Hyperspherical Local Margin Retraction for Zero-Shot Instance-Wise Machine Unlearning

Hongyi Lyu
Macquarie University

Xuyun Zhang*
Macquarie University

Xiaoxiao Chi*
Macquarie University

Guanfeng Liu
Macquarie University

Amin Beheshti
Macquarie University

Abstract

Zero-shot machine unlearning aims to remove the influence of specified training instances from a pretrained model when the retain set is unavailable. Instance-wise unlearning further operates at the level of individual instances, where a forget instance may induce instance-specific representation and margin deviations that are entangled with class-relevant local structure. However, without retain instances, separating these deviations from the structure that retraining would preserve is difficult, exposing retain–forget entanglement in representation space. We propose HLMR (Hyperspherical Local Margin Retraction), a novel zero-shot instance-wise unlearning approach that formulates unlearning as coupled geometric and margin retraction in hyperspherical latent space. HLMR uses hyperspherical geodesic probes to construct a local reference in representation space, preserving the structure supported by this reference while unlearning instance-specific margin excess. Empirical evaluation across multiple datasets and architectural settings shows that HLMR aligns more closely with the retraining oracle in unlearning utility and membership-inference scores than state-of-the-art zero-shot baselines.

1 Introduction

Can a deep neural network forget specific training instances when the retain set is unavailable? This question captures the setting of zero-shot instance-wise machine unlearning, where the unlearning algorithm is given a pretrained model and the target forget set, but has no access to the retain set [13, 36]. This setting arises when privacy requirements, copyright restrictions, data-retention limits, or deployment constraints prevent access to the remaining training data. It is further motivated by privacy regulations such as the GDPR, where the right to erasure is commonly associated with the right to be forgotten and motivates mechanisms for removing the influence of requested data in the model [43]. Compared with retain-data-dependent unlearning, zero-shot unlearning is substantially more challenging: the model must remove the requested information without access to retain instances, which are typically used to regularize the update, restore utility, and determine what information should remain.

Retraining from scratch can provide the oracle reference for instance-wise unlearning, but it is infeasible in the zero-shot setting because no retain set is available at all. Approximate unlearning methods instead seek to modify the pretrained model so that its behavior approaches that of an oracle retrained model [30, 25]. The core challenge is that, without retain instances, the learner cannot determine which part of the forget instance's local original-class support is instance-specific and which part would still be preserved after unlearning. Recent analysis shows that higher entanglement

*Corresponding authors: xuyun.zhang@mq.edu.au, xiaoxiao.chi@students.mq.edu.au.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

between forget and retain sets makes unlearning harder, because unlearning the forget set can inadvertently erase information from the retain set [48]. A successful zero-shot instance-wise method must therefore avoid over-unlearning [29], where retain-supported structure is unnecessarily damaged.

Retain-dependent methods assume that a retain set can be used to guide the update, distill preserved knowledge, or identify entangled retain instances [48, 12, 23, 25, 19, 32], but this assumption fails to hold in the zero-shot unlearning setting. Existing zero-shot unlearning methods rely only on the pretrained model and the forget instances. For example, the ADV+IMP method uses adversarial examples and parameter-importance estimates [6]; SULI refines unlearning through layered soft-label adjustment [35]; JiT smooths local input-output behavior [22]; and MU-MIS suppresses instance contribution through output-level sensitivity signals [10]. These methods are efficient and practically appealing, but their unlearning heuristics are not designed to estimate the neighborhood-supported local representation geometry around the forget instance. Consequently, they do not directly answer the key question in zero-shot instance-wise unlearning: To what extent should the forget instance be retracted with respect to local original-class support when no retain set is available?

For intuition, consider deleting one apple image from a fruit classifier. The model retrained without that image still recognizes apples, because the remaining apple images support this knowledge; given only the pretrained model and the image to delete, the question is *where forgetting should stop*. HLMR perturbs the image's normalized feature along nearby directions (*probes*: feature directions, not images or retained data) and lets the frozen pretrained model score how strongly each probe favors "apple" over the strongest competing class (its *margin*). A low quantile of the probe margins gives a *baseline*, which says how far the margin should be reduced, and nearby low-margin probes give an *anchor*, which says where the feature should move. The update moves the feature toward the anchor and penalizes only the margin above the baseline: for a margin of 0.8 and a baseline of 0.5, the excess is 0.3, and the penalty vanishes once the margin reaches 0.5.

We formulate the update as a local retraction in hyperspherical representation space, using the geometry around the forget point as its reference. By probing the geometry around the forget direction under the pretrained model, one can estimate how much original-class support is locally maintained without directly accessing retain instances. This estimate provides the basis for retracting the point-specific excess while preserving neighborhood-supported local geometry.

Based on this idea, we propose **Hyperspherical Local Margin Retraction** (HLMR), a zero-shot instance-wise unlearning framework in hyperspherical latent space. HLMR constructs geodesic probes around the forget direction and uses teacher-evaluated margins on these probes to estimate a neighborhood baseline. These probes are not treated as substitute retain instances; they expose how the pretrained teacher's directional margin field varies under small geodesic perturbations around the forget representation. A support level that is stable across this local probe neighborhood is used as a conservative proxy for neighborhood-supported structure, whereas excess support concentrated at the forget direction is treated as point-supported. It further derives a geometric anchor from nearby lower-margin probes within a bounded geodesic neighborhood and measures the gap between the forget instance and the baseline as a retraction criticality score. The unlearning update combines two coupled objectives: a geometric retraction that moves the current forget representation toward the anchor, and a margin retraction that reduces the forget point's original-class margin toward the neighborhood-supported baseline. This design operationalizes zero-shot instance-wise unlearning as a local retraction problem: it reduces the forget point's excess original-class support relative to a probe-derived neighborhood baseline, while empirically limiting unnecessary distortion of nearby class-relevant structure. Here, "neighborhood-supported" denotes local support estimated from teacher-evaluated geodesic probes without directly observing retain instances. Our contributions are summarized as follows:

- We formulate zero-shot instance-wise unlearning as a local retain–forget entanglement problem in representation space, where point-supported representation protrusion and excess original-class support must be separated from neighborhood-supported local geometry.
- We propose HLMR, a hyperspherical latent-space unlearning framework that uses teacher-evaluated geodesic probes to estimate a neighborhood-supported margin baseline and retraction anchor using only the pretrained model and forget instances.
- We introduce coupled geometric and margin retraction objectives with an adaptive criticality score, and empirically show across multiple benchmark–architecture settings that HLMR aligns more closely with the retrained model than state-of-the-art zero-shot baselines.

2 Related Work

Approximate unlearning with retain data. Beyond exact machine unlearning [11, 46] and SISA-style partition retraining [4], most approximate unlearning methods modify a pretrained model to approximate the behavior of an oracle retrained model. Second-order and influence-inspired methods use Hessian or Fisher information to approximate the effect of removing data [25, 26], while saliency and parameter-importance-based methods such as SSD [23] and SalUn [19] identify parameters most responsible for unlearning. Uncertainty-based methods, such as LoTUS [40], use a subset of retain instances and an unseen set to preserve utility. Distillation-based methods such as SCRUB [32] and Bad-Teacher [12] balance unlearning and preservation through teacher–student objectives, and recent variants formulate the tradeoff through bargaining or metric-learning objectives, such as DUCK [14] and MUNBa [45]. RUM [48] partitions the forget set by difficulty and applies per-subset unlearning. Building on this, Cheng et al. [9] explicitly partition the retain set into adjacent and remote subsets and protect the two groups separately. SCAR [3] avoids direct access to the retain set by using out-of-distribution data for preserving model utility. All of these methods use retained or additional data to regularize updates, reconstruct preserved knowledge, or identify subsets to protect. This data requirement is incompatible with the zero-shot setting considered here.

Zero-shot unlearning. An increasing number of works remove the retain-data dependence and rely only on the pretrained model and the forget set. We organize this literature by the type of unlearning heuristic it constructs. Output-space pushing methods such as "Negative gradient" [42] maximize the forget set loss directly, but their global parameter modification is prone to over-unlearning [29]. Input-space smoothing is exemplified by JiT [22], which minimizes a local Lipschitz objective in the neighborhood of each forget instance. Output-sensitivity suppression underlies MU-MIS [10], which uses logit-wise input gradients as an instance-contribution proxy. Layered self-distillation is used by SULI [35], which sorts forget instances by entropy and redistributes their soft labels. Adversarial proxy generation is used in ADV+IMP [6], which constructs adversarial neighbors as retain surrogates, and a related line uses generative or boundary-based proxy synthesis [8, 7]. These methods have driven significant practical progress, but their unlearning heuristics are constructed through input smoothing, output suppression, relabeling, adversarial neighborhoods, or proxy distillation. They therefore do not explicitly define a retain-free local reference in hyperspherical representation space, nor do they estimate an original-class support level from the teacher's local directional margin field. HLMR addresses this more specific gap by using teacher-evaluated geodesic probes to build a local representation-space reference for retracting point-concentrated margin excess without directly using retain instances.

Instance-wise difficulty and retain-forget entanglement. Forget instances are not equally easy to unlearn. Recent analyses show that unlearning difficulty depends on the relationship between forget and retain data, including gradient alignment, semantic proximity, and instance memorization [48, 18, 20, 21]. RUM formalizes this phenomenon through retain-forget entanglement and shows that highly entangled forget instances require different treatment from outliers [48]; related work further partitions retain instances by their entanglement with the forget set and protects them differently during unlearning [9]. These studies provide important evidence that instance-wise unlearning is governed by local retain-forget structure, but they presuppose access to retain instances or retain gradients. In the zero-shot unlearning setting, this information is unavailable, so the learner cannot directly identify which local structure is induced by the forget instance and which structure would remain supported by its neighborhood. This motivates our focus on estimating neighborhood-supported local geometry.

Hyperspherical and directional representation learning. Hyperspherical embeddings and angular-margin classifiers characterize class separation through angular distances between normalized features and class directions. SphereFace introduces angular-margin constraints on a hypersphere [34]. CosFace and ArcFace further formulate discriminative classification through cosine and additive angular margins, respectively [44, 16]. The associated geometric tools, including tangent spaces, exponential and logarithmic maps, and Riemannian optimization, are standard for modeling local motion on curved representation spaces [2]. More broadly, manifold-learning methods have long emphasized that local neighborhoods preserve meaningful geometric structure, as reflected in locally linear embedding and neighborhood-preserving embedding objectives [38, 28]. HLMR builds on these ideas and applies them to a different problem: estimating neighborhood-supported

local representation geometry for zero-shot instance-wise unlearning using geodesic probes and teacher-evaluated margins in hyperspherical latent space.

3 Preliminaries

Pretrained classifier and hyperspherical latent directions. Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$ denote the original training set, where $y_i \in \{1, \dots, C\}$. We consider a pretrained classifier f_θ with parameters $\theta = (\phi, W)$, consisting of a feature extractor $h_\phi : \mathcal{X} \to \mathbb{R}^d$ and a linear classifier $W = [w_1, \dots, w_C]^\top \in \mathbb{R}^{C \times d}$. For an input x, let $z = h_\phi(x) \in \mathbb{R}^d$ denote its latent representation. HLMR operates on the normalized latent direction $u = z/\|z\|_2 \in \mathbb{S}^{d-1}$, where $\mathbb{S}^{d-1} = \{u \in \mathbb{R}^d : \|u\|_2 = 1\}$. We similarly normalize each classifier vector as $\hat{w}_c = w_c/\|w_c\|_2$ and define the directional class score $s_c(u; \theta) = \langle u, \hat{w}_c \rangle$. We use s_c as the optimization surrogate for the local directional decision structure on $\mathbb{S}^{d-1}$. Under a bias-free normalized classifier head, this score is exactly order-equivalent to the classifier logits. Under the standard linear heads used in our experiments, s_c is an approximate directional surrogate: the raw biased logits can rank classes differently. This surrogate is appropriate for HLMR because the method explicitly targets angular original-class support around the forget direction. We discuss the exact order-equivalence conditions, empirical score–logit compatibility checks before and after unlearning, and numerical safeguards in Appendix B and Appendix C.

Hyperspherical local geometry. The normalized latent directions naturally lie on the unit hypersphere, which provides the explicit local geometric structure used by our method. For two directions $u, v \in \mathbb{S}^{d-1}$, we measure local proximity by the geodesic distance $d_{\mathbb{S}}(u, v) = \arccos(\langle u, v \rangle)$. The tangent space at u is denoted by $T_u\mathbb{S}^{d-1} = \{\xi \in \mathbb{R}^d : \langle \xi, u \rangle = 0\}$, and the orthogonal projection of a vector $a \in \mathbb{R}^d$ onto this tangent space is $\Pi_u(a) = a - \langle a, u \rangle u$. We further use the standard logarithmic and exponential maps $\mathrm{Log}_u(\cdot)$ and $\mathrm{Exp}_u(\cdot)$ (explicit forms in Appendix F) to move between the hypersphere and its local tangent space.

Zero-shot instance-wise unlearning. For exposition, we first consider a singleton forget set $\mathcal{D}_f = \{(x_f, y_f)\} \subset \mathcal{D}$, and let $\mathcal{D}_r = \mathcal{D} \setminus \mathcal{D}_f$ denote the corresponding retain set. The extension to a multi-instance forget set is given in Appendix K, where per-instance local quantities are computed separately and aggregated only at the loss level. The ideal reference is the oracle retrained model obtained by removing the forget instance and retraining on $\mathcal{D}_r$: $\theta_{\mathrm{oracle}} = \arg\min_\theta \frac{1}{|\mathcal{D}_r|} \sum_{(x,y) \in \mathcal{D}_r} \ell(f_\theta(x), y)$, where $\ell(\cdot, \cdot)$ denotes the standard classification loss. In the zero-shot setting, however, $\mathcal{D}_r$ is unavailable at unlearning time; the algorithm only has access to the pretrained teacher θ^{pre} and the forget instance (x_f, y_f). The objective is therefore to produce an updated model θ_u whose behavior approximates the oracle retraining effect without using any retain data [13]; the empirical retrained reference $\theta^\star$ used in our experiments is described in Appendix M.2.

Point-supported excess and neighborhood-supported local geometry. A forget instance can give rise to two distinct local effects. First, it induces a point-supported representation protrusion in hyperspherical latent space and excess original-class support concentrated at the forget direction. Second, it also lies within neighborhood-supported local representation geometry that would remain after the forget instance is removed, because nearby retained instances still support the same class-relevant structure. Under zero-shot unlearning, the true retain neighborhood is unavailable, so the learner cannot directly determine which part of the local original-class support is instance-specific and which part would still be preserved by retraining. Operationally, we treat point-supported margin excess as the positive gap between the forget point's original-class margin and a conservative local baseline estimated from nearby geodesic probes. We treat geometric protrusion as the deviation of the current forget direction from a weighted low-margin local anchor in the same probe neighborhood. These operational definitions lead directly to the lower-quantile baseline, retraction criticality, geometric anchor, and coupled losses in Section 4. In this paper, *Retraction* is used descriptively to mean a directional-margin pull-back of the forget representation and its original-class support toward a neighborhood-supported local reference on $\mathbb{S}^{d-1}$.

4 Methodology

Given a pretrained teacher θ^{pre} and a forget instance (x_f, y_f), HLMR constructs a student-centered but teacher-evaluated local geometry around the current forget direction. The probe neighborhood provides an operational proxy for local class support. From this neighborhood, HLMR derives a conservative local margin baseline and a bounded retraction anchor, which define a geometric retraction loss and a margin retraction loss. In addition, HLMR uses a parameter-space stabilizer on the last trainable representation block to reduce optimization drift. The first two terms are probe-derived local objectives, whereas the third term is a regularizer. Additional geometric definitions and implementation details are in Appendix F–J.

Teacher-evaluated local probes around a detached current center. Since true retained neighbors are unavailable in the zero-shot setting, we form the local center from the current unlearning model, but evaluate all probe margins under the frozen pretrained teacher. At refresh step t, the detached current forget direction is

$$u_{f,t}^{\mathrm{ctr}} = \mathrm{stopgrad}\left(\frac{h_{\phi_t}(x_f)}{\|h_{\phi_t}(x_f)\|_2}\right) \in \mathbb{S}^{d-1}. \tag{1}$$

We sample tangent perturbations $\xi_{k,t} \overset{\text{i.i.d.}}{\sim} P_r$ in $T_{u_{f,t}^{\mathrm{ctr}}}\mathbb{S}^{d-1}$, where P_r draws a random tangent direction with bounded radius $\|\xi_{k,t}\|_2 \le r < \pi$, and map them to probe directions

$$u_{f,t}^{(k)} = \mathrm{Exp}_{u_{f,t}^{\mathrm{ctr}}}(\xi_{k,t}), \qquad k = 1, \dots, K.$$

Probe margins are evaluated under the frozen pretrained teacher θ^{pre}, yielding a teacher-evaluated, student-centered local geometry: the center moves with the optimization, while the margin field stays fixed at θ^{pre}. This construction assumes that the teacher's directional margin field is locally informative within a small geodesic neighborhood around the current forget direction; large probe radii violate this locality assumption even when the hyperspherical operators remain mathematically well-defined. The periodic-refresh schedule is described in Appendix I.

