts are positive/non-tied seeds.

T	$K{=}2$	$K{=}3$	$K{=}4$	$K{=}6$	$K{=}8$	$K{=}12$	best	shape
2,400	11.03	—	**15.86**	—	14.54	11.04	4	interior
4,800	21.38	—	**28.56**	—	25.47	20.29	4	interior
9,600	37.43	**42.03**	40.65	38.43	35.56	30.82	3	interior
19,200	**50.39**	49.87	47.52	43.85	41.48	34.85	2	monotone
38,400	**54.52**	52.74	51.01	48.28	45.01	39.55	2	monotone
76,800	**56.02**	53.88	52.42	51.00	—	—	2	monotone

T	$K{=}2$	$K{=}3$	$K{=}4$	$K3-K2$	$K4-K2$	$K3-K4$
9,600	38.05	41.18	41.19	$+3.13$ (8/8)	$+3.14$ (8/8)	-0.01 (5/8)
38,400	53.93	52.63	51.00	-1.30 (1/8)	-2.92 (0/8)	$+1.62$ (7/8)

The full fresh comparison family retains the seen-class CUB late non-confirmation. Class-disjoint CUB has early levels 9.61/11.96% for K2/K4; its absolute accuracy is not comparable to seen-class retrieval.

Table 8: Complete fresh replication family, retaining unresolved and adverse outcomes.

endpoint	predicted	fresh seeds	seeds	exact p
CIFAR-100, 9,600	$+$	$+3.14$	8/8	.0078
CIFAR-100, 38,400	$-$	-2.92	0/8	.0078
CUB-200 seen-class, 2,400	$+$	$+2.59$	8/8	.0078
CUB-200 seen-class, 38,400	$-$	-3.43	2/8	.29
CUB-200 class-disjoint, 2,400	$+$	$+2.35$	8/8	.0078
CUB-200 class-disjoint, 38,400	$-$	-3.21	0/8	.0078

The previous LR and schedule blocks are shown separately. The half-rate native contrast at 9,600 is unresolved; the previous LR study's lower grid boundary motivated the extension below.

Table 9: Previous LR contrasts and terminal-budget linear-decay control, seeds 100–107. Superscript a marks short-budget reads from a later-registered block on the same discovery seeds. The nine-cell LR family cannot attain $.05/9$ with eight pairs.

rate	2,400	4,800	9,600	19,200	38,400
$0.5\times$	$+7.04$ (8/8)^a	$+6.37$ (8/8)^a	$+0.95$ (6/8, sign p=.29, permutation p=.031)	-2.63 (0/8)	-3.32 (0/8)
$1\times$	$+4.83$ (8/8)	$+7.18$ (8/8)	$+3.22$ (8/8)	-2.88 (0/8)	-3.52 (0/8)
$2\times$	—	—	$+3.35$ (8/8)	-2.81 (0/8)	-3.29 (0/8)

T	4,800	9,600	19,200	38,400
decay	$+6.11$ (8/8)	$+4.61$ (8/8)	-0.78 (1/8, p=.070)	-2.93 (0/8)
constant	$+7.18$	$+3.22$	-2.88	-3.52

Table 10: Scratch-ResNet native replication, seeds 100–107. Both original two-cell-family sign tests give $p = .0078$; these are reported percentile, not newly recomputed t, intervals.

Updates	$F-K_2$	95% percentile interval	Positive/non-tied
9,600	$+4.64$	$[4.07, 5.14]$	8/8
38,400	-2.76	$[-3.81, -1.78]$	0/8

Other transfer failures. The pass-matched cross-dataset design does not align training states. Its predicted external negative interaction is not established on Aircraft (Maji et al., 2013) and is opposite on CUB. The earlier pretrained seen-class CUB comparison changes several recipe axes and does not isolate pretraining. A continuation fork and class-adaptive allocation fail or remain unresolved on their original targets; neither is a validated selector. Their complete exploratory records remain in the audit archive.

Table 11: Pass-matched native differences across datasets. † marks results below the prespecified ten-times-chance floor. The last two pass ranges reflect different dataset sizes at common update budgets, not uncertainty intervals.

passes	CIFAR-100	FGVC-Aircraft	CUB-200
5.1	+4.83 (8/8)	+0.56 (8/8)†	+0.70 (7/8)†
10.2	+7.18 (8/8)	—	—
20.5	+3.22 (8/8)	+2.23 (8/8)†	+1.04 (8/8)
41.0	−2.88 (0/8)	+4.51 (8/8)†	+3.04 (8/8)
81.9	−3.52 (0/8)	+1.54 (5/8)	+5.13 (8/8)
154–276	—	−1.36 (2/8)	+0.34 (5/8)
615–1,106	—	+0.23 (6/8)	−3.69 (0/8)

B.2 Design, selection and uncertainty of the new LR block

The new block trains native K_2 and F at rates $\{.125, .25, .5, 1\} \times 10^{-3}$ on seeds 1200–1211: $2 \times 4 \times 12 = 96$ trajectories to 38,400 updates. The five reads are 2,400/4,800/9,600/19,200/38,400 updates, or 5.12/10.24/20.48/40.96/81.92 source passes. The recipe and hold-out match Appendix A. Intermediate reads restore all RNG states; the source reports execution checks against separate constant-LR runs. Within each arm and seed, all four LRs receive the same 38,400 batches. No checkpoints are saved in this block.

At budget T, let $H_{A,r,j}(T)$ be the integer R@1 hit count. The selection rule is

$$r_A^{(-s)}(T) = \arg\max_{r \in \mathcal{R}} \sum_{j \neq s} H_{A,r,j}(T), \qquad D_s^{\mathrm{LOSO}}(T) = R_{F,r_F^{(-s)},s}(T) - R_{K_2,r_{K_2}^{(-s)},s}(T),$$

with exact ties resolved in favor of the higher LR. The scored seed is excluded from its own selection, but folds share 10 of their 11 selection seeds and reuse the hold-out queries. Selected differences need not be independent. B1/B2 sign-formula p values and t intervals are nominal summaries, not selection-aware inference. The source records identify the complete per-rate outcomes in the research repository. The companion contains the reported levels, primary seed differences, and selection-bootstrap summaries, rather than a complete per-arm/per-rate raw array export.

Four comparisons were registered as one family at $.05/4$: LOSO at 4,800 and 38,400 (B1/B2), and fixed $.25\times$ at the same budgets (B3/B4). The early endpoint was changed from 9,600 to 4,800 using earlier discovery results before any of these 96 runs existed; this is prospective within this block, not an outcome-independent choice across the research program. The fixed-rate B3/B4 differences do not share a data-selected LR and provide the primary inferential support here. All four source verdicts are reported in Table 2; the inferential interpretation distinguishes the two types.

All rate-specific levels. Table 16 gives every native and alternative-objective mean level, including the lower-rate extension. The native four-rate levels are also stored in `data/lr_arm_rate_levels.csv`.

All contrasts and the sampled sign change. Entries are mean differences in pp, followed by positive/non-tied seeds. Non-primary budgets are descriptive. The first negative sampled mean occurs at 9,600 for the two lower fixed rates and 19,200 for the two higher rates. This does not estimate an exact continuous crossover. At 4,800 both arms choose .125 in all LOSO folds. At 9,600, K2 chooses .125 and F chooses .25 in every fold. At 19,200, K2 chooses .25 throughout; F chooses .25 four times and .5 eight times. At 38,400, K2 chooses .25 eleven times and .5 once; F again chooses .25 four times and .5 eight times.

Table 12: All native rate-specific and selected differences across budgets; primary cells are identified separately.

LR multiplier	2,400	4,800	9,600	19,200	38,400
.125	+4.89 (12/12)	+3.17 (11/12)	-1.61 (0/12)	-3.51 (0/12)	-3.89 (0/12)
.25	+5.31 (12/12)	+3.75 (12/12)	-0.15 (6/12)	-3.32 (0/12)	-3.70 (0/12)
.5	+6.77 (12/12)	+5.76 (12/12)	+1.90 (12/12)	-2.42 (0/12)	-3.47 (0/12)
1	+4.97 (12/12)	+6.86 (12/12)	+3.41 (12/12)	-2.30 (0/12)	-2.72 (1/12)
LOSO	+4.89 (12/12)	+3.17 (11/12)	-0.80 (2/12)	-3.62 (0/12)	-4.04 (0/12)

Seed-level primary contrasts. Table 15 combines the original four LR comparisons with the three new Mbar comparisons without counting the reused early values twice. The post-hoc paired temporal contrast B4–B3 has mean -7.4517, standard deviation 1.6349, and 95% t interval $[-8.4904, -6.4129]$. It is not an additional prespecified test.

Grid boundaries and execution record. On the original four-rate grid, empirical late maxima are interior (K2 at .25, F at .5), while early maxima reach the lower boundary. The subsequent six-rate extension in Appendix B.4 makes both native early maxima interior without changing any selected early seed value. Neither grid certifies continuous/global optimality. LR also changes AdamW's effective decay. The original block records 96 completed runs, no guard failures, and one final evaluation after pre-execution code fixation. Its native-only scope is complemented by the completed Mbar and Lin LR extensions below.

