FORGOTTEN COMPARED TO WHAT? A CALIBRATED AUDIT OF CLASS UNLEARNING*

Anonymous authors
Paper under double-blind review

ABSTRACT

Machine unlearning aims to make a trained model behave like one retrained without designated data. Whether it succeeds depends on what we measure. A model may stop predicting a removed class even though its scores still distinguish that class. Yet finding such a signal is not enough: a model trained without the class may also recognize its images through generalization. The relevant question is therefore: forgotten compared to what? We introduce a calibrated audit that measures recovery from an unlearned model relative to models trained without the removed class, at a fixed operating point. Our main test, woken recall, measures recall after calibrating a single logit offset on retained data at a nominal 5% false-positive rate. We use the existing classifier and choose the offset using only retained-class examples. In a CIFAR-10 study of 480 checkpoints from 16 fixed unlearning configurations, models with near-zero forget accuracy often have higher removed-class recall than retain-only references matched on retained accuracy. A frozen-feature probe gives a separate, more supervised view and leads to the same conclusion. Recovery has to be read against a stated reference and operating point; forget accuracy alone cannot provide that comparison.

1 INTRODUCTION

Machine unlearning aims to remove designated training data from an already trained model without retraining from scratch (Cao & Yang, 2015; Bourtoule et al., 2021). Methods now alter gradients, parameters, decision boundaries, teachers, or representations (Golatkar et al., 2020; Chundawat et al., 2023a; Chen et al., 2023; Kurmanji et al., 2023; Fan et al., 2024; Foster et al., 2024), but share an ideal reference: a model trained without the removed data.

Evaluation is therefore part of the problem. Class-unlearning studies commonly report *forget accuracy* and *retain accuracy*, while recent work has questioned what such measurements, verification tests, representation analyses, and normalization-sensitive evaluations establish (Thudi et al., 2022; Zhang et al., 2024a; Hayes et al., 2025; Kim et al., 2026; Gao et al., 2026; Yu et al., 2026; Kalani et al., 2026). A checkpoint can look forgotten under one measurement and different under another.

Consider a classifier asked to forget cats. It can stop predicting *cat* while still assigning cat images systematically stronger cat scores than retained images. A change to the class bias alone can create this behavior without changing the features or the ordering of scores (Zheng et al., 2026). Zero forget accuracy can therefore hide class discrimination that is still present in the checkpoint. Finding recoverable discrimination is only half of the comparison. A model trained without cats can learn features from other classes that also separate cat images. Recovery therefore does not tell us whether unlearning left something unusual behind unless we compare it with a model that never received the cat training examples. This is the question in our title: *forgotten compared to what?*

The audit compares a frozen unlearned checkpoint, the target, with a retain-only reference. The auditor applies the same recovery procedure to both at a specified operating point. *Excess recovery* is the target's recall minus the reference recall. We call the resulting removed-class recall *woken recall*: the fraction of held-out removed-class images detected after a single logit offset calibrated on retained images at a nominal 5% false-positive rate. We fit no new classifier and use only retained images

*Code is available at `https://anonymous.4open.science/r/forgotten-3F5C`

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.

to choose the offset. Calibrating each model separately gives both the same nominal false-positive budget and makes each model's woken recall invariant to constant shifts in its removed-class logit.

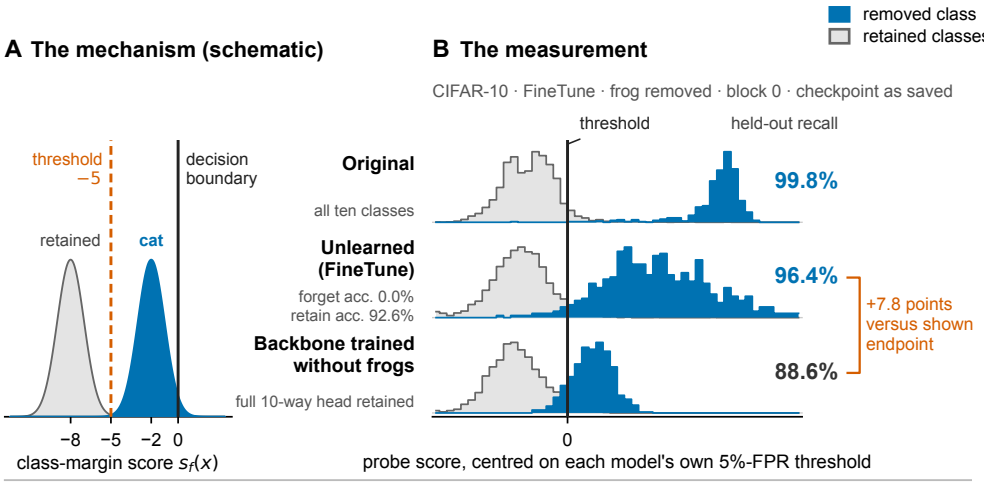

Figure 1: **Score suppression and recovery relative to retraining.** (A) Schematic margins show how a scalar correction exposes ordering below the prediction boundary. (B) FineTune on frog removal has zero forget accuracy and 96.4% probe recall, compared with 88.6% for the displayed retain-only endpoint. The probe gap is $+7.8$ points against that endpoint and $+9.8$ against 13 utility-matched references. Figure 2 shows the comparison using existing margins for the same checkpoint.

For the FineTune frog-removal checkpoint in Figure 1, forget accuracy is zero, but the existing classifier margin gives a +37.8-percentage-point gap over matched references (Figure 2). A separately fitted frozen-feature probe gives a +9.8-percentage-point gap. Because the readouts use different supervision, each is compared with references evaluated under the same protocol.

A completed retraining run is the obvious reference. If unlearning damages retained accuracy, an intermediate retain-only checkpoint provides a comparison at similar utility. We use both: their conclusions are nearly identical at high utility but diverge as utility falls.

We study existing scores and frozen-feature probes in 480 CIFAR-10/ResNet-18 checkpoints from 16 fixed configurations. We also examine reference choice and BatchNorm recalibration, with additional experiments on CIFAR-100 and a pretrained ViT.

2 A REFERENCE-RELATIVE AUDIT OF CLASS UNLEARNING

2.1 LOOKING BELOW THE WINNING LABEL

Let $D = D_\mathrm{r} \cup D_\mathrm{f}$, where D_f contains class-c training examples and D_r the retained examples. With training and update randomness explicit, unlearning produces $f_u = \mathcal{U}(\mathcal{A}(D;\xi), D_\mathrm{f}, D_\mathrm{r}; \zeta)$, whereas retraining produces $g = \mathcal{A}(D_\mathrm{r}; \xi')$. The ideal target compares their distributions over randomness (Bourtoule et al., 2021); our audit compares a declared recovery behavior.

For logits $f(x) \in \mathbb{R}^K$ and prediction $\hat{y}_f(x) = \arg\max_k f_k(x)$, write

$$F(f) = \Pr_{X_c}[\hat{y}_f(X_c) = c], \qquad R(f) = \Pr_{(X_r,Y_r)}[\hat{y}_f(X_r) = Y_r]. \tag{1}$$

Here $X_c \sim \mu_c$ is a held-out removed-class example and $(X_r, Y_r) \sim \mu_r$ a retained example. Forget accuracy records whether c wins, but not how close it was to winning. We therefore compare the removed logit with its strongest competitor:

$$s_f(x) = f_c(x) - \max_{k \neq c} f_k(x). \tag{2}$$

The classifier predicts c when this *class margin* is positive.

Figure 1A gives the simplest case. Suppose every cat image has margin -2 and every retained image has margin -8. Forget accuracy is zero, yet a threshold at -5 separates the two groups perfectly. Equivalently, adding 5 to the cat logit makes the existing classifier recover that distinction. The question is now quantitative: how far may we move this threshold before ordinary retained images begin to look like the removed class?

2.2 Fixing the cost of recovery

A recovery number is meaningless without fixing that cost. If we increase the removed-class logit enough, every image will eventually be predicted as the removed class. We therefore choose the correction from retained images and constrain its false-positive rate (FPR).

Let $\widehat{\tau}_{f,\alpha}$ be the $(1-\alpha)$ quantile of margins on retained calibration images. At our primary $\alpha = 0.05$, this is the 95th percentile. For removed-class evaluation set S_c^{eval}, define

$$\widehat{W}_\alpha(f) = \frac{1}{|S_c^{\mathrm{eval}}|} \sum_{x \in S_c^{\mathrm{eval}}} \mathbf{1}\{s_f(x) > \widehat{\tau}_{f,\alpha}\}, \qquad \delta_{f,\alpha} = -\widehat{\tau}_{f,\alpha}. \tag{3}$$

We call $\widehat{W}_\alpha(f)$ *woken recall*. Adding $\delta_{f,\alpha}$ to the removed logit makes the label win exactly when its margin crosses the calibrated threshold. Retained images alone choose this scalar; removed-class examples are used only for held-out recall. No feature direction or classifier weight is fitted.

Calibration turns Figure 1A into a controlled measurement: we ask how much recovery appears while allowing a declared fraction of retained images to cross the boundary. We report forget and retain accuracy before correction. Sweeping the threshold gives the ROC; Appendix E.1 reports the reversed-score route. Woken recall gives a well-defined recovery value for one checkpoint, but that value alone does not tell us whether the checkpoint is unusual. A model trained without the removed class may also achieve substantial recall. The reference supplies that comparison.

2.3 Recovery compared with retraining

For every audited target, we run the same calibration on retain-only references. Thresholds are chosen separately because different models can have different score scales and offsets, but the nominal FPR is shared. Let $\mathcal{G}$ be a declared reference set for the same removed class and normalization condition. We define

$$\widehat{G}_{\mathrm{woken},\alpha}(f;\mathcal{G}) = \widehat{W}_\alpha(f) - \operatorname*{median}_{g \in \mathcal{G}} \widehat{W}_\alpha(g). \tag{4}$$

A ten-point gap means that, at the same nominal FPR, the target recovers ten percentage points more of the held-out class than the reference median. This subtraction is essential: our retain-only references themselves support removed-class recognition (Figure 1), so absolute recovery alone does not establish excess recovery over retraining.

Which retain-only model should we use? A fully trained model is the natural endpoint because it represents completed retraining. But an unlearning update can lower retained accuracy. Comparing such a target with a much stronger endpoint mixes two differences: forgetting and utility. We therefore add intermediate checkpoints from the retain-only training trajectories and, when possible, compare at similar retained accuracy.

Each removed class has a *reference ladder* containing three retain-only training seeds and 29 checkpoints per seed. Write $\widehat{R}$ for test retain accuracy and $\widehat{H}_{a,\alpha}$ for held-out recall from readout a, with $\widehat{H}_{\mathrm{bias},\alpha} = \widehat{W}_\alpha$. For saved or BN-recalibrated condition v, define

$$\widehat{\mathcal{N}}_{c,v}(q) = \{g_{c,s,e,v} : |\widehat{R}(g_{c,s,e,v}) - q| \le 2 \text{ percentage points}\},$$

where s and e index reference seed and epoch. The matched gap is

$$\widehat{G}_{a,\alpha}(f_v) = \widehat{H}_{a,\alpha}(f_v) - \operatorname*{median}_{g_v \in \widehat{\mathcal{N}}_{c,v}(\widehat{R}(f_v))} \widehat{H}_{a,\alpha}(g_v). \tag{5}$$

We require at least five checkpoints and compute each target's gap before aggregation. Utility matching does not imply equivalent representations (Appendix A.2), so endpoint and matched

Table 1: **Two readouts available to the auditor.** Both readouts use frozen checkpoints, the same retained calibration images, and disjoint recovery evaluation images. Each reference receives its own readout and threshold under the corresponding protocol. Counts are for one CIFAR-10 deletion.

	Woken recall	Frozen-feature probe
Score	Existing class margin	Fitted linear feature score
Fitting examples	None	350 removed, 3,150 retained
Calibration	1,350 retained images	1,350 retained images
Adaptation	One logit offset	Scaler, probe, threshold

comparisons remain distinct. Throughout, hats denote empirical quantities in formal definitions. For readability, figures, tables, and descriptive result text suppress hats when the empirical interpretation is unambiguous. The class margin looks below the winning label, calibration fixes the cost of recovery, and retraining supplies the baseline. Because calibration depends on score ordering rather than absolute class bias, pushing the removed logit downward cannot erase the measured signal.

2.4 Suppressing the class bias cannot defeat the audit

A constant shift of the removed logit translates every margin and its retained calibration threshold by the same amount:

$$s_{f^\delta}(x) - \widehat{\tau}_{f^\delta,\alpha} = s_f(x) - \widehat{\tau}_{f,\alpha}.$$

Woken recall and the ROC therefore remain unchanged.

Proposition 1 (Recovery invariance under class bias shifts)**.** *Let $f_c^\delta = f_c + \delta$ and $f_k^\delta = f_k$ for $k \neq c$. On a finite audit set $S = S_f \cup S_r$ containing the removed and retained evaluation examples, with finite logits, a sufficiently negative δ yields zero empirical forget accuracy without reducing empirical retain accuracy or changing the feature extractor. For every threshold t,*

$$\{x : s_{f^\delta}(x) > t + \delta\} = \{x : s_f(x) > t\}.$$

Thus a threshold sweep, its ROC curve, and its ranking AUC are invariant to the bias shift.

Appendix A proves the result, including calibrated woken recall and fixed-reference gaps. If a bias edit changes ordinary retain accuracy, rematching can select a different reference neighborhood. The proposition therefore isolates the point illustrated in Figure 1: making the removed label stop winning is not enough to change the ordering that woken recall measures. The next section turns this audit into an experimental protocol and applies the same logic to a second, supervised feature readout.

3 Experimental design

Models and fixed configurations. The systematic study uses CIFAR-10 (Krizhevsky, 2009) with a CIFAR-adapted ResNet-18 (He et al., 2016). We reserve 5,000 validation images before training. Three originals train on the remaining 45,000 images with seeds 0, 1, and 2; each defines a *block*. For each deleted class, three retain-only models train on 40,500 images with the same 200-epoch recipe and ten-way head. The removed output receives no positive training examples, although retained examples and weight decay can update it.

The 16 methods span teacher, gradient, parameter-editing, boundary, and representation approaches (Chundawat et al., 2023a; Golatkar et al., 2020; Chen et al., 2023; Tarun et al., 2024), including SalUn (Fan et al., 2024), SSD (Foster et al., 2024), SCRUB (Kurmanji et al., 2023), and POUR-P (Le et al., 2026). Cat/block-0 validation selects one configuration for each of 15 methods; POUR-P uses its closed-form configuration. Each fixed setting transfers unchanged to ten classes and three blocks, yielding 480 checkpoints (Appendices B, F).

Audit data. Output accuracies use all 10,000 test images. Disjoint recovery subsets assign 3,500 images to probe fitting, 1,500 to threshold calibration, and 5,000 to evaluation. Woken recall uses 1,350 retained calibration images to set its offset. Targets and references share partitions; recovery never selects unlearning configurations.

Supervised feature recovery. We fit a binary logistic probe to frozen penultimate features (Alain & Bengio, 2017). Each checkpoint receives a separate scaler and probe, fitted on the same labeled examples from both groups, followed by threshold calibration on retained data and held-out evaluation (Table 1). Reference backbones exclude the class during training; their probes receive the removed-class labels at audit time. The probe can thus learn which features learned from retained data distinguish the class. Appendix B specifies the logistic regression, and Figure 5 shows its workflow.

The readouts differ in adaptation and supervision: the linear-head margin test uses an intersection of fixed halfspaces, whereas the probe learns one halfspace. Their decision families are not generally nested (Appendix A.3); we report them separately.

Reference support and analysis history. Matching uses all 9,000 retained test images, including the recovery evaluation negatives. Comparator selection therefore shares evaluation data even though probe fitting and threshold calibration remain disjoint from recovery evaluation. The class-specific 87-checkpoint ladders support 372 saved and 382 recalibrated targets. We report all supported outcomes and the descriptive $F \leq 1\%$, $R \geq 90\%$ cohort, which represents eleven methods as saved and eight after recalibration. Shared images and reference trajectories make counts dependent.

The study was specified around the frozen-feature probe; cross-route analyses are post hoc. Woken gaps use matched checkpoints, while endpoint, trajectory, CIFAR-100, and ViT checks use the probe.

Paired normalization control. BatchNorm (BN) (Ioffe & Szegedy, 2015) running statistics can become inconsistent with activations after unlearning (Kalani et al., 2026). We evaluate saved checkpoints and repeat the audit after BN recalibration on retained data (Izmailov et al., 2018). For learned parameters θ and buffers ϕ, the intervention is $\mathcal{B}_{D_r}(\theta, \phi) = (\theta, \widehat{\phi}(\theta; D_r))$. Targets and references receive the same pass over retained data, followed by score extraction, probe refitting, threshold calibration, and reference matching. Appendix B specifies the pass.

4 Results

4.1 Zero forget accuracy can hide substantial excess recovery

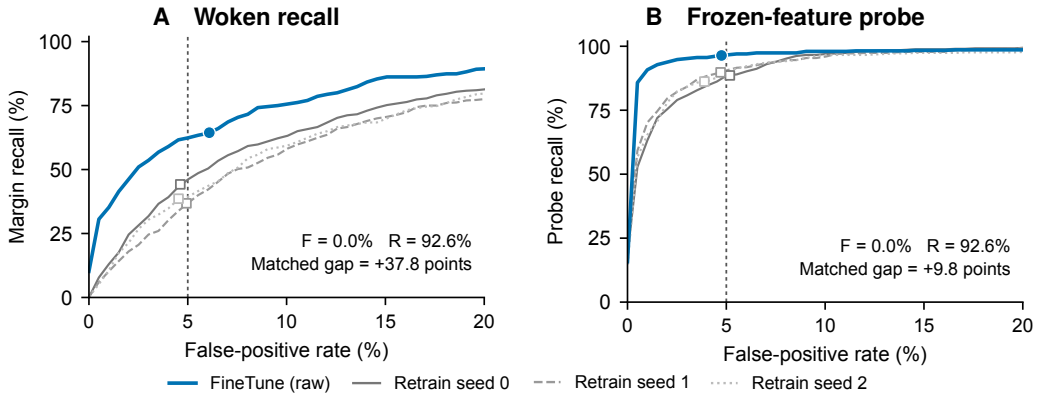

Figure 2: **One calibrated offset exposes excess recovery.** FineTune, frog removal, block 0, as saved: $F = 0\%$, $R = 92.6\%$. The existing margin yields a $+37.8$-point matched reference gap; a separately fitted probe yields $+9.8$. Gray curves show three fully trained endpoints, while the annotations use the matched ladder. Markers show calibrated recall at realized FPR; the vertical line marks nominal 5%. Figure 8 includes both normalization conditions.

We begin with the checkpoint from Figure 1. FineTune removes every correct frog prediction while retaining 92.6% accuracy on the other classes. By forget accuracy, there is nothing left to recover. Woken recall gives a different reading. At the same nominal 5% FPR used for retain-only references, the checkpoint recovers 37.8 percentage points more frog images than its matched reference median (Figure 2). After BN recalibration, forget accuracy remains zero and the gap increases to $+41.4$ points.