Directional original-class margin. We measure local class preference using the normalized directional score introduced in Section 3. For the forget label y_f, we define the original-class margin field as

$$m_{y_f}(u;\theta) = s_{y_f}(u;\theta) - \max_{j \neq y_f} s_j(u;\theta). \tag{2}$$

This normalized directional margin depends only on the latent direction and classifier orientation, and is therefore consistent with the hyperspherical geometry on which our method operates. As discussed in Appendix B, it serves as a directional surrogate for local decision structure. The raw biased-logit margin also depends on feature and classifier norms and on class biases.

Near-neighborhood lower-quantile baseline. We answer the question "to what local support level should the forget point be retracted?" by estimating a conservative neighborhood-supported margin level from the near probes, evaluated under θ^{pre}. Let $N_f \subset \{1, \dots, K\}$ denote the index set of the K_{near} nearest probes to $u_{f,t}^{\mathrm{ctr}}$, where $1 \le K_{\mathrm{near}} \le K$. We define

$$m_{f,t}^{\mathrm{pre}} = m_{y_f}(u_{f,t}^{\mathrm{ctr}};\theta^{\mathrm{pre}}), \qquad m_{f,t}^{(k)} = m_{y_f}(u_{f,t}^{(k)};\theta^{\mathrm{pre}}).$$

The neighborhood-supported baseline is the empirical lower q-quantile over the near-probe set:

$$m_{f,t}^{\mathrm{base}} = \mathrm{Quantile}_q\left(\{m_{f,t}^{(k)} : k \in N_f\}\right), \qquad q \in (0,1). \tag{3}$$

We use a lower quantile because probe directions still influenced by the forget-centered protrusion can elevate the mean and over-preserve original-class support; the choice of q, the quantile convention, and a comparison against alternatives are discussed in Appendix G.

Local retraction criticality and adaptive retraction strength. Not all forget instances exceed their neighborhood-supported baseline to the same extent. We quantify this positive local excess by

$$\omega_{f,t} = \mathbf{1}[m_{f,t}^{\mathrm{pre}} > 0] \cdot \mathrm{clip}\left(\frac{[m_{f,t}^{\mathrm{pre}} - m_{f,t}^{\mathrm{base}}]_+}{m_{f,t}^{\mathrm{pre}} + \delta}, 0, \omega_{\max}\right), \tag{4}$$

where $[a]_+ = \max(a, 0)$, $\delta > 0$ prevents numerical blow-up for small positive margins, and $\omega_{\max} > 0$ caps extreme values. This definition activates adaptive amplification only when the pretrained teacher assigns a positive original-class directional advantage to the forget direction. The criticality is a bounded update-strength score relative to the probe-derived local reference; it is not an estimate of the instance-wise retraining response.

We convert $\omega_{f,t}$ into an instance-wise retraction multiplier $\alpha_{f,t} = 1 + \kappa\omega_{f,t}$, and set

$$\tau_{f,t} = \min(\tau_0\alpha_{f,t}, \tau_{\max}), \qquad \beta_{f,t} = \beta_0\alpha_{f,t},$$

where $\tau_0, \beta_0 > 0$ and $\kappa \geq 0$ are base coefficients and $\tau_{\max} < r < \pi$ keeps the geometric anchor within the intended local probe radius.

Neighborhood-supported retraction anchor. The desired target should be a direction derived from probes that are simultaneously nearby and less supportive of the original class. We assign each probe a joint softmax weight

$$\pi_{k,t} = \frac{\exp\left(-\frac{d_{\mathbb{S}}^2(u_{f,t}^{(k)}, u_{f,t}^{\mathrm{ctr}})}{\sigma^2} - \frac{m_{f,t}^{(k)}}{T_m}\right)}{\sum_{j=1}^{K} \exp\left(-\frac{d_{\mathbb{S}}^2(u_{f,t}^{(j)}, u_{f,t}^{\mathrm{ctr}})}{\sigma^2} - \frac{m_{f,t}^{(j)}}{T_m}\right)}, \tag{5}$$

where $\sigma > 0$ is an angular scale and $T_m > 0$ is a margin temperature. The first term favors nearby probes, while the second favors low-margin probes. The aggregated tangent direction is

$$\bar{\xi}_{f,t} = \sum_{k=1}^{K} \pi_{k,t}\xi_{k,t} \in T_{u_{f,t}^{\mathrm{ctr}}}\mathbb{S}^{d-1}.$$

The retraction anchor is

$$u_{f,t}^{\mathrm{anc}} = \mathrm{Exp}_{u_{f,t}^{\mathrm{ctr}}}\left(\tau_{f,t}\frac{\Pi_{u_{f,t}^{\mathrm{ctr}}}(\bar{\xi}_{f,t})}{\|\Pi_{u_{f,t}^{\mathrm{ctr}}}(\bar{\xi}_{f,t})\|_2 + \epsilon}\right). \tag{6}$$

The anchor uses probe information only for direction selection; its step magnitude is independently controlled by $\tau_{f,t} \leq \tau_{\max} < r$. When $\|\bar{\xi}_{f,t}\|_2 < \delta_{\mathrm{anc}}$, we set $u_{f,t}^{\mathrm{anc}} = u_{f,t}^{\mathrm{ctr}}$ for numerical stability.

Local geometric retraction loss. Under the current model parameters θ, let $u_f(\theta) = h_\phi(x_f; \theta)/\|h_\phi(x_f; \theta)\|_2$. We relocate the current forget representation toward the (detached) retraction anchor via the squared geodesic distance

$$\mathcal{L}_{\mathrm{geo}}(\theta) = d_{\mathbb{S}}^2\big(u_f(\theta),\, \mathrm{stopgrad}(u_f^{\mathrm{anc}})\big). \tag{7}$$

This term targets the instance-induced local protrusion at the representation level. Because $u_{f,t}^{\mathrm{anc}}$ is constructed within a bounded $\tau_{f,t}$-ball around the refreshed center $u_{f,t}^{\mathrm{ctr}}$, and the local geometry is refreshed periodically, optimization is kept within a local spherical neighborhood, which reduces the risk of antipodal ill-conditioning.

Local margin retraction loss. Representation relocation alone is insufficient if the forget point still retains an excessive original-class directional advantage after being pulled toward the anchor. We therefore penalize only the positive excess over the neighborhood-supported baseline:

$$\mathcal{L}_{\mathrm{ret}}(\theta) = \beta_{f,t}\left[m_{y_f}(u_f(\theta); \theta) - \mathrm{stopgrad}(m_{f,t}^{\mathrm{base}})\right]_+^2. \tag{8}$$

This loss is active only when the current margin exceeds the local baseline, thereby reducing instance-specific excess without explicitly pushing the forget direction below the neighborhood-supported support level. The non-smooth points (non-unique competing class) are handled with standard subgradients (Appendix E).

Restricted update scope and auxiliary parameter stabilizer. We split $\theta = (\theta_{\mathrm{frz}}, \theta_{\mathrm{rep}}, W)$ into frozen earlier layers, the last trainable representation block, and the classifier head, and update only $(\theta_{\mathrm{rep}}, W)$. To reduce representation-block drift without pulling the classifier head back toward the original decision boundary, we use an auxiliary parameter-space stabilizer on θ_{rep}:

$$\mathcal{L}_{\mathrm{param}}(\theta) = \frac{1}{d_{\mathrm{rep}}} \left\| \mathrm{vec}(\theta_{\mathrm{rep}}) - \mathrm{vec}(\theta_{\mathrm{rep}}^{\mathrm{pre}}) \right\|_2^2, \tag{9}$$

where d_{rep} is the number of scalar parameters in the trainable representation block. The classifier head W is deliberately not anchored, allowing the local decision boundary to retract freely. This term acts as a trust-region stabilizer.

Final objective. The total objective within a refresh interval is

$$\mathcal{L}_{\mathrm{total}}(\theta; \eta_t) = \mathcal{L}_{\mathrm{geo}}(\theta; \eta_t) + \mathcal{L}_{\mathrm{ret}}(\theta; \eta_t) + \lambda_{\mathrm{param}} \mathcal{L}_{\mathrm{param}}(\theta), \tag{10}$$

where $\lambda_{\mathrm{param}} \geq 0$. The detached local state is $\eta_t = (u_{f,t}^{\mathrm{ctr}}, m_{f,t}^{\mathrm{base}}, \omega_{f,t}, \tau_{f,t}, \beta_{f,t}, u_{f,t}^{\mathrm{anc}})$. Within each refresh interval, η_t is treated as fixed, so optimization minimizes a piecewise-stationary surrogate. Across refresh steps, scalar quantities are optionally EMA-smoothed, while the directional anchor is reconstructed from the refreshed tangent geometry, as detailed in Appendix I.

5 Experiments

Experimental setup. We evaluate zero-shot instance-wise unlearning on three benchmarks with increasing scale and architectural diversity: CIFAR-10 with ResNet-18 [31, 27], CIFAR-100 with ResNet-50 [31, 27], and Tiny-ImageNet with ViT-S/16 [33, 15, 17]. The forget set is sampled in a class-stratified manner so that its class prior matches the original training distribution; zero-shot methods access only the pretrained model and the forget set during unlearning. All results are reported as mean and standard deviation over five unlearning seeds. Preprocessing, training schedules, split construction, and implementation settings are in Appendix M.

Evaluation protocol and metrics. We assess approximate unlearning empirically through retraining alignment, utility preservation, and membership-inference attacks. Following prior work, we evaluate each method from three complementary perspectives: unlearning quality, model utility, and runtime efficiency [24, 13, 32, 48]. In the instance-wise setting, these metrics must be interpreted relative to the oracle retrained model. A forgotten instance can still be correctly classified after successful unlearning if its class-relevant information is supported by retained instances. Therefore, good zero-shot instance-wise unlearning is defined as oracle-aligned unlearning: the post-unlearning model should behave similarly to the oracle retrained model on the forget set, while preserving utility on the retain set and the test set.

For utility, we report **Test acc.** (Test set accuracy) and **Retain acc.** (Retain set accuracy), which measure classification accuracy on the test set and retain set, respectively. We also report **Forget acc.** (Forget set accuracy), the classification accuracy on the forget set. Unlike class-wise unlearning, forget set accuracy is not interpreted as an isolated objective in the instance-wise setting; instead, we compare it with the retrained model. A substantially higher value than the oracle indicates under-unlearning, while a substantially lower value indicates over-unlearning [29] or abnormal behavior. For unlearning quality, we use membership inference attacks to measure whether the forget instances remain distinguishable from non-members. The loss-based membership inference attack (**Loss-MIA**) follows the loss-threshold attack of Yeom et al. [47, 32]: it uses the per-instance cross-entropy loss to distinguish the forget set from a class-matched test set. We further evaluate the stronger likelihood-ratio membership inference attack (**LiRA-MIA**) [5, 32], which uses shadow models to estimate in/out likelihoods and serves as a confirmatory privacy evaluation. For efficiency, we report **Runtime**, the wall-clock time in seconds required to complete the unlearning procedure under the same hardware and software environment. Additional details on the evaluation protocol, model selection, and fairness control are provided in Appendix L.

Baselines. We compare our HLMR with state-of-the-art zero-shot unlearning baselines and the retraining reference. **Retrain** trains the same architecture from scratch on the retain set and serves as the gold standard. **Negative Gradient** [42] (NegaGrad) maximizes the classification loss on the forget

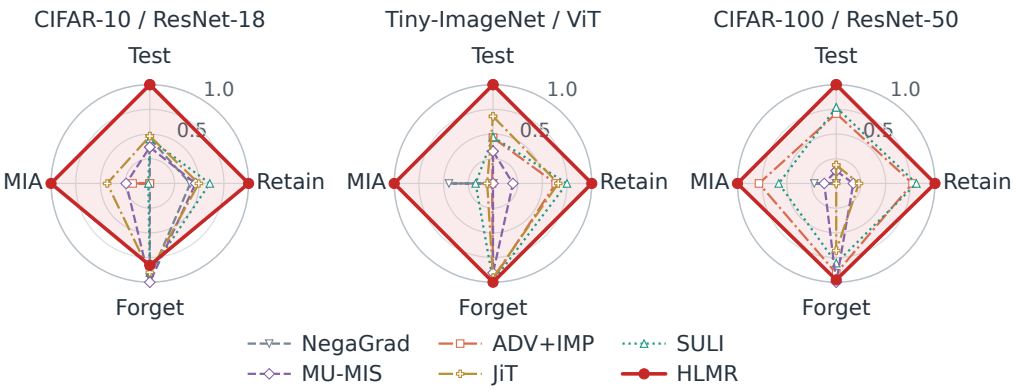


Figure 1: **Relative retraining alignment across benchmarks.** HLMR aligns most consistently with the retrained model. For each dataset and metric, g is the absolute method–retraining mean gap averaged over three forget ratios. Radial scores $s = (g_{\max} - g)/(g_{\max} - g_{\min})$ use extrema across the six methods on that axis. Higher is better; $s = 1$ denotes the smallest observed gap. MIA denotes Loss-MIA.

set to push the model away from fitting the revoked instances. **ADV+IMP** [6] combines adversarial neighbor generation with importance-weighted parameter updates. **MU-MIS** [10] suppresses the contribution of instances through sensitivity-based objectives. **SULI** [35] performs self-unlearning through entropy-based layered iterations and soft-label redistribution. **JiT** [22] reduces local sensitivity around forget instances by minimizing a local Lipschitz-style objective. Detailed hyperparameters and implementation choices for each baseline are given in Appendix N. HLMR and the zero-shot baselines operate under the same information budget, namely the pretrained model and the forget set; Table 19 compares the information budget and deletion task of related method families.

6 Results and Discussion

Overview. Since HLMR optimizes a hyperspherical directional margin while evaluation uses the raw-logit classifier, we additionally verify post-unlearning score–logit compatibility in Appendix B. After HLMR unlearning, top-1 agreement between directional-score and raw-logit predictions remains above 0.94, and both Pearson and Spearman margin correlations remain above 0.94 across all evaluated benchmark–architecture settings. This indicates that the directional-margin objective remains empirically aligned with the raw-logit decision margin after updating the classifier head. Figure 1 summarizes retraining alignment across the three benchmarks. Table 1 summarizes CIFAR-100/ResNet-50 and Tiny-ImageNet/ViT performance, covering both CNN and attention-based architectures. CIFAR-10 results and per-ratio breakdowns are deferred to Appendix O. See the ablation subsection below and Appendix P for the complete ablation study.

Unlearning utility. Table 1 shows that HLMR achieves the strongest zero-shot utility while remaining closest to the retrained reference. Across all nine benchmark–ratio settings (Appendix O), HLMR delivers the highest test and retain accuracy among zero-shot methods, improving on the strongest baseline by about 4 percentage points on average; this is consistent with the design goal of preserving neighborhood-supported structure.

Unlearning quality. In instance-wise unlearning, forget accuracy should not be interpreted as a quantity to minimize in isolation. A forgotten instance can remain correctly classified when its class-relevant evidence is also supported by retained instances; the appropriate goal is therefore to match the behavior of the retrained oracle. HLMR has the smallest Loss-MIA gap in Table 1 and is closest to the oracle in eight of the nine settings, two of them by margins within one standard deviation of the runner-up (Appendix O). This indicates that its forgotten instances lose the membership evidence retained by the original model more consistently than competing zero-shot methods. The likelihood-ratio membership inference attack provides a stronger unlearning quality evaluation [32].