B.3 Selection-aware bootstrap sensitivity

This post-hoc reanalysis resamples twelve seed positions with replacement 10,000 times (NumPy seed 20260926), then repeats the selection rule within each replicate. For each scored position only that position is excluded: a second copy of the same original seed can remain in the selection set. This is the reported bootstrap convention, not evaluation on fresh independent seeds. Its percentile ranges quantify sensitivity under that procedure. The difference between the observed and bootstrap-average estimates does not establish a universal direction of LOSO bias.

Table 13: Native four-rate selection reanalysis. Bootstrap ranges are 2.5–97.5 percentiles; they are distinct from the nominal t intervals and do not provide Mbar uncertainty.

Updates	Observed LOSO	Bootstrap LOSO	In-sample best	Bootstrap in-sample
4,800	$+3.17$	$[2.44, 3.70]$	$+3.17$	$[2.44, 3.70]$
9,600	-0.80	$[-1.44, 0.19]$	-0.80	$[-1.42, -0.14]$
19,200	-3.62	$[-4.00, -2.62]$	-3.29	$[-3.80, -2.62]$
38,400	-4.04	$[-4.59, -2.73]$	-3.61	$[-4.08, -2.98]$

B.4 Mbar LR study and the lower native rates

The extension deliberately reuses seeds 1200–1211 to pair Mbar with the already observed K2/F controls. It adds 48 Mbar runs on $\{.125, .25, .5, 1\} \times 10^{-3}$ through 38,400 updates, twelve selected-rate K2 replays with saved checkpoints, and 48 short K2/F runs at $.0625\times$ and $.03125\times$ through 9,600. All 108 runs complete without recorded failure. The K2 replays match the original metrics and recorded intermediate rows on all twelve seeds. The design is internally prespecified before the added runs, but not a fresh-seed confirmation. The original four-cell sign family is retained below; selected intervals remain nominal and no equivalence margin was specified.

Table 14: The full extension family (threshold .0125). A3 uses the head-based LR choices from A1. Ext-B1 is source B1, prefixed to distinguish the extension; its values equal original LR-B1 seed by seed. Sign-formula p values are .3877, .00049, .3877, .00635.

ID	Comparison / readout	Updates	Mean [95% t]	$n_+/12$	Decision
A1	$\bar{M} - K_2$, hold-out head	38,400	-0.40 $[-1.07, 0.27]$	4	Unresolved
A2	$\bar{M} - F$, hold-out head	38,400	$+3.64$ $[2.96, 4.32]$	12	Met
A3	$\bar{M} - K_2$, test encoder	38,400	-0.92 $[-1.75, -0.10]$	4	Unresolved
Ext-B1	$F - K_2$, hold-out head	4,800	$+3.17$ $[2.41, 3.92]$	11	Met

Table 15: All available primary seed differences (pp). LR-B1/B2 are native LOSO at 4,800/38,400; LR-B3/B4 are fixed .25× at those budgets. A1/A2/A3 are defined in Table 14. Ext-B1 equals LR-B1 on every seed. A1 minus A2 equals LR-B2.

Seed	LR-B1	LR-B2	LR-B3	LR-B4	A1	A2	A3
1200	$+4.08$	-4.44	$+3.22$	-3.42	$+0.26$	$+4.70$	-1.17
1201	$+4.14$	-5.94	$+5.24$	-5.94	-0.60	$+5.34$	-2.22
1202	$+3.36$	-3.50	$+3.22$	-3.90	-0.90	$+2.60$	$+0.77$
1203	$+2.98$	-3.42	$+4.22$	-1.64	$+1.20$	$+4.62$	-1.84
1204	$+4.04$	-5.32	$+3.00$	-5.32	-0.90	$+4.42$	-2.47
1205	$+2.72$	-5.42	$+3.94$	-3.92	-2.74	$+2.68$	$+1.02$
1206	-0.22	-3.94	$+3.52$	-3.94	-0.66	$+3.28$	-1.21
1207	$+3.42$	-2.96	$+5.46$	-3.18	-0.30	$+2.66$	-1.62
1208	$+3.24$	-1.40	$+3.72$	-1.94	$+0.36$	$+1.76$	-1.13
1209	$+3.24$	-4.68	$+4.06$	-4.68	-0.94	$+3.74$	$+0.34$
1210	$+2.80$	-4.96	$+2.26$	-4.26	-0.68	$+4.28$	$+0.74$
1211	$+4.18$	-2.44	$+3.12$	-2.30	$+1.12$	$+3.56$	-2.29

Mbar selects .5× on seeds 1202, 1205, and 1210 and .25× otherwise. K2 selects .5× only on seed 1211. F selects .25× on 1201, 1204, 1206, and 1209 and .5× otherwise. Both native arms select .125× on every seed at 4,800. The new low-rate runs do not change those choices, so Ext-B1 is a search-boundary check, not additional independent replication. The per-budget native and Mbar trajectories plotted in Figure 2 are retained as a complete five-budget table in the companion (`supplement/companion_tables/selected_trajectory.tex`); Lin counterparts appear in Table 20. The all-seed in-sample best-vs-best Mbar–K2 differences at these budgets are $+6.07, +4.29, +2.43, +0.06, -0.22$; the corresponding Mbar–F values are $+1.18, +1.13, +3.23, +3.35, +3.39$. These are optimistically selected on the evaluation seeds and are not substitutes for the LOSO comparison.

Common-rate descriptive check. Subtracting the reported rounded arm means gives the following late fixed-rate differences. Each comparison uses one common LR, rather than an LR chosen by LOSO. These are an additional descriptive calculation from aggregate levels, with at most 0.01 pp rounding discrepancy from an unrounded paired mean. The supplied summaries do not determine their per-rate uncertainty or seed sign counts.

Readouts of checkpoints selected at the late endpoint. In the following table LR selection always uses 38,400-update hold-out head R@1, including when an earlier checkpoint is evaluated. Thus its 9,600 hold-out value ($+3.07$) differs from the per-budget-selected value ($+2.43$) above. A1 and A3 use different splits and cannot be subtracted to isolate a readout interaction. The late same-test head–encoder mean difference is approximately $+0.51$ pp, but its paired interval is not supplied by the added report. As a descriptive same-rate comparison, Mbar at each seed's K2-selected rate gives late R@1 differences of 0.00 $[-0.61, 0.61]$ on hold-out head, -0.19 $[-0.75, 0.36]$ on test head, -1.29 $[-1.78, -0.80]$ on test encoder, and -1.17 $[-1.88, -0.45]$ on hold-out encoder. The complete early counterparts and counts remain in the companion transcription. These are K2-selected rates, not an independently prespecified common fixed rate.

Table 16: All native and alternative-objective mean hold-out head R@1 levels (percent), twelve matched seeds. Native low-rate runs stop at 9,600. All six-rate native short-budget maxima are interior; Mbar and Lin each have four rates and remain at the lower boundary at 2,400/4,800.

Arm	LR multiplier	2,400	4,800	9,600	19,200	38,400
K2	0.03125	22.10	29.05	35.94	—	—
K2	0.0625	25.09	34.05	42.21	—	—
K2	0.125	25.94	36.30	45.96	51.02	53.56
K2	0.25	23.59	34.82	45.32	52.61	55.71
K2	0.5	17.29	29.12	42.09	51.75	55.58
K2	1	10.97	21.37	37.75	50.05	54.10
F	0.03125	24.81	30.79	35.27	—	—
F	0.0625	28.72	36.11	40.19	—	—
F	0.125	30.83	39.47	44.35	47.51	49.68
F	0.25	28.91	38.56	45.16	49.29	52.01
F	0.5	24.06	34.88	43.99	49.33	52.10
F	1	15.94	28.23	41.16	47.75	51.39
Mbar	0.125	32.00	40.59	47.16	50.85	53.47
Mbar	0.25	30.05	40.35	48.39	52.67	55.49
Mbar	0.5	24.75	36.59	46.46	52.66	55.46
Mbar	1	16.27	29.45	43.75	51.19	54.47
L_{in}	0.125	30.09	38.94	46.24	51.07	54.05
L_{in}	0.25	28.17	38.24	46.30	52.53	55.96
L_{in}	0.5	22.94	34.64	44.88	51.79	55.74
L_{in}	1	17.12	29.25	42.37	51.10	54.81

Table 17: Common-rate differences computed from the displayed mean levels in Table 16; these are not new runs or additional prespecified comparisons.

Common rate	$\bar{M} - F$ (approx.)	$\bar{M} - K_2$ (approx.)	Updates
.125×	$+3.79$	-0.09	38,400
.25×	$+3.48$	-0.22	38,400
.5×	$+3.36$	-0.12	38,400
1×	$+3.08$	$+0.37$	38,400

Table 18: All reported metrics and splits for late-head-selected Mbar–K2 checkpoints. Intervals are nominal 95% paired-t. Full sign counts are transcribed in the companion summary record.