Table 2: **Excess from existing scores across the CIFAR-10 study.** Nominal 5% FPR; gaps use matched references. Median gaps are in points. The last row gives conservative lower bounds on *jointly* positive woken and probe gaps in the $F \leq 1\%$, $R \geq 90\%$ cohort, derived from aggregate counts in Appendix E.2. These bounds are not exact subgroup counts or confidence limits.

Measurement	As saved	BN recalibrated
Supported targets / all targets	372/480	382/480
Positive woken gaps / supported	347/372	347/382
Median woken gap	$+49.45$	$+55.10$
Positive probe gaps / supported	347/372	358/382
Median probe gap	$+10.25$	$+13.00$
Both gaps positive / supported	328/372	335/382
Both positive in low-forget, high-utility cohort	$\geq 244/264$	$\geq 166/189$

Woken gaps are positive in 347/372 supported saved checkpoints and 347/382 after BN recalibration, with median gaps of $+49.45$ and $+55.10$ points (Table 2). Woken recall learns no new feature direction: retained images choose only a scalar correction to existing scores.

The same pattern survives where conventional evaluation looks strongest. Among checkpoints with both low forget accuracy and high retained utility, at least 244/264 saved and 166/189 recalibrated outcomes have positive gaps under both woken recall and the feature probe. Thus the phenomenon is not driven only by methods that achieve forgetting by destroying the classifier.

The two recovery routes are not interchangeable. Their gaps differ sharply in Table 2, revealing what the probe adds.

4.2 A LEARNED READOUT CHANGES THE AMOUNT OF RECOVERY

Woken recall asks what can be recovered through the classifier that already exists. The frozen-feature probe receives considerably more help: it learns a new direction from labeled removed and retained examples. On the FineTune checkpoint above, this fitted readout reaches 96.4% recall, but its matched reference is already highly recoverable. The resulting excess is therefore only $+9.8$ points, compared with $+37.8$ from the existing margin. This apparent reversal is useful. Fully trained retain-only endpoints have median woken recall of 36.4% as saved and 26.9% after recalibration, whereas their probe recalls are 88.4% and 82.0%. The probe raises recovery for the target, but also for a backbone trained without the removed class. Raw recovery can therefore increase while *excess* recovery decreases. This is precisely why the reference cannot be omitted.

Probe fitting also depends on labeled-data budget: in an airplane-class check, increasing the fitting set from 1,750 to 4,000 images reduces the median gap from 10.2 to 7.8 points across 35 outcomes (Appendix E.5). Woken recall has no fitting stage.

Despite these differences, the two routes usually agree where retained utility is high. The probe gap is positive in 263/264 saved checkpoints with $F \leq 1\%$ and $R \geq 90\%$, and in all 189 such checkpoints after recalibration. Woken-only disagreements are concentrated at lower utility: at most one saved woken-only positive and no recalibrated woken-only positive has retain accuracy above 90% (Appendix E.2). This agreement leaves a more difficult possibility. Perhaps the positive probe gaps come from how we chose the retraining reference rather than from the audited checkpoints themselves.

4.3 NEAR RETRAINING-LEVEL UTILITY, THE REFERENCE CHOICE BARELY CHANGES THE RESULT

The main probe comparison uses retain-only checkpoints within two retained-accuracy points of each target. This controls for utility, but it also raises an obvious concern: an intermediate checkpoint is not the gold-standard endpoint usually meant by retraining. For the high-utility checkpoints that matter most, replacing the matched reference with completed retraining changes little. Among the $F \leq 1\%$, $R \geq 90\%$ cohort, matched references give positive probe gaps in 263/264 saved outcomes and 189/189 recalibrated outcomes. Fully trained reference medians give exactly the same positive

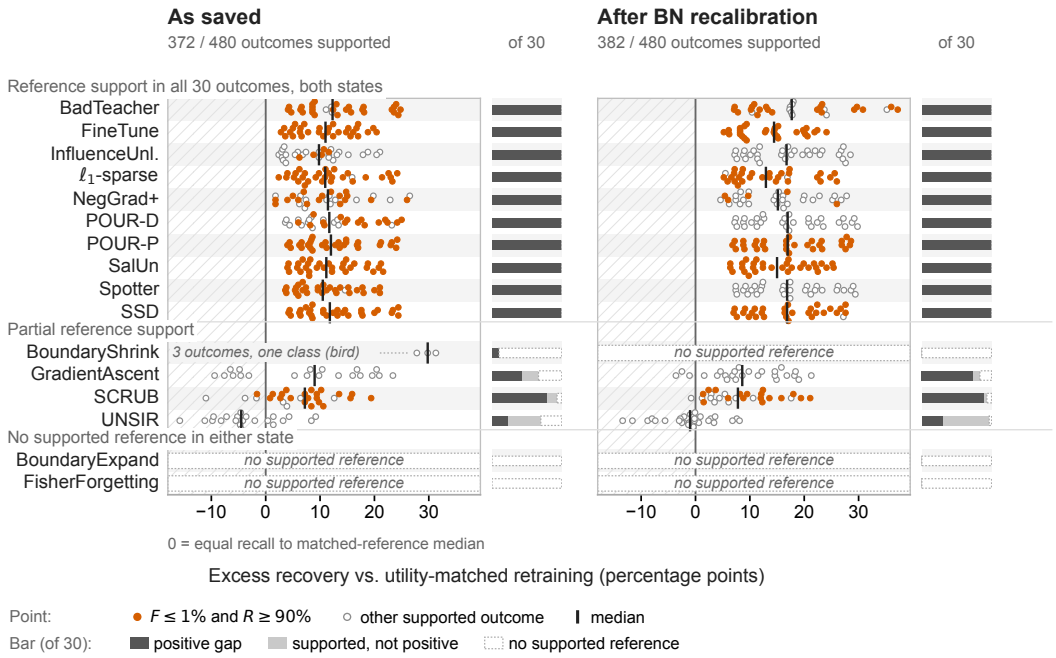

Figure 3: **The supervised feature comparison across methods.** Each point is a supported class/block outcome at its matched probe gap. Filled points satisfy $F \leq 1\%$, $R \geq 90\%$; bars show positive gaps, nonpositive gaps, and unsupported outcomes out of 30. Each normalization condition determines its own support and cohort.

Table 3: **High-utility probe excess survives reference choice.** Fixed cohorts satisfy $F \leq 1\%$, $R \geq 90\%$. Cells give positive counts and median recall gaps in points. Endpoint rows use the median or maximum recovery of three fully trained references without matching.

Reference	Saved: n_+/n (gap)	BN: n_+/n (gap)
Matched checkpoints	263/264 (+10.7)	189/189 (+14.0)
Fully trained median	263/264 (+9.0)	189/189 (+13.0)
Fully trained maximum	260/264 (+7.8)	187/189 (+12.2)

counts (Table 3). Even comparing each target with the highest-recovery fully trained reference leaves 260/264 and 187/189 gaps positive.

Ten methods have positive probe gaps in every class/block pair in both normalization conditions (Figure 3). FineTune, SalUn, and POUR-P are especially clean cases: all thirty outcomes in both conditions have exactly zero forget accuracy, retain accuracy above 90%, and positive probe gaps. Equal-seed weighting and one- or three-point matching windows preserve the high-utility signs. Held-out retain-only trajectories provide a scale for ordinary reference variation: their median high-utility gap is only $+0.1$ points, and the audited gap exceeds the corresponding 95th-percentile control scale in 241/264 saved and 185/189 recalibrated cohort outcomes.

So why keep the matched reference ladder at all? Figure 4 gives the answer.

At retained accuracy above 90%, the fully trained endpoint exceeds the local matched baseline by only 0.8 median points as saved and 0.5 after recalibration. Below 85%, those differences grow to 13.4 and 9.6 points. In that regime, replacing the matched comparator with the endpoint can materially reduce the measured gap. Coverage also becomes sparse: only 45/153 saved and 47/145 recalibrated low-utility targets have a supported local comparison. The reference constructions agree when utility is preserved but diverge as utility deteriorates, which also affects how checkpoint perturbations should be interpreted.

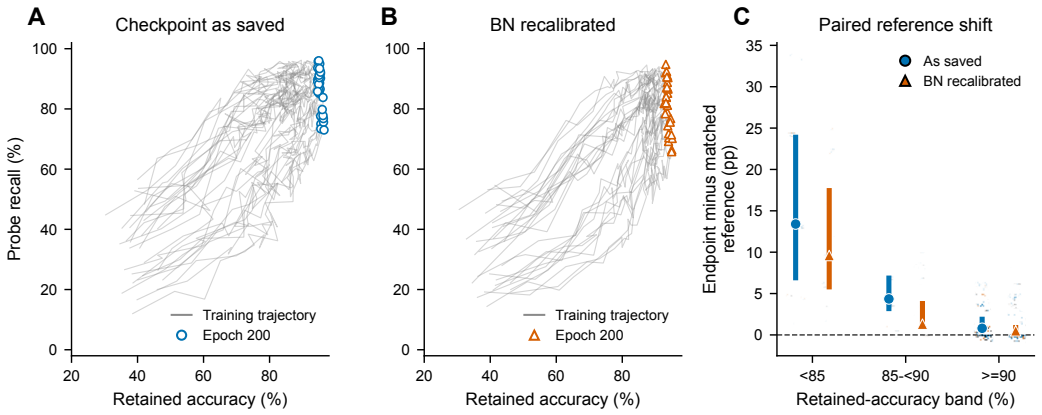

Figure 4: **Probe baselines differ more at lower utility.** (A, B) Thirty retain-only trajectories connected in epoch order; hollow markers denote fully trained endpoints. (C) Endpoint minus matched reference probe recall for supported targets. Large markers and bars show band medians and interquartile ranges.

4.4 Normalization can move both sides of the comparison

BatchNorm recalibration provides a useful stress test because it changes running statistics while leaving learned parameters fixed. For several methods, that small intervention radically changes the prediction-level appearance of forgetting. Spotter's median forget accuracy rises from 0% to 93.2%, POUR-D's from 1.0% to 95.25%, and InfluenceUnlearning's from 7.65% to 93.75%, while retain accuracy remains above 93.5%.

Yet the recovery comparison does not simply disappear. Woken gaps remain widespread after recalibration (Table 2). More importantly, recalibration moves the reference as well as the target. Among 365 outcomes supported in both conditions, median target probe recall falls by 0.4 points, while matched reference recall falls by 4.0 points. The paired probe gap consequently rises by 3.5 points. These marginal medians need not subtract exactly.

A larger post-recalibration gap may not mean the target contains more recoverable information: in the typical pair, the reference moved more. The audit therefore has to treat checkpoint state as part of the reference specification, not merely as a preprocessing detail. The same logic applies to the operating point itself. Changing the false-positive budget changes both target and reference recovery, sometimes by very different amounts.

4.5 The audit is informative only where the reference leaves room

On CIFAR-10, tightening the probe's nominal FPR from 5% to 1% lowers recall for both sides, but much more for the matched references. Median probe gaps therefore rise from $+10.25$ to $+29.4$ points as saved and from $+13.0$ to $+33.2$ after recalibration. These are not larger estimates of the same effect: they are measurements at a different operating point (Appendix E.3).

CIFAR-100 makes the role of the reference visible in another way. SalUn gives similar raw probe recall after rabbit and willow-tree deletion, roughly 66–70%. Yet matched references recover only 51% of rabbit images and 74% of willow-tree images. The resulting gaps are $+15/+17/+17$ points for rabbit but $-4/-4/-6$ for willow tree across the three blocks. All six checkpoints have zero observed forget accuracy and at most 1.80 points of retain loss. Similar target recall therefore leads to opposite conclusions once the class-specific reference is included.

The pretrained ViT pushes this problem to its limit. At the primary 5% FPR, retain-only reference probes already achieve near-ceiling recovery, leaving only 0.2 points of median headroom. A reference with 99.8% recall permits at most $+0.2$ points of positive excess, regardless of what the target contains. At stricter 2% and 1% budgets, more headroom appears; under those settings SCRUB reduces the paired original gap in all nine outcomes, by median amounts of 2.0 and 2.6 points.

The ViT result exposes a limit of the measurement: when the reference saturates the readout, a small gap cannot distinguish successful removal from lack of headroom. CIFAR-100 shows the converse problem: high absolute recovery need not imply excess. Recovery becomes interpretable only after fixing the reference and operating point.

4.6 Discussion

The audit separates target recovery, retain-only recovery, and the operating point at which they are compared. Woken recall reads existing classifier scores; a feature probe asks a different question after fitting a new readout. Neither recovery value is meaningful without its corresponding reference.

Reference choice matters more as retained accuracy falls, and normalization can move both sides. An audit should report pre-correction forget/retain accuracy, target/reference recall, nominal/realized FPR, and the reference construction.

Scope. We evaluate one transferred configuration per method. Woken results use matched references, while systematic endpoint and trajectory controls use probes. Utility matching does not identify equal representations, and repeated calibration/probe-refitting uncertainty remains unmeasured. The extensions cover two CIFAR-100 classes and one shared-pretraining ViT setting. The measurements concern accessible class discrimination, not causal memorization or membership privacy.

5 Related Work

Approximate Machine Unlearning methods span several strategies, including selective parameter updates (Golatkar et al., 2020), teacher-based objectives (Chundawat et al., 2023a), decision-boundary manipulation (Chen et al., 2023), alternating forget–retain optimization (Kurmanji et al., 2023), weight-saliency and parameter-dampening approaches (Fan et al., 2024; Foster et al., 2024), and representation-oriented methods (Le et al., 2026). Related work has also developed zero-shot and fast unlearning procedures (Chundawat et al., 2023b; Tarun et al., 2024), while deep regression unlearning extends the problem beyond classification (Tarun et al., 2023). Our study draws from these methodological families.

Class bias manipulation changes removed-label predictions while preserving feature representations (Zheng et al., 2026). We use that operation as a recovery test: retained images calibrate one offset, and the same test on retain-only references measures excess detection. Gao et al. (2026) examine feature–classifier misalignment through newly fitted linear classifiers. Mirage (Yu et al., 2026) compares binary probe accuracy with retraining and studies geometric and layerwise behavior. Representation analysis, restoration, and relearning provide further views of recovery (Kim et al., 2026; Yong et al., 2026; Jang et al., 2026; Dehghani et al., 2026; Ha et al., 2026; Lee et al., 2026; Le et al., 2026). Our audit places recovery from existing scores alongside supervised probe recovery, with each readout calibrated and compared against its own reference. Appendix A.5 contrasts the operations.

Confusion, membership inference, and benchmarking address other aspects of removal (Goel et al., 2022; Hayes et al., 2025; Naderloui et al., 2025; Deeb & Roger, 2024; Grimes et al., 2024). Verification depends on the observable being tested (Thudi et al., 2022; Zhang et al., 2024a); relative kernel tests compare distributions statistically (Ribero et al., 2026). Our calibration fixes an image-level detection operating point. BN recalibration (Izmailov et al., 2018; Kalani et al., 2026) checks its dependence on checkpoints.

6 Conclusion

A model can stop predicting a removed class while exposing more class information than a model trained without it. Woken recall measures this difference through existing scores after retained-data calibration; feature probes provide a separate supervised view. Across the CIFAR-10 study, excess recovery remains common among low-forget, high-utility checkpoints. The comparison also shows its own limits: reference choice matters when utility falls, and saturated references can leave little room to measure excess. Unlearning claims should therefore state what is recoverable, at what operating point, and compared with which retain-only reference.

REPRODUCIBILITY STATEMENT

Appendices A through H provide proofs, protocols, classwise results, reference and calibration checks, implementation provenance, and both extensions. Post hoc analyses of the main study are recorded as dated protocol amendments and reported in Appendices D and E. The accompanying source includes model-level exports, configurations, split identifiers, and figure/table scripts. The 480-checkpoint CIFAR-10/ResNet-18 study, 36-checkpoint pretrained ViT extension, and five-method, 30-checkpoint CIFAR-100 comparison are reported separately. Appendix H identifies the extension's frozen protocols and result files; its raw and BN evaluations are paired measurements, not additional training replications.

ETHICS STATEMENT

The study evaluates class removal on public image-classification benchmarks. Recovery procedures access model logits or representations, and normalization recalibration additionally uses retained data and local checkpoint access. The experiments concern class-level behavior and do not establish privacy guarantees for individual training examples.

AI USE STATEMENT

AI tools assisted with experiment planning, implementation, mathematical checks, manuscript preparation, and figure generation. The authors take responsibility for the experimental procedures, analyses, and final claims.

REFERENCES

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. In *International Conference on Learning Representations (ICLR), Workshop Track*, 2017. URL https://openreview.net/forum?id=HJ4-rAVtl.

Lucas Bourtoule, Varun Chandrasekaran, Christopher A. Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. Machine unlearning. In *IEEE Symposium on Security and Privacy (S&P)*, pp. 141–159, 2021. doi: 10.1109/SP40001.2021.00019.

Yinzhi Cao and Junfeng Yang. Towards making systems forget with machine unlearning. In *IEEE Symposium on Security and Privacy (S&P)*, pp. 463–480. IEEE, 2015. doi: 10.1109/SP.2015.35.

Min Chen, Weizhuo Gao, Gaoyang Liu, Kai Peng, and Chen Wang. Boundary unlearning: Rapid forgetting of deep networks via shifting the decision boundary. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7766–7775, 2023.

Vikram S. Chundawat, Ayush K. Tarun, Murari Mandal, and Mohan Kankanhalli. Can bad teaching induce forgetting? Unlearning in deep networks using an incompetent teacher. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 7210–7217, 2023a. doi: 10.1609/aaai.v37i6.25879.

Vikram S. Chundawat, Ayush K. Tarun, Murari Mandal, and Mohan Kankanhalli. Zero-shot machine unlearning. *IEEE Transactions on Information Forensics and Security*, 18:2345–2354, 2023b. doi: 10.1109/TIFS.2023.3265506.

Aghyad Deeb and Fabien Roger. Do unlearning methods remove information from language model weights? *arXiv preprint arXiv:2410.08827*, 2024. URL https://arxiv.org/abs/2410.08827.

Zahra Dehghani, Pablo Piantanida, and Mohammadhadi Shateri. Source-free class relearning: Diagnosing forgetting in class unlearning. *arXiv preprint arXiv:2609.02018*, 2026. URL https://arxiv.org/abs/2609.02018.

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

Chongyu Fan, Jiancheng Liu, Yihua Zhang, Eric Wong, Dennis Wei, and Sijia Liu. SalUn: Empowering machine unlearning via gradient-based weight saliency in both image classification and generation. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/abs/2310.12508.

Jack Foster, Stefan Schoepf, and Alexandra Brintrup. Fast machine unlearning without retraining through selective synaptic dampening. In *AAAI Conference on Artificial Intelligence*, volume 38, pp. 12043–12051, 2024. doi: 10.1609/aaai.v38i11.29092.