Table 1: Compact utility–quality summary on CIFAR-100 / ResNet-50 and Tiny-ImageNet / ViT. Each entry averages the corresponding metric over the three forget ratios reported in Appendix O. **HLMR has the highest test and retain accuracy and the smallest Loss-MIA gap among the evaluated zero-shot methods in this averaged summary.**

	CIFAR-100 / ResNet-50				Tiny-ImageNet / ViT			
Method	Test acc. (↑)	Retain acc. (↑)	Forget acc.	Loss-MIA gap (↓)	Test acc. (↑)	Retain acc. (↑)	Forget acc.	Loss-MIA gap (↓)
Original	83.29	99.97	100.00	0.13	85.05	99.98	100.00	0.10
Retraining	82.67	99.96	82.49	0.00	84.85	99.88	83.61	0.00
NegaGrad	64.41 ± 7.02	81.62 ± 9.04	79.83 ± 8.19	0.06 ± 0.02	63.23 ± 19.95	75.61 ± 24.43	69.59 ± 23.50	0.05 ± 0.19
ADV+IMP	75.61 ± 0.41	94.66 ± 0.29	82.51 ± 1.10	0.02 ± 0.04	71.78 ± 2.42	89.74 ± 2.20	84.14 ± 2.37	0.06 ± 0.05
SULI	76.59 ± 0.26	95.33 ± 0.29	83.25 ± 2.08	0.04 ± 0.02	71.91 ± 2.19	92.29 ± 3.04	84.58 ± 3.55	0.05 ± 0.01
MU-MIS	66.43 ± 2.61	84.54 ± 3.49	82.22 ± 2.91	0.07 ± 0.03	69.24 ± 5.39	80.10 ± 6.25	86.11 ± 3.94	0.06 ± 0.04
JiT	67.36 ± 0.30	85.51 ± 0.36	83.09 ± 0.72	0.08 ± 0.03	75.77 ± 0.20	90.32 ± 0.29	85.37 ± 0.32	0.06 ± 0.01
HLMR	**80.23 ± 0.57**	**98.51 ± 0.52**	82.14 ± 0.82	**0.01 ± 0.02**	**81.70 ± 0.27**	**97.88 ± 0.38**	84.70 ± 0.70	**0.03 ± 0.01**

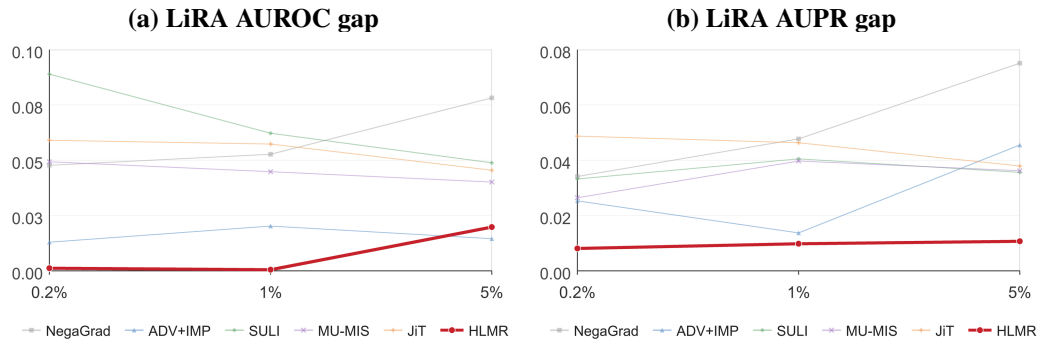


Figure 2: LiRA-MIA three-point profile on CIFAR-100 / ResNet-50. HLMR empirically reduces the membership-inference heuristic and aligns with the retraining oracle most consistently among evaluated zero-shot baselines. The y-axis is the absolute gap to retraining, so lower is better.

As shown in Figure 2, HLMR attains the smallest AUROC gap at the 0.2% and 1% ratios and the smallest AUPR gap at all three ratios; at 5%, ADV+IMP attains a marginally smaller AUROC gap (0.0145 vs. 0.0198) at $\sim 16\times$ the runtime. The complete table is in Appendix O.

Validating the probe-derived local reference. On CIFAR-100/ResNet-50, we compare the probe baseline with the same lower quantile computed over the 128 nearest same-class retained embeddings. Across random, feature-clustered, and boundary-conditioned forget sets, Pearson correlations range from 0.776 to 0.879 and Spearman correlations from 0.783 to 0.900: the probe baseline tracks a local margin statistic of real retained examples. The probe-relative excess has near-zero correlation with the instance-wise margin change under retraining; it measures deviation from the local reference and controls retraction strength, and does not estimate the retraining response (Appendix S.3).

Structured forget sets and local retention. We further evaluate class-balanced 1% random, feature-clustered, and boundary-conditioned forget sets (500 images each), each with its own retrained reference, using one configuration frozen before any evaluation outcome is observed. On both targeted structured sets, mean absolute Test and Retain gaps remain below one percentage point, Forget gaps below four points, and Loss-MIA AUC gaps below 0.052. On the same-class retained images nearest to the feature cluster (Adjacent-Retain), HLMR achieves $98.72 \pm 0.32\%$ accuracy, the highest among the unlearning methods and closest to retraining (100%; Appendix S.2).

Sensitivity to directional-field distortion. HLMR depends on the quantitative fidelity of the teacher's directional-margin field. Rotating only the teacher classifier directions used for the probes on the boundary-conditioned set leaves the sign agreement of probe margins at 0.999 for rotation 0.8, yet Forget accuracy falls from 54.88% to 35.97%, against 58.80% for retraining; at $\pi/2$, utility nearly collapses (Appendix S.4). Sign consistency alone is thus insufficient for a useful local target.

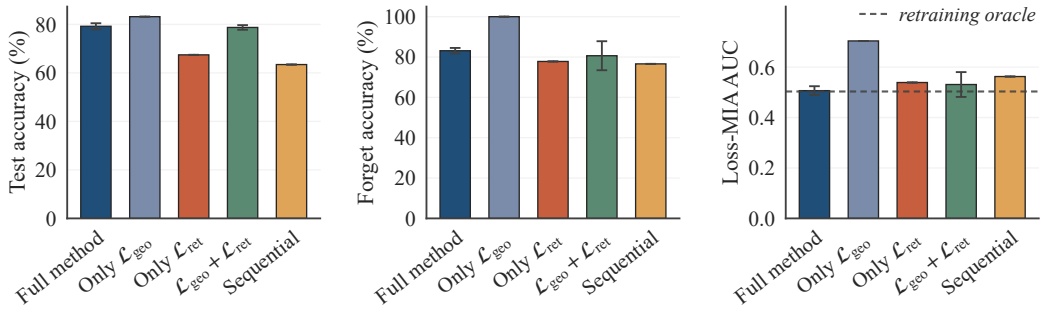


Figure 3: **Three-layer framework isolation: HLMR's full objective is the only configuration aligned with the retraining oracle.** *Only* $\mathcal{L}_{\text{geo}}$ leaves the boundary intact and forgetting fails; *only* $\mathcal{L}_{\text{ret}}$ damages utility.

Table 2: Compact ablation summary on CIFAR-100 / ResNet-50, forget ratio = 1%. Rows include the canonical baseline and two representative ablations: disabling geometry EMA and disabling dynamic geometry refresh. The completed table, hyperparameter sweeps, and design-choice ablations are in Appendix P (Table 11).

Configuration	Test acc. (%)	Retain acc. (%)	Forget acc. (%)	Sensitivity gap	Loss-MIA AUC
Canonical baseline	81.11 ± 0.25	98.37 ± 0.77	85.12 ± 0.38	-24.17 ± 1.97	0.509 ± 0.037
Disable geometry EMA	81.61 ± 0.28	99.60 ± 0.04	95.32 ± 0.61	-19.01 ± 2.19	0.591 ± 0.006
Disable dynamic geometry refresh	82.58 ± 0.23	99.88 ± 0.03	100.00 ± 0.00	-32.03 ± 2.19	0.684 ± 0.006

Unlearning efficiency. Detailed runtimes and the complexity discussion are in Appendix O. The runtime of most zero-shot unlearning methods is fairly similar, with the exception of ADV+IMP, which incurs a high computational cost because of the generation of adversarial examples.

Ablation study. We run an ablation study on CIFAR-100 / ResNet-50 (forget ratio of 1%), with five seeds per configuration. Each experiment overrides one or two fields of the canonical hyperparameter block while keeping all other settings fixed. We report test, retain, and forget accuracy, the sensitivity gap (SG; Appendix P.1), and Loss-MIA AUC. Table 2 highlights two representative configurations; the complete table and per-step training dynamics are deferred to Appendix P.

Figure 3 shows that HLMR performs best when geometric retraction, margin retraction, dynamic geometry refresh, and the auxiliary parameter stabilizer are used together. The detailed ablations in Appendix P.3 further clarify the division of roles: the main unlearning heuristic is carried by the coupled geometry–margin mechanism and by periodically refreshing the local geometry, whereas the parameter stabilizer serves as an auxiliary trust-region regularizer that reduces optimization drift. These ablations measure component roles and operating ranges under one configuration; the viable small-radius regime is $r \in [0.05, 0.20]$, with degradation at 0.50 and collapse at 1.00 (Appendix P.6).

7 Conclusion

We formulate zero-shot instance-wise unlearning as local margin retraction implemented through coupled geometric and margin objectives in hyperspherical latent space. By probing a small geodesic neighborhood around the forget direction under the pretrained teacher, HLMR estimates a lower-quantile margin baseline, an instance-wise retraction criticality, and a retraction anchor, and combines geometric and margin retraction losses to remove the instance-induced local protrusion and excess original-class support without consulting any retain data. Additional diagnostics show that the probe baseline tracks a retain-derived local margin statistic and characterize where this local reference remains reliable. Limitations and broader impact are discussed in Appendix Q and Appendix R.

Acknowledgments

Hongyi Lyu is supported by an International Macquarie University Research Excellence Scholarship (iMQRES, allocation number 20257340). The authors declare no competing interests.

References

[1] S. M. Ahmed, U. Y. Basaran, D. S. Raychaudhuri, A. Dutta, R. Kundu, F. F. Niloy, B. Guler, and A. K. Roy-Chowdhury. Towards source-free machine unlearning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 4948–4957, 2025.

[2] G. Bécigneul and O.-E. Ganea. Riemannian adaptive optimization methods. In *International Conference on Learning Representations*, 2019.

[3] J. Bonato, M. Cotogni, and L. Sabetta. Is retain set all you need in machine unlearning? restoring performance of unlearned models with out-of-distribution images. In *European Conference on Computer Vision*, pages 1–19. Springer, 2024.

[4] L. Bourtoule, V. Chandrasekaran, C. A. Choquette-Choo, et al. Machine unlearning. In *IEEE Symposium on Security and Privacy (SP)*, 2021.

[5] N. Carlini, S. Chien, M. Nasr, S. Song, A. Terzis, and F. Tramer. Membership inference attacks from first principles. In *2022 IEEE symposium on security and privacy (SP)*, pages 1897–1914. IEEE, 2022.

[6] S. Cha, S. Cho, D. Hwang, H. Lee, T. Moon, and M. Lee. Learning to unlearn: Instance-wise unlearning for pre-trained classifiers. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 11186–11194, 2024.

[7] H. Chen, T. Zhu, X. Yu, and W. Zhou. Zero-shot machine unlearning with proxy adversarial data generation. *arXiv preprint arXiv:2507.21738*, 2025.

[8] M. Chen, W. Gao, G. Liu, K. Peng, and C. Wang. Boundary unlearning. *arXiv preprint arXiv:2303.11570*, 2023.

[9] J. Cheng, P. Liu, Q. Li, and C. Zhang. Machine unlearning under retain-forget entanglement. *arXiv preprint arXiv:2603.26569*, 2026.

[10] X. Cheng, Z. Huang, W. Zhou, Z. He, R. Yang, Y. Wu, and X. Huang. Remaining-data-free machine unlearning by suppressing sample contribution. *arXiv preprint arXiv:2402.15109*, 2024.

[11] S. B. R. Chowdhury, K. Choromanski, A. Sehanobish, A. Dubey, and S. Chaturvedi. Towards scalable exact machine unlearning using parameter-efficient fine-tuning. *arXiv preprint arXiv:2406.16257*, 2024.

[12] V. S. Chundawat, A. K. Tarun, M. Mandal, and M. Kankanhalli. Can bad teaching induce forgetting? unlearning in deep networks using incompetent teachers. In *AAAI*, 2023.

[13] V. S. Chundawat, A. K. Tarun, M. Mandal, and M. Kankanhalli. Zero-shot machine unlearning. *IEEE Transactions on Information Forensics and Security*, 18:2345–2354, 2023.

[14] M. Cotogni, J. Bonato, L. Sabetta, F. Pelosin, and A. Nicolosi. DUCK: Distance-based unlearning via centroid kinematics. *arXiv preprint arXiv:2312.02052*, 2023.

[15] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. IEEE, 2009.

[16] J. Deng, J. Guo, N. Xue, and S. Zafeiriou. ArcFace: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019.

[17] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[18] C. Fan, J. Liu, A. Hero, and S. Liu. Challenging forgets: Unveiling the worst-case forget sets in machine unlearning. In *ECCV*, 2024.

[19] C. Fan, J. Liu, Y. Zhang, E. Wong, D. Wei, and S. Liu. SalUn: Empowering machine unlearning via gradient-based weight saliency in both image classification and generation. In *International Conference on Learning Representations*, 2024.

[20] V. Feldman. Does learning require memorization? a short tale about a long tail. *ACM Symposium on Theory of Computing (STOC)*, 2020.

[21] V. Feldman and C. Zhang. What neural networks memorize and why: Discovering the long tail via influence estimation. *NeurIPS*, 2020.

[22] J. Foster, K. Fogarty, S. Schoepf, Z. Dugue, C. Öztireli, and A. Brintrup. An information theoretic approach to machine unlearning. *arXiv preprint arXiv:2402.01401*, 2024.

[23] J. Foster, S. Schoepf, and A. Brintrup. Fast machine unlearning without retraining through selective synaptic dampening. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 12043–12051, 2024.

[24] A. Ginart, M. Guan, G. Valiant, and J. Y. Zou. Making ai forget you: Data deletion in machine learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

[25] A. Golatkar, A. Achille, and S. Soatto. Eternal sunshine of the spotless net: Selective forgetting in deep networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9304–9312, 2020.

[26] C. Guo, T. Goldstein, A. Hannun, and L. Van Der Maaten. Certified data removal from machine learning models. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 3832–3842. PMLR, 2020.

[27] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[28] X. He, D. Cai, S. Yan, and H.-J. Zhang. Neighborhood preserving embedding. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, volume 2, pages 1208–1213. IEEE, 2005.

[29] H. Hu, S. Wang, J. Chang, H. Zhong, R. Sun, S. Hao, H. Zhu, and M. Xue. A duty to forget, a right to be assured? exposing vulnerabilities in machine unlearning services. *arXiv preprint arXiv:2309.08230*, 2023.

[30] Z. Izzo, M. A. Smart, K. Chaudhuri, and J. Zou. Approximate data deletion from machine learning models. In *International conference on artificial intelligence and statistics*, pages 2008–2016. PMLR, 2021.

[31] A. Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

[32] M. Kurmanji, P. Triantafillou, J. Hayes, and E. Triantafillou. Towards unbounded machine unlearning. *Advances in neural information processing systems*, 36:1957–1987, 2023.

[33] Y. Le and X. Yang. Tiny imagenet visual recognition challenge. *CS231N Course Project Report*, 2015.

[34] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song. SphereFace: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 212–220, 2017.

[35] H. Lyu, X. Zhang, H. Hu, S. Wang, C. He, and L. Qi. Fine-grained and efficient self-unlearning with layered iteration. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*, pages 7643–7651, 2025.

[36] T. T. Nguyen, T. T. Huynh, Z. Ren, P. L. Nguyen, A. W.-C. Liew, H. Yin, and Q. V. H. Nguyen. A survey of machine unlearning. *ACM Transactions on Intelligent Systems and Technology*, 16(5):1–46, 2025.

[37] X. Pennec. Intrinsic statistics on riemannian manifolds: Basic tools for geometric measurements. *Journal of Mathematical Imaging and Vision*, 25:127–154, 2006.

[38] S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *science*, 290(5500):2323–2326, 2000.

[39] R. Shokri, M. Stronati, C. Song, and V. Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pages 3–18. IEEE, 2017.

[40] C. N. Spartalis, T. Semertzidis, P. Daras, and E. Gavves. LoTUS: Large-scale machine unlearning with a taste of uncertainty. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

[41] A. K. Tarun, V. S. Chundawat, M. Mandal, and M. Kankanhalli. Fast yet effective machine unlearning. *IEEE Transactions on Neural Networks and Learning Systems*, 35(9):13046–13055, 2024.

