SEPARABILITY DETERMINES FEASIBILITY, BUT NOT THE RANK: THE CEILING OF PARAMETER-SPACE UNLEARNING AND THE UNPREDICTABILITY OF ITS RANK HYPERPARAMETER

Anonymous authors
Paper under double-blind review

ABSTRACT

SEMU-style SVD-truncation unlearning hits a ceiling on CNN classifiers: over the whole range of the rank hyperparameter γ on CIFAR-100 whole-class forgetting the selected rank stays at 1 and forget accuracy remains pinned at 98.00%, and neither full rank nor extending to all layers restores forgetting. The geometric root of this failure is a class-mean common factor shared by the decision-layer gradients: under the unified convention $G \approx (\sum_i \delta_i) \otimes \bar{h}$, forget and retain inherit the same $\bar{h}$ — on the random-subset experiment the two sides' mean features differ by a measured 0.75° — leaving rank-1 updates no selective foothold. At the same depth, activations become separable once globally centered: their AUC margin over an in-distribution null ($+0.14$ to $+0.42$ for whole classes, ≤ 0 for random subsets) yields an *ex ante* feasibility criterion — calibrated per configuration, with no cross-configuration threshold — whose verdicts stay consistent on ImageNet and ViT experiments. We further prove that the projection rank k^* has no static predictor: with activations, weights, and projection basis fixed, changing only the fine-tuning process moves k^* from 4 to beyond 10 — rank selection is necessarily a deployment-time scan; the criterion's practical value lies precisely in confining this scan cost to feasible problems.

1 INTRODUCTION

Machine unlearning lacks an *ex ante* verdict: the same method can differ by an order of magnitude across forget sets, and nothing predicts the difference.

We turn the question into a measurable geometric one. The core observation: **separability depends on which space you measure in**. At the classification layer the forget and retain gradient sums share one feature common factor, while activations at the very same depth *are* separable once global centering removes that factor (Figure 1). The former explains the ceiling of parameter-space methods; the latter yields an ex ante criterion, and the gap between them is this paper's main result.

Contributions (single problem axis: when is unlearning feasible — mechanism $\rightarrow$ criterion $\rightarrow$ boundary).

1. **Mechanism: the ceiling stems from the class-mean feature common factor, not from truncation** (§3). Forget and retain share the same class-mean feature $\bar{h}$, so the rank-1–2 updates that SEMU's (Sendera et al., 2025) procedure actually executes have no selective foothold; the gradient-side quantities (α_g^{sgn}, L_k/R_k) carry no feasibility information, hence γ tuning, full rank, and layer extension all fail — the ceiling is *measured* to be unremovable within the rank dimension (§3.3).
2. **Criterion: globally centered activations give an ex ante verdict via the AUC margin over an in-distribution null** (§4). Whole classes $+0.14$–$+0.42$, random subsets ≤ 0; all configurations of Tables 6/11 agree with the verdict (§4.2); calibration is once per configuration — no cross-configuration threshold exists (§4.2, §7).

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.

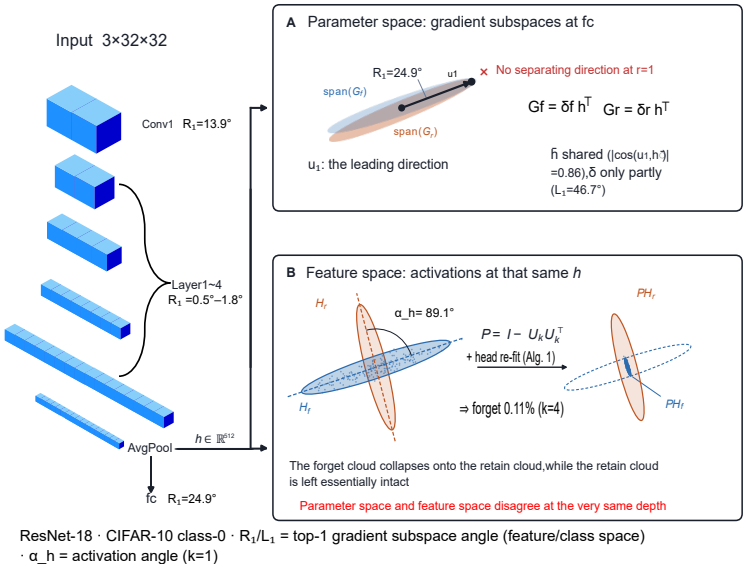

Figure 1: Same depth, two spaces (operational order: Algorithm 1). **A**: at the classifier the forget and retain gradient sums share the top feature direction ($R_1^{(\mathrm{fc})} = 24.9°$: shared $\bar{h}$, partial δ alignment, $L_1 = 46.7°$) — rank 1 has no separating direction. **B**: at the same depth activations sit $\alpha_h = 89.1°$ apart against the $\approx 89.8°$ null; P_k removes the leading directions — the forget cloud collapses onto the intact retain cloud, head re-fit completes the forgetting (0.11%, $k = 4$). §3.1 convention (model/dataset in the panel footer).

Table 1: Notation.

α_g^{sgn}; L_k/R_k	signed angle of forget/retain gradient sums, $\in [0°, 180°]$; leading-k principal angles, $\in [0°, 90°]$, of the left (class-)/right (feature-space) singular subspaces of the gradient-sum matrix $G \in \mathbb{R}^{C\times D}$ (§3.1)	$k/k^*/r$; γ; τ_k	projection rank / minimum rank reaching the verdict FA / SEMU-γ rank (three distinct); evr threshold; FA threshold (Prop. §6.1; calib. F; verdict §4.2)
α_h; AUC; MIA; Ovl	angle of centered activation subspaces; energy-ROC (receiver-operating-characteristic) area; membership-inference attack rate; normalized overlap depth (A.1; MIA: §5.3, protocol D.3)	C, D; n_f, n_r; T	classes; feature dim ($= d$: 512/192); per-side sample counts ($\neq$ seed count n); fine-tuning steps (T=0/10 arms, §6.1/F)
null / p95 / margin	worst-case null in the same configuration (20 repeats $\times$ three spectral shapes) and its 95th percentile; margin $=$ AUC $-$ p95 (A.1)	FA/FA$_{\mathrm{oracle}}$	forget accuracy / retrained-oracle counterpart

$g = \delta h^\top$: per-sample gradient $=$ error factor $\delta_i \otimes$ feature h; $\bar{h}$: global mean feature; U_k, $P_k = I - U_k U_k^\top$: projection basis and projector; H_f/H_r: activation matrices; W, b: head. R@evr99: R_k at the evr $= 0.99$ rank. Energy curves $\phi(k)$, $\rho(k)$ are appendix-local (A.1). Appendices C.2, D.1, E.4, E.5, F use the older notation UA $=$ FA.

3. **Boundary: within the process class that freezes the backbone and updates only the classification head, the rank k^* has no static predictor** (§6). The criterion decides whether this scan is worth paying for; the process decides which rank the scan reaches — deployment-time scanning is irreplaceable.

Supporting findings (not listed as contributions): the $88.6°$ of ViT-Tiny is a common-factor packing regime (shared mean $\rightarrow$ class-private codebook), not a backbone-architecture name (§3.3, E.1); the step size is the only lever that can push FA down, and what it pushes down is a non-selective rewrite of both sides (§6 corollary); the criterion holds equally on CIFAR, ImageNet, and ViT-B/L ($n = 1$, band level, A.3(f)).

2 Related Work

SEMU (Sendera et al., 2025): the direct object of this paper. SEMU performs truncated SVD of each layer's gradient matrix G, updates with the retained low-rank factors, and selects the rank by an explained-variance threshold γ (§3.3 shows it degenerates to $r = 1$ over the whole range). Its Theorem 4.1 asserts truncated SVD is the **best Frobenius approximation of** G — *about approximating G, silent on whether the subspace is fit for forgetting*: an update forgets only if it acts more strongly on forget than retain data. Our paper measures that gap: (i) the family's rank-1 updates land on the feature common factor shared by both sides (§3.2; *form* decides nothing, the direction does, §5.4 arms B/C); (ii) truncation level is ablated in full (§3.3, B.1); (iii) approximation optimality does not transmit to forgetting optimality.

Method map, and where each family stands relative to our criterion (per-method mechanism in C.2). The *projection-plugin family* (LOFT (Fang et al., 2026), POUR (Le et al., 2026), and, beyond vision, global-toxic-subspace suppression (Duan et al., 2026)): our relation is diagnostic, not competitive — we propose no operator, we explain why same-family parameter-space operations fail. The *linear-erasure family* (INLP (Ravfogel et al., 2020), LEACE (Belrose et al., 2023)): **LEACE strictly dominates our operator on closed-form linear erasure** — the direct reason the operator is not claimed as a contribution. *Continual-learning orthogonalization* (GPM (Saha et al., 2021)): the same machinery run in reverse to *cause* forgetting — direction selection decides the use, separability decides feasibility. *Exact and certified unlearning* (Cao & Yang, 2015; Bourtoule et al., 2021; Guo et al., 2020): our $|\mathrm{FA} - \mathrm{FA}_{\mathrm{oracle}}|$ convention (§3.1) is the empirical counterpart of certified indistinguishability, deliberately **without** ε, δ guarantees (end of §5.4). *Failure mechanisms and reweighting* (Mavrothalassitis et al., 2025; Wu et al., 2022): our increment grounds "separability" in a specific layer and space.

On the evaluation side, two contemporaneous works: output-level metrics are a systematic overestimate of unlearning success, and auditing must take a retrained model as the representation-space reference (Yong et al., 2026); the error of auditing via membership-inference attacks (MIA) is statistically unobservable, so a missed attack cannot serve as evidence of forgetting (Sun et al., 2026) — consistent with the reading limits we impose on MIA in §5.3.

3 The Parameter-Space Ceiling

3.1 Setup and measurement convention

CIFAR-10/100 with ResNet-18 (feature dimension $d = 512$ before the classifier) and ViT-Tiny ($d = 192$, test 85.8%); whole-class and random-subset forget scenarios. Symbols: Table 1..

Reference. All results are scored against a retrained oracle; **FA** (forget accuracy) denotes accuracy on the forget set, subscript oracle its value on the retrained model (Appendices C.2, D.1, E.4, E.5 and F use the older spelling UA, synonymous). The oracle is always retrained **from random initialization** on D_r for 182 epochs, never a fine-tune of the original (epoch-0 training accuracy 27.99% against the original's $\approx 94\%$ rules out weight loading). The three whole-class oracles sit at FA 0.00%, retain 100.00%, test 94.18–95.11%; retain being saturated, cross-method comparison uses the test column. The structure — FA exactly 0, retain saturated, MIA 1.000 — reproduces on an independently retrained ResNet-34 oracle (test 92.19%; C.1): "0.00% is a match, not an overshoot" does not rest on one model.

Three conventions that change conclusions. *First* (centering): the projection basis is estimated under the *pooled* mean of both sides — computed once from the pooled activations and subtracted from all activations, at estimation and inference alike, so the centering reference does not vary with the forget set. Switching to forget-set self-centering erases exactly the class-mean offset itself and the verdict flips with it (§4.3; mechanism in §5.2); our paper fixes the pooled convention throughout. *Second* (holdout and sample size): separability measures hold out 50% of the forget set — the top-k directions are fit on one half (in-sample) and the energy curves are scored on the other (out-of-sample; no-holdout inflation to AUC 0.58: A.1). Activations use 4096 samples per side; configurations with $n_f < d$ (a CIFAR-100 class: $450 < 512$) leave the directions underdetermined and are flagged, never read against full-sample rows. *Third* (gradient-side reporting): decision-layer geometry is reported *only* through the two quantities — α_g^{sgn} and L_k/R_k (k additionally under

Table 2: Decision-layer gradient geometry under the unified §3.1 convention; α_g^{sgn}: signed sum-angle. Full column sets and the secondary experiments: Table 10; last column: R@evr99 (rank chosen by the SEMU energy rule, evr $= 0.99$). †: conv rows use SEMU's per-input-channel mean proxy (chan-mean); not read against the 3D per-channel convention (baselines and floors of the 9-dimensional right space: A.3h).

Experiment	α_g^{sgn} (archived)	R_1	R@evr99
C10 whole class · r18 fc	126.0°	24.9°	$\mathbf{5.1}^\circ$ (k=2)
C10 whole class · r18 trunk layer1–4†	99.1–126.6°	$\mathbf{0.5}^\circ$–$\mathbf{1.8}^\circ$	$\leq 1.8^\circ$ (k=1)
ViT-Tiny · C10 head (from scratch)	96.6°	$\mathbf{88.6}^\circ$	88.6° (k=1)

the SEMU energy rule) — because for $C \leq D$ heads the full-rank left singular subspaces of both gradient sums span $\mathbb{R}^C$ and the full-rank principal angle is *identically* 0° independent of the data (random-matrix control 0.000°/0.006°): it cannot carry separability information. Both quantities are read against their own baselines: the right-space dimension varies with the layout ($\mathbb{R}^D$ at the head, only $k_h k_w = 9$ under the trunk chan-mean layout), so absolute angles never compare cross-layout (baselines: A.3h). Measurement data must also be augmentation-free and fixed-order, with per-sample sums (cross-entropy reduction='sum'; audit chain: A.3(f)(g)); all CIFAR-side values in Table 2, Table 12, Tables 13 and 14 are under these final conventions.

3.2 Gradient geometry: structural readings

Reading (i): α_g^{sgn} has no stable discriminating power. It separates neither architectures nor feasibility and moves with the training-sufficiency and forget-composition settings rather than with structure, so it cannot serve as a cross-setting comparison quantity (sweep and small-sample-artifact exclusion: Table 12 in C.1, and A.3(d)(e)).

Reading (ii): the coincidence is strongest in the trunk — exactly at SEMU's working rank. Following a clear layer-wise trend, trunk convolutions give $R_1 = 0.5^\circ$–1.8°, *inside* both the random-reslice null band (0.9°–2.1°) and the within-class half-split floor (0.5°–1.6°): under rank-1 saturation of the 9-dimensional chan-mean right space, "sharing one learned semantic direction" is indistinguishable from "any two subsets sharing a model-pinned direction". Reproduces on C100 and r34; fc heads span 24.9°–58.2°. Mechanism: the class-mean feature common factor, $G \approx (\sum_i \delta_i) \otimes \bar{h}$ — forget and retain share the same $\bar{h}$, a *global*-mean feature, so forget composition does not move the dominant direction. This predicts SEMU's total non-forgetting on random subsets (FA $\geq 98\%$; SEMU-convention readings: §3.3 and the ResNet-34 recomputation table in §C.1) — exactly what a non-selective update does under a shared common factor (A.3h(iv)). SEMU updates *all* linear and conv layers, and $\gamma \in [0.50, 0.99]$ degenerates to rank 1–2 (§3.3), so its conv path lands inside the trunk band: **even a separable 100-class head gives truncated updates no selective foothold in the trunk** (from-scratch ViT cross-check: §3.3).

3.3 The ceiling is unremovable within the rank dimension

The three rank-side routes fail **as a group**. **Truncation is not the bottleneck:** on CIFAR-10 whole class, $\gamma = 0.95$ ($r = 1$) gives FA 13.78% and full rank ($r = 10$, all classifier directions) 13.44% — no truncation, still no forgetting. γ **is not an interpretable factor:** the gradient spectrum is extremely concentrated at its head ($\sigma_1/\sigma_2 \approx 13\times$ on C10, $\approx 50\times$ on C100; this is the pre-forgetting *gradient* spectrum, not to be read against D.1's activation spectra (3.38/5.62); source: B.1, reproduction audit: C.1c), so on CIFAR-100 whole classes *every* $\gamma \in [0.50, 0.99]$ gives $r = 1$ with FA pinned at 98.00%, while random subsets of the same data give $r = 36$ at $\gamma = 0.95$ — the map is data-decided and differs by an order of magnitude across scenarios. **Layer extension has vanishing returns:** (CIFAR-10 class-0) classifier-only $20.87\% \rightarrow$ all 21 layers 20.07%; random subsets stay $\approx 99.4\%$. **This is not a shortage of tuning skill — the barrier lies in the rank dimension itself:** within the head-only class the rank admits no static predictor (§6); step size is the one remaining degree of freedom (corollary in §6).