Yichen Gao, Altay Unal, Akshay Rangamani, and Zhihui Zhu. An illusion of unlearning? assessing machine unlearning through internal representations. In *Proceedings of the 29th International Conference on Artificial Intelligence and Statistics*, volume 300 of *Proceedings of Machine Learning Research*, pp. 4114–4122, 2026. URL https://proceedings.mlr.press/v300/gao26e.html.

Shashwat Goel, Ameya Prabhu, Amartya Sanyal, Ser-Nam Lim, Philip Torr, and Ponnurangam Kumaraguru. Towards adversarial evaluations for inexact machine unlearning. *arXiv preprint arXiv:2201.06640*, 2022. URL https://arxiv.org/abs/2201.06640.

Aditya Golatkar, Alessandro Achille, and Stefano Soatto. Eternal sunshine of the spotless net: Selective forgetting in deep networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9301–9309, 2020. doi: 10.1109/CVPR42600.2020.00932.

Keltin Grimes, Collin Abidi, Cole Frank, and Shannon Gallagher. Gone but not forgotten: Improved benchmarks for machine unlearning. *arXiv preprint arXiv:2405.19211*, 2024. URL https://arxiv.org/abs/2405.19211.

SeungBum Ha, Saerom Park, and Sung Whan Yoon. Unlearning's blind spots: Over-unlearning and prototypical relearning attack. In *International Conference on Machine Learning (ICML)*, 2026. URL https://arxiv.org/abs/2506.01318.

Jamie Hayes, Ilia Shumailov, Eleni Triantafillou, Amr Khalifa, and Nicolas Papernot. Inexact unlearning needs more careful evaluations to avoid a false sense of privacy. In *2025 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pp. 497–519, 2025. doi: 10.1109/SaTML64287.2025.00034.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 448–456, 2015. URL https://proceedings.mlr.press/v37/ioffe15.html.

Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. In *Conference on Uncertainty in Artificial Intelligence*, pp. 876–885. AUAI Press, 2018. URL https://arxiv.org/abs/1803.05407.

Yurim Jang, Jaeung Lee, Dohyun Kim, Jaemin Jo, and Simon S. Woo. Suppression or deletion: A restoration-based representation-level analysis of machine unlearning. *arXiv preprint arXiv:2602.18505*, 2026. URL https://arxiv.org/abs/2602.18505.

Aaryaman Kalani, Murari Mandal, Dhruv Kumar, Mohan Kankanhalli, and Yash Sinha. The BatchNorm illusion: Diagnosing normalization artifacts in machine unlearning evaluation. *arXiv preprint arXiv:2609.08901*, 2026. URL https://arxiv.org/abs/2609.08901.

Yongwoo Kim, Sungmin Cha, and Donghyun Kim. Are we truly forgetting? a critical re-examination of machine unlearning evaluation protocols. *Engineering Applications of Artificial Intelligence*, 167:113785, 2026. doi: 10.1016/j.engappai.2026.113785. URL https://arxiv.org/abs/2503.06991.

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

Meghdad Kurmanji, Peter Triantafillou, Jamie Hayes, and Eleni Triantafillou. Towards unbounded machine unlearning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 1957–1987, 2023.

Anjie Le, Can Peng, Yuyuan Liu, and J. Alison Noble. POUR: A provably optimal method for unlearning representations via neural collapse. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026. URL https://arxiv.org/abs/2511.19339.

Jaewon Lee, Yongwoo Kim, and Donghyun Kim. Erase at the core: Representation unlearning for machine unlearning. *arXiv preprint arXiv:2602.05375*, 2026. URL https://arxiv.org/abs/2602.05375.

Nima Naderloui, Shenao Yan, Binghui Wang, Jie Fu, Wendy Hui Wang, Weiran Liu, and Yuan Hong. Rectifying privacy and efficacy measurements in machine unlearning: A new inference attack perspective. In *Proceedings of the 34th USENIX Security Symposium*, pp. 5545–5564. USENIX Association, 2025. URL https://www.usenix.org/conference/usenixsecurity25/presentation/naderloui.

Mónica Ribero, Antonin Schrab, and Arthur Gretton. Regularized f-divergence kernel tests. *arXiv preprint arXiv:2601.19755*, 2026. URL https://arxiv.org/abs/2601.19755.

Andreas Steiner, Alexander Kolesnikov, Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. How to train your ViT? Data, augmentation, and regularization in vision transformers. *Transactions on Machine Learning Research*, 2022. URL https://openreview.net/forum?id=4nPswr1KcP.

Ayush K. Tarun, Vikram S. Chundawat, Murari Mandal, and Mohan Kankanhalli. Fast yet effective machine unlearning. *IEEE Transactions on Neural Networks and Learning Systems*, 35(9):13046–13055, 2024. doi: 10.1109/TNNLS.2023.3266233.

Ayush Kumar Tarun, Vikram Singh Chundawat, Murari Mandal, and Mohan Kankanhalli. Deep regression unlearning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 33921–33939. PMLR, 2023.

Anvith Thudi, Hengrui Jia, Ilia Shumailov, and Nicolas Papernot. On the necessity of auditable algorithmic definitions for machine unlearning. In *USENIX Security Symposium*, pp. 4007–4022, 2022. URL https://www.usenix.org/conference/usenixsecurity22/presentation/thudi.

Ross Wightman. PyTorch image models. https://github.com/rwightman/pytorch-image-models, 2019.

Teresa Pui Yee Yong, Win Kent Ong, and Chee Seng Chan. Erased, but not gone: Output forgetting is not true forgetting. *arXiv preprint arXiv:2606.25001*, 2026. URL https://arxiv.org/abs/2606.25001.

Zhenyu Yu, Yangchen Zeng, Chunlei Meng, Guangzhen Yao, and Shuigeng Zhou. Do vision models truly forget? new findings from representation-level certification of visual unlearning in vertical federated learning. *arXiv preprint arXiv:2605.20282*, 2026. URL https://arxiv.org/abs/2605.20282.

Binchi Zhang, Zihan Chen, Cong Shen, and Jundong Li. Verification of machine unlearning is fragile. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 58717–58738, 2024a. URL https://proceedings.mlr.press/v235/zhang24h.html.

Ruiqi Zhang, Licong Lin, Yu Bai, and Song Mei. Negative preference optimization: From catastrophic collapse to effective unlearning. In *First Conference on Language Modeling (COLM)*, 2024b. URL https://openreview.net/forum?id=MXLBXjQkmb.

Weidong Zheng, Kongyang Chen, Yuanwei Guo, and Yatie Xiao. Classification-head bias in class-level machine unlearning: Diagnosis, mitigation, and evaluation. *arXiv preprint arXiv:2605.08730*, 2026. URL https://arxiv.org/abs/2605.08730.

A PROPERTIES OF THE RECOVERY AUDIT

NOTATION

Symbol	Meaning
D	Full training dataset, with $D = D_{\mathrm{r}} \cup D_{\mathrm{f}}$.
$D_{\mathrm{r}}, D_{\mathrm{f}}$	Retained and removed/forget training examples.
c	Removed class.
$\mathcal{A}, \mathcal{U}$	Training algorithm and unlearning procedure.
ξ, ζ	Training and unlearning randomness.
f_u	Model produced by unlearning; f denotes a generic audited classifier.
g	Retain-only reference model/checkpoint.
$X_c \sim \mu_c$	Held-out removed-class example and its population.
$(X_r, Y_r) \sim \mu_r$	Retained example/label pair and retained population.
$F(f), R(f)$	Population forget and retain accuracy.
$\widehat{R}(f)$	Empirical/test retain accuracy.
$s_f(x)$	Class margin $f_c(x) - \max_{k \neq c} f_k(x)$.
α	Nominal false-positive-rate operating point.
$\widehat{\tau}_{f,\alpha}$	Retained-calibration threshold for model f at operating point α.
$\widehat{W}_\alpha(f)$	Empirical woken recall of f at operating point α.
$\delta_{f,\alpha}$	Additive removed-logit correction, $-\widehat{\tau}_{f,\alpha}$.
$\mathcal{G}$	Declared retain-only reference set.
a	Readout route, e.g. bias/woken, reversed, or probe.
$\widehat{H}_{a,\alpha}(f)$	Empirical held-out recall from readout a at operating point α.
v	Checkpoint normalization condition (saved or BN recalibrated).
$\widehat{\mathcal{N}}_{c,v}(q)$	Utility-matched reference neighborhood for class c, condition v, and retain accuracy q.
$\widehat{G}_{a,\alpha}(f_v)$	Matched reference-relative recall gap for readout a.
$\widehat{G}_{\mathrm{woken},\alpha}(f; \mathcal{G})$	Woken-recall gap relative to a declared reference set.
$a^*(f_v)$	Readout selected by calibration recall for the auxiliary diagnostic.
$\widehat{G}_{\mathrm{sel},\alpha}(f_v)$	Reference-relative gap for the calibration-selected diagnostic.
$\mathcal{S}$	Declared finite comparator subset of a matched neighborhood.
$\widehat{\mathcal{I}}_{a,\alpha}(f_v; \mathcal{S})$	Reference-sensitivity interval over comparator set $\mathcal{S}$.
$C(f)$	ROC curve/set obtained by thresholding the class margin.
$T(f)$	Tie-aware rank AUC of the class-margin score.
h, w_k, b_k	Feature representation and weights/biases of a linear classifier head.
θ, ϕ	Learned parameters and normalization buffers.
$\mathcal{B}_{D_{\mathrm{r}}}(\theta, \phi)$	BN recalibration intervention using retained data.
$\mathcal{L}_r, \mathcal{L}_f, \mathcal{J}_\lambda$	Retain-only loss, negative forget term, and combined negative-training objective.
λ, ρ	Positive weighting/regularization constants in Appendix A.7.
$p_M(x)$	Frozen-feature probe decision score for model M.
$A(M)$	Tie-aware rank statistic comparing deleted training images with fresh same-class images.
B, P, U, Q	Outcome sets used in the cross-route counting arguments of Appendix E.2.
H_u, H_{ref}	Target/unlearned probe recall and matched reference recall in summary tables.
H_{probe}	Probe recall.
G_{probe}	Probe recall gap relative to the stated reference.
G_{end}	Gap relative to the median fully trained endpoint reference.
H_o, H_t	Original-model and target-model probe recall in ViT fixed-reference controls.
G_o, G_t	Original-model and target-model gaps in ViT fixed-reference controls.
$\Delta_\alpha(u, o)$	Fixed-reference target-minus-original recall/gap change at operating point α.
$\bar{w}_c$	Normalized removed-class classifier row used in the POUR-P projection.
L_R	Retained-accuracy loss relative to the corresponding original model in the CIFAR-100 extension.
n_s, n_+	Number of supported outcomes and number with strictly positive gaps.

A.1 PROOF OF PROPOSITION 1

A bias edit translates every margin by the same amount. For any input x,

$$s_{f^\delta}(x) = (f_c(x) + \delta) - \max_{k \neq c} f_k(x) = s_f(x) + \delta.$$

On the finite audit set $S = S_f \cup S_r$ from Proposition 1, choose

$$\delta_* = -\max\{0, \max_{x \in S} s_f(x)\} - 1.$$

Then $\delta_* < 0$ and $s_f(x) + \delta_* < 0$ everywhere on S. The removed label never maximizes the logits, so empirical forget accuracy and the rate of predicting c on retained examples both equal zero. Because retained examples have labels other than c, lowering the class-c logit cannot turn a correct retained prediction into an error; empirical retain accuracy is therefore nondecreasing. The edit leaves feature parameters, normalization buffers, and feature vectors unchanged.

For any threshold t, $\{s_{f^\delta} > t + \delta\} = \{s_f > t\}$. Sweeping the thresholds therefore recovers the same family of decision sets. Translation preserves rankings and ties, establishing invariance of the ROC and tie-aware rank AUC. If a translation-equivariant order-statistic rule chooses threshold τ on the original calibration scores, it chooses $\tau + \delta$ on the shifted scores. The resulting held-out decision set is identical. Equivalently, the additive logit correction needed to reach that decision set changes by $-\delta$.

Reference-relative comparisons. A class bias shift also leaves the feature probe's inputs unchanged. These invariances describe the audited model's readings. If the edit changes ordinary retain accuracy, reference lookup can select a different neighborhood. Fixed-reference recovery gaps therefore inherit the invariance, whereas gaps recomputed after retain accuracy matching need not. For an invariant utility statistic, one could use

$$R_{-c}(f) = \Pr_{(X_r, Y_r) \sim \mu_r} \left[\arg\max_{k \neq c} f_k(X_r) = Y_r \right],$$

which excludes the removed logit. Our reported comparisons use ordinary retain accuracy, as specified in Section 3.

A.2 Retained utility does not identify feature recovery

Proposition 2 (Retained utility does not identify feature recovery)**.** *Without structural assumptions on the representation and head, equal retained accuracy does not imply equal feature recoverability. There exist two models with identical logits on every input, and hence identical F and R, for which a linear feature score has AUC $1/2$ in one model and AUC 1 in the other. Conversely, changing only the head can change R while leaving every feature-only audit input unchanged.*

Proof. The claim is existential, so two elementary constructions suffice. Take a removed population concentrated at x_c and a retained population concentrated at a distinct point x_r. For the first construction, let $h_0(x) = 0$ for every input and let a head ψ_0 produce any fixed logit vector. Every linear score on h_0 is constant, so its tie-aware AUC is $1/2$. Now set

$$h_1(x_c) = 1, \qquad h_1(x_r) = 0,$$

and let ψ_1 ignore its input and return exactly the same fixed logits as ψ_0. The two composed classifiers have identical logits on every input and therefore identical forget and retain accuracy. The linear score $h_1(x)$, however, ranks every removed-class input above every retained input, so its AUC is 1. Thus even identical output functions, a stronger condition than equal retained accuracy, do not identify feature recoverability.

For the converse, hold a feature extractor h fixed and choose two heads that make different retained predictions. For example, one head can read a retained-label coordinate in h, while another permutes the corresponding output rows. Their retained accuracies can differ, while the inputs supplied to any feature-only audit are identical because both expose the same $h(x)$. Hence retained utility can change without a change in feature-only recoverability. These constructions establish only non-identification; they do not say that the empirical reference ladder realizes either extreme.

A.3 Decision regions of the two readouts

For a linear multiclass head $f_k(h) = w_k^\top h + b_k$, the margin detector satisfies

$$s_f(h) > \tau \quad \Longleftrightarrow \quad (w_c - w_k)^\top h + b_c - b_k > \tau \quad \text{for every } k \neq c.$$

Its positive region is an intersection of halfspaces whose normals are fixed by the trained head. A binary linear probe uses a single fitted halfspace. Varying the one margin threshold does not in general produce a family nested inside the probe family, nor does a fixed margin family contain all probe halfspaces. The restriction in woken recall is operational: the auditor keeps the existing head and chooses one scalar from retained calibration examples. The probe instead learns a direction and intercept from labeled examples of both groups. Neither readout is guaranteed to dominate the other's recall or reference-relative gap.

A.4 Interpreting rankings and recovery curves

The examples below show what score ordering can and cannot establish about discrimination and agreement with retraining. We therefore report the existing-score and fitted-feature readouts separately.

Chance AUC can conceal a nonlinear score relationship. Let $s(X_r) = 0$, and let $s(X_c)$ be -1 or $+1$ with equal probability. The oriented ranking has AUC $= 0.5$, but $|s(x)| > 1/2$ perfectly detects class membership. AUC measures the ordering induced by the score.

Reversed ordering has multiple possible sources. The scores $s(X_r) = -1$ and $s(X_c) = -2$ give AUC $= 0$. Such ordering can result from generalization, changed readout weights, or subsequent class-targeted updates. It does not identify which training examples were used.

Output agreement leaves unobserved representations unconstrained. Two networks can implement the same output map while one carries additional hidden features that its head ignores. Their margin curves agree, but a probe may distinguish their representations. Conversely, matching two summaries, such as AUC and recall at 5% FPR, leaves the rest of the curves unconstrained. Reporting each route and its full curve preserves these distinctions.

A.5 Comparison with existing diagnostics

Table 4 compares the operations measured by related diagnostics. The woken route uses a fixed class margin and retained images to set its offset. Probe routes fit a new decision function from labeled positives and negatives. Their recovery gaps answer different questions, even when both are compared with retraining. The route disagreements in Table 19 use identical target checkpoints and local reference sets; within each normalization condition, the support cohort is the same.

Table 4: **Readout and reference choices in closely related evaluations.** Entries describe the cited diagnostics; they are not an experimental ranking of audit procedures.

Study	Readout or quantity	Comparison and operating point
Gao et al. (2026)	Newly fitted linear classifier; nearest-class-center accuracy and feature/classifier alignment.	Output and feature-level accuracies, including retain-only retraining.
Mirage (Yu et al., 2026)	Binary linear probe classification accuracy with geometric and layerwise diagnostics.	Probe accuracy gap to retraining; near-ceiling example reported.
Zheng et al. (2026)	Class bias shifts and statistics of bias parameters.	Output changes under bias manipulation and bias-oriented evaluation metrics.
This audit	Woken recall from a scalar offset calibrated on retained data; separately fitted feature probe.	Matched recall gaps for both readouts at common nominal FPR; systematic endpoint sensitivity and fixed-reference control results for the probe.

A.6 Calibration-selected recovery diagnostic

The fixed-route comparisons apply each readout separately to every model and reference. An auxiliary diagnostic allows the audited model to select among the bias, reversed-score, and probe routes. Each route first receives its own threshold at the nominal FPR budget. Recall on the threshold calibration subset then chooses

$$a^*(f_v) = \arg \max_{a \in \{\text{bias,reversed,probe}\}} \widehat{H}^{\text{cal}}_{a,\alpha}(f_v). \tag{6}$$

The choice is frozen before measuring evaluation recall. For this diagnostic, the reference ladder is read on the *same route chosen by the audited model*:

$$\widehat{G}_{\text{sel},\alpha}(f_v) = \widehat{H}^{\text{eval}}_{a^*(f_v),\alpha}(f_v) - \operatorname*{median}_{g_v \in \widehat{\mathcal{N}}_{c,v}(\widehat{R}(f_v))} \widehat{H}^{\text{eval}}_{a^*(f_v),\alpha}(g_v).$$

Each reference retains its own calibrated threshold for that route. The support condition remains at least five matched checkpoints within the retain accuracy window. Calibration selects the readout for this diagnostic; its held-out recall and gap can lie above or below the corresponding measurements from the fixed probe protocol.

The route choice is target-conditioned: the target selects a route, while references are evaluated on that choice. Held-out evaluation prevents choosing the route on evaluation recall but does not make these roles symmetric. The reported study does not include a calibration check of this selector using retain-only models on both sides. Accordingly, this auxiliary result supplies no model-level false-alarm calibration and is not used to establish the fixed-route findings. The ordinary woken route uses no removed-class calibration labels; the selector does use them to compare routes.

A.7 Negative training around a retain-only solution

A model can already minimize its retain-only objective and still move under a negative training update. Such updates increase loss on removed examples while controlling retained behavior. Following related analyses of gradient ascent optimization (Zhang et al., 2024b), we examine a fixed-weight negative term at a stationary retain-only solution.

Let $\mathcal{L}_r(\theta)$ be a differentiable retain-only objective including regularization, under a fixed forward map convention, and suppose $\nabla\mathcal{L}_r(\theta_g) = 0$. Define

$$\mathcal{L}_f(\theta) = -\mathbb{E}_{x\sim D_f} \log p_\theta(c \mid x), \qquad \mathcal{J}_\lambda(\theta) = \mathcal{L}_r(\theta) - \lambda\mathcal{L}_f(\theta), \quad \lambda > 0.$$

Then

$$\nabla\mathcal{J}_\lambda(\theta_g) = -\lambda\nabla\mathcal{L}_f(\theta_g).$$

Whenever the forget gradient is nonzero, the retain-only solution is not stationary for $\mathcal{J}_\lambda$. For an unconstrained class bias and finite softmax logits,

$$\left.\frac{\partial\mathcal{J}_\lambda}{\partial b_c}\right|_{\theta_g} = \lambda\mathbb{E}_{D_f}[1 - p_{\theta_g}(c \mid x)] > 0.$$

Zero argmax accuracy does not eliminate this gradient. The calculation concerns a persistent negative loss with fixed λ; early stopping, constraints, or a vanishing coefficient change the optimization problem.

A two-point example. Let $p(c \mid x) = \sigma(b + wx)$, with retained examples at $x = 0$ assigned to the other class and removed examples at $x = 1$ assigned to c. For $\rho > 0$, let

$$\mathcal{L}_r(w, b) = \log(1 + e^b) + \frac{\rho}{2}(w^2 + b^2).$$

Its unique minimizer has $w_g = 0$ and $b_g < 0$, since $\sigma(b_g) + \rho b_g = 0$. It has perfect retain accuracy, zero forget accuracy, and tied scores on the two input types. The negative training objective is

$$\mathcal{J}_\lambda(w, b) = \mathcal{L}_r(w, b) + \lambda \log\sigma(b + w).$$

This function is continuous and coercive because the quadratic term dominates its negative tail that grows at most linearly. It therefore admits a global minimizer. At any stationary point,

$$\rho w + \lambda[1 - \sigma(b + w)] = 0,$$

which requires $w < 0$, and

$$\sigma(b) + \rho b + \lambda[1 - \sigma(b + w)] = 0,$$

which requires $b < 0$. Every global minimizer therefore preserves zero forget accuracy and perfect retain accuracy while ranking the removed input below the retained input. Its ordinary AUC equals zero and its reversed AUC equals one. The input already distinguished the two classes; negative training changes how the output score expresses that distinction.

B Implementation details and reference models

This appendix describes the CIFAR-10 study. Appendices G and H specify the extension protocols and their differences.

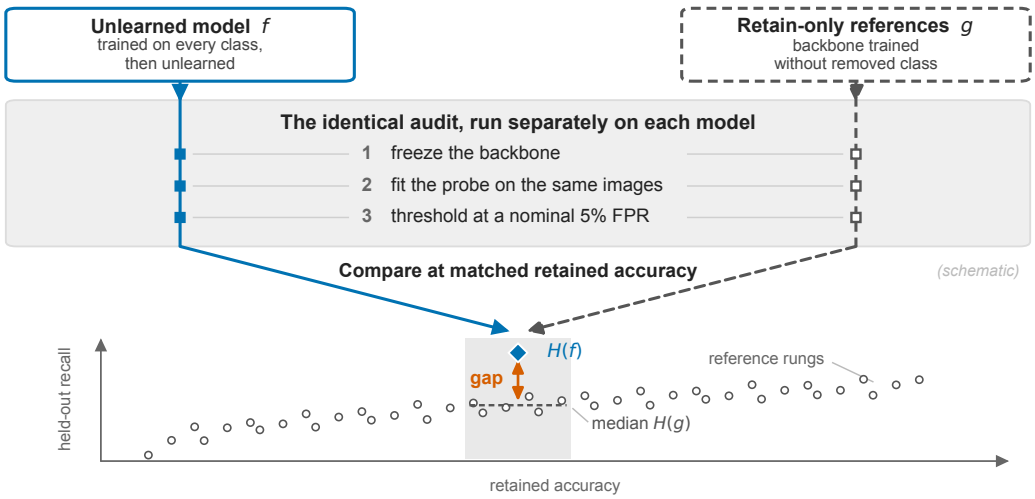

Figure 5: **Probe workflow and local reference comparison.** Each model has a separately fitted probe and calibrated threshold. The lower schematic illustrates utility matching; completed endpoints provide a separate comparison. Woken recall uses existing logits in place of fitting a probe, followed by its own threshold calibration on retained data.

B.1 Training, validation, and recovery data

The experiments use one split of CIFAR-10 throughout original model training, reference training, configuration selection, and unlearning. We hold out 500 examples per class from the 50,000-image training set using seed 12345. This leaves 45,000 training examples and 5,000 validation examples. Each class-deletion run uses 4,500 removed-class training examples and 40,500 retained-class training examples. Validation examples are excluded from training originals, references, and unlearned models.

Classification accuracy and recovery use different portions of the official 10,000-image test set. Forget accuracy uses all 1,000 removed-class test examples, and retain accuracy uses all 9,000 retained-class test examples. For recovery, a fixed stratified split with seed 12345 allocates 5,000 examples to calibration and 5,000 to evaluation. A separate stratified split with seed 12346 divides the calibration half: 70% for fitting the probe and 30% for calibrating its threshold and selecting an auxiliary recovery route. Table 5 gives the resulting counts. The recovery evaluation half is disjoint from both fitting and threshold calibration. Test-based recovery fitting is an audit of a frozen checkpoint; it never selects unlearning hyperparameters or changes that checkpoint's learned parameters.

Table 5: **Data allocation for one class deletion.** The final three rows partition the official test set for recovery. Classification accuracy uses their union.

Partition	Removed class	Retained classes	Total
Model training	4,500	40,500	45,000
Configuration validation	500	4,500	5,000
Recovery probe fitting	350	3,150	3,500
Recovery threshold calibration	150	1,350	1,500
Recovery evaluation	500	4,500	5,000

Test set usage. The audit fits and calibrates its probe on official test images. One fixed stratified split with seed 12345 divides the 10,000 test images into a 5,000-image calibration half and a 5,000-image evaluation half. The calibration half supplies 3,500 images for fitting the probe and 1,500 for setting its threshold; the evaluation half supplies the 500 removed-class positives and 4,500 retained negatives that carry recall and realized FPR. The three recovery subsets are disjoint by construction, so evaluation images do not fit the probe or its threshold. Classification accuracy, however, uses the full official test set. Primary reference matching therefore uses retain accuracy over all 9,000

retained test images, including the 4,500 recovery evaluation negatives. The held-out claim applies to recall and realized FPR, not to the utility value used for matching. The same split and fitting budget are used for every model, every reference checkpoint, every route, and both normalization conditions. Unlearning configurations were selected on the cat validation split in block 0 before any test measurement (Section 3). Recovery data do not select an unlearning configuration or update a checkpoint.

B.2 Original models and retain-only references

We use the CIFAR ResNet-18 architecture with a 3×3 stem and no initial max pooling. Three originals are trained independently with seeds 0, 1, and 2; each original defines one experimental block. Original and reference training use SGD with learning rate 0.1, momentum 0.9, weight decay 5×10^{-4}, Nesterov momentum, batch size 128, and cosine annealing over 200 epochs. Training augmentation uses random crops with padding 4 and horizontal flips, followed by per-channel normalization.

For each of ten removed classes and each of three seeds, a reference model is trained from scratch on the corresponding 40,500 retained training examples. We save epochs 1 through 9 and every tenth epoch through 200, giving 29 snapshots per continuous training run, 30 reference runs, and 870 reference checkpoints. The 30 epoch-200 checkpoints are the fully trained retain-only endpoints. Intermediate checkpoints provide candidate matched retain-only comparisons at lower retain accuracy and share their parent training trajectory.

B.3 Configuration selection and transfer

The frozen selection pool contains 127 candidates across fifteen methods, including each added method's default anchor. Fifteen methods receive a selected configuration; POUR-P is a closed-form method with no tuning. All candidates are evaluated on the cat validation split in block 0. The chosen configuration is then transferred unchanged to all ten deletion classes and all three blocks. Appendix F records the implementation sources and adaptations.

Selection uses only validation forget and retain accuracy. If any candidate satisfies $F_{\text{val}} \leq 1\%$ and $R_{\text{val}} \geq 85\%$, we select the feasible candidate with highest retain accuracy (`OK`). Otherwise, if some candidate satisfies the forget criterion, we select the one with highest retain accuracy among those candidates (`TRADEOFF-ONLY`). If none satisfies the forget criterion, we minimize forget accuracy, then maximize retain accuracy (`NOT-REACHED`). Remaining ties break by candidate identifier. The final swept configurations comprise ten `OK`, four `TRADEOFF-ONLY`, and one `NOT-REACHED` selection. These statuses summarize cat/block-0 validation outcomes. The $F \leq 1\%, R \geq 90\%$ region in the figures describes test outcomes after the selected configurations have been evaluated.

B.4 Recovery scores and calibration

The primary 5% budget fixes the threshold tradeoff for this experiment. It was not chosen after evaluation and does not prescribe a deployment policy. Realized FPR is the frequency of flagged retained images, not the probability of falsely declaring that unlearning failed. Precision also depends on the prevalence of removed-class images where the detector is used.

The implementation defines the margin as

$$s_f(x) = f_c(x) - \max_{k \neq c} f_k(x),$$

so larger scores indicate stronger removed-class evidence. The ordinary head route thresholds s_f; the reversed route thresholds $-s_f$. The feature route fits a linear binary probe on frozen penultimate features using the probe fitting subset. Thresholds target a 5% false-positive rate on calibration negatives. We apply the same score construction and calibration budget to unlearned models and reference checkpoints and evaluate recall and realized FPR on the held-out recovery half. AUC uses the rank-based Mann-Whitney statistic, assigning average ranks to tied scores. The displayed ROC curves interpolate empirical curves onto 200 FPR grid points.

Probe fitting. We extract flattened features from global average pooling in the final ResNet block and fit a separate scaler and probe for each model and normalization condition. `StandardScaler`

estimates feature means and scales only from the 3,500-example probe fitting subset. The code explicitly calls `LogisticRegression` with $C = 1$ and `max_iter=2000`, then saves its decision function scores as float32. The remaining settings use library defaults: mean centering and variance scaling, L-BFGS optimization with an ℓ_2 penalty, a fitted intercept, tolerance 10^{-4}, uniform sample weights, and an unset random state. The fitting data and fitted scaler are held fixed when scoring the threshold calibration and evaluation subsets. The backbone remains frozen.

Threshold convention. Each route sets its threshold to the 95th percentile of its 1,350 retained-class scores in the threshold calibration subset. The implementation calls `np.percentile` without overriding its linear interpolation default. It stores the negative threshold as an offset δ and predicts the removed class exactly when $s(x) + \delta > 0$; equality is negative. This rule sets the nominal 5% FPR budget; realized evaluation FPRs quantify its transfer to held-out images.

The experimental designation of the probe and the timing of the cross-route analysis are stated in Section 3. For the auxiliary calibration-selected route, recall on the threshold calibration subset selects among the ordinary head, reversed head, and probe. Exact calibration recall ties favor the ordinary head, then the reversed head, then the probe. The selected route is then evaluated once. Its reference is measured on the *same route selected for the unlearned model*, giving a recovery comparison for that calibration-selected readout. A post hoc maximum over evaluation recalls supplies a descriptive envelope.

B.5 Reference pooling and support

For each removed class and normalization condition, we pool all 29 epochs from all three reference seeds, giving up to 87 checkpoints. This pool supplies the references for all three unlearning blocks of that class.

For a model with retain accuracy q, the supported neighborhood contains every checkpoint g satisfying $|\widehat{R}(g) - q| \leq 2$ percentage points, including the boundary. We report a gap only when at least five matched retain-only checkpoints lie in that neighborhood. The reference is the median recovery recall of those checkpoints, and the minimum and maximum describe their spread. Models with fewer than five local checkpoints receive an unsupported label. For these models, the exports retain the five nearest checkpoints as contextual reference values.

We compute each probe gap by subtracting that row's matched reference recall from its probe recall. Method summaries take the median of these *differences per row* over supported rows. When support is partial, supported-subset retain accuracy describes the same rows as the gap; the all-row retain median describes the full study. Reference minimum-to-maximum ranges describe checkpoint variation, and checkpoint counts measure local support.

B.6 BatchNorm recalibration

The saved condition uses each checkpoint as produced by its method. The recalibrated condition resets running statistics in BN1d and BN2d modules, sets `momentum=None`, and places only those modules in training mode. All other modules remain in evaluation mode. We perform one forward pass over all 40,500 retained training examples with gradients disabled, then return the entire model to evaluation mode. The same intervention is applied to reference checkpoints, and feature probes and recovery measurements correspond to that normalization condition.

The loader uses batch size 512, no shuffling, no augmentation, and `drop_last=False`. Images undergo tensor conversion and the base model's channel normalization. Retained examples appear in ascending original training set index order; the seed-12345 validation carve-out fixes set membership. Thus the pass contains 79 full batches and a final batch of 52 examples. No removed-class or validation example enters this pass. The loader has a configurable worker count and defaults to three workers.

PyTorch cumulative averaging gives all 80 batches equal weight, including the final 52-example batch alongside 79 batches of 512. The resulting values define a specified normalization buffer intervention, not exact dataset moments. The frozen Gen-2 exports were generated at revision `6779173`; their contemporaneous guard was a signed scalar sum over learned parameters, which can miss cancelling changes. The elementwise named parameter comparison is a retrospective strengthening and was

not recorded for every frozen export. On one manifest-verified FineTune checkpoint, it found all 62 parameter tensors identical, all 60 expected BN buffers changed, and no unexpected non-BN buffer change. We determine reference support again from retain accuracy in each condition. Sensitivity to the ordering and construction of the 512-example batches has not been measured, so the BN results should be interpreted as this declared intervention.

B.7 Inventory and reproducibility

Table 6: **Experimental inventory.** Saved and BN-recalibrated conditions are paired evaluations of the same unlearning checkpoints. Fully trained retain-only endpoints are separate from the unlearning study.

Item	Count
Original training runs / blocks	3
Methods / fixed method configurations	16
Deleted classes per block	10
Unlearning checkpoints	480
Saved and BN unlearning evaluation rows	960
Retain-only reference training runs	30
Reference checkpoints	870
Fully trained reference checkpoints (epoch 200)	30
Supported unlearning comparisons, saved / BN	372 / 382
Unsupported unlearning comparisons, saved / BN	108 / 98

The accompanying exports record the frozen revision, split and checkpoint hashes, selected configurations, upstream commits, and measurements per model. The source package includes the inputs and scripts for reproducing summaries, prevalence counts, and figures. Analyses specified after the study was frozen are recorded as dated amendments to that protocol, and are reported in Appendix E.

B.8 Recorded audit settings and reporting fields

Table 7 collects the CIFAR-10 audit settings specified above. The same settings apply to targets and references within a route. The reported code uses explicit calls and the listed library defaults; package version numbers for the frozen scoring environment are not specified in this manuscript. A version-locked environment and a single verified reproduction command are not supplied here. The method manifests and extension protocols remain necessary for inherited optimizer settings and implementation-specific choices.

A recovery record should identify the checkpoint, removed class, original block, normalization condition, and readout, together with the fitting and calibration budgets and split identifiers. Report forget/retain accuracy before the intervention, target and reference recall, nominal and realized FPR, the paired gap, and whether the reference is an endpoint or a matched neighborhood. For a neighborhood, record the qualifying checkpoint identities, distinct seeds, epoch range, and support count. For control comparisons, retain the target's reference identities and report original recall, target recall, both gaps, and reference headroom. This reporting specification distinguishes observed recovery, reference sensitivity, and audit precision without merging the routes into a single success criterion.

B.9 Audit cost

The completed ViT training logs record cumulative wall-clock times on an NVIDIA RTX A6000. A ten-epoch all-class run took a median 479.5 seconds over three runs (range 478.9–484.6 seconds). A ten-epoch class-excluded reference run took a median 444.0 seconds over the nine logged runs (range 437.2–445.8 seconds). Each reference run produced twenty saved rungs during one continuous training trajectory, so those rungs are not twenty separate training runs. Reference training is reusable across targets with the same class and block; unlearning and its target scoring are work per target.

The frozen ResNet and ViT scoring logs do not record separate wall-clock durations for feature extraction, probe fitting and threshold calibration, or BN recalibration. Feature extraction takes a

Table 7: **Recorded CIFAR-10 audit configuration.** These settings describe the frozen measurements. Extension differences are documented in Appendices G and H.

Component	Setting
Recovery partition	Seeds 12345 and 12346; 3,500 fitting, 1,500 threshold calibration, and 5,000 evaluation images.
Woken score	$f_c(x) - \max_{k \neq c} f_k(x)$; no fitted feature direction.
Probe	Fit-only standardization; logistic regression with $C = 1$, L-BFGS, ℓ_2 penalty, intercept, tolerance 10^{-4}, at most 2,000 iterations, uniform sample weights, and unset random state.
Stored probe scores	Float32 decision function scores.
Primary threshold	95th percentile of 1,350 retained calibration scores, linear interpolation; strict $>$ rule, ties negative.
Endpoint reference	Median recall of three epoch-200 retain-only references per class and normalization condition. The reported study-wide median- and maximum-endpoint sensitivity results use probe recall.
Conditional reference	All rungs within inclusive ± 2 retain accuracy points; at least five rungs, equally weighted. Epochs 1 through 9 and every tenth epoch through 200, across three seeds.
BN intervention	Reset BN1d/BN2d running statistics, `momentum=None`; one pass over retained data without gradients, batch size 512, ascending indices, no shuffle or augmentation, final partial batch retained.

forward pass per checkpoint and can be cached. Probe fitting and calibration are repeated per model, route, and normalization condition. BN recalibration takes one forward pass on retained data per ResNet checkpoint and does not apply to the LayerNorm ViT. Because of hardware contention, the available ViT training times are representative measurements rather than portable throughput estimates.

C Coverage across methods, classes, and blocks

The study evaluates the same sixteen fixed configurations on ten removed classes and three independently trained original models. A supported outcome has at least five matched retain-only checkpoints within two percentage points of its retain accuracy. A positive outcome has a strictly positive probe gap for that model. Table 9 reports both quantities within each block; Figure 6 shows class-level variation.

C.1 Accuracy within supported comparisons

Table 10 gives forget and retain accuracy for the outcomes contributing to each method's gap. For BoundaryShrink, only three saved-condition outcomes have reference support. Their median forget and retain accuracies are 97.8% and 90.2%, compared with 0% and 29.8% across all its outcomes. All three remove bird, one in each block. Its positive gap therefore describes a supported subset with high forget accuracy and coverage of one class.

C.2 The low-forget, high-utility slice

Table 11 reports the conjunction of accuracy criteria and positive probe gaps. These criteria describe already evaluated checkpoints. Eligibility is determined separately from the accuracy measured in each normalization condition.

The low-forget, high-utility slice contains 88/89, 89/89, and 86/86 positive saved-condition gaps in blocks 0, 1, and 2; after recalibration the counts are 62/62, 65/65, and 62/62. Block 0 supplied configuration selection. The counts are similar across blocks, and we treat that comparison as descriptive. The single nonpositive gap in the saved $F \leq 1\%, R \geq 90\%$ slice is SCRUB on automobile in block 0: $F = 0\%$, $R = 91.86\%$, probe recall 93.6%, reference recall 95.2%, and gap -1.6 points. FineTune, SalUn, and POUR-P each satisfy exact zero forget accuracy, retain accuracy above 90%, and positive probe gaps in all thirty outcomes in both conditions. The accompanying CSV provides all 960 model-level measurements and their reference support.

Table 8: **Output accuracy alongside excess recovery.** Each method has 30 outcomes per condition. F_{all} and R_{all} summarize all outcomes; F_s, R_s, and G_{probe} summarize the same n_s supported outcomes. n_+/n_s counts positive gaps. Accuracies are percentages; gaps are percentage points. N/A marks no supported outcome; blank aggregate cells omit accuracy summaries.

	As saved				BN recalibrated			
Method	F_{all}/F_s	R_{all}/R_s	G_{probe}	n_+/n_s	F_{all}/F_s	R_{all}/R_s	G_{probe}	n_+/n_s
BadTeacher	0.0/0.0	94.4/94.4	+12.30	30/30	0.0/0.0	90.3/90.3	+17.70	30/30
BoundaryExpand	20.1/N/A	28.9/N/A	N/A	0/0	40.4/N/A	24.3/N/A	N/A	0/0
BoundaryShrink	0.0/97.8	29.8/90.2	+29.80	3/3	1.4/N/A	50.3/N/A	N/A	0/0
FineTune	0.0/0.0	92.6/92.6	+11.00	30/30	0.0/0.0	91.9/91.9	+14.45	30/30
FisherForgetting	0.0/N/A	11.1/N/A	N/A	0/0	0.0/N/A	11.1/N/A	N/A	0/0
GradientAscent	0.0/0.0	76.0/77.0	+9.00	13/20	0.0/0.0	83.9/84.4	+8.60	22/25
InfluenceUnlearning	7.7/7.7	93.6/93.6	+9.80	30/30	93.8/93.8	93.5/93.5	+16.75	30/30
L1Sparse	0.0/0.0	91.8/91.8	+10.95	30/30	0.0/0.0	91.0/91.0	+12.95	30/30
NegGradPlus	0.9/0.9	91.0/91.0	+11.45	30/30	2.2/2.2	91.0/91.0	+15.15	30/30
POUR-D	1.0/1.0	95.1/95.1	+11.70	30/30	95.2/95.2	93.5/93.5	+16.95	30/30
POUR-P	0.0/0.0	95.3/95.3	+12.00	30/30	0.0/0.0	94.0/94.0	+16.95	30/30
SalUn	0.0/0.0	94.2/94.2	+11.15	30/30	0.0/0.0	93.0/93.0	+15.00	30/30
SCRUB	0.0/0.0	92.0/92.2	+7.20	24/28	0.0/0.0	90.2/90.3	+7.80	27/28
Spotter	0.0/0.0	95.1/95.1	+10.50	30/30	93.2/93.2	93.5/93.5	+16.85	30/30
SSD	0.0/0.0	95.2/95.2	+11.80	30/30	0.0/0.0	93.9/93.9	+16.85	30/30
UNSIR	0.0/0.0	75.1/76.1	-4.50	7/21	0.0/0.0	78.7/78.8	-1.00	9/29
all supported outcomes			+10.25	347/372			+13.00	358/382
retain ≥ 90			+10.70	314/315			+15.50	302/302
and forget ≤ 1			+10.70	263/264			+14.00	189/189

Table 9: **Positive gaps and reference support by original model block.** Each cell gives positive/supported outcomes out of ten removed classes. A $0/0$ cell means no supported comparison. The same class-specific reference pool serves all three blocks. BoundaryShrink's three $1/1$ cells are the same removed class, bird, so its row records one class seen in three blocks rather than three classes.

	As saved			BN recalibrated		
Method	Block 0	Block 1	Block 2	Block 0	Block 1	Block 2
BadTeacher	10/10	10/10	10/10	10/10	10/10	10/10
BoundaryExpand	0/0	0/0	0/0	0/0	0/0	0/0
BoundaryShrink	1/1	1/1	1/1	0/0	0/0	0/0
FineTune	10/10	10/10	10/10	10/10	10/10	10/10
FisherForgetting	0/0	0/0	0/0	0/0	0/0	0/0
GradientAscent	4/6	5/8	4/6	6/7	7/8	9/10
InfluenceUnlearning	10/10	10/10	10/10	10/10	10/10	10/10
L1Sparse	10/10	10/10	10/10	10/10	10/10	10/10
NegGradPlus	10/10	10/10	10/10	10/10	10/10	10/10
POUR-D	10/10	10/10	10/10	10/10	10/10	10/10
POUR-P	10/10	10/10	10/10	10/10	10/10	10/10
SalUn	10/10	10/10	10/10	10/10	10/10	10/10
SCRUB	7/9	8/10	9/9	8/9	10/10	9/9
Spotter	10/10	10/10	10/10	10/10	10/10	10/10
SSD	10/10	10/10	10/10	10/10	10/10	10/10
UNSIR	1/6	3/7	3/8	4/10	2/10	3/9

D CALIBRATION AND REFERENCE MEASUREMENTS

D.1 REALIZED FALSE-POSITIVE RATES

Threshold calibration targets a nominal 5% FPR using calibration negatives. Realized FPR measures transfer of this calibration to separate evaluation images. Table 12 reports the median and range for each route and evaluation condition. The ranges describe variation across models.

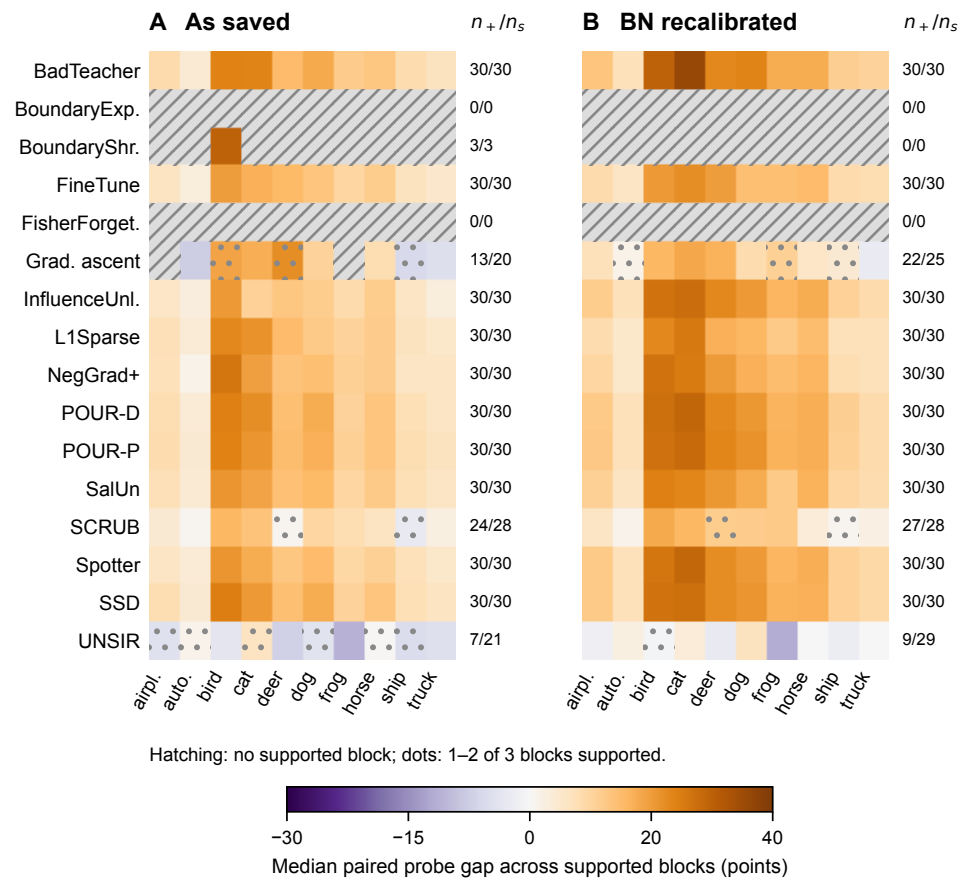

Figure 6: **Class-level excess feature recovery.** Each cell summarizes the supported probe gaps per model for one method and removed class across the three original model blocks. The color scale represents the median gap in percentage points; reference coverage distinguishes fully supported, partially supported, and missing cells. Each median aggregates individual paired gaps against matched retain-only checkpoints.

The probe has median evaluation FPR 4.978% on unlearned models as saved and 4.944% after recalibration. The fully trained references have medians 4.800% and 4.478%. Thus recall comparisons use a common calibration target, with the realized rates shown alongside the results. In the ROC figures, markers show each calibrated operating point at its measured FPR and recall.

D.2 Fully trained retain-only endpoints

These epoch-200 models are the endpoints displayed in Figure 8. Endpoint gaps use completed references, while matched ladder gaps use local neighborhoods that can include earlier checkpoints; the main text reports the constructions separately. All sixty saved/BN control rows have exact zero forget accuracy. Both the classifier margin and the feature probe recover omitted-class images from these references.

The ladder contains 870 checkpoints in each normalization condition. Reference matching uses full-precision retain accuracy to determine membership in the inclusive two-point window. Summaries aggregate the per-model gaps and their support counts.

Table 10: **Accuracy of the reference-supported subset.** All medians in a method/condition column use exactly the n_s outcomes supporting the probe gap. The coverage denominator is thirty. Missing cells have no supported outcomes.

	As saved			BN recalibrated		
Method	F_s	R_s	$n_s/30$	F_s	R_s	$n_s/30$
BadTeacher	0.0	94.4	30/30	0.0	90.3	30/30
BoundaryExpand	N/A	N/A	0/30	N/A	N/A	0/30
BoundaryShrink	97.8	90.2	3/30	N/A	N/A	0/30
FineTune	0.0	92.6	30/30	0.0	91.9	30/30
FisherForgetting	N/A	N/A	0/30	N/A	N/A	0/30
GradientAscent	0.0	77.0	20/30	0.0	84.4	25/30
InfluenceUnlearning	7.7	93.6	30/30	93.8	93.5	30/30
L1Sparse	0.0	91.8	30/30	0.0	91.0	30/30
NegGradPlus	0.9	91.0	30/30	2.2	91.0	30/30
POUR-D	1.0	95.1	30/30	95.2	93.5	30/30
POUR-P	0.0	95.3	30/30	0.0	94.0	30/30
SalUn	0.0	94.2	30/30	0.0	93.0	30/30
SCRUB	0.0	92.2	28/30	0.0	90.3	28/30
Spotter	0.0	95.1	30/30	93.2	93.5	30/30
SSD	0.0	95.2	30/30	0.0	93.9	30/30
UNSIR	0.0	76.1	21/30	0.0	78.8	29/30

Table 11: **Positive probe gaps among low-forget, high-utility outcomes.** The numerator counts positive supported gaps and the denominator counts all outcomes meeting the stated accuracy conditions; all outcomes in these slices have reference support.

	$F \leq 1\%,\ R \geq 90\%$		$F = 0,\ R \geq 90\%$	
	As saved	BN	As saved	BN
Positive / eligible	263/264	189/189	219/220	182/182
Methods represented	11	8	10	8

D.3 Uncertainty at the unit of replication

The study is a crossed design, so its rows are dependent. Within one (method, removed class) cell, the three blocks share the removed class, data partition, and pooled reference ladder. Among the 118 cells with a supported saved-condition gap in all three blocks, gap correlations are $r = 0.923$, 0.966, and 0.930 for the three block pairs. These correlations quantify dependence within the study; we draw no conclusion from their size about generalization across original models.

As a design choice, we do not resample blocks. The three blocks within a cell differ in their parent original model, which is replicated three times across the study, while sharing the other factors above. Treating these blocks as design-level replications would overstate the replication structure. Accordingly, we report no block-clustered interval.

We take the (method, removed class) cell as the resampling unit. With 20,000 draws and seed 20260910, the pooled saved-condition median of $+10.25$ points has a 95% interval of $[8.5, 11.6]$ over 129 cells, and the recalibrated median of $+13.00$ has $[11.3, 16.0]$ over 130 cells. Resampling the ten removed classes instead gives $[6.4, 14.8]$ and $[9.0, 19.8]$. For the proportion of supported rows with a positive gap, cell clustering gives design effects of 2.27 and 2.49, corresponding to effective sample sizes near 164 and 153 rather than 372 and 382. At the cell level, 118 of 129 saved-condition cells and 122 of 130 recalibrated cells have a positive median gap.

The two resampling schemes make different conditional assumptions. Cells within a removed class share that class, the data partition, the three original models, and the pooled reference ladder; cells within a method share one frozen configuration. The cell bootstrap treats cells as exchangeable under those shared factors. The class bootstrap treats the ten classes as exchangeable while still conditioning on the three original models and single partition. Aggregating rows to cells removes within-cell dependence but leaves dependence between cells.

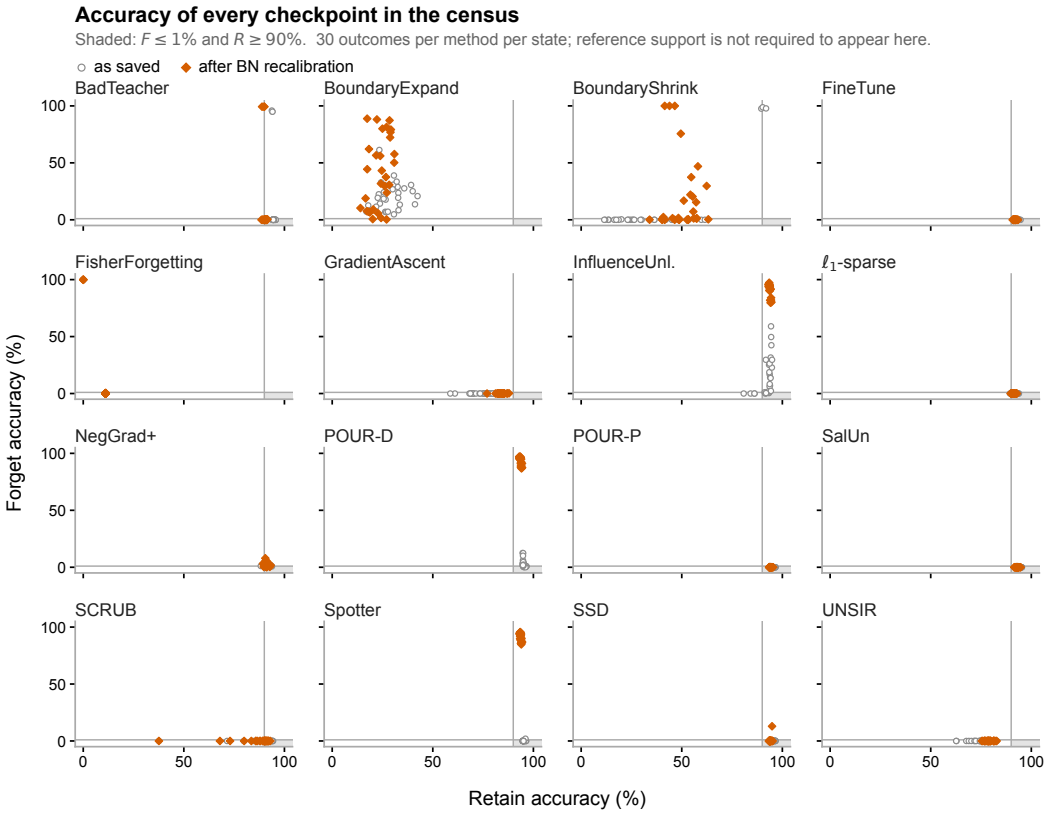

Figure 7: **Forget and retain accuracy of every checkpoint in the study, by method.** Each panel holds that method's 30 class/block outcomes in each normalization condition. Reference support is not required, so methods with no supported gap in Figure 3 are still plotted. The shaded strip is the descriptive region $F \leq 1\%$ and $R \geq 90\%$.

Table 12: **Realized evaluation FPR at a nominal 5% calibration target.** All values are percentages. The unlearned group contains 480 checkpoints, the full reference ladder contains 870 checkpoints, and the fully trained endpoint group contains thirty epoch-200 checkpoints per normalization condition.

Group	Route	As-saved median [min, max]	BN median [min, max]
Unlearned	Probe	4.978 [3.444, 6.800]	4.944 [3.511, 7.111]
Unlearned	Bias sweep	5.089 [3.400, 6.800]	5.111 [3.289, 6.689]
Unlearned	Reversed	5.000 [3.533, 7.200]	5.133 [2.933, 7.356]
Reference ladder	Probe	5.000 [3.333, 7.356]	4.889 [3.289, 6.956]
Reference ladder	Bias sweep	5.156 [3.533, 7.511]	5.356 [3.578, 7.200]
Reference ladder	Reversed	4.867 [3.289, 7.222]	5.000 [3.467, 7.800]
Epoch-200 references	Probe	4.800 [3.667, 6.044]	4.478 [3.800, 5.667]
Epoch-200 references	Bias sweep	5.022 [3.956, 6.044]	5.133 [3.844, 6.533]
Epoch-200 references	Reversed	4.800 [3.778, 5.756]	4.933 [3.822, 6.222]

These bootstrap intervals describe sampling variability over method/class cells within this CIFAR-10 study, with one architecture and one data partition. They are the paper's only interval estimates. We report no hypothesis test or p-value, and the intervals do not resolve individual method signs, especially for GradientAscent and BoundaryShrink (Section 4.3). This post hoc analysis resamples frozen gaps per row and computes no new model quantity.

Table 13: **Recovery from thirty fully trained retain-only endpoints.** Each value is a median across ten removed classes and three training seeds. Accuracies and recalls are percentages; recovery uses a nominal 5% FPR.

