OPTING OUT SHOULD NOT BE DETECTABLE: FRESH-IDENTITY INDISTINGUISHABILITY FOR IDENTITY UNLEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Unlearning is evaluated by what a model can still retrieve, not by whether the deletion request is visible. We define *fresh-identity indistinguishability* (FII) relative to an adversary class and test for its violation with the paired detection-AUC gap between a released encoder and its same-seed retrain oracle on identical identities and splits: a positive gap witnesses a violation. At 100 opted-out Market-1501 identities, an adversary holding no training image reaches AUC 0.802 ± 0.020 against the oracle's 0.622 ± 0.081 (gap $+0.180$, 5/5 seeds), naming 17% of requesters at a 1% false-accusation rate; one unknowingly holding some training images reaches 0.968 ± 0.008 against 0.717 ± 0.019 (gap $+0.252$, TPR 0.588). The mark replicates on ResNet-18 and ConvNeXt-T encoders and on CUHK03-NP, and a detector fit on one opt-out set reads it off disjoint identities (0.801 against 0.591 on their oracle, 20 of 20 pairs), off other architectures and off the other dataset, and it stays detectable through a retrieval service that returns only top-10 ranks. It is a tighter-than-never-seen cluster: forgotten identities remain more self-similar than people the model never saw. Across 39 operating points of five unlearning methods, none keeps retain mAP at or above 0.95, brings forget mAP to at most 0.02 above the oracle's and passes the mixed-adversary FII test; the methods that preserve utility best, `neggrad_plus` and `finetune`, leak on every seed under the mixed adversary. Our evidence is three academic benchmarks, not face retrieval.

1 INTRODUCTION

A deletion request is not anonymous. When a person invokes a right to erasure (European Parliament and Council of the European Union, 2016) the operator learns who they are; a deletion process should not itself disclose who asked, and we treat that as a privacy desideratum. It concerns the released artifact, not the request log: the published model should carry no mark identifying the people removed from it. The unlearning literature asks whether forgotten data can be recovered or its membership inferred (Shokri et al., 2017; Carlini et al., 2022); we ask whether the removal itself is visible.

We name the violated property *fresh-identity indistinguishability* (FII): after a person has been unlearned, they should be statistically interchangeable, in everything the released model exposes, with a person the model never saw. Section 3 gives it a testable form as a paired gap to a same-seed retrain oracle, because the retrain oracle itself is separable from never-seen identities, so absolute indistinguishability is not the appropriate retrain-relative reference.

Two adversaries. The *mixed* adversary's probes are images withheld from training while the reference half of each identity is images the model trained on, so it unknowingly holds training images of its targets. The *fresh* adversary draws probe and reference from the withheld set alone, so no image it touches was seen by any model; never-seen identities get the same image budget. On Market-1501 with a ResNet-50/BNNeck embedding, 100 identities removed and five seeds, the fresh adversary separates the opted-out identities from never-seen ones at AUC 0.802 ± 0.020 against a retrain oracle at 0.622 ± 0.081: a paired gap of $+0.180$, significant on 5/5 seeds, with TPR 0.170 at 1% FPR. The

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.

mixed adversary reaches 0.968 ± 0.008 against 0.717 ± 0.019, gap $+0.252$, TPR 0.588 (per seed 0.468 to 0.720). Table 1 reports both.

We make five empirical observations. The attack scales and replicates: fresh 0.756 / mixed 0.914 at 20 identities and 0.802 / 0.968 at 100 on Market-1501, and the same gap on a ResNet-18 (0.794 against 0.603) and a ConvNeXt-T (0.770 against 0.614) trained on the same identities, on CUHK03-NP with a ResNet-50 (0.752 against 0.574), and on Tiny-ImageNet as a class-level replication whose oracle is consistent with chance, every one on 3/3 seeds under the fresh adversary. It transfers: a detector fitted on one seed's opt-out set reads disjoint identities at 0.801 against 0.591 on their retrain oracle (20 of 20 ordered pairs, fresh adversary), other architectures and the other dataset, and the attack survives a service returning only top-10 ranks (Section 6); the detector is supervised, and this is its label-free evidence. It has a mechanism: the forgotten identities remain a tighter cluster than never-seen people; pairwise own-image similarity alone separates them at AUC 0.89 where the oracle reads 0.52 (Section 6). Published unlearning does not remove it: at $n = 100$ over five seeds, `neggrad_plus` holds retain mAP 0.9992 and leaks at AUC 0.942, `finetune` holds 0.959 and leaks at 0.908, both 5/5 under the mixed adversary, and across 39 operating points of five methods no configuration keeps retain mAP at or above 0.95, brings forget mAP to at most 0.02 above the oracle's and passes the mixed-adversary FII test (Section 7).

Two conventions kept this hidden: probes drawn from training images put the retrain oracle at AUC 0.398, below chance, so nothing could be calibrated against it; and single-run reporting hides that selected operating points sit inside bistable bands (Section 8). Both are measurement choices, not facts about unlearning.

2 Related work

Membership inference. Membership inference asks whether a sample was in the training set (Shokri et al., 2017; Yeom et al., 2018; Carlini et al., 2022). Unlearning evaluation borrows the instrument: a method passes when a membership attack on forgotten samples does no better than on never-trained ones (Triantafillou et al., 2024), a convention Hayes et al. (2024) argue is too weak. User-level membership inference already lifts the unit from a sample to a person: Li et al. (2022) ask whether any image of a person trained a re-identification embedding, and FACE-AUDITOR (Chen et al., 2023) audits few-shot face recognisers from a set of a person's images and their similarity statistics. Our detector shares those inputs; the question differs: not whether a person was ever represented in training, but whether, after a deletion procedure, the released encoder remains distinguishable from the exact-retraining counterfactual on photographs no model saw. Taking the retrained model as the reference against which approximate unlearning is judged follows exact unlearning (Bourtoule et al., 2021) and Thudi et al. (2022); we adopt the same reference for detectability. The nearest prior detectability claim reads unlearning traces off language model outputs (Chen et al., 2026b), where the trace names the procedure, not the person.

Evaluating unlearning. IDI scores from intermediate features how much forgotten information survives (Jeon et al., 2026); LoTUS removes the oracle from the loop with a retrain-free Jensen–Shannon divergence (Spartalis et al., 2025). Both ask how much forgotten material remains; neither asks whether the request itself is legible from outside. Deep Unlearn benchmarks eighteen methods (Cadet et al., 2025), SCRUB among them (Kurmanji et al., 2023); none reports a fixed configuration bistable under the run seed alone (Section 8).

Identity unlearning in retrieval. MUFAC and MUCAC benchmark forgetting identities in facial age and attribute classifiers (Choi & Na, 2023); TIU-ReID removes a target identity from a re-identification model and scores its retrievability (Chen et al., 2026a); Damnatio Memoriae disperses or re-maps selected identities in a face embedding (Öztürk et al., 2026); De-ReID shifts an unlearned person's embedding along that person's own image variations (Peng et al., 2025); embedding dispersion scatters a forgotten face identity's cluster across the hypersphere (Zakharov, 2026). All five score success as suppression of the forgotten identity. We measure the complementary quantity, which under the mixed adversary at 100 identities orders the other way: the methods that suppress least leak most (Section 7). We are not aware of prior work asking whether the deletion request itself is inferable from a released embedding, or calibrating that inference against a same-seed retrain oracle. The nearest antecedents are Joshi & Wang (2026b), who measure how non-face identity link-

age accumulates across cumulative releases of data, checkpoints and metadata, and Joshi & Wang (2026a), who audit what a forget set leaves in released embedding banks and find that on Market-1501 even the retrain-from-scratch anchor still retrieves a removed identity, the anchor collapse that the paired gap of Section 3 is designed to absorb. Neither asks whether an identity's deletion status is inferable from the released encoder, which is this paper's question.

3 THREAT MODEL AND THE FRESH-IDENTITY INDISTINGUISHABILITY CRITERION

What the adversary has. The adversary queries the released encoder for embeddings and holds a few images of the target, the access an embedding API grants; Section 6 weakens this to services that return only top-k scores or ranks. It has no weights, gradients, activations or way to retrain, and no knowledge of whether the target opted out. It does need one labelled opt-out set to fit a detector, obtained once, not necessarily from the target's model: the detector transfers across disjoint identities, across architectures and across datasets, and works through a ranks-only retrieval service (Section 6).

Two adversaries. The primary detection experiments are reported under two adversaries. *Mixed*: probes are the images withheld from training, and the identity's reference half is its remaining images, which the model trained on, so the adversary unknowingly holds some training images of the target. *Fresh*: probe and reference both come from the withheld set, so no image the adversary touches was seen by any model; never-seen identities get the same budget, a median of about four images split $2/2$, so the two groups are compared on identical evidence. Fresh leads, mixed is stronger, and *no training images* describes the fresh adversary alone. We withhold 30% of each forget identity's images: $118/119/93$ probes per seed at 20 identities, $484/457/498$ at 100.

The criterion. Write $\mathrm{AUC}_d(f)$ for the detector's area under the curve when it separates the forget identities of matched draw d from their matched never-seen identities in the embedding of encoder f. For every draw we score two encoders on identical identities, identical probe and reference splits and identical features: the released encoder, and a *retrain oracle* trained from the same seed on the retain set alone, the model that would have existed had the request been honoured by retraining. The reported quantity is the paired gap

$$\Delta_d = \mathrm{AUC}_d(\text{released}) - \mathrm{AUC}_d(\text{oracle}), \tag{1}$$

For an adversary class $\mathcal{A}$, a release satisfies *fresh-identity indistinguishability* (FII) with respect to $\mathcal{A}$ when no $A \in \mathcal{A}$ reads the deletion request from the released embedding more easily than from a model that never trained on the person, $\sup_A \Delta^A \leq 0$. Our detector is one member of any such class, so its gap is a witness rather than a certificate: $\Delta > 0$ shows that FII fails for every class containing it, while $\Delta = 0$ means only that the release passes this operational test, with these ten statistics, this sample size and this adversary; a more expressive attacker could separate what a linear one cannot. We test $\Delta = 0$ with a paired permutation in which one label shuffle is applied to both encoders at once, and report $\bar{\Delta}$ with its spread on every row of every table. The absolute form of the property, indistinguishability from a person the model never saw at all, would require the released encoder to sit at the never-seen null floor. The retrain oracle itself fails that form: at 100 identities it reads AUC 0.719 ± 0.028 against a null floor of 0.481 ± 0.059, which is chance. Retraining is the exact-unlearning reference, the distribution against which a deletion request is honoured, so a criterion it fails is the wrong yardstick for approximate methods that are measured against it. The retrain-relative gap is the operational target: it charges a method only for what it leaves behind beyond retraining, and whatever the probe protocol and the negative population do to the raw estimator enters both arms identically and cancels.

Setup. Market-1501 (Zheng et al., 2015): ResNet-50 with BNNeck, 2048-d (Luo et al., 2019), 751 identities, 20 or 100 forgotten. Tiny-ImageNet (Le & Yang, 2015): a class-level replication without Market-1501's count and camera covariates, ResNet-18, 200 classes, 20 forget, 130 retain, 50 never-seen, no cameras and 500 images per class, so no covariate exists. The Market-1501 ResNet-50 headline uses five seeds; every other setting three; seeds vary draw and init. The same identities, holdouts and draws are reused to train a ResNet-18 (512-d) and a ConvNeXt-T (768-d)

Table 1: Market-1501 detection under both adversaries. Each forget identity is matched to a never-seen identity on image count and camera count and subsampled to match, five draws per seed; five seeds at $n = 100$, three at $n = 20$; spread is across seeds. The paired gap is $\bar{\Delta}$ of Eq. (1) and significance is the paired permutation test of Section 3. Under the mixed adversary at $n = 100$ the never-seen versus never-seen null floor is 0.481 ± 0.059, which is chance. TPR at 1% FPR is not expressible on the $1/20$ grid of the 20-identity design.

	mixed, n=100	fresh, n=100	mixed, n=20	fresh, n=20
Released original, AUC	0.968 (0.008)	**0.802** (0.020)	0.914 (0.019)	**0.756** (0.014)
Retrain oracle, AUC	0.717 (0.019)	0.622 (0.081)	0.565 (0.068)	0.525 (0.032)
Paired gap $\bar{\Delta}$	$+0.252$	$\mathbf{+0.180}$	$+0.349$	$\mathbf{+0.231}$
Significant seeds	5/5	5/5	3/3	3/3
TPR @ 1% FPR	0.588	**0.170**	–	–
TPR @ 5% FPR	0.840	0.393	–	–

with the same recipe, and CUHK03-NP (Li et al., 2014; Zhong et al., 2017) (767 training identities, 10 images each, two cameras) is run with the ResNet-50 at 100 forgotten identities. Appendix A holds Table 4.

4 THE DETECTOR AND ITS CALIBRATION

Per-identity geometry. Each scored identity splits into a probe half and a reference half, defined per adversary in Section 3. Probes are scored against their own reference half and a gallery of all scored reference halves plus 300 retain and 300 test distractor identities, fixed per seed. Ten scalars summarise the geometry (Appendix C): `pos_mean`, `pos_max`, `xcam_pos_mean`, `hardneg`, `margin`, `top1`, `top10_mean`, `negtop10_mean`, `centroid_dispersion` and `pair_sim_mean`. The same ten transfer unchanged to Tiny-ImageNet.