Updates / readout	R@1	R@5	MRR@10
9,600 / hold-out head	$+3.07$ $[1.83, 4.31]$	$+2.27$ $[1.45, 3.09]$	$+2.73$ $[1.70, 3.75]$
9,600 / test head	$+2.67$ $[1.52, 3.82]$	$+1.88$ $[1.12, 2.65]$	$+2.30$ $[1.33, 3.27]$
9,600 / test encoder	$+1.86$ $[1.37, 2.35]$	$+1.75$ $[1.24, 2.26]$	$+1.81$ $[1.40, 2.22]$
9,600 / hold-out encoder	$+1.56$ $[1.19, 1.94]$	$+1.59$ $[1.18, 1.99]$	$+1.64$ $[1.28, 1.99]$
38,400 / hold-out head	-0.40 $[-1.07, 0.27]$	-0.03 $[-0.59, 0.53]$	-0.27 $[-0.81, 0.27]$
38,400 / test head	-0.41 $[-0.86, 0.04]$	-0.10 $[-0.58, 0.39]$	-0.30 $[-0.70, 0.10]$
38,400 / test encoder	-0.92 $[-1.75, -0.10]$	-0.35 $[-0.88, 0.19]$	-0.70 $[-1.36, -0.04]$
38,400 / hold-out encoder	-0.70 $[-1.48, 0.09]$	-0.19 $[-0.68, 0.31]$	-0.56 $[-1.24, 0.11]$

B.5 COMPLETED LIN LR EXTENSION

The Lin extension adds 48 trajectories on the four-rate grid and reuses seeds 1200–1211, the native controls, and the selected K2 checkpoint replays. Lin uses the same intervention and matched K4 batches as the reference-rate study. Runs reach 38,400 updates with intermediate reads at 2,400/4,800/9,600/19,200 and checkpoints at 9,600/38,400. For each scored seed, choose the rate maximizing the other eleven seeds' 38,400 hold-out R@1 hit counts, breaking ties toward the higher rate. Lin selects $.25\times$ except on seed 1210 ($.5\times$). The report records 48 completions, zero failures, and exact reproduction of all twelve selected K2 replays.

The design was fixed before these Lin runs, but the control outcomes were already observed. It is a registered, reused-seed extension, not confirmatory replication. Its original three-cell rule applies the two-sided sign formula at $.05/3$; Table 19 retains all decisions. C1/C3 are two-sided; C2 predicts a positive difference. The nominal selected summaries do not provide calibrated selection inference or powered equivalence. The source's C2 description refers to positive repulsion; the manuscript interprets the implemented change using Eq. (4): it removes the entire within-positive KL term.

Table 19: All three Lin LR primary comparisons at 38,400 updates. C3 uses the same head-selected rates as C1. Numerical seed differences are supplied in the companion CSV; C1 minus C2 equals native LR-B2 for every seed.

ID	Comparison / readout	Mean [95% t]	Positive/non-tied	p_{sign}	Decision
C1	$L_{\text{in}} - K_2$, hold-out head	$+0.08$ $[-0.49, 0.64]$	6/12	1.0000	Unresolved
C2	$L_{\text{in}} - F$, hold-out head	$+4.11$ $[3.53, 4.69]$	12/12	.00049	Met
C3	$L_{\text{in}} - K_2$, test encoder	-0.49 $[-1.06, 0.08]$	3/12	.1460	Unresolved

Rate and budget sensitivity. Table 16 gives all Lin levels. At 9,600, 19,200 and 38,400, the largest mean is at $.25\times$, inside the tested grid; the first two budgets instead select the lowest tested rate. The late $.25\times$ versus $.5\times$ difference is only $+0.22$ $[-0.76, 1.21]$, so interior maxima do not identify a unique continuous optimum. The late common-rate Lin–F differences are $+4.38, +3.95, +3.64, +3.42$ pp at $.125, .25, .5, 1\times$, each positive on all twelve seeds. Lin–K2 is $+0.49, +0.25, +0.17, +0.71$, respectively. These fixed-rate contrasts complement the selected procedures, without replacing their original rules.

The next table reselects at each budget. It is distinct from reading early checkpoints at rates selected for 38,400 updates. Every non-primary budget and the Lin–Mbar comparisons are descriptive. The earlier Lin–Mbar deficit and late unresolved difference are both retained.

Table 20: Per-budget LR selection for Lin and all comparators, nominal 95% paired-t intervals. These are twelve training seeds, not 48 independent rate–seed pairs.

Updates	$L_{\text{in}} - K_2$	$L_{\text{in}} - F$	$L_{\text{in}} - \bar{M}$
2,400	$+4.15$ $[+3.62, +4.68]$	-0.74 $[-1.32, -0.16]$	-1.92 $[-2.28, -1.56]$
4,800	$+2.64$ $[+1.91, +3.37]$	-0.53 $[-1.09, +0.03]$	-1.65 $[-2.34, -0.96]$
9,600	-0.01 $[-0.55, +0.54]$	$+0.79$ $[+0.16, +1.42]$	-2.44 $[-3.09, -1.78]$
19,200	-0.09 $[-0.85, +0.68]$	$+3.53$ $[+2.93, +4.14]$	$+0.43$ $[-0.26, +1.12]$
38,400	$+0.08$ $[-0.49, +0.64]$	$+4.11$ $[+3.53, +4.69]$	$+0.48$ $[-0.08, +1.04]$

All metrics for late-head-selected checkpoints. Here all rate choices are made at 38,400 on hold-out head R@1. Official-test images are separate from the LR-selection split but were previously examined in the research program. C1 (hold-out head) and C3 (test encoder) cannot be subtracted to isolate a readout interaction. At the same official-test split the mean head–encoder difference is approximately $+1.10$ pp, but its paired interval is not identified by the marginal summaries.

Additional descriptive records. The companion retains every fixed-rate Lin–K2, Lin–F and Lin–Mbar trajectory, in-sample envelopes, and same-rate checkpoint comparison. In particular, Lin–Mbar summaries over 48 rate–seed pairs reuse twelve training seeds: the source's pair-level t inter-

Table 21: All Lin–K2 metrics at late-head-selected rates. R@1/R@5 are percentage points and MRR@10 is multiplied by 100. The companion preserves sign counts and nominal sign-formula values.

Updates / readout	R@1	R@5	MRR@10
9,600 / hold-out head	$+0.89\ [-0.45, +2.23]$	$+1.40\ [+0.57, +2.23]$	$+1.16\ [+0.05, +2.27]$
9,600 / test head	$+1.74\ [+0.68, +2.80]$	$+1.59\ [+0.93, +2.25]$	$+1.65\ [+0.78, +2.53]$
9,600 / test enc.	$+0.49\ [+0.17, +0.81]$	$+0.73\ [+0.26, +1.21]$	$+0.61\ [+0.32, +0.89]$
9,600 / hold-out enc.	$+0.06\ [-0.46, +0.58]$	$+0.44\ [-0.13, +1.02]$	$+0.28\ [-0.17, +0.73]$
38,400 / hold-out head	$+0.08\ [-0.49, +0.64]$	$+1.25\ [+0.84, +1.66]$	$+0.52\ [+0.06, +0.98]$
38,400 / test head	$+0.61\ [-0.05, +1.27]$	$+1.16\ [+0.61, +1.70]$	$+0.85\ [+0.29, +1.41]$
38,400 / test enc.	$-0.49\ [-1.06, +0.08]$	$-0.11\ [-0.53, +0.32]$	$-0.34\ [-0.79, +0.11]$
38,400 / hold-out enc.	$-0.64\ [-1.25, -0.03]$	$-0.19\ [-0.67, +0.30]$	$-0.50\ [-1.01, +0.02]$

vals are not independent-seed uncertainty estimates and are not used for inference here. At 38,400, Lin's LOSO improvement over its own $1\times$ run is $+0.81\ [0.27, 1.35]$. Comparing the old 900–911 block to the new 1200–1211 block would also change seeds and is not a paired estimate of tuning benefit. Only LR is varied; temperature, batch size, and weight decay remain fixed, and LR changes both the update scale and effective AdamW decay.

C Objectives, KL intervention and readout-specific decisions

C.1 Algebra and gradient roles

For $U \sim \mathrm{Unif}(P)$, define $L_Q(U) = \log S - z_U$ and $L_M(U) = \log(Z + e^{z_U}) - z_U$. In Q, an unsampled positive has coefficient $e^{z_p}/S > 0$; in M it has zero coefficient. Finite target sums give $\mathbb{E}\nabla L_Q = \nabla L_F$ and $\mathbb{E}\nabla L_M = \nabla L_{\bar{M}}$ at a fixed state. With $A = \sum_p e^{z_p}$, $Z = \sum_n e^{z_n} > 0$, $S = Z + A$ and $\widetilde{q}_p = e^{z_p}/A$,

$$L_F - L_{\mathrm{in}} = \log(A/m) - \frac{1}{m}\sum_p z_p = \mathrm{KL}(u\|\widetilde{q}).$$

The deleted KL derivative is $\widetilde{q}_p - 1/m$ and sums to zero over positives. Both objectives have negative coefficient e^{z_n}/S and summed positive coefficient $-Z/S$, as in Eq. (5). Individual positive coefficients are $e^{z_p}/S - 1/m$ for F and $e^{z_p}/S - \widetilde{q}_p < 0$ for L-in. Removing KL therefore redistributes attraction as well as eliminating locally repulsive positive coefficients. These are fixed-logit identities, not a prediction of parameter updates or trained retrieval.