[42] A. Thudi, G. Deza, V. Chandrasekaran, and N. Papernot. Unrolling sgd: Understanding factors influencing machine unlearning. In *2022 IEEE 7th European Symposium on Security and Privacy (EuroS&P)*, pages 303–319. IEEE, 2022.

[43] P. Voigt and A. Von dem Bussche. The eu general data protection regulation (gdpr). *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, 10(3152676):10–5555, 2017.

[44] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5265–5274, 2018.

[45] J. Wu and M. Harandi. Munba: Machine unlearning via nash bargaining. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4754–4765, 2025.

[46] R. Yang, M. He, Z. He, Y. Qiu, and X. Huang. Muso: achieving exact machine unlearning in over-parameterized regimes. *Machine Learning*, 114(8):176, 2025.

[47] S. Yeom, I. Giacomelli, M. Fredrikson, and S. Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st computer security foundations symposium (CSF)*, pages 268–282. IEEE, 2018.

[48] K. Zhao, M. Kurmanji, G.-O. Bărbulescu, E. Triantafillou, and P. Triantafillou. What makes unlearning hard and what to do about it. *Advances in Neural Information Processing Systems*, 37:12293–12333, 2024.

A Additional Notations

For clarity, K denotes the number of geodesic probes and $K_{\mathrm{near}} \leq K$ denotes the number of nearest probes used to construct the local baseline. We use $T_m > 0$ for the margin-temperature in probe weighting and N_{step} for the number of optimization steps. The probe weights are denoted by π_k, while ϱ denotes the forget ratio in experiments. Unless otherwise stated, all local retraction quantities are stop-gradient detached within each refresh interval.

B Directional Surrogate Score and Score–Logit Compatibility

The directional class score $s_c(u;\theta) = \langle u, \hat{w}_c \rangle$ used in the main text is invariant to feature and classifier norms and is aligned with the hyperspherical geometry on which HLMR operates. However, it does not in general coincide with the raw biased logit

$$\ell_c(x) = w_c^\top z + b_c = \|z\|_2 \|w_c\|_2 \langle u, \hat{w}_c \rangle + b_c.$$

Thus, s_c is order-equivalent to ℓ_c only when the classifier head is bias-free and the class weight vectors are norm-balanced, or when a normalized classifier with a shared scale is used. Under the standard biased linear heads used in our experiments, s_c serves as a geometry-consistent directional surrogate for raw logits. The diagnostic below verifies that this surrogate remains empirically compatible with the raw-logit classifier after HLMR updates the classifier head.

Post-unlearning score–logit compatibility. Because HLMR optimizes the normalized directional score $s_c(u;\theta) = \langle u, \hat{w}_c \rangle$, whereas standard classifiers make predictions using raw logits $\ell_c(x) = w_c^\top z + b_c$, we evaluate whether the directional surrogate remains compatible with the raw-logit classifier after HLMR updates the classifier head. This diagnostic is not used during unlearning, hyperparameter selection, checkpoint selection, or membership-inference evaluation.

We report four quantities. Weight-norm CV is the coefficient of variation of classifier row norms $\{\|w_c\|_2\}_{c=1}^C$. Bias / linear ratio measures the average magnitude of the bias term relative to the linear logit contribution. Top-1 agreement compares $\arg\max_c s_c(u;\theta)$ with $\arg\max_c \ell_c(x)$. Pearson and Spearman correlations are computed between the true-label directional margin

$$m_y^{\mathrm{dir}}(x) = s_y(u;\theta) - \max_{j \neq y} s_j(u;\theta)$$

and the true-label raw-logit margin

$$m_y^{\mathrm{raw}}(x) = \ell_y(x;\theta) - \max_{j \neq y} \ell_j(x;\theta).$$

These quantities measure whether the hyperspherical margin optimized by HLMR continues to track the raw-logit decision margin after unlearning.

Diagnostic split protocol. For this diagnostic, we use five independent class-stratified random forget splits generated by split seeds $42, 43, 44, 45, 46$. These seeds are not cross-validation folds and do not correspond to independently trained original models. For each seed, we rebuild the forget/retain split by shuffling class-wise indices with the corresponding split seed and load the matching HLMR checkpoint whose filename encodes the same architecture, forget ratio, and seed. The forget-split statistics therefore reflect variability across independently sampled forget sets. The test-split statistics are computed on the same held-out test set, but vary across seeds because each seed produces a different forget split and hence a different HLMR-unlearned checkpoint. Thus, the reported standard deviations quantify split-induced checkpoint variability with the pretrained model held fixed.

Table 3 shows that HLMR does not materially break the empirical alignment between the directional surrogate and the raw-logit classifier after unlearning. The classifier weight norms remain reasonably balanced, with weight-norm CV at most 0.080, and the bias contribution remains small relative to the linear logit term, with bias / linear ratio at most 0.013. Top-1 agreement between directional-score and raw-logit predictions remains above 0.94 on both forget and test splits across all benchmark–architecture settings. More importantly, both Pearson and Spearman correlations between directional margins and raw-logit margins remain above 0.94, indicating that the margin geometry optimized by HLMR remains strongly aligned with the raw-logit decision margin used for evaluation. These diagnostics support the hyperspherical directional margin as an empirical surrogate for local decision structure after unlearning under the biased linear heads tested here.

Table 3: **Post-unlearning compatibility between directional scores and raw logits.** Statistics are computed after HLMR unlearning at the 1% forget ratio and reported as mean $\pm$ standard deviation over five independently sampled class-stratified forget splits generated by split seeds 42–46. We report diagnostics on both the forget split and the test split. The table checks whether updating the classifier head materially changes the empirical alignment between the hyperspherical directional surrogate optimized by HLMR and the raw-logit classifier used for evaluation.

Setting	Split	Weight-norm CV ↓	Bias / linear ratio ↓	Top-1 agreement ↑	Pearson ↑	Spearman ↑
CIFAR-10 / ResNet-18	Forget	0.031 ± 0.000	0.013 ± 0.000	0.989 ± 0.008	0.967 ± 0.004	0.970 ± 0.005
CIFAR-10 / ResNet-18	Test	0.031 ± 0.000	0.013 ± 0.000	0.984 ± 0.003	0.969 ± 0.002	0.971 ± 0.002
CIFAR-100 / ResNet-50	Forget	0.080 ± 0.001	0.007 ± 0.000	0.966 ± 0.011	0.952 ± 0.007	0.942 ± 0.009
CIFAR-100 / ResNet-50	Test	0.080 ± 0.001	0.007 ± 0.000	0.962 ± 0.003	0.972 ± 0.001	0.968 ± 0.002
Tiny-ImageNet / ViT-S/16	Forget	0.038 ± 0.000	0.003 ± 0.000	0.943 ± 0.013	0.976 ± 0.001	0.973 ± 0.002
Tiny-ImageNet / ViT-S/16	Test	0.038 ± 0.000	0.003 ± 0.000	0.970 ± 0.002	0.993 ± 0.000	0.990 ± 0.000

Directional-margin field regularity. For fixed unit-normalized classifier directions, the teacher's directional margin field of Eq. (2) satisfies, for all $u, v \in \mathbb{S}^{d-1}$,

$$\left|m_y(u;\theta^{\mathrm{pre}}) - m_y(v;\theta^{\mathrm{pre}})\right| \le 2\|u-v\|_2 \le 2d_{\mathbb{S}}(u,v). \tag{11}$$

Indeed, each linear score obeys $|\langle u-v, \hat{w}_j\rangle| \le \|u-v\|_2$ by Cauchy–Schwarz, and the difference between two maxima is bounded by the largest difference between their corresponding scores. Applying these two bounds to the original-class score and its strongest competitor proves the first inequality; the second holds because the chord length $\|u-v\|_2 = 2\sin\big(d_{\mathbb{S}}(u,v)/2\big)$ does not exceed the geodesic distance. Competitor switches can make the field non-differentiable, but preserve this continuity and Lipschitz bound. For paired probe margins perturbed by at most ε, the corresponding empirical quantiles also differ by at most ε, because their order statistics and interpolated quantiles obey the same uniform bound.

These properties establish numerical regularity. A useful local reference additionally requires quantitative directional fidelity and semantic locality. A globally applied positive temperature scale changes probabilities while leaving normalized classifier directions and HLMR's probe targets unchanged; probability calibration and fidelity of this directional field are therefore distinct. The controlled field-corruption experiment in Appendix S.4 tests the latter.

C Numerical Safeguards

We collect here the numerical safeguards used in the implementation that are referenced from the main text.

Inner products and Exp/Log maps. Inner products $\langle u, v\rangle$ are clamped to $[-1, 1]$ before $\arccos$; small thresholds replace divisions by $\sin\omega$ or $\|\xi\|_2$ when these quantities are near zero. The local logarithmic map is well-defined whenever $v \neq -u$, and the exponential map is injective for $\|\xi\|_2 < \pi$; the probe radius r is restricted to $r < \pi$. The canonical ablation uses $r = 0.20$, and the separate radius sweep extends to $r = 1.00$ (Appendix P.6).

Criticality and adaptive step size. For a fixed positive numerator, the unstabilized criticality ratio underlying Eq. (4) grows as the positive teacher margin approaches zero. We therefore use a stabilizer $\delta > 0$ in the denominator (replacing $|m_f^{\mathrm{pre}}|$ by $\max(m_f^{\mathrm{pre}}, 0) + \delta$ so the ratio is well-defined for $m_f^{\mathrm{pre}} \le 0$) and a hard cap $\omega_{\max}$. The adaptive step size τ_f is correspondingly bounded by $\tau_{\max} < r < \pi$ via $\tau_f = \min(\tau_0\,\alpha_f, \tau_{\max})$, which keeps the geometric retraction step inside the local geodesic ball used for probing. Without these safeguards, configurations with small $|m_f^{\mathrm{pre}}|$ can drive ω_f to two-digit magnitudes and cause β_f to dominate $\mathcal{L}_{\mathrm{ret}}$ at training time.

Anchor zero-aggregate branch. When the weighted tangent direction $\bar{\xi}_f$ in Eq. (6) has near-zero magnitude, both numerator and denominator of $\Pi(\bar{\xi}_f)/\|\Pi(\bar{\xi}_f)\|_2$ become unreliable. We therefore use the explicit branch

$$u_f^{\mathrm{anc}} = \begin{cases} \mathrm{Exp}_{u_f^{\mathrm{ctr}}}\big(\tau_f\,\Pi_{u_f^{\mathrm{ctr}}}(\bar{\xi}_f)/(\|\Pi_{u_f^{\mathrm{ctr}}}(\bar{\xi}_f)\|_2 + \varepsilon)\big), & \|\bar{\xi}_f\|_2 \ge \delta_{\mathrm{anc}}, \\ u_f^{\mathrm{ctr}}, & \|\bar{\xi}_f\|_2 < \delta_{\mathrm{anc}}, \end{cases}$$

with a small numerical threshold $\delta_{\mathrm{anc}} > 0$. The fallback $u_f^{\mathrm{anc}} = u_f^{\mathrm{ctr}}$ implies a vanishing geometric retraction loss for that step, which is the desired behavior in the absence of a clear retraction direction.

D Geometric Interpretation of the Anchor

The anchor in Eq. (6) uses the tangent direction from a first-order weighted manifold-mean construction, with a separately controlled step length. Within the local injectivity radius, $\mathrm{Log}_{u_f^{\mathrm{ctr}}}(u_f^{(k)}) = \xi_k$, so the aggregated tangent vector $\bar{\xi}_f = \sum_k \pi_k \xi_k$ lies in $T_{u_f^{\mathrm{ctr}}}\mathbb{S}^{d-1}$ in exact arithmetic; the explicit projection $\Pi_{u_f^{\mathrm{ctr}}}(\bar{\xi}_f)$ is retained only for numerical robustness. Mapping $\bar{\xi}_f$ back to the sphere by $\mathrm{Exp}_{u_f^{\mathrm{ctr}}}$ is the first-order intrinsic-mean construction of [37], applied to a probe ensemble whose weights π_k favor low-margin nearby directions. Two design choices distinguish HLMR from a plain weighted Fréchet/Karcher mean: (i) the magnitude of the retraction step is controlled separately by τ_f (not by $\|\bar{\xi}_f\|$), so probes contribute only their direction; and (ii) the anchor is evaluated under the frozen teacher θ^{pre} and stop-gradient detached, so it functions as a fixed external retraction target throughout each refresh interval.

E Optimization Details

The directional margin $m_{y_f}(u;\theta) = s_{y_f}(u;\theta) - \max_{j \neq y_f} s_j(u;\theta)$ is differentiable when the active competing class $j^\star(u) = \arg\max_{j \neq y_f} s_j(u;\theta)$ is unique; in that regime $\nabla_u m_{y_f} = \hat{w}_{y_f} - \hat{w}_{j^\star}$. At points where $j^\star$ is non-unique, $\max$ admits only a subdifferential, and the hinge-square in Eq. (8) is correspondingly piecewise-differentiable. We use the standard subgradients of $\max$ and $[\cdot]_+$ that automatic differentiation frameworks provide; this is well-defined almost everywhere and matches the convention used by the multiclass margin losses.

F Additional Geometric Definitions

For completeness, we summarize the hyperspherical operators used in the main text. Given $u, v \in \mathbb{S}^{d-1}$, the geodesic distance is $d_{\mathbb{S}}(u,v) = \arccos(\langle u, v\rangle)$. The tangent space at u is $T_u\mathbb{S}^{d-1} = \{\xi \in \mathbb{R}^d : \langle \xi, u\rangle = 0\}$, and the orthogonal projection of a vector $a \in \mathbb{R}^d$ onto this tangent space is $\Pi_u(a) = a - \langle a, u\rangle u$. We use the standard logarithmic and exponential maps on the unit hypersphere:

$$\mathrm{Log}_u(v) = \frac{\omega}{\sin\omega}\big(v - \cos(\omega)u\big), \qquad \omega = d_{\mathbb{S}}(u,v), \tag{12}$$

and

$$\mathrm{Exp}_u(\xi) = \cos(\|\xi\|_2)\, u + \sin(\|\xi\|_2)\, \frac{\xi}{\|\xi\|_2}, \qquad \xi \in T_u\mathbb{S}^{d-1}. \tag{13}$$

In implementation, these expressions are evaluated with standard numerical safeguards: inner products are clamped to $[-1, 1]$ before $\arccos$, and small thresholds replace divisions by $\sin\omega$ or $\|\xi\|_2$ when these quantities are near zero. The local logarithmic map is well-defined whenever $v \neq -u$, and the exponential map is injective for $\|\xi\|_2 < \pi$; the probe radius r used by HLMR is restricted to $r < \pi$, and both the canonical radius ($r = 0.20$) and the sensitivity sweep (up to $r = 1.00$) satisfy this condition (Appendix P.6).

The probe sampling law P_r used in HLMR draws tangent perturbations as $\xi = \rho\hat{\xi}$ with $\hat{\xi} \sim \mathrm{Uniform}(T_u\mathbb{S}^{d-1} \cap S^{d-2})$ (generated by drawing a Gaussian vector and projecting onto $T_u\mathbb{S}^{d-1}$, then ℓ_2-normalizing) and $\rho \sim \mathrm{Uniform}(0, r]$. This factorizes the probe distribution into a uniform direction in tangent space and a uniform radius bounded by $r < \pi$, so that all probes remain inside the local geodesic ball.

G Why a Lower-Quantile Baseline?

Quantile convention. Throughout this paper, $\mathrm{Quantile}_q(\cdot)$ in Eq. (3) denotes the empirical lower q-quantile under the Hyndman–Fan type-7 instance-quantile convention (the default in NumPy and PyTorch); this fixes the linear-interpolation rule between order statistics so that the baseline is reproducible across implementations.

Comparison of the lower quantile, mean, and minimum. The neighborhood-supported baseline is intended to approximate a conservative local support level that remains sustainable after the instance-induced excess effect of the forget instance is removed. When probes near the forget-centered protrusion have elevated margins, their contributions raise the mean and preserve more of the local excess in the reference. By contrast, using the minimum is overly unstable and unduly sensitive to sampling noise. The empirical lower q-quantile balances these effects: it limits the contribution of elevated margins and reduces sensitivity to low outliers [38, 28].