ViT cross-validation: gradient separability is not a sufficient condition. From-scratch ViT-Tiny pushes gradient-side separability to its ceiling ($R_1 = 88.6^\circ$ against a within-class floor of only 2.3°) yet SEMU moves $\mathrm{FA}_{\mathrm{train}}$ only $96.32\% \rightarrow 89.34\%$ (ideal reference 0.00%; test segment of the same run $86.7\% \rightarrow 76.9$–77.1%, B.1); at full rank (truncation damage exactly 0) the outcome is the **worst** of four settings, 89.86% — removing it even slightly backfires.

Exception geometry: a packing boundary of the common factor, not an architecture name. The $88°$ ceiling appears only for from-scratch $\times$ Tiny after sufficient training (73% still in the CNN band; an independent seed-7 run reproduces 87–$90°$ at $\geq 85\%$). The same Tiny recipe with from-scratch ViT-B reaches $R_1 = 39.6°$ at test 86.5%, matching pretrained Base ($39.1°$) — not a ViT-family or no-pretraining generic. The two layers separate: same-architecture pretraining pulls the mean-feature angle $120° \to 56°$ and $|\cos(u_1, \bar{h})|$ (u_1: the leading feature-space direction of the forget-side gradient sum) $0.027 \to 0.66$, while the class cone / R_1 only reach $73°$. Per-class head gradients are rank-1 codewords on both CNN and ViT; what changes is the cone opening angle, not a low-rank shared-subspace phase transition. SEMU therefore still fails on Tiny not because the common factor vanished, but because it became class-private μ_c instead of a shared $\bar{h}$ — the update still has no forget/retain selectivity (exclusion chain and instruments: E.1).

4 Feature Space: Separability and the Ex Ante Criterion

4.1 Why collinear gradients do not imply collinear activations

The classifier gradient $g = \delta h^\top$ is an outer product of two factors, and §3.2 Reading (ii) shows the gradient-side inseparability is dominated by the **class-mean feature common factor** $\bar{h}$: $G \approx (\sum_i \delta_i) \otimes \bar{h}$, both forget and retain sides sharing the same $\bar{h}$, with SEMU's rank-1 updates landing on it. The activation criterion, under **global-mean centering**, precisely removes this common factor: the centered activation space retains class-specific structure, and the criterion measures there. Inseparability in parameter space and separability in feature space at the very same depth are thus **two faces of one fact: whether anything is separable depends on whether the common factor has been removed**. When the forget mean is already orthogonal to global $\bar{h}$ (from-scratch Tiny), global-mean centering still yields a feasible criterion and SEMU still fails — the geometric form of "separability is not sufficient." This also previews §4.3: centering on the forget set's own mean erases the class-specific offsets along with it, and the verdict flips.

4.2 Separability is reported as a margin over the null

We measure the activation principal angle α_h and the AUC of the energy ROC at the projection site — after average pooling on ResNet, the CLS token on ViT; definitions in A.1.

Both α_h and the AUC are finite-sample statistics: absolute values are not comparable across settings and a fixed 0.5 is not a usable baseline for all scenarios — everything is reported relative to the null, and the decision quantity is always the margin $=$ AUC $-$ null p95 (the p95 is the worst-case null over 20 repeats $\times$ three spectral-decay shapes under the same configuration, measured by splitting an i.i.d. sample in half and running the full pipeline; A.1). The p95 is used only within its own configuration (here 0.5080 for the CIFAR-10 $(4096, 4096, 512)$ cell; the size series inherits it under B.2's "AUC insensitive to n_f" license; per-scenario nulls: Table 6). On the angle side: **the no-signal anchor of α_h is $\approx 89.8°$, not $0°$**, so a single absolute angle carries no separability conclusion; separability enters separately through the normalized overlap depth Ovl (A.1). Figure 2 gives the full decay of α_h with k across the four scenarios (whole-class curves at null magnitude for $k = 1$, the random subset hugging $0°$ from the start); MIA is post-unlearning, §5.3.

"Margin > 0 as a feasibility criterion holds on all tested whole-class and random-subset scenarios" (no exception in Tables 6/11); the boundary-band rows newly measured (E.4) are instead decided by per-configuration flip intervals; the usage protocol closes this subsection. The criterion is computed by the model holder: its inputs (H_f, H_r) are data the screened operator already holds, adding no access requirement; one forward pass plus an SVD at request time leaves only scalars to retain (null p95, margin) — raw activations can be disposed of with D_f. Neither gradient-side quantity is a feasibility criterion: α_g^{sgn} has no consistent mapping onto feasibility (§3.2 Reading (i), all four scenarios), and R_1's feasible and infeasible sides overlap completely, composition-independent (§3.2 Reading (ii)) — gradient geometry answers mechanism, not feasibility. α_h and the AUC margin make the distinction (Table 6; A.1): positive for every whole class, ≤ 0 for the random subset — **feature space makes the distinction that parameter space cannot.** Ovl gives the same ordering (0.007/0.042/0.291/0.910) without taking a column.

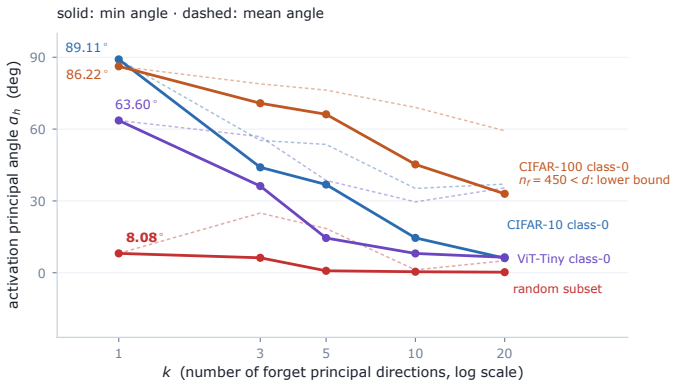

Figure 2: Activation angle α_h versus truncation rank k, four headline scenarios, against their in-distribution nulls. Whole-class curves start at null magnitude and decay only with k; the random subset sits near 0° at all k — no rank separates it.

Random subsets: four sizes, one consistent rejection. Across $n_f = 500/1000/2000/4500$ the AUCs span only 0.008 (0.4952–0.5029), all under the same null p95 0.5080 (margins -0.013 to -0.005; full per-size table, independent-draw difference, and synchronized-collapse detail: Table 11) — **AUC is insensitive to forget-set size**. Even at the best rank $k = 1$, FA is still 98.60–99.36%: infeasibility comes from forget-set **composition** (uniform across classes), not **size** — deleting 500 samples (1.1%) already leaves no usable k.

What follows is a within-configuration usage protocol, not a universal threshold. The ≈ 0.15 zero-observation void ($+0.140/-0.005$) behind the original one-sided rule (skip at margin ≤ 0) has been **filled** by the composition-axis dilution experiment (E.4, 2×2 four arms; verdict $\tau_k = 0.5\%$, retain drop ≤ 1.0pp): the four arms' half-open flip-calibration intervals (upper end of the infeasible side, lower end of the feasible side) have *empty intersection* (per-arm endpoints: E.4 R1/4; trigger 0.10 withdrawn: $18/22$ feasible grids < 0.10). **Usage (all within configuration). (i) Calibrate:** run the null battery once per new configuration (A.1, zero training), then reuse; **(ii) Decide:** margin $\leq -0.03 \rightarrow$ scan exempted (in-domain at worst under-scan — feasible-row floor -0.0188, never mis-skipped; no out-domain guarantee); margin $> +0.05 \rightarrow$ the scan is worth paying for, but **the rank comes only from the scan of the actual unlearning run** (§6 proposition: no static predictor can replace this act). **(iii) Boundary band** $(-0.03, +0.05]$**:** the no-verdict region, ranking scan priority only (the true flip point is unknown pre-deployment and drifts with the arm); direction — within a fixed regime, the more coherent the forget composition, the lower the flip point. **Scope:** the criterion constrains only operators that depend on the forget activation's principal directions (§6.2); verdicts for the other families lie outside its output surface.

4.3 The centering convention decides whether the criterion holds

This consistency relies on estimating the basis under the **global mean**: self-centering erases the class-mean offset and flips the verdict — on ViT-Tiny, a **false negative** (AUC $0.6516 \rightarrow 0.4537$, below random, an *actually solvable* problem rejected; to reach FA 0.00% it takes $k = 2$ under self-centering vs $k = 1$ globally); on ResNet/CIFAR-10 it leaves FA 94.15% ($k = 3$) vs global 1.00% (an $\approx 94\times$ efficacy gap; an observation, not a universal law). **We therefore use global-mean centering throughout**; Table 7 data and k grids: A.2; ablation reversal: §5.2.

5 Projection as a Diagnostic Tool

5.1 Method and positioning

Collect forget activations H_f at the projection site, take their top-k principal directions U_k under global-mean centering, build $P_k = I - U_k U_k^\top$ in front of the classifier, then fine-tune the head on the retain set. The full procedure (the backbone trained *before* the request) is given in Algorithm 1. **Positioning.** It claims no new method (§2); its role is a **discriminating tool**: if a **linear** feature-space

operation completes the forgetting while parameter-space truncated SVD — equally linear, same depth — cannot, the root cause localizes to directional separability, not capacity or expressiveness.

5.2 Ablation: what actually completes the forgetting

Arms (head-level interventions; SEMU's family-internal ones live in §3.3/§6 under different letters). All share one checkpoint, retain set, and fine-tuning objective/epochs on CIFAR-10 whole class. **A**: projection only, head **not** fine-tuned (does projection alone forget?). **B**: **no** projection, head fine-tuned on retain only (does re-fitting alone forget? k-free, one $k = 4$ row). **C**: random projection of the **same rank** plus fine-tuning (does the direction matter?). Full and C use $n = 3$ forgetting-*behavior* seeds (split and checkpoint fixed; the seed varies only the fine-tuning stream); A is a single reading without fine-tuning. No significance test is added — a t-test on three points has almost no power — so effect size carries the argument: at $n = 3$, mean$\pm$std bounds every seed reading within $\pm 1.42\times$ std of the mean, the worst-case bounds of the $k = 3$ and $k = 4$ columns (0.76% vs. 0.17%) do not overlap and $k^* = 4$ is seed-stable; the k=4 verdict gap is ≈ 99.6 percentage points (per-seed raw values: E.2).

Table 3: Ablation arms (forget accuracy FA %, CIFAR-10 whole class, global centering — the unified convention). B is convention-invariant; the self-mean panel is in A.2. The full column is mean$\pm$std over behavior seeds 0/1/2 (C: three-seed mean); $3/3$ per-seed agreement (E.2).

k	full FA	A proj-only FA	A retain	C rand FA
1	79.79 ± 0.32	97.51	99.45	99.70
3	1.00 ± 0.17	99.49	96.92	99.70
4	$\mathbf{0.11 \pm 0.04}$	**99.73**	95.65	99.70
10	0.00 ± 0.00	100.00	**12.76**	99.70
B (head-only, k=4): FA **99.71**				

Under the global convention, projection and fine-tuning are jointly necessary. A never forgets at any k and its retain collapses with k (12.76% at $k = 10$): an unfine-tuned projection is not harmless. B and C also do not forget. Forgetting happens only when the direction comes from the forget-activation principal subspace *and* is paired with retain fine-tuning. The A-retain divergence across conventions stems from one mechanism: global U_k contains the class-mean direction, along which the head reads much discriminative energy; severing it without re-fitting collapses retain (§4.3). Mechanism: on retain-only data the forgotten class never appears as a target, so cross-entropy keeps suppressing its logits — **fine-tuning is part of forgetting**; projection removes directions along which the class stays separable *while suppressed*. **Under self-mean centering the causality reverses**: at $k = 4$ full and arm A both reach FA 0.00% — projection alone completes the forgetting, fine-tuning merely repairs retain, reproducing prior work's conclusion under its own convention; **centering changes the causal mechanism, not just efficacy and the criterion (§4.3)**.

5.3 External validation: membership inference on ViT

An independent measure confirms the accuracy-based criterion: on ViT-Tiny, membership inference (Shokri et al., 2017) reaches **MIA** $= 1.000$ **at every** $k \in \{1, 3, 5, 10\}$ for the full arm, while the C random-projection arm stays at 0.199–0.208 with $\mathrm{FA}_{\mathrm{train}}$ 79.2–80.1% — directional separability does not equal forgetting; fine-tuning is a component of it (the mechanism of the Proposition). Protocol and the complete $(k, \text{arm}, \mathrm{FA}_{\mathrm{train}}, \text{MIA})$ table and its reading limits (a saturated MIA is not a privacy certificate; ViT experiments have no retrained oracle) are in **D.3**.

5.4 Boundary distinctions

Inserting a projection is not re-parameterizing the head. WP is algebraically a single linear map, but only a vanishing part of the head re-parameterization's **reachable set** can forget: B (99.71%) and C (99.70%) never reach the full method's 0.11% at $k = 4$ (Table 3) — **expressiveness** versus **reachability**.

Folding: less inspectable, not less leaky. Folding removes the module from the inference path and is numerically exact, but an adversary also holding the original published model can still recover U_k:

the exploitation bar is raised; the leak is not eliminated (threat model and reconstruction detail: C.2).

Claim convention. This section claims that the **activation subspaces** of forget and retain are separable — a readout-pathway-level claim; removal of D_f's influence from the whole model is **not** claimed (argued in E.5; a declared known cost, §7). The core argument uses only the "readout path" notion.

6 Infeasibility of the Rank Dimension

6.1 Structural barrier and the rank proposition

§3.3 leaves a pattern — γ tuning, full rank, and layer extension all fail, identically across datasets and depth. Separability is decidable from pre-unlearning quantities (§4), but **the rank cannot be statically predicted** — formalizing "static" as a typing constraint on the predictor's domain yields the definition and proposition (notation, assumptions, threshold calibration: §F; fully formal version, nontriviality, and scope: G).

Static predictor (definition). A predictor Q of k^* is **static** if its domain excludes the unlearning process A: Q may take as input any quantity determined by (H_f, H_r, W, b) alone, but no information contributed by A — process hyperparameters, the data and augmentation stream its fine-tuning consumes, random seeds. The definition **places no computational restriction on** Q: static statistics of arbitrary complexity, and any computable combination of them, are in range — the failure stems not from choosing bad criteria but from the typing constraint that *the input excludes the process*. All five static rank criteria examined here characterize only P_k and necessarily fail for this one reason (five-criterion comparison and correction attempt: D.1, Fig. 3 right).

Proposition 1 (Nonexistence of a static rank predictor)**.** *Under (i) a frozen backbone with only the classifier updated, (ii) U_k determined by H_f alone (true of the closed-form construction; false of learned projections (LOFT and the like), which are outside the scope), and (iii) a fixed threshold τ_k, there is no static predictor $Q : (H_f, H_r, W, b) \mapsto \mathbb{N}$ such that $Q(H_f, H_r, W, b) = k^*(A)$ for all unlearning processes A satisfying the three assumptions.*

Proof. Q static $\Rightarrow$ its value is independent of $A \Rightarrow k^*$ cannot depend on A; yet the two arms of §5.2 share one (H_f, H_r, W, b) and one P_k — and they are not an arbitrary hyperparameter pair: the $T = 0$ arm has the same form as POUR-P's one-shot projection (different direction source; G), and the $T = 10$ arm is the standard use of the fine-tuning family (G) — with $k^* > 10$ at $T = 0$ and $k^* = 4$ at $T = 10$: contradiction (Figure 3 left; full nontriviality argument: G).

Scope. The proposition denies only "a single static Q consistent for all processes"; it does not deny that k^* may be recoverable once A is fixed (a predictor containing A, or simulating it, costs a scan). Assumption (ii) fails for learned projections (LOFT and the like), so for them this proposition simply does not transfer (G). **Status of assumption (i):** the frozen backbone is not an incidental simplification of the proof but the condition under which the question is well posed — k^* takes as argument the threshold-crossing behavior under P_k, and P_k must be given by pre-unlearning quantities. The three-case split for unfrozen-backbone procedures and the claims left unproven are given in G; generalizing the proposition there joins learned projections as an open question. **Consequence:** deployment still needs one forget-accuracy scan — but separability decides in advance whether that scan is worth paying for (usage protocol, §4.2).