Measurement	As saved	BN recalibrated
Forget accuracy	0.00	0.00
Retain accuracy	95.19	93.70
Bias sweep recall	36.40	26.90
Frozen-feature probe recall	88.40	82.00
Probe realized FPR	4.800	4.478

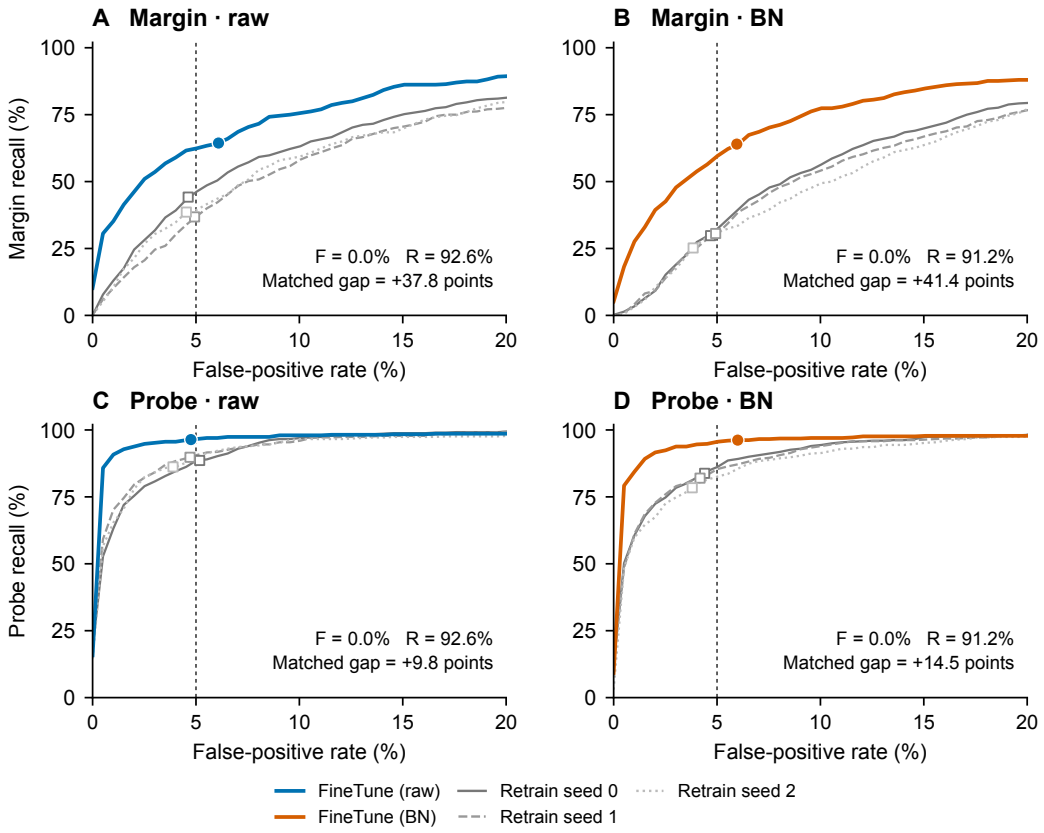

Figure 8: **Woken and probe recovery with their own references.** Curves show FineTune, frog removal, block 0, before and after BN recalibration, alongside three fully trained epoch-200 references per condition. Frog was not used for configuration selection. The annotations labeled "Matched gap" are matched ladder gaps from Equation 5, not differences from a single plotted endpoint. Woken gaps are $+37.8$ points as saved and $+41.4$ after recalibration; probe gaps are $+9.8$ and $+14.5$. Markers show realized FPR and recall at calibrated thresholds; the vertical line denotes nominal 5% FPR.

D.4 Precision of individual recovery comparisons

All reported comparisons use fixed audit partitions. The intervals in Appendix D.3 resample measured probe gaps; they do not propagate changes in fitting images, calibrated thresholds, or evaluation partitions. Recall resolution is 0.2 points with 500 positives in CIFAR-10 and ViT, and two points with 50 positives in CIFAR-100. Resolution alone does not quantify the uncertainty of a paired gap: that also depends on which images the target and reference detect and on variation in calibration and fitting.

Repeated calibration and probe refitting sensitivity have not been measured. A paired analysis would assign identical alternative image partitions to each target, original, and reference, then recalibrate the fixed margins and separately refit the probes. Keeping reference identities fixed would isolate readout variation; changing the matching rule would address comparator selection separately. Repeated partitions share images, so their spread would measure partition sensitivity unless a justified resampling design supplied statistical coverage. These are possible extensions of the current fixed-partition measurements.

D.5 SENSITIVITY TO THE REFERENCE NEIGHBORHOOD

For a declared finite comparator set $\mathcal{S} \subseteq \widehat{\mathcal{N}}_{c,v}(\widehat{R}(f_v))$, reference sensitivity can be written as

$$\widehat{\mathcal{I}}_{a,\alpha}(f_v; \mathcal{S}) = \left[\widehat{H}_{a,\alpha}(f_v) - \max_{g \in \mathcal{S}} \widehat{H}_{a,\alpha}(g),\ \widehat{H}_{a,\alpha}(f_v) - \min_{g \in \mathcal{S}} \widehat{H}_{a,\alpha}(g) \right]. \tag{7}$$

Its lower endpoint is positive exactly when the target exceeds every comparator in $\mathcal{S}$. The interval describes variation across these measured references; it is not a statistical uncertainty interval.

The empirical checks in this subsection use probe recall. Each supported row's reference is the median probe recall of matched retain-only checkpoints inside its ± 2-point retain accuracy window. Such a neighborhood can compare undertraining with update-induced damage. To measure sensitivity within the observed neighborhood, we recompute each gap against its highest-recovery checkpoint, which produces the smallest gap available from that neighborhood. The median spread between the lowest- and highest-recovery checkpoints is 7.0 points; the widest spread is 25.4 points in the saved condition and 22.2 after recalibration.

Of the 705 rows with a positive gap, 678 remain positive under that substitution: 335 of 347 in the saved condition (96.5%) and 343 of 358 after recalibration (95.8%). The 27 rows that fall to zero or below (3.8%) are concentrated in UNSIR, with 14, and SCRUB, with 11. Among the ten methods positive in all 30 outcomes, one row flips. This check uses only checkpoints from the same pooled retain-only trajectories and was specified after the study was frozen.

We also vary the matching bandwidth while leaving every measured checkpoint and probe score fixed. Table 14 shows that the aggregate result changes little from one to three retain accuracy points. Among rows supported at both the two-point setting and an alternative, no saved-condition sign changes occur. The BN condition has no sign changes at one point and four at three points. In the low-forget, high-retain cohort, no sign changes occur under either alternative in either condition. Widening to three points adds 12 supported saved rows and 6 supported BN rows, so coverage changes should be distinguished from changes on a common cohort.

Table 14: **Probe gap sensitivity to the retain accuracy matching bandwidth.** Each cell reports positive gaps over supported outcomes and the median gap in percentage points. The two-point row is the primary specification.

	Saved		BN recalibrated	
Bandwidth	n_+/n	Median	n_+/n	Median
± 1 point	325/342	$+10.5$	334/344	$+14.6$
± 2 points	347/372	$+10.25$	358/382	$+13.0$
± 3 points	355/384	$+10.1$	363/388	$+12.9$

Paired decomposition of the BN change. To separate movement of the audited model from movement of its reference, we pair rows by method, removed class, and block, then retain only rows supported in both conditions. Table 15 writes the gap change as a target-recall change minus a matched reference change. Across 365 common-support pairs, recalibration lowers the target median by 0.4 points and the reference median by 4.0 points, while the paired gap rises by 3.5 points. The reference therefore moves more than the audited model in the typical pair. Because these are separate marginal medians, their displayed values need not subtract exactly to the median paired gap change.

The operating cohort itself changes: 264 saved-condition rows and 189 recalibrated rows qualify, with 188 in both, one entering after recalibration, and 76 leaving. The separately reported cohort medians of $+10.7$ and $+14.0$ points therefore are not a paired effect.

Table 15: **Paired decomposition of the BN probe gap change.** Shifts are values after BN recalibration minus saved values in percentage points. The last column counts positive, zero, and negative paired gap shifts.

Cohort	n	Target shift	Reference shift	Gap shift	$+/0/-$
Common reference support	365	-0.4	-4.0	$+3.5$	334 / 3 / 28
$F \leq 1\%, R \geq 90\%$ in both	188	-0.8	-4.1	$+3.15$	179 / 2 / 7

D.6 Reference construction and held-out trajectory controls

The probe reference checks pool matched retain-only checkpoints across three training seeds. The equal-seed check first takes the median recovery among matched rungs within each reference seed, then takes the median of those three seed medians. It requires all three seeds and at least five pooled rungs. Restricting to the common cohort supported by both procedures leaves 370 saved-condition outcomes. Their pooled median gap is $+10.35$ points and their equal-seed median is $+10.25$; the median absolute change per row is 0.30 points, and one row changes sign. After BN recalibration, the common cohort has 376 outcomes, the medians change from $+13.15$ to $+13.05$, the median absolute change is 0.20, and three rows change sign.

Figure 4 shows why the local comparison matters. At retain accuracy of at least 90%, all 315 saved and 302 recalibrated outcomes have reference support. Their neighborhoods contain medians of 15 and 19 rungs, all three seeds are represented, and the median within-neighborhood recovery ranges are 6.8 and 7.0 points. The fully trained endpoint differs from the local reference by only 0.8 and 0.5 median points in this band. Below 85% retain accuracy, however, support falls to 45/153 and 47/145 outcomes, and the difference between endpoint and matched references grows to 13.4 and 9.6 points. We leave unsupported rows unsupported rather than substituting the endpoint.

The bandwidth and weighting checks retain neighborhoods selected from test retain accuracy. A validation-based matching check has not been performed, and per-neighborhood seed counts and epoch ranges are not summarized for the low-utility band. Five nearby rungs can come from correlated parts of a trajectory. A separate validation-matching analysis would select neighborhoods using retained validation accuracy, preserve frozen unlearning configurations and recovery scores, and report coverage changes separately from changes in gaps on common-support targets. Endpoint comparisons require no such utility-based reference selection.

Table 16: **Probe reference checks in the low-forget, high-retain cohort.** The cohort has $F \leq 1\%$ and $R \geq 90\%$. Each cell reports positive gaps over eligible outcomes and the median excess. Recall gaps are percentage points; the AUC gap is dimensionless. Matched-checkpoint rows condition on retain accuracy. The fully trained rows use endpoint models without retain accuracy matching.

	Saved		BN recalibrated	
Reference or measure	n_+/n	Median	n_+/n	Median
Matched checkpoints	263/264	$+10.7$	189/189	$+14.0$
Equal-seed checkpoints	263/264	$+10.7$	189/189	$+14.3$
Fully trained median	263/264	$+9.0$	189/189	$+13.0$
Fully trained max recovery	260/264	$+7.8$	187/189	$+12.2$
Matched excess AUC	263/264	$+0.0175$	189/189	$+0.0208$

The fully trained endpoint comparisons in Table 3 are intentionally unmatched: they compare with the epoch-200 retain-only models without conditioning on retain accuracy. A leave-one-trajectory-out check treats each fully trained model as an audit target and uses the other trajectories as references. All 30 endpoints are supported in each condition. The saved-condition gaps have median $+0.55$ points and range from -3.4 to $+5.0$; the recalibrated gaps have median $+0.45$ and range from -4.8 to $+5.2$.

For a second scale check, each checkpoint on one retain-only trajectory is treated as the target and matched against checkpoints from the other trajectories. Table 17 summarizes these held-out trajectory controls by retain accuracy band. Every band includes checkpoints from all 30 trajectories.

Table 17: **Held-out trajectory control gaps by retain accuracy band.** Gaps are differences in probe recall, in percentage points, between a retain-only target checkpoint and matched checkpoints from the other trajectories. The controls are grouped by the target's retain accuracy.

	Saved			BN recalibrated		
Retain accuracy	n	$q_{.95}$	$q_{.975}$	n	$q_{.95}$	$q_{.975}$
$R < 85\%$	165	6.32	7.76	102	5.59	6.55
$85\% \leq R < 90\%$	193	4.28	4.64	281	3.90	4.60
$R \geq 90\%$	283	4.20	4.59	263	4.38	4.98

Within the high-utility band, the control gap has median $+0.1$ points in both conditions. The main audited gap exceeds the corresponding 95th-percentile control scale in 241 of 264 saved-condition cohort outcomes and 185 of 189 recalibrated outcomes. These are descriptive comparisons. The controls are dependent because trajectories contribute overlapping target checkpoints and reference neighborhoods. We report no confidence interval or p-value for them, and they are not deletion tests.

E Post hoc analyses

The analyses in this appendix reuse the 480 study checkpoints and reference ladder. The route, budget, deleted-versus-fresh, and readout capacity analyses were specified after the study was frozen and recorded as protocol amendments. The utility bounds below follow algebraically from their reported counts. They are deductions, not additional checkpoint evaluations.

E.1 Recovery by readout route

Woken recall is the ordinary classifier route, called the bias sweep in the result tables. Thresholding s_f gives held-out recall $\widehat{W}_\alpha(f)$ at nominal FPR α as defined empirically in Equation 3. Varying the threshold traces

$$C(f) = \left\{ \left(\Pr[s_f(X_r) > t], \Pr[s_f(X_c) > t] \right) : t \in \overline{\mathbb{R}} \right\}. \tag{8}$$

Its area counts how often a removed-class image outranks a retained image, with ties counted by half:

$$T(f) = \Pr[s_f(X_c) > s_f(X_r)] + \tfrac{1}{2} \Pr[s_f(X_c) = s_f(X_r)]. \tag{9}$$

The reversed route thresholds $-s_f$. Its AUC is $1 - T(f)$, while recall at a specified FPR requires its own calibration. A constant class bias edit preserves the margin ROC and decisions under a correspondingly shifted threshold for a fixed reference set (Proposition 1). A rematched gap can change if the edit changes retain accuracy enough to select a different neighborhood.

All three fixed routes use the same supported rows: the probe-supported and bias-supported sets have empty symmetric difference in both conditions, with 372 saved rows and 382 after recalibration. Table 18 therefore compares readouts on identical targets and reference identities. The fitting protocols and reference recalls remain route-specific.

Table 18: **Median gap by readout route on the same supported rows.** Gaps are medians of differences per row against each model's matched retain-only checkpoints, in percentage points. n_+ counts strictly positive gaps. The selected route is the calibration-selected diagnostic of Appendix A.6; each model chooses its route by calibration recall.

	As saved		BN recalibrated	
Route	Median gap	n_+/n_s	Median gap	n_+/n_s
Frozen-feature probe	$+10.25$	347/372	$+13.00$	358/382
Bias sweep	$+49.45$	347/372	$+55.10$	347/382
Reversed score	-0.20	35/372	-0.30	36/382
Calibration-selected (diagnostic)	$+11.50$	347/372	$+15.20$	358/382

The equal saved-condition positive counts do not identify the same outcomes. Both fixed gaps are positive on 328 rows, only the probe is positive on 19, only woken recall is positive on 19, and neither

Table 19: **Gap agreement and the location of woken-only positives.** Post hoc signs on identical supported targets and reference sets at nominal 5% FPR; nonpositive means $G \leq 0$. The final rows are bounds derived from the probe counts by utility, not a new row-level analysis (Appendix E.2). Each condition has its own support.

Woken gap	Probe gap	Saved ($n = 372$)	BN ($n = 382$)
Positive	Positive	328	335
Positive	Nonpositive	19	12
Nonpositive	Positive	19	23
Nonpositive	Nonpositive	6	12
Woken only, $R \geq 90\%$		≤ 1	0
Woken only, $R < 90\%$		18–19	12

is positive on 6. After recalibration, the counts are 335, 23, 12, and 12. These signs use the fixed audit partitions. Their sensitivity is discussed in Appendix D.4. The target-conditioned calibration-selected diagnostic matches the probe's sign on every supported row.

The readouts use different score constructions and different amounts of fitting supervision. Their reference baselines also differ: fully trained references have median woken recall of 36.4% as saved and 26.9% after recalibration, compared with probe recall of 88.4% and 82.0% (Table 13). These are endpoint summaries; Table 18 uses each target's matched reference baseline. The results describe recovery from existing scores and supervised feature recovery under their respective fitting protocols.

E.2 Cross-route agreement and retained utility

At nominal 5% FPR, write $G_{\text{woken}} = G_{\text{bias},0.05}$. Let B be the supported outcomes with $G_{\text{woken}} > 0$, P those with $G_{\text{probe}} > 0$, and U those with $R \geq 90\%$, within one normalization condition. Table 19 gives $|B \setminus P| = 19$ as saved and 12 after recalibration. Table 8 gives $|U \setminus P| = 1$ and 0. Therefore

$$|(B \setminus P) \cap U| \leq |U \setminus P|,$$

so at most one saved woken-only positive and no recalibrated woken-only positive has $R \geq 90\%$. The other 18–19 saved outcomes and all 12 recalibrated outcomes lie below 90%. The same upper bounds hold after also imposing $F \leq 1\%$. These deductions locate the disagreements relative to the paper's descriptive utility threshold; they do not distinguish moderate loss from severe degradation.

The only possible saved high-utility candidate is SCRUB, automobile deletion, block 0: $F = 0\%$, $R = 91.86\%$, probe recall 93.6%, matched probe reference recall 95.2%, and probe gap -1.6 points (Appendix C). The aggregate counts do not establish its woken gap sign, so we report *at most one*, not one.

The same counts also bound agreement in the low-forget, high-utility cohort $Q = \{F \leq 1\%, R \geq 90\%\}$. Because there are only 19 probe-only positives in the entire saved supported set and 23 after recalibration,

$$|B \cap P \cap Q| \geq |P \cap Q| - |P \setminus B|.$$

At least $263 - 19 = 244$ of the 264 saved cohort outcomes and $189 - 23 = 166$ of the 189 recalibrated cohort outcomes thus have both gaps positive. These are conservative bounds derived from reported totals, not exact subgroup counts or confidence limits. They support the operational value of recovery from existing scores where a probe also reports excess.

Scope of the cross-route counts. Table 19 reports signs on fixed audit partitions and matched reference sets. Exact subgroup identities, recall components of discordant rows, and their endpoint-relative gaps are not tabulated. The endpoint, equal-seed, and held-out trajectory results in Appendix D.6 concern probe recovery. The study has not measured the corresponding woken reference control distributions or paired calibration sensitivity. Consequently, the cross-route counts describe observed agreement and disagreement under the stated protocol, while the utility bounds follow from those counts alone.

E.3 SENSITIVITY TO THE FALSE-POSITIVE BUDGET

Amendment A4 recomputes the probe gap at a 1% false-positive budget on the same calibration split, applied to every model and reference checkpoint. On the same supported rows and matched retain-only checkpoints, the median gap changes from $+10.25$ to $+29.40$ points in the saved condition and from $+13.00$ to $+33.20$ after recalibration. Both recalls fall as the budget tightens. The median paired drop per row from the 5% to the 1% budget is 5.8 and 6.7 points for unlearned models, and 24.5 and 26.9 points for matched references. The reference ladder has median probe AUC 0.963 across 870 checkpoints, but its scores lie closer to the negative distribution, so its recall falls more under the stricter threshold. We report 1% as a second operating point, not as a larger estimate of the 5% effect.

Table 20: **The probe gap at five false-positive budgets.** Recalls are medians over supported rows; H_u is the unlearned models' probe recall and H_{ref} is the median recall of their matched retain-only checkpoints. Support is unchanged at every budget: 372 rows in the saved condition and 382 after recalibration, with the same checkpoints matched in every case. Gaps are percentage points; recalls are percentages. The prespecified 5% budget is the primary operating point.

	As saved			BN recalibrated		
Budget	G_{probe}	H_u	H_{ref}	G_{probe}	H_u	H_{ref}
0.5%	$+37.45$	87.6	50.4	$+41.90$	86.3	44.8
1%	$+29.40$	91.3	60.2	$+33.20$	90.2	55.0
2%	$+20.40$	93.9	71.8	$+24.50$	93.4	67.0
5% (primary)	$+10.25$	97.0	85.8	$+13.00$	96.4	81.4
10%	$+4.00$	98.6	93.7	$+6.20$	98.2	91.0

The matched reference set is identical at every budget because matching uses retain accuracy, which does not depend on the recovery threshold. None of the 960 rows changes its matched-checkpoint count. The reference's median recall drop therefore comes from thresholding a fixed set of checkpoints. Within the reference-supported saved cohort, median realized evaluation FPR tracks the target at every budget from 1% upward (1.000%, 1.978%, 4.956%, and 9.978%). The 4.956% value is computed over 372 supported rows; the 4.978% value in Section 4.5 uses all 480 unlearned checkpoints. The recall drop is larger than this threshold-transfer variation.

At the 0.5% budget, 5.9% of saved-condition rows and 7.6% of recalibrated rows realize less than half the target rate, compared with 2.2% and 2.1% at 1%. We include the 0.5% row for completeness and draw no conclusion from it.

The finite threshold calibration set contains 1,350 retained negatives for each removed class. At 1%, a quantile is determined by roughly the upper 14 of those scores; at 5%, it is determined by roughly the upper 68. This resolution tradeoff motivates treating the 5% setting as a declared audit convention and the stricter rows as sensitivity analyses. It is not a reconstruction of the historical choice of 5%.

E.4 DELETED EXAMPLES VERSUS FRESH IMAGES OF THE SAME CLASS

The study measures class-level recovery on held-out test images. A separate post hoc study, recorded in Amendment A5, tests whether a model treats the 4,500 removed-class *training* images differently from fresh images of the same class. For each model, let $p_M(x)$ be the decision function score from its separately fitted frozen-feature probe. We compute

$$A(M) = \Pr\left[p_M(X_{\text{train-forget}}) > p_M(X_{\text{test-forget}})\right] + \tfrac{1}{2}\Pr\left[p_M(X_{\text{train-forget}}) = p_M(X_{\text{test-forget}})\right],$$

the rank agreement between the deleted training images and that model's removed-class evaluation half. $A = 0.5$ means chance rank discrimination between deleted and fresh images under p_M; the score distributions need not coincide. The statistic is threshold free and invariant to any strictly increasing rescaling, so threshold placement and recall ceilings do not affect it. We report each model's excess over the median A of its matched retain-only checkpoints, under the same support rule, and the corresponding excess for the original model from the same block. The analysis uses the

probe route in the saved condition; its supported rows are the 372 supported saved-condition rows of the main study.

Retain-only references have $A = 0.5002$: they never saw the deleted images, and the statistic is near chance for them. The original models have $A = 0.5665$ and median excess 0.0655, positive in all 30 original/class probe combinations. These combinations come from three independently trained original networks and ten removed classes. The export repeats them across 16 method rows, producing 480 entries, but only 29 distinct original A values. Unlearned models have $A = 0.5403$ and median excess 0.0425, positive in 351 of 372 supported rows (94.4%). These are the existing matched reference results.

For the same 372 supported rows, we also pair each unlearned model with its original from the same block and use the unlearned row's reference baseline for both. The paired excess change is then $[A(u) - A_{\text{ref}}] - [A(o) - A_{\text{ref}}] = A(u) - A(o)$. Its median is -0.02585; 287 rows decrease, 31 are unchanged, and 54 increase. This paired score difference avoids changes caused solely by choosing a different reference neighborhood. It still does not identify a causal mechanism.

Table 21: **Excess rank agreement between deleted training images and fresh images of the same class, by method.** Saved condition, probe route. Each value is the median over supported outcomes of $A(M)$ minus the median for matched retain-only checkpoints. D_f marks access to actual deleted examples during the update. Original signal records use beyond the common initialization: T denotes a frozen original model teacher and S denotes a statistic or direction derived from the original model. BB marks a backbone update. These fields describe the evaluated implementation.

Method	D_f	Original signal	BB	Excess A	n_+/n_s
POUR-D	Yes	T+S	Yes	$+0.0690$	30/30
POUR-P	No	S	No	$+0.0650$	30/30
SSD	Yes	S	Yes	$+0.0630$	30/30
BoundaryShrink	Yes	T	Yes	$+0.0610$	3/3
InfluenceUnlearning	Yes	S	Yes	$+0.0570$	30/30
Spotter	Yes	T	Yes	$+0.0570$	30/30
BadTeacher	Yes	T	Yes	$+0.0520$	30/30
SalUn	Yes	S	Yes	$+0.0420$	30/30
NegGradPlus	Yes	–	Yes	$+0.0405$	30/30
FineTune	No	–	Yes	$+0.0290$	30/30
L1Sparse	No	–	Yes	$+0.0270$	30/30
SCRUB	Yes	T	Yes	$+0.0195$	25/28
GradientAscent	Yes	–	Yes	$+0.0155$	16/20
UNSIR	No	S	Yes	-0.0030	7/21
BoundaryExpand	Yes	–	Yes	N/A	0/0
FisherForgetting	No	S	Yes	N/A	0/0
Original models (reference point)	–	–	–	$+0.0655$	30/30

POUR-P provides an internal check on the estimator. Its update leaves the feature extractor unchanged, and deterministic separate probe fitting on identical feature inputs yields the same scores as its original from the same block. Its raw $A(M)$ is therefore identical in all 30 rows. Its reported matched excess can still differ because the update changes retain accuracy slightly and hence changes the reference window. The paired change with the same reference above remains zero for POUR-P.

Interpretation and limits. The statistic detects a dataset-level score difference between the 4,500 deleted training images and fresh images of the same class, beyond decodability of the class concept. For methods that do not revisit deleted examples, the difference cannot arise from direct optimization on those examples during the update. This removes one attribution route, but it does not establish per-example memorization or causal training influence. Methods that do revisit deleted examples can also introduce sample-specific effects during unlearning. Across all methods, the design cannot separate influence inherited from original training from influence introduced or changed during the update. Two methods contribute no supported outcome, and BoundaryShrink contributes one removed class.

E.5 Readout capacity and labeled budget

The prespecified probe protocol is linear and uses a fixed fitting set, so we swept both capacity and budget while keeping the other settings unchanged. We applied each readout and budget to the

unlearned models and every matched reference checkpoint. This sweep covers one removed class, airplane, across the three blocks, giving 35 supported outcomes; every recomputed score was checked bit for bit against the frozen cache, and the baseline cell reproduces its 48 study rows exactly. It is a limited readout budget check, not evidence that a nonlinear probe must increase or decrease a target–reference gap elsewhere.

Neither tested nonlinear readout improves separation in this subset. The linear probe has median held-out AUC 0.9919 on unlearned models and 0.9805 on matched references. A network with one hidden layer reaches 0.9890 and 0.9773, and a distance-weighted nearest-neighbor rule reaches 0.9631 and 0.9478. Both are lower on models and references. Their recall gaps of 9.2 and 17.0 points, compared with the probe's 7.6, do not test a stronger readout. We also report rank-based excess because it does not depend on the calibrated threshold.

The labeled budget was swept over four fitting budgets, $1{,}750$, $3{,}500$, $4{,}000$ and $4{,}500$ images, with the readout held linear and the threshold split held at a common 500 images, the largest budget leaving fitting and threshold sets that exactly partition the calibration half. Across those four the median gap was 10.2, 8.1, 7.8 and 7.8 points and the rank-based excess 0.0153, 0.0132, 0.0129 and 0.0119: the gap reached 7.8 at $4{,}000$ fitting images and remained there at $4{,}500$, while the rank-based excess fell across all four. The gap the main text reports is measured at one budget, and this comparison is confined to the airplane cell's three blocks and 35 supported outcomes.

Threshold data matters separately. The same probe at $3{,}500$ fitting examples gives a median gap of 7.6 points when its threshold is calibrated on $1{,}500$ images and 8.1 when calibrated on 500: half a point of difference from the threshold split alone, with the fitting set identical. The budget comparison above holds that split fixed at 500 for that reason.

F Method implementation and fidelity

Before including an implementation in the study, we compare it with a pinned upstream source. The manifest marks nine methods as *verified* and seven as *verified with documented deviation*; Table 22 gives the adaptations. Fifteen methods use the cat/block-0 validation selection described in Appendix B. POUR-P uses its closed-form default.

The fixed-transfer results include configurations that miss a validation criterion. Unsupported study rows have no recovery comparison. Poor utility under one selected configuration does not show that other configurations would fail, and we have not tested whether class-specific tuning would restore utility or change excess recovery.

F.1 Documented adaptations

BoundaryExpand preserves the extra-class training and subsequent head slicing while adapting the source's AllCNN feature/classifier split to ResNet-18. BoundaryShrink uses the base model's normalization constants when converting adversarial inputs to pixel space, clamping, and renormalizing. FisherForgetting estimates its Fisher information on a seeded subset of 5,000 retained examples, and the selected configuration uses noise cap 0.05. GradientAscent uses five unlearning epochs; its selected learning rate is 2×10^{-4}.

NegGradPlus uses the source objective and retain-driven update schedule, reshuffling the forget batches at each iterator restart. Spotter preserves the source's loss terms, adversarial precomputation, and frozen BN statistics while keeping the normalization used to train our original models. UNSIR places the model in evaluation mode during noise optimization. Its selected configuration uses impair learning rate 0.005, five repair epochs, and the full retained training split.

F.2 Projection and normalization buffer fidelity

POUR-P constructs $P = I - \bar{w}_c \bar{w}_c^\top$ from the normalized removed-class classifier row $\bar{w}_c$, projects retained classifier rows, and zeroes the removed-class row and bias. POUR-D uses a frozen original teacher and regresses student features toward projected teacher features on the removed-class training split. It uses the source's L2 regression objective for 20 epochs, with the classifier and BN running statistics frozen; validation selects learning rate 0.02.

Table 22: **Implementation provenance for the sixteen-method study.** V denotes verified and D denotes verified with a documented deviation. Commit prefixes identify the pinned upstream versions. Configurations follow the validation selection described in Appendix B.

Method	Upstream / commit	Status	Implemented operation or adaptation
BadTeacher	Bad Teaching / `f1aa988`	V	Mixed competent/incompetent teacher targets; student trained as in the source.
FineTune	Unlearn-Sparse / `518d87a`	V	Retain-only cross-entropy fine-tuning.
InfluenceUnlearning	Unlearn-Sparse / `518d87a`	V	WoodFisher perturbation and gradient scaling.
L1Sparse	Unlearn-Sparse / `518d87a`	V	Retain fine-tuning with decaying ℓ_1 regularization.
POUR-D	Representation Unlearning / `774ba5f`	V	Projected-teacher feature regression; frozen classifier and BN statistics.
POUR-P	Representation Unlearning / `774ba5f`	V	Closed-form classifier projection; zero removed-class row and bias.
SCRUB (Kurmanji et al., 2023)	SCRUB / `98ffb11`	V	Alternating forget maximization and retain minimization with the source KL loss.
SSD (Foster et al., 2024)	Selective Synaptic Dampening / `cdfdc0e`	V	Squared-gradient importance and bounded parameter dampening.
SalUn (Fan et al., 2024)	Unlearn-Saliency / `a6e89fd`	V	Saliency masking, random-label forget pass, and true-label retain pass.
BoundaryExpand	Boundary Unlearning / `19f9f23`	D	Adapt AllCNN head expansion and slicing to ResNet-18's final layer.
BoundaryShrink	Boundary Unlearning / `19f9f23`	D	Use base-model normalization for adversarial clamping.
FisherForgetting	Unlearn-Sparse / `518d87a`	D	Estimate Fisher information on a seeded 5,000-example retained subset; expose the noise cap for selection.
GradientAscent	Unlearn-Sparse / `518d87a`	D	Forget cross-entropy ascent; five local epochs and a selected learning rate.
NegGradPlus	SCRUB / `98ffb11`	D	Same retain/forget objective; reshuffle forget batches when restarting the iterator.
Spotter	Spotter / `17d3529`	D	Use the original model's CIFAR-10 normalization.
UNSIR	Fast Machine Unlearning / `252bc92`	D	Explicit evaluation mode during noise construction; fraction of retained data selected by validation.

The checkpoint audit finds BN running statistics identical to the corresponding original in all thirty checkpoints for FisherForgetting, InfluenceUnlearning, POUR-D, POUR-P, SSD, and Spotter. The normalization control measures how these checkpoints respond when retained data update those buffers.

The complete manifest provides upstream repository names, entry points, deviations, and checkpoint provenance for the included methods.

Table 23: **Cat/block-0 validation outcomes for the selected fixed-transfer configurations.** `OK` meets $F_{\text{val}} \leq 1\%$ and $R_{\text{val}} \geq 85\%$. `TRADEOFF-ONLY` meets the forget target but misses the retain floor, while `NOT-REACHED` misses the forget target. POUR-P has no search status because its released configuration is closed form; its validation outcome is still reported. Selection status and implementation fidelity are separate claims. Accuracies are percentages.

Method	Selection status	F_{val}	R_{val}
BadTeacher	`OK`	0.0	96.0
BoundaryExpand	`NOT-REACHED`	27.8	30.8
BoundaryShrink	`TRADEOFF-ONLY`	0.0	37.8
FineTune	`OK`	0.0	93.8
FisherForgetting	`TRADEOFF-ONLY`	0.0	11.1
GradientAscent	`TRADEOFF-ONLY`	0.0	80.5
InfluenceUnlearning	`OK`	0.4	92.2
L1Sparse	`OK`	0.0	94.6
NegGradPlus	`OK`	0.6	93.1
POUR-D	`OK`	0.0	96.7
POUR-P	N/A (closed form)	0.0	96.7
SCRUB	`OK`	0.0	94.2
SSD	`OK`	0.0	96.8
SalUn	`OK`	0.0	96.0
Spotter	`OK`	0.0	96.4
UNSIR	`TRADEOFF-ONLY`	0.0	79.8

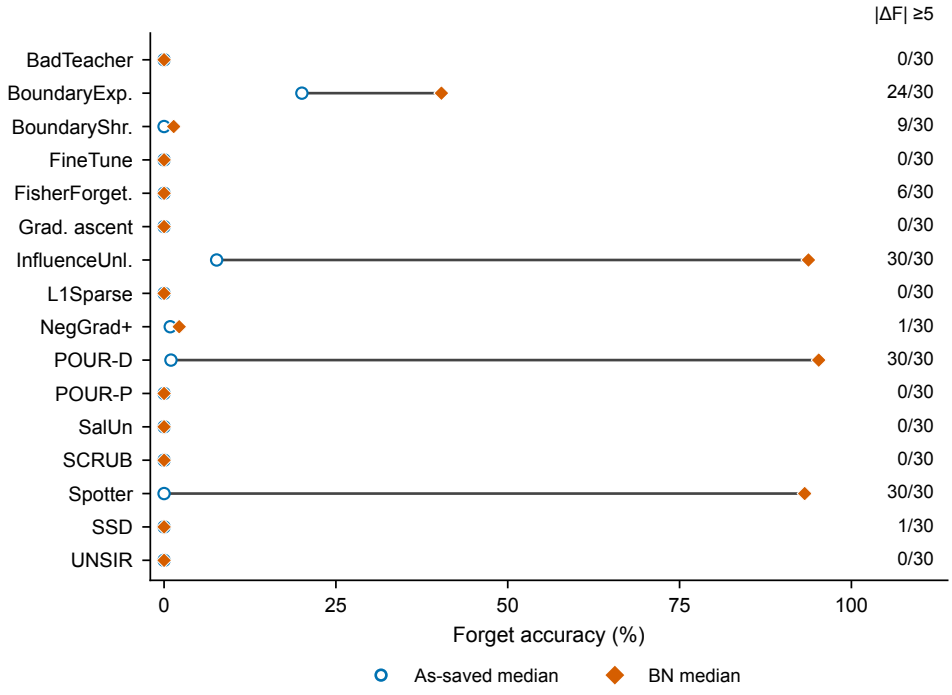

Figure 9: **Normalization control reveals differences in predictions.** Forget accuracy before and after BN recalibration on retained data for all methods. Endpoints show separate saved-condition and BN medians over the same thirty checkpoints per method. The count on the right records paired absolute forget accuracy changes of at least five points. Recalibration changes normalization buffers while holding learned parameters fixed. Figure 3 reports the separate probe comparison against matched retain-only checkpoints.

G VISION TRANSFORMER EXTENSION

This extension measures frozen-feature probe recovery, including reference headroom and paired original model contrast. Its conclusions concern that fitted readout; recovery from existing margins has not been evaluated here. An earlier ViT pilot trained its all-class original on all 50,000 CIFAR-10 training images before introducing the validation split, and its candidate grid was informed by test results. We exclude that pilot. The reported extension is the frozen rerun specified in `VIT_CLEAN_PROTOCOL.md` and executed at code revision `f05a2e2`. Model selection uses only cat deletion in block 0. The selected SalUn, SCRUB, and CF-k configurations transfer unchanged to automobile, cat, and ship deletion across three seeds. FineTune is a fixed sanity control.

Original and reference training. Every run constructs the ViT-S/16 model (Dosovitskiy et al., 2021) using `vit_small_patch16_224.augreg_in21k` from `timm` (Wightman, 2019), with AugReg ImageNet-21k pretrained weights (Steiner et al., 2022) and a fresh ten-way linear head, then fine-tunes the full model. Each class-excluded reference restarts from that pretrained initialization; it does not continue the corresponding all-class CIFAR-10 model. Originals use 45,000 training images, and references use the 40,500 retained images. The same 5,000 validation images are excluded from both. Training uses AdamW with learning rate 10^{-4}, weight decay 0.05, per-step cosine decay over ten epochs, batch size 64, and dropped final partial batches. Images are resized to 224×224, randomly flipped horizontally, converted to tensors, and normalized with ImageNet mean $(0.485, 0.456, 0.406)$ and standard deviation $(0.229, 0.224, 0.225)$. Evaluation omits the flip. The implementation and run records are in `scripts/vit_clean_train.py` and `checkpoints/vit_clean_ext/*/train_log.jsonl`.

Reference ladder and probe. Each class/block reference run saves rungs at steps 316, 632, ..., 6320, exactly twice per epoch for ten epochs. The final checkpoint alias is not a twenty-first matching rung. Matching stays within the same class and block and uses the local median of at least five rungs within ± 2 retain accuracy points; all 36 target rows have 16 to 20 matches. The probe feature is the 384-dimensional class token after the transformer's final LayerNorm and token pooling, before the ten-way linear head; `fc_norm` is the identity. The scaler, logistic probe ($C = 1$, `max_iter=2000`), and 3,500/1,500/5,000 fit, threshold, and evaluation partitions match Appendix B. The feature and scoring code is in `scripts/vit_clean_score.py` and `scripts/vit_clean_recovery.py`. ViT-S/16 uses LayerNorm, so the BN intervention does not apply.