H Riemannian Gradient on the Hypersphere

For optional gradient-based interpretation, if $\nabla_u m_{y_f}(u;\theta)$ denotes the Euclidean gradient of the directional margin field with respect to the normalized latent direction u, then its Riemannian counterpart on the unit hypersphere is given by tangent projection:

$$\operatorname{grad}_{\mathbb{S}} m_{y_f}(u;\theta) = \Pi_u\big(\nabla_u m_{y_f}(u;\theta)\big)\,. \tag{14}$$

The margin retraction loss in Eq. (8) is a hinge-square in the directional margin and is optimized using the standard subgradients of $\max$ and $[\cdot]_+$. The norm of $\operatorname{grad}_{\mathbb{S}} m_{y_f}(u;\theta)$ measures local margin sensitivity along feasible directions on the unit hypersphere.

I Periodic Refresh of Local Retraction Geometry

A static local target gradually becomes mismatched to the evolving forget representation during optimization. For this reason, our implementation periodically refreshes the local retraction geometry during training, while keeping the margin field fixed at the pretrained teacher θ^{pre} (*teacher-evaluated, student-centered geometry*). At each refresh step t, we (i) recompute the (detached) current forget direction $u_{f,t}^{\mathrm{ctr}}$, (ii) sample tangent perturbations $\xi_{k,t} \sim P_r$ and form probe directions $u_{f,t}^{(k)} = \operatorname{Exp}_{u_{f,t}^{\mathrm{ctr}}}(\xi_{k,t})$, (iii) re-evaluate the near-neighborhood lower-quantile baseline $m_{f,t}^{\mathrm{base}}$, (iv) update the criticality $\omega_{f,t}$ and the adaptive coefficients $\tau_{f,t} = \min(\tau_0\,\alpha_{f,t}, \tau_{\max})$, $\beta_{f,t} = \beta_0\,\alpha_{f,t}$, and (v) reconstruct the retraction anchor $u_{f,t}^{\mathrm{anc}}$. All resulting quantities $\eta_t = (u_{f,t}^{\mathrm{ctr}}, m_{f,t}^{\mathrm{base}}, \omega_{f,t}, \tau_{f,t}, \beta_{f,t}, u_{f,t}^{\mathrm{anc}})$ are stop-gradient detached and act as fixed external constants until the next refresh.

To suppress refresh-to-refresh oscillation, we apply EMA smoothing only to the scalar refresh quantities

$$\zeta_{f,t} = (m_{f,t}^{\mathrm{base}}, \omega_{f,t}, \tau_{f,t}, \beta_{f,t}).$$

Concretely, for each scalar $a \in \{m_f^{\mathrm{base}}, \omega_f, \tau_f, \beta_f\}$,

$$a_{f,t}^{\mathrm{ema}} = \mu a_{f,t-\Delta}^{\mathrm{ema}} + (1-\mu) a_{f,t}^{\mathrm{raw}}, \qquad \mu \in [0,1), \tag{15}$$

where Δ is the refresh interval. The directional anchor $u_{f,t}^{\mathrm{anc}}$ is not EMA-averaged directly on the sphere; instead, it is reconstructed from the smoothed $\tau_{f,t}^{\mathrm{ema}}$ and the freshly computed weighted tangent direction $\bar{\xi}_{f,t}$.

J Implementation-Specific Instantiation

The main text uses the abstract notation θ_{rep} to denote the last trainable representation block. In our ResNet18 implementation, this corresponds to the final residual stage, while the classifier head is implemented as the final linear layer. Accordingly, optimization is restricted to the last representation block and classifier head, whereas earlier layers remain frozen. The layer-local parameter-anchoring term in Eq. (9) is applied only to the trainable representation block θ_{rep} and not to the classifier head W, allowing the decision boundary to retract freely at the classifier level.

K Algorithmic Summary

Algorithm 1 summarizes the proposed training procedure. The extension to a forget set $\mathcal{D}_f$ is performed at the loss level: for each instance $i \in \mathcal{D}_f$, HLMR maintains its own detached local retraction quantities

$$\eta_t^{(i)} = (u_{i,t}^{\mathrm{ctr}}, m_{i,t}^{\mathrm{base}}, \omega_{i,t}, \tau_{i,t}, \beta_{i,t}, u_{i,t}^{\mathrm{anc}}),$$

Algorithm 1 HLMR with Teacher-Evaluated, Student-Centered Geometry Refresh

Initialize pretrained teacher θ^{pre} (frozen) and trainable unlearning model θ
Collect the forget instance (x_f, y_f)
for $t = 1, \ldots, N_{\mathrm{step}}$ **do**
  if t is a refresh step **then**
    Compute detached current center $u_{f,t}^{\mathrm{ctr}} = \mathrm{stopgrad}(h_{\phi_t}(x_f)/\|h_{\phi_t}(x_f)\|_2)$
    Sample $\xi_{k,t} \sim P_r$, form probes $u_{f,t}^{(k)} = \mathrm{Exp}_{u_{f,t}^{\mathrm{ctr}}}(\xi_{k,t})$
    Evaluate margins $m_{f,t}^{(k)} = m_{y_f}(u_{f,t}^{(k)}; \theta^{\mathrm{pre}})$, $m_{f,t}^{\mathrm{pre}} = m_{y_f}(u_{f,t}^{\mathrm{ctr}}; \theta^{\mathrm{pre}})$
    Update scalar EMA states and reconstruct $u_{f,t}^{\mathrm{anc}}$
  end if
  Compute $\mathcal{L}_{\mathrm{geo}}$, $\mathcal{L}_{\mathrm{ret}}$, and $\mathcal{L}_{\mathrm{param}}$ using the most recent η_t
  Update $(\theta_{\mathrm{rep}}, W)$ by minimizing $\mathcal{L}_{\mathrm{total}}(\theta; \eta_t)$
end for
return unlearned model θ_u

computed separately under the frozen teacher. The set-level objective is

$$\mathcal{L}_{\mathrm{set}}(\theta) = \frac{1}{|\mathcal{D}_f|} \sum_{i \in \mathcal{D}_f} \left(\mathcal{L}_{\mathrm{geo}}^{(i)}(\theta) + \mathcal{L}_{\mathrm{ret}}^{(i)}(\theta) \right) + \lambda_{\mathrm{param}} \mathcal{L}_{\mathrm{param}}(\theta). \tag{16}$$

The parameter stabilizer acts in the ambient parameter space and is shared across instances.

This design preserves the zero-shot constraint because no retain instances or cross-instance retain neighborhoods are introduced. It also makes the method scalable, since each forget instance contributes a local retraction objective that can be averaged within a mini-batch. However, this extension does not explicitly model interactions among nearby forget instances. When many forget instances form a dense same-class cluster, their local anchors become correlated and the averaged objective induces stronger shared-structure erosion. We therefore treat interaction-aware multi-instance retraction as a limitation and evaluate large-ratio settings empirically in Appendix O.

L Additional Evaluation and Fairness Details

Zero-shot setting. We use *zero-shot unlearning* to denote the setting in which the unlearning algorithm has no access to the retain set and can use only the target model and forget set during the unlearning procedure. Some methods generate proxy examples from the information available at request time. HLMR does not use such substitutes; it operates only on the pretrained model and the forget set. The request-time information budget also meets the definition of *retain-free* or *remaining-data-free* unlearning. The broader *source-free* setting [1] additionally permits external substitutes for the original data: SCAR [3], for example, uses out-of-distribution images. Learned-noise methods such as Min-Max [13] generate noise from the model alone and differ in the deletion unit (Table 19). HLMR evaluates tangent perturbations of the forget representations under the pretrained teacher. These directions serve as probes of the local margin field; they are never used as input or training instances.

Evaluation protocol. All methods are evaluated from the same pretrained checkpoint, the same forget/retain split, the same unlearning random seeds, and the same test, forget, and retain evaluation loaders. The same membership-inference evaluation pipeline is used for all methods. During unlearning, zero-shot methods can access only the pretrained model and the forget set; retain instances are used only for evaluation after the unlearning procedure has finished. This protocol compares unlearning algorithms with data splits, model initialization, and evaluation procedures held fixed.

Model selection and fairness control. Hyperparameters are selected from the original papers, public implementations, and a bounded search around recommended values. To define matched forgetting strength *without consulting the retrained oracle or the retain set during model selection*, we use the held-out validation set to estimate the original model's generalization accuracy and select

unlearning configurations whose forget accuracy approaches this validation accuracy. Because the retrained oracle is itself a model that has never seen the forget set and is expected to generalize at this level, validation-accuracy matching serves as an oracle-free proxy for retraining alignment, providing an objective stopping criterion while avoiding any leakage from the oracle into the unlearning procedure. A configuration that preserves high test or retain accuracy while leaving forget accuracy close to the original model is not counted as a favorable utility–forgetting tradeoff. Retraining is used only as the oracle reference for evaluation *after* unlearning has finished, not as an attainable requirement for zero-shot methods.

M Additional Experimental Details

M.1 Datasets, Tasks, and Forget-Retain Splits

Compute resources. All experiments were run on RTX 5090 GPUs with 32 GB GPU memory and AMD 9950X3D CPU with 64 GB memory. The software environment used Python 3.10.6, PyTorch 2.11.0, torchvision 0.26.0, and CUDA 12.8. Runtime reported in Tables 5, 6, and 7 is the wall-clock time for the core unlearning computation and excludes original training, retraining, membership-inference attack evaluation, plotting, and logging. The total compute budget includes training the original and retrained reference models, running all unlearning baselines over five seeds, and evaluating utility and membership-inference metrics.

Datasets. We evaluate on CIFAR-10, CIFAR-100, and Tiny-ImageNet. CIFAR-10 contains $50{,}000$ training images and $10{,}000$ test images from 10 classes at 32×32 resolution. CIFAR-100 has the same number of training and test images but contains 100 fine-grained classes. Tiny-ImageNet contains $100{,}000$ training images and $10{,}000$ validation images from 200 classes. For Tiny-ImageNet, we resize the original 64×64 images to 224×224 to match the patch-grid input expected by ViT-S/16.

For CIFAR-10, we normalize images using channel mean $(0.4914, 0.4822, 0.4465)$ and standard deviation $(0.2023, 0.1994, 0.2010)$. Training uses random crop with four-pixel padding and random horizontal flip. For CIFAR-100, we use mean $(0.5071, 0.4867, 0.4408)$ and standard deviation $(0.2675, 0.2565, 0.2761)$, with the same crop-and-flip augmentation. For Tiny-ImageNet, we use ImageNet normalization statistics $(0.485, 0.456, 0.406)/(0.229, 0.224, 0.225)$ and random horizontal flip during training. The Tiny-ImageNet validation split is reorganized into an ImageFolder-compatible directory structure because its official validation labels are provided in a single annotation file.

For random instance-wise unlearning, the forget set $\mathcal{D}_f$ is sampled from $\mathcal{D}$ in a per-class stratified manner. For each class c, we draw $\lfloor |\mathcal{D}^{(c)}| \cdot \rho \rceil$ examples without replacement, where ρ is the forget ratio. This keeps the marginal class prior of $\mathcal{D}_f$ aligned with the original dataset and avoids trivial class-imbalance shortcuts. The forget ratios are $\{1\%, 5\%, 10\%\}$ for CIFAR-10, $\{0.2\%, 1\%, 5\%\}$ for CIFAR-100, and $\{0.2\%, 1\%, 5\%\}$ for Tiny-ImageNet.

M.2 Original and Oracle Model Training

For each dataset and forget ratio, we cache two reference checkpoints: the original model trained on the full training set $\mathcal{D}$ and the oracle retrained model trained from scratch on $\mathcal{D}_r$. The oracle retrained model instantiates the empirical gold standard $f_{\theta^\star}$ against which approximate unlearning methods are evaluated.

On CIFAR-10, we train a ResNet-18 with a 512-dimensional penultimate representation and a ten-way linear classifier. The model is trained for 40 epochs with batch size 256 using SGD with learning rate 0.1, momentum 0.9, cross-entropy loss, and a **ReduceLROnPlateau** schedule with factor 0.1 and patience 3.

On CIFAR-100, we use a ResNet-50 adapted to 32×32 inputs. The default 7×7 stride-2 convolution is replaced by a 3×3 stride-1 convolution, and the max-pooling layer is replaced by an identity mapping. The classifier head is replaced with a 100-way linear layer. The model is trained for 50 epochs with batch size 128 using Nesterov SGD with learning rate 0.1, momentum 0.9, weight decay 5×10^{-4}, and a per-batch cosine annealing schedule.

On Tiny-ImageNet, we use the pretrained ViT-S/16 and replace the original classifier head with a fresh 200-way head. The model is fine-tuned for 20 epochs with batch size 64 using AdamW with

Table 4: Datasets, architectures, and forget ratios used in our experiments.

Dataset	Model	Classes	Forget ratios / Num of $\mathcal{D}_f$
CIFAR-10	ResNet-18	10	$1\%, 5\%, 10\%$: $\{500, 2500, 5000\}$
CIFAR-100	ResNet-50	100	$0.2\%, 1\%, 5\%$: $\{100, 500, 2500\}$
Tiny-ImageNet	ViT-S/16	200	$0.2\%, 1\%, 5\%$: $\{200, 1000, 5000\}$

learning rate 5×10^{-5} and weight decay 0.05, together with a per-batch cosine annealing schedule. The lower learning rate is used to preserve pretrained representations while adapting the head and late blocks to Tiny-ImageNet.

M.3 Evaluation Metrics and Implementation

All metrics are computed using deterministic evaluation loaders without data augmentation or shuffling. We report mean and standard deviation over five unlearning seeds for each fixed forget/retain split.

The accuracy metrics are computed on three splits. **TEST_ACC** is the accuracy on the held-out test set, **RETAIN_ACC** is the accuracy on $\mathcal{D}_r$, and **FORGET_ACC** is the accuracy on $\mathcal{D}_f$. To quantify oracle alignment on the forget set, we also compute the forget accuracy gap:

$$\Delta_F = \left| F_{\mathrm{acc}}(\theta_u) - F_{\mathrm{acc}}(\theta^\star) \right|,$$

where θ_u is the unlearned model and $\theta^\star$ is the oracle retrained model. A smaller Δ_F indicates closer agreement with the oracle retrained behavior on the forget set.

For the loss-based membership inference attack, **LOSS_MIA** uses per-instance cross-entropy loss as the attack feature. A stratified k-fold logistic regression classifier is trained to distinguish forgotten instances from class-matched non-members. We report the attack area under the receiver operating characteristic curve as the main privacy proxy. This protocol follows standard membership inference evaluation and prior unlearning practice [39, 47, 32].

For the likelihood-ratio membership inference attack, we follow Carlini et al. [5]. The attack scores are computed using a shared pool of shadow models trained with the same architecture and training recipe. For each evaluation setting, attack scores are computed on the forget examples and a class-matched non-member set from the test split. When fewer than 512 forget examples are available, we use all available forget examples and sample an equal number of class-matched non-members; otherwise, we use 512 forget examples and 512 non-members. We train 24 shadow models and the same shadow-model pool and attack pipeline are used for all unlearning methods within a dataset–architecture setting. In the tables, **LiRA-MIA** includes area under the receiver operating characteristic curve, area under the precision-recall curve, and true-positive rates at low false-positive-rate operating points.

Runtime is reported as wall-clock time in seconds for the core unlearning computation. Original training, retraining, membership-inference attack evaluation, plotting, logging, and checkpoint input/output are excluded. Method-specific operations required by an unlearning algorithm are included in that algorithm's runtime.

N Baseline Implementation Details

RETRAIN trains the same architecture from scratch on $\mathcal{D}_r$ using the same optimizer, learning-rate schedule, batch size, and number of epochs as the original model. It serves as the empirical oracle model $f_{\theta^\star}$ and provides the reference behavior for oracle-aligned evaluation.

GRADIENT ASCENT (GA), also known as NegGrad, performs unlearning by maximizing the cross-entropy loss on $\mathcal{D}_f$. In implementation, the sign of the loss is flipped so that each update pushes the model away from fitting the revoked instances.

MU-MIS suppresses instance contribution using a sensitivity-gap objective [10]. The implementation minimizes a higher-order objective that suppresses the gradient norm of the per-example loss on $\mathcal{D}_f$. A ratio-based early-stopping rule discards unstable updates when the normalized gradient norm

overshoots its initial value. For ViT-S/16, we use a per-pixel gradient normalization mode so that the objective is not inflated by the higher input resolution.

SULI performs self-unlearning through entropy-guided layered processing and soft-label redistribution. The forget set is first partitioned according to the entropy of the original model's prediction [35]. A student model is then distilled from a modified teacher distribution in which the top-1 logit is suppressed before the softmax. Buckets are processed sequentially so that instances with different forgetting difficulty are handled in a layered manner. For high-resolution inputs, an inner-batch-size cap is used to control peak memory.