Estimator. Features are standardised per fold and classified by L_2-regularised logistic regression ($C = 1$, liblinear) (Pedregosa et al., 2011; Fan et al., 2008). Scores are out-of-fold decision values from stratified 5-fold cross-validation repeated 5 times, pooled before AUC. Each seed contributes 5 matched draws; the paired permutation test on $\bar{\Delta}$ applies one label shuffle to both encoders and to all five draws at once and averages the five null gaps, 2,000 replicates, so the null is that of the reported statistic (Ojala & Garriga, 2010) (Appendix C gives the audit against the within-draw null the scripts first used), and each seed's AUC is the mean of its five draws, with dispersion the standard deviation across seeds. Scores are out-of-fold on the scored identities, so the attack is supervised: the adversary needs some identities' opt-out status to accuse others, a cost paid once (Section 6).

Three calibrators. The same-seed retrain oracle of Equation 1. The null floor: never-seen against never-seen at matched n through the identical pipeline in the oracle's embedding, 0.481 ± 0.059 at 100 identities, which is chance. A covariate-only detector over six bookkeeping features, touching no embedding: 0.397 ± 0.057, so per-identity bookkeeping separates the groups in the direction opposite to membership and cannot supply the attack (Appendix B).

5 OPT-OUT IS DETECTABLE FROM RELEASED EMBEDDINGS

An adversary holding no training image of any target separates 100 opted-out identities from 100 never-seen ones at AUC 0.802 ± 0.020 against the same-seed retrain oracle's 0.622 ± 0.081: a paired gap of $+0.180$, significant on 5/5 seeds. At a 1% false-accusation rate that adversary names 17.0% of requesters (TPR 0.170, per-seed 0.082 to 0.253) and 39.3% at 5% (per-seed 0.318 to 0.474). The mixed adversary, who unknowingly holds some of the target's training images among the reference half, is stronger: AUC 0.968 ± 0.008 against an oracle at 0.717 ± 0.019, gap $+0.252$, 5/5 seeds, TPR 0.588 at 1% FPR (per-seed 0.468 to 0.720) and 0.840 at 5%. Table 1 reports both adversaries at both scales; Appendix C gives the detection curves.

Table 2: The released original against its same-seed retrain oracle in every setting, held-out probes, 100 forgotten identities; five seeds for the first row, three for the others; draw-level permutation test. CUHK03-NP's mixed row is saturated by the protocol's split asymmetry, for the oracle as much as for the released model, and the paired gap records that correctly.

	mixed adversary				fresh adversary			
Setting	AUC	oracle	gap	sig.	AUC	oracle	gap	sig.
Market-1501, ResNet-50	0.968 ± 0.008	0.717	$+0.252$	5/5	0.802 ± 0.020	0.622	$+0.180$	5/5
Market-1501, ResNet-18	0.980 ± 0.005	0.726	$+0.255$	3/3	0.794 ± 0.022	0.603	$+0.192$	3/3
Market-1501, ConvNeXt-T	0.928 ± 0.012	0.698	$+0.229$	3/3	0.770 ± 0.021	0.614	$+0.156$	3/3
CUHK03-NP, ResNet-50	0.999 ± 0.001	0.985	$+0.014$	0/3	0.752 ± 0.026	0.574	$+0.178$	3/3

Twenty identities. The ordering holds at the smaller scale used elsewhere in the paper: fresh gap $+0.231$, mixed $+0.349$, $3/3$ seeds each (Table 1). An operating point at 1% FPR is not expressible at this scale.[1]

Other encoders and a second identity dataset. The same identities, holdouts and draws train a ResNet-18 and a ConvNeXt-T; CUHK03-NP is run with the ResNet-50 at 100 forgotten identities (Table 2). Under the fresh adversary the gap reproduces everywhere: ResNet-18 0.794 against 0.603 (gap $+0.192$, 3/3), ConvNeXt-T 0.770 against 0.614 ($+0.156$, 3/3), CUHK03-NP 0.752 against 0.574 ($+0.178$, 3/3). CUHK03-NP also shows why the criterion is a paired gap and not a level: with two cameras and ten images per identity, the mixed protocol's asymmetric splits separate the groups on geometry alone, so the retrain oracle itself reads 0.985 and the released encoder 0.999, a gap of $+0.014$ (0/3). A criterion stated as distance from chance would have called that a leak; the paired gap reads it as nothing beyond retraining.

A class-level replication without the covariates. Tiny-ImageNet is a class-level proxy, not identity data and not evidence of generalization to identities; it is here because every class has exactly 500 training images and there are no cameras, so a detector reading per-identity bookkeeping has nothing to read. With a ResNet-18 over 200 classes the released original separates forgotten from never-seen classes at AUC 0.9998 ± 0.0003, paired gap 0.4808, $3/3$ seeds, and the oracle reads 0.519 ± 0.136, consistent with chance at $n = 3$ (Appendix D).

In persons: of 100 opt-out identities, at one false accusation in a hundred the fresh adversary names 17, the mixed adversary 59, and a from-scratch retrain 6.

6 What the signal is, and what it is not

We next test whether the signature transfers, survives restricted-output access, and has a feature-level structure.

It transfers across identities, architectures and datasets. We fit the detector on one seed's forget identities, pooling that seed's five matched draws, and apply the frozen scaler and classifier elsewhere, always twice: to the target's released encoder and to the target's retrain oracle, the control that shows what the detector has learned (Figure 1, Appendix H). Across the five Market-1501 seeds the twenty ordered pairs of disjoint identity sets read 0.801 on the released encoder against 0.591 on the oracle under the fresh adversary, a gap of $+0.210$ on 20 of 20 pairs, and 0.988 against 0.568 under the mixed one. A detector fitted on the ResNet-50 reads a ResNet-18 trained on a disjoint seed's identities at 0.798 against 0.570 and a ConvNeXt-T at 0.763 against 0.604 (fresh; 6 of 6 pairs each). A Market-1501 fit reads CUHK03-NP at 0.709 against 0.559 and a CUHK03-NP fit reads Market-1501 at 0.621 against 0.489 (fresh, all nine pairs positive); under the mixed adversary the same pairs read 0.995 and 0.888 on the released encoders. The signature is generic to prior training exposure, not to particular people, a backbone or a corpus: these transfers read the released

[1] With 100 negatives the empirical false-positive grid resolves in steps of $1/100$, so TPR at 1% FPR is set by a single order statistic of the negative scores and TPR at 0.1% FPR is unestimable; the 20-identity design resolves only $1/20$. We therefore report TPR at 1% with its per-seed spread and never as a point.

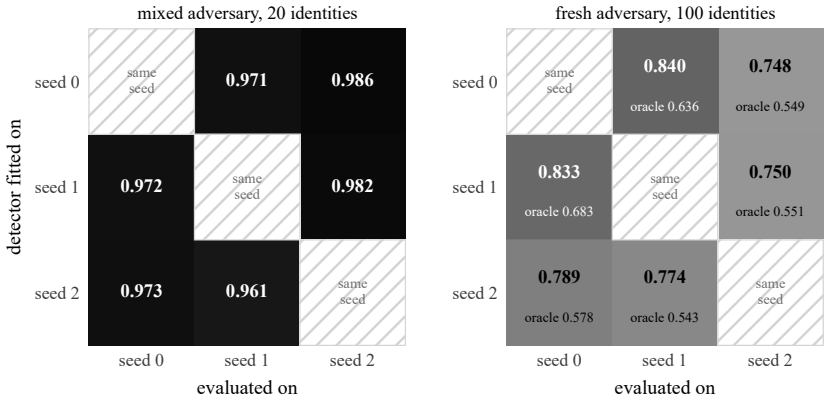


Figure 1: A detector fitted on one seed's opt-out identities and evaluated, unchanged, on a disjoint seed's. The forget sets share no identity, so transfer means the detector reads a generic prior-training signature rather than anything about particular people. Left: mixed adversary at 20 identities, mean 0.974 over six ordered pairs. Right: fresh adversary at 100 identities, three seeds, where no adversary holds a training image; each cell gives the transferred detector's AUC on the target seed's released encoder and, beneath it, on that seed's retrain oracle. The five-seed, cross-architecture and cross-dataset transfers are tabulated in Appendix H.

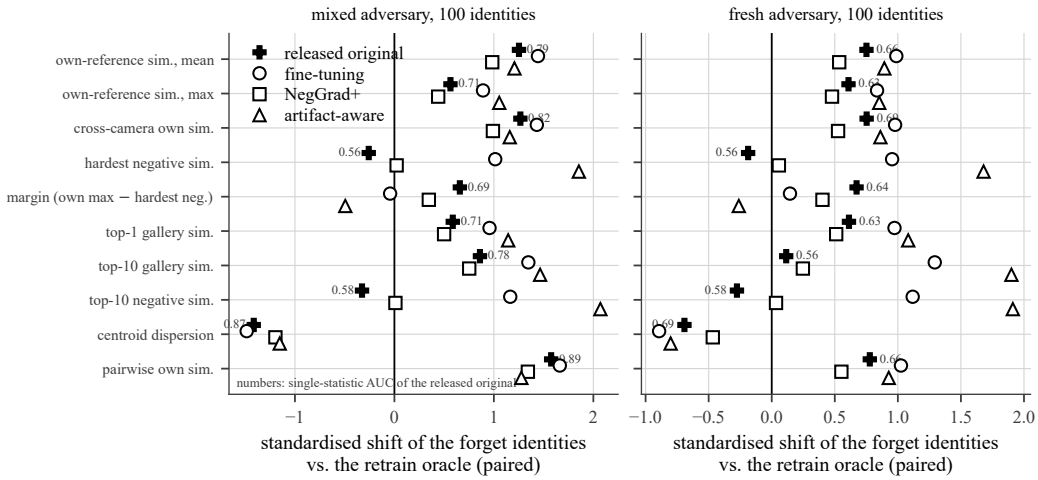


Figure 2: What the detector reads. For each of the ten statistics, the paired shift of the forget identities between the released encoder and the same-seed retrain oracle, in units of the oracle's within-group spread, 100 identities, three seeds, five draws each; numbers are the single-statistic AUC of the released original. Forgotten identities remain more self-similar and less dispersed than never-seen people, a cluster the model once fitted, while their similarity to other people barely moves; the utility-preserving methods keep that shape and `artifact_aware` moves the negatives instead.

originals, and Section 7 shows utility-preserving unlearning leaves it in place. One labelled opt-out set from another tested model, identity split or dataset therefore suffices as supervision.

It survives a service that returns no embedding. A pairwise-score API reproduces the ten statistics exactly, since all are cosine similarities among images the adversary holds; we simulate two weaker services on the same draws and gallery, one returning the ten best-matching entries with scores, one returning only their order. With scores the released original reads 0.840 against the oracle's 0.623 under the mixed adversary (gap $+0.216$, 3/3 seeds) and 0.694 against 0.591 under the fresh one (3/3); with ranks alone, 0.737 against 0.623 (3/3) and 0.635 against 0.578 (2/3). Both ser-

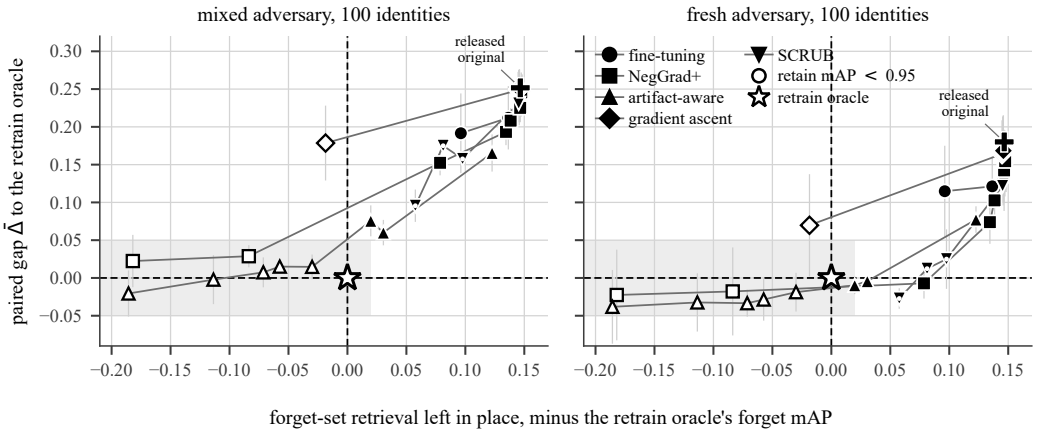


Figure 3: The utility, forgetting and detectability frontier. Each method's sweep, connected in order of forget-set retrieval; x is forget mAP minus the retrain oracle's (five-seed mean), y the paired gap to the same-seed oracle (the star), mean over surviving seeds, bars the seed spread of the gap. Filled markers keep retain mAP at or above 0.95, hollow ones do not. The shaded box holds configurations whose forget mAP is at most 0.02 above the oracle's and whose paired gap lies within ± 0.05; the target of Section 7 adds no rejection under the mixed adversary. No filled marker enters the mixed-adversary box; the one inside the fresh box, `artifact_aware` at $\beta = 10$, is rejected on the left.

vices preserve the mixed-adversary method ordering of Section 7; under the fresh adversary neither leaves a method-level gap significant (Appendix H).

It is a tighter cluster, not a displaced one. Figure 2 decomposes the signature statistic by statistic: the paired shift of the forget identities between released encoder and oracle, in units of the oracle's within-group spread, with each statistic's own separating power. Under the released original the forgotten identities' images are more similar to one another than a never-seen person's are, pairwise own-image similarity 1.6 standard deviations above the oracle (single-statistic AUC 0.89 where the oracle reads 0.52), their centroid less dispersed (-1.4) and their cross-camera similarity higher ($+1.3$), while their distance to other people barely moves (hardest negative -0.3). The geometry is consistent with memorisation residue: a cluster the model once fitted stays tighter than the oracle's after the person is gone. `finetune` and `neggrad_plus` leave that residue in place ($+1.7$ and $+1.3$ on pairwise similarity), consistent with their higher detectability; `artifact_aware`, whose loss scatters forget images toward other people, moves the negatives instead ($+2.1$ on the top-10 negative similarity) and leaves the cluster statistics near zero. Under the fresh adversary every shift halves and keeps its sign (Appendix H).

Nulling the single forgotten-versus-never-seen direction, Gaussian noise and PCA truncation leave the attack in place until retrieval is destroyed; the signature is distributed across the embedding, not concentrated where it is easiest to remove (Appendix I).

7 Utility-preserving unlearning retains the mark

We apply four published unlearning methods and one of our own to each 100-identity original at the operating points our sweep selects and score every released encoder under both adversaries. `finetune` and `grad_ascent` (Golatkar et al., 2020; Thudi et al., 2022), `neggrad_plus` and `scrub` (Kurmanji et al., 2023) are as published. `artifact_aware` is a diagnostic baseline of ours, related to the identity-shift and embedding-dispersion methods of Peng et al. (2025) and Zakharov (2026) but not one of them, so we define it: each step minimises retain cross-entropy plus $\beta \, \| f(x_f) - t \|^2$, where $f(x_f)$ is a forget image's embedding and t is drawn independently per image from a buffer of retain embeddings, so forgotten identities are scattered across the retain manifold rather than collapsed to a point (8 epochs, lr 0.05, $\beta = 20$; Appendix G lists every operating point). The forget loader holds the forget identities' training images only, never the withheld probes: the controller deletes what it holds, and the adversary's images were never held (an earlier run of these

Table 3: Market-1501, 100 forget identities, five seeds, held-out probes, corrected protocol (the forget loader holds the forget identities' training images only). AUC is the mean over seeds ($\pm$ s.d.); *gap* is the paired difference $\bar{\Delta}$ against the retrain oracle on the same seed and the same identity draws; *sig.* counts seeds on which the draw-level paired permutation test rejects at $\alpha = 0.05$, with $^-$ marking a rejection below the oracle. Encoders with retain mAP below 0.50 are collapsed and not scored.

	utility		mixed adversary			fresh adversary		
Model	forget	retain	AUC	gap	sig.	AUC	gap	sig.
original	0.971	0.9994	0.968 ± 0.008	$+0.252$	5/5	0.802 ± 0.020	$+0.180$	5/5
`finetune`	0.921	0.959	0.908 ± 0.066	$+0.192$	5/5	0.737 ± 0.083	$+0.115$	4/5
`neggrad_plus`	0.971	0.9992	0.942 ± 0.008	$+0.225$	5/5	0.764 ± 0.029	$+0.142$	5/5
`artifact_aware`	0.795	0.941	0.731 ± 0.029	$+0.015$	0/5	0.603 ± 0.063	-0.019	0/5
`grad_ascent`	0.806	0.889	0.895 ± 0.059	$+0.179$	5/5	0.691 ± 0.075	$+0.070$	4/5
`scrub`				collapsed	5/5			
retrain oracle	0.825	0.996	0.717 ± 0.019	–	–	0.622 ± 0.081	–	–

operating points had fed the withheld images to the methods; Appendix H reports both). A run whose encoder collapses, retain mAP below 0.50, is a dead model rather than a private one: we report it as collapsed and do not score it. We do not retune at 100 identities by design: the question is whether a tuned operating point survives a larger deletion request; Appendix G reports the 20-identity study on which the operating points were tuned.

Under the mixed adversary the survivors separate (Table 3). `neggrad_plus` holds retain mAP at 0.9992 and forget mAP at 0.971, and is detected at AUC 0.942, a paired gap of $+0.225$ over the same-seed oracle on 5/5 seeds. `finetune` holds retain 0.959 and forget 0.921, and is detected at 0.908, gap $+0.192$, 5/5. `grad_ascent`, which collapsed at these settings when the withheld images were in its loader, now survives on 5 of five seeds at retain 0.889 and leaks at 0.895 (gap $+0.179$, 5/5); `scrub` collapses on every seed. `artifact_aware` drives forget retrieval to 0.795, below the oracle's 0.825, at retain 0.941, and reads 0.731 with a gap of $+0.015$ that the paired test rejects on 0/5 seeds; under the fresh adversary it reads 0.603, gap -0.019, 0/5. Under the fresh adversary `finetune` reads 0.737, gap $+0.115$, 4/5, and `neggrad_plus` 0.764, gap $+0.142$, 5/5.

No operating point reaches the three-way target. A single operating point per method invites the objection that a better one exists. We therefore sweep 39 configurations across the five methods, each method's strength knobs on both sides of its operating point, three seeds each and five at each method's operating point, under the corrected protocol, and score every surviving configuration's utility and paired gap (Figure 3; the full table is Appendix H). The target is retain mAP at or above 0.95, forget mAP at most 0.02 above the oracle's 0.825, and a paired gap within ± 0.05 with no rejection under the mixed adversary, the more powerful one (Appendix H). No configuration meets it. Of the 9 configurations with forget mAP at most 0.02 above the oracle's, one keeps retain mAP above 0.95 (`artifact_aware` at $\beta = 10$, 0.956) and it still rejects on 2/3 seeds under the mixed adversary with gap $+0.076$ (under the fresh adversary its gap is -0.010, 0/3); the other eight fall below 0.95, and every configuration that keeps retain utility above 0.95 is detected under the mixed adversary, the least so at gap $+0.060$ (`artifact_aware`, 2/3 seeds). Along each method's knob the large drops in the paired gap arrive with drops in retain utility, and below the oracle's retrieval the gap goes negative: the encoder has been damaged past what retraining would have done.

Our pre-specified hypothesis is falsified by these rows. We named over-suppression as the tell before the sweep ran, the claim being that driving forget retrieval below the oracle is what makes a release legible.[2] The opposite holds: the mark is left by the methods that modify least, and it fades where retrieval has been paid for it; the same ordering appears on Tiny-ImageNet (Appendix D) and at 20 identities (Appendix G). Detectability tracks residual forget retrieval, an observation and not a threshold, and not a circular one: the oracle retrieves these identities at forget mAP 0.825 yet sits at TPR 0.064 at 1% FPR, so retrieval level and membership signal are separable.

[2]Pre-specified analysis plan, Stage-0 discussion and Rule 5, reproduced in the supplementary material.

8 OPERATING POINTS ARE SEED-SENSITIVE

Single-run reporting hides how unstable these methods are at a fixed setting, the published one where a paper gives it. We ran 216 unlearning runs on Market-1501, 27 method-by-learning-rate cells of eight run seeds, with the forget set, the original checkpoint, the retain budget and every hyperparameter fixed so that only batch order and method-internal sampling vary. Every method has a collapse cliff and a band before it in which the identical configuration both survives and destroys the encoder. SCRUB's published operating point, $\mathrm{lr} = 0.01$, sits at $7/8$ collapse, and its lone survivor has already fallen to retain mAP 0.8206. Appendix J gives the full map.

9 SCOPE AND WHAT WE DO NOT CLAIM

Our evidence is three academic benchmarks, Market-1501 (Zheng et al., 2015), CUHK03-NP (Li et al., 2014; Zhong et al., 2017) and Tiny-ImageNet (Le & Yang, 2015), and three ImageNet-pretrained backbones trained by one recipe; none is a face-retrieval corpus or a deployed system, and we claim nothing about operational identity search. The black-box services are simulated from the released embeddings on the paper's own gallery, not measured on a deployed index.

The detector is supervised; its label-free use rests on transfer with the retrain-oracle control (Section 6), which is strong within Market-1501 and across the three backbones and weaker across datasets, where a CUHK03-NP fit reads Market-1501 at 0.621 against 0.489. The fresh adversary, which holds no training image, is the stricter one and is underpowered by construction, about four images per identity, so its method-level verdicts are ordering evidence beside the mixed adversary's. The frontier sweep's detection test uses 500 replicates.

`artifact_aware`'s non-leak at its operating point (0/5 and 0/5) sits at retain mAP 0.941 and forget 0.795, below the oracle's 0.825 (Section 7). Applied instead to never-seen identities it moved the real forget identities' gap by -0.326 without seeing a forget image, and a control on the same checkpoint shows the effect is not global: twenty retain identities against never-seen fall only -0.094 ($p = 0.22$) while the forget identities and the ghosts fall together. Its non-leak is real, bought with retrieval, and not specific to the opt-out identities.

No multiple-comparison correction across method families, and no test of the oracle against the null floor. TPR at 1% FPR is a single order statistic over 100 negatives, spanning 0.468 to 0.720 per mixed-adversary seed, and TPR at 0.1% is unestimable at this identity count. The defences we break are post-hoc transforms of a released embedding; we test no training-time defence and propose none.

10 CONCLUSION

A fresh adversary, holding no training image, separates opted-out from never-seen identities at AUC 0.802 against a retrain oracle at 0.622 (paired gap $+0.180$, 5/5 seeds); the stronger mixed adversary, unknowingly holding training images, reaches 0.968 against 0.717 (Section 5). The gap reproduces on a ResNet-18, a ConvNeXt-T and CUHK03-NP; a detector fitted once reads disjoint identities, other architectures and the other dataset above what it reads on the oracle; the signature survives a ranks-only retrieval service; and the mark is consistent with memorisation residue, a cluster that stays tighter than a stranger's (Section 6). Across 39 operating points of five unlearning methods no configuration keeps retain mAP at or above 0.95, brings forget mAP to at most 0.02 above the oracle's and passes the mixed-adversary FII test; `neggrad_plus` and `finetune`, with the best conventional scorecards, leak on every seed under the mixed adversary (Section 7). The pre-specified over-suppression hypothesis is falsified: the largest gaps belong to the least-modifying methods.

Report detection as the paired gap to a same-seed retrain oracle on withheld probes under a stated adversary, with the oracle's forget-set retrieval beside forget and retain utility, a null floor at the reported n, and a seed-stability cell and collapse fraction at the published operating point. Name the gallery behind any retrieval ceiling (Appendix F).

None of the procedures tested here reaches the oracle's forget-set retrieval and closes the gap under both adversaries without paying retrieval for it.

ETHICS STATEMENT

We study a privacy attack on public academic person re-identification and image-classification benchmarks. We collect no new human-subject data, make no attempt to associate benchmark identity labels with real people, submit no real erasure request and query no deployed system. Market-1501 and CUHK03-NP are used for person retrieval; we perform no face recognition. The black-box services of Section 6 are simulations over our own galleries. The attack is dual-use: it can reveal whether an identity was represented in training and later removed. The detector is a logistic regression over retrieval statistics that any holder of a released encoder could rebuild, and we release the implementation for reproducibility and defensive auditing; it requires one labelled opt-out set, which limits casual misuse but not a determined adversary. We make no legal claim that FII is required by any deletion regime; we treat non-disclosure of deletion status as a privacy desideratum. We propose no defence, and name the methods that leak so they are not certified safe on utility alone.

REPRODUCIBILITY STATEMENT

The supplementary material holds the experiment scripts, configurations, seeds, every result JSON cited in the paper and the scripts that regenerate its tables and figures from them; dataset copies, checkpoints, the training and unlearning harness and the trained-model manifests are omitted from the anonymous submission and will be released with the camera-ready version. The appendix gives hyperparameters, the collapse rule and per-cell run counts. An independent re-implementation (`exp9-fresh-reference/independent_eval.py`), with its own matching, splits, RNG and scikit-learn's AUC and no import from the detector code, reproduces the first three-seed run of the 100-identity mixed-adversary numbers (the five-seed headline of Table 1 supersedes them), independent value first: original 0.969/0.971, oracle 0.711/0.719, null 0.466/0.481, covariate-only 0.433/0.397 (same side of chance), TPR at 5% FPR 0.853/0.866. TPR at 1% FPR reads 0.727 against 0.616, just above the pipeline's per-seed span of 0.556 to 0.720; at $n = 100$ that operating point is set by a single negative order statistic, so we treat it as unstable and report it with its spread rather than as a point. The pipeline's raw script folded the null and covariate AUCs as $\max(a, 1 - a)$ but not the models'; all reported values are the unfolded recomputation in `exp5-scale100/results/estimator_check.json`, which the independent evaluator confirms. The scaled-up study (five seeds, the frontier sweep, the ResNet-18, ConvNeXt-T and CUHK03-NP encoders, the black-box services and the mechanism decomposition) uses the same manifests; its draw construction is a single shared module, and the 20-identity mixed-adversary transfer it reproduces to four decimals (0.974) is the check that it matches the original pipeline.

USE OF LARGE LANGUAGE MODELS

Large language models were used in the manner of an external reviewer: to polish and aid the writing, to give reviewer-style critiques of successive drafts, and to argue against the paper's claims as a devil's advocate, so that weak claims could be tightened or withdrawn. They had no part in designing, running or analysing the experiments; all experiments, code, results and reported numbers are the authors' own. The authors reviewed every suggestion and take full responsibility for the paper.