Corollary: the one remaining lever is the step size, and it is not selective. lr $1\mathrm{e}{-5} \to 1\mathrm{e}{-3}$ drives FA $14.53 \to 0.02 \to 0.00\%$ while retain falls $97.73 \to 96.47 \to 69.41\%$ and test $91.57 \to 90.32 \to 65.56\%$ (C.1c); with no separating direction, force rewrites both sides proportionally, so the cost curve of "FA $= 0$" *is* the behavioral signature of §3.2 Reading (ii) — against feat-proj's retain $+0.01$pp / test -0.04pp: not stronger or weaker on one axis, but the two consequences of whether a direction is selective. (The corollary shares the proposition's domain — see **Scope**.)

6.2 Baseline cross-check

The cross-check against published methods sits directly in the body (Table 4): one checkpoint, one forget set (CIFAR-10 class 0), original-paper default hyperparameters (untuned, none an upper

bound), $n = 3$ mean $\pm$ std (feat-proj prints its 3-seed mean; std: Table 3); only fine-tuning (FT), gradient ascent (GA), and feat-proj are tabulated; the full nine rows (BE and FT_prune included): D.2; FT_prune bimodal warning: E.2.

Table 4: Same-setting cross-check (FA/Retain/Test %; MIA rate; full nine rows: D.2).

Method	FA (%)	Retain	Test	MIA
Oracle	0.00	100.00	94.47	1.000
FT	86.14±9.04	99.97±0.02	94.82±0.10	0.139±0.090
GA	11.97±0.15	95.44±0.15	89.04±0.18	0.880±0.002
feat-proj (k=4)	**0.11**	**99.48**	**94.43**	**0.999**

The criterion's predictive force is direction-specific — it constrains only operators using the forget activations' principal directions — and both outcomes are measured: at $+0.282$ (whole class) feat-proj reaches oracle; at -0.009 (random subset — boundary band, exemption not granted) it collapses forget and retain at all k (§4.2) — prediction, not weakness.

Use. An SEMU-style pipeline whose update direction is the forget activations' principal directions queries the criterion at request time (one forward pass plus a rank-k truncated SVD; rules: final paragraph of §4.2). Fine-tuning/weight/boundary families do not consume that direction — no jurisdiction (per row, D.2); learned projections (LOFT) lie outside assumption (ii). The interface is a configuration-level go/no-go pre-screen, not a general parameter oracle; the paper's own random-subset negative rows (-0.013 to -0.005) are boundary-band rows — the scan is paid; savings concentrate deep-negative.

7 Limitations and Future Work

First: the criterion's calibration domain is narrow. The four arms' flip intervals have empty common intersection (no universal threshold); the boundary drifts with backbone, dataset and regime with no law yet; the skip rule's out-domain margin is untested. *Future work*: widen the calibration, quantify drift from composition coherence, and derive an extrapolable margin.

Second: the criterion does not exempt scanning, nor does the proposition outgrow its quantifier domain. Rank selection still requires a deployment-time scan with the actual process; non-existence covers only the backbone-frozen, head-only family; whether a static predictor exists with an unfrozen backbone is open (three cases: G).

Third: projection is a diagnostic, not complete model forgetting. Only the head readout pathway is removed (backbone frozen); folding lowers inspectability without eliminating leakage (§5.4, C.2) — a declared weakening vs. the ideal definition. Future: backbone-layer diagnosis, unidentifiable-increment directions.

Convention memo. Main-table error bars cover only forgetting-process randomness ($n = 3$, fixed splits; E.2). Scale experiments remain $n = 1$ (no d_{CLS} sweep; ImageNet validation-only, E.1): band level and direction only (A.3(f)), no magnitude extrapolation. §6.2 is not a leaderboard — defaults, none at their upper bound; remaining negative results and the falsification list: E.3.

8 Conclusion

Mechanism. The decision-layer gradient is dominated by a class-mean feature common factor shared by both sides, so low-rank updates lack exactly the forget/retain selectivity (§3.2); the ceiling cannot be lifted along the rank dimension (truncation, full rank, layer extension, §3.3), and the one remaining lever, step size, rewrites both sides non-selectively (§6). Packing regimes differ between CNNs and narrow from-scratch ViT-Tiny (shared mean vs. class-private codebook); the conclusion stands.

Criterion. Global centering plus the null-relative margin gives a configuration-level ex ante screen (deep-negative exempt, boundary ranked; §4.2); scale verdicts band-level only (A.3(f)).

Boundary. The required rank comes from scanning with the actual process — within its domain, no static predictor exists (§6); projection cuts the readout pathway, not complete forgetting (§5.4). Mechanism boundary: §3.2; measurement conventions are part of the conclusion.

REFERENCES

Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. LEACE: Perfect linear concept erasure in closed form. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Martin Bertran, Shuai Tang, Michael Kearns, Jamie Morgenstern, Aaron Roth, and Zhiwei Steven Wu. Reconstruction attacks on machine unlearning: Simple models are vulnerable. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Lucas Bourtoule, Varun Chandrasekaran, Christopher A. Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. Machine unlearning. In *IEEE Symposium on Security and Privacy (S&P)*, 2021.

Yinzhi Cao and Junfeng Yang. Towards making systems forget with machine unlearning. In *IEEE Symposium on Security and Privacy (S&P)*, 2015.

Min Chen, Zhikun Zhang, Tianhao Wang, Michael Backes, Mathias Humbert, and Yang Zhang. When machine unlearning jeopardizes privacy. In *ACM CCS*, 2021.

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

Remi Denton, Wojciech Zaremba, Joan Bruna, Yann LeCun, and Rob Fergus. Exploiting linear structure within convolutional networks for efficient evaluation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2014.

Zenghao Duan, Zhiyi Yin, Zhichao Shi, Liang Pang, Shaoling Jing, Zihe Huang, Jiayi Wu, Yu Yan, Jingcheng Deng, Huawei Shen, and Xueqi Cheng. Projecting out the malice: A global subspace approach to LLM detoxification. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2026.

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

Kun Fang, Qinghua Tao, Junxu Liu, Yaxin Xiao, Qingqing Ye, Jian Sun, and Haibo Hu. Machine unlearning in low-dimensional feature subspace. *arXiv preprint arXiv:2601.22456*, 2026.

Chuan Guo, Tom Goldstein, Awni Hannun, and Laurens van der Maaten. Certified data removal from machine learning models. In *International Conference on Machine Learning (ICML)*, 2020. PMLR 119:3832–3842.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

Jinghan Jia, Jiancheng Liu, Parikshit Ram, Yuguang Yao, Gaowen Liu, Yang Liu, Pranay Sharma, and Sijia Liu. Model sparsity can simplify machine unlearning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

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

Mengyao Ma, Shuofeng Liu, Jason Xue, Surya Nepal, and Guangdong Bai. ReTrace: Reinforcement learning-guided reconstruction attacks on machine unlearning. In *International Conference on Learning Representations (ICLR)*, 2026.

Ioannis Mavrothalassitis, Pol Puigdemont, Noam Itzhak Levi, and Volkan Cevher. Ascent fails to forget. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020.

Gobinda Saha, Isha Garg, and Kaushik Roy. Gradient projection memory for continual learning. In *International Conference on Learning Representations (ICLR)*, 2021.

Marcin Sendera, Łukasz Struski, Kamil Kśiążek, Kryspin Musiol, Jacek Tabor, and Dawid D. Rymarczyk. SEMU: Singular value decomposition for efficient machine unlearning. In *International Conference on Machine Learning (ICML)*, 2025. PMLR v267.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *IEEE Symposium on Security and Privacy (S&P)*, 2017.

Sidak Pal Singh and Dan Alistarh. WoodFisher: Efficient second-order approximation for neural network compression. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Jialong Sun, Zeming Wei, Jiaxuan Zou, Jiacheng Gong, Guanheng Wang, Chengyang Dong, Jialong Li, and Bo Liu. Statistical MIA: Rethinking membership inference attack for reliable unlearning auditing. *arXiv preprint arXiv:2602.01150*, 2026.

Ga Wu, Masoud Hashemi, and Christopher Srinivasa. PUMA: Performance unchanged model augmentation for training data removal. In *AAAI Conference on Artificial Intelligence*, 2022. DOI 10.1609/aaai.v36i8.20846.

Hao Xuan and Xingyu Li. RUB: Evaluating residual knowledge in unlearned models. In *CVPR 2026 Workshop on Machine Unlearning for Vision (MUV)*, pp. 8550–8559, 2026.

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

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations (ICLR)*, 2017.

REPRODUCIBILITY STATEMENT

Every quantity reported in the paper is computed by code shipped with the anonymized supplementary code package, whose README maps each script to the section it supports; the package contains no datasets, weights, or result files, and the result files named in the appendices are regenerated by the drivers. The measurement conventions, null-calibration procedure, and per-experiment protocols are stated in full in §3.1, §4.1–§4.2, §5.1–§5.2 and Appendix A.1/A.3, with the SEMU source entries in B.1, the arm protocol in §5.2 (full panel in A.2), the dilution study in E.4, and the formalization, threshold calibration, and complete proof in F/G.

AI USE STATEMENT

In this work, we used generative AI tools (large-language-model assistants) for the following tasks with **required disclosure**: *assist in the writing of proofs* (typesetting and organizing the formalization in Appendix F and the proof write-up in Appendix G; all mathematical content originates from the authors); *design or provide feedback on research methodology or experiments* (composition-axis dilution protocol of Appendix E.4, per-configuration null-calibration procedure, and adversarial cross-checks of reported numbers). We have not used generative AI tools for *generate synthetic data sets*, *clean and reformat data sets*, *help develop theoretical models or conceptual frameworks*, *formulate mathematical claims*, *provide critical ingredients for proving mathematical claims*, or *propose or refine hypotheses*.

Additionally, we used generative AI tools for tasks with **recommended disclosure**: *create or edit software code* (experiment drivers for the measurements, ablation arms, baselines, external validation and the dilution study; including the anonymized code release, *creation of artifacts*); *create or*

modify scientific figures; *draft parts of the research paper*; *edit the research paper to improve readability*; *suggest a structure for the research paper* (section and appendix reorganization); *summarize or analyse existing literature* and *identify relevant literature* (verification of cited entries); and *format references*. We have not used generative AI tools to *discover research topics or identify gaps*, *transcribe recordings*, or *propose a title or keywords*.

We have reviewed all AI-assisted work. AI-assisted prose was checked against the measured results and rewritten or approved line by line by the authors; AI-assisted proof write-ups were verified against the formal definitions, and the non-triviality construction was re-derived independently. AI-generated or AI-edited code was executed under the authors' supervision, and every number reported in this paper was regenerated from run logs and cross-checked by multi-seed replication and the independent-machine re-measurement audit (Appendix A.3). Literature summaries were verified against the original sources.

We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

A Measurement Convention and Audit

A.1 Measurement definitions and null control (fills §3.1, §4.2)

Energy ROC and AUC. Let $q_1 \dots q_d$ be the principal directions of H_f (the columns of U_k) under global-mean centering (descending variance). Define

$$\varphi(k) = \frac{\sum_{j \le k} \|H_f q_j\|^2}{\|H_f\|^2} \qquad \rho(k) = \frac{\sum_{j \le k} \|H_r q_j\|^2}{\|H_r\|^2} \tag{1}$$

where $\varphi(k)$ is the forget energy removed and $\rho(k)$ the retain energy damaged by the first k forget directions. Both curves share one basis (both use the forget-side directions), are monotone, and start at 0 (when $n_f < d$, φ saturates early and the curve stops before $(1, 1)$). Plotting (ρ, φ) traces a ROC-type curve whose area is the AUC. An $\mathrm{AUC} \approx 0.5$ means that removing the forget principal directions drains retain energy at the same rate as forget energy — no projection at any rank separates.

Holdout. U_k is fit on half the forget set; φ is scored on the other half. Without the holdout, φ is in-sample and ρ is out-of-sample, and by Ky Fan maximality even with H_f, H_r i.i.d. we get $\mathbb{E}[\varphi - \rho] > 0$ — that is the PCA generalization gap, not separability; the no-holdout in-distribution null can reach AUC 0.58, more than half of most margins in Table 6.

Null control. Split i.i.d. samples in half under each scenario's own (n_f, n_r, d) and run the full pipeline, 20 repeats $\times$ three spectral shapes (flat /moderate /steep), take the worst-case p95. ResNet scenarios: $n_f = n_r = 4096$, $d = 512$; CIFAR-100: $n_f = 450 < d$ (its margin should be read as an upper bound); ViT: $d = 192$. The α_h null (principal angles between the top-k principal directions of the centered H_f and H_r) is measured the same way (89.7°–90.0°), and $\mathrm{Ovl} = (\alpha_h^{\mathrm{null}} - \alpha_h)/\alpha_h^{\mathrm{null}}$ is normalized by it.

Sampling and centering. Activations are always 4096 per side (ResNet: after avg-pool; ViT: CLS); the projection basis and deployment apply the same global-mean shift (consequences in §4.3).

Table 5: Energy curves $\varphi(k)/\rho(k)$ and AUC (this table replaces a standalone energy-ROC figure). Each cell is (φ, ρ).

Scenario	k=1	k=3	k=5	k=10	curve end	AUC	null p95
CIFAR-10 whole class	$(0.693, 0.217)$	$(0.810, 0.347)$	$(0.886, 0.466)$	$(0.975, 0.685)$	reaches $(1, 1)$	0.7902	0.5080
CIFAR-100 whole class	$(0.819, 0.017)$	$(0.863, 0.041)$	$(0.882, 0.066)$	$(0.911, 0.127)$	rank-deficient, stops at $\rho = 0.82$ (rank 225)	0.9496	0.5282
CIFAR-10 random subset	$(0.156, 0.148)$	$(0.376, 0.374)$	$(0.580, 0.585)$	$(0.956, 0.957)$	reaches $(1, 1)$, hugs the diagonal	0.4986	0.5080
ViT-Tiny whole class (**self-mean**, for reference)	$(0.144, 0.095)$	$(0.317, 0.238)$	$(0.409, 0.457)$	$(0.532, 0.703)$	reaches $(1, 1)$, below the diagonal	0.4537	0.5113
ViT-Tiny whole class (**global**, this paper)	$(0.578, 0.213)$	—	—	—	only the k=1 point and AUC persisted under global convention	**0.6516**	0.5113

Reading: for whole classes, φ already far exceeds ρ at $k = 1$ (CIFAR-100 most of all: 0.819 vs 0.017 — the first direction is almost pure forget); for the random subset the two track each other throughout (gap ≤ 0.001); the two ViT rows pose the same question under different conventions — the curve-level display of §4.3's criterion flip. AUC is reported under the closing convention of Table 6; the rank-deficient CIFAR-100 row was additionally measured under an "open, observed-domain only" convention, and the two agree to three decimals (0.950) — truncation of the domain does not inflate the area.

Table 6: Separability main table (cited in §4.2; four headline scenarios; gradient columns under the §3.1 unified convention; margin = AUC − null p95).

Scenario	α_g^{sgn}	R_1	α_h	AUC	null p95	margin
CIFAR-10 whole class (ResNet-18)	126.0°	24.9°	89.11°	0.7902	0.5080	$+\mathbf{0.282}$
CIFAR-100 whole class (ResNet-18)	95.7°	54.4°	86.22°	0.9496	0.5282	$+\mathbf{0.421}$
ViT-Tiny whole class (classifier)	96.6°	88.6°	63.60°	0.6516	0.5113	$+\mathbf{0.140}$
CIFAR-10 random subset (ResNet-18)	60.1°	19.4°	$\mathbf{8.08}^\circ$	0.4986	0.5080	$-\mathbf{0.009}$

A.2 Centering convention: criterion comparison (Table 7) and full k grid

Table 7: Centering convention × criterion. Switching to the forget set's own mean erases the class-mean offset: on ViT-Tiny the criterion flips to a **false negative** ($0.6516 \to 0.4537$, rejecting an actually solvable problem), while on ResNet the efficacy at $k = 3$ drops from 1.00% to 94.15% forget accuracy (an $\approx 94\times$ observation on ResNet/CIFAR-10, **not a universal law**). The paper uses global-mean centering throughout.

Arch	global AUC	self AUC	flips?	global k^*	self k^*	k=3 FA global /self
ResNet-18 / CIFAR-10	0.7902	0.7566	no (both ≥ 0.60)	k=4 (k=3: $\mathbf{1.00}\%$)	k=4 (k=3: $\mathbf{94.15}\%$)	$\mathbf{1.00}/\mathbf{94.15}$ ($\approx 94\times$)
ViT-Tiny /CIFAR-10	$\mathbf{0.6516}$ feasible	$\mathbf{0.4537}$ reject	**yes, false negative**	$\mathbf{k=1}$ (FA 0.00, retain 96.67)	$\mathbf{k=2}$ (FA 0.00, retain 96.39)	0.00/0.00 (both fully forget)

The ViT row is decisive: under self-mean, AUC $= 0.4537$ sits below chance and the criterion rejects an **actually solvable** problem (under that convention $k = 1$ still leaves $\mathrm{FA}_{\mathrm{train}}$ 0.74%, reaching 0.00 only at $k = 2$, retain 96.39 vs baseline 96.17; under global centering AUC $= 0.6516$ and $k = 1$ already gives forget 0.00, retain 96.67). **Whether AUC serves as a criterion is thus a question about the centering convention, not about AUC.** On ResNet, self-mean leaves 94.15% forget accuracy at $k = 3$ where global centering gives 1.00% — the
$approx 94$
$times$ efficacy gap is an observation on ResNet/CIFAR-10, **not a universal law**; on ViT, centering changes almost only the correctness of the verdict, not whether forgetting succeeds. The convention also changes the causal mechanism the ablation reveals (§5.2).

Table 7 reports the criterion contrast; the last column reports only the k=3 efficacy contrast, and the body (§4.3) keeps the conclusion only. Below: same checkpoint, same k grid, centering as the only change. Both columns are **single-seed (run_seed $= 1$) readings**; the global column differs from the $n = 3$ means of §5.2/§6.2 by ≤ 0.16pp (seed-to-seed noise); on those readings the gap is $94.02/1.16 \approx 81\times$, same order of magnitude.

Centering effect (full-method FA grid, same readings as before): global centering decays smoothly (k=1: $79.73\% \to$ k=3: 1.16%); under self-mean FA sits at $\mathbf{94.02}\%$ at k=3 and falls off a cliff thereafter. The "forget information stored across layers" phase-transition narrative is therefore an artifact of the centering convention, and this paper does not adopt it.

Self-mean counterpart of the §5.2 ablation (same checkpoint and protocol as the §5.2 global panel, centering as the only change; B is convention-invariant at FA 99.71):

Table 8: Ablation arms under self-mean centering (FA %, CIFAR-10 whole class).

k	full FA	A proj-only FA	A retain
1	95.52	90.76	99.49
3	94.15	93.62	99.39
4	**0.00**	**0.00**	99.28
10	0.00	0.00	99.48

C's self-mean readings equal its global ones to the displayed digits (99.70 at every k).

Self-mean ablation panel (consolidated from §5.2). At $k = 4$: full FA 0.00 / arm A 0.00 / A retain 99.28; arm C stays flat at $\approx 99.70\%$ for every k under both conventions (the SVD direction is necessary); arm B is convention-invariant (99.71%).

A.3 Scale-up experiments and convention-audit details (cloud RTX 4090; supplying §3.2, §4.2, §7)

All experiments seed 2. Environment torch 2.2.2+cu121, timm 1.0.29.

(a) Scope of the convention-change audit. Only the blast radius of the convention change is recorded here. **Unaffected**: α_h/AUC/margin/Ovl and null controls (explicit k-truncation), conv-layer 3D per-channel principal angles, the unified-convention α_g^{sgn} readings of every experiment, and every behavioral result (§3.3, §5, §6, Table 20, B.1, C.1).

(b) ViT-B/L scale experiments (CIFAR-10 class-0; timm augreg_in21k_ft_in1k pretrained, fine-tuned $30/20$ epochs under this paper's protocol).

Experiment	ft test	α_g^{sgn}[head]	block-angle range	AUC (k^*, rule)	$\alpha_h(k{=}1)$	feat-proj $k{=}1/3$: FA /retain /test
ViT-Base (86M)	98.48%	110.18°	38.8°–110.2°	0.7781 ($k{=}3$, youden)	81.55°	$0.00/99.62, 99.61/98.70$
ViT-Large (304M)	99.06%	117.68°	69.5°–111.7°	0.8383 ($k{=}4$, youden)	89.90°	$0.00/100.00, 100.00/99.04$

Baseline: forget_train $99.4/100.0$, retain $99.46/100.00$. Zero-cost projection reproduces the ViT-Tiny shape at $53\times/35\times$ parameters; α_g^{sgn} stays high, but per §3.2, R_1 ($39.1^\circ/30.7^\circ$) is back inside the CNN interval — "the exception survives" holds only under the non-discriminative reading.

(c) ImageNet validation-only experiments (ILSVRC2012 val 50k; forget = all 50 images of Bouvier; probe class = sheepdog).

Experiment	α_g^{sgn}[fc]	AUC	null p95	margin	FA at $k{=}1$–10	retain (val) pre→post	probe pre→post
ResNet-50	92.17°	0.8154	0.5347	$+\mathbf{0.2807}$	all k: 0.00	$80.10 \to 84.93$	$88.0 \to 92.0$
ConvNeXt-Tiny	92.49°	0.8208	0.5264	$+\mathbf{0.2944}$	all k: 0.00	$81.86 \to 87.97$	$86.0 \to 90.0$

$\alpha_h(k{=}1/3/5/10)$: r50 $70.6^\circ/61.4^\circ/54.8^\circ/49.5^\circ$; cnxt $89.8^\circ/85.3^\circ/83.8^\circ/81.2^\circ$. Convention note: the α_g^{sgn} column holds unified-convention re-measured values (other columns are activation/behavioral and do not depend on the accumulation convention); the null is a Bouvier in-class $25/25$ split (a small-sample null at 50 images per class — a declared caveat); no oracle, no MIA (the 50-per-class regime; decision axis see (f)); the retain "increase" ($+4.8$pp/$+6.1$pp) is the post-projection head re-fit pulling an underfit validation retain back up, not a capability gain beyond baseline; all neighbor top-10 classes do not drop.

(d) The n_f confound verdict. Question: is ImageNet's 92° a small-n_f artifact at $n_f = 50$ or a 1000-class structural effect? On cached features, grow the forget set from the anchor class to $m = 1 \ldots 64$ classes nested ($n_f : 50 \to 3200$):

m (n_f)	1 (50)	2 (100)	4 (200)	8 (400)	16 (800)	32 (1600)	64 (3200)
r50 nested α_g^{sgn}	91.9	92.4	91.5	93.7	92.8	92.7	95.2

(a small-sample artifact would collapse toward 0° as n_f grows; measured: flat and rising with class heterogeneity — **the artifact branch is excluded**; the random arms and the retain-side symmetric control are equally flat, full curves equally flat). Convention note: the sweep curves are on the cloud experiments' batch-mean convention; the anchor $m = 1$ experiment's sum re-measurement shifts by $\leq 0.25^\circ$, so the verdict stands. Same question, same answer on the CIFAR side: C10 class-0 downsampled $5000 \to 50$ walks $145.6^\circ \to 153.8^\circ$; C100 from $500 \to 50$ walks $92.4^\circ \to 96.2^\circ$ — n_f **is no lever under either regime**; the ImageNet–CIFAR difference is a structural effect of 1000 classes /feature heterogeneity.

(e) Breed replication (the forget parameterization). Bouvier is no special case: golden retriever, boxer (dalmatian has an empty directory — skipped) $\times$ two models; probe class unchanged. All six α_g^{sgn} readings (unified re-measurement) fall inside the 91.7°–92.7° narrow band, every margin $> +0.25$ judges feasible, and $k = 1$ already gives FA $= 0.00$ with retain undiminished — **both the criterion and the conclusions are breed-independent**.

(f) Boundaries. All cloud experiments (b)–(e) are single-seed (seed 2) and lack retrained oracles; the decision axis is criterion–outcome consistency + neighbor specificity: margin > 0.25, zero-cost FA $= 0.00$ at all k, and no neighbor top-10 class drops.

(g) Second audit of the measurement convention: from "identically zero" to "reproducible". The first version of the unified script family on the CIFAR experiments still used the main pipeline's marked loader (training augmentation + shuffle + CE batch-mean). Three control experiments exposed two independent defects. **First (batch-partition dependence)**: batch-mean accumulation up-weights the trailing partial batch, and its membership varies with the shuffle stream — across

three controlled shuffle seeds, fc's R_1 spread is $0.36°$ (spectrum concentrated, head stable) while conv1's is $23.1°$ (spectrum flat, top-1 underdetermined).

Second (augmentation RNG — the main cause): the marked loader's RandomCrop/Flip consume the global RNG per __getitem__, so **re-shuffling equals re-drawing the dataset**: between two shuffles of the same checkpoint the gradient sums differ relatively by 15%, and they still fail to coincide even under per-sample summation. Diagnostic control: the cloud experiments (clean eval + fixed order) reproduce across machines to $0.01°$ (ViT-Tiny local $96.59°/88.6°$ vs cloud $96.62°/88.6°$) — clean + fixed-order eliminates both defects.

The unified convention (§3.1): clean transform + fixed order + CE reduction='sum'; all 20 CIFAR experiments re-measured; the 2×2 LN arms and ViT-Tiny experiment supplemented. **Convention shift on the ImageNet/ViT-scale experiments (proxy bound, superseded by exact re-measurement)**: the batch-mean $\leftrightarrow$ per-sample-sum proxy experiment had bounded α_g^{sgn} at $\leq 0.24°$ and R_1 at r50 $\leq 3.1°$ / convnext $\leq 13°$; post-hoc re-measured shifts are sgn $\leq 0.25°$, R_1 r50 $\leq 2.9°$ / convnext $-8.7°$ (cf. the convention note in (d)) and ViT-B/L five-column difference $\leq 0.7°$ — no scale conclusion (the $92°$ narrow band, the flat/upward n_f sweep, the breed replication) is affected by the convention shift (exact values: Table 10 and A.3(c)/(e)).

(h) Bases for the gradient-side angles: the null battery. Motivation: §3.2 and (g) established that absolute angles cannot be read across settings, yet the body had still been using the absolute value of R_1 as structural evidence — the activation side had long been equipped with a null, the gradient side had none. Each experiment receives four bases, 20 repeats, seed 2, entirely under the §3.1 convention: **iso** (an i.i.d. Gaussian pair of matched shape — the dimensional ceiling), **stiefel** (keep each side's U and singular values, redraw the right basis — the spectrum-matched independence ceiling), **split** (a same-distribution half-split with true labels at the experiment's own (n_f, n_r) — the α_h null recipe ported to the gradient side), **halfclass** (two halves of the forget class — the within-class reproducibility floor). Define **separation depth** $= (\mathrm{real} - \mathrm{halfclass})/(\mathrm{iso} - \mathrm{halfclass})$.

Table 9: R_1 of every experiment against its own four null bases (deg). †: chan-mean trunk layout, right space $\mathbb{R}^9$. "partial": ViT-Tiny reaches forget 89.34% (§3.3).

Experiment (right-space dim D)	real R_1	iso	stiefel	split	halfclass	depth	feasible
trunk layer1–4† (9) · C10 r18	0.54–1.75	66.8–77.1	71.0 (all layers)	0.9–2.1	0.6–1.4	$-0.01 \ldots 0.01$	yes
trunk layer1–4† (9) · C100 r18	0.21–1.30	66.8–77.1	71.0 (all layers)	1.1–1.6	0.6–1.6	$-0.02 \ldots 0.00$	**no**
trunk layer1–4† (9) · C10 r34	1.04–1.48	66.8–77.1	71.0 (all layers)	1.1–1.7	0.5–1.5	$-0.00 \ldots 0.01$	yes
C10 r18 fc (512)	24.92	88.24	88.23	36.07	12.58	0.16	yes
C10 r34 fc (512)	28.45	88.24	88.23	55.88	18.26	0.15	yes
C100 r18 fc (512)	54.39	88.69	87.85	25.64	21.11	0.49	**no**
C10 random 4500 fc (512)	19.40	88.24	88.23	38.82	33.73	-0.26	**no**
ViT-Tiny head (192)	88.61	86.97	87.60	28.20	2.35	1.02	partial
ViT-Base head (768)	39.11	88.31	88.04	8.84	0.85	0.44	yes
ViT-Large head (1024)	30.74	88.94	88.37	36.07	0.21	0.34	yes

Four conclusions. **(i) The low trunk R_1 is rank-1 saturation in a 9-dimensional right space**: the evr $= 0.99$ rule selects $k = 1$ at every layer (R@evr99 equals R_1 bit for bit), the real value falls inside the split-null band, and at layer3/4 it is even below the within-class floor — "forget and retain share a learned semantic direction" is indistinguishable here from "any two subsets share a direction the model pins down". **The verdict is not a single-experiment coincidence**: switching dataset (C100) and depth (r34), all 12 (layer, experiment) cells have real $\leq$ split p95 (8 of them also below split p5) and 11 of 12 below the floor p95 — the sole exception, r34·layer2 (real $1.48°$ vs floor p95 $0.75°$, over by $0.73°$), still sits inside its own split band $0.6°$–$2.1°$ (depth 0.013); the per-layer iso ceilings are **identical across the three experiments** ($77.1/75.8/66.8/73.9$), because they depend only on the layout shape, not on the weights. The only shape with an exception is conv1: its evr99 selects $k = 4$ (R@evr99 $= 3.4°/1.8°/2.2°$, against the $3.6°/1.8°/2.1°$ stored in Table 10, difference $\leq 0.2°$), but its real values ($13.8°/19.7°/13.0°$) likewise fall inside their floor bands (means $18.6°/13.8°/12.8°$) with $|\mathrm{depth}| \leq 0.09$ — the non-selectivity conclusion is unchanged. The trunk gradients' information sits on the class-space side and **moves with class count, not depth**: L_1 is $53.3°$–$80.0°$ on C10 (r18) and $57.1°$–$79.9°$ (r34), rising to $80.9°$–$89.8°$ on C100. The layer ordering also replicates: trunk real $<$ conv1's evr99 truncation value $<$ fc (C100: $1.30° < 1.83° < 54.4°$; r34: $1.48° < 2.18° < 28.5°$) — the trunk$\rightarrow$fc layer-band ordering of Reading (ii) is not a C10-r18 artifact.

(ii) The high-dimensional dilution hypothesis is falsified: the iso ceiling moves only from $86.97°$ to $88.94°$ across $D = 192 \to 1024$ (less than $2°$) while the real spread to be explained is $64°$; and the experiment labeled "diluted" (ViT) is the lower-dimensional one.

(iii) ViT-Tiny's $88.6°$ is class-specific separation at the ceiling: within-class floor $2.35°$, per-class pairwise gradient angles $82.1°$–$88.3°$, per-class $\mathrm{evr}_{\mathrm{top1}} = 0.981$–$0.990$, $\sigma_1/\sigma_2 = 13.4$; 98.2% of its forget-side energy is carried by **its own** class-mean factor ($\cos^2 = 0.982$) while being orthogonal to the **global** $\bar{h}$ ($|\cos(u_1, \bar{h})| = 0.027$) — on this experiment the common-factor mechanism degenerates into the class-specific case, consistent with §4.1 (the criterion removes the global $\bar{h}$). A side check supporting §3.2 (Tiny ceiling replicated by seed 7; pretrained contrast still $n = 1$): **the median per-class gradient angle falls from $86.3°$ (ViT-Tiny) to $45.0°/39.8°$ (pretrained ViT-B/L), and the C10 r18 fc class-cone median is exactly $41.6°$** — "pretrained ViTs return to the CNN band" (§3.2) replicates on a new instrument, and what moves is the **across-class** separation (the within-class floor stays $0.12°$–$2.29°$), i.e. the exception narrows through the loss of class-specific directions, not through larger within-class noise.

(iv) The angle between the two sides' mean features measures $G \approx (\sum_i \delta_i) \otimes \bar{h}$ directly: random subset $\mathbf{0.75}°$ ($|\cos| = 0.9999$), C10 whole class $54.6°$, C100 $62.3°$, ViT-Tiny $119.9°$ ($\cos = -0.499$). The top-1 direction's $|\cos(u_1, \bar{h})|$ per experiment is 0.86 (C10 fc) / 0.60 (C100 fc) / 0.81 (random subset) — the composition changes, the dominant direction does not.

Audit and limits: the validation arm's real R_1 differs from the archived runs by $\leq 0.024°$; the C10 r18 trunk experiment reproduces all three layer bands of Table 10 (conv1 row max difference $0.21°$, inside the conv-backward atomics nondeterminism band; fc rows $\leq 0.05°$); trunk weight loading is hard-checked (missing $= 0$ and unexpected $= 0$).

Each of the three fc experiments was measured by two independent implementations (closed-form head gradients vs full-graph backprop): real values bit-identical, bases agreeing to iso $\leq 0.50°$, floor $\leq 1.38°$, depth ≤ 0.022 — the nulls themselves are reproducible.

Repeat widths stratify per base, and this paper argues only on the tight side: split is tight on the trunk but very wide on head experiments; the halfclass floor is tight on the trunk and the ViT experiments but wide on CNN fc experiments; iso is somewhat wide on heads, and on the trunk the 9-dimensional right space pushes the angle distribution far from the concentration point — so "trunk real inside or below the split band" and the band-level saturation verdict (11 of 12 cells, see (i)) are usable, while head-experiment split values and fc-experiment depths are ordinal references only and iso serves normalization only; no significance claim is made on the loose side. All per-experiment bands and p5/p95 are archived.

A.4 Extension of Table 2 to the pretrained-ViT and ImageNet experiments

Table 10: Table 2, continued: decision-layer gradient geometry for the eight secondary CIFAR experiments (conv1/C100/ ResNet-34/random; the body table keeps only the §3.2 anchors) and the pretrained-ViT and ImageNet experiments, same unified §3.1 convention and column meanings as Table 2.

Experiment	α_g^{sgn}	L_1	R_1	L_2	R_2	R@evr99
C10 whole class · r18 fc	$126.0°$	$46.7°$	$24.9°$	$20.8°$	$5.1°$	$\mathbf{5.1}°$ (k=2)
C10 whole class · r18 trunk layer1–4†	99.1–$126.6°$	53.3–$80.9°$	$\mathbf{0.5}°$–$\mathbf{1.8}°$	50.2–$63.8°$	$\mathbf{0.3}°$–$\mathbf{1.1}°$	$\leq 1.8°$ (k=1)
ViT-Tiny · C10 head (from scratch)	$96.6°$	$\mathbf{89.0}°$	$\mathbf{88.6}°$	$72.4°$	$72.2°$	$88.6°$ (k=1)
C10 whole class · r18 conv1†	$114.2°$	$61.1°$	$13.9°$	$44.1°$	$12.6°$	$\mathbf{3.6}°$ (k=4)
C100 whole class · r18 fc	$95.7°$	$83.5°$	$54.4°$	$82.6°$	$53.6°$	$54.4°$ (k=1)
C100 whole class · r18 conv1†	$106.2°$	$65.4°$	$19.6°$	$51.1°$	$17.9°$	$\mathbf{1.8}°$ (k=4)
C10 whole class · r34 fc	$125.6°$	$38.8°$	$28.4°$	$3.3°$	$15.9°$	$15.4°$ (k=3)
C10 whole class · r34 conv1†	$54.2°$	$45.4°$	$12.8°$	$33.6°$	$12.7°$	$\mathbf{2.1}°$ (k=4)
C10 random 4500 · r18 fc	$60.1°$	$51.5°$	$19.4°$	$39.7°$	$19.3°$	$9.6°$ (k=6)
C100 random 4500 · r18 fc	$83.9°$	$83.7°$	$\mathbf{15.0}°$	$81.7°$	$12.9°$	$3.5°$ (k=59)
C10 random 4500 · r34 fc	$77.9°$	$35.9°$	$48.5°$	$11.7°$	$23.1°$	$10.3°$ (k=5)
ViT-Base · C10 head (IN pretrained)	$110.2°$	$52.8°$	$39.1°$	$28.4°$	$25.0°$	$\mathbf{39.1}°$ (k=1)
ViT-Large · C10 head (IN pretrained)	$117.7°$	$13.3°$	$30.7°$	$7.6°$	$30.0°$	$\mathbf{30.7}°$ (k=1)
ImageNet · r50 fc (Bouvier)	$92.2°$	$86.9°$	$49.2°$	$86.8°$	$48.3°$	$48.3°$ (k=2)
ImageNet · convnext_tiny fc	$92.5°$	$86.3°$	$78.7°$	$84.3°$	$73.4°$	$73.4°$ (k=2)

Algorithm 1 SVD-PROJECTION UNLEARNING AT ONE SITE — the procedure evaluated throughout this paper. The backbone is trained *before* the request arrives (the Require prerequisite); the only rank-dependent choices are lines 2–4.

Require: backbone ϕ with head (W, b), trained on $D_r \cup D_f$ before any request arrives; forget set D_f; retain set D_r; energy ratio evr; epochs T; step size η

Ensure: re-fitted head $(\hat{W}, \hat{b})$ whose readout pathway for D_f is removed

1: cache features $H_f \leftarrow \phi(D_f)$, $H_r \leftarrow \phi(D_r)$
2: $\bar{h} \leftarrow$ pooled mean of the rows of H_f and H_r; $\bar{H}_f \leftarrow H_f - \bar{h}$, $\bar{H}_r \leftarrow H_r - \bar{h}$ ▷ pooled global-mean centering (§3.1/§4.3 convention); the same shift is applied to features at inference
3: $(\sigma_i, q_i) \leftarrow$ singular pairs of $\bar{H}_f$; $k \leftarrow \min\{j : \sum_{i \le j} \sigma_i^2 \ge \text{evr} \sum_i \sigma_i^2\}$ ▷ SEMU energy rule, its Eqs. 11–12
4: $P_k \leftarrow I - U_k U_k^\top$ ▷ spanned by top-k *feature* directions
5: fold $W' \leftarrow W P_k$ ▷ exact: logit diff $O(10^{-13})$ (§5.4); inference sees one module
6: fine-tune (W', b) on $(P_k \bar{H}_r,\ y_r)$ with cross-entropy, T epochs at step η ▷ projection *alone* does not forget: FA 97.51–100.00 at T=0 (Table 3, arm A)
7: **return** $(\hat{W}, \hat{b})$

Reading notes: pretrained ViT-B/L return to the CNN feature-space band ($R_1 = 39.1°/30.7°$), and the two ImageNet experiments sit inside the $91.7°$–$92.7°$ α_g^{sgn} band with $R_1 = 49.2°/78.7°$; all re-measurement and convention-audit details for these rows are in A.3(b)(c)(e)(g).

A.5 THE FULL SCREENING PROCEDURE (ALGORITHM 1, REFERENCED BY §5.1)

This appendix is the single place that pins the temporal order: training, forget request, projection, head re-fit.

B SEMU INTERVENTIONS AND COMPOSITION-AXIS VERIFICATION

B.1 SEMU: TRUNCATION, FULL RANK, LAYER-BY-LAYER (SOURCE ENTRIES FOR §3.3)

Method core. SEMU's update takes the form $A + U_r R V_r^\top$ (its Eq. 13) — **formally a LoRA-style low-rank adapter** (Hu et al., 2022); low-rank weight factorization for compression has an earlier lineage (Denton et al., 2014). The rank r is selected by the explained-variance threshold γ (its Eqs. 11–12), which the original calls "an interpretable factor determining the approximation rank." Its Theorem 4.1 is the classical Eckart–Young–Mirsky result: the rank-r truncated SVD is the best Frobenius approximation of G — what §2 reads as "about approximation, not about forgetting."

§3.3's readings correspond to code-package artifacts: $\gamma = 0.95$ (forget 13.78%) and full rank (13.44%) are the evr-0.95 and full-rank keys of the class-0 SEMU log (gradient spectra σ_1/σ_2 recorded in the corresponding runs' analysis directories). Layer-cascade source (§3.3's layer-extension row): layer_cascade_results.json — class0 fc-only $20.87\% \rightarrow$ adding conv1 (all 21 layers) 20.07%, intermediate steps drifting by ≤ 0.27pp; the random arm's FA stays 99.40–99.44% at every step (baseline 99.40%). One within-family extra row: on ResNet-34, CIFAR-100 whole class at $\text{evr} = 1.0$ (full rank $r = 100$) still leaves FA pinned at 98.00% — "full rank still does not forget" reproduces isomorphically across datasets and across depth.

Replication-convention note (the attention branch of the SEMU original). The original code carries a separate explained_variance_ratio_attention for keys containing "attn", defaulting to 1.0 — attention layers are **un-truncated** in the original default; our ViT-Tiny replication has no such branch, a single global evr applying to every Linear/Conv2d (qkv/proj included). The discrepancy is bracketed, not untested: the truncated arms (evr $\in [0.5, 0.95]$, attention also cut to $k = 1$) and the full-rank arm (attention likewise un-truncated, covering the original default's attention configuration) give test-segment FA 76.9–77.1 vs. 77.9 (retain ≈ 96.3 for both; the $96.32 \rightarrow 89.34/89.86$ quoted in §3.3 is the FA_{train} column of the same runs, and the two columns agree in sign); the original default's hybrid sits between them with same-signed endpoints — "ViT-Tiny forgets only partially" is insensitive to the branch. The CNN arms (ResNet-18/34) contain no attention layers; no other result is affected.

B.2 RANDOM SUBSETS AT FOUR SIZES

Table 11: Random subsets at four sizes ($n = 3$ forgetting runs with seeds; same checkpoints as Table 6; null p95 0.5080 throughout).

n_f	AUC	null p95	margin	k=1 FA	k=10 FA	k=10 retain	k=10 test
500	0.4968	0.5080	**−0.011**	98.60	24.67	33.54	32.10
1000	0.5017	0.5080	**−0.006**	99.20	24.37	30.01	28.96
2000	0.5029	0.5080	**−0.005**	99.28	25.17	28.87	27.77
4500	0.4952	0.5080	**−0.013**	99.36	25.90	27.28	26.69

All four sizes reject (margin < 0) and all four fail to forget: even the most favorable rank $k = 1$ leaves FA at 98.60–99.36%, while $k = 10$ collapses forget, retain and test jointly — the per-size detail behind §4.2's composition-not-size verdict. The 0.0034 gap between Table 6's 0.4986 and this table's 4500-row 0.4952 is two independent 4096-sample draws of the same forget set, not a size effect — both under p95.

C BASELINE GRIDS, HEAD-LEVEL ATTRIBUTION, AND RELATED-WORK MECHANISMS

C.1 HEAD-LEVEL ISOLATION OF DECISION-LAYER ATTRIBUTION, AND WITHIN-FAMILY RESNET-34 REPLICATION

(a0) Training-sufficiency control. Seven from-scratch checkpoints (49.08%–87.01%, plus the official fully-trained point 94.38%, the C10 row of Table 2) measured under the §3.1 unified convention (full GPU run):

Table 12: Gradient geometry along the from-scratch accuracy axis (fc; accuracy column sorted ascending).

test acc. (from scratch)	49.08	64.43	72.42	80.69	83.19	84.84	87.01	94.38 (official)
R_1	27.4°	30.0°	25.3°	32.2°	22.1°	32.8°	36.4°	24.9°
L_1	75.8°	85.4°	70.5°	78.1°	72.4°	89.7°	84.4°	**46.7°**
α_g^{sgn}	77.4°	84.9°	71.5°	78.7°	73.8°	90.2°	85.5°	**126.0°**

Three points. **First**: across a 38-point accuracy span R_1 stays in 22.1°–36.4° and the official point 24.9° sits inside the band — the **training-sufficiency confound is excluded**; the feature common factor is established early and does not move with accuracy (trunk $R \leq 1.8°$ already at the official point).

Second: L_1 drops moderately with convergence (70.5°–89.7° → 46.7°) — δ-alignment is a process quantity, but lands moderate, not near-zero.

Third: α_g^{sgn} is flat through the undertrained band (71.5°–90.2°) and jumps to 126.0° at the official point — the same lesson as the recipe control in (a′) below: it cannot be compared across settings.

(a) Head-level replacement (the six head_* experiments). Design invariant: one ResNet-18 trunk (checkpoint 0), only the decision layer swapped and the new head retrained (all data incl. class-0; first four rows 15 epochs, last two rows 20 epochs, SGD lr 0.05 — **the original recipe and seed**, reproduction self-check: six test-accuracy rows match the original cloud re-measurement archive row by row to ±0.01pp, first column below).

Three points. **First**, the self-check row is digit-identical to the main anchor (C10-r18 experiment of Table 2: 126.0/46.7/24.9) — the unified convention is deterministic. **Second**, the six R_1 values span 23.6°–48.6°: pooling type /head dimension /head depth barely move it (23.6–30.0, in-band), **the pre-head LayerNorm doubles it to 48.6° — still below the C100 whole-class anchor's own** 54.4° (Table 2): LN is an in-band magnitude effect, not a qualitative break "into the ViT zone." MLP head's unnormalized hidden layer 28.8°: the mechanistic reading of unnormalized linear layers recursively sharing the common factor survives. α_g^{sgn} spans 116.6°–126.2°, mutually indistinguishable (Table 2's overall verdict).

Table 13: Decision-layer gradient geometry under head variants (unified §3.1 convention; C10 class 0 vs retain, seed 2).

Head variant	test acc	α_g^{sgn}	L_1	R_1(fc)
avgpool+fc (as-is, no retrain; self-check)	94.38%	126.0°	46.7°	24.9°
avgpool+fc (retrained; recipe control)	94.56%	126.2°	48.3°	30.0°
maxpool+fc	94.59%	126.2°	44.2°	27.7°
flatten+fc ($512\times4\times4 \to 10$)	94.54%	120.2°	43.3°	27.5°
avgpool→LayerNorm→fc (pre-head norm)	94.52%	116.6°	48.2°	**48.6°**
MLP head ($512 \to 256$+ReLU $\to 10$, no norm)	94.56%	124.0°	50.2°	23.6° (hidden layer avgpool.2: 28.8°)

Third: the LN arm's forward half (does adding LN push a CNN head into the ViT zone) is answered in Second; the reverse half (does the ViT exception survive LN removal) is in (a′). **Boundary honestly stated**: whether §4's activation criterion holds under LN/MLP heads was not measured.

(a′) ViT − LayerNorm two-way arm (head retrained with the same recipe as (a)). Original design question: does the ViT exception **rely on** the final LayerNorm? One frozen ViT-Tiny trunk (best_v2), head only retrained (20ep SGD lr 0.05, cached features); orig is the no-retrain self-check. α_g^{sgn} and R_1 measured under the §3.1 unified convention (clean eval, fixed order, per-sample sums); clean test /head-accuracy self-check: orig 85.80% (reproduction anchor), the two retrained arms 85.78%/85.61%, head accuracy 97.2%/97.1% — heads healthy, readings interpretable:

Arm	final LN	α_g^{sgn}[head]	R_1	clean test
orig (self-check: reproduction anchor)	yes	96.59°	88.3°	85.80%
ln_retrained (recipe control)	yes	118.65°	74.8°	85.78%
noln_retrained (causal arm)	**removed (Identity)**	102.83°	**79.7°**	85.61%

Three points. **First (neither direction is a switch)**: combined with (a)'s forward half — ViT head − LN: $R_1 = 79.7°$, only 4.9° from the LN-keeping same-recipe control (74.8°) and far from returning to the CNN band — **pre-head normalization is neither sufficient nor necessary for the exception; the two-way exclusion holds**. **Second**: the recipe-control arm proves itself necessary again — retraining alone moves R_1 from 88.3° to 74.8° (−13.5°), opposite in sign and larger than the "LN effect"; without reporting ln_retrained, noln's reading would be misattributed to a small positive effect.

Third: after two-way exclusion of normalization the exception remains; same-architecture controls recast it as two-layer geometry (mean misalignment vs. inter-class codebook angle), not trunk feature construction — the finished from-scratch ViT-B ($R_1 = 39.6°$ at 86.5%) and the exclusion chain: E.1.

(b) ResNet-34 within-family replication: full recomputation of Table 6's rows. ResNet-34 pretrained from scratch 182 epochs (CIFAR-10 best SA 95.0%, CIFAR-100 75.9%); the entire pipeline changes only the *arch* switch, seed 2 / class-0 and random-4500 protocols unchanged.

Table 14: The three ResNet rows of §4.2 Tables 6/11 recomputed on ResNet-34 (ex ante diagnosis → ex post verification).

Scenario·model	$\alpha_g^{\mathrm{sgn}}/R_1$ (unified)	$\alpha_h(k{=}1)$	AUC	margin	Ovl	SEMU (lr 1e-5) forget %	feat-proj ($n{=}3$)
C10 class-0 · r18	126.0°/24.9°	89.11°	0.7902	+0.282	0.007	evr0.95: 13.78; full rank: 13.44	k=4: FA 0.11, retain +0.01pp
C10 class-0 · r34	125.6°/28.4°	89.27°	0.8487	**+0.341**	0.005	evr0.95: 17.40 (retain 97.22); full rank: 13.87	k=4: FA 0.00, retain 99.52 (base 99.50), MIA 1.000
C100 class-0 · r18	95.7°/54.4°	86.22°	0.9496	+0.421	0.042	$\gamma \in [0.50, 0.99] \to r{=}1$: const 98.00	best k=1: FA 1.33
C100 class-0 · r34	96.0°/58.2°	85.70°	0.9718	**+0.444**	0.046	evr0.5/0.95 → $r{=}1$: 98.00; **full rank** $r{=}100$**: still** 98.00	k≥ 3: FA 0.00, retain 97.57 (= baseline)
C10 random4500 · r18	60.1°/19.4°	8.08°	0.4986	−0.009	0.910	FA 98.0, no forgetting	k=1: 99.36; k=10: 25.90/27.28 (1.38pp apart, collapse together)
C10 random4500 · r34	77.9°/48.5°	11.93°	0.5031	**−0.005**	0.867	FA 99.47, MIA 0.005 (no forgetting at all)	k=1: 99.43; k=10: 13.24/13.69 (0.45pp apart)

(null p95 reuses r18's measured value at the same (n_f, n_r, d) configuration: both sides' margins ($\geq +0.34$ vs ≤ -0.005) dwarf any plausible null drift, so the approximation does not affect the verdict; the C100 random row was not run.)

Remaining within-family controls (CIFAR-10 protocol):

Three readings. **First**, both sides of the criterion co-occur on a third model: feasible side zero-cost attainment, infeasible side "collapse at all k with gap < 1.4pp," directional necessity (C arm flat), "projection alone does not forget, and applied alone it is harmful" (A arm).

Second, the $\gamma \to r$ map even depends on the model — evr0.95 corresponds to $r = 1$ on r18 and $r = 2$ on r34 — one more line for "γ is not an interpretable factor" (§3.3).

Table 15: Oracle, baselines, ablations: isomorphic replication on r34.

Quantity	r18	r34
oracle (from-scratch 182ep: FA /retain /test /MIA)	0.00/100.00/94.47/1.000	0.00/100.00/92.19/1.000
FT /IU /GA /RL /BS /BE FA	79.2/57.4/12.2/2.6/20.3/23.0	67.6/39.9/9.8/1.8/21.2/20.4
baseline retain (same column order)	99.9/99.2/95.3/99.9/98.8/98.9	99.9/99.0/94.5/99.9/98.1/98.0
feat-proj k=1 (class-0) FA	79.79	74.50 (threshold shape recurs)
ablation A (T=0) k=4/k=10 retain	99.73 no-forget / collapse to 12.76	99.80 no-forget /collapse to 23.10
ablation C (random same rank) along k	99.70 flat	99.93 flat

Third, the c100 side's r34 reading: full-rank r=100 SEMU still gives 98.00% (source entry in B.1); and arm A (projection only) already pushes FA to 9.33% at $k = 1$ (the CIFAR-10 counterpart reading is 99.49%) — when class energy concentrates in the first principal direction, the projection alone can sever most of the readout, but it does not converge to zero along k (23.11% at $k = 4$, rebounding to 49.33% at $k = 10$), failing to constitute completed forgetting under any τ_k. The SEMU row shares r18's regime (lr $= 1\text{e}{-5}$); step curves in (c).

(c) SEMU step-size sensitivity and reproduction protocol (r18_semuL_* runs). evr $= 0.95$, class-0, seed 2, current code scanning the rescaled-step lr:

Table 16: SEMU step-size curve (ResNet-18).

unlearn lr	forget	retain	test	MIA
$1\text{e}{-5}$	14.53	97.73	91.57	0.855
$1\text{e}{-4}$	**0.02**	96.47	90.32	1.000
$1\text{e}{-3}$	0.00	**69.41**	**65.56**	—

lr $= 1\text{e}{-5}$ reproduces §3.3's evr0.95 reading ($13.78/97.89/91.70$); $1\text{e}{-4}$ is the point where the step begins to bite — pushing FA to 0.02 costs retain -3.0pp and test -4.2pp relative to oracle; $1\text{e}{-3}$ collapses outright. In a subspace without directional selectivity, the more force, the more forget and retain are rewritten in one proportion — step size is not a fourth escape route but the behavioral signature of §3.2 Reading (ii) (working rank $r = 1$ on the shared feature factor). **Reproducibility audit**: all §3.3 anchors regenerated with the current code (evr0.95 $\rightarrow$ 14.53 vs paper 13.78; full rank $\rightarrow$ 14.02 vs paper 13.44) — same regime, run-level difference ≤ 0.8pp: §3.3 is self-certifying with the code package; r34 counterparts in Table 14 (lr $= 1\text{e}{-5}$ same regime).

C.2 Per-method mechanisms in the related-work map (fills §2)

Projection-plugin family. LOFT (LOw-dimensional FeaTure subspace) studies unlearning in a low-dimensional feature subspace and learns a projection plugin (Fang et al., 2026); POUR fixes its projection direction through Neural Collapse (Le et al., 2026) — the $T = 0$ arm of §6's two-arm construction has the same one-shot "project once, train nothing after" form as POUR-P, with a different direction source. One adjacent member beyond vision: the GLOSS line suppresses a global toxic subspace in LLM FFN weights (Duan et al., 2026), evidence that the projection idea crosses modalities and architectures. Division of labor with this paper: the family proposes operators; this paper proposes a criterion for when operators can work.

Linear erasure, line by line against our operator. INLP removes attribute information by iterative nullspace projection (Ravfogel et al., 2020); LEACE gives the closed-form optimal linear eraser (Belrose et al., 2023). Our operator is the restricted special case "global-centered top-k right-singular directions of $H_f \rightarrow$ fold into the head $\rightarrow$ re-fit on retain" (rank chosen by the separability criterion); LEACE optimizes in closed form over *all* linear maps, hence the strict domination on the closed-form-linear-erasure point. Objectives differ as well: linear non-detectability versus closeness to a retrained reference plus the feasibility question.

Continual-learning orthogonalization. GPM computes each task's gradient subspace from an activation SVD and orthogonalizes new-task gradients against it to *prevent* forgetting (Saha et al., 2021); this paper uses the same machinery to *cause* it — the tool is neutral, direction selection decides the use.

Exact and certified unlearning. Exact unlearning retrains from scratch on $D \setminus D_f$ (Cao & Yang, 2015), its cost argued in SISA (Bourtoule et al., 2021); certified removal demands indistinguishabil-

ity from a never-trained model (Guo et al., 2020). This paper's $|\mathrm{UA} - \mathrm{UA}_{\mathrm{oracle}}|$ convention (§3.1) is the empirical counterpart, deliberately without ε, δ guarantees.

Failure mechanisms and reweighting. Mavrothalassitis et al. (2025) prove statistical dependence steers gradient ascent away from the retrained solution; PUMA maintains utility by influence modeling and reweighting over the remaining data, one of the geometry-free family of §6.2 (Wu et al., 2022).

Leakage-channel literature. Unlearning leaves traces in the model (Chen et al., 2021); the pre/post parameter difference can reconstruct deleted data (Bertran et al., 2024); the RUB benchmark's mapping attack (UMA) recovers forgotten content through residual knowledge inside unlearned models (Xuan & Li, 2026); ReTrace guides reconstruction with residual rewards on large deep networks (Ma et al., 2026). Weight-level residue is a general property of unlearning operations, not projection-specific — even a retrained model keeps one — which is the basis of the §5.4 positioning sentence.

Folding-arm reconstruction detail (§5.4 expansion). After folding $W' = WP_k$ the inference path has no separate module (measured max logit difference ≈ 1.1–1.4×10^{-13} across $k \in \{1, 3, 5, 10\}$, argmax agreement 1.0000). Threat model, two-sided: an adversary holding **only** the folded model cannot identify U_k (k dimensions hidden inside the $(d - C)$-dimensional orthogonal complement); an adversary also holding the original published model **can** — the row space of $W - W' = WU_kU_k^\top$ is exactly $\mathrm{span}(U_k)$, recovered by a single SVD at principal angle $0°$ with no approximation — and the latter is the realistic case, since forgetting is applied to already-published models. The increment over the known attack surface is the **form** (rank $\leq k$, here 1–5, simple algebra): the exploitation bar rises; the leak does not vanish.

D Rank-Criteria Detail and External Verification

D.1 Rank-criteria detail (fills §6)

Table 17: Criterion → prediction vs measurement.

Criterion	Computed on	Predicts (C10 / C100)	Measured k^*	Mode of failure
Principal-angle threshold 30°	angle between U_f, U_r	18 / —	4	over-selects 18×
Energy Youden point	$\arg\max[\varphi(k) - \rho(k)]$	2 / —	4	under-selects
Discriminative score (through classifier pre-fine-tune)	$\mathrm{UA}(P_k$ alone)	1 / —	4	under-selects
Head-alignment $\mathrm{E}(k)$	distribution of $\langle w_0, q_i \rangle$	$k_{90} = 1$ (both)	4/1	no discrimination
Spectral concentration	$r_{\mathrm{eff}}, \sigma_1/\sigma_2, k_{95}$	8/28	4/1	anti-correlated

Table 18: Spectral concentration (activation spectrum; a different quantity from §3.3's gradient spectra — the two must not be read against each other; k_p: smallest rank whose leading directions capture $p\%$ of activation spectral energy; $r_{\mathrm{eff}} = \exp(H)$: effective rank from the spectral-entropy H of the normalized singular values).

Scenario	dims	$\varphi(1)$	r_{eff}	r_{eff}/d	σ_1/σ_2	k_{95}	k_{99}	measured k^*
CIFAR-10 class-0	512	0.693	56.8	0.111	**3.38**	8	25	4
CIFAR-100 class-0	225	0.819	99.0	0.440	**5.62**	28	225	**1**
CIFAR-10 random	512	0.156	59.4	0.116	**1.17**	9	33	none
ViT class-0 (self-mean)	192	0.144	138.2	0.720	**1.18**	97	157	2

σ_1/σ_2 is the only scale-free, cross-row-readable column here: the two failing configurations (random; ViT self-mean's false negative) tie at 1.17–1.18 while feasible ones sit at 3.38–5.62 — **a second ex ante feasibility signal, independent of AUC**. But *no* column predicts k^*: CIFAR-100 needs only $k = 1$ yet has the flattest spectrum ($k_{95} = 28$, $k_{99} = 225$, $r_{\mathrm{eff}}/d = 0.440$ — the highest of all rows). Energy concentration and discriminative concentration fully decouple.

Table 19: Head-alignment readout distribution $\mathrm{E}(k)$ (share of the classifier's class-0 readout weights falling on the top-k directions of H_f; global centering).

Scenario	$\mathrm{E}(1)$	$\mathrm{E}(4)$	k_{90}	k_{99}	measured k^*
CIFAR-10 class-0	0.947	**0.957**	1	57	4
CIFAR-100 class-0	0.945	0.946	1	207	**1**

The anticipated contrast (CIFAR-100 concentrated in top-1, CIFAR-10 spread over ≈ 4 directions) **does not materialize**: both place $\geq 94.5\%$ of head weights on the single leading direction, their

$\mathrm{E}(k)$ curves nearly coincide, and even k_{99} is anti-ordered (57 vs 207); no threshold on $\mathrm{E}(k)$ recovers the correct ordering. In this table, $\mathrm{E}(4) = 0.957$ — removing 95.7% of the readout weights still does not produce forgetting; the raw and margin variants agree to three decimals.

Correction attempt (tested; predictive power not restored). Scoring *after* brief head re-fitting also fails: at probe lengths $10/20/50$ epochs, $\mathrm{UA}(k{=}4)$ is $96.78\%/89.38\%/32.76\%$ (Figure 3, right) — a trend consistent with the fine-tuning-dynamics account of forgetting, still nowhere near the real method's 0.11% at the same k and epoch count. The probe re-fits on 4096 cached clean activations while the method fine-tunes on $40{,}500$ augmented images; reproducing the trajectory *is* the scan.

D.2 Protocol and hyperparameters (supplying §6.2)

Common protocol: CIFAR-10 class-0 as the forget set; baselines and feat-proj share one checkpoint; seed 2 (baselines additionally run seeds 1/3 for the $n = 3$ mean $\pm$ std reported in Table 20); the oracle is the 182-epoch from-scratch retrain (§3.1); feat-proj uses global centering and $n = 3$ forgetting-run seeds. All baseline hyperparameters are the **original-paper /public-implementation defaults, untuned per method** (Fan et al., 2024) — the two tables of this appendix test exactly the prediction "does it rely on activation separability?" and do not constitute a leaderboard; full per-method hyperparameters and commands reproduce from the code package.

Method	Provenance	Mechanism
FT	standard baseline	fine-tune on retain data only
GA	failure mechanism analyzed by Mavrothalassitis et al. (2025)	gradient ascent on the forget set
RL	random-label training convention per Zhang et al. (2017)	randomize forget labels
IU (Wfisher)	WoodFisher (Singh & Alistarh, 2020)	second-order Fisher-approximate down-weighting
BS /BE	Boundary Unlearning (Chen et al., 2023)	boundary shrink /expand
FT_prune	Jia et al. (2023)	L1-sparse fine-tuning

Full result table (the body's Table 4 is a four-row trim of this one — the RL, IU, BS, BE and FT_prune rows appear only in this full table; mean $\pm$ std over behavior seeds 1/2/3):
Table 20: Baseline cross-validation, full view (ResNet-18, class 0, baselines: $n = 3$ seeds, mean $\pm$ std). No entry is tuned to its upper bound (upper bound: C.1).

Method	FA (%)	Retain	Test	MIA	Consistent with prediction?
Oracle	0.00	100.00	94.47	1.000	✓ whole-class oracle is 0; perfect forgetting
FT	86.14±9.04	99.97±0.02	94.82±0.10	0.139±0.090	✓ geometry-free: retain-only fine-tune; default reading (C.1)
GA	11.97±0.15	95.44±0.15	89.04±0.18	0.880±0.002	✓ geometry-free: more aggressive ascent; lower FA, retain damaged
RL	1.10±1.08	99.92±0.05	94.78±0.09	0.989±0.011	✓ geometry-free: random labels destroy forget supervision
IU (Wfisher)	68.65±22.39	99.44±0.40	94.13±0.49	0.313±0.224	✓ geometry-free: Fisher-approximate down-weighting, not using the activation leading directions
BS (boundary shrink)	20.36±0.09	99.01±0.26	93.34±0.03	0.796±0.001	✓ geometry-free: boundary shrink
BE (boundary expand)	23.21±0.40	99.14±0.28	93.36±0.04	0.768±0.004	✓ geometry-free: boundary expand, same family
FT_prune (l1-sparse)	0.00±0.00	39.91±40.73	38.42±38.62	0.333±0.471	✓ geometry-free: FA zero, but 2/3 seeds collapse globally (§6.2 reading)
feat-proj ($k{=}4$)	**0.11**	**99.48**	**94.43**	**0.999**	✓ **geometry-dependent: criterion predicts feasible; projection reaches oracle level, retain intact**

Reading. The FT row is **not** evidence that projection out-forgets fine-tuning — tuned FT behaves no differently from feat-proj (C.1). What the table does say is on the **mechanism axis**: the criterion judges this scenario feasible ($+0.282$); the one geometry-dependent entry lands on the oracle ($0.11/99.48/94.43/0.999$ vs. $0.00/100.00/94.47/1.000$), while every geometry-free method's position is given in the last column; seed stds ≤ 0.4 for GA/BS/BE — the separation is not a lucky seed. The large stds of the FT and IU rows are read, per the E.2 convention, as instability of the methods themselves; between-method differences of the order of those stds are not taken from their means. FT_prune must be read with its warning: FA 0.00 is bimodal (per-seed detail and the lesson in E.2) — the distinction is distribution shape, not loss magnitude. On random subsets the criterion rejects (§4.2); baselines were not run there, so they are not tabulated.

D.3 ViT MIA details and reading caveats (expanded from §5.3)

Protocol. SVC attack (RBF kernel, $C = 3$, gamma='auto'), attack-training set = random 50% of retain, test set = full test set, target set = full forget set (target_train=None convention). **Reading caveats (referenced from §5.3 and §6.2):** MIA correctness lies in $[0, 1]$; 0.5 corresponds to an uninformative attack over features, 1.000 to forget points being inseparable from unseen test points in attack features — **not an absolute privacy certificate**: a globally collapsed model also pushes the attack to saturation (Table 20's GA row, the GA and FT_prune rows on ViT), which is why MIA is always reported jointly with FA / retain / test, never alone. 1.000 is the ResNet-side retrained-oracle value; ViT has no retrained oracle, so a ViT 1.000 reads as attacker-blindness only, not as equivalence to a ViT retrain.

Table 21: ViT-Tiny CIFAR-10 whole class: $(k, \text{ablation}, \text{FA}_{\text{train}}, \text{MIA})$ under the protocol and caveats above.

k	Ablation	FA_{train} (%)	MIA correctness
1	full	0.00	**1.000**
3	full	0.00	**1.000**
5	full	0.00	**1.000**
10	full	0.00	**1.000**
1	C random	80.10	0.199
3	C random	79.42	0.206
5	C random	79.16	0.208
10	C random	79.24	0.208

E LIMITATION DETAILS

E.1 THE VIT SPLIT: HYPOTHESIS-EXCLUSION CHAIN AND MEASUREMENT PROPERTY

Exclusion chain. The split itself ($R_1 \approx 89°$ only in from-scratch ViT-Tiny; pretrained ViT-B/L back in the CNN band, $R_1 = 39.1°/30.7°$; Table 2 continued, A.4) already excludes two old candidates. (i) *Pre-head LayerNorm — excluded in both directions* (C.1a/a′): adding LN to a CNN head doubles R_1 but stops at $48.6°$, short of the C100 anchor ($54.4°$) and outside the ViT region; deleting LN from the ViT-Tiny head gives $R_1 = 79.7°$, $4.9°$ from the LN-retaining recipe-control arm ($74.8°$) and far from the CNN band. Pre-head normalization is neither sufficient nor necessary. (ii) *High-dimensional dilution* (A.3h(ii)).

Table 12 cannot exclude ViT undertraining: those seven accuracy points are ResNet-18; the ViT-Tiny trajectory is the opposite — $R_1 = 38°$ at 73% (CNN band), pinning at 87–89° only at 85.8–86.7%. Same-architecture pretraining is done (vit_tiny_patch16_224.augreg_in21k_ft_in1k $\to$ CIFAR-10 FT, test 97.5%): mean-feature angle $120° \to 56°$ (CNN band), forget $|\cos(u_1, \bar{h})|$ $0.027 \to 0.66$, cone median and R_1 only to 73° — pretraining repairs the shared mean first and does not finish the cross-class private axes.

Do not write the remainder as trunk feature construction. The mechanism boundary is in §3.2. The intermediate variable is a representation-organization regime: every class's head gradient is a rank-1 codeword on all models ($\text{evr}_{\text{top1}} \approx 1$); the anomaly is how codewords pack. From-scratch narrow ViT unpacks class directions from a low-rank bundle to near-orthogonality (seed 7: PR ≈ 3.2, $R_1 = 60°$ at ep10; PR ≈ 8.6, $R_1 = 89°$, cone median $84°$ at ep100; forget/retain mean $130°$). Forget-vs-others and retain-only cones share the same distribution, so this is a global codebook, not a forget-conditioned private axis. Pretraining repairs mean alignment ($120° \to 56°$) while cone/R_1 only reach 70°/73° — two separable layers.

From-scratch ViT-B (same Tiny recipe, seed 2, pretrained=False) at test 86.5% has $R_1 = 39.6°$, mean angle $39.6°$, cone median $53°$, matching pretrained Base ($39.1°$ / $34.6°$ / $48°$) — width alone is enough to stay in the CNN band; the ep20 $17.8°$ (67.9%) is undertrained onset, not a width conclusion. Class-mean effective rank is 8.5 / 8.5 / 7.6 on FS-Tiny / PT-Tiny / PT-B, so CNN-like R_1 comes from a tighter cone and μ_F not orthogonal to the retain span, not a low-rank shared subspace. $\eta_F \approx 1$ on all three experiments (forget is always a mean outer product); the discriminators are η_R (0.012 / 0.063 / 0.30) and $\angle(u_1, \bar{h}_{\text{global}})$. Independent seed 7 reproduces 87–90° at $\geq 85\%$: Tiny $88°$ is not a measurement error, and is demoted to an appendix narrow boundary.

Second (measurement property). The ImageNet experiments are validation-only (50/class, no oracle, no MIA): their decision axis is criterion–outcome consistency, not oracle equivalence, and the train-split upgrade was not executed. The accumulation-convention re-measurement chain is closed (A.3f/g). No sequential or multi-class forgetting requests are covered.

E.2 FULL STATISTICAL CONVENTION

feat-proj main tables use $n = 3$ forgetting-run seeds (behavioral seeds: the data split and the checkpoint are fixed, only the fine-tuning stream varies); the error bars cover **only the randomness of the forgetting process**. The data-split seed is fixed, so oracle and split variances are not included; capturing those requires training an oracle per split seed. **Why no significance tests**: a t-test on three points has almost no power and cannot separate a bimodal from a unimodal distribution (the

FT_prune row below is exactly such a case), so the paper argues in magnitudes of difference instead — at $n = 3$, mean±std bounds every seed reading within $\pm 1.42 \times \mathrm{std}$ of the mean (a conservative bound holding under both the sample and the population std convention); the $k = 3/k = 4$ column gap of 0.89 pt that §5.2 relies on is $\approx 5\times$ the within-column std (0.17), and $k^* = 4$ does not flip with the choice of seed; the Series-D dilution runs of Appendix E.4 agree verdict-for-verdict across all three seeds (Reading 1 there). The Table 20 baselines were originally single-seed and are now extended to $n = 3$ — and the extension itself produced a material correction: the single-seed FT_prune reading proved to be an unrepresentative mode of a bimodal distribution — only seed 2 keeps the published $97.51/93.03/\mathrm{MIA} = 1.000$ reading, the other two collapse globally (retain ≈ 11.1, MIA 0; §6.2 reading, D.2) — so seed variance here is not a courtesy metric. ViT scenarios are $n = 1$: FA is already 0.00% for all $k \geq 1$ there, so multi-seeding would not change the verdict, but strictly its seed variance is unmeasured.

Per-seed raw values, listed in full rather than summarized and transcribed from the run logs feat_proj_sweep.json and baseline_multiseed.json. (Convention note: the ±std entries of the main-text Table 3 are computed at full log precision under the ddof$= 1$ sample convention, while Table 20 uses ddof$= 0$; the per-seed values below are recorded to two decimals, so recomputing ±std from the two-decimal values moves the last digit — the $\pm 1.42 \times \mathrm{std}$ bound above holds under both conventions.)

Per-seed FA (%) of the full-method and C columns of Table 3 (behavioral seeds 0/1/2; verdict rule FA $\leq \tau_k = 0.5\%$):

k	full			verdict	C		
	s0	s1	s2		s0	s1	s2
1	80.13	79.73	79.51	not reached	99.69	99.71	99.71
3	1.02	1.16	0.82	not reached	99.69	99.71	99.71
4	0.11	0.16	0.07	met FA $\leq \tau$	99.69	99.71	99.71
10	0.00	0.00	0.00	met FA $\leq \tau$	99.69	99.71	99.71

The per-seed verdict order agrees across all three seeds $(3/3;\ k^* = 4$ for each): the $k = 3$ observed floor 0.82 and the $k = 4$ observed ceiling 0.16 do not intersect; arm A has no fine-tuning and a single reading (97.51/99.49/99.73/100.00), arm B is a single 99.71 (independent of k), and the single C value of the main table is the three-seed mean.

Per-seed FA (%) for the seven method rows of Table 20 (behavioral seeds 1/2/3; the mean±std column matches the main text and D.2 exactly):

Method	s1	s2	s3	mean±std
FT	98.91	79.18	80.33	86.14 ± 9.04
GA	11.78	12.16	11.98	11.97 ± 0.15
BE	22.89	22.96	23.78	23.21 ± 0.40
RL	0.00	2.58	0.73	1.10 ± 1.08
IU	99.91	57.42	48.62	68.65 ± 22.39
BS	20.31	20.29	20.49	20.36 ± 0.09
FT_prune	0.00	0.00	0.00	0.00 ± 0.00

(The FT_prune collapse sits in the retain column: only seed 2 keeps retain 97.51, the other two ≈ 11.1 — see above.)

E.3 Complete list of falsified explorations

All measured and not true (CIFAR-10/100, ResNet-18/34), listed so others do not repeat them:

- **Activation principal angles do not predict retain cost**: at every feasible k on both architectures the retain cost is $\approx$ zero, so there is no cost gradient to predict.
- **Ovl predicts neither rank nor cost**: three classes spanning $18\times$ in Ovl (automobile/ship/airplane) need $k = 4/5/10$ to reach FA 0.00%, anti-ordered with Ovl (C.1).
- **The class-mean angle is not a proxy for semantic proximity**: correlation with α_h is -0.417, the wrong sign.

- **Confusion rates are correctly signed but weak**: correlation -0.233.

E.4 THE CRITERION-BOUNDARY VOID, FILLED: COMPOSITION-AXIS DILUTION EXPERIMENTS AND CROSS-ARM FAILURE OF THE FLIP

Background and design. The original criterion left a void of ≈ 0.15 with zero observations (feasible-side minimum $+0.140$, infeasible-side maximum -0.005; whole-class/random-subset scenarios only, all verified inside the deployment regime, at both ends of Table 3 and B.2). Following the §7 First plan, this section **constructs samples in the middle band of the composition axis**, as a 2×2 **full-factorial four-arm** design (cifar10/cifar100 $\times$ ResNet-18/ResNet-34): forget fixed at $n_f = 4096$ with m class-0 samples, the rest filled at random with other classes; retain draws 4096 from the remaining classes and excludes the fillers (**Series D**: three independent draws per composition point, m taken inside the class-0 pool 4500/450 over cifar10 $\in \{4096, \ldots, 256\}$ and cifar100 $\in \{450, 225, 150, 112, 80, 56\}$); **Series B** leaves the residual class-0 inside retain (retain-side contamination; $m = 4050/2700/900 \times k \in \{1, 4, 10\}\times$ three seeds); **Series S** is pure class-0 at $n_f = 64$–2048 with per-size nulls. Feasibility verdicts and margin rows share the same draw-11 composition; the verdict rule is $\mathrm{UA}_{\mathrm{cls}} \leq \tau = 0.5\%$ **and** RA drop ≤ 1.0 pt, k swept to 30 (cifar100 legacy grids to 14), all three seeds passing; 1089 Series-D projection + head-fine-tune runs and 27 Series-B runs. Margins are re-measured on clean cached activations under the §4.2 convention (null: worst-case p95 of three spectral decays $\times$ 20 trials; the $(4096, 4096, 512)$ cell reproduces the stored value 0.5080, deviation $4.8 \cdot 10^{-5}$; cifar100 configurations use self-paired nulls). **Regime caliber:** SVD projection and head fine-tuning on cached features through a frozen backbone — k^* is not comparable with Table 3 (train-mode BN drift + augmented FT): in this regime whole-class forget reaches $\mathrm{UA}_{\mathrm{cls}} = 0.00$ already at $k = 1$, where the published arm has $k^* = 4$. Flip positions and Series-B failures are mapped first within this regime; the bracket rows are then spot-checked under the deployment pipeline over six scenarios — the regime moves brackets, in arm-dependent directions (Reading 5). Anchor self-check (whole-class cached AUC, cifar10 arms): r18 $= 0.7830$ (the published 0.7902 comes from the augmented-activation path, independent of the subset convention); r34 $= 0.8479$ (margin $+0.3398$).

Reading 1: the void is filled continuously, but the four arms' flip intervals intersect empty. The dilution margins descend along m (recovering only past the trough). r18/c10: $+0.2754\,(4096) \rightarrow +0.1360\,(2048) \rightarrow +0.0827\,(1024) \rightarrow +0.0730\,(896) \rightarrow +0.0611\,(768) \rightarrow +0.0483\,(640) \rightarrow +0.0416\,(576) \rightarrow +0.0293\,(512) \rightarrow -0.0055\,(384) \rightarrow -0.0262\,(256)$ (the $+0.0110$ at $m = 448$ is a margin-only point). The r34/c10 ladder reproduces the shape, its margins **everywhere higher** than r18 at shared verdict points ($+0.3403$ vs $+0.2754$ at the whole-class end, down to $+0.0325$ vs $+0.0293$ at 512; the margin-only 448 is the exception, r34 $+0.0070 <$ r18 $+0.0110$), **yet its flip lands on negative margin** (still feasible at -0.0188, $m = 384$; infeasible from -0.0204, $m = 352$) — margin level and flip position decouple across arms. Both cifar100 arms compress margins and flips around 0.000 (r18: $+0.0310\,(450)/+0.0113\,(225)/+0.0062\,(150)/+0.0003\,(112)/-0.0039\,(80)/-0.0059\,(56)$; r34 same shape, Table 23). cifar10 margins bottom out at $m = 256$ (-0.0262/-0.0331) and recover to -0.0070 at $m = 0$ (128/64/0 are margin-only, no verdicts) — a few coherent components are *less* separable than pure noise. The verdict tables (Table 22, Table 23) use $\tau = 0.5\%$, $\mathrm{TOL} = 1.0$ pt; per-seed verdicts agree across all three seeds; margin-only points $m = 448/128/64/0$ are omitted.

For the rule "skip the scan if margin $\leq t$", arm-wise correctness requires t inside that arm's calibration half-open interval [infeasible-side upper end, feasible-side lower end): r18/c10 $[+0.0293, +0.0416)$, r34/c10 $[-0.0204, -0.0188)$, r18/c100 $[+0.0003, +0.0062)$, r34/c100 $[-0.0043, +0.0005)$. **The four-way intersection is empty** (the cifar100 pair still overlaps on $[+0.0003, +0.0005)$ but the cifar10 arms pull it empty). The sharpest contradiction needs only two same-dataset rows: on cifar10, r34 at $m = 384$ is **feasible at negative margin** -0.0188 while r18 at $m = 512$ is **infeasible at positive margin** $+0.0293$; on cifar100, the two arms at $m = 112$ differ by 0.0002 in margin with opposite verdicts. The margin orders feasibility within an arm but the ordering fails across arms, and across all 33 verdict grids of this paper — the measured domain — no feasible t exists. We therefore **retract the suggested trigger value** 0.10: 18 of the 22 feasible grids have margin below 0.10 and would be mis-skipped by it, so it is not conservative. The old "skip if ≤ 0" one-sided rule held on the eight Table 6/11 rows (seven configurations; the 4500 set appears twice as independent draws) but is refuted in-band by a negative-margin feasible row; deep in the negative region it remains safe: every feasible row has margin ≥ -0.0188, so a $t \leq -0.03$ skip rule

Table 22: cifar10 composition-axis verdict ladder (bold: knife-edge rows).

m	r18/c10 margin	verdict	r34/c10 margin	verdict
4096	$+0.2754$	feasible ($k^* = 1$)	$+0.3403$	feasible ($k^* = 1$)
2048	$+0.1360$	feasible ($k^* = 1$)	$+0.1542$	feasible ($k^* = 1$)
1024	$+0.0827$	feasible ($k^* = 1$)	$+0.0919$	feasible ($k^* = 1$)
896	$+0.0730$	feasible ($k^* = 1$)	$+0.0838$	feasible ($k^* = 1$)
768	$+0.0611$	feasible ($k^* = 1$)	$+0.0708$	feasible ($k^* = 1$)
640	$+0.0483$	feasible ($k^* = 1$)	$+0.0567$	feasible ($k^* = 1$)
576	$+0.0416$	feasible ($k^* = 2$)	—	—
512	$\mathbf{+0.0293}$	**infeasible**	$+0.0325$	feasible ($k^* = 2$)
384	-0.0055	infeasible	$\mathbf{-0.0188}$	**feasible** ($k^* = 4$)
352	—	—	-0.0204	**infeasible**
320	—	—	-0.0268	infeasible
256	-0.0262	infeasible	-0.0331	infeasible

Table 23: cifar100 composition-axis verdict ladder (bold: the $m = 112$ pair).

m	r18/c100 margin	verdict	r34/c100 margin	verdict
450	$+0.0310$	feasible ($k^* = 1$)	$+0.0333$	feasible ($k^* = 1$)
225	$+0.0113$	feasible ($k^* = 1$)	$+0.0125$	feasible ($k^* = 1$)
150	$+0.0062$	feasible ($k^* = 4$)	$+0.0066$	feasible ($k^* = 3$)
112	$\mathbf{+0.0003}$	**infeasible**	$\mathbf{+0.0005}$	**feasible** ($k^* = 20$)
80	-0.0039	infeasible	-0.0043	infeasible
56	-0.0059	infeasible	-0.0072	infeasible

at worst under-scans and never mis-skips a feasible row within the measured domain (no guarantee outside the domain). This is **isomorphic to the §6 proposition** — two parallel impossibilities over different objects (there: a static predictor $(H_f, H_r) \mapsto k^*$; here: a real threshold pre-screening verdicts): non-calibratability propagates in parallel from rank selection to threshold pre-screening.

Reading 2: the margin is blind to retain-side composition (Series B). With forget pure class-0 ($m = 4050/2700/900$, residual class-0 kept inside retain), the forget-side margins are all high ($+0.2702/+0.3073/+0.3653$; the full six-point chain $+0.2702$–$+0.3680$, at the small-m end even exceeding the whole-class anchor $+0.2750$), yet **all 27 runs fail** under the paper's caliber ($3\,m \times 3\,k \times 3$ seeds), through complementary failure channels: at $m = 4050$, $\mathrm{UA}_{\mathrm{cls}} = 0.00$ (cleanly forgotten) but the RA drop is $+1.04$–$+1.13$, beyond tolerance — decidable only at tolerance ≥ 1.5 pt; at $m = 2700/900$ the drops also exceed tolerance (up to $+7.56$), and at large k class-0 becomes **unforgettable** ($k = 10$: $\mathrm{UA}_{\mathrm{cls}} = 56.4\%/95.9\%$, seed spread ≤ 0.25 pt) — the retain-side residual class-0 (unevaluated-subset accuracy 95.89%, equal to the "forgotten" 10% subset's 95.89%) is re-learned by the head fine-tuning. The margin is a forget-side statistic, undefended against retain-side contamination: scenarios it deems "worth scanning" can fail through channels entirely unrelated to it.

Reading 3: sample size is not the confound (Series S). Pure class-0 at $n_f = 64$–2048 (per-size nulls in the $n_f < d = 512$ segment) gives margins $+0.3560$–$+0.4288$; $n_f = 64$ — 1.4% of the cached 4500-sample class pool — still $+0.4070$. Consistent with B.2 (AUC insensitive to forget size): separability is carried by composition, not by sample count.

Reading 4: flip positions are TOL-robust at the paper's caliber; cross-arm shifts occur only outside it. Re-adjudicating all 1089 Series-D runs over $\tau \in \{0.11, 0.25, 0.5, 1, 2\}\% \times \mathrm{TOL} \in \{0.5, 1, 1.5, 2\}$ pt: at $\tau = 0.5\%$ (the paper's caliber) **all four arms' flip brackets are invariant across the entire TOL column**. Shifts appear only outside the caliber and only in some cells: r18/c10 moves one notch down at $\tau \geq 1\%$ and $\mathrm{TOL} \geq 1$ pt (the whole row at $\tau = 2\%$); r18/c100 moves one notch down at $\tau \geq 1\%$ and $\mathrm{TOL} \geq 1$ pt (the $\mathrm{TOL} = 0.5$ column never moves); r34/c10 moves one notch up at $\tau \leq 0.25\%$ and $\mathrm{TOL} \leq 1$ pt; r34/c100 is unchanged across the entire 5×4 grid. The knife-edge row is r18 $m = 512$ (nearest pass at $k = 4$: $\mathrm{UA}_{\mathrm{cls}} = 0.59$, exceeding τ by 0.09 pt, drop exactly 1.00 pt at the TOL edge), but forcing $\mathrm{UA} \leq 0.5$ requires $k \geq 5$, where the drop jumps to ≥ 9.6 pt — the verdict does not ride on boundary luck.

Reading 5: bracket positions move with the fine-tuning regime, which is itself an arm (six-scenario spot check). Re-running the bracket rows under the Table 3 deployment pipeline (train-

mode BN drift + augmented FT through the frozen trunk, $k \leq 6$, the body's verdict rule, baseline $k = 0$ in the same regime, single seed): both cifar10 arms move one notch **down** — r18 $m = 512$ flips feasible ($k^* = 2$, UA $= 0.00$, drop $+0.02$; judged infeasible in the clean regime) and r34 $m = 352$ flips feasible ($k^* = 4$, drop $+0.73$); the $m = 640$ and $m = 384$ rows are unchanged from the clean regime ($k^* = 1/k^* = 4$). cifar100-r18 moves **up**: $m = 150$ does not pass within the window ($\mathrm{UA}_{\mathrm{cls}}$ stalls at 0.67% for $k = 4$–6 — one sample of the 150-row, the lower edge of τ's granularity; the window scanned no passing k, recorded honestly as *not passed in window*, not judged infeasible); $m = 112$ stays infeasible. The direction of movement is again arm-dependent (cifar10 looser, cifar100 tighter). The no-universal-t conclusion survives in-frame on measured rows: r34/c10 feasible at -0.0204 vs. r18/c100 infeasible at $+0.0062$ — the two brackets remain disjoint. Reading 5 is a six-scenario single-seed regime-boundary check, not a full re-mapping: the body verdict tables stay at the clean-regime caliber. Its impact on the conclusion: the fine-tuning regime joins the non-calibratability dimensions — cross-configuration, cross-composition, **cross-regime** static trigger numbers do not exist within the measured domain. (The deployment regime's extra RA collapse at large k — e.g., r18 $m = 640$, $k = 6$: drop $+10.66$ — also caps k on the deployment side, consistent with §6's "the scan is irreplaceable.")

Conclusion. Filling the void upgrades "threshold calibration is impossible" from *evidence absence* to *measured non-calibratability*: the verdict boundary moves ≈ 0.06 along the margin axis across backbone and dataset (the four flip brackets span -0.0204 to $+0.0416$), with inconsistent directions (the deeper arm is r34 on cifar10, r18 on cifar100); superimposing the fine-tuning regime (Reading 5) shifts the brackets another notch, again arm-dependent — in both regimes the measured feasible-row margin floor is -0.0204. The surviving uses of the criterion are two: per-configuration calibration (this section's $1089 + 27$ runs are one complete boundary map per depth configuration on cifar10/cifar100), and the direction law — the more coherent the forget composition, the lower the flip point; plus one deployment corollary: the deep-negative skip ($t \leq -0.03$) never mis-skips a feasible row within the measured domain, in either regime (no guarantee outside the domain). Static trigger numbers across configurations, regimes, or compositions are no longer offered as an option.

E.5 Claim scope and relation to the standard definition (fills §5.4)

The projection modifies only the classifier while the feature extractor is frozen, so the claim object is **activation-subspace separability** (geometry and privacy): removing the leading directions of H_f drops class accuracy to oracle level with retain unharmed (§4, §5.2); the direction must come from that subspace (arm C falsifies random directions); MIA independently confirms the same layer (§5.3). Removal of D_f's influence from the whole model is **not** claimed — the readout path is what is removed, intrinsic to the feature-projection route (§7, Third).

Under the standard reading — exact unlearning requires equivalence to a model that never saw D_f — **no approximate method qualifies, including every baseline in §6.2**, and the leakage literature (C.2) shows weight-level residue is a universal property rather than projection-specific: even a retrained model's parameter difference sits on a known reconstruction surface (Chen et al., 2021; Bertran et al., 2024; Ma et al., 2026). Feasibility is therefore operationalized as a three-layer conjunction — behavioral (UA approaches the oracle), membership (MIA indistinguishable from unseen data), readout (W' no longer routes the class) — and the diagnostic conclusions hold within this convention; the weaker-than-ideal convention is a declared, known cost, and its truth depends on no complete-removal claim.

F Proposition Formalization: Notation, Assumptions, Threshold Calibration

Notation. An **unlearning process** $A = (\text{fine-tune epochs } T, \text{lr}, \text{schedule}, \text{batch order})$ specifies how the classifier is fine-tuned on $(P_k H_r, y_r)$, yielding $W_A(k)$. Let $\mathrm{UA}(k; A)$ be the accuracy of $W_A(k) \circ P_k$ on (H_f, y_f), and $k^*(A) = \min\{k : \mathrm{UA}(k; A) \leq \tau_k\}$. This appendix keeps the older name UA; it denotes the same quantity written FA in the main text.

Assumptions. (i) The backbone ϕ is frozen; only the classifier is updated. (ii) U_k is determined by H_f alone — true of our closed-form construction, false of learned projections (LOFT and the like), which therefore lie outside the proposition's scope. (iii) The threshold τ_k is fixed; we take 0.5%

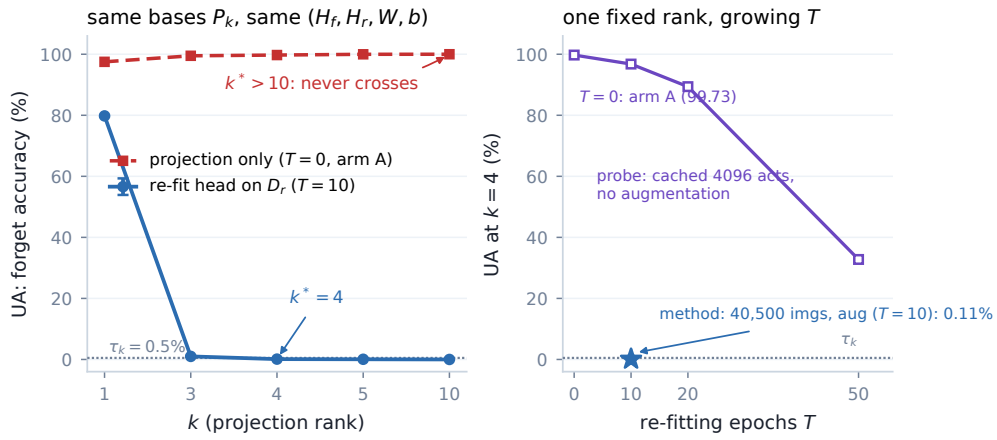

Figure 3: Proposition 1 in data Left: $\mathrm{UA}(k)$, two arms on one checkpoint — the full method ($T = 10$) crosses $\tau_k = 0.5\%$ at $k^* = 4$; pure projection ($T = 0$) never crosses. Right: $\mathrm{UA}(k{=}4)$ falls with fine-tuning budget T toward the real method's 0.11%, which a 50-epoch cached-activation probe does not reach — why a cheap proxy cannot replace the scan.

— $\mathrm{UA}(k{=}4) = 0.11\%$ crosses, $\mathrm{UA}(k{=}3) = 1.00\%$ does not, so $k^* = 4$ for the full method. k^* translates with τ_k, but the proposition holds for every $\tau_k \in (0.11\%, 1.00\%)$.

The static-predictor definition, the proposition statement, and the proof are in §6; the fully formal version, the non-triviality argument, and the scope discussion are in G.

G Proof, Non-Triviality, and Scope

Complete proof of Proposition 1. Suppose such a static Q exists. Take any A_1, A_2 acting on the same (H_f, H_r, W, b); since Q is static, its value is independent of A, hence $k^*(A_1) = Q(\cdot) = k^*(A_2)$. It therefore suffices to exhibit one pair with $k^*(A_1) \neq k^*(A_2)$. **The construction was already delivered by §5.2**: ablation A ($T = 0$) and the full method ($T = 10$) share one checkpoint, the same H_f, H_r, W, b and the same P_k, differing only in T — yet at $k = 4$ their forget accuracies are 99.73% vs. 0.11%, and at $k = 10$, 100.00% vs. 0.00%. Hence under $\tau_k = 0.5\%$, $k^*(T{=}0) > 10$ while $k^*(T{=}10) = 4$ (Figure 3, left). □

Why this is not trivial (full argument). The point is that P_k *is* a function of pre-unlearning quantities, by assumption (ii). If forgetting were equivalent to applying P_k, k^* would be predictable. The content of the proposition is precisely that **forgetting is not equivalent to applying** P_k: fine-tuning is a component of forgetting, and the two act in opposite directions — projection weakens the readout path, fine-tuning rebuilds or suppresses it. Geometrically, k^* is decided by the local loss-landscape geometry the fine-tuning trajectory traverses, a joint product of (P_k, A) that a snapshot fixes only together with A; the two-arm construction makes this concrete as "same snapshot, different trajectories." Nor are the two arms an arbitrary hyperparameter pair: $T = 0$ — apply the projection once, train nothing after — has the same form as the one-shot classifier-weight projection of POUR-P (Le et al., 2026) (directions' source differs, C.2; the training regime does not), and $T = 10$ is the fine-tuning family's usage — two published deployment modes, not post-hoc tuning.

Scope (full discussion). The proposition proves the nonexistence of a static predictor. It does **not** deny that k^* can be recovered once A is fixed — but the predictor must include A in its input or pay to simulate A, and then the cost is the scan; none of the five criteria does. It **does not address** learned projections: assumption (ii) fails for LOFT and other learned projections (the subspace and the objective are learned jointly, so the very problem form is ill-posed for them), so the argument simply **does not transfer** to them (nor does any claim that they are predictable), and their rank/capacity selection remains open. Nor does the proposition cover procedures that unfreeze the backbone, in three cases: with no projection the rank question is not posed; if the projection basis is determined by the updated activations the input is already a process quantity; if the basis is fixed by the old features while the backbone is further updated on top of it, the quantifier does not cover the case and nonexistence is unproven — yet the same two-arm contrast remains unfavorable (the backbone

trajectory only adds to k^*'s process dependence), so generalizing the proposition joins learned projections as an open problem. A scan does give the correct k (CIFAR-10 $k = 4$, CIFAR-100 $k = 1$); what fails is predicting it statically.