Unlearning configurations. All four methods start from the corresponding all-class CIFAR-10 original, and all unlearning loaders use batch size 64. CF-k freezes the model, unfreezes the last k Conv2d or Linear modules in reverse traversal without reinitializing them, and fine-tunes those parameters on retained data. With $k = 1$ on this ViT, only the linear head is trainable. Exact overrides, configuration IDs, validation outcomes, and hashes are recorded in `vit_clean_ext/frozen_configs.json`; inherited optimizer details are defined in `src/methods/`.

Frozen ViT unlearning configurations. LR denotes the initial learning rate. Schedules described as constant have milestones beyond the reported run.

Method / ID	Reported setting
SalUn / `vce00`	SGD, LR 10^{-4}, two epochs, saliency fraction 0.5, momentum 0.9, weight decay 5×10^{-4}, constant LR.
SCRUB / `vce05`	Adam, LR 10^{-5}, three epochs with forget-set maximization in each, classification weight 1, distillation weight 0.5, temperature 4, weight decay 0.01, constant LR.
CF-k / `vce02`	$k = 1$, SGD, LR 0.1, five epochs, momentum 0.9, weight decay 5×10^{-4}, Nesterov, gradient clip 10, cosine warm restarts with $\eta_{\min} = 0.005$, $T_0 = 1$, and $T_{\text{mult}} = 2$.
FineTune / `vce_control`	Full-model retain-only SGD, LR 10^{-3}, three epochs, momentum 0.9, weight decay 5×10^{-4}, constant LR.

Table 24: **ViT-S/16 setting with ImageNet-21k pretraining.** F_{all} and R_{all} are medians over all nine transferred outcomes per method. Operating rows satisfy $F \leq 1\%$ and $R \geq 90\%$; recovery columns summarize only those rows. H_{ref} is the matched reference median within the same block, and $G_{\text{probe}} = H_{\text{probe}} - H_{\text{ref}}$. CF-k did not reach the validation forgetting target, and FineTune is not treated as an unlearning result.

Method	Status	F_{all}	R_{all}	Operating	H_{probe}	H_{ref}	G_{probe} (n_+/n)
CF-k	NOT-REACHED	37.0	98.4	0/9	–	–	–
FineTune	control	98.4	98.1	0/9	–	–	–
SCRUB	OK	0.0	97.9	9/9	99.2	99.8	−0.6 (2/9)
SalUn	OK	0.6	98.2	5/9	99.8	99.8	0.0 (2/5)

Realized probe FPR. Thresholds target 5% FPR on the threshold calibration split. On the separate evaluation split, the 36 ViT target checkpoints have median probe FPR 5.322% and range $[4.022, 6.133]\%$. The 14 low-forget, high-retain rows from validation-feasible methods have median 5.422% and range $[4.378, 6.044]\%$. Across the 180 reference rungs, the median is 4.956% and the range is $[3.689, 5.889]\%$. Target values are recorded in `vit_clean_ext/results/recovery.csv`; reference values are computed from the frozen `scores/*__probe.npy` arrays with the same saved partitions and scoring functions.

The two validation-feasible methods produce 14 low-forget, high-retain outcomes. Absolute recovery is near ceiling, but only 4 of 14 gaps are positive and the median gap is -0.1 points. The widespread positive probe excess in the ResNet study does not carry over to this pretrained ViT setting under the fitted probe audit. Near-ceiling reference recall leaves little headroom for positive recall gaps at this operating point. For an individual comparison with reference recall of 99.8%, the largest possible positive recall gap is 0.2 percentage points. The 99.8% value in Table 24 is a cohort median, so that bound does not apply to every row. The small ViT gaps therefore do not establish deletion or representation equivalence, and this extension does not isolate architecture from pretraining and other protocol differences.

The shared ImageNet-21k initialization further limits the deletion claim: the removed class is excluded from the CIFAR-10 fine-tuning stage, not necessarily from the full pretraining history. This is a prospective frozen rerun, not a test-naive preregistration. Earlier development informed the finite candidate families, but the reported selection and transfer did not use test results.

Fixed-reference comparison. For the target u and its original o, we use the target's reference set for both comparisons. Writing empirical probe recall and gap at operating point α as H_α, G_α,

$$\Delta_\alpha(u, o) = G_\alpha(u) - G_\alpha(o) = H_\alpha(u) - H_\alpha(o). \tag{10}$$

The reference term cancels, so the paired change reflects the two measured recalls. Each model retains its own probe and calibrated threshold. A negative Δ_α records lower target recall; the original–reference gap gives the starting contrast. Reference headroom, $100 - H_{\text{ref}}$ in points, is the largest possible positive gap for that reference. Table 25 summarizes the fixed operating cohort across budgets.

Table 25: **ViT control contrast depends on the operating point.** The 5% budget is primary; 2% and 1% are post hoc. Median paired probe gaps are in points, with positive counts in parentheses. Original/bias comparisons reuse nine probes across 36 reference sets. SalUn and SCRUB keep the same 5 and 9 cohort targets.

FPR	Ref. headroom	Original / bias	SalUn	SCRUB	Both methods
5%	0.2	+0.2 (24/36)	0.0 (2/5)	−0.6 (2/9)	−0.1 (4/14)
2%	0.5	+0.4 (28/36)	+0.2 (3/5)	−1.4 (3/9)	−0.1 (6/14)
1%	1.0	+0.8 (36/36)	+0.6 (4/5)	−1.4 (2/9)	−0.6 (6/14)

Fixed-reference sensitivity. We evaluate the three existing originals and bias-only controls at 5% FPR (primary), with 2% and 1% as post hoc budgets. Before evaluation, the control sets the removed-class head bias to -1000. Across 10,000 audit images per class/seed pair, only that scalar

changes: removed-class predictions disappear, while backbone features, frozen probe scores, and other-class logits remain exactly unchanged. `scripts/vit_clean_verify_controls.py` records the checks. No backbone training or unlearning retuning is performed.

Each original/control uses the corresponding target's existing reference rungs, without rematching. These 36 fixed-reference comparisons reuse nine class/seed probes and are not necessarily utility matched or independent. Checkpoints, fitted target/reference probes, partitions, cohorts, and reference sets stay fixed. Thresholds use each model's calibration negatives only; recall and realized FPR use the separate evaluation split. Results and hashes are under `vit_clean_ext/results/sensitivity/`; the row-level CSV is `vit_fixed_reference_sensitivity_rows.csv`.

Table 26: **ViT fixed-reference sensitivity.** Each method row covers all nine class/seed outcomes, including those outside Table 24's operating cohort. Original and bias-only rows reuse those probes across 36 fixed reference sets. Columns report median original, target, and reference recall (H_o, H_t, H_{ref}), paired original and target gaps (G_o, G_t), reference headroom ($100 - H_{\text{ref}}$), positive target-gap counts, and realized target/reference FPR. Gaps and headroom are computed per comparison before aggregation.

FPR	Target	H_o	H_t	H_{ref}	G_o	G_t	Headroom	n_+/n	FPR$_t$/FPR$_{\text{ref}}$
5%*	Original	99.6	99.6	99.8	0.2	0.2	0.2	24/36	5.556/5.022
5%	Bias only	99.6	99.6	99.8	0.2	0.2	0.2	24/36	5.556/5.022
5%	SalUn	99.6	99.8	99.8	0.2	0.2	0.2	5/9	5.422/5.022
5%	SCRUB	99.6	99.2	99.8	0.2	−0.6	0.2	2/9	5.200/5.067
5%	CF-k	99.6	99.6	99.8	0.2	0.2	0.2	6/9	5.556/5.022
5%	FineTune	99.6	99.6	99.8	0.2	0.2	0.2	6/9	5.178/5.022
2%	Original	99.6	99.6	99.5	0.4	0.4	0.5	28/36	2.133/1.933
2%	Bias only	99.6	99.6	99.5	0.4	0.4	0.5	28/36	2.133/1.933
2%	SalUn	99.6	99.6	99.5	0.4	0.2	0.5	7/9	2.089/1.933
2%	SCRUB	99.6	97.6	99.5	0.4	−1.4	0.5	3/9	2.200/1.933
2%	CF-k	99.6	99.6	99.6	0.4	0.4	0.4	7/9	2.133/1.933
2%	FineTune	99.6	99.6	99.5	0.4	0.4	0.5	7/9	2.022/1.933
1%	Original	99.6	99.6	99.0	0.8	0.8	1.0	36/36	1.067/0.856
1%	Bias only	99.6	99.6	99.0	0.8	0.8	1.0	36/36	1.067/0.856
1%	SalUn	99.6	99.6	99.0	0.8	0.8	1.0	8/9	1.111/0.856
1%	SCRUB	99.6	97.2	99.0	0.8	−1.4	1.0	2/9	1.178/0.856
1%	CF-k	99.6	99.6	99.0	0.8	0.8	1.0	9/9	1.067/0.844
1%	FineTune	99.6	99.6	99.0	0.8	0.8	1.0	9/9	1.044/0.856

*Primary budget; 2% and 1% are post hoc. Recalls/FPRs are %; gaps/headroom are points.

At 5% FPR, references leave 0.2 points of median headroom. Original and bias-only gaps have a $+0.2$-point median, although cat comparisons already show $+2.8$. Original gap medians rise to $+0.4$ and $+0.8$ at 2% and 1%. SCRUB reduces the paired original gap in all nine outcomes at both stricter budgets (median reductions of 2.0 and 2.6 points), whereas other methods have positive median gaps across their nine outcomes. Among Table 24's 14 operating rows, 6 of 14 gaps are positive at each stricter budget, with medians of -0.1 and -0.6. These fixed-partition measurements show limited headroom at 5% and method-dependent changes in recall. They do not estimate precision under probe refitting or new calibration splits (Appendix D.4). Stricter-budget separation indicates operating point sensitivity; shared pretraining remains a plausible, untested explanation for saturation.

H CIFAR-100 EXTENSION

This extension measures probe recovery at finer class granularity. The comparison using existing scores is confined to the CIFAR-10 study.

H.1 Scope, selection, and audit data

CIFAR-100 has the same total image count as CIFAR-10, but a 100-class label space.[1] This extension tests finer class granularity rather than larger dataset volume. We audit rabbit (class 65) and willow tree (class 96), each across three independently trained original ResNet-18 blocks. Unlearning configuration choices use house and apple validation deletion tasks, never the two audit classes. FineTune is a fixed-default comparison. The five reported methods yield six class/block outcomes each and share the same frozen class-specific retain-only reference ladders.

The audit fits a separate linear logistic probe ($C = 1$) to each checkpoint. A stratified partition assigns 35 test images per class to fitting, 15 to threshold calibration, and 50 to evaluation (Table 27). Thresholds use calibration negatives at a nominal 5% FPR; recall and realized FPR use the disjoint evaluation set. Targets and references receive the same images and budgets. The total split sizes match CIFAR-10, but the removed-class fitting and evaluation counts are ten times smaller.

Table 27: **CIFAR-100 audit allocation for one class deletion.** Fitting, threshold calibration, and evaluation are disjoint. Forget and retain accuracy in this extension use the evaluation subset.

Partition	Removed class	Retained classes	Total
Probe fitting	35	3,465	3,500
Threshold calibration	15	1,485	1,500
Recovery evaluation	50	4,950	5,000

Each deleted class has three retain-only reference seeds, with 29 saved checkpoints per seed and a fully trained endpoint at epoch 200. The 87 rungs are snapshots of three training trajectories, not independent trained models. Conditional reference recall is the median over rungs within ± 2 percentage points of the target's *evaluation-subset* retain accuracy; at least five rungs are required. This differs from the main study, which matches on all retained test images. Probe fitting and threshold calibration remain disjoint from evaluation, but the utility used to select reference neighborhoods is measured on the evaluation negatives. Endpoint gaps instead subtract the median recall of the three fully trained references, without utility matching.

For each block, retained accuracy loss is the original model's accuracy minus the unlearned model's accuracy on the same retained evaluation images. Negative loss denotes an improvement. We report this loss continuously and distinguish a preservation tier allowing at most two points of loss from the looser ten-point tolerance used to contextualize FineTune. Neither replaces the CIFAR-10 cohort definition nor serves as a universal criterion for acceptable unlearning. Raw targets use raw references and originals; BN-recalibrated targets use their recalibrated counterparts.

H.2 Evaluated implementations

The runs are CIFAR-100 adaptations of the named approaches, not exact end-to-end reproductions of their published benchmark settings. SalUn (Fan et al., 2024) uses the validation-selected configuration `SalUn__t1_00`: a fixed-label joint update, mask threshold 0.3, learning rate 0.01, ten epochs, batch size 256, and update seed 0. CF-k uses the selected CIFAR-100 adaptation. BadTeacher (Chundawat et al., 2023a) uses the lower learning rate selected on the validation tasks, and FineTune retains its fixed default. Their executable settings are recorded in the corresponding frozen protocols.

The reported SSD (Foster et al., 2024) port aligns the full-data importance loader with the source's class-grouped retained examples followed by forget examples, using batch size 64. SSD squares each minibatch's mean gradient, so both batching and ordering affect the importance estimate. An earlier port used mixed full-data batches of size 256; those results are excluded and are not attributed to the published method. The corrected run's saved candidates, checkpoints, scores, and result rows were checked for consistency. A fixed-default GradientAscent run did not meet the joint forgetting and utility criteria on these cells and was not carried forward into the broader class run; this characterizes that configuration rather than the method as a whole.

[1] Official dataset description: `https://www.cs.toronto.edu/~kriz/cifar.html`.

H.3 MULTI-METHOD RESULTS AND REFERENCE COVERAGE

Table 28 reports the raw condition. All five methods attain zero observed forget accuracy in all six cells, each evaluated on 50 removed-class images. CF-k, BadTeacher, and SSD each preserve retained accuracy to within 0.79 points and show positive conditional gaps in all six outcomes. SalUn also preserves utility, but positive excess appears only for rabbit. Reporting methods separately retains this difference rather than treating their shared-image outcomes as independent evidence.

Table 28: **CIFAR-100 results on two held-out deletion classes and three original model blocks.** Raw condition, nominal 5% FPR. Loss and gaps are percentage points. $n_s/6$ reports reference coverage separately from positive gaps n_+/n_s; median gaps use only supported cells. All rows are the evaluated adaptations described above.

Method	$F = 0$	Retain loss range	Loss ≤ 2	$n_s/6$	n_+/n_s	Median gap
SalUn	6/6	$[0.73, 1.80]$	6/6	6/6	3/6	$+5.5$
CF-k	6/6	$[-0.18, 0.51]$	6/6	6/6	6/6	$+19.5$
BadTeacher	6/6	$[-0.06, 0.79]$	6/6	6/6	6/6	$+16.0$
SSD	6/6	$[-0.28, 0.65]$	6/6	6/6	6/6	$+14.5$
FineTune	6/6	$[6.46, 8.73]$	0/6	2/6	2/2	$+9.0$

FineTune reaches zero forget accuracy with 6.46–8.73-point retained accuracy loss. Only two raw cells have at least five local references; both gaps are positive. The other four conditional gaps remain unsupported. Fully trained endpoint comparisons are available and positive for all six raw cells, but do not match utility. After BN recalibration all six FineTune cells have reference support and five gaps are positive. These comparisons are reported alongside the lower-loss methods rather than treating a larger utility loss alone as evidence of a broken implementation.

H.4 SALUN: CLASS-DEPENDENT EXCESS AND CONTROLS

All six raw SalUn checkpoints have zero observed forget accuracy and at most 1.80-point retained accuracy loss. Their neighborhoods contain 11–15 rungs, with all three reference seeds represented. Table 29 shows that absolute recall is similar across the two classes, but the reference-relative reading differs. Rabbit recall is 66–68%, above a 51% matched reference; willow tree recall is 68–70%, below a 74% reference. Fully trained endpoints agree with the matched reference signs in every cell. The pooled median gap of $+5.5$ points therefore summarizes a class split, not uniformly positive recovery.

Table 29: **Raw SalUn recovery by class and block on CIFAR-100.** All six cells have $0/50$ correct removed-class predictions. R is retain accuracy, L_R its loss against the original from the same block, H_u probe recall, and H_{ref} matched reference recall. G_{end} uses the median of three epoch-200 endpoints. Accuracies, recalls, and FPRs are percentages; losses and gaps are points.

Class	Block	R	L_R	H_u	FPR	H_{ref}	Rungs	G	G_{end}
Rabbit	0	76.79	0.73	66	5.43	51	12	$+15$	$+16$
Rabbit	1	77.41	1.35	68	4.75	51	12	$+17$	$+18$
Rabbit	2	77.31	1.25	68	4.83	51	12	$+17$	$+18$
Willow tree	0	76.44	1.11	70	6.51	74	15	-4	-4
Willow tree	1	76.93	1.80	70	4.75	74	12	-4	-4
Willow tree	2	77.68	0.95	68	4.63	74	11	-6	-6

After BN recalibration on retained data, all six SalUn comparisons remain supported. Rabbit gaps are $+18$, $+10$, and $+21$ points for blocks 0–2; willow tree gaps are $+2$, 0, and -4. Absolute retained accuracy is approximately 72–74%, rather than 76–78% in the raw condition, so comparisons use BN-recalibrated references and originals. Realized FPR ranges from 4.22–6.34% after recalibration, versus 4.63–6.51% raw. The common nominal 5% budget does not make evaluation FPR identical across models.

A fixed bias-only shift drives the raw original's forget accuracy from 62–72% to zero in every class/block cell while leaving penultimate features and probe scores unchanged. Median raw probe recall across the three blocks remains 78% for rabbit and 82% for willow tree. This is a constructed non-removal control, not an unlearning method.

H.5 Artifacts and limits of the extension

The extension files reside under `revision_analysis/cifar100_24h_extension/`. The SalUn/reference/control bundle contains `recovery.csv`, `controls.csv`, `references.csv`, `summary.json`, `PROTOCOL.md`, `extension.py`, and `test_extension.py`. Multi-method, selected BadTeacher, and corrected SSD rows are in `multimethod/recovery.csv`, `rescue/recovery.csv`, and `ssd_rescue/recovery.csv`, respectively, with their protocols and executable settings in the same directories. These exports retain raw and BN conditions separately.

The SalUn/reference/control verification records cover 183 checkpoint hashes, 186 score file identities and hashes, and recomputation of 384 result rows: 348 reference rows, 24 control rows, and 12 SalUn rows. This inventory concerns that bundle, not all later multi-method artifacts. Checkpoint and score consistency checks support reproducibility; they do not provide statistical independence or additional class coverage.

With 50 evaluation positives, each model's recall changes in two-point steps; medians over reference checkpoints can have finer increments. The same evaluation images recur across blocks, and reference rungs share training trajectories. Configuration selection is separate from the two audit deletion classes, but these observations do not estimate excess-recovery prevalence over all 100 classes. They show that substantial excess can coexist with small utility loss beyond CIFAR-10, while SalUn's class-dependent signs illustrate why absolute recovery and nonpositive gaps require a reference-aware reading.