REFERENCES

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

Xavier F. Cadet, Anastasia Borovykh, Mohammad Malekzadeh, Sara Ahmadi-Abhari, and Hamed Haddadi. Deep unlearn: Benchmarking machine unlearning for image classification. In *2025 IEEE 10th European Symposium on Security and Privacy (EuroS&P)*, pp. 939–962. IEEE, 2025. doi: 10.1109/EuroSP63326.2025.00058. arXiv:2410.01276.

Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramèr. Membership inference attacks from first principles. In *2022 IEEE Symposium on Security and Privacy (SP)*, pp. 1897–1914. IEEE, 2022. doi: 10.1109/SP46214.2022.9833649. arXiv:2112.03570.

Jingong Chen, Dongyoun Kim, Chulwoo Pack, Jun Huang, and Kwanghee Won. TIU-ReID: Target identity unlearning for person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 8506–8515, 2026a. Workshop on Machine Unlearning for Vision (MUV).

Min Chen, Zhikun Zhang, Tianhao Wang, Michael Backes, and Yang Zhang. FACE-AUDITOR: Data auditing in facial recognition systems. In *32nd USENIX Security Symposium (USENIX Security 23)*. USENIX Association, 2023.

Yiwei Chen, Soumyadeep Pal, Yimeng Zhang, Qing Qu, and Sijia Liu. Unlearning isn't invisible: Detecting unlearning traces in LLMs from model outputs. In *International Conference on Learning Representations (ICLR)*, 2026b. arXiv:2506.14003.

Dasol Choi and Dongbin Na. Towards machine unlearning benchmarks: Forgetting the personal identities in facial recognition systems. *arXiv preprint arXiv:2311.02240*, 2023. AAAI 2024 Workshop on Privacy-Preserving Artificial Intelligence (PPAI).

European Parliament and Council of the European Union. Regulation (EU) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation). Official Journal of the European Union L 119, pp. 1–88, 2016. Article 17: right to erasure.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874, 2008.

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. arXiv:1911.04933.

Jamie Hayes, Ilia Shumailov, Eleni Triantafillou, Amr Khalifa, and Nicolas Papernot. Inexact unlearning needs more careful evaluations to avoid a false sense of privacy. *arXiv preprint arXiv:2403.01218*, 2024.

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.

Dongjae Jeon, Wonje Jeung, Taeheon Kim, Albert No, and Jonghyun Choi. An information theoretic evaluation metric for strong unlearning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 22173–22181, 2026. doi: 10.1609/aaai.v40i26.39373.

Tirth Joshi and Honggang Wang. Unlearning is not deletion: Auditing residual information in released vision artifacts. In *ECCV 2026 Workshop on Unlearning and Model Editing (U&ME)*, 2026a. To appear.

Tirth Joshi and Honggang Wang. Compositional non-face re-identification pressure under cumulative vision releases. In *Computer Vision – ECCV 2026, Part VIII*, Lecture Notes in Computer Science, pp. 150–167, Cham, 2026b. Springer. doi: 10.1007/978-3-032-37026-6_9.

Meghdad Kurmanji, Peter Triantafillou, Jamie Hayes, and Eleni Triantafillou. Towards unbounded machine unlearning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023. arXiv:2302.09880.

Ya Le and Xuan Yang. Tiny ImageNet visual recognition challenge. CS 231n course report, Stanford University, 2015. URL `https://cs231n.stanford.edu/reports/2015/pdfs/yle_project.pdf`.

Guoyao Li, Shahbaz Rezaei, and Xin Liu. User-level membership inference attack against metric embedding learning. *arXiv preprint arXiv:2203.02077*, 2022. ICLR 2022 Workshop on Privacy, Accountability, Interpretability, Robustness, Reasoning on Structured Data (PAIR^{2}Struct).

Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. DeepReID: Deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 152–159, 2014. doi: 10.1109/CVPR.2014.27.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A ConvNet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11976–11986, 2022. doi: 10.1109/CVPR52688.2022.01167.

Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of tricks and a strong baseline for deep person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2019.

Markus Ojala and Gemma C. Garriga. Permutation tests for studying classifier performance. *Journal of Machine Learning Research*, 11(62):1833–1863, 2010.

Ünsal Öztürk, Vedrana Krivokuća Hahn, Sushil Bhattacharjee, and Sébastien Marcel. Damnatio memoriae: Adversarially and selectively forgetting identities in the embedding space of face recognition models. *arXiv preprint arXiv:2609.27115*, 2026.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12: 2825–2830, 2011.

Yi-Xing Peng, Yu-Ming Tang, Kun-Yu Lin, Qize Yang, Jingke Meng, Xihan Wei, and Wei-Shi Zheng. Person de-reidentification: A variation-guided identity shift modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29331–29341, 2025.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, pp. 3–18. IEEE, 2017. doi: 10.1109/SP.2017.41.

Christoforos N. Spartalis, Theodoros Semertzidis, Efstratios Gavves, and Petros Daras. LoTUS: Large-scale machine unlearning with a taste of uncertainty. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10046–10055, 2025.

Anvith Thudi, Gabriel Deza, Varun Chandrasekaran, and Nicolas Papernot. Unrolling SGD: Understanding factors influencing machine unlearning. In *2022 IEEE 7th European Symposium on Security and Privacy (EuroS&P)*, pp. 303–319. IEEE, 2022. arXiv:2109.13398.

Eleni Triantafillou, Peter Kairouz, Fabian Pedregosa, Jamie Hayes, Meghdad Kurmanji, Kairan Zhao, Vincent Dumoulin, Julio Jacques Junior, Ioannis Mitliagkas, Jun Wan, Lisheng Sun Hosoya, Sergio Escalera, Gintare Karolina Dziugaite, Peter Triantafillou, and Isabelle Guyon. Are we making progress in unlearning? findings from the first NeurIPS unlearning competition. *arXiv preprint arXiv:2406.09073*, 2024.

Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st Computer Security Foundations Symposium (CSF)*, pp. 268–282. IEEE, 2018. doi: 10.1109/CSF.2018.00027.

Mikhail Zakharov. Face identity unlearning for retrieval via embedding dispersion. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*, pp. 1277–1286, March 2026.

Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 1116–1124, 2015. doi: 10.1109/ICCV.2015.133.

Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li. Re-ranking person re-identification with k-reciprocal encoding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1318–1327, 2017. doi: 10.1109/CVPR.2017.389.

A EXPERIMENTAL SETUP

Table 4 fixes both pipelines. Tiny-ImageNet is chosen over a second person-re-identification benchmark because every class has exactly 500 training images and there are no cameras, so the matched-population problem cannot arise and a covariate-only detector has nothing to read. It replicates the finding at the level of classes rather than people and is not evidence of generalisation across identities.

Table 4: Experimental setup. Every released model trains for 60 epochs, and a 30% per-identity image holdout supplies the adversary's probes under both adversaries of Section 3. Per seed we train an original on retain + forget and a retrain oracle on retain only; the five unlearning methods are applied to the original at both Market-1501 forget-set sizes and on Tiny-ImageNet.

	Market-1501 (Zheng et al., 2015)	Tiny-ImageNet (Le & Yang, 2015)
Encoder	ResNet-50 + BNNeck, 2048-d (Luo et al., 2019)	ResNet-18, 512-d penultimate features
Task	open-set person retrieval, 12,936-image gallery	200-way classification, retrieval on features
Identities	751 train: 20 forget / 731 retain (100 / 651 at the scaled setting)	200 classes: 20 forget / 130 retain / 50 never-seen
Never-seen pool	test-split identities, untrained on by every model	the 50 held-out classes
Probe set	118 / 119 / 93 withheld images per seed (484 / 457 / 498 at 100 identities)	3,000 withheld images per seed

B MATCHED NEGATIVES AND THE COVARIATE CONTROL

Matched negatives. Each forget identity is paired with a never-seen identity ranked by shortfall in image and camera count, one penalty per missing image and three per missing camera, the five best-ranked candidates taken and those tied at minimal shortfall drawn from at random, then subsampled to the target's image and camera counts. A fixed bank of 300 identities per split supplies retrieval distractors, and the test-split bank is excluded from the negative pool. Never-seen identities are drawn from the Market-1501 test split, so *never seen* means unseen by every model in this study rather than sampled from a separate corpus. Under the fresh adversary the matched never-seen identity also receives the same per-identity image budget as its forget partner, a median of about four images split evenly between probe and reference.

The covariate control, and its sign. On Market-1501 a detector given per-identity bookkeeping alone, image count, camera count, images per camera and the maximum, minimum and standard deviation of the per-camera counts, never touching an embedding, reaches 0.397 ± 0.057 against a never-seen-versus-never-seen null floor of 0.481 ± 0.059, which is chance. Both sit below 0.5, and the sign is the whole of the matter. Bookkeeping does separate the two groups, a displacement of $|0.397 - 0.5| = 0.10$ over 15 draws is not noise, but it separates them in the direction opposite to membership, so a detector reading only metadata would systematically accuse the wrong group. The covariate confound therefore cannot account for any part of the observed detection: it pushes against the attack rather than supplying it, and matched-population bookkeeping makes the attack harder, not easier. A criterion defined as distance from 0.5 would read this below-chance metadata detector as a strong one.

Why the comparison is paired. Raw AUC is not a stable yardstick across protocols. Under the conventional protocol, with probes drawn from training images, the retrain oracle's own AUC is 0.398 ± 0.069, below chance, against 0.565 ± 0.068 under held-out probes at the same forget-set size. A same-seed oracle absorbs both effects: whatever the probe protocol and the negative population do to the estimator, they do to both encoders, so the paired gap Δ of Equation 1 and not the raw AUC is what this paper reports.

C Detector definitions, split asymmetry and significance

C.1 The ten per-identity statistics

Let an identity's probe half be P and its reference half R, all embeddings unit-norm, similarity the inner product. Three statistics describe probe-to-own-reference agreement: `pos_mean` and `pos_max`, the mean and maximum similarity over $P \times R$, and `xcam_pos_mean`, the mean over those pairs whose two images come from different cameras. Four describe the identity against the gallery, which holds every scored identity's reference half plus 300 distractor identities drawn from the retain split and 300 from the test split, with the identity's own reference entries masked out of the negatives: `hardneg`, the mean over probes of the largest negative similarity; `top1` and `top10_mean`, the mean over probes of the largest and of the mean of the ten largest gallery similarities; and `negtop10_mean`, the same restricted to negatives. One is the retrieval margin, $\texttt{margin} = \texttt{pos_max} - \texttt{hardneg}$. Two describe the identity's own manifold over all of its embeddings: $\texttt{centroid_dispersion} = 1 - \|\bar{e}\|$ for the mean $\bar{e}$ of unit-norm embeddings e, and `pair_sim_mean`, the mean similarity over all within-identity pairs. The gallery is drawn once per seed under a fixed generator and held constant across the five matched draws. Nothing in the battery is specific to person re-identification, and the same ten are computed unchanged on Tiny-ImageNet, where the cross-camera statistic degenerates because the dataset has no camera labels.

C.2 Split asymmetry under the mixed adversary

Under the mixed adversary the two arms are probed differently. A forget identity is scored on a $30/70$ split whose probe half is the withheld set, drawn uniformly per identity without regard to camera; a never-seen identity, which has no withheld set, is split $50/50$ by permuting, stably sorting by camera and taking alternate elements, so its split is camera-stratified and the forget identity's is not. We do not correct this, because the correction is the oracle. The same-seed retrain oracle is scored on the same identities under the same splits and the same features, so the protocol asymmetry enters both arms of Equation 1 identically and cancels in Δ. It is the paired gap, not the raw AUC, that is attributable to the released encoder. The fresh adversary removes the asymmetry at its source: probe and reference halves are both drawn from the withheld set and never-seen identities receive the same per-identity image budget, so the two groups are compared on identical evidence, at the cost of roughly four images per identity instead of seventeen.

C.3 Significance and matching

Significance in the 20-vs-20 sweep is a two-sided label-permutation test with 500 permutations and 2000-sample bootstrap intervals over identities (Ojala & Garriga, 2010). Significance in every comparison that produces a main table is the paired permutation on Δ described in Section 4: one label shuffle applied to both encoders and to all five draws of a seed, the five null gaps averaged, 2,000 replicates. A significant entry therefore means *different from the same-seed retrain oracle*, never *above chance*. The scripts that produced the first draft of every table built the null differently: 300 shuffles within each draw, the 1,500 single-draw null gaps pooled and the five-draw mean compared against that pool. A single-draw gap is more variable than a five-draw mean, so that null was wider than the statistic's and the test conservative. Table 5 recomputes both from the identical feature matrices: every rejection under the pooled null survives the draw-level null, and the entries it failed to reject that the draw-level test rejects are `artifact_aware` mixed at 100 (0/3 to 1/3), `finetune` fresh at 100 (1/3 to 3/3), `artifact_aware` fresh at 100 (0/3 to 1/3), `finetune` fresh at 20 (0/3 to 3/3), `neggrad_plus` fresh at 20 (0/2 to 1/2), `artifact_aware` fresh at 20 (0/3 to 3/3), `grad_ascent` fresh at 20 (2/3 to 3/3), `scrub` fresh at 20 (0/1 to 1/1). All Market-1501 tables report the draw-level test. Tiny-ImageNet (Table 6) retains the pooled test, whose rejections stand for the same reason and whose one non-rejection, `grad_ascent` at gap -0.025, is not a claim the paper makes. At $n = 100$ no bootstrap is run; dispersion is the standard deviation over the five draws.

Negatives are never-seen identities chosen greedily to match each positive on image count and camera count and then subsampled to it. The matching is one-sided: a negative is cut down to its positive but never padded up, so residual imbalance survives by construction, and the covariate detector of Appendix B measures which way it cuts.

Table 5: Permutation audit. For every scored Market-1501 entry: the observed paired gap $\bar{\Delta}$, the p-value under the within-draw pooled null the scripts first used, and under the draw-level null the paper reports. Both are computed from the same feature matrices; 2,000 replicates, so the smallest attainable p is 0.0005. Collapsed encoders are not scored.

		seed 0			seed 1			seed 2		
adversary	model	gap	p pooled	p draw	gap	p pooled	p draw	gap	p pooled	p draw
100 forget identities										
mixed	original	+0.217	0.0004	0.0005	+0.260	0.0001	0.0005	+0.279	0.0001	0.0005
mixed	finetune	+0.173	0.0021	0.0005	+0.242	0.0003	0.0005	+0.135	0.0146	0.0005
mixed	neggrad_plus	+0.195	0.0006	0.0005	+0.234	0.0005	0.0005	+0.233	0.0002	0.0005
mixed	artifact_aware	+0.005	0.9073	0.8596	+0.032	0.4711	0.2874	+0.073	0.1057	0.0120
fresh	original	+0.130	0.0126	0.0005	+0.147	0.0062	0.0005	+0.234	0.0001	0.0005
fresh	finetune	+0.080	0.1012	0.0060	+0.121	0.0191	0.0010	+0.065	0.2362	0.0425
fresh	neggrad_plus	+0.044	0.3297	0.1154	+0.056	0.2456	0.0515	+0.105	0.0304	0.0015
fresh	artifact_aware	−0.058	0.2455	0.0465	−0.023	0.6424	0.4003	+0.005	0.9209	0.8526
20 forget identities										
mixed	original	+0.382	0.0033	0.0005	+0.372	0.0033	0.0005	+0.292	0.0082	0.0005
mixed	finetune	+0.296	0.0109	0.0005	+0.281	0.0120	0.0010	+0.254	0.0192	0.0010
mixed	neggrad_plus	+0.088	0.5022	0.3188	+0.105	0.3670	0.1749		collapsed	
mixed	artifact_aware	+0.093	0.3273	0.1364	+0.048	0.6111	0.4803	+0.079	0.4048	0.2044
mixed	grad_ascent	+0.316	0.0109	0.0005	+0.344	0.0111	0.0005	+0.217	0.0477	0.0070
mixed	scrub		collapsed			collapsed		+0.121	0.2059	0.0675
fresh	original	+0.243	0.0278	0.0005	+0.245	0.0351	0.0020	+0.206	0.0362	0.0005
fresh	finetune	+0.152	0.1230	0.0070	+0.137	0.1898	0.0360	+0.180	0.0761	0.0030
fresh	neggrad_plus	+0.089	0.4922	0.2549	+0.148	0.2179	0.0415		collapsed	
fresh	artifact_aware	+0.155	0.1715	0.0160	+0.163	0.1789	0.0320	+0.132	0.2596	0.0440
fresh	grad_ascent	+0.243	0.0345	0.0005	+0.196	0.1066	0.0130	+0.239	0.0288	0.0005
fresh	scrub		collapsed			collapsed		+0.181	0.0872	0.0050

One correction. The raw scoring script folded the null and covariate AUCs as $\max(a, 1-a)$ while leaving model AUCs unfolded; all quantities were recomputed unfolded on identical draws. Reported values are unfolded throughout: the null moved from 0.542 to 0.481 and the covariate detector from 0.603 to 0.397; model and oracle values are unchanged.

Power. At 20 against 20 the permutation null's 97.5th percentile lands between 0.685 and 0.725, so non-significance there bounds effects above about AUC 0.70. Operating points are stated only at $n = 100$, where the FPR grid is $1/100$.

D TINY-IMAGENET REPLICATION

Tiny-ImageNet is the replication without Market-1501's count and camera covariates of Section 5: a class-level proxy, not identity data, and not evidence that the attack generalizes across identities. Each seed draws 200 classes, 20 forget, 130 retain and 50 never-seen, and trains a ResNet-18 for 60 epochs to roughly 86% accuracy, with about 3,000 held-out images per seed reserved as probes. Because every class has exactly 500 training images and the dataset carries no camera metadata, the matched-population problem that Market-1501 poses cannot arise: the never-seen controls need no covariate matching, and the feature battery drops its camera statistic outright.

The released original separates forgotten classes from never-seen ones (Figure 5) at AUC 0.9998 ± 0.0003, per seed $1.000/1.000/0.9995$, with paired gap 0.4808 and permutation $p = 0.0080/0.0020/0.0007$, and reaches TPR 1.00 at 5% FPR on all three seeds. The retrain oracle reads 0.519 ± 0.136, per seed $0.618/0.5755/0.3635$, which is consistent with chance at $n = 3$; its permutation p of 1.0 is arithmetic rather than evidence, since the oracle is the reference against which gaps are measured and its own gap is exactly zero, so that row supports no significance claim in either direction. The wide seed spread makes this a weak control, and we report it as one.

Two consequences follow. The signal is not a property of person re-identification, of metric learning with a BNNeck head, or of Market-1501's camera structure. And the oracle's return to chance once

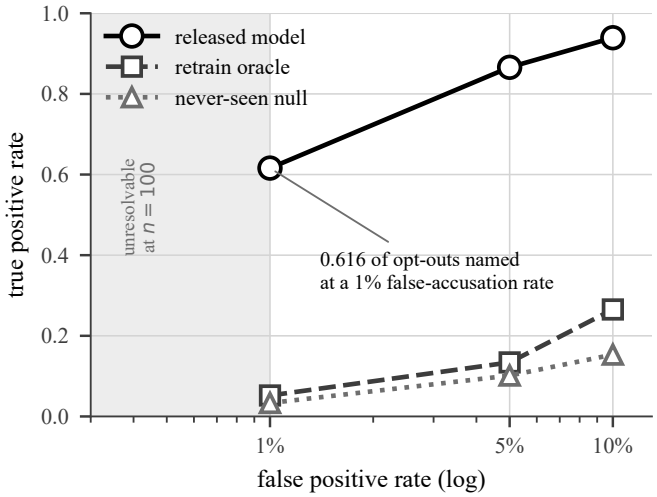


Figure 4: Detection power at 100 opt-out identities against 100 matched never-seen identities under the mixed adversary, held-out probes, mean of three seeds. Markers are the three false-positive rates the design resolves; with 100 negatives nothing below 1% is measurable and that region is left blank. The fresh adversary's operating points are given in Section 5.

the covariate is removed by construction is what a membership-only signal predicts, which is why the calibrated original-versus-oracle comparison, rather than either raw level, is the quantity reported on Market-1501.

D.1 Unlearning methods on Tiny-ImageNet

Table 6: Tiny-ImageNet, held-out probes, 20 forget classes, three seeds. The oracle's permutation p is 1.0 by construction, since it is its own reference, so its row reads as consistent with chance at $n = 3$ rather than as a test. `scrub` emits non-finite activations on all three seeds. Significance here uses the within-draw pooled null, which is conservative (Appendix C).

Model	forget mAP	retain mAP	AUC	gap	sig.
original	0.4647	0.9180	0.9998 ± 0.0003	$+0.481$	3/3
`finetune`	0.4413	0.9122	0.9998	$+0.481$	3/3
`neggrad_plus`	0.1269	0.6823	0.880 ± 0.025	$+0.361$	2/3
`artifact_aware`	0.0567	0.2263	0.810 ± 0.132	$+0.291$	2/3
`grad_ascent`	0.0215	0.0301	0.494 ± 0.029	-0.025	0/3
`scrub`	n/a	non-finite	not scored		
retrain oracle	0.0978	0.9077	0.519 ± 0.136	–	–

The under-modification result reproduces exactly (Table 6). `finetune` moves forget mAP from 0.4647 to 0.4413, holds retain mAP at 0.9122, and is detected at 0.9998 on every seed. The ordering does not reproduce, and cannot be tested here. The Market-1501 operating points do not transfer: `scrub` diverges on all three seeds, `grad_ascent` lands at retain mAP 0.0301 and `artifact_aware` at 0.2263, so the only utility-preserving rows are the original and `finetune`. `neggrad_plus` does reach the oracle's forget level (0.1269 against 0.0978) and is still detected at 0.880 on $2/3$ seeds, but at retain mAP 0.6823 that is not a comparison between released models a controller would ship. Confirming the ordering here would have required a run holding retain mAP above 0.90 at a forget mAP at or below the oracle's 0.0978, and our record contains none. The reading that suppression below the oracle's retrieval level accompanies the disappearance of the mark is therefore untested on Tiny-ImageNet, not confirmed by it.

D.2 THE DELETION CEILING

Tiny-ImageNet separates the ceiling from detectability. As a class-level replication it has no cameras and 500 images per class, so the covariate detector has nothing to read there. Its retrain oracle strips forgotten classes to mAP 0.0978 from the original's 0.4647, where on Market it removes almost nothing ($0.9887 \rightarrow 0.8738$, Appendix F): headroom 0.367 against 0.126, roughly three times the room to forget. The two mAPs come from different retrieval protocols and should not be divided into a ratio, but within each dataset the comparison is like for like and the direction is not in doubt. Detection of the original is 0.9998 ± 0.0003 either way ($3/3$ seeds significant). A cramped ceiling is therefore not what makes opting out visible, and the ceiling itself is a property of metric-learned person embeddings rather than of open-set retrieval in general. What this does not settle is whether the oracle marks the detectability threshold on Tiny-ImageNet, because there every method that reduced forget retrieval also damaged retain utility, to 0.68 or below against the original's 0.918.

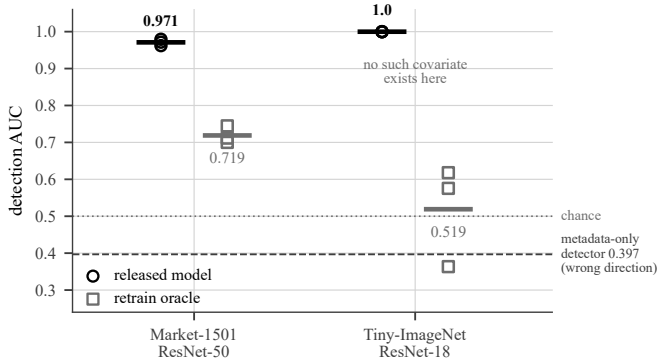


Figure 5: The same protocol on Tiny-ImageNet, a class-level replication in which every class has identical image counts and no camera metadata, so the population covariate of Market-1501 cannot be constructed. Detection is unaffected and the oracle returns to chance. The dashed rule is a detector given per-identity bookkeeping and no embedding: it sits below chance, separating the groups in the direction opposite to membership.

E CROSS-SEED TRANSFER AND THE NULLING DEFENCE

The nulling study runs at 20 forget identities under the mixed adversary on Market-1501 (Figure 1, Table 7); transfer at three seeds is in Table 8; the five-seed, cross-architecture and cross-dataset transfers are in Appendix H.3.

F DETECTABILITY IS NOT AN ARTIFACT OF A CRAMPED FORGETTING CEILING

The Market-1501 retrain oracle still retrieves forgotten identities at forget mAP 0.8738, leaving headroom, original minus oracle, of 0.1262 ± 0.0235 against the standard 12,936-image gallery, with per-seed values 0.106, 0.152 and 0.120. That headroom is a property of the gallery, not of unlearning. Padding the gallery with test-split images, whose identities are disjoint from training and therefore add distractors but no true positives, raises headroom to 0.1664, 0.1997, 0.2333 and 0.2645 ± 0.0418 while the oracle's forget mAP falls from 0.8738 to 0.6828 (Table 9). The rise is monotone in each of the three seeds separately and has not begun to plateau at 29,870 images, $2.3\times$ the standard gallery and the largest the test split permits. A ceiling quoted without its gallery is not a measurement, and the single value 0.106 that earlier drafts built on is the low end of the seed band rather than its centre.

Tiny-ImageNet, where the oracle leaves roughly three times the headroom and detection of the original is unchanged, is in Appendix D.

Table 7: The signal is identity-generic and distributed. **Upper:** detector fitted on the row seed's forget identities, five matched draws pooled, evaluated on the column seed's disjoint forget identities with no refitting. **Lower:** nulling the single direction separating forgotten from never-seen centroids in the released original embedding, with the publisher granted the true partition. Mean rows are computed across the three seeds.