For the averaged per-target objective,

$$\partial_{z_p} L_{\bar{M}} = -\frac{Z}{m(Z + e^{z_p})}, \qquad \partial_{z_n} L_{\bar{M}} = \frac{e^{z_n}}{m}\sum_p \frac{1}{Z + e^{z_p}}.$$

For two positives p, q within the same state, the ratios of attractive coefficient magnitudes are

$$\frac{|\partial_{z_p} L_{\mathrm{in}}|}{|\partial_{z_q} L_{\mathrm{in}}|} = e^{z_p - z_q}, \qquad \frac{|\partial_{z_p} L_{\bar{M}}|}{|\partial_{z_q} L_{\bar{M}}|} = \frac{Z + e^{z_q}}{Z + e^{z_p}}.$$

Thus the two objectives have opposite score-dependent allocations even though every positive coefficient is negative. This is a consequence of the existing objective definitions, not a learned-dynamics theorem; their total negative scale also differs. It preserves relative weights among negatives at a fixed state but changes their scale. Fclip instead uses $\widetilde{g}_p = \min(e^{z_p}/S - 1/m, 0)$ while retaining $g_n = e^{z_n}/S$. On an open region with a clipped positive, $\partial\widetilde{g}_p/\partial z_n = 0$ but $\partial\widetilde{g}_n/\partial z_p = -p_n p_p$. The total positive coefficient becomes $-Z/S - \sum_p \max(p_p - 1/m, 0)$, rather than $-Z/S$. This need not be an integrable scalar-loss gradient. Its failure concerns that intervention, not the necessity of every repulsion-related mechanism.

C.2 Objective–estimator grid and fresh clipping control

The following tables preserve the original fourth-corner levels and contrast readouts. The fourth corner was added on already-observed control seeds, not as a wholly fresh factorial. Grid1/Grid2

Table 22: The four objective–estimator contrasts with paired-t sensitivity intervals, seeds 108–115. The late Mbar–M sign decision remains unmet despite its positive mean interval.

Matched K4 contrast	9,600 updates	38,400 updates
$F - Q$	$+6.78\ [6.05, 7.50]$	$+2.08\ [1.30, 2.86]$
$\bar{M} - M$	$+5.75\ [4.92, 6.57]$	$+1.58\ [0.79, 2.37]$
$M - Q$	$+3.50\ [2.44, 4.55]$	$+4.24\ [3.19, 5.29]$
$\bar{M} - F$	$+2.47\ [1.48, 3.45]$	$+3.74\ [3.18, 4.30]$

meet the original $.05/3$ criteria; Grid3, late $\bar{M} - M$, does not. These are the original grid B1/B2/B3, renamed only for unambiguous cross-reference. Define $H = (F - Q) - (\bar{M} - M)$.

Table 23: Objective–estimator contrasts at K4, seeds 108–115; absolute levels are in the companion table G_5_1. Grid1–3 are the original three-cell family. $H = (F - Q) - (\bar{M} - M)$ is descriptive. Intervals in the detailed table are the original percentile intervals.

contrast	T	Recall@1	bootstrap interval	seeds	exact p	registered	Recall@5	MRR@10
$\bar{M} - F$	38,400	$+3.74$	$[3.31, 4.19]$	8/8	.0078	Grid1, primary	$+0.93$ (7/8)	$+2.63$ (8/8)
$\bar{M} - M$	9,600	$+5.75$	$[5.09, 6.36]$	8/8	.0078	Grid2	$+4.13$ (8/8)	$+4.98$ (8/8)
$\bar{M} - M$	38,400	$+1.58$	$[0.95, 2.17]$	7/8	.070	Grid3, unresolved	$+0.32$ (6/8)	$+1.07$ (7/8)
$\bar{M} - F$	9,600	$+2.47$	$[1.70, 3.19]$	8/8	.0078	descriptive	$+1.16$ (7/8)	$+1.92$ (8/8)
$\bar{M} - K_2$	9,600	$+5.60$	$[4.72, 6.56]$	8/8	.0078	descriptive	$+4.44$ (8/8)	$+5.14$ (8/8)
$\bar{M} - K_2$	38,400	$+0.82$	$[0.10, 1.46]$	6/8	.29	descriptive	$+0.39$ (6/8)	$+0.64$ (7/8)
H	9,600	$+1.03$		7/8	.070	descriptive		
H	38,400	$+0.50$		4/8	1.0	descriptive		

The fresh 116–123 ladder retains all six predictions, including the adverse Fclip result and unresolved late native-reference comparisons.

Table 24: Complete fresh objective/clipping family, seeds 116–123. Mean and bootstrap intervals retain their distinct definitions.

cell	contrast	T	registered	Δ	t interval	bootstrap	seeds	sign p	status
L1	$F - K_2$	9,600	> 0	$+3.67$	$[2.71, 4.62]$	$[2.96, 4.40]$	8/8	0.0078	held
L2	$F - K_2$	38,400	< 0	-3.08	$[-4.26, -1.91]$	$[-4.13, -2.35]$	0/8	0.0078	held
L3	$\bar{M} - F$	38,400	> 0	$+3.60$	$[2.75, 4.45]$	$[2.96, 4.27]$	8/8	0.0078	held
L4	$F_{\text{clip}} - F$	38,400	> 0	-1.73	$[-2.86, -0.59]$	$[-2.64, -0.92]$	0/8	0.0078	**against**
L5	$\bar{M} - K_2$	38,400	two-sided	$+0.52$	$[-0.49, 1.52]$	$[-0.25, 1.32]$	6/8	0.2891	unresolved
L6	$\bar{M} - K_2$	76,800	two-sided	$+0.26$	$[-0.41, 0.94]$	$[-0.16, 0.86]$	5/8	0.7266	unresolved

A CUDA allocation failure before any history write was rerun with the identical command/seed on another device. Late outcomes were accessed before final judging, after registration and judge fixation; the subsequent narrative was outcome-informed. Runtime history confirms clipping was exercised, unlike a placeholder zero in the initial treatment record. The auxiliary Q–M temperature block does not contain $\bar{M}$ and cannot establish temperature invariance of the full grid.

C.3 Shared-denominator intervention and every readout

Seeds 900–911 compare K2, F, $\bar{M}$ and L-in, with shared per-step hashes for K4 procedures. Saved 9,600/38,400 checkpoints reproduce stored hold-out metrics under the original TF32 feature-extraction setting. The later full-float32 probe extraction and its TF32 sensitivity check are described in Appendix C.5. Table 25 gives all six primary cells and their original $.05/6$ sign decisions. The combined matrix retains all contrasts and readouts at both budgets. The companion preserves the complete original matrices, including absolute levels.

Table 25: All six prespecified Lin-block comparisons, seeds 900–911. Y1–Y4 meet the original sign criterion at $.05/6$; Y5/Y6 do not.

ID	Contrast	Readout / updates	Mean	95% t	$n_+/12$
Y1	$F - K_2$	Hold-out head / 9,600	$+2.78$	$[2.02, 3.54]$	12
Y2	$F - K_2$	Hold-out head / 38,400	-2.77	$[-3.30, -2.23]$	0
Y3	$F - K_2$	Test encoder / 38,400	-1.34	$[-1.73, -0.94]$	0
Y4	$L_{\text{in}} - F$	Hold-out head / 38,400	$+3.56$	$[3.05, 4.06]$	12
Y5	$\bar{M} - L_{\text{in}}$	Hold-out head / 38,400	-0.25	$[-0.77, 0.28]$	4
Y6	$\bar{M} - K_2$	Test encoder / 38,400	-0.52	$[-0.84, -0.20]$	2

Table 26: All six fixed-recipe contrast readouts at both budgets, seeds 900–911. Entries are mean pp (positive/non-tied seeds). Complete absolute levels are retained in the companion tables G_16_2 and G_16_3.

Updates	Contrast	Hold-out head	Test head	Test encoder	Hold-out encoder
9,600	$F - K_2$	$+2.78$ (12/12)	$+3.00$ (12/12)	$+3.22$ (12/12)	$+2.70$ (12/12)
9,600	$\bar{M} - K_2$	$+5.19$ (12/12)	$+5.47$ (12/12)	$+4.60$ (12/12)	$+4.07$ (12/12)
9,600	$L_{\text{in}} - K_2$	$+3.47$ (12/12)	$+4.15$ (12/12)	$+2.43$ (12/12)	$+1.89$ (12/12)
9,600	$\bar{M} - F$	$+2.41$ (12/12)	$+2.46$ (12/12)	$+1.39$ (12/12)	$+1.38$ (11/12)
9,600	$L_{\text{in}} - F$	$+0.69$ (10/12)	$+1.15$ (11/11)	-0.79 (1/12)	-0.80 (1/12)
9,600	$\bar{M} - L_{\text{in}}$	$+1.72$ (12/12)	$+1.31$ (11/12)	$+2.18$ (12/12)	$+2.18$ (11/12)
38,400	$F - K_2$	-2.77 (0/12)	-2.92 (0/12)	-1.34 (0/12)	-1.61 (1/12)
38,400	$\bar{M} - K_2$	$+0.54$ (8/12)	$+0.17$ (7/12)	-0.52 (2/12)	-0.89 (1/12)
38,400	$L_{\text{in}} - K_2$	$+0.79$ (8/11)	$+0.83$ (10/12)	-0.37 (3/12)	-0.97 (2/12)
38,400	$\bar{M} - F$	$+3.31$ (12/12)	$+3.09$ (12/12)	$+0.82$ (11/12)	$+0.72$ (10/12)
38,400	$L_{\text{in}} - F$	$+3.56$ (12/12)	$+3.74$ (12/12)	$+0.97$ (12/12)	$+0.64$ (9/12)
38,400	$\bar{M} - L_{\text{in}}$	-0.25 (4/12)	-0.65 (4/12)	-0.15 (6/12)	$+0.08$ (8/12)

Y5's non-rejection does not establish identical mechanisms. Its post-hoc TOST is $p = .0047$ at ± 1 and .156 at $\pm .5$. Y6 has a descriptive mean-t $p = .0039$ and one-point TOST $p = .0032$, but does not meet the primary sign criterion. These analyses do not overwrite it. A subsequent paired reanalysis directly estimates the head–encoder differences below, replacing the earlier indirect sign-count bound as the main readout evidence. The local KL identity is the basis for the intervention claim, not a unique repulsion-only mediation argument. Two implementation-check tolerances were adjusted after initial float32/thread checks; they were not performance-judge tolerances. Shared-device times do not certify latency equivalence.

C.4 Direct paired readout analyses

For each seed, compute $I_{A-B} = (R_A - R_B)_{\text{head}} - (R_A - R_B)_{\text{enc}}$ on the same query split and the same checkpoint within each procedure. The full reanalysis below is post hoc and unadjusted. The six contrasts derive from three linearly independent contrast directions and should not be counted as six independent confirmations. The previously reported indirect early sign-count bound is superseded by these direct intervals; its proof and checks are retained in the author revision archive. At

Table 27: All head-minus-encoder R@1 interactions, reference-rate Lin block (seeds 900–911). Intervals are direct paired-t intervals, not differences of marginal interval endpoints.

Contrast	Test 9,600	Test 38,400	Hold-out 9,600	Hold-out 38,400
$L_{\text{in}} - F$	$+1.94$ $[1.44, 2.44]$	$+2.78$ $[2.10, 3.45]$	$+1.49$ $[0.83, 2.16]$	$+2.92$ $[2.08, 3.75]$
$\bar{M} - F$	$+1.08$ $[0.45, 1.70]$	$+2.28$ $[1.78, 2.77]$	$+1.03$ $[0.32, 1.74]$	$+2.59$ $[1.77, 3.41]$
$L_{\text{in}} - K_2$	$+1.73$ $[1.24, 2.21]$	$+1.20$ $[0.63, 1.76]$	$+1.58$ $[0.80, 2.35]$	$+1.76$ $[1.13, 2.39]$
$\bar{M} - K_2$	$+0.86$ $[0.39, 1.33]$	$+0.69$ $[0.12, 1.27]$	$+1.11$ $[0.45, 1.77]$	$+1.43$ $[0.91, 1.95]$
$F - K_2$	-0.21 $[-0.70, 0.27]$	-1.58 $[-2.11, -1.05]$	$+0.08$ $[-0.47, 0.63]$	-1.16 $[-2.02, -0.30]$
$\bar{M} - L_{\text{in}}$	-0.86 $[-1.41, -0.32]$	-0.50 $[-0.87, -0.14]$	-0.46 $[-0.90, -0.03]$	-0.33 $[-0.82, 0.17]$

the late test endpoint, $I_{F-K_2} = -1.58$: relative to K2, F is worse at the head than at the encoder. Consequently $I_{L_{\text{in}}-F} = I_{L_{\text{in}}-K_2} - I_{F-K_2} \simeq 1.20-(-1.58) = 2.78$. This distinguishes the gain interaction from the native-reference interaction without attributing their whole difference causally to the alternative objective. Table 4 gives the test Lin–K2 metric intervals. Their hold-out counterparts are retained in the companion table `lin_metrics.tex`: late hold-out head R@1/R@5/MRR@10 are $+0.79/+1.12/+0.95$, and hold-out encoder values are $-0.97/+0.10/-0.56$, with all intervals and counts preserved.

C.5 Completed post-hoc linear-probe evaluation

Features and fitting protocol. This analysis uses the 48 reference-rate runs on seeds 900–911, each at 9,600 and 38,400 updates (96 checkpoints). The encoder input is its raw, unnormalized 256-d pooled feature feeding the projector. The head input is the 128-d L2-normalized embedding. No backbone weights are updated. On the 45,000-image training partition, a full-batch linear classifier minimizes mean cross-entropy plus $\lambda\|W\|_F^2/2$; the bias is unpenalized. Fitting uses zero initialization, float32 Adam ($\beta_1 = .9, \beta_2 = .999, \epsilon = 10^{-8}$), initial LR .01, cosine decay to zero over 4,000 steps, and seed zero.

For each feature type, choose one λ from $\{10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$ using mean hold-out accuracy over eight predefined late checkpoints (four arms, seeds 900 and 901), with ties favoring larger λ. Both feature types select 10^{-3}, an interior value; the selected-value file was written before test evaluation. The same value is then used for all fits, including early checkpoints. Test top-1 uses the 10,000 official-test images. This post-hoc analysis includes all contrasts regardless of sign and does not replace the original retrieval decisions. The early checkpoints inherit a regularization value selected at the late endpoint. The LR-selected training checkpoints are not probed.

Table 28: Absolute test linear-probe top-1 (percent; seed standard deviations in parentheses), twelve reference-rate seeds.

Updates / features	K2	F	$\bar{M}$	L_{in}
9,600 / encoder	58.26 (0.53)	60.98 (0.29)	61.81 (0.39)	60.50 (0.42)
9,600 / head	51.80 (0.55)	54.48 (0.68)	56.29 (0.49)	55.16 (0.64)
38,400 / encoder	64.93 (0.51)	64.41 (0.32)	65.21 (0.35)	64.84 (0.48)
38,400 / head	62.95 (0.56)	61.09 (0.56)	63.11 (0.46)	63.00 (0.64)

Table 29: All encoder probe contrasts at both budgets. Entries are mean percentage points and unadjusted 95% paired-t intervals; $n_+/n_{\neq 0}$ excludes exact ties. Together with Table 30, this gives all 24 contrasts.

Contrast	9,600 [95% t]	$n_+/n_{\neq 0}$	38,400 [95% t]	$n_+/n_{\neq 0}$
$F - K_2$	$+2.72$ $[+2.39, +3.05]$	12/12	-0.52 $[-1.01, -0.03]$	2/12
$\bar{M} - K_2$	$+3.55$ $[+3.15, +3.95]$	12/12	$+0.28$ $[-0.16, +0.72]$	7/12
$L_{\text{in}} - K_2$	$+2.24$ $[+1.91, +2.57]$	12/12	-0.09 $[-0.56, +0.38]$	7/12
$\bar{M} - F$	$+0.83$ $[+0.56, +1.10]$	12/12	$+0.80$ $[+0.50, +1.11]$	11/12
$L_{\text{in}} - F$	-0.48 $[-0.69, -0.27]$	2/12	$+0.43$ $[+0.12, +0.74]$	10/12
$\bar{M} - L_{\text{in}}$	$+1.31$ $[+0.94, +1.69]$	12/12	$+0.37$ $[-0.004, +0.75]$	7/11

Multiplicity and interpretation. The 24 probe comparisons are unadjusted and include dependent contrasts: within a budget and feature type, six pairwise contrasts arise from only three linearly independent differences. Agreement in the mean sign with retrieval occurs in 21/24 cells and is descriptive, not 24 independent replications. The late encoder Lin–Mbar ordering is especially uncertain: Mbar–Lin is $+0.37$ $[-0.004, 0.75]$. Small differences from K2 are point estimates, not evidence of equivalence within half a percentage point. The weak late native encoder-probe sign result ($p = .0386$) is also not a multiplicity-corrected confirmation of the retrieval reversal.

Table 30: All head probe contrasts at both budgets. Entries are mean percentage points and unadjusted 95% paired-t intervals; $n_+/n_{\neq 0}$ excludes exact ties. Together with Table 29, this gives all 24 contrasts.

Contrast	9,600 [95% t]	$n_+/n_{\neq 0}$	38,400 [95% t]	$n_+/n_{\neq 0}$
$F - K_2$	$+2.68$ $[+2.38, +2.97]$	12/12	-1.86 $[-2.48, -1.25]$	0/12
$\bar{M} - K_2$	$+4.49$ $[+3.97, +5.02]$	12/12	$+0.16$ $[-0.29, +0.61]$	6/12
$L_{\text{in}} - K_2$	$+3.36$ $[+3.03, +3.69]$	12/12	$+0.05$ $[-0.41, +0.51]$	6/12
$\bar{M} - F$	$+1.82$ $[+1.26, +2.37]$	12/12	$+2.02$ $[+1.61, +2.44]$	12/12
$L_{\text{in}} - F$	$+0.69$ $[+0.20, +1.17]$	11/12	$+1.91$ $[+1.45, +2.38]$	12/12
$\bar{M} - L_{\text{in}}$	$+1.13$ $[+0.65, +1.61]$	12/12	$+0.11$ $[-0.46, +0.68]$	6/12

Protocol deviations and numerical sensitivity. The fitting protocol specified CPU execution, but the completed run used two NVIDIA A100 GPUs in float32 with TF32 disabled for matrix multiplication and convolution. The number of workers, thread count, feature-cache location, and output filename also differed; the analysis remained post hoc. The new raw-encoder/normalized-head feature specification here follows the completed report rather than the earlier progress summary. Full-float32 extraction did not exactly reproduce retrieval values previously computed with TF32 convolution. Re-extracting with the original TF32 setting reproduced all 96 stored checkpoint readouts.

The report repeats all 96 fits on TF32-extracted features at the same selected λ. All 24 mean directions agree with the main probe; the largest mean change is 0.0083 pp. Zero exclusion by the nominal interval agrees in 23/24 cells: late encoder Mbar–Lin changes from $[-0.004, 0.75]$ to $[+0.004, 0.76]$. This sensitivity reinforces treating that near-zero comparison descriptively. A repeat of the K2/900 late fit reproduces weights and predictions; a CPU comparison for that checkpoint has zero test-prediction disagreements. These single-checkpoint checks do not establish universal CPU/GPU identity.

The completed research report also records a separate verification of regularization selection, twenty checkpoint refits, contrast statistics, and feature hashes. Those execution checks were reported by the experiment authors, not rerun during this manuscript revision. The companion supplies five late encoder contrast arrays and the full summary transcription; it does not contain the temporary feature caches or fitted classifier weights.

C.6 Prospective late-endpoint non-inferiority and equivalence

The NI block uses seeds 600–623 for K2/$\bar{M}$, with F on the first 12, at the shared 10^{-3} LR, $V = 1$, and hold-out head R@1. The one-point margin was informed by earlier results and fixed before these new runs. Its complete level and decision tables follow.

The two endpoints are 76,800 and 153,600, not all times up to that budget. The weak single-view head results meet the one-point NI/TOST criteria. Twelve-seed path contrasts cannot be added to reconstruct the 24-seed $\bar{M} - K_2$ mean. Narrower post-hoc margins remain unresolved in part: at one-sided .025, $\pm.75$ is met at 76,800 ($p = .023$), not at 153,600 ($p = .039$); $\pm.5$ fails on the upper side at both endpoints. This leaves room for a small advantage; it does not establish half-point inferiority or exact equality.

Table 31: Complete non-inferiority and equivalence family on hold-out head R@1 at the reference LR. K2/Mbar use 24 seeds; F and contrasts involving F use only the first 12.

cell	contrast	updates	test	Δ	interval	p	status
N1	$\bar{M} - K_2$	76,800	mean > -1	$+0.15$	lower bound -0.44	0.0002	held
N2	$\bar{M} - K_2$	153,600	mean > -1	$+0.28$	lower bound -0.24	1.8×10^{-5}	held
E1	$\bar{M} - K_2$	76,800	within ± 1	$+0.15$	$[-0.44, 0.74]$	0.0033	held
E2	$\bar{M} - K_2$	153,600	within ± 1	$+0.28$	$[-0.24, 0.81]$	0.0046	held
R1	$F - K_2$	153,600	sign, < 0	-3.07	$[-3.87, -2.27]$, 0/12	0.0005	held

	38,400	76,800	153,600
$\bar{M} - K_2$ ($n = 24$)	$+0.43$ $[-0.02, 0.87]$ (15/24)	$+0.15$ $[-0.44, 0.74]$ (12/23)	$+0.28$ $[-0.24, 0.81]$ (16/24)
$F - K_2$ ($n = 12$)	-2.23 $[-2.79, -1.67]$ (0/12)	-2.80 $[-3.62, -1.98]$ (0/12)	-3.07 $[-3.87, -2.27]$ (0/12)
$\bar{M} - F$ ($n = 12$)	$+2.90$ [2.22, 3.57] (12/12)	$+3.40$ [2.71, 4.09] (12/12)	$+3.47$ [2.64, 4.30] (12/12)
K_2 level	53.97	55.84	56.76
F level	51.42	52.97	53.51
$\bar{M}$ level	54.40	56.00	57.04

D PRETRAINED RETRIEVAL AND DUPLICATE CONTROL

D.1 CUB DEVELOPMENT AND LONG-BUDGET COMPARISONS

CUB initially tests LRs $10^{-5}, 3 \times 10^{-5}, 10^{-4}$ and temperatures .05/.1 on development seeds 0–1 at 150 passes. A declared edge extension adds 3×10^{-6}. Each procedure's own development selection, and a separate shared selector, choose that lower edge at .05. Selected seeds 300–311 and shared-recipe seeds 320–327 therefore differ in seeds, not hyperparameters. Their complete primary decisions are below; complete trajectories are retained in the companion (metric pairs are R@1 / MAP@R).

Table 32: Selected CUB block: original primary decisions; complete trajectories are in companion table G_10_2. Metric pairs are R@1 / MAP@R. At 75 passes Mbar–K2 R@1 is -0.33, with two positive seeds and one exact tie (2/11).

cell	contrast	registered	Δ	t interval	seeds	sign p	status
P1	$\bar{M} - F$	> 0	$+2.97$	[2.57, 3.37]	12/12	0.0005	held
P2	$\bar{M} - K_2$	two-sided	$+0.03$	$[-0.29, 0.35]$	5/12	0.7744	unresolved

Table 33: Shared-recipe CUB block: original decisions; full trajectories are in companion table G_10_4; metric pairs are R@1 / MAP@R. Distinct final seeds have identical hyperparameters to the selected block.

cell	contrast	registered	Δ	t interval	seeds	sign p	status
C1	$\bar{M} - F$	> 0	$+2.71$	[2.32, 3.10]	8/8	0.0078	held
C2	$\bar{M} - K_2$	two-sided	-0.49	$[-0.83, -0.15]$	0/7	0.0156	resolved, negative

The shared-recipe native-reference decision has seven negatives and one exact tie. All 24 development cell means have lower 150-pass MAP@R than at 30 passes; this motivates the additional development-selected stopping analysis below. Post-hoc pooling gives $\bar{M} - K_2 = -.18$ $[-.42, .07]$ at 150 passes and $-.35$ $[-.65, -.05]$ at 75, without changing either original verdict or identifying a tuning-policy effect.

D.2 POST-HOC DEVELOPMENT-SELECTED STOPPING

At each procedure's already selected LR and temperature, choose the checkpoint maximizing mean development MAP@R across seeds 0 and 1, with ties within 10^{-9} going to the shorter read. The candidate pass budgets are 30, 75, and 150. All procedures choose 30 passes. Development and full-fitting datasets have different sizes, so checkpoint update indices differ; the full-fitting 30-pass checkpoint is 1,833 updates. This analysis chooses stopping using development data, not final-test performance, but was specified after the original fixed-endpoint study. Both seed blocks and both metrics are reported below.

Table 34: Development-selected CUB stopping: all three contrasts, both metrics and both seed blocks. Original 150-pass results are retained in Tables 32–33 and the complete companion transcription.

Block / contrast	30-pass R@1 [95% t]	30-pass MAP@R [95% t]	R@1 $n_+/n_{\neq 0}$
Selected / $\bar{M} - F$	$+1.52$ $[1.18, 1.86]$	$+1.16$ $[1.09, 1.22]$	12/12
Selected / $\bar{M} - K_2$	-0.34 $[-0.62, -0.07]$	-0.01 $[-0.12, 0.11]$	1/12
Selected / $F - K_2$	-1.86 $[-2.22, -1.50]$	-1.16 $[-1.29, -1.03]$	0/12
Shared / $\bar{M} - F$	$+1.24$ $[0.86, 1.63]$	$+1.07$ $[0.97, 1.16]$	8/8
Shared / $\bar{M} - K_2$	-0.07 $[-0.53, 0.39]$	$+0.15$ $[-0.01, 0.32]$	4/7
Shared / $F - K_2$	-1.32 $[-1.81, -0.82]$	-0.91 $[-1.05, -0.77]$	0/8

D.3 SHORT-BUDGET CUB

Development exploration preceded final seeds 400–415. All eight primary R@1/MAP@R predictions are kept, including the unmet early native-positive prediction. Multiple budget reads reuse trajectories. All trajectory pairs are R@1 / MAP@R; the short-budget source table has been reordered without changing values.

Table 35: Short-budget CUB: all primary metric comparisons. Complete trajectories are in companion table G_11_2. Metric pairs are consistently R@1 / MAP@R; numbers in parentheses are positive-seed counts.

cell	contrast	passes	registered	metric	Δ	t interval	seeds	sign p	status
SB1	$F - K_2$	1	> 0	MAP@R	$+0.15$	$[0.01, 0.30]$	10/16	0.4545	unresolved
SB1	$F - K_2$	1	> 0	Recall@1	$+0.30$	$[0.03, 0.56]$	11/16	0.2101	unresolved
SB2	$F - K_2$	24	< 0	MAP@R	-1.01	$[-1.13, -0.90]$	0/16	3.1×10^{-5}	held
SB2	$F - K_2$	24	< 0	Recall@1	-1.49	$[-1.74, -1.25]$	0/16	3.1×10^{-5}	held
SB3	$\bar{M} - F$	24	> 0	MAP@R	$+1.13$	$[1.09, 1.16]$	16/16	3.1×10^{-5}	held
SB3	$\bar{M} - F$	24	> 0	Recall@1	$+1.39$	$[1.20, 1.58]$	16/16	3.1×10^{-5}	held
SB4	$\bar{M} - K_2$	1	> 0	MAP@R	$+0.63$	$[0.51, 0.76]$	16/16	3.1×10^{-5}	held
SB4	$\bar{M} - K_2$	1	> 0	Recall@1	$+0.75$	$[0.47, 1.04]$	15/16	0.0005	held

D.4 SOP AND NONDUPLICATE-POSITIVE TRAINING

SOP expands an exploratory LR grid from two to four values and chooses the interior 3×10^{-5} at $\tau = .05$ before final seeds 500–511. Primary R@1 and MAP@R families remain separate. Full primary outcomes follow; companion trajectories use R@1 / MAP@R pairs with positive-seed counts in parentheses.

Items with fewer than four images require repeated source images at K4: 4,019 of 11,318 training items. The no-duplicate block uses the remaining 7,299 items/49,632 images, new seeds 520–531 and 5,170 updates (ten passes over that filtered population). This retains 83.34% of images, not 83.34% of items. Its full primary family follows.

The objective improvement persists without duplicate-positive items. Changed population and seeds prevent attributing the change in effect size uniquely to duplicates. In both SOP blocks, the late native-reference comparison remains unresolved rather than proving equality.

Table 36: SOP: complete primary comparisons; trajectories are in companion table G_12_2; metric pairs are R@1 / MAP@R.

cell	contrast	passes	registered	metric	Δ	t interval	seeds	sign p	status
SOP1	$F - K_2$	1	< 0	Recall@1	-3.88	$[-4.38, -3.39]$	0/12	0.0005	held
SOP2	$F - K_2$	10	< 0	Recall@1	-5.65	$[-5.97, -5.34]$	0/12	0.0005	held
SOP3	$\bar{M} - F$	10	> 0	Recall@1	$+5.61$	$[5.35, 5.86]$	12/12	0.0005	held
SOP4	$\bar{M} - K_2$	1	> 0	Recall@1	$+0.20$	$[-0.18, 0.58]$	8/12	0.3877	unresolved
SOP5	$\bar{M} - K_2$	10	two-sided	Recall@1	-0.05	$[-0.34, 0.24]$	6/12	1.0000	unresolved
SOP1	$F - K_2$	1	< 0	MAP@R	-3.56	$[-4.01, -3.12]$	0/12	0.0005	held
SOP2	$F - K_2$	10	< 0	MAP@R	-5.67	$[-6.06, -5.28]$	0/12	0.0005	held
SOP3	$\bar{M} - F$	10	> 0	MAP@R	$+5.64$	$[5.39, 5.90]$	12/12	0.0005	held
SOP4	$\bar{M} - K_2$	1	> 0	MAP@R	$+0.22$	$[-0.21, 0.64]$	9/12	0.1460	unresolved
SOP5	$\bar{M} - K_2$	10	two-sided	MAP@R	-0.03	$[-0.36, 0.31]$	6/12	1.0000	unresolved

Table 37: SOP without duplicate-positive items: complete primary family; training population and seeds differ from the original SOP block.

cell	contrast	passes	registered	metric	Δ	t interval	seeds	sign p	status
ND1	$F - K_2$	10	< 0	Recall@1	-4.77	$[-5.17, -4.37]$	0/12	0.0005	held
ND2	$\bar{M} - F$	10	> 0	Recall@1	$+4.69$	$[4.52, 4.87]$	12/12	0.0005	held
ND3	$\bar{M} - K_2$	10	two-sided	Recall@1	-0.08	$[-0.43, 0.28]$	6/12	1.0000	unresolved
ND1	$F - K_2$	10	< 0	MAP@R	-4.99	$[-5.44, -4.53]$	0/12	0.0005	held
ND2	$\bar{M} - F$	10	> 0	MAP@R	$+4.81$	$[4.64, 4.98]$	12/12	0.0005	held
ND3	$\bar{M} - K_2$	10	two-sided	MAP@R	-0.18	$[-0.64, 0.28]$	5/12	0.7744	unresolved

E Input and view boundaries

E.1 Strong-input extension

The original five-cell strong-input family is retained below. The positive late native difference at 38,400 contradicts the registered prediction; the negative 76,800 endpoint was registered as an extension after that observation. These are separate stages, not one successful initial prediction. At 76,800 $\bar{M} - K_2 = +1.55$ is descriptive. The early objective contrast remains unresolved.

Table 38: Strong-input family and separately specified extension. S5 was a conditional prediction from an earlier observed progress-based rule, not a validated universal selector. † means one exact tie: the displayed 6/8 counts six positives among eight pairs, while the sign test excludes the tie and uses seven.

contrast	T	status	Δ R@1	t interval	seeds	sign p	R@5	MRR@10
$F - K_2$	9,600	S2, held	$+8.98$	$[7.95, 10.00]$	8/8	0.0078	$+12.51$ (8/8)	$+10.29$ (8/8)
$F - K_2$	38,400	S3, **failed**	$+2.09$	$[1.21, 2.96]$	8/8	0.0078	$+2.32$ (8/8)	$+2.25$ (8/8)
$F - K_2$	76,800	S5, held (rule)	-2.38	$[-3.00, -1.76]$	0/8	0.0078	-0.46 (2/8)	-1.58 (0/8)
$\bar{M} - K_2$	9,600	desc.	$+9.10$	$[8.09, 10.11]$	8/8	0.0078	$+12.30$ (8/8)	$+10.30$ (8/8)
$\bar{M} - K_2$	38,400	desc.	$+5.14$	$[4.12, 6.17]$	8/8	0.0078	$+3.59$ (8/8)	$+4.50$ (8/8)
$\bar{M} - K_2$	76,800	desc.	$+1.55$	$[1.10, 2.00]$	8/8	0.0078	$+0.72$ (6/8†)	$+1.22$ (8/8)
$\bar{M} - F$	9,600	desc.	$+0.13$	$[-0.53, 0.78]$	4/8	1.0000	-0.21 (3/8)	$+0.01$ (4/8)
$\bar{M} - F$	38,400	S1, primary, held	$+3.06$	$[1.97, 4.14]$	8/8	0.0078	$+1.27$ (7/8)	$+2.25$ (8/8)
$\bar{M} - F$	76,800	S4, held	$+3.93$	$[3.04, 4.82]$	8/8	0.0078	$+1.18$ (7/8)	$+2.81$ (8/8)

E.2 TWO-VIEW TRANSFER

This block uses CIFAR ResNet-18 with a 3x3 stem/no max-pool, a 512–512–128 head, strong 32-pixel views, $B_{\text{src}} = 200$, $V = 2$, $\tau = .1$ and constant-rate AdamW for 18,000 updates. Seeds 700–711 are new, with reads at 5/10/20/40/80 passes. Views, batch, encoder, input and temperature change jointly relative to the main recipe. Table 40 gives all four original decisions. T1 is opposite the early prediction; T4's positive mean interval does not replace its unmet family sign criterion. The supplied results contain no $\bar{M}(K2, V2)$ arm for a same-objective K comparison.

Table 39: Two-view ResNet-18: all trajectory reads, including early negative Mbar–K2 values.

	5 passes	10 passes	20 passes	40 passes	80 passes
$F - K_2$	-1.17 (1/12)	-1.82 (1/11)	-1.27 (2/12)	-2.32 (0/12)	-2.04 (0/12)
$\bar{M} - F$	$+0.24$ (8/12)	$+0.83$ (10/12)	$+2.18$ (12/12)	$+3.81$ (12/12)	$+2.84$ (12/12)
$\bar{M} - K_2$	-0.93 (3/12)	-0.99 (3/12)	$+0.90$ (9/12)	$+1.49$ (11/12)	$+0.80$ (10/12)
K_2 level	16.59	26.90	39.21	54.10	62.55
F level	15.42	25.09	37.93	51.78	60.51
$\bar{M}$ level	15.66	25.91	40.11	55.59	63.35

Table 40: All four two-view primary comparisons. Family threshold .0125; sign-formula $p = .00635, .00049, .00049, .03857$, respectively. T4 is a prespecified two-sided comparison, not an unregistered positive finding.

ID	Contrast	Passes	Mean	$n_+/12$	Original decision
T1	$F - K_2$	5	-1.17	1	Opposite predicted positive
T2	$F - K_2$	80	-2.04	0	Predicted negative met
T3	$\bar{M} - F$	80	$+2.84$	12	Predicted positive met
T4	$\bar{M} - K_2$	80	$+0.80$	10	Unresolved

E.3 PAIRED VIEW-CONSTRUCTION INTERVENTION

The separate block keeps the strong 64-pixel conv recipe, $B_{\text{src}} = 96$ and $\tau = .05$, changing only the configured view count on seeds 800–811. X3 tests the paired difference between native effects, rather than inferring interaction from significance labels.

Table 41: Complete paired view-construction family and resource/readout checks, seeds 800–811. Changing views changes encoded work as well as positive relations.

cell	contrast	updates	registered	Δ	t interval	seeds	sign p	status
X1	$F - K_2$, one view	9,600	> 0	$+8.81$	$[8.16, 9.46]$	12/12	0.0005	held
X2	$F - K_2$, two views	9,600	two-sided	$+0.18$	$[-0.66, 1.01]$	6/12	1.0000	unresolved
X3	$(F - K_2)_{1v} - (F - K_2)_{2v}$	9,600	> 0	$+8.63$	$[7.73, 9.53]$	12/12	0.0005	held

	4,800	9,600
$F - K_2$, one view	$+5.90$ (12/12)	$+8.81$ (12/12)
$F - K_2$, two views	-0.23 (7/12)	$+0.18$ (6/12)
$(F - K_2)_{1v} - (F - K_2)_{2v}$	$+6.12$ (12/12)	$+8.63$ (12/12)
K_2: two views minus one	$+19.47$ (12/12)	$+22.33$ (12/12)
F: two views minus one	$+13.35$ (12/12)	$+13.70$ (12/12)
level, K_2, one view	9.13	17.22
level, F, one view	15.02	26.03
level, K_2, two views	28.60	39.55
level, F, two views	28.37	39.73

X2's post-hoc ± 1 TOST $p = .027$ meets nominal .05, not $.05/3$. Adding a view also doubles encoded work, anchors and negative candidates and changes BN inputs. Thus it is not a pure positive-

count intervention. Equal encoded presentations at unequal update counts is only a descriptive resource check. No late view-interaction experiment is supplied.

F SENSITIVITY, RESOURCE ACCOUNTING AND AUDIT

Earlier probe and secondary metrics. Three earlier seed pairs were retrospectively evaluated by head/encoder retrieval and multinomial logistic regression on the official test set after checkpoint verification. Feature-type regularization was selected on the hold-out; these are not the newer Y3/Y6 comparisons. The full source also reports an encoder-R@5 readout with mixed seed directions. Fresh CIFAR's late native R@5 and fresh disjoint CUB's late R@5 fail their original sign targets even though MRR is negative. These metric boundaries are not overwritten by the primary R@1 findings.

Seed and class uncertainty. The companion table E_13_1 preserves the complete post-hoc seed-and-class sensitivity matrix. Its shared-class and pigeonhole-bootstrap methods retain a fixed gallery and require the original seed-by-class arrays for reproduction. They do not certify new training classes or a newly sampled gallery. In particular, the shared-CUB native-reference sign is sensitive to the uncertainty model. These intervals do not replace the original tests or the separate prospective NI block. With the joint bootstrap, the fixed-endpoint selected-CUB Mbar–K2 interval is $[-0.49, 0.58]$ and the shared-CUB interval is $[-1.12, 0.10]$. SOP Mbar–F remains positive at $[5.32, 5.88]$. These are checks of the fixed evaluation population, not new-class generalization guarantees.

Fixed-recipe pooled sensitivity. A post-hoc random-effects analysis combines four non-overlapping CIFAR seed blocks at 38,400 updates, the shared 10^{-3} LR, weak single-view input, and hold-out head R@1. Only these matching protocols are pooled. The block means for Mbar–K2 are $+0.82$ (8 seeds), $+0.52$ (8), $+0.43$ (24), and $+0.54$ (12). The reported DerSimonian–Laird estimate is $+0.53$ pp with normal interval $[0.25, 0.81]$, estimated $\tau^2 = 0$, and $I^2 = 0$. The modified HKSJ interval is $[0.07, 0.98]$, and the prediction interval is $[-0.09, 1.14]$. Seed pooling gives $+0.53$ $[0.24, 0.81]$, 35/52 positive. This is evidence for a small positive fixed-recipe head difference, not a result for LR-selected procedures or other representations. The separate five-block native pool is -2.66 $[-2.95, -2.38]$. A reported NI seed-by-class sensitivity gives $[-0.58, 0.87]$ and $[-0.40, 0.96]$ at its two endpoints; these are post-hoc checks, not replacements for the original TOST.

Implementation cost. A weak-input single-A100 benchmark reports F/$\bar{M}$ at 15.2 ms/step, Q at 24.1 and M at 35.3, each with 927 MiB peak allocation. Two short repetitions discard the first 30 steps and summarize 269 intervals. Separate loop measurements do not give an exact additive time decomposition. Pretrained SOP reports 8.06 GB allocated/13.41 GB reserved, but timing clusters by device load. L-in/F's 27.85/27.81 ms shared-device times are not controlled latency equivalence. Source-image matching does not equalize encoded work, tuning cost or end-to-end wall time.

Registration and source audit. A historical research-program snapshot, prior to the 96-run LR block, records 52 judges, 352 predictions, 121 failed/unresolved predictions and 2,619 runs. Registration categories are 45 before block runs, three after partial outcomes and four pre-analyses of existing runs. Judge histories are 19 sealed-before-data/run-once, one identical-code rerun, eight partial-block reads, ten edited after some data (six affecting verdicts or guards) and fourteen written after data existed. These are not all clean prospective confirmations. The 96-run native LR block, the 108-run Mbar/lower-rate extension, the 48-run Lin extension, and the post-hoc probe are listed separately in Table 42. The historical census is not relabeled as an updated total; a complete current census needs the full repository export.

The full research-program census also records 75 failed or unresolved predictions that were not described as failures in the earlier narrative. The old grouping-policy control has an adverse implemented-policy contrast and an unresolved randomized replacement; it is not used as a mechanism of native K. Query-transition detail, geometry, resource-changing decompositions, policy replay and complete correction/judge histories are preserved in the author audit archive. Their exclusion from the submission supplement reflects the current questions, not favorable signs. The comparison families used in this paper retain contrary and unresolved cells; the extra records are not treated as additional confirmations.

G EVIDENCE PROVENANCE AND COMPANION CONTENTS

The table distinguishes new-seed studies, additions to observed controls, and post-hoc analyses. Internal pre-execution fixation does not establish independently verifiable external preregistration. The companion transcribes the completed result reports, including all Lin LR tables and all 24

Table 42: Family-level evidence map. This is not a reconstructed program-wide census or a claim that every historical evaluation was prospectively fixed.

Family	Seeds / unit	Evidence status	Role
Native replication	108–115; later 116–123	Independent seed blocks; late ladder values read before judging	Native reversal; adverse clipping retained
Lin intervention	900–911 (12)	New-seed six-cell family	Fixed-recipe objective and native contrasts
Native LR	1200–1211 (12)	New seeds; four-cell family; early endpoint informed by prior data	Fixed-rate and LOSO comparisons
Mbar / low-rate addition	1200–1211 (12 reused)	Fixed before added runs; controls already observed	LR-selected comparison and boundary extension
Lin LR extension	1200–1211 (12 reused)	Three-cell design fixed before 48 added runs; observed controls reused	LR-selected Lin comparison
Linear probe	900–911; 96 checkpoints	Post-hoc; fit protocol fixed, GPU deviation disclosed	Frozen-feature classification
NI / equivalence	600–623 (24); F on 12	Margin and endpoints fixed before new runs	Reference-rate hold-out head equivalence
CUB / SOP	Separate final seed blocks	Per-procedure development choice before final fitting	Class-disjoint objective comparisons
Strong / two-view	Separate blocks	Original failures and later extension distinguished	Scope of native and objective effects
Paired readouts	Existing Lin checkpoints	Post-hoc; same-query paired analysis	Readout interaction and other metrics
CUB stopping	Existing development/final reads	Post-hoc, using development MAP@R	Stopping sensitivity, both seed blocks
Bootstrap / pooling	Existing LR or matching blocks	Post-hoc; explicit resampling/pooling rule	Selection and fixed-recipe sensitivity

probe contrasts. It supplies numeric seed arrays for the native primary LR contrasts, the Mbar and Lin primary extensions, and five late encoder-probe contrasts. These arrays support recomputation of their means, intervals, sign-formula values, and matched contrast identities. Other paired interactions, stopping analyses, and bootstrap/pooling results are reported analyses rather than newly re-estimated raw-array results.

The completed Lin LR and probe report supersedes the earlier progress snapshot. Probe protocol deviations are reported in Appendix C.5. The historic trainer, complete executed configurations, split manifests, raw per-rate/per-readout arrays, and feature caches are not included in this companion. The numerical transcriptions and independent loss implementation provide an audit of the supplied results, not a full training reproduction package.