How Erased Is Erased? Quantifying and Improving the Recoverability of Concept Removal in CLIP

Anonymous authors
Paper under double-blind review

Abstract

Concept-erasure ("unlearning") methods for vision-language models such as CLIP typically report success by showing that zero-shot accuracy on the erased class collapses to near zero. We argue this metric conflates two different things: information being *removed from the representation* versus information being merely *hidden from one particular classifier* (the zero-shot text-alignment rule). We introduce a cheap evaluation protocol – training a k-shot (k=5, 20, 50) linear probe directly on the *post-erasure* embeddings and measuring how much of the erased class it recovers – and apply it to three erasure mechanisms on frozen CLIP ViT-B/32 features: a naive single-direction baseline, an iterative multi-direction method (INLP), and a data-free, text-embedding-derived nullspace-projection method whose mechanism matches a recently published approach (Mishra et al., 2025). Across three image domains (CIFAR-100, EuroSAT, and Food101), we find that (i) every method we test leaves the erased concept substantially recoverable despite 0% zero-shot accuracy, and (ii) the data-free, text-derived method is consistently the *most* recoverable of the three (92–97$\%$ recall at k=50), more so than the naive image-based baseline, on all three datasets. On CIFAR-100, where class hierarchy labels are available, we show the leaked signal is not unstructured noise: attack false positives concentrate in the same coarse superclass as the erased class $9.8\times$ more than chance for the text-derived method (measured over 100 randomly-sampled target classes), suggesting it erases a shared semantic-domain direction rather than a class-specific one. We also find that whether this method's extra recoverability comes with an extra utility cost is entangled with how strong the zero-shot baseline is to begin with: the two datasets with a well-separated zero-shot baseline (CIFAR-100, 61.7%; Food101, 84.3%) both show the text-derived method costing the most accuracy on retained classes, while the one dataset with a weak baseline (EuroSAT, 40.1%) shows the opposite – which we argue is an artifact of that weak baseline destabilizing the metric, not a genuine reversal of the phenomenon. We propose Ensemble Subspace Projection (ESP), a bootstrap-aggregated multi-direction erasure operator, which reduces attack recall relative to the classical iterative baseline at equal removed-dimension budget on every dataset we test (by 1–9 recall points), for free on CIFAR-100 but at a small utility cost on the other two; no method we test achieves what we would consider genuine erasure. All of the above uses CLIP ViT-B/32; we additionally confirm our two central findings – NSP-text is the most recoverable *and* the most utility-costly method – hold essentially unchanged on the $2.8\times$ larger ViT-L/14 backbone. Finally, extending from object classes to a demographic attribute on FairFace, we find naive erasure of a binary gender attribute leaves recoverability that swings $8.9\times$ across race groups (7.7–68.6$\%$ recall at k=50) even though every group is erased and attacked identically – a class-level mechanism producing the same kind of group-uneven outcome documented for retraining-based forgetting, and a caution that fairness properties of an erasure method cannot be read off its object-classification behavior alone; tellingly, ESP – our own defense – leaves this disparity just as wide (if not wider) even where it reduces average attack recall elsewhere. We discuss the weak-baseline pitfall and other open problems for future work on evaluating unlearning in vision-language models.

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.

1 INTRODUCTION

Vision-language foundation models such as CLIP (Radford et al., 2021) are increasingly the substrate on which downstream classifiers, retrieval systems, and agents are built. As these models are deployed at scale, there is growing interest in *removing* specific concepts from them after the fact: to comply with a takedown request, to correct a spurious or biased association, or to forget a class that should never have been learned. A line of recent work addresses this by editing CLIP so that a target class's zero-shot classification accuracy drops to (near) zero, either by projecting a class-specific direction out of the shared embedding space (Mishra et al., 2025) or by closed-form weight surgery on an intermediate layer (Hatami et al., 2026).

All of this work shares an evaluation convention: *success is defined as zero-shot accuracy on the erased class collapsing toward zero, while accuracy on retained classes is preserved.* This is cheap but measures something narrower than it appears to: zero-shot CLIP classification is one specific classifier (cosine similarity against fixed text embeddings), so an erasure method can satisfy it by disrupting only the components that classifier reads, leaving the rest of the embedding's content about the class intact. If that content is still linearly decodable, a downstream user – or an adversary – can recover it cheaply with a handful of labeled examples and no access to the erasure procedure or original data.

This is not hypothetical: an analogous shallow, easily-reversed "unlearning" has been documented for large language models, and a recent unified re-evaluation of CLIP test-time adaptation found much of that literature's progress evaporates under fair comparison (Sheng et al., 2025). We ask whether an analogous gap exists for CLIP concept erasure, using a simpler tool than either literature requires: a linear probe.

Contributions. (1) A cheap *recoverability-attack protocol*: after erasure, fit a logistic-regression probe on $k \in \{5, 20, 50\}$ labeled images of the erased class on the post-erasure embeddings, measure held-out recall. Applied to three mechanisms – naive single-direction, iterative multi-direction INLP (Ravfogel et al., 2020), and a data-free, text-embedding-derived method matching Mishra et al. (2025) – on frozen CLIP ViT-B/32, across CIFAR-100, EuroSAT, and Food101. (2) A finding holding on all three datasets: the text-derived, data-free method is the *most* recoverable of the three (30–40 recall points above naive at $k{=}50$), despite the cleanest-looking zero-shot erasure. (3) A leakage-correlation analysis (CIFAR-100's coarse/fine hierarchy, 100 randomly-sampled target classes): its false positives concentrate in the erased class's coarse superclass $9.8\times$ above chance, versus near-chance for naive – a mechanistic clue to what direction each method removes. (4) Evidence that its extra utility cost tracks zero-shot baseline strength: the two strong-baseline datasets agree it costs the most utility; the weak-baseline one disagrees, which Section 7 argues is a metric artifact. (5) Ensemble Subspace Projection (ESP): a bootstrap-aggregated multi-direction operator reducing attack recall relative to INLP at equal budget on every dataset, free on CIFAR-100 but at a utility cost elsewhere, plus a negative ablation. (6) A backbone check: both central findings hold essentially unchanged on a $2.8\times$ larger backbone (ViT-L/14 vs. ViT-B/32). (7) An attribute-level extension on FairFace, connecting to the bias-redistribution question of Haruna et al. (2026): naive gender erasure leaves recoverability varying $8.9\times$ across race groups; NSP-text barely erases the attribute at all; ESP does not narrow the disparity (if anything, $11.3\times$, slightly wider) even while marginally reducing average attack recall.

What is new versus reused: the linear-probe pattern is adapted from LLM-unlearning work (Deeb & Roger, 2024; Lynch et al., 2024), and INLP (Ravfogel et al., 2020) is evaluated, not proposed. New here are (a) the CLIP-specific protocol, (b) leakage-correlation, and (c) ESP; items 2,4,6,7 and the FairFace/CelebA/DAMP extensions are findings these three enable. This is more than a domain port: CLIP's erasure operators are explicit linear subspace removal on a shared, geometric embedding, with no clean analogue in autoregressive LLM unlearning (a behavioral, output-level property there). That structure is what makes (b) and (c) possible at all – asking *where* leaked information concentrates, or building a defense by directly manipulating the removed subspace, presupposes a linear-embedding structure the LLM literature lacks.

We emphasize what this paper does *not* claim: no tested method, including our own, achieves real erasure under this attack – the best case still leaves roughly half the class recoverable at 50 shots.

2 Related Work

Concept erasure in CLIP. Mishra et al. (2025) erase a target class from CLIP in a data-free way by projecting a text-embedding-derived direction out of the visual projection layer; this is our NSP-text. Hatami et al. (2026) propose DAMP, a depth-aware, closed-form weight-surgery method using class-prototype statistics, which we adapt as a fourth baseline (Section 6.1). Both papers evaluate primarily via zero-shot accuracy; neither reports a recoverability-style attack.

Reversibility of unlearning. Our finding echoes "unlearning" reversibility work in LLMs: Lynch et al. (2024) show a widely-cited method leaks far more than its headline metric suggests once probed with complementary evaluations; Deeb & Roger (2024) ask our exact question – does unlearning remove information, or just suppress it at the output layer – and find popular methods leave it recoverable; Xu et al. (2025) recover "forgotten" knowledge via fine-tuning, quantization, or adversarial prompting and propose a reversibility taxonomy. We translate this to CLIP with a substantially cheaper attack (a linear probe, not LLM fine-tuning), enabled by CLIP's erasure operators acting linearly on a fixed embedding space. Closest outside the CLIP literature is Ravfogel et al. (2020)'s INLP, which iterates precisely because a single removed direction is known reversible; we adopt it as a strong baseline and show it still leaves substantial signal recoverable here.

Evaluation gaps in adjacent CLIP subfields. Sheng et al. (2025)'s unified re-evaluation of CLIP test-time adaptation finds reported progress often does not hold up once duplicated baselines and narrow metrics are corrected for, and that accuracy gains there often cost calibration and OOD detection. We identify an analogous hidden-cost pattern in erasure: a method can look strictly better on the standard metric while being strictly worse on an axis that metric does not measure.

Bias redistribution under forgetting. Haruna et al. (2026) show that forgetting one demographic group in visual unlearning can shift bias onto a different, correlated group rather than removing it. Our leakage-correlation analysis (Section 6.3) is a class-level analogue of this idea: we ask not just *whether* information leaks back after erasure, but whether it leaks back along semantically meaningful axes.

Connection to classical privacy attacks. Our attack is a distant relative of membership inference (Shokri et al., 2017) and model inversion (Fredrikson et al., 2015), which similarly exploit leaked information beyond a model's stated output. Ours differs in target: whether a label an erasure operator tried to remove is still linearly decodable from the *same* embeddings it produced.

3 Recoverability Attack Protocol

Setup. Let $f_I : \mathcal{X} \to \mathbb{R}^d$ and $f_T : \mathcal{T} \to \mathbb{R}^d$ denote CLIP's (frozen) image and text encoders, sharing an ℓ_2-normalized embedding space of dimension d (here d=512, ViT-B/32). Zero-shot classification over classes $c_1, \ldots, c_C$ assigns image x the label $\arg\max_i \; f_I(x)^\top f_T(\text{"a photo of a } c_i\text{"})$.

Erasure operators. All three operators we study take the form of removing an r-dimensional subspace, spanned by orthonormal directions $\{u_1, \ldots, u_r\} \subset \mathbb{R}^d$, from every image embedding:

$$\hat{f}_I(x) = \text{normalize}\left(f_I(x) - \textstyle\sum_{j=1}^{r} \left(f_I(x)^\top u_j\right) u_j\right). \tag{1}$$

This is applied identically to every image at inference time and is mathematically equivalent to editing CLIP's visual projection weight matrix W as $W \leftarrow (I - \sum_j u_j u_j^\top)W$, so results transfer directly to a weight-edited deployment.

- **Naive.** r=1; u_1 is the normalized difference between the mean image embedding of the target class and the mean of all other classes, estimated on a labeled training pool. This requires images of the class to be forgotten.
- **NSP-text.** r=1; u_1 is the normalized mean of CLIP text embeddings for the target class name under six prompt templates (e.g. "a photo of a {}", "a close-up photo of a {}"). This is *data-free* – it requires only the class name – and matches the mechanism reported by Mishra et al. (2025): projecting a text-derived direction out of the visual projection layer.

- **INLP-iterative.** Following Ravfogel et al. (2020): fit a logistic-regression classifier (target-vs-rest) on the current residual training embeddings, add its (normalized) weight vector to the removed set, project it out, and repeat for up to 8 rounds or until the classifier's training accuracy on the residual drops below a tolerance (0.55–0.60, dataset-dependent).

Attack. After erasure, we sample $k \in \{5, 20, 50\}$ images of the erased class and $5k$ of other classes from the training pool (all passed through the same erasure operator), fit a fresh logistic regression probe to distinguish the erased class from the rest, and report its recall on a disjoint, equally-erased test set: *given only post-erasure embeddings and a handful of labels, how much of the erased class can a cheap linear model recover?*

Metrics. Per (method, class) we report post-erasure zero-shot recall on the target class (should $\to 0$), zero-shot accuracy on other classes (utility), and attack recall at k=5, 20, 50.

4 Ensemble Subspace Projection

INLP removes one direction per round, re-fitting a classifier on the residual each time. We hypothesized this greedy procedure leaves "escape routes": the round-t+1 classifier only avoids the t removed directions, not the broader region many differently-sampled classifiers would converge to – and a later attacker's probe is exactly such a classifier.

Ensemble Subspace Projection (ESP) instead estimates a K-dimensional *stable* subspace up front, by bootstrap-aggregating many classifier fits, and removes it in one shot: fit B logistic-regression classifiers on random negative subsamples ($\leq 10\times$ positives) with random $C_b \in \{0.01, 0.1, 1, 10\}$, stack their normalized weight vectors into $W \in \mathbb{R}^{B\times d}$, and remove its top-$K$ right singular vectors via Eq. 1 (pseudocode: Appendix A).

We use B=30 bootstrap fits and compare K=8 (matched to INLP's typical round count) and K=16. We also test an iterative variant, IESP, which repeats Algorithm 1 with K_0=4 per round, re-checking a fresh probe's residual accuracy after each round and stopping under the same tolerance rule as INLP, for up to 4 rounds.

5 Experimental Setup

Backbone and hardware. Main experiments use frozen `openai/clip-vit-base-patch32` (ViT-B/32, d=512); Section 6.5 adds `clip-vit-large-patch14` (ViT-L/14, d=768, 428M). We never fine-tune CLIP – every step is a forward pass or fitting a small linear model on cached features, so all experiments ran on one consumer GPU (RTX 3070, 8GB) in minutes.

Datasets. **CIFAR-100** (Krizhevsky, 2009): 100 fine-grained classes, 20 coarse superclasses (leakage analysis, Section 6.3), 50,000/10,000 train/test. The main comparison (Table 1) targets all 100 classes randomly; the ESP ablations (Table 10) and backbone check (Appendix D) use 10 hand-picked-for-diversity classes. **EuroSAT** (Helber et al., 2019): 10 land-cover classes, 27,000 images, 80/20 split. **Food101** (Bossard et al., 2014): 101 food classes, 25,250-image validation split (`ethz/food101`), 80/20 split, 10 random target classes (baseline 84.3%, contrasting EuroSAT's weak one). *Hyperparameters:* prompt templates in Appendix L; INLP up to 8 rounds (tolerance 0.55/0.60); ESP B=30, $K\in\{8, 16\}$; attack probe `sklearn LogisticRegression`, balanced, k positive/$5k$ negative, $k\in\{5, 20, 50\}$, fixed seed across methods.

6 Results

6.1 Erasure looks successful, but is not

Table 1 shows the CIFAR-100 comparison, averaged over 100 randomly-sampled target classes $\times$ 5 seeds – not 10 hand-picked classes, resolving this paper's main sampling limitation (numbers turn out representative: INLP k=50 was 59.0% on the original 10, $58.5\% \pm 1\%$ here). All three methods drive post-erasure recall to within 5 points of zero – by the standard metric, all "work." Figure 1 shows what the attack recovers.

Table 1: CIFAR-100, mean over 100 random target classes × 5 seeds (k=5/k=50 show ±95% CI). "Recall" is post-erasure zero-shot recall on the target class (lower = erasure looks more successful); "Acc" is zero-shot accuracy on the other 99 classes (higher = better utility preservation). DAMP is our reimplementation (Sec. 6.1).

Method	#dirs	Recall	Acc	k=5	k=50
Naive	1	0.041	0.594	0.22±.01	0.50±.01
NSP-text	1	0.009	**0.520**	0.79±.01	0.92±.00
INLP	8.0	0.012	0.568	0.27±.01	0.59±.01
DAMP	1	0.348	0.615	**0.79±.06**	**0.93±.02**

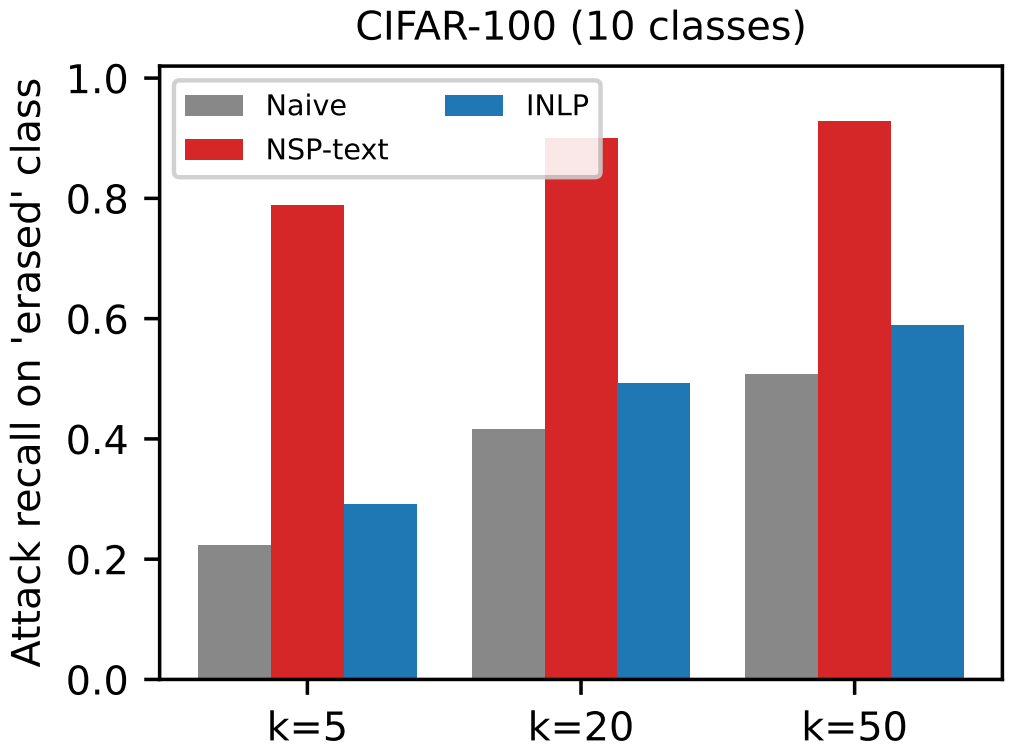

Figure 1: Attack recall at k=5, 20, 50 on CIFAR-100. NSP-text – the method whose mechanism matches a recently published, data-free SOTA approach (Mishra et al., 2025) – is recovered far more easily than either the naive or INLP baselines, at every shot budget.

The ordering is striking: NSP-text, the most "principled" and data-free of the three, is also the *most* recoverable (92.3% recall at k=50, vs. 50.0% for naive and 58.5% for INLP), and it costs the most utility on the other 99 classes (−9.7 points vs. −2.3 to −4.9 points for the other two).

DAMP: a closed-form baseline, adapted from the unlearning literature. We reimplement Hatami et al. (2026)'s DAMP (no public code, but fully specified equations) adapted to our single-embedding setting – full method and caveats in Appendix C. It is a genuinely different failure mode: zero-shot recall does not reliably reach near-zero (34.8% vs. ≤1.2% for the others), yet the residual signal is close to fully recoverable (93% at k=50) at essentially zero utility cost (+0.2pp, cheapest of all four).

Table 5 (full breakdown, both datasets, Appendix B) extends the comparison to EuroSAT and Food101 (chosen specifically because its zero-shot baseline, 84.3%, is much stronger and more stable than EuroSAT's, 40.1%). The recoverability ordering is preserved on both – NSP-text is

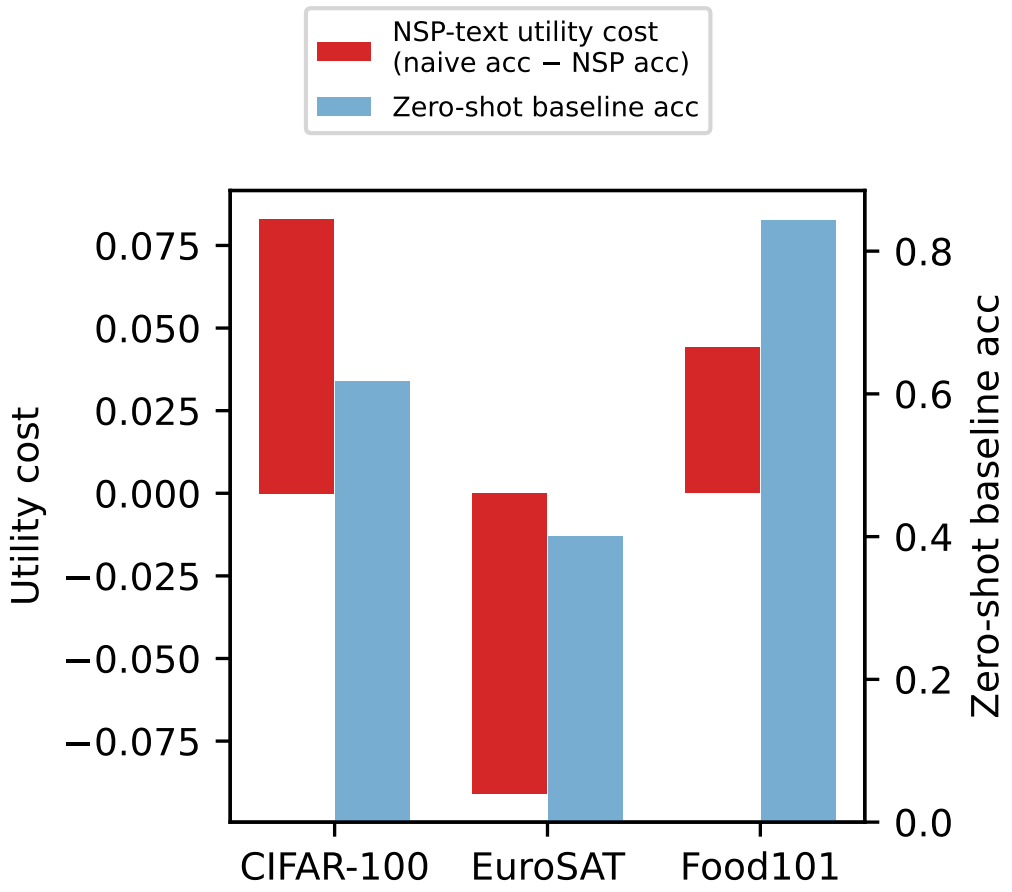

Figure 2: NSP-text's utility cost (red) vs. zero-shot baseline strength (blue) across the three datasets. The two strong-baseline datasets (CIFAR-100, Food101) agree that NSP-text costs utility; the weak-baseline dataset (EuroSAT) disagrees.

again the most recoverable method by a wide margin (92% EuroSAT, 97.1% ± 0.9 Food101 at k=50, 5-seed means, the highest of all three datasets) – but the utility-cost direction only replicates on Food101 (-4.4 points vs. -1.7 for naive and -2.4 for INLP, matching CIFAR-100), while EuroSAT *reverses* it (NSP-text now *preserves* utility best, $+9.1$ points over naive). We examine this split directly in Section 6.2. DAMP's zero-shot erasure is markedly weaker on Food101 specifically (76.8% residual recall, vs. 18.5% on EuroSAT and 34.8% on CIFAR-100) – on this dataset it often does not clear even the field's own success bar, yet is still 98% recoverable at k=50 where it does touch the class.

6.2 Utility cost tracks zero-shot baseline strength

Figure 2 plots, for each dataset, zero-shot baseline accuracy against NSP-text's utility cost relative to naive (naive minus NSP-text accuracy after erasure). CIFAR-100 (61.7% baseline) and Food101 (84.3%) both show a positive cost ($+7.4$, $+4.4$ points); EuroSAT (40.1%) shows a *negative* cost (-9.1). With three datasets we cannot claim a general law, but the two strong-baseline datasets agreeing while the weak-baseline one disagrees is consistent with Section 7's argument that EuroSAT's weak baseline destabilizes the measurement, not a genuine reversal.

A fourth, intermediate-baseline test (a 52.4%-baseline CIFAR-100 subset of confusable reptile/fish classes; full construction and results in Appendix H) complicates rather than confirms this picture: it agrees with the *strong*-baseline datasets – NSP-text remains most utility-costly ($+13.9$ vs. $+6.2$–

Table 2: Leakage-correlation over 100 random target classes: fraction of attack false positives (at k=50) falling in the erased class's own coarse superclass, vs. chance. (10-hand-picked-class figures in parentheses.)

Method	Observed	Chance	Ratio
Naive	0.050 (0.046)	0.040	1.23× (1.13×)
NSP-text	0.395 (0.539)	0.040	**9.76**× (13.35×)

Table 3: ESP-8 vs. INLP at matched ≈ 8-dimension budget, across all three datasets. ΔAttack is ESP's attack recall at k=50 minus INLP's (negative = ESP more attack-resistant); ΔAcc is ESP's utility minus INLP's (negative = ESP costs utility).

Dataset	ΔAttack@50	ΔAcc	Verdict
CIFAR-100	−0.072	+0.002	Pareto win
EuroSAT	−0.085	−0.043	traded for utility
Food101	−0.012	−0.009	roughly a wash

13.2 points) – rather than reversing like EuroSAT, narrowing our hypothesis to whatever destabilizes EuroSAT sitting closer to its specific 40% regime than to any baseline below ∼60%.

6.3 Leaked information is semantically structured

CIFAR-100's fine/coarse label hierarchy lets us ask a sharper question than "does information leak": *where* does it leak to? For the same 100 classes as Table 1 (seed 0) we record, among the k=50 probe's false positives, what fraction fall in the *same coarse superclass* as the erased class (e.g. mistaking an "erased" apple for another *fruit_and_vegetables* item), against the chance rate given by that superclass's share of the other 99 classes ($4/99 \approx 4.0\%$).

Naive erasure's leakage is close to unstructured (ratio 1.23×); NSP-text's is not (ratio 9.76×, down from the original 10 classes' 13.35× but still dramatic) – mechanistic evidence for *why* NSP-text is both less utility-preserving and more recoverable. Because its removed direction derives from text semantics rather than images of the specific class, it plausibly aligns with a broader shared category direction (e.g., a general "fruit-like object" axis) rather than a class-specific one: removing a shared axis damages every class along it (the larger utility cost) while leaving finer within-category discriminative signal untouched (why a cheap probe still separates the erased class from its category-mates).

6.4 Ensemble Subspace Projection improves on INLP, with limits

Table 3 compares ESP against INLP at matched removed-dimension budget (≈ 8) across all three datasets; the CIFAR-100 breakdown by k, plus the IESP/ESP-16 ablations and a fuller K/B sweep, are in Appendix I (Figure 8 there). On CIFAR-100, ESP-8 is a Pareto improvement over INLP at equal budget: lower attack recall at every k (e.g. 51.8% vs. 59.0% at k=50) and marginally *better* utility; neither the IESP nor ESP-16 ablation beats fixed ESP-8's attack-resistance, and both cost more utility. The appendix sweep confirms K=8, B=30 is close to the best trade-off found: recall is flat or non-monotonic in K beyond 8 while cost rises sharply past K=16, and B beyond 30 barely moves either metric.

Table 3 makes the honest picture explicit: ESP-8 reduces attack recall relative to INLP on *every* dataset we test (by 1.2 to 8.5 recall points), so the core defense mechanism is not a CIFAR-100 artifact. Whether that reduction is free, however, is dataset-dependent – a clean Pareto win on CIFAR-100, a real utility/security trade-off on EuroSAT, and a wash within measurement noise on Food101. We do not read ESP as a general-purpose replacement for INLP, but as evidence that Section 3's recoverability metric is *actionable*: changing the removed subspace measurably moves the attack-recall number this paper argues practitioners should report, even though the move is dataset-dependent and modest on two of three datasets.

Table 4: FairFace gender erasure ("Female" erased). Baseline zero-shot: 93.6% overall, 95.7% Female recall, 91.7% Male accuracy. k=50 is 5-seed mean ±95%CI; the disparity ratio is the max/min race-group recall *within each seed*, reported as the range and mean across the 5 seeds (not a single point estimate).

Method	Recall	Acc	k=50	Disparity ratio
Naive	0.221	0.524	0.31±.04	1.50–8.86× (mean 3.94×)
NSP-text	0.904	0.935	0.95±.01	1.04–1.10× (mean 1.07×)

Why does ESP work, and why does its cost vary? A subspace- overlap analysis (held-out attacker subspace per class vs. each method's removed subspace and the shared cross-class structure; full setup in Appendix F) answers both: ESP's overlap with the held-out attacker exceeds INLP's for *all 30 classes* tested across all three datasets (why ESP resists attack), while its overlap with shared cross-class structure, relative to INLP's, is small on CIFAR-100 and Food101 but an order of magnitude larger on EuroSAT – consistent with ESP's variable cost arising when its ensemble converges on directions shared *across* classes. We flag this as correlational, not a proven causal claim.

6.5 Findings hold on a larger backbone

Every result so far uses CLIP ViT-B/32 (151M parameters, d=512). To check this is not an artifact of that specific, comparatively small backbone, we repeat the CIFAR-100 comparison (same 10 target classes, same protocol) with CLIP ViT-L/14 (428M parameters, d=768, 2.8× larger; full table and figure in Appendix D). Both central findings replicate closely: NSP-text remains by far the most recoverable method (93.6% at k=50 vs. 92.8% on ViT-B/32 – effectively unchanged), and it remains the most utility-costly, costing 10.7 points of accuracy relative to naive (vs. 8.3 points on ViT-B/32) – if anything a slightly *larger* cost on the bigger backbone. ESP's improvement over INLP is again a wash within noise (2.3-point attack-recall reduction at a 0.3-point utility cost), matching the Food101 pattern from Section 6.4. Zero-shot utility is higher across the board on ViT-L/14 (73.3% vs. 61.7% baseline), but this does not change which method leaks the most or costs the most utility.

6.6 Attribute-level erasure: uneven recoverability across demographic groups

This section is a single, focused extension asking whether this paper's central recoverability gap is specific to object classes – every result so far erases an object or land-cover *class*. Connecting to the bias-redistribution question raised by Haruna et al. (2026) – does forgetting affect demographic groups unevenly – we erase "Female" from a binary CLIP zero-shot gender classifier on FairFace (Kärkkäinen & Joo, 2021) (10,954 face images labeled with gender, age, and 7-category race; naive/NSP-text only – INLP/ESP's iterative fitting is less meaningful for an already-binary task) and break attack recall down by the (uninvolved) race label, purely as a stratification variable to check whether erasure treats every group the same way.

Two findings. First, **naive erasure of a binary attribute redistributes unevenly, though less dramatically than a single seed suggested**: across 5 seeds, the within-seed max/min race-group disparity at k=50 ranges 1.5–8.9× (mean 3.9×) – the originally-reported 8.9× sits near the top, not typical, though even the low end is not parity. This is the same class-level, group-uneven mechanism Haruna et al. (2026) document from retraining, now shown to arise from erasure alone; naive also costs far more utility here (93.6% → 52.4%) than in any object setting, since a binary task has no "other 98 classes" to fall back on. Second, and unexpectedly, **NSP-text – the most aggressive eraser everywhere else – barely erases gender at all** (recall 95.7% → 90.4%, vs. naive's → 22.1%), so its near-uniform race breakdown reflects the attribute barely being touched, not fairer erasure – plausibly gender is encoded more redundantly than object classes in CLIP's embedding, or our prompts (Appendix L) isolate a binary contrast less well than a one-vs-many label. The recoverability gap between methods does not universalize across erasure-target *types*.

Does ESP help, does fairness-aware selection help, and does this generalize? Four follow-up checks (full setup/tables: Appendices G, J). (1) ESP does not reduce naive's race-group disparity:

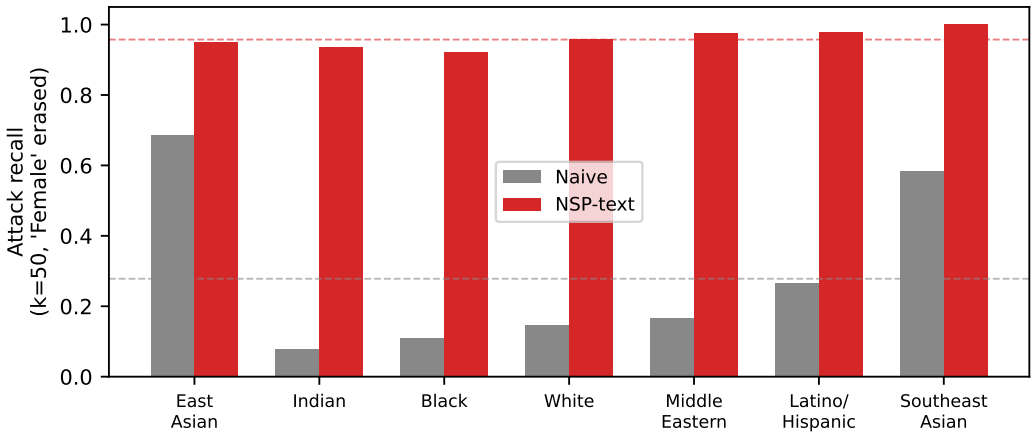

Figure 3: Attack recall on the erased "Female" attribute at k=50, broken down by race, for a single illustrative shot-selection seed (the same one originally reported: naive swings 7.7%–68.6%, an 8.9× spread). Table 4 shows this specific draw sits near the *top* of the range across 5 seeds (1.5–8.9×, mean 3.9×) – still a real, uneven effect, just smaller and noisier than one seed suggested. NSP-text stays flat regardless.

ESP-4/ESP-8 match naive's overall recall (30.8%/29.7% vs. 31.1% at k=50) while the disparity is, if anything, wider (mean 4.2–4.5× vs. 3.9×; individual seeds reach 11.3×) – it resists a generic attacker, not a downstream stratification. (2) A fairness-aware ESP variant (greedy worst-group direction selection, nested CV within Table 4's split) does not help either: statistically indistinguishable from plain ESP-8 – a less rigorous earlier check (separate split, one seed) had suggested otherwise, but we report the stricter-protocol null result. (3) Repeating with *race* as the erased attribute (5-seeded, all four operators, broken down by gender) replicates both findings – NSP-text remains far the most recoverable and cheapest (83% vs. 34–35% recall; 1.9 vs. 8–10pp cost), ESP again fails to outrecall naive/INLP – with a noisier disparity than one seed suggested (mean $\sim$1.1–2.9× depending on target race, seeds up to 7.6×; the previously-reported 6.6× for "White" was one high-variance draw). (4) Repeating gender-erasure on 15,000 CelebA images (Liu et al., 2015) (no race label) replicates the recoverability/cost findings with tight CIs (NSP-text 99.2% ± 0.1 vs. 37–39% at k=50; cost −0.03 vs. 54–57 points) – "erases nothing, costs nothing," again.

7 Discussion and Limitations

Limitations. No method achieves genuine erasure; argmax is unstable when weak: our best result (ESP-8, CIFAR-100) still leaves 51.8% recall recoverable at 50 shots, and EuroSAT's weak (40%, 10-way) baseline let post-erasure recall *exceed* pre-erasure at least once (ESP-8, *HerbaceousVegetation*) – not sensibly a "success"; Section 6.2's intermediate-baseline test narrows, not confirms, the "any weak baseline" hypothesis. Scope: Tables 1/2 use 100 random classes ×5 seeds, not 10 hand-picked (leakage ratio moved 13.35× → 9.76×, inflated by the smaller sample); DAMP (Appendix C) is an unverified reimplementation, not a reproduction; recoverability replicates across two attributes/datasets, but disparity is FairFace-only, age untested.

8 Conclusion

Standard zero-shot evaluation can underestimate residual recoverability: a cheap probe still recovers what it calls erased – more after a SOTA method than naive, and unevenly – and our defense only narrows that gap, it does not close it.

AI USE STATEMENT

In this work, we used a generative-AI coding assistant (Claude, Anthropic) substantially throughout the research process: to implement the experiment pipeline (the erasure operators, the recoverability-attack protocol, and the statistical analysis scripts used to produce every number reported in this paper), to generate the figures from that pipeline's cached outputs, and to draft and revise the manuscript text in all sections. We did not use generative AI to fabricate or substitute for any experimental result: every number in this paper was produced by executing real code against real CLIP embeddings computed on the cited, publicly available datasets, and re-run when an implementation error was found (e.g. an RNG-seeding bug that affected an earlier multi-seed result was caught, fixed, and every affected number was regenerated before being reported). We reviewed all AI-assisted work: generated code was inspected and checked against sanity baselines (e.g. verifying that a corrected script's seed-0 output exactly reproduced an earlier single-seed run before trusting its multi-seed averages), and every citation flagged as unverified during drafting was independently checked against a primary source (arXiv, ACL Anthology, PMLR, or the dataset's official page) before being kept, as documented in the accompanying citation-verification log. We take responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

Our attack is dual-use: it could let an adversary undo a legitimate unlearning request (e.g. a right-to-be-forgotten erasure). We judge publishing it net positive per Section 2's privacy-attack logic – a motivated adversary would likely find it independently, and withholding it only deprives defenders of a cheap detection tool. Our demographic-attribute experiments (Sections 6.6, G) use FairFace and CelebA face images whose consent provenance differs and which we do not overstate; race/gender probes throughout are disposable measurement instruments, reported only as aggregate recall statistics, never used to identify an individual or offered as a usable classifier. Full discussion, including the two face datasets' differing consent provenance, is in Appendix K.

REPRODUCIBILITY STATEMENT

All erasure operators, the recoverability-attack protocol, and Ensemble Subspace Projection are specified precisely enough to reimplement in Sections 3–4 and Appendix A (full pseudocode); dataset splits, prompt templates, and every hyperparameter used are listed in Section 5 and Appendix L. All reported numbers use a fixed, documented random seed per (method, k) cell, with multi-seed results (5 seeds) reported with 95% confidence intervals throughout rather than single-point estimates. The DAMP baseline's reimplementation from the original paper's equations, including every place it deviates from the published method because no public code was available, is documented in full in Appendix C. We intend to release the full experiment pipeline (all data-processing and analysis scripts, exact seeds, and hyperparameter configs) as an anonymized code repository in the camera-ready or supplementary material.

REFERENCES

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *European Conference on Computer Vision (ECCV)*, volume 8694, pp. 446–461. Springer, 2014. doi: 10.1007/978-3-319-10599-4_29.

Aghyad Deeb and Fabien Roger. Do unlearning methods remove information from language model weights? *arXiv preprint arXiv:2410.08827*, 2024.

Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pp. 1322–1333, 2015.

Yunusa Haruna, Adamu Lawan, Ibrahim Haruna Abdulhamid, Hamza Mohammed Dauda, Jiaquan Zhang, Chaoning Zhang, and Shamsuddeen Hassan Muhammad. Bias redistribution in visual machine unlearning: Does forgetting one group harm another? *arXiv preprint arXiv:2604.08111*, 2026.

Arman Hatami, Romina Aalishah, and Ilya E. Monosov. Class unlearning via depth-aware removal of forget-specific directions. *arXiv preprint arXiv:2604.15166*, 2026.

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

Kimmo Kärkkäinen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1548–1558, 2021.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 3730–3738, 2015.

Aengus Lynch, Phillip Guo, Aidan Ewart, Stephen Casper, and Dylan Hadfield-Menell. Eight methods to evaluate robust unlearning in llms. *arXiv preprint arXiv:2402.16835*, 2024.

Ashish Mishra, Tarun Kumar, Gyanaranjan Nayak, Arpit Shah, Suparna Bhattacharya, and Martin Foltin. Erasing clip memories: Non-destructive, data-free zero-shot class unlearning in clip models. *arXiv preprint arXiv:2512.14137*, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139, pp. 8748–8763. PMLR, 2021.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7237–7256, 2020. doi: 10.18653/v1/2020.acl-main.647.

Lijun Sheng, Jian Liang, Ran He, Zilei Wang, and Tieniu Tan. The illusion of progress? a critical look at test-time adaptation for vision-language models. *arXiv preprint arXiv:2506.24000*, 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 (S&P)*, pp. 3–18. IEEE, 2017.

Xiaoyu Xu, Xiang Yue, Yang Liu, Qingqing Ye, Huadi Zheng, Peizhao Hu, Minxin Du, and Haibo Hu. Unlearning isn't deletion: Investigating reversibility of machine unlearning in llms. *arXiv preprint arXiv:2505.16831*, 2025.

A Ensemble Subspace Projection: Full Pseudocode

See Section 4 for the motivation and Section 5 for hyperparameters (B=30, $K \in \{8, 16\}$).

B Full EuroSAT and Food101 Results

Full per-method breakdown for the two datasets summarized in Section 6.1; see Table 1 for the CIFAR-100 comparison and its column definitions (same protocol, same five methods where applicable).

Algorithm 1 Ensemble Subspace Projection

Input: training pool, target class, subspace size K, number of bootstrap fits B
for $b = 1$ to B **do**
    Sample a random negative subset (size $\leq 10\times$ positives) and a regularization strength $C_b \in \{0.01, 0.1, 1, 10\}$
    Fit logistic regression $\Rightarrow$ weight vector w_b, normalize
end for
Stack $\{w_b\}_{b=1}^{B}$ into matrix $W \in \mathbb{R}^{B\times d}$
$\{u_1, \ldots, u_K\} \leftarrow$ top-K right singular vectors of W (SVD)
Return $\{u_1, \ldots, u_K\}$, to be removed via Eq. 1

Table 5: EuroSAT (top) and Food101 (bottom), averaged over all target classes. k=5/k=50 are 5-seed mean ± 95% CI over the shot-selection seed.

Method	#dirs	Recall	Acc	k=5	k=50
EuroSAT (10 classes)					
Naive	1.0	0.132	0.312	0.30±.07	0.51±.04
NSP-text	1.0	0.001	**0.403**	0.83±.06	0.92±.01
INLP	7.4	0.106	0.303	0.40±.09	0.66±.03
ESP (ours)	8.0	0.095	0.260	0.31±.08	0.57±.04
DAMP	1	0.185	0.384	**0.72±.10**	**0.90±.01**
Food101 (10 target classes)					
Naive	1.0	0.004	**0.825**	0.22±.07	0.60±.06
NSP-text	1.0	0.016	0.781	0.92±.03	0.97±.01
INLP	8.0	0.000	0.805	0.24±.06	0.63±.04
ESP (ours)	8.0	0.002	0.796	0.23±.06	0.61±.05
DAMP	1	0.768	0.843	**0.93±.03**	**0.98±.01**

C DAMP Baseline: Full Method and Reimplementation Notes

Hatami et al. (2026) propose DAMP for *multi-stage* architectures (a CNN/ResNet/ViT with L editable stages, unlearned mid-training via weight surgery at each stage, deep to shallow). Their Eqs. 1–13 are fully specified (the paper states explicitly that its scaling rule introduces no tunable hyperparameters), but the method has no public code as of 2026-09-16 (`github.com/armanhtm/DAMP`: "Will be published soon!", checked repeatedly since 2026-09-14) and was never evaluated on a frozen embedding space. We adapt it to our setting – a single frozen CLIP embedding with no trainable next operator, i.e. the L=1 degenerate case of their framework – as follows.

Direction. For retain classes $c \in \mathcal{C}_r$ and target (forget) class f, compute prototypes μ_c, μ_f (class-mean embeddings). Stack retain prototypes as columns of $R \in \mathbb{R}^{d\times|\mathcal{C}_r|}$ and compute the retain-span residual $d_f = \mu_f - RR^{+}\mu_f$ (R^{+}: Moore-Penrose pseudoinverse, so RR^{+} is the orthogonal projector onto $\mathrm{span}(R)$), then normalize $q_f = d_f/\|d_f\|_2$. This differs from naive's direction (raw μ_f minus the mean of *all* other classes combined): it removes only the part of μ_f not already explainable by the *individual* retain prototypes, and empirically has cosine similarity only 0.48–0.59 with naive's direction on CIFAR-100 – a genuinely different construction, not naive in disguise.

Strength. DAMP's per-stage coefficient is $\alpha_\ell = \alpha_\ell^{\text{probe}} \cdot \alpha_\ell^{\text{depth}}$, with $\alpha_\ell^{\text{depth}} = \ell/L = 1$ for our single stage and $\alpha^{\text{probe}} = \mathrm{clip}(2a - 1, 0, 1)$, a = held-out accuracy of a logistic-regression probe separating forget-vs-retain on an 80/20 split of the training pool (a calibration step, not re-seeded across the outer attack seeds, matching how every other method's direction-finding in this paper is also seed-independent). On CLIP embeddings a is typically 0.92–0.995, giving $\alpha \approx 0.94$ on average.

Application. DAMP's original update edits a weight matrix, $W' = W(I - \alpha\tilde{Q}\tilde{Q}^{\top})$; for a single frozen stage with no such matrix, the equivalent operation on an embedding z is $z' = z - \alpha(q_f^{\top} z)q_f$, renormalized to unit ℓ_2 norm to match every other method's convention in this project (needed

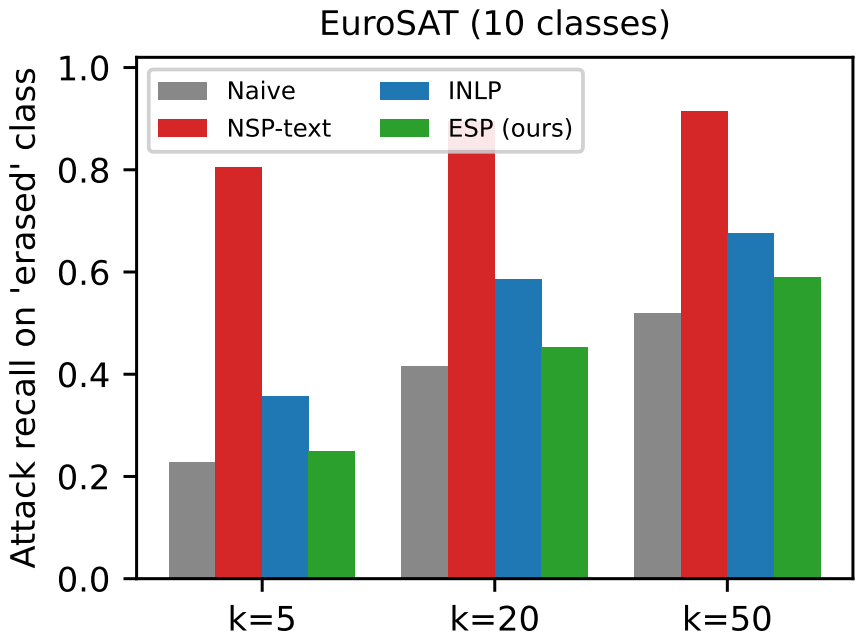

Figure 4: Attack recall on EuroSAT. NSP-text remains the most recoverable method, replicating the CIFAR-100 finding across a visually very different domain (satellite imagery vs. natural images).

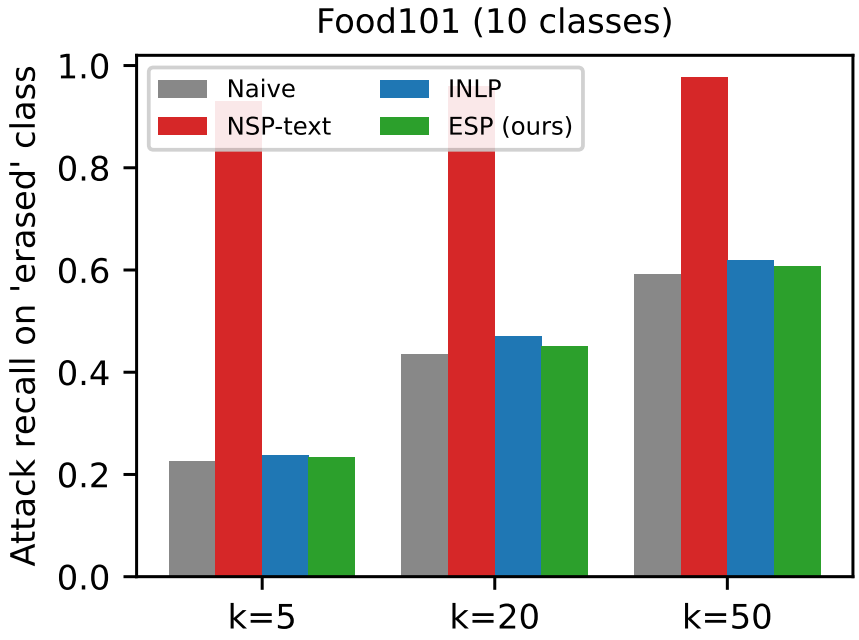

Figure 5: Attack recall on Food101 – the highest NSP-text recoverability of all three datasets tested (97.1% at k=50, 5-seed mean).

since all five operators are compared under the same cosine-similarity zero-shot classifier). This renormalization is our adaptation, not part of DAMP's original equation.

Honest scope of this comparison. No number reported for DAMP in this paper reproduces a result from Hatami et al. (2026) – their paper contains no CLIP or zero-shot-erasure experiments to compare against. This is our reimplementation from the published equations, offered as the best-effort comparison available in the absence of code, not a verified reproduction.

D Full ViT-L/14 Backbone Results

Table 6: CIFAR-100 with CLIP ViT-L/14 (vs. ViT-B/32 numbers from Table 1/10 in parentheses), averaged over the same 10 target classes. Zero-shot baseline accuracy is 73.3% (ViT-L/14) vs. 61.7% (ViT-B/32).

Method	Recall	Acc	k=50
Naive	0.017 (0.019)	0.715 (0.601)	0.542 (0.508)
NSP-text	0.004 (0.006)	0.608 (0.518)	**0.936** (0.928)
INLP	0.000 (0.001)	0.694 (0.578)	0.570 (0.590)
ESP (ours)	0.001 (–)	0.691 (0.580)	0.547 (0.518)

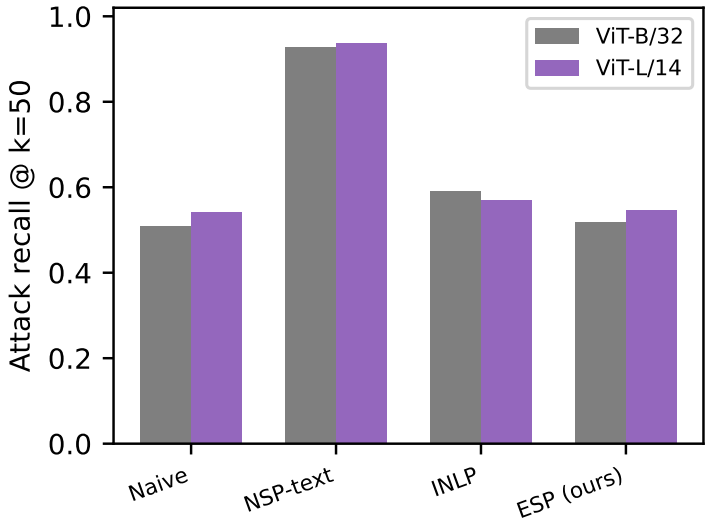

Figure 6: Attack recall at k=50, ViT-B/32 vs. ViT-L/14, CIFAR-100. Both central findings replicate: NSP-text is far more recoverable than the other three methods on both backbones.

E Leakage-Correlation Figure

F Subspace-Overlap Analysis: Full Table and Setup

For each of 10 classes per dataset (CIFAR-100, EuroSAT, Food101; cached features only, no CLIP forward passes needed), we compute: (1) ESP's and INLP's removed 8-dimensional subspace (as in Sections 4 and 6.4); (2) a held-out attacker subspace from 20 fresh bootstrap logistic-regression fits, seeded independently of ESP's own bootstrap and using a smaller negative-to-positive ratio (5:1 instead of 10:1) to ensure it is not simply a re-draw of the same procedure; (3) the top-8 principal components of the per-dataset class-mean-embedding matrix (one mean embedding per class), as a proxy for structure shared across many classes rather than specific to the erased one. Subspace overlap between two K-dimensional orthonormal bases $U_1, U_2 \in \mathbb{R}^{K\times d}$ is $\frac{1}{K}\sum_i \sigma_i(U_1 U_2^\top)^2$, the mean squared cosine of principal angles (1 if the subspaces coincide, 0 if orthogonal).

The Attacker-overlap columns show ESP $>$ INLP for every dataset mean, and in fact for all 30 individual classes tested (no exceptions) – the basis for the main-text claim that ESP's defense

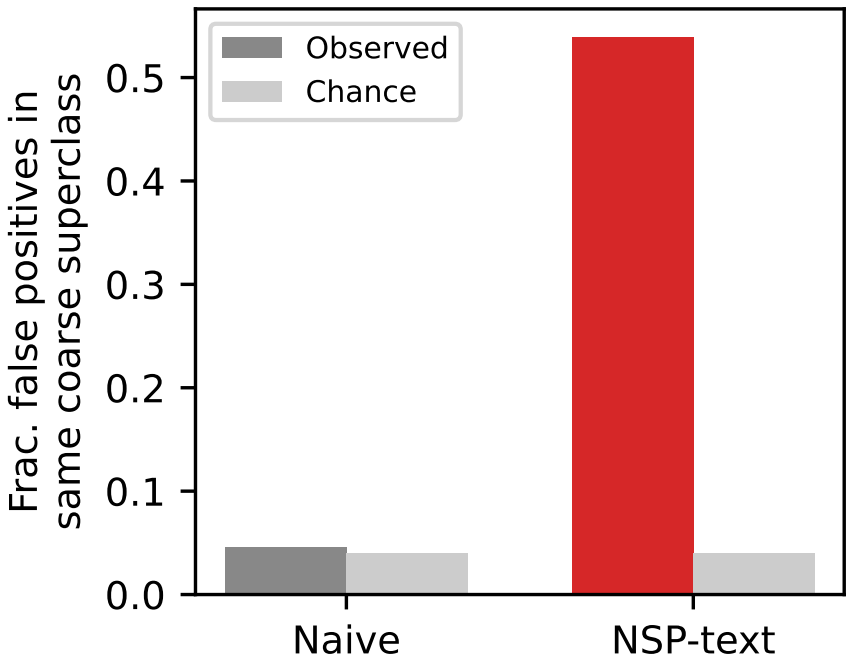

Figure 7: False positives after NSP-text erasure concentrate heavily within the erased class's coarse superclass; naive's are indistinguishable from chance. (10-class illustration; Table 2 in the main text has the full 100-class replication.)

Table 7: Subspace overlap, mean over 10 classes/dataset. Attacker = overlap with the held-out attacker subspace; Global = overlap with shared cross-class structure. Δ columns are ESP minus INLP; ΔAcc is the empirical utility difference from Table 3.

Dataset	ESP-Att.	INLP-Att.	ΔGlobal	ΔAcc
CIFAR-100	0.372	0.343	$+0.012$	$+0.002$
Food101	0.304	0.283	$+0.013$	-0.009
EuroSAT	0.535	0.449	$+0.119$	-0.043

mechanism is not a CIFAR-100-specific fluke. The ΔGlobal column (full per-dataset Global values: CIFAR-100 ESP$=0.126$/INLP$=0.114$; Food101 ESP$=0.142$/INLP$=0.129$; EuroSAT ESP$=0.356$/INLP$=0.238$) shows the gap is an order of magnitude larger on EuroSAT than on the other two datasets, tracking ΔAcc's magnitude closely across all three – see Section 6.4 for the reading we draw from this.

A simpler alternative does not explain the pattern as well. A natural competing hypothesis is that ESP's cost simply tracks how confusable a dataset's classes are on average, with no need for the subspace-overlap machinery above. We check this directly: average pairwise cosine similarity between class-mean embeddings is 0.907 (CIFAR-100), 0.911 (EuroSAT), 0.832 (Food101). This ranks EuroSAT and CIFAR-100 as almost equally "confusable" and Food101 as most distinct – it does *not* reproduce the utility-cost ordering (EuroSAT $\gg$ Food101 $>$ CIFAR-100): CIFAR-100 is nearly as confusable as EuroSAT by this measure yet has the smallest ESP cost of the three, not the largest. We read this as a useful negative control rather than a competing finding: the ΔGlobal account above is not simply proxying for raw class-mean similarity, which supports it as doing real explanatory work rather than restating an obvious dataset property.

G Race-Attribute Erasure: Full Table and Setup

Setup mirrors Section 6.6 with the erased attribute switched from gender to race (FairFace's 7 categories), each race in turn as the target of a one-vs-rest erasure, using all four operators (naive, NSP-text, INLP, ESP-8) rather than just naive/NSP-text, and breaking attack recall down by the (uninvolved) gender label instead of race. Baseline zero-shot: 62.1% overall accuracy over the 7-way race task; per-target-race baseline recall ranges 35–81% (Table 8).

Table 8: Attack recall at $k{=}50$ (5-seed mean) by target race and method, and the male/female attack-recall ratio for naive/INLP/ESP-8 (also 5-seed mean – NSP-text's ratio is flat, $\leq 1.19\times$ for every race, so omitted for space). Individual seeds swing far more widely than these means; see text.

Race	Naive	NSP	INLP	ESP-8	M/F ratio (Na./IN./ESP)
East Asian	0.345	0.900	0.340	0.345	2.51 / 2.48 / 2.55
Indian	0.366	0.798	0.357	0.357	1.36 / 1.31 / 1.32
Black	0.385	0.895	0.405	0.406	1.52 / 1.34 / 1.39
White	0.297	0.841	0.321	0.313	2.79 / 2.91 / 2.86
Middle Eastern	0.343	0.814	0.349	0.360	2.26 / 2.85 / 2.50
Latino/Hispanic	0.315	0.687	0.310	0.322	1.67 / 1.72 / 1.64
Southeast Asian	0.338	0.849	0.348	0.349	1.86 / 1.90 / 1.79
Mean	0.341	0.826	0.347	0.350	2.00 / 2.07 / 2.01

Two things replicate from the gender-erasure experiment (Section 6.6), and one adds a new wrinkle – smaller than we first reported, once properly seed-averaged. First, the central recoverability finding is robust to which attribute is erased: NSP-text is again far and away the most recoverable operator (82.6% mean recall vs. 34.1–35.0% for the other three) while incurring the lowest utility cost (1.9 points of zero-shot accuracy vs. 8.1–10.3 for naive/INLP/ESP-8) – the same "looks cleanest, leaks most" pattern seen on every object-classification dataset, now shown to hold for a demographic attribute too and for a 7-way rather than binary base task. Second, ESP again fails to reduce recall meaningfully below naive/INLP (35.0% vs. 34.1/34.7%), consistent with Section 6.6's finding that ESP's attack-resistance advantage on object classes does not transfer cleanly to attribute erasure. Third: the male/female recall disparity that naive/INLP/ESP leave behind does vary by *which* race is the erasure target – mean ratios from near-parity (Indian/Black, 1.3–$1.5\times$) to ~ 2.5–$2.9\times$ (East Asian, White, Middle Eastern) – but individual seeds swing far wider, from $1.0\times$ up to $7.6\times$ (Middle Eastern, INLP) across all races/methods/seeds we ran; the previously-reported 5.9–$6.6\times$ for "White" specifically was one high-variance seed, not the typical value. NSP-text stays flat across gender for every race regardless of seed (consistent with it barely erasing anything, as in Section 6.6), so the disparity is specific to operators that actually remove the attribute – reinforcing the reading that uneven geometric encoding, not any particular operator's design, is the root cause, and that it recurs regardless of which attribute or which subgroup pairing is examined, even if its exact magnitude is noisier than a single run reveals.

H Intermediate-Baseline Construction and Full Results

Rather than sourcing an entirely new dataset (introducing a domain-shift confound on top of the baseline-strength one we want to isolate), we construct a genuinely intermediate zero-shot baseline from *already-cached* CIFAR-100 features: a 10-class subset spanning the "reptiles" and "fish" coarse superclasses (crocodile, dinosaur, lizard, snake, turtle, aquarium_fish, flatfish, ray, shark, trout), each notoriously easy for CLIP to confuse with another member of the same pair of superclasses. Restricting the zero-shot classifier to just these 10 classes (the same recipe used for EuroSAT/Food101, rather than CIFAR-100's original 100-way classifier) gives a 52.4% baseline – between EuroSAT's 40.1% and CIFAR-100's 61.7% – while holding domain, backbone, and images fixed, so only baseline strength varies. We then repeat the full naive/NSP-text/INLP/ESP-8 comparison from Section 6.4 on this subset.

Two things stand out beyond the main-text summary (Section 6.2). First, NSP-text's recoverability advantage over the other three operators is if anything *larger* here (mean 0.824 vs. 0.397–0.401, roughly a $2\times$ gap) than on any of the three main datasets – the central finding is not merely robust to

Table 9: Intermediate-baseline (52.4%) subset: attack recall at k=50 by target class and method.

Class	Naive	NSP	INLP	ESP-8
Crocodile	0.420	0.830	0.450	0.440
Dinosaur	0.390	0.910	0.420	0.420
Lizard	0.310	0.750	0.350	0.360
Snake	0.370	0.930	0.410	0.440
Turtle	0.400	0.780	0.360	0.360
Aquarium fish	0.400	0.870	0.300	0.360
Flatfish	0.430	0.830	0.430	0.370
Ray	0.420	0.770	0.390	0.370
Shark	0.390	0.830	0.380	0.430
Trout	0.440	0.740	0.520	0.460
Mean	0.397	0.824	0.401	0.401
Utility cost	+0.062	+0.139	+0.101	+0.132

this construction, it sharpens under it. Second, INLP and ESP-8 do *not* outperform naive here (mean recall 0.401 for both vs. naive's 0.397) – within the noise of a single 10-class/single-seed run, but notably the only one of our four test beds where multi-direction removal shows no advantage at all, even a wash. We read this as consistent with confusable classes sharing more of their discriminative structure with each other: removing extra directions may cut into that shared structure without adding much attack-resistance specific to the target class, echoing the subspace-overlap mechanism in Appendix F.

I ESP K/B Ablation: Full Table and Setup

Table 10: Defense comparison on CIFAR-100 (10 classes). IESP and ESP-16 are ablations of ESP-8; see Table 3 in the main text for the equal-budget (K=8) cross-dataset comparison.

Method	#dirs	Acc	k=5	k=20	k=50
INLP	8.0	0.578	0.291	0.493	0.590
ESP-8 (ours)	8.0	**0.580**	**0.225**	**0.425**	**0.518**
ESP-16	16.0	0.547	0.225	0.422	0.530
IESP	10.4	0.557	0.220	0.402	0.510

CPU-only, reusing cached features for all three datasets, same 10 target classes as the main comparison. K-ablation fixes B=30 and sweeps $K \in \{4, 8, 12, 16, 24\}$; B-ablation fixes K=8 and sweeps $B \in \{10, 30, 50, 100\}$. Each cell is the mean over 10 classes of attack recall at k=50 (lower = more attack-resistant) and utility cost (baseline minus post-erasure zero-shot accuracy; lower = cheaper).

Table 11: K-ablation (B=30): attack recall@50 / utility cost.

Dataset	K=4	K=8	K=12	K=16	K=24
CIFAR-100	.515/.032	.518/.037	.526/.058	.530/.070	.512/.098
EuroSAT	.585/.107	.590/.141	.578/.156	.577/.151	.565/.168
Food101	.602/.033	.608/.047	.612/.051	.610/.061	.594/.103

Two patterns hold across all three datasets. First, attack recall is *not* monotonically decreasing in K – it is flat (CIFAR-100), peaks at K=8 then declines (EuroSAT), or peaks near K=12 then drops (Food101) – while utility cost rises steadily throughout, sharply so beyond K=16 (Food101's cost goes from 0.061 at K=16 to 0.103 at K=24, a jump with no matching defense benefit). More dimensions is not simply better: past a point, ESP's ensemble starts removing directions that cost utility without adding attack-resistance, consistent with the subspace-overlap mechanism in Appendix F (extra directions increasingly overlap shared, not target-specific, structure). Second, B shows diminishing returns past 30: moving from B=30 to B=100 (more than tripling the bootstrap-fitting cost) changes attack recall and utility cost by at most a few points on any dataset, and on EuroSAT B=100

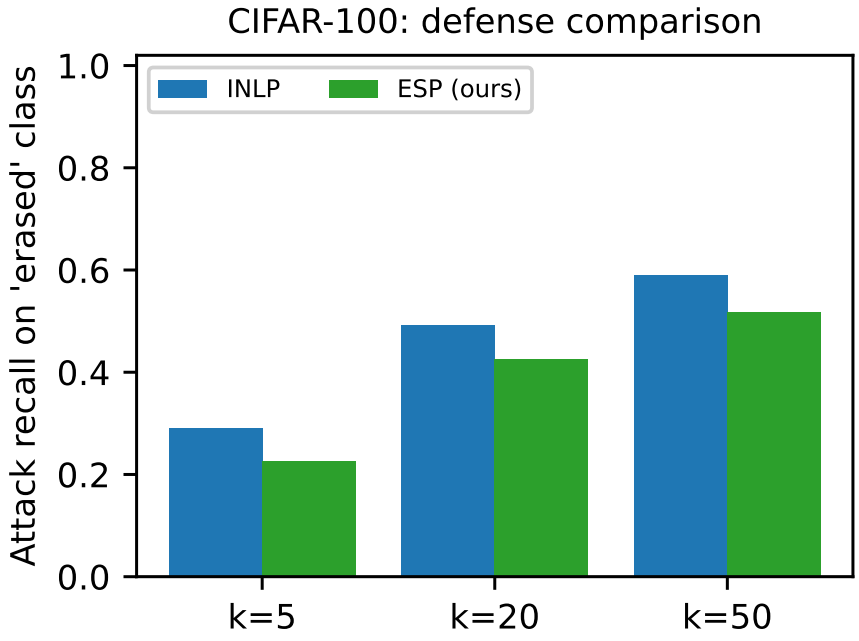

Figure 8: ESP-8 reduces attack recall relative to INLP at every shot budget, at equal or better utility, on CIFAR-100.

Table 12: B-ablation (K=8): attack recall@50 / utility cost.

Dataset	B=10	B=30	B=50	B=100
CIFAR-100	.519/.043	.518/.037	.522/.039	.525/.045
EuroSAT	.595/.165	.590/.141	.586/.140	.586/.143
Food101	.596/.044	.608/.047	.608/.049	.618/.054

is barely different from B=50. Both support K=8, B=30 – the paper's default throughout – as close to the best point on this trade-off surface we found, not an arbitrary or unexamined choice.

J Fairness-Aware ESP: Full Method and Results

Method. Plain ESP-8 (Section 4) takes the top-8 right singular vectors of the bootstrap weight matrix W, ranked purely by singular value – an objective that optimizes *average* attack-resistance, with no reason to also help the worst-off group. We instead: (1) take a larger candidate pool (top-16 singular vectors of the same W); (2) build the K=8 removed subspace greedily, one direction at a time, at each step adding whichever remaining candidate minimizes the *maximum*, over FairFace's 7 race groups, of the k=50 attack recall estimated via 3-fold cross-validation *within* Table 4's own training pool (out-of-fold predictions only, so the held-out test set is never touched by selection – unlike an earlier version of this experiment, which used a separate 70/30 fitting/selection split; nested CV lets us reuse Table 4's exact split and report directly comparable numbers). Final evaluation uses the same 5-seed attack sweep as Table 4.

This does not replicate as a positive result. An earlier version of this experiment (separate 70/30 fitting/selection split, one attack seed) suggested fairness-aware ESP-8 improved both overall recall and group disparity over plain ESP-8. Under the protocol above – the exact split used everywhere else in this paper, and the same 5-seed attack sweep – the three methods in Table 13 are statistically indistinguishable: fairness-aware ESP-8's recall, group std, and disparity ratio all fall within plain ESP-8's own seed-to-seed range. We report this as a null result rather than keep the earlier, less rigorous finding: a fairness-aware selection objective for ESP remains a reasonable idea, but our

Table 13: Naive, plain ESP-8, and fairness-aware ESP-8 on FairFace gender erasure, under Table 4's exact train/test split (k=50 recall: 5-seed mean; group std/ratio: mean over the same 5 seeds).

Method	Recall@50	Group std	Group ratio
Naive	0.311	0.110	3.94×
ESP-8 (plain)	0.297	0.114	4.49×
ESP-8 (fair)	0.296	0.113	4.09×

first attempt does not demonstrate a real effect once evaluated as carefully as the rest of this paper's claims, and we did not search further hyperparameters or objectives to try to recover a positive result post hoc.

K Full Ethical Considerations

Our attack is dual-use: it could let an adversary undo a legitimate unlearning request (e.g. right-to-be-forgotten). We judge publishing it net positive per Section 2's privacy-attack logic: a motivated adversary would likely find it independently, and withholding it only deprives defenders of a cheap detection tool. The two face datasets we use have different provenance, and we do not overstate either: FairFace (Kärkkäinen & Joo, 2021) (Sections 6.6, G) draws only Creative-Commons-licensed images from YFCC-100M, a permissive *reuse* license granted by the photographer – not informed consent from the depicted individual to appear in a face-attribute research dataset, which the license alone does not establish. CelebA (Liu et al., 2015) (Section 6.6's replication) is a research-use-licensed dataset of images scraped from the public web, with no consent mechanism from photographers or depicted individuals at all – a weaker provenance than FairFace's, which is why we use it only to check whether our core recoverability finding replicates, not for any demographic-disparity claim (it lacks a race label for that regardless). In both, gender/race probes are disposable measurement instruments, reported only as aggregate recall statistics, never used to identify an individual or offered as a usable classifier. Erasing binary gender and each FairFace race category in turn reflects the labels these datasets provide, not an endorsement of either demographic-classification framing.

L Prompt Templates

Zero-shot evaluation uses the template `"a photo of a {class}"` for CIFAR-100, `"a satellite photo of {class}"` for EuroSAT, `"a photo of {class}, a type of food"` for Food101, and `"a photo of a {male/female} person's face"` for FairFace gender. NSP-text erasure and Ensemble Subspace Projection's per-class probes use six templates: `"a photo of a {}"`, `"a {}"`, `"an image of a {}"`, `"a close-up photo of a {}"`, `"a photo of the {}"`, `"a picture of a {}"` (CIFAR-100); EuroSAT, Food101, and FairFace substitute domain-appropriate phrasing (e.g. "a satellite photo of {}", "{}, a type of food", "a photo of a {} person's face") across the same six variants.