JiT reduces local sensitivity around each forget instance through a local Lipschitz-style objective [22]. For each forget instance, Gaussian neighbors are drawn and the model is updated to reduce local input-output variation around the instance. We use the implementation's default small learning rate and a single pass over $\mathcal{D}_f$ unless otherwise specified.

ADV+IMP follows the instance-wise Learning to Unlearn setting [6]. It alternates between constructing adversarial neighbors of the forget instances using PGD and applying an importance-weighted parameter update that encourages disagreement on those neighbors. The outer loop terminates when the forget accuracy falls below the prescribed threshold or when the maximum number of iterations is reached.

All baselines and HLMR use the same pretrained original checkpoint, the same forget/retain split, the same five-seed evaluation protocol, and the same metric pipeline. This matched protocol compares unlearning algorithms with the data split, architecture, training recipe, and measurement procedure held fixed.

Baselines excluded from this comparison. We evaluated only zero-shot methods that operate strictly on the pretrained model and the forget set. SCAR [3] requires an auxiliary out-of-distribution corpus as a preservation surrogate, and LoTUS [40] requires a subset of retain instances and an unseen set; both use auxiliary resources that violate the strict zero-shot protocol under which we report all results. Including them would conflate the effect of the unlearning algorithm with that of the auxiliary signal, and a fair comparison would require an additional axis of ablation over the auxiliary data source; we therefore defer this comparison to future work.

Request-time information budget. The zero-shot instance-wise interface is $\mathcal{A} : (\theta^{\mathrm{pre}}, \mathcal{D}_f) \mapsto \theta_u$. Each unlearning call receives the pretrained model and the requested forget set, without retain examples, external auxiliary/out-of-distribution (OOD)/unseen/calibration data, or stored training-derived statistics. Signals generated solely from these inputs remain within the budget: HLMR's latent probes and ADV+IMP's adversarial proxies are both internally generated. HLMR probes are latent directions used to evaluate the teacher's local margin field.

Comparison scope. Table 19 separates information access, model formulation, and deletion unit. These distinctions explain the main matched panel without equating all source-free or retain-free protocols. The method of Ahmed et al. [1] matches the request-time information budget; its Hessian-based formulation addresses linear and mixed-linear models. SCAR and LoTUS use additional preservation or calibration resources. Min-Max [13] and UNSIR [41] target class-identity deletion, whereas instance-wise deletion must preserve other records from the requested record's class. Adapting those methods would change the model formulation, deletion unit, or preservation objective and would constitute an additional comparison.

O Full Unlearning Results

Full raw results corresponding to the main-paper summaries are reported below. Values are mean $\pm$ standard deviation over five unlearning seeds where applicable. Runtime is reported in seconds and excludes original training, retraining, attack evaluation, plotting, logging, and checkpoint input/output.

Runtime and complexity. HLMR adds periodic latent probe scoring and geometry refresh to standard gradient-based unlearning updates. With forget batch size B, probe count K, feature dimension d, class count C, update steps S, and refresh interval R, the additional refresh scoring cost

scales approximately as $(S/R)\,BKCd$, plus the usual forward and backward costs on the trainable layers.

O.1 Full Accuracy, Loss-MIA, and Runtime Results

The full accuracy, loss-MIA, and runtime results tables are Tables 5, 6, and 7.

O.2 Full LiRA-MIA Results

The likelihood-ratio membership inference attack is reported separately from the main accuracy and runtime tables to keep both sets of results readable. The main paper uses LiRA as confirmatory privacy evidence; the low-false-positive-rate operating points are included here for completeness. And full LiRA-MIA tables are Tables 8, 9, and 10.

P Detailed Ablation Study

This appendix expands the ablation summary in Section 6. We provide the complete 24-configuration results table (Section P.2), the full component- and design-choice analyses (Sections P.3–P.4), the two remaining hyperparameter sweep figures (Section P.6), and per-step training dynamics for two notable runs (Section P.7).

P.1 Experimental Protocol

All ablations are run on **CIFAR-100 / ResNet-50** at a fixed forget ratio of $\rho = 1\%$ (per-class stratified, 500 forget instances in total). To isolate the effect of each ablation knob, we use a single forget/retain split and vary only the unlearning seed across $\{42, 43, 44, 45, 46\}$, yielding five independent unlearning runs per configuration. The canonical baseline corresponds to our standard hyperparameter block: steps $= 100$, probe_count $= 128$, probe_radius $= 0.2$, $\lambda_{\text{param}} = 10^{-3}$, $\beta_0 = 19$, margin_weight $= 19.0$, geometry_mode $=$ hybrid, optimizer = Adam, lr $= 10^{-4}$, trainable scope = **layer3+layer4+fc**, and checkpoint selection = **best_score**. Every other experiment overrides one or two fields on top of this canonical baseline; everything else is held fixed.

We report five metrics per run. Test accuracy (T_{acc}), retain-set accuracy (R_{acc}), and forget set accuracy (F_{acc}) are computed on deterministic, non-augmented eval loaders. The Sensitivity Gap (SG) is the dataset-level scalar

$$\text{SG} \;=\; \mathbb{E}_{x\in\mathcal{D}_f}\big[\|\nabla_x \log p_\theta(y \mid x)\|^2\big] \;-\; \mathbb{E}_{x\in\mathcal{D}_r}\big[\|\nabla_x \log p_\theta(y \mid x)\|^2\big],$$

We compute Loss-MIA AUC with five-fold cross-validation. Higher T and R indicate better utility; we assess forgetting by comparing F and Loss-MIA AUC with retraining. An AUC near 0.5 indicates near-chance discrimination between members and nonmembers by this loss-based attack. As $|\text{SG}|$ approaches zero, the forget and retain sets have similar mean squared input-gradient norms.

P.2 Full Results Table

Table 11 reports the full $24 \times 5 = 120$-run matrix. The configurations highlighted in Table 2 of the main paper appear in their respective experimental groups below; remaining rows complete the component drop-one ablations, the design-choice variants, and the hyperparameter sweep configurations.

P.3 Component-wise Contribution

We evaluate the contribution of each component by individually disabling it (the component drop-one rows in Table 11). The findings cleanly partition into *load-bearing* versus *auxiliary* mechanisms.

Geometry refresh and EMA are load-bearing. Disabling dynamic geometry refresh is catastrophic for forgetting: F_{acc} rises to 100.00% across all five seeds, and Loss-MIA AUC balloons from the canonical baseline value of 0.509 to 0.684. Test accuracy remains 82.58%, slightly above the canonical baseline. Without refresh, the retraction targets stay fixed as the model parameters change. Removing the geometry EMA is less severe but consistent in direction: F_{acc} jumps to 95.32%, MIA

Table 5: Full accuracy, loss-based membership-inference, and runtime results on CIFAR-10 / ResNet-18. Accuracy columns are percentages. Values are mean ± standard deviation over five unlearning seeds where applicable. Runtime is core unlearning compute time in seconds.

Forget ratio	Method	Test acc.	Forget acc.	Retain acc.	Loss-MIA	Runtime
1% (500)	Original	92.56	100.00	99.38	0.5472 ± 0.0061	–
1% (500)	Retraining	91.58	92.40	99.06	0.4920 ± 0.0098	–
1% (500)	NegaGrad	87.04 ± 0.47	91.56 ± 0.95	91.88 ± 0.65	0.5387 ± 0.0113	2.15
1% (500)	ADV+IMP	87.67 ± 1.26	91.28 ± 2.36	94.72 ± 1.37	0.5200 ± 0.0056	12.61
1% (500)	SULI	87.24 ± 1.04	92.72 ± 1.58	94.48 ± 1.11	0.5180 ± 0.0086	0.26
1% (500)	MU-MIS	85.67 ± 0.07	92.12 ± 0.18	92.72 ± 0.08	0.5157 ± 0.0183	2.85
1% (500)	JiT	87.05 ± 0.08	92.08 ± 0.33	93.68 ± 0.17	0.5153 ± 0.0191	1.58
1% (500)	HLMR	90.14 ± 0.22	91.14 ± 0.32	97.03 ± 0.38	0.5063 ± 0.0123	1.67
5% (2500)	Original	92.56	100.00	99.38	0.5472 ± 0.0061	–
5% (2500)	Retraining	91.64	93.12	99.40	0.4959 ± 0.0053	–
5% (2500)	NegaGrad	85.70 ± 1.86	92.38 ± 2.06	91.45 ± 2.12	0.5285 ± 0.0068	13.52
5% (2500)	ADV+IMP	84.15 ± 1.20	91.03 ± 1.40	90.25 ± 1.25	0.5241 ± 0.0085	135.82
5% (2500)	SULI	86.12 ± 0.32	92.51 ± 0.42	93.25 ± 0.30	0.5368 ± 0.0029	0.92
5% (2500)	MU-MIS	85.89 ± 0.04	92.02 ± 0.08	91.93 ± 0.04	0.5301 ± 0.0082	16.33
5% (2500)	JiT	85.69 ± 0.10	92.52 ± 0.10	92.21 ± 0.12	0.5288 ± 0.0065	6.27
5% (2500)	HLMR	89.11 ± 0.60	92.05 ± 0.75	95.52 ± 0.66	0.5109 ± 0.0104	7.06
10% (5000)	Original	92.56	99.30	99.39	0.5472 ± 0.0061	–
10% (5000)	Retraining	91.13	91.92	98.92	0.4968 ± 0.0033	–
10% (5000)	NegaGrad	84.59 ± 1.76	91.85 ± 1.66	91.14 ± 1.83	0.5322 ± 0.0097	29.56
10% (5000)	ADV+IMP	76.41 ± 1.06	88.30 ± 1.55	80.82 ± 1.11	0.5447 ± 0.0039	610.56
10% (5000)	SULI	83.87 ± 0.17	92.92 ± 0.19	90.93 ± 0.23	0.5437 ± 0.0034	4.14
10% (5000)	MU-MIS	84.02 ± 0.03	91.90 ± 0.08	90.96 ± 0.05	0.5390 ± 0.0042	19.59
10% (5000)	JiT	85.07 ± 0.06	90.89 ± 0.06	90.42 ± 0.06	0.5290 ± 0.0041	13.08
10% (5000)	HLMR	88.98 ± 0.44	91.92 ± 0.58	94.43 ± 0.74	0.5213 ± 0.0106	15.51

Table 6: Full accuracy, loss-based membership-inference, and runtime results on Tiny-ImageNet / Vision Transformer. Accuracy columns are percentages. Values are mean ± standard deviation over five unlearning seeds where applicable. Runtime is core unlearning compute time in seconds.

Forget ratio	Method	Test acc.	Forget acc.	Retain acc.	Loss-MIA	Runtime
0.2% (200)	Original	85.05	100.00	99.98	0.6130 ± 0.0084	–
0.2% (200)	Retraining	84.45	82.00	99.98	0.5325 ± 0.0206	–
0.2% (200)	NegaGrad	81.25 ± 0.60	84.90 ± 1.47	98.31 ± 0.48	0.5267 ± 0.0316	5.74
0.2% (200)	ADV+IMP	81.11 ± 0.30	84.30 ± 1.75	99.16 ± 0.07	0.4650 ± 0.0305	5.74
0.2% (200)	SULI	71.81 ± 2.89	83.42 ± 4.32	95.95 ± 3.72	0.4500 ± 0.0175	21.85
0.2% (200)	MU-MIS	66.38 ± 6.82	84.50 ± 4.72	76.08 ± 9.02	0.5767 ± 0.0725	7.57
0.2% (200)	JiT	80.39 ± 0.17	84.90 ± 0.42	97.08 ± 0.14	0.4783 ± 0.0113	2.00
0.2% (200)	HLMR	83.47 ± 0.09	83.50 ± 0.71	99.43 ± 0.05	0.4982 ± 0.0292	2.89
1% (1000)	Original	85.05	100.00	99.98	0.6130 ± 0.0084	–
1% (1000)	Retraining	85.10	84.00	99.88	0.5113 ± 0.0112	–
1% (1000)	NegaGrad	48.58 ± 27.27	54.62 ± 30.55	57.61 ± 32.47	0.4683 ± 0.2123	4.75
1% (1000)	ADV+IMP	68.57 ± 1.84	84.72 ± 2.04	88.51 ± 2.48	0.5467 ± 0.0053	305.65
1% (1000)	SULI	76.79 ± 0.23	85.91 ± 0.66	93.96 ± 0.18	0.5227 ± 0.0137	4.50
1% (1000)	MU-MIS	70.43 ± 3.89	87.66 ± 3.04	82.12 ± 3.87	0.5628 ± 0.0120	8.44
1% (1000)	JiT	74.99 ± 0.24	85.28 ± 0.18	87.76 ± 0.40	0.5503 ± 0.0138	4.65
1% (1000)	HLMR	81.86 ± 0.14	85.74 ± 0.51	98.91 ± 0.04	0.5197 ± 0.0137	7.03
5% (5000)	Original	85.05	100.00	99.98	0.6130 ± 0.0084	–
5% (5000)	Retraining	85.01	84.84	99.78	0.4946 ± 0.0046	–
5% (5000)	NegaGrad	59.87 ± 21.21	69.24 ± 26.85	70.91 ± 27.13	0.5821 ± 0.2373	52.38
5% (5000)	ADV+IMP	65.65 ± 3.76	83.39 ± 3.10	81.54 ± 2.88	0.5649 ± 0.0826	1513.21
5% (5000)	SULI	67.14 ± 2.44	84.42 ± 4.32	86.95 ± 3.72	0.5598 ± 0.0055	11.16
5% (5000)	MU-MIS	70.91 ± 5.04	86.16 ± 3.89	82.11 ± 4.56	0.5729 ± 0.0081	41.51
5% (5000)	JiT	71.94 ± 0.17	85.94 ± 0.30	86.11 ± 0.27	0.5704 ± 0.0071	36.47
5% (5000)	HLMR	79.77 ± 0.43	84.85 ± 0.83	95.30 ± 0.66	0.5414 ± 0.0055	32.25

Table 7: Full accuracy, loss-based membership-inference, and runtime results on CIFAR-100 / ResNet-50. Accuracy columns are percentages. Values are mean ± standard deviation over five unlearning seeds where applicable. Runtime is core unlearning compute time in seconds.

Forget ratio	Method	Test acc.	Forget acc.	Retain acc.	Loss-MIA	Runtime
0.2% (100)	Original	83.29	100.00	99.97	0.6432 ± 0.0088	–
0.2% (100)	Retraining	82.61	82.00	99.96	0.5050 ± 0.0317	–
0.2% (100)	NegaGrad	67.14 ± 8.72	81.80 ± 9.68	84.84 ± 11.21	0.5833 ± 0.0380	0.40
0.2% (100)	ADV+IMP	80.32 ± 0.12	82.80 ± 0.84	98.78 ± 0.04	0.4767 ± 0.0554	6.90
0.2% (100)	SULI	81.68 ± 0.16	82.80 ± 3.11	98.91 ± 0.03	0.4967 ± 0.0340	0.13
0.2% (100)	MU-MIS	66.25 ± 4.52	81.40 ± 5.04	87.40 ± 6.04	0.5700 ± 0.0552	2.86
0.2% (100)	JiT	70.28 ± 0.41	84.40 ± 1.14	90.68 ± 0.49	0.5933 ± 0.0429	0.38
0.2% (100)	HLMR	82.08 ± 0.21	82.00 ± 1.00	99.85 ± 0.04	0.4970 ± 0.0342	1.46
1% (500)	Original	83.29	100.00	99.97	0.6432 ± 0.0088	–
1% (500)	Retraining	82.01	82.60	99.95	0.5033 ± 0.0138	–
1% (500)	NegaGrad	66.56 ± 5.39	82.28 ± 6.95	84.27 ± 7.47	0.5533 ± 0.0087	1.09
1% (500)	ADV+IMP	78.86 ± 0.61	81.88 ± 1.37	96.97 ± 0.16	0.5327 ± 0.0191	46.29
1% (500)	SULI	76.00 ± 0.26	83.04 ± 1.56	97.91 ± 0.14	0.5533 ± 0.0169	0.60
1% (500)	MU-MIS	67.16 ± 0.12	82.64 ± 0.09	84.52 ± 0.15	0.5760 ± 0.0187	4.10
1% (500)	JiT	67.36 ± 0.26	82.68 ± 0.44	83.99 ± 0.34	0.5821 ± 0.0059	2.82
1% (500)	HLMR	81.14 ± 0.30	81.80 ± 0.68	99.05 ± 0.12	0.5090 ± 0.0099	3.35
5% (2500)	Original	83.29	100.00	99.97	0.6432 ± 0.0088	–
5% (2500)	Retraining	83.39	82.88	99.97	0.5167 ± 0.0097	–
5% (2500)	NegaGrad	59.52 ± 6.55	75.42 ± 7.71	75.74 ± 7.99	0.5753 ± 0.0092	6.90
5% (2500)	ADV+IMP	67.65 ± 0.35	82.86 ± 1.03	88.23 ± 0.48	0.5311 ± 0.0301	299.38
5% (2500)	SULI	72.09 ± 0.34	83.91 ± 0.91	89.17 ± 0.48	0.5716 ± 0.0057	2.88
5% (2500)	MU-MIS	65.89 ± 0.10	82.62 ± 0.13	81.69 ± 0.09	0.5866 ± 0.0107	29.68
5% (2500)	JiT	64.45 ± 0.19	82.19 ± 0.21	81.86 ± 0.17	0.5821 ± 0.0059	15.02
5% (2500)	HLMR	77.48 ± 0.91	82.61 ± 0.74	96.63 ± 0.89	0.5304 ± 0.0104	18.48

Table 8: Full LiRA-MIA results on CIFAR-10 / ResNet-18. Values are mean ± standard deviation over five unlearning seeds where applicable. AUROC and AUPR are area under the receiver operating characteristic and precision-recall curves; TPR columns report true-positive rates at fixed false-positive rates.

Forget ratio	Method	LiRA AUROC	LiRA AUPR	TPR@1% FPR	TPR@0.1% FPR
1% (500)	Original	0.5916 ± 0.0062	0.2575 ± 0.0056	0.0208 ± 0.0041	0.0012 ± 0.0011
1% (500)	Retraining	0.5094 ± 0.0137	0.0503 ± 0.0025	0.0060 ± 0.0034	0.0000 ± 0.0000
1% (500)	NegaGrad	0.5323 ± 0.0132	0.0531 ± 0.0028	0.0180 ± 0.0061	0.0020 ± 0.0018
1% (500)	ADV+IMP	0.5435 ± 0.0138	0.0569 ± 0.0033	0.0100 ± 0.0057	0.0020 ± 0.0021
1% (500)	SULI	0.5438 ± 0.0136	0.0580 ± 0.0035	0.0140 ± 0.0054	0.0020 ± 0.0018
1% (500)	MU-MIS	0.5270 ± 0.0142	0.0516 ± 0.0027	0.0060 ± 0.0036	0.0020 ± 0.0022
1% (500)	JiT	0.5384 ± 0.0131	0.0541 ± 0.0030	0.0200 ± 0.0068	0.0000 ± 0.0009
1% (500)	HLMR	0.5058 ± 0.0137	0.0499 ± 0.0024	0.0060 ± 0.0043	0.0000 ± 0.0009
5% (2500)	Original	0.5916 ± 0.0062	0.2575 ± 0.0056	0.0208 ± 0.0041	0.0012 ± 0.0011
5% (2500)	Retraining	0.4961 ± 0.0061	0.1970 ± 0.0033	0.0088 ± 0.0019	0.0008 ± 0.0006
5% (2500)	NegaGrad	0.5482 ± 0.0064	0.2264 ± 0.0047	0.0160 ± 0.0029	0.0012 ± 0.0007
5% (2500)	ADV+IMP	0.5460 ± 0.0062	0.2208 ± 0.0042	0.0116 ± 0.0022	0.0004 ± 0.0005
5% (2500)	SULI	0.5649 ± 0.0063	0.2320 ± 0.0046	0.0144 ± 0.0029	0.0008 ± 0.0006
5% (2500)	MU-MIS	0.5394 ± 0.0064	0.2148 ± 0.0040	0.0080 ± 0.0022	0.0008 ± 0.0006
5% (2500)	JiT	0.5480 ± 0.0063	0.2245 ± 0.0044	0.0128 ± 0.0027	0.0004 ± 0.0005
5% (2500)	HLMR	0.5150 ± 0.0064	0.2006 ± 0.0037	0.0080 ± 0.0024	0.0008 ± 0.0007
10% (5000)	Original	0.5916 ± 0.0062	0.2575 ± 0.0056	0.0208 ± 0.0041	0.0012 ± 0.0011
10% (5000)	Retraining	0.5007 ± 0.0051	0.3332 ± 0.0039	0.0116 ± 0.0018	0.0004 ± 0.0003
10% (5000)	NegaGrad	0.5570 ± 0.0051	0.3799 ± 0.0050	0.0172 ± 0.0026	0.0022 ± 0.0011
10% (5000)	ADV+IMP	0.5517 ± 0.0050	0.3637 ± 0.0044	0.0094 ± 0.0018	0.0014 ± 0.0006
10% (5000)	SULI	0.5614 ± 0.0048	0.3741 ± 0.0045	0.0100 ± 0.0028	0.0010 ± 0.0006
10% (5000)	MU-MIS	0.5412 ± 0.0050	0.3588 ± 0.0044	0.0118 ± 0.0019	0.0018 ± 0.0007
10% (5000)	JiT	0.5415 ± 0.0048	0.3628 ± 0.0043	0.0112 ± 0.0021	0.0012 ± 0.0007
10% (5000)	HLMR	0.5212 ± 0.0049	0.3437 ± 0.0044	0.0114 ± 0.0020	0.0008 ± 0.0007

Table 9: Full LiRA-MIA results on Tiny-ImageNet / Vision Transformer. Values are mean ± standard deviation over five unlearning seeds where applicable. AUROC and AUPR are area under the receiver operating characteristic and precision-recall curves; TPR columns report true-positive rates at fixed false-positive rates.

Forget ratio	Method	LiRA AUROC	LiRA AUPR	TPR@1% FPR	TPR@0.1% FPR
0.2% (200)	Original	0.5493 ± 0.0222	0.3051 ± 0.0171	0.0000 ± 0.0069	0.0000 ± 0.0000
0.2% (200)	Retraining	0.4841 ± 0.0233	0.2686 ± 0.0140	0.0000 ± 0.0000	0.0000 ± 0.0000
0.2% (200)	NegaGrad	0.5766 ± 0.0273	0.3972 ± 0.0295	0.0450 ± 0.0173	0.0050 ± 0.0116
0.2% (200)	ADV+IMP	0.5493 ± 0.0222	0.3051 ± 0.0171	0.0000 ± 0.0000	0.0000 ± 0.0000
0.2% (200)	SULI	0.5166 ± 0.0235	0.2848 ± 0.0164	0.0000 ± 0.0000	0.0000 ± 0.0000
0.2% (200)	MU-MIS	0.6478 ± 0.0220	0.4279 ± 0.0300	0.0350 ± 0.0221	0.0000 ± 0.0000
0.2% (200)	JiT	0.5230 ± 0.0251	0.3125 ± 0.0214	0.0150 ± 0.0124	0.0000 ± 0.0000
0.2% (200)	HLMR	0.5157 ± 0.0230	0.2715 ± 0.0136	0.0000 ± 0.0000	0.0000 ± 0.0000
1% (1000)	Original	0.5697 ± 0.0193	0.5892 ± 0.0183	0.0176 ± 0.0109	0.0078 ± 0.0050
1% (1000)	Retraining	0.5041 ± 0.0188	0.4981 ± 0.0144	0.0059 ± 0.0049	0.0020 ± 0.0026
1% (1000)	NegaGrad	0.5661 ± 0.0174	0.5746 ± 0.0165	0.0195 ± 0.0070	0.0098 ± 0.0056
1% (1000)	ADV+IMP	0.5907 ± 0.0159	0.5987 ± 0.0163	0.0449 ± 0.0123	0.0098 ± 0.0134
1% (1000)	SULI	0.5677 ± 0.0179	0.5522 ± 0.0165	0.0215 ± 0.0073	0.0078 ± 0.0047
1% (1000)	MU-MIS	0.5709 ± 0.0172	0.5755 ± 0.0167	0.0176 ± 0.0093	0.0078 ± 0.0057
1% (1000)	JiT	0.5783 ± 0.0178	0.5760 ± 0.0177	0.0254 ± 0.0085	0.0098 ± 0.0054
1% (1000)	HLMR	0.5209 ± 0.0194	0.5139 ± 0.0166	0.0872 ± 0.0055	0.0020 ± 0.0029
5% (5000)	Original	0.5733 ± 0.0189	0.5777 ± 0.0176	0.0176 ± 0.0107	0.0000 ± 0.0000
5% (5000)	Retraining	0.4612 ± 0.0195	0.4724 ± 0.0151	0.0039 ± 0.0058	0.0000 ± 0.0000
5% (5000)	NegaGrad	0.4861 ± 0.0187	0.5020 ± 0.0163	0.0059 ± 0.0067	0.0000 ± 0.0000
5% (5000)	ADV+IMP	0.5733 ± 0.0189	0.5777 ± 0.0176	0.0176 ± 0.0107	0.0000 ± 0.0000
5% (5000)	SULI	0.5454 ± 0.0174	0.5362 ± 0.0165	0.0078 ± 0.0108	0.0000 ± 0.0000
5% (5000)	MU-MIS	0.5333 ± 0.0184	0.5293 ± 0.0172	0.0059 ± 0.0081	0.0020 ± 0.0029
5% (5000)	JiT	0.5645 ± 0.0190	0.5635 ± 0.0185	0.0098 ± 0.0072	0.0000 ± 0.0000
5% (5000)	HLMR	0.4912 ± 0.0176	0.4861 ± 0.0176	0.0073 ± 0.0088	0.0000 ± 0.0000

Table 10: Full LiRA-MIA results on CIFAR-100 / ResNet-50. Values are mean ± standard deviation over five unlearning seeds where applicable. AUROC and AUPR are area under the receiver operating characteristic and precision-recall curves; TPR columns report true-positive rates at fixed false-positive rates.

Forget ratio	Method	LiRA AUROC	LiRA AUPR	TPR@1% FPR	TPR@0.1% FPR
0.2% (100)	Original	0.5775 ± 0.0337	0.2310 ± 0.0319	0.0100 ± 0.0222	0.0100 ± 0.0103
0.2% (100)	Retraining	0.5403 ± 0.0312	0.1841 ± 0.0179	0.0000 ± 0.0007	0.0000 ± 0.0000
0.2% (100)	NegaGrad	0.5880 ± 0.0334	0.2183 ± 0.0264	0.0100 ± 0.0205	0.0000 ± 0.0016
0.2% (100)	ADV+IMP	0.5274 ± 0.0380	0.2094 ± 0.0303	0.0100 ± 0.0170	0.0100 ± 0.0103
0.2% (100)	SULI	0.4514 ± 0.0315	0.1509 ± 0.0158	0.0000 ± 0.0155	0.0000 ± 0.0000
0.2% (100)	MU-MIS	0.5896 ± 0.0294	0.2105 ± 0.0229	0.0100 ± 0.0187	0.0000 ± 0.0042
0.2% (100)	JiT	0.5993 ± 0.0309	0.2328 ± 0.0311	0.0200 ± 0.0185	0.0100 ± 0.0104
0.2% (100)	HLMR	0.5415 ± 0.0339	0.1923 ± 0.0240	0.0000 ± 0.0134	0.0000 ± 0.0103
1% (500)	Original	0.5222 ± 0.0197	0.5436 ± 0.0191	0.0160 ± 0.0097	0.0000 ± 0.0056
1% (500)	Retraining	0.5081 ± 0.0187	0.4951 ± 0.0156	0.0060 ± 0.0051	0.0020 ± 0.0024
1% (500)	NegaGrad	0.5608 ± 0.0162	0.5428 ± 0.0164	0.0080 ± 0.0078	0.0000 ± 0.0006
1% (500)	ADV+IMP	0.4879 ± 0.0181	0.5088 ± 0.0163	0.0280 ± 0.0107	0.0060 ± 0.0051
1% (500)	SULI	0.4459 ± 0.0180	0.4546 ± 0.0125	0.0060 ± 0.0040	0.0000 ± 0.0022
1% (500)	MU-MIS	0.5529 ± 0.0191	0.5347 ± 0.0178	0.0140 ± 0.0071	0.0020 ± 0.0032
1% (500)	JiT	0.5654 ± 0.0173	0.5414 ± 0.0168	0.0120 ± 0.0066	0.0020 ± 0.0035
1% (500)	HLMR	0.5085 ± 0.0190	0.4853 ± 0.0143	0.0060 ± 0.0044	0.0020 ± 0.0027
5% (2500)	Original	0.5477 ± 0.0180	0.5669 ± 0.0179	0.0156 ± 0.0142	0.0039 ± 0.0041
5% (2500)	Retraining	0.5174 ± 0.0167	0.5120 ± 0.0149	0.0117 ± 0.0061	0.0020 ± 0.0031
5% (2500)	NegaGrad	0.5956 ± 0.0184	0.5871 ± 0.0189	0.0352 ± 0.0174	0.0059 ± 0.0054
5% (2500)	ADV+IMP	0.5319 ± 0.0181	0.5575 ± 0.0171	0.0352 ± 0.0134	0.0039 ± 0.0071
5% (2500)	SULI	0.5662 ± 0.0160	0.5476 ± 0.0166	0.0137 ± 0.0084	0.0000 ± 0.0008
5% (2500)	MU-MIS	0.5575 ± 0.0164	0.5482 ± 0.0157	0.0117 ± 0.0077	0.0020 ± 0.0041
5% (2500)	JiT	0.5629 ± 0.0174	0.5499 ± 0.0163	0.0176 ± 0.0102	0.0039 ± 0.0050
5% (2500)	HLMR	0.5372 ± 0.0158	0.5226 ± 0.0151	0.0014 ± 0.0044	0.0020 ± 0.0021

Table 11: HLMR ablation on CIFAR100 / ResNet50 at 1% forget, 5 seeds. Values are mean ± std across seeds. Higher test and retain accuracy indicate better utility; forget accuracy is compared with retraining. An AUC near 0.5 indicates near-chance discrimination by this loss-based membership attack.

Group	Configuration	Test acc. (%)	Retain acc. (%)	Forget acc. (%)	Sensitivity gap	Loss-MIA AUC
Baseline	Canonical baseline	81.11 ± 0.25	98.37 ± 0.77	85.12 ± 0.38	-24.17 ± 1.97	0.509 ± 0.037
Component drop	Disable anchor regularization	78.73 ± 0.96	98.04 ± 0.76	80.64 ± 7.19	-24.46 ± 2.83	0.531 ± 0.049
Component drop	Disable geometry EMA	81.61 ± 0.28	99.60 ± 0.04	95.32 ± 0.61	-19.01 ± 2.19	0.591 ± 0.006
Component drop	Disable dynamic geometry refresh	82.58 ± 0.23	99.88 ± 0.03	100.00 ± 0.00	-32.03 ± 2.19	0.684 ± 0.006
Component drop	Use a single probe	79.14 ± 0.75	98.53 ± 0.60	81.80 ± 4.02	-23.74 ± 1.98	0.495 ± 0.042
Component drop	Disable locality kernel	80.65 ± 0.38	99.35 ± 0.14	88.64 ± 1.46	-22.00 ± 1.40	0.469 ± 0.010
Design choice	Checkpoint: minimum-excess step	78.69 ± 1.07	97.84 ± 0.66	91.04 ± 0.83	-8.88 ± 2.35	0.563 ± 0.006
Design choice	Checkpoint: final step	76.23 ± 0.25	96.55 ± 0.20	70.60 ± 2.46	-30.17 ± 2.47	0.596 ± 0.005
Design choice	Optimizer: SGD	79.96 ± 0.35	99.09 ± 0.19	98.40 ± 0.45	-3.36 ± 1.26	0.615 ± 0.007
Design choice	Trainable scope: classifier only	83.14 ± 0.08	99.97 ± 0.00	100.00 ± 0.00	3.55 ± 0.23	0.659 ± 0.001
Design choice	Trainable scope: all parameters	78.04 ± 1.27	97.62 ± 1.04	80.52 ± 6.48	-29.11 ± 3.61	0.530 ± 0.050
Probe radius sweep	Probe radius = 0.05	79.80 ± 0.35	99.00 ± 0.09	82.76 ± 1.61	-22.28 ± 0.77	0.452 ± 0.022
Probe radius sweep	Probe radius = 0.10	79.35 ± 0.62	98.60 ± 0.32	82.56 ± 5.07	-24.25 ± 2.64	0.507 ± 0.043
Probe radius sweep	Probe radius = 0.20	81.11 ± 0.25	98.37 ± 0.77	85.12 ± 0.38	-24.17 ± 1.97	0.509 ± 0.037
Probe radius sweep	Probe radius = 0.50	73.97 ± 1.79	93.68 ± 1.44	61.96 ± 6.38	-27.46 ± 4.53	0.605 ± 0.040
Probe radius sweep	Probe radius = 1.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	221506.57	0.497 ± 0.003
Probe count sweep	Probe count = 16	78.14 ± 1.16	97.67 ± 1.01	78.60 ± 4.83	-25.39 ± 1.72	0.545 ± 0.046
Probe count sweep	Probe count = 64	79.18 ± 1.31	98.39 ± 1.10	84.52 ± 5.88	-24.90 ± 3.53	0.505 ± 0.045
Probe count sweep	Probe count = 256	79.35 ± 1.04	98.56 ± 0.62	84.12 ± 6.50	-23.94 ± 1.84	0.508 ± 0.043
Probe count sweep	Probe count = 512	80.21 ± 0.76	99.05 ± 0.41	88.80 ± 3.14	-21.36 ± 2.72	0.505 ± 0.017
Anchor	Anchor regularization = 1e-4	78.57 ± 0.68	98.05 ± 0.56	79.80 ± 3.65	-26.31 ± 2.78	0.515 ± 0.048
Anchor	Anchor regularization = 1e-3	81.11 ± 0.25	98.37 ± 0.77	85.12 ± 0.38	-24.17 ± 1.97	0.509 ± 0.037
Anchor	Anchor regularization = 1e-2	78.34 ± 1.12	97.81 ± 0.75	77.20 ± 4.39	-26.28 ± 2.27	0.538 ± 0.055
Anchor	Anchor regularization = 1e-1	79.25 ± 0.91	98.44 ± 0.69	83.00 ± 6.59	-23.96 ± 3.71	0.529 ± 0.042