Cross-seed transfer (AUC), fit → evaluated			
Fit on	eval s0	eval s1	eval s2
s0	–	0.9708	0.9855
s1	0.9717	–	0.9820
s2	0.9733	0.9609	–
mean over six ordered pairs: **0.974**			

	Detection AUC		Forget mAP		Retain mAP	
Seed	before	nulled	before	nulled	before	nulled
s0	0.9035	0.9030	0.9957	0.9958	0.9997	0.9997
s1	0.9025	0.8955	0.9715	0.9713	0.9999	0.9999
s2	0.9360	0.9255	0.9990	0.9979	0.9988	0.9986
mean	0.9140	0.9080	0.9887	0.9883	0.9995	0.9994

G THE TWENTY-IDENTITY STUDY

The operating points used throughout the paper were selected on a 63-configuration sweep at 20 forget identities per seed, and the 20-identity experiment is reported here in full because it is the design in which the main text's operating points were tuned and the design whose power limits motivate the move to 100 identities. All rows use held-out probes and the mixed adversary of Section 3. Under the fresh adversary at this scale the original reads AUC 0.756 ± 0.014 against an oracle at 0.525 ± 0.032, a paired gap of $+0.231$ significant on $3/3$ seeds. The methods under the same adversary, on their non-collapsed seeds, read `grad_ascent` 0.751 (gap $+0.226$, $3/3$), `finetune` 0.681 (gap $+0.156$, $3/3$), `artifact_aware` 0.675 (gap $+0.150$, $3/3$), `neggrad_plus` 0.627 (gap $+0.119$, $1/2$) on its two surviving seeds and `scrub` 0.739 (gap $+0.181$, $1/1$) on its one surviving seed: every surviving method is separated from its own oracle at this scale, at about four images per identity, and the method-level comparison is made at 100 identities, where the utility of the same encoders is measured on five times the forget images.

Operating points. Selected on the 20-identity sweep and held fixed at 100, each epoch drawing the same 3,840-image retain budget: `finetune` 2 epochs, lr 0.07; `grad_ascent` 2 epochs, lr 0.001; `neggrad_plus` 2 epochs, lr 0.005, $\alpha = 0.9$; `scrub` 5 epochs, lr 0.01, three maximisation steps per epoch; `artifact_aware` 8 epochs, lr 0.05, $\beta = 20$.

Thin rows and how they arose. Three cells were discarded by the collapse rule of Section 8: `scrub` on seeds 0 and 1 (retain mAP 0.0029 and 0.0027) and `neggrad_plus` on seed 2 (0.0328). `scrub` therefore contributes one seed here and `neggrad_plus` two, which is why the main text rebases on the 100-identity experiment, where every surviving method contributes five seeds. Survivors are not random draws: they are selected for having trained least destructively, which biases them toward higher retrieval and, by this paper's own thesis, toward more leakage. Because the gap is paired per seed it is not the difference of the AUC column from the oracle row: `scrub` survives only on seed 2, where the oracle reads 0.6435 rather than the three-seed mean.

The power bound at this scale. At 20 versus 20 identities the permutation null on detector AUC reaches 0.685–0.711 at its 97.5th percentile across draws, and 0.725–0.733 once best-feature selection is priced in (supplementary material). Every $0/n$ row in Table 10 therefore bounds the effect at about AUC 0.70 rather than excluding it. Two of the three non-rejecting methods have point es-

Table 8: Cross-seed transfer with the retrain-oracle control. The detector is fitted on the first seed's forget identities, all five draws pooled, and applied unchanged to the second seed's disjoint identities, once on that seed's released encoder and once on its retrain oracle; AUC is averaged over the second seed's five draws. The gap is what the transferred detector reads beyond the oracle. Under the fresh adversary neither seed's adversary holds a training image.

fit → eval	released	oracle	gap
fresh adversary, 100 identities			
s0 → s1	0.840	0.636	+0.204
s0 → s2	0.748	0.549	+0.199
s1 → s0	0.833	0.683	+0.149
s1 → s2	0.750	0.551	+0.199
s2 → s0	0.789	0.578	+0.211
s2 → s1	0.774	0.543	+0.231
mean	0.789	0.590	+0.199
fresh adversary, 20 identities			
s0 → s1	0.821	0.524	+0.296
s0 → s2	0.876	0.674	+0.202
s1 → s0	0.886	0.666	+0.220
s1 → s2	0.912	0.711	+0.201
s2 → s0	0.881	0.656	+0.225
s2 → s1	0.860	0.625	+0.235
mean	0.873	0.643	+0.230
mixed adversary, 100 identities			
s0 → s1	0.992	0.581	+0.411
s0 → s2	0.996	0.589	+0.406
s1 → s0	0.989	0.609	+0.381
s1 → s2	0.995	0.581	+0.414
s2 → s0	0.985	0.598	+0.388
s2 → s1	0.988	0.565	+0.422
mean	0.991	0.587	+0.404

Table 9: Deletion headroom against gallery size on Market-1501. Headroom is original forget mAP minus retrain-oracle forget mAP under an identical retrieval protocol, with the extra gallery entries drawn from the identity-disjoint test split so that they act as pure distractors. Three seeds.

Gallery size	Headroom	Oracle forget mAP
12,936 (standard)	0.1262 ± 0.0235	0.8738
14,936	0.1664 ± 0.0332	0.8292
17,936	0.1997 ± 0.0422	0.7875
22,936	0.2333 ± 0.0342	0.7382
29,870	$\mathbf{0.2645 \pm 0.0418}$	0.6828

timates below that bound; `scrub`'s single surviving seed at 0.765 does not. This is the reason the ordering claim is made at 100 identities and not here.

The ordering, and where it is not monotone. Across the seven models of Table 10 detection tracks residual forget retrieval at Spearman $\rho = 0.857$ ($p = 0.014$), with the oracle on the curve. The relation is not monotone. `artifact_aware` has the lowest forget mAP in the table and a higher AUC than `neggrad_plus`, and `scrub` (0.9406) and `neggrad_plus` (0.8800) both retain more of the forget set than the oracle does (0.8738) while failing to reject. A rule of the form "above the oracle leaks, at or below it does not" is contradicted by two of the six method rows, which is why the main text states the relation as an observation and reports the paired gap rather than a retrieval threshold.

Table 10: Market-1501, held-out probes, mixed adversary, 20 forget identities, collapse-filtered. AUC is the mean over valid seeds ($\pm$ s.d. where $n > 1$); *gap* is the paired difference against the retrain oracle on the same seed and the same identity draws, and *sig.* counts seeds on which the paired permutation test rejects at $\alpha = 0.05$.

Model	n	forget mAP	retain mAP	AUC	gap	sig.
original	3	0.9887	0.9995	0.9140 ± 0.0191	$+0.349$	3/3
`grad_ascent`	3	0.9634	0.9986	0.8575 ± 0.0182	$+0.292$	3/3
`finetune`	3	0.9856	0.9977	0.8422 ± 0.0476	$+0.277$	3/3
`scrub`	1	0.9406	0.9915	0.7650	$+0.122$	0/1
`neggrad_plus`	2	0.8800	0.9672	0.6225 ± 0.0177	$+0.097$	0/2
`artifact_aware`	3	0.7074	0.9430	0.6383 ± 0.0748	$+0.073$	0/3
retrain oracle	3	0.8738	0.9994	0.5652 ± 0.0680	–	–

Scale dependence of the collapse. `grad_ascent` is the healthiest method in Table 10 at retain mAP 0.9986. Rerun at 100 forget identities with identical hyperparameters and every forget image in its loader, the withheld probes included, it destroyed the encoder on all three seeds (forget mAP 0.057, retain mAP 0.061); with the loader restricted to the training images (Appendix H.1) it survives on 5 of five seeds at retain 0.889 and leaks. `scrub` collapses at 100 identities under both loaders. The stability cliff documented in Section 8 is therefore scale-dependent, loader-dependent and seed-dependent: a configuration that survives the run seed at one forget-set size is not thereby a configuration at another.

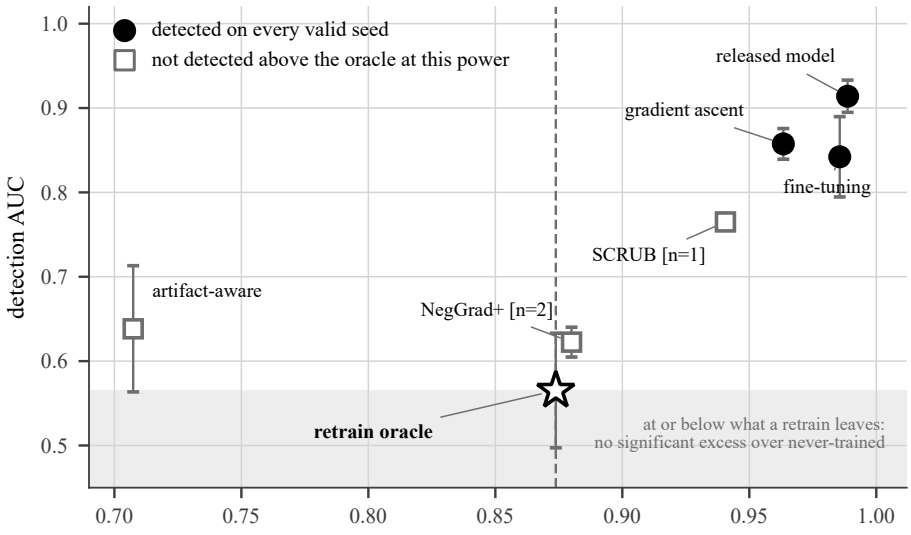


Figure 6: Detection against how much forget-set retrieval each method leaves in place, 20 identities, mixed adversary, with the retrain oracle marked on both axes. Filled circles are detected on every valid seed; open squares are not detected above the oracle at this power. Runs whose encoder collapsed are excluded. The 100-identity table and frontier in the text are the primary evidence; this figure shows the same ordering at the smaller scale.

H THE SCALED-UP STUDY

Everything in this appendix reuses the Market-1501 manifests of the main study: forget draws and 30% holdouts are seed-derived, so identities, probe sets and matched draws are identical across protocols and backbones. Detection uses the draw-level permutation test throughout (2,000 replicates for the settings tables, 500 for the frontier sweep and the service simulations).

H.1 Protocol correction and additional seeds

The first run of the operating points built the forget loader from every image of the forget identities, the withheld probes included, so a method with a forget loss optimised directly on the images the fresh adversary later probed with (`finetune`, which touches only the retain set, and the original and oracle rows were unaffected). Table 11 places the two runs side by side; the corrected run also adds seeds 3 and 4. The direction of the flaw is conservative for a leak claim, since a method pushed the probe images away, and the corrected verdicts are the same in kind: the two utility-preserving methods leak on every seed under the mixed adversary, and `artifact_aware` passes the FII test at a retain cost. `grad_ascent`, which had collapsed on every seed with the larger forget loader, survives on 5 of five seeds with it removed and leaks.

Table 11: Operating points under the two forget-loader protocols, 100 identities. Left: withheld probes inside the forget loader, three seeds, as first run. Right: training images only, five seeds. Utility is forget / retain mAP; gap and significant seeds under each adversary.

	probes in loader (3 seeds)						training images only (5 seeds)					
	forget	retain	mixed gap	sig.	fresh gap	sig.	forget	retain	mixed gap	sig.	fresh gap	sig.
`finetune`	0.908	0.958	+0.183	3/3	+0.088	3/3	0.921	0.959	+0.192	5/5	+0.115	4/5
`neggrad_plus`	0.949	0.9992	+0.221	3/3	+0.068	1/3	0.971	0.999	+0.225	5/5	+0.142	5/5
`artifact_aware`	0.789	0.937	+0.037	1/3	−0.025	1/3⁻	0.795	0.941	+0.015	0/5	−0.019	0/5
`grad_ascent`	0.057	0.061	collapsed 3/3				0.806	0.889	+0.179	5/5	+0.070	4/5

H.2 Other encoders and CUHK03-NP

ResNet-18 (512-d) (He et al., 2016) and ConvNeXt-T (768-d) (Liu et al., 2022) replace the ResNet-50 stem in the same BNNeck recipe (last-stride trick for the ResNets, none for ConvNeXt), ImageNet-pretrained, 60 epochs, same optimiser and losses; CUHK03-NP (detected boxes, new protocol, Market-1501 layout) is trained with the ResNet-50 recipe unchanged on 767 identities of about 10 images from two cameras, 100 forgotten, 30% withheld. Table 2 in the main text gives the released-original rows. Applying the Market-1501 ResNet-50 operating points unchanged to the other encoders is a test of the operating points rather than of the methods: on ConvNeXt-T every method but `neggrad_plus` collapses at those learning rates, on ResNet-18 `artifact_aware` and `scrub` collapse, on CUHK03-NP only `scrub` does (Table 12), which is the scale- and recipe-dependence of Section 8 seen from another side.

H.3 Transfer with the retrain-oracle control

Table 13 lists every transfer setting. Same-identity pairs share the forget identities and change only the encoder; disjoint pairs change identities as well; the cross-dataset rows pool all nine ordered seed pairs.

H.4 Black-box services

Each probe is submitted to a simulated service over the same gallery the white-box detector used (the distractor bank plus every scored identity's reference half); the service returns the $k = 10$ best entries with scores, or their order only, and the adversary knows which entries are its own reference images. Eight statistics are computed per probe and averaged per identity in each variant (score-based: top-1 and top-10 mean score, best own and best non-own score and their margin, own count, cross-camera own score, mean own score; rank-based: reciprocal rank of the best own entry, own at rank 1, own within rank 5, own fraction of the top ten, rank of the first non-own entry, cross-camera reciprocal rank, own absent, nearest neighbour same identity). Table 14 and Figure 7 give the results.

H.5 The frontier sweep

Table 15 lists every configuration of Figure 3: the selected operating point first, followed by the additional sweep configurations, with utility averaged over the surviving seeds and the paired gap

Table 12: The Market-1501 operating points applied unchanged to other encoders and to CUHK03-NP; surviving seeds in parentheses, utility as forget / retain mAP, paired gap with significant seeds under each adversary.

	forget	retain	mixed gap (sig.)	fresh gap (sig.)
ResNet-18, Market-1501				
`finetune` (3/3)	0.962	0.999	+0.231 (3/3)	+0.151 (3/3)
`neggrad_plus` (3/3)	0.960	0.999	+0.234 (3/3)	+0.152 (3/3)
`artifact_aware`			collapsed 3/3	
`grad_ascent` (3/3)	0.846	0.917	+0.212 (3/3)	+0.119 (3/3)
`scrub`			collapsed 3/3	
ConvNeXt-T, Market-1501				
`finetune`			collapsed 3/3	
`neggrad_plus` (3/3)	0.734	0.886	+0.034 (0/3)	−0.054 (2/3)
`artifact_aware`			collapsed 3/3	
`grad_ascent`			collapsed 3/3	
`scrub`			collapsed 3/3	
ResNet-50, CUHK03-NP				
`finetune` (3/3)	0.956	1.000	+0.014 (0/3)	+0.166 (3/3)
`neggrad_plus` (3/3)	0.944	0.998	+0.013 (0/3)	+0.175 (3/3)
`artifact_aware` (3/3)	0.429	0.789	−0.004 (0/3)	+0.006 (0/3)
`grad_ascent` (3/3)	0.918	0.993	+0.014 (0/3)	+0.179 (3/3)
`scrub` (2/3)	0.417	0.833	−0.000 (0/2)	+0.058 (1/2)

Table 13: Detector fitted on the first setting's released original, evaluated unchanged on the second setting's released encoder and on its retrain oracle; means over the pairs named, with the number of pairs whose gap is positive.

	mixed adversary			fresh adversary		
fit → evaluated	released	oracle	gap (pos.)	released	oracle	gap (pos.)
Market ResNet-50 → Market ResNet-50, disjoint seeds (20 pairs)	0.988	0.568	+0.419 (20/20)	0.801	0.591	+0.210 (20/20)
ResNet-50 → ResNet-18, same identities	0.996	0.580	+0.417 (3/3)	0.835	0.618	+0.217 (3/3)
ResNet-50 → ResNet-18, disjoint identities	0.996	0.577	+0.419 (6/6)	0.798	0.570	+0.227 (6/6)
ResNet-50 → ConvNeXt-T, same identities	0.958	0.596	+0.362 (3/3)	0.814	0.657	+0.157 (3/3)
ResNet-50 → ConvNeXt-T, disjoint identities	0.956	0.595	+0.361 (6/6)	0.763	0.604	+0.159 (6/6)
ConvNeXt-T → ResNet-50, disjoint identities	0.991	0.601	+0.390 (6/6)	0.788	0.581	+0.207 (6/6)
ResNet-18 → ConvNeXt-T, disjoint identities	0.948	0.590	+0.357 (6/6)	0.760	0.601	+0.160 (6/6)
Market-1501 → CUHK03-NP (all 9 pairs)	0.995	0.627	+0.368 (9/9)	0.709	0.559	+0.150 (9/9)
CUHK03-NP → Market-1501 (all 9 pairs)	0.888	0.668	+0.220 (9/9)	0.621	0.489	+0.132 (9/9)

with significant seeds under each adversary. The oracle's forget mAP is 0.825. No configuration satisfies retain mAP ≥ 0.95, forget mAP $\leq$ oracle $+0.02$ and an unrejected gap within ± 0.05 under the mixed adversary.

I POST-HOC DEFENCES

A deployer who accepts Section 5 has one move left that does not involve retraining: perturb the embedding before releasing it. We test the two perturbations that cost nothing to apply, additive Gaussian noise and PCA truncation, on seed 0's released embedding at 20 forget identities under the mixed adversary of Section 3, and recompute retain mAP on the same perturbed embedding the adversary attacks, so that no defence can buy a lower AUC by degrading a representation nobody would ship. Nulling the leading forget-versus-unseen direction was already excluded in Section 6:

Table 14: Detection through a retrieval service, 100 identities, three seeds: AUC, the same-seed oracle under the same service, and the paired gap with significant seeds.

		mixed adversary			fresh adversary		
service	model	AUC	oracle	gap (sig.)	AUC	oracle	gap (sig.)
top-10 scores only	original	0.840 ± 0.019	0.623	$+0.216$ (3/3)	0.694 ± 0.023	0.591	$+0.102$ (3/3)
	finetune	0.730 ± 0.078	0.623	$+0.107$ (3/3)	0.640 ± 0.075	0.591	$+0.049$ (1/3)
	neggrad_plus	0.746 ± 0.014	0.623	$+0.123$ (3/3)	0.615 ± 0.036	0.591	$+0.023$ (0/3)
	artifact_aware	0.626 ± 0.045	0.623	$+0.003$ (0/3)	0.591 ± 0.045	0.591	-0.000 (0/3)
top-10 ranks only	original	0.737 ± 0.019	0.623	$+0.114$ (3/3)	0.635 ± 0.024	0.578	$+0.057$ (2/3)
	finetune	0.703 ± 0.045	0.623	$+0.080$ (2/3)	0.609 ± 0.053	0.578	$+0.031$ (0/3)
	neggrad_plus	0.707 ± 0.040	0.623	$+0.084$ (2/3)	0.592 ± 0.053	0.578	$+0.013$ (0/3)
	artifact_aware	0.626 ± 0.023	0.623	$+0.003$ (0/3)	0.582 ± 0.060	0.578	$+0.004$ (0/3)

Table 15: Every configuration of the frontier sweep, 100 identities, three seeds per configuration and five at each method's operating point; *valid* counts non-collapsed seeds.

method	configuration	valid	forget	retain	held-out forget	mixed gap (sig.)	fresh gap (sig.)
artifact_aware	epochs=8, lr=0.05, beta=20	5/5	0.795	0.941	0.784	$+0.015$ (0/5)	-0.019 (0/5)
artifact_aware	epochs=4, lr=0.05, beta=3	3/3	0.947	0.994	0.911	$+0.165$ (3/3)	$+0.077$ (3/3)
artifact_aware	epochs=4, lr=0.05, beta=10	3/3	0.844	0.956	0.821	$+0.076$ (2/3)	-0.010 (0/3)
artifact_aware	epochs=8, lr=0.05, beta=10	3/3	0.855	0.972	0.840	$+0.060$ (2/3)	-0.004 (0/3)
artifact_aware	epochs=12, lr=0.05, beta=20	3/3	0.767	0.937	0.762	$+0.015$ (0/3)	-0.029 (0/3)
artifact_aware	epochs=8, lr=0.05, beta=30	3/3	0.753	0.913	0.745	$+0.007$ (0/3)	-0.033 (0/3)
artifact_aware	epochs=12, lr=0.05, beta=30	3/3	0.711	0.897	0.726	-0.002 (0/3)	-0.032 (1/3)
artifact_aware	epochs=16, lr=0.05, beta=50	3/3	0.639	0.805	0.650	-0.021 (0/3)	-0.038 (1/3)
finetune	epochs=2, lr=0.07	5/5	0.921	0.959	0.895	$+0.192$ (5/5)	$+0.115$ (4/5)
finetune	epochs=2, lr=0.01	3/3	0.973	1.000	0.948	$+0.243$ (3/3)	$+0.153$ (3/3)
finetune	epochs=2, lr=0.03	3/3	0.972	1.000	0.948	$+0.239$ (3/3)	$+0.145$ (3/3)
finetune	epochs=2, lr=0.05	3/3	0.971	0.999	0.946	$+0.234$ (3/3)	$+0.137$ (3/3)
finetune	epochs=2, lr=0.1	collapsed on every seed					
finetune	epochs=2, lr=0.15	collapsed on every seed					
finetune	epochs=4, lr=0.07	3/3	0.961	0.997	0.933	$+0.212$ (3/3)	$+0.121$ (3/3)
finetune	epochs=8, lr=0.05	3/3	0.970	0.999	0.947	$+0.229$ (3/3)	$+0.132$ (3/3)
grad_ascent	epochs=2, lr=0.001	5/5	0.806	0.889	0.777	$+0.179$ (5/5)	$+0.070$ (4/5)
grad_ascent	epochs=2, lr=0.0001	3/3	0.970	1.000	0.945	$+0.251$ (3/3)	$+0.168$ (3/3)
grad_ascent	epochs=2, lr=0.0002	3/3	0.970	1.000	0.944	$+0.251$ (3/3)	$+0.168$ (3/3)
grad_ascent	epochs=2, lr=0.0003	3/3	0.970	1.000	0.944	$+0.250$ (3/3)	$+0.167$ (3/3)
grad_ascent	epochs=2, lr=0.0005	3/3	0.969	1.000	0.942	$+0.249$ (3/3)	$+0.164$ (3/3)
grad_ascent	epochs=1, lr=0.001	3/3	0.970	1.000	0.943	$+0.250$ (3/3)	$+0.166$ (3/3)
grad_ascent	epochs=5, lr=0.0001	3/3	0.970	1.000	0.944	$+0.250$ (3/3)	$+0.167$ (3/3)
neggrad_plus	epochs=2, lr=0.005, alpha=0.9	5/5	0.971	0.999	0.945	$+0.225$ (5/5)	$+0.142$ (5/5)
neggrad_plus	epochs=2, lr=0.002, alpha=0.9	3/3	0.972	1.000	0.947	$+0.246$ (3/3)	$+0.155$ (3/3)
neggrad_plus	epochs=2, lr=0.008, alpha=0.9	3/3	0.959	0.998	0.924	$+0.193$ (3/3)	$+0.074$ (1/3)
neggrad_plus	epochs=2, lr=0.005, alpha=0.7	3/3	0.741	0.935	0.713	$+0.029$ (0/3)	-0.018 (1/3)
neggrad_plus	epochs=2, lr=0.005, alpha=0.5	collapsed on every seed					
neggrad_plus	epochs=2, lr=0.002, alpha=0.5	3/3	0.643	0.869	0.649	$+0.022$ (0/3)	-0.023 (1/3)
neggrad_plus	epochs=2, lr=0.008, alpha=0.7	collapsed on every seed					
neggrad_plus	epochs=4, lr=0.005, alpha=0.9	3/3	0.903	0.991	0.862	$+0.152$ (3/3)	-0.007 (0/3)
neggrad_plus	epochs=2, lr=0.003, alpha=0.8	3/3	0.963	0.999	0.927	$+0.208$ (3/3)	$+0.103$ (3/3)
scrub	epochs=5, lr=0.01, max_steps=3	collapsed on every seed					
scrub	epochs=5, lr=0.003, max_steps=3	2/3	0.922	0.992	0.869	$+0.158$ (2/2)	$+0.025$ (1/2)
scrub	epochs=5, lr=0.005, max_steps=3	2/3	0.882	0.975	0.834	$+0.096$ (2/2)	-0.027 (0/2)
scrub	epochs=5, lr=0.007, max_steps=3	collapsed on every seed					
scrub	epochs=5, lr=0.005, max_steps=1	3/3	0.970	1.000	0.943	$+0.231$ (3/3)	$+0.122$ (3/3)
scrub	epochs=5, lr=0.005, max_steps=5	collapsed on every seed					
scrub	epochs=2, lr=0.01, max_steps=3	1/3	0.906	0.969	0.853	$+0.175$ (1/1)	$+0.012$ (0/1)

AUC $0.914 \rightarrow 0.908$ across three seeds at retain mAP $0.9995 \rightarrow 0.9994$. The mark is not one direction, so deleting one does not remove it.

Noise does nothing at any level (Table 16). Across $\sigma \in [0.02, 0.4]$ retain mAP stays inside $[0.9997, 0.9998]$ and AUC inside $[0.9035, 0.9255]$. PCA truncation costs nothing down to 128 components and almost nothing at 32 (AUC 0.8955 at retain mAP 0.9927). Only 8 components move the attack, to 0.8090, and they buy that 9.5 points of AUC with 59 points of retain mAP. The comparison that settles the question is the retrain oracle on the same seed, the same twenty identities, the same draws and the same detector: AUC 0.5220 at retain mAP 0.9997. Deleting the data reaches a

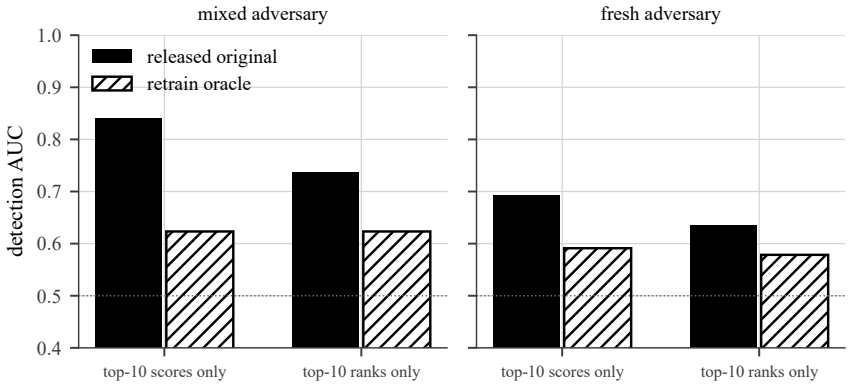


Figure 7: Detection through a retrieval service that returns only its ten best entries with scores, or their order alone; released original against the same-seed retrain oracle under the same service, 100 identities, three seeds.

Table 16: Post-hoc defences on the released embedding (seed 0, 20 forget identities, mixed adversary). AUC is the detector's area under the curve averaged over the protocol's draws, not the paired gap of Equation 1; the undefended row is the seed-0 entry of Table 7. Retain mAP is measured on the perturbed embedding itself. Noise σ is relative to the mean embedding norm. The last row is not a defence: it is the retrain oracle on the same seed, identities, draws and detector, and it is the point every row above is trying to reach.

family	setting	attack AUC	retain mAP
released embedding	no defence	0.9035	0.9997
Gaussian noise	$\sigma = 0.02$	0.9035	0.9997
	$\sigma = 0.05$	0.9035	0.9997
	$\sigma = 0.1$	0.9055	0.9997
	$\sigma = 0.2$	0.9045	0.9997
	$\sigma = 0.4$	0.9255	0.9998
PCA truncation	2048 components	0.9035	0.9997
	512	0.9120	0.9998
	128	0.9270	0.9999
	32	0.8955	0.9927
	8	0.8090	0.4095
retrain oracle	reference, not a defence	0.5220	0.9997

detectability the best post-hoc defence misses by 0.29, and pays none of the retrieval cost. Figure 8 draws this as a frontier. The band a defence would have to reach is empty, and the only point inside it is the one that required not training on the images in the first place.

Two entries move the wrong way, to 0.9255 at $\sigma = 0.4$ and to 0.9270 at 128 components, and both sit inside the three-seed spread of the undefended quantity, 0.9025 to 0.9360 (Table 7), so we do not read them as effects of the perturbation. The count-related channel of the feature battery, which perturbations acting on directions leave untouched, cannot explain a rise: the covariate-only detector runs below chance (Appendix B). Tiny-ImageNet removes that channel by construction, every class there having the same number of images, and the attack is stronger there, not weaker.

This closes the two zero-cost post-hoc routes and nothing else. These single-seed numbers support only the negative claim, that no perturbation removed the mark; they are not evidence about its size. Training-time defences, differentially private optimisation, explicit suppression of the count channel, per-identity calibration of gallery composition, are untested, and a deployer willing to retrain at all has the oracle itself available.

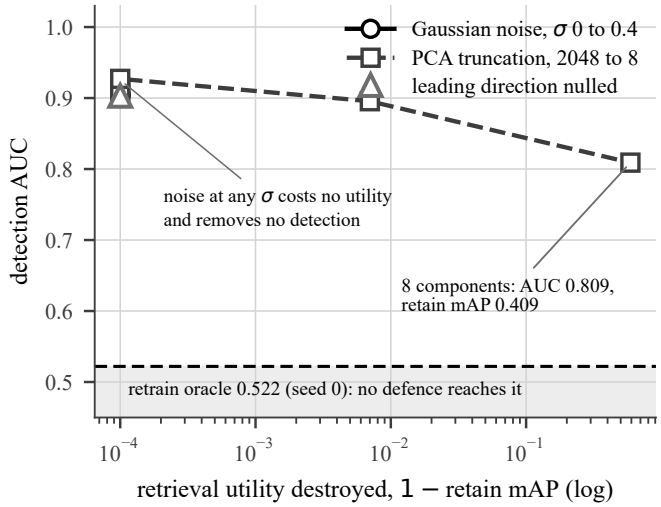


Figure 8: Each post-hoc defence against the retrieval utility it destroys, seed 0, 20 identities, mixed adversary. The shaded band is at or below the retrain oracle's detectability, where a defence would have to land to work. Nothing reaches it: the only perturbation that meaningfully lowers detection, truncating to eight principal components, has already taken retained retrieval from 0.9997 to 0.410.

J STABILITY OF THE SELECTED OPERATING POINTS

This appendix gives the full stability study summarised in Section 8. The design fixes the forget set, the original checkpoint, the retain budget and every hyperparameter, and varies only the run seed, which controls batch order and method-internal sampling. Each cell is eight runs; the study is 27 cells and 216 runs on Market-1501, and no run errored.

The collapse map. Figure 9 and Table 17 give collapse counts and survivor utility for every cell. SCRUB and NegGrad+ die above $\mathrm{lr} = 0.01$; `artifact_aware` survives to 0.10 and dies at 0.20, roughly an order of magnitude higher, and is already bistable at 0.02. Each method has a band in which the identical configuration both survives and destroys the encoder. SCRUB's spans $\mathrm{lr} = 0.007$ to 0.010, a factor of 1.4 in learning rate, wide enough that careful tuning does not step over it. SCRUB's published operating point, $\mathrm{lr} = 0.01$, sits at $7/8$ collapse: a single run there is close to a coin flip, and the survivor is what gets reported.

Soft failure usually precedes hard failure. NegGrad+ survivors average retain mAP 0.7993 at $\mathrm{lr} = 0.006$, where nothing collapses, then 0.7657 at 0.007 and 0.6712 at 0.008; `artifact_aware` survivors average 0.7549 at 0.10 without a single collapse in the cell. For these two methods a utility check catches the approaching cliff. SCRUB gives no such warning: its survivors hold 0.9947 at 0.007 and 0.9791 at 0.009, where seven of eight runs are already dead. A healthy-looking run is not evidence of a healthy setting, even when its retain mAP is indistinguishable from the original's.

The cliff moves with forget-set size. At 20 identities `grad_ascent` was the healthiest method in the study, at retain mAP 0.9986 and forget mAP 0.9634. At 100 identities, with the same hyperparameters, five times the forget images and the withheld probes inside its loader, it destroyed the encoder on all three seeds (forget mAP 0.057, retain mAP 0.061); with the loader restricted to the training images it survives on 5 of five seeds at retain 0.889 (Appendix H.1). SCRUB collapses on all three seeds at that scale as well, at forget mAP 0.003 and retain mAP 0.004. Both are reported as collapsed rather than scored. An operating point validated at one forget-set size carries no guarantee at another, so a stability cell has to be rerun whenever the size of the deletion request changes.

Cross-dataset confirmation. The fragility is not a Market-1501 artifact. Carried unchanged to Tiny-ImageNet, the same operating points make SCRUB emit non-finite activations on all three seeds and leave `grad_ascent` at retain mAP 0.0301.

Table 17: Collapse fraction (runs with retain mAP below 0.50, out of eight) and mean retain mAP of the surviving runs, for every method-by-learning-rate cell. An empty utility entry means no run survived. Learning rates not run for a method are omitted.

Method	Learning rate	Collapsed	Survivor retain mAP
SCRUB	0.003	0/8	0.9993
	0.005	0/8	0.9988
	0.006	0/8	0.9954
	0.007	3/8	0.9947
	0.008	4/8	0.9942
	0.009	7/8	0.9791
	0.010	7/8	0.8206
	0.012	8/8	
	0.015	8/8	
	0.020	8/8	
	0.040	8/8	
NegGrad+	0.002	0/8	0.9978
	0.005	0/8	0.9674
	0.006	0/8	0.7993
	0.007	3/8	0.7657
	0.008	7/8	0.6712
	0.009	8/8	
	0.010	8/8	
	0.012	8/8	
	0.015	8/8	
	0.020	8/8	
	0.040	8/8	
`artifact_aware`	0.02	1/8	0.9581
	0.05	0/8	0.9374
	0.10	0/8	0.7549
	0.20	8/8	
	0.40	8/8	

The collapse rule and the cells it discards. We discard any run with retain mAP below 0.50. The threshold is not load-bearing: across the 216 runs no run lies between 0.3504 and 0.5333, and collapsed runs sit at 0.003 to 0.084 against 0.94 to 0.999 for healthy ones. At 20 identities the rule discards three cells, SCRUB s_0 at 0.0029, SCRUB s_1 at 0.0027 and `neggrad_plus` s_2 at 0.0328, which is why those two methods contribute one and two seeds to the 20-identity table. At 100 identities it discards SCRUB on every seed, and `grad_ascent` only when the withheld probes were in its loader.

Survivorship. The surviving runs in a bistable cell are not random draws from the configuration. A survivor is a run that happened to train least destructively, which we expect to bias it toward higher residual retrieval and, by this paper's own thesis, toward more leakage. We have not measured that bias and its magnitude is unknown. If its direction is as we argue, it inflates rather than suppresses the surviving methods' apparent detectability, so it cannot manufacture the non-significant cells we report; the rows with one or two surviving seeds are upper bounds on those methods' safety rather than measurements of it.

K Probe protocol and which method a benchmark accuses

Scoring a forget identity with frames the original model trained on conflates the deletion request with memorisation of those frames, and it moves the one reference an audit needs. Under training-image probes the retrain oracle, a model that never saw the forget identities at all, scores 0.363, 0.354 and 0.478 across the three seeds (0.398 ± 0.069); an independent anchor in the same sweep lands at

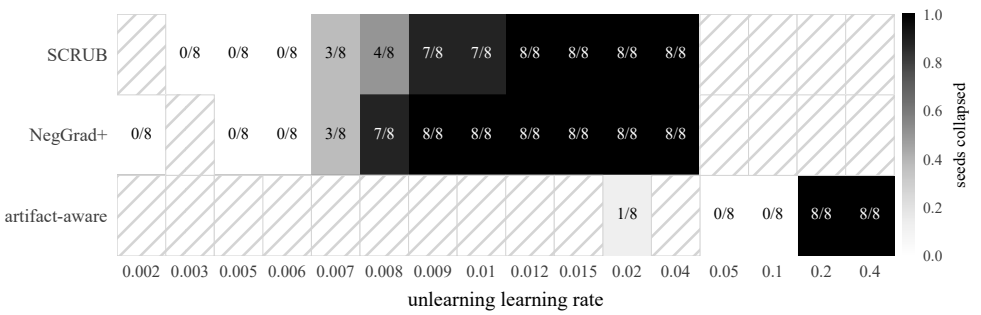


Figure 9: Unlearning stability with the model, the forget set and every hyperparameter held fixed, so only the run seed varies; each cell gives how many of eight runs destroyed the encoder. All three methods pass from never collapsing to always collapsing across a narrow band of learning rate, and inside it the identical configuration does both. SCRUB's published operating point, $\mathrm{lr} = 0.01$, lies inside its band. Hatched cells were not run.

0.435 with no draw significant. Under held-out probes the same oracle sits above chance on every seed: 0.522, 0.530 and 0.6435 (0.565 ± 0.068). An AUC below 0.5 says the forget identities are harder to retrieve than their matched controls, so a criterion stated as distance from chance reads the training-probe oracle as evidence pointing the other way. Every probe in this paper is drawn from images no model trained on.

The mis-placed reference also moves the accusation. In the training-probe sweep the methods flagged in all five negative-set draws were plain fine-tuning (AUC 0.855) and SCRUB (0.8125); `grad_ascent` (0.690) and `artifact_aware` (0.5975) were flagged in one draw of five, and `neggrad_plus` (0.4675) in none. Under held-out probes the ordering inverts: `neggrad_plus` and `finetune` leak on every seed under the mixed adversary at 100 identities, `artifact_aware` does not, and SCRUB's cells are invalid because the encoder collapsed. The two protocols differ in their operating points as well as their probes, and the counts run over negative-set draws in the first case and over seeds in the second, so this is not a controlled swap. It establishes the narrower claim the paper needs: which method a benchmark indicts is a property of its probe set, and a protocol that mis-places the oracle cannot audit unlearning at all.

L Scope: the runs behind the caveats

L.1 The ghost-unlearning control

To ask whether a method's effect on the detector is targeted at the opt-out identities, we ran each method at its selected operating point on twenty *ghost* identities drawn from the test split, people no model ever trained on, and then re-measured the paired gap of the *real* forget identities against the untouched original encoder. One seed, twenty identities, five matched draws, 200 permutations. A method whose detector-level effect is specific to the forget set should leave this gap intact (Table 18).

Table 18: Applying each method to identities no model trained on, and re-measuring the real forget identities' paired gap. Two methods erase most of the real signature, down to roughly the oracle's level, without ever being shown a forget image.

method run on ghosts	real-forget gap vs. original	p	real-forget AUC (model / original)
artifact_aware	-0.326	0.007	0.578 / 0.904
neggrad_plus	-0.378	0.005	0.526 / 0.904
finetune	-0.112	0.14	0.792 / 0.904
grad_ascent	-0.093	0.22	0.810 / 0.904
scrub	-0.090	0.16	0.814 / 0.904

artifact_aware and neggrad_plus remove most of the real forget identities' signature although no forget image entered either run. At these operating points the erasure is therefore not specific to the

deleted people; it is consistent with the retain-only fine-tuning both perform. This is why Section 9 reports that artifact_aware is not detected above the oracle at its operating point and attributes that to nothing. A control on the same checkpoint shows the effect is not global: twenty retain identities scored against never-seen fall only -0.094 ($p = 0.22$) while the forget identities and the ghosts fall together. The retain term shields the identities it names and every other trained identity flattens, so the non-leak is real but not specific to the opt-out identities; whether neggrad_plus behaves this way at 100 identities, where it leaks on $5/5$ seeds, is untested.

One further observation from the same run, which we record rather than interpret: scrub is the only method that leaves a significant mark *on the ghosts themselves*, gap $+0.226$ at $p = 0.014$. Unlearning a person the model never learned makes that person detectable.