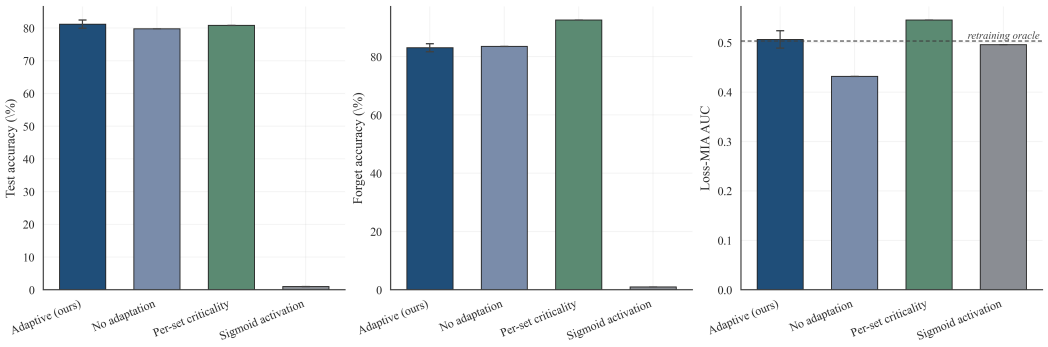


Figure 4: **Adaptive criticality ablation.** Removing per-instance adaptation pushes Loss-MIA below the retraining oracle (over-erasure); per-set criticality over-corrects above; sigmoid activation collapses the model (gray, control). The proposed adaptive criticality is the only configuration aligned with the oracle (dashed line, Loss-MIA $= 0.5033$).

AUC to 0.591, and the selection score becomes non-monotonic (Section P.7). Together, these two ablations establish that the temporal smoothing and dynamic re-estimation of the local geometry are the two mechanisms that carry the unlearning heuristic.

Parameter-anchor regularization, probe diversity, and locality kernel are auxiliary. Disabling the parameter-anchor term ($\lambda_{\text{param}} = 0$) leaves all forgetting metrics within seed noise ($\Delta F_{\text{acc}} = -4.5$ against $\sigma(F_{\text{acc}}) \approx 7$; $\Delta\text{AUC} = +0.022$), but costs 2.4 points of test accuracy. The single-probe ablation ($\text{probe_count} = 1$) stays within seed noise of the canonical baseline on all forgetting metrics ($\Delta F_{\text{acc}} = -3.3$) at a 2.0-point test-accuracy cost, suggesting that the iterative training dynamics (100 steps with periodic geometry refresh) act as an implicit ensembler that compensates for low instantaneous probe count. The uniform-locality ablation lowers T_{acc} by 0.5 points and raises F_{acc} by 3.5 points relative to the canonical baseline. We thus characterize these three components as *regularizers*: they nudge the optimization, but the geometric machinery does the work. We expect their effect to grow at larger forget ratios where the gradient interference between forget and retain objectives is stronger.

Lower-quantile margin baseline is the only choice that aligns with the oracle. Among baseline alternatives (Figure 5), the mean over-erases (MIA 0.45, below the oracle); the minimum is overly aggressive on local margins and damages utility ($T_{\text{acc}} = 75.8\%$); the class prototype baseline simply never forgets ($F_{\text{acc}} = 100\%$, MIA 0.71). The lower-quantile choice $q = 0.10$ is the only configuration whose Loss-MIA lies within 0.01 of the retraining oracle. The accompanying q-sweep (Figure 6) confirms this trend: smaller q tracks the oracle more closely; $q \in [0.10, 0.25]$ form a stable plateau, while $q = 0.50$ already starts to over-erase. This empirical pattern is consistent with the qualitative analysis in Appendix G: a lower-quantile estimator avoids inheriting the instance-induced bias that mean and high-q estimators absorb from the forget-centered protrusion.

Weighted log-map anchor outperforms simpler alternatives. Figure 7 compares the proposed anchor (a distance- and margin-weighted log-map of probe directions, Eq. (6)) with three natural simplifications. A uniform average over low-margin probes (no distance weighting) drifts in MIA. The class prototype anchor reproduces the failure mode observed for the prototype baseline: it preserves too much original-class support and barely forgets ($F_{\text{acc}} = 95\%$, MIA 0.60). Selecting the single lowest-margin probe (no aggregation) under-erases more than the weighted average. The weighted log-map is the only construction whose Loss-MIA aligns with the retraining oracle. Together with Figure 5, this reveals a consistent failure mode of *prototype-based* zero-shot targets across both the baseline and anchor design dimensions, supporting teacher-derived local probes for the baseline and anchor.

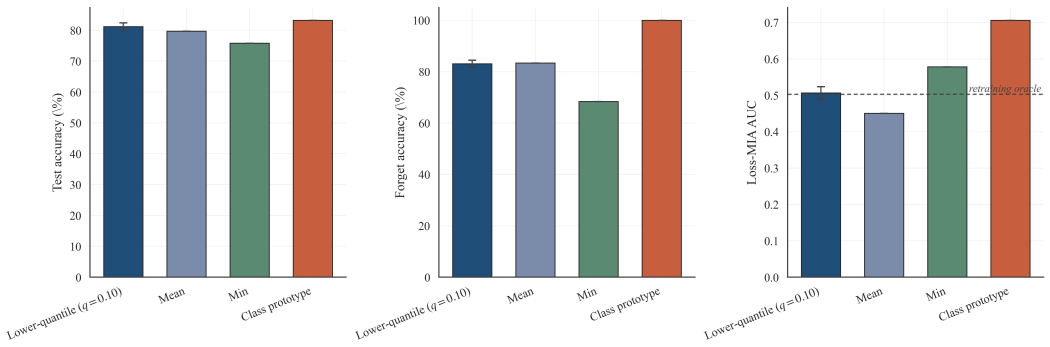


Figure 5: **Margin baseline design.** The lower-quantile baseline (q=0.10) is the only configuration whose Loss-MIA aligns with the retraining oracle. Mean over-erases, min damages utility, class prototype fails to forget. Dashed line: retraining oracle.

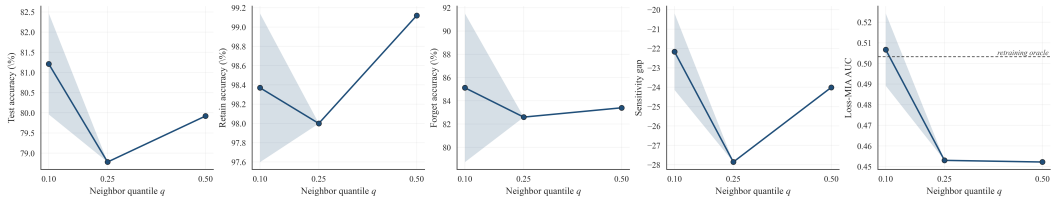


Figure 6: **Neighbor quantile q sweep.**

P.4 Design-choice Ablations

We validate our discrete design decisions against natural alternatives (the design-choice rows in Table 11).

Trainable scope. Restricting updates to the classifier (scope =**fc**) achieves the highest test accuracy (83.14%) but completely fails to forget ($F_{\text{acc}} = 100.00\%$, SG $= +3.55$): the classifier alone cannot absorb enough parametric change to reshape the forget set's local geometry. Conversely, training all parameters (scope =**all**) modestly increases forgetting ($F_{\text{acc}} = 80.52\%$) but inflates utility variance ($\sigma(T_{\text{acc}}) = 1.27$ vs. 0.08 for the classifier-only setting), confirming that our default **layer3+layer4+fc** sits at the empirically chosen sweet spot.

Checkpoint selection. Comparing the three checkpoint modes is informative. Selecting on minimum margin excess (**best_excess**) aligns poorly with the true utility-forgetting tradeoff: F_{acc} rises to 91% and SG shrinks in magnitude to -8.88, indicating premature stopping. Our default **best_score** provides the most favorable point on this tradeoff curve.

Optimizer. Replacing Adam with SGD yields virtually no forgetting ($F_{\text{acc}} = 98.40\%$, SG $= -3.36$), demonstrating that Adam is *not* a convenience choice but essential at our learning rate and step budget: the HLMR loss has heterogeneous curvature across its geometric retraction, margin retraction, and parameter-anchor terms that benefits substantially from per-parameter adaptive scaling.

P.5 Mechanism-driven Design Ablations

This subsection complements the component drop-one analysis in Appendix P.3 and the framework isolation in Section 6 (Figure 3) by dissecting three design choices that follow from the three-layer formulation: (i) the form of the per-instance retraction criticality ω_f, (ii) the construction of the neighborhood-supported margin baseline m_f^{base}, and (iii) the construction of the retraction anchor u_f^{anc}. All experiments use the canonical CIFAR-100 / ResNet-50 / $\rho = 1\%$ setting and are evaluated

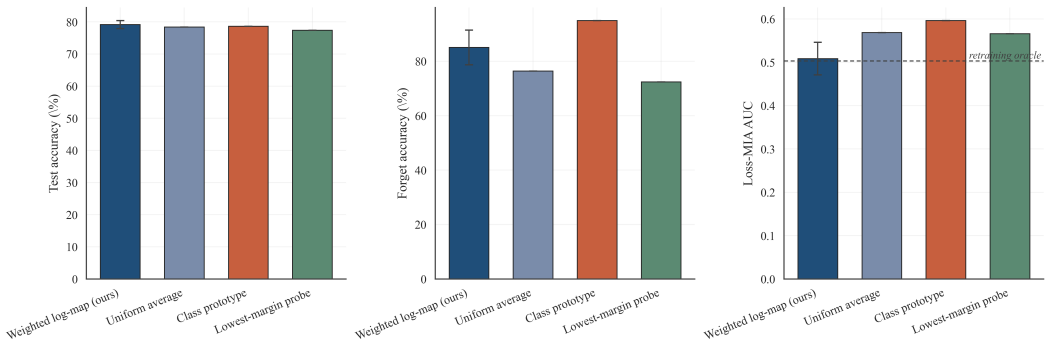


Figure 7: **Retraction anchor design.** The proposed weighted log-map anchor is the only construction whose Loss-MIA aligns with the retraining oracle. Class prototype anchors fail to forget, mirroring the prototype-baseline failure in Figure 5. The single lowest-margin probe and uniform average both drift in MIA.

against the retraining oracle (Loss-MIA $= 0.5033$), shown as a dashed reference line in each figure. Throughout this subsection we treat closeness to the oracle in Loss-MIA as the primary alignment criterion; values *below* the oracle indicate over-erasure, while values above indicate under-erasure.

Adaptive criticality is necessary to align with the retraining oracle. Disabling adaptation (Figure 4, $\alpha_f \equiv 1$) drives Loss-MIA to 0.43, well below the oracle level: a uniform forgetting strength pushes *every* forget instance harder than retraining would, producing over-erasure that also leaks into representation drift. Replacing the per-instance criticality with a per-set average over-corrects in the opposite direction (MIA 0.55). A sigmoid reparameterization of ω_f destabilizes optimization completely, collapsing the model to a constant prediction; we show this failed configuration in gray as a control. Only the proposed ReLU-gated, magnitude-preserving criticality keeps Loss-MIA closest to the oracle, supporting the per-instance form used in Eq. (4).

Cross-figure consistency. Two cross-figure observations strengthen these conclusions. First, prototype-based zero-shot targets fail in both roles: the class prototype baseline (Figure 5) and the class prototype anchor (Figure 7) both produce $F_{\text{acc}} \geq 95\%$ and Loss-MIA ≥ 0.60, indicating that aggregating to the class level destroys the local geometry the framework is designed to exploit. Second, configurations that over-erase Loss-MIA (mean baseline, no-adaptation criticality, high-q baseline) consistently land near MIA ≈ 0.43–0.45, whereas under-erasing configurations (min baseline, prototype variants) land at MIA ≥ 0.58, with the proposed defaults occupying the narrow band around the oracle. These results support *retraction alignment*: reducing Loss-MIA below the oracle increases the alignment gap.

P.6 Hyperparameter Sensitivity

We sweep three hyperparameters: probe radius, probe count, and anchor regularization weight. Their sensitivities are reported below.

Probe radius and locality. Sweeping $\text{probe_radius} \in \{0.05, 0.10, 0.20, 0.50, 1.00\}$ (Figure 8) reveals a viable small-radius regime on $[0.05, 0.20]$, noticeable utility degradation at 0.50, and complete model collapse at 1.00 ($T = R = F = 1\%$, SG $\sim 2.2 \times 10^5$).

On the unit hypersphere, a tangent displacement of length $\|\xi\|_2 < \pi$ moves the probe through that angle. At $r = 1.0$, a full-radius probe is $\approx 57.3^\circ$ from the forget direction:

$$\text{Exp}_{u_f^{\text{ctr}}}(\xi) = \cos(1)\, u_f^{\text{ctr}} + \sin(1)\, \frac{\xi}{\|\xi\|_2} \approx 0.54\, u_f^{\text{ctr}} + 0.84\, \frac{\xi}{\|\xi\|_2}.$$

At this radius, sampling extends beyond the small neighborhood used to estimate local class support. A larger radius expands the region from which HLMR selects the nearest probes for the baseline m_f^{base} and weights probes for the anchor u_f^{anc}. This loss of locality helps explain the observed

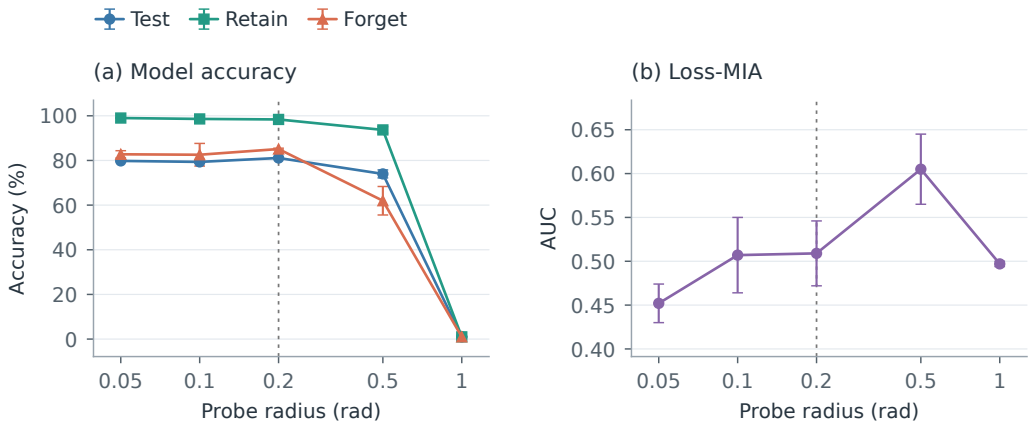


Figure 8: **Probe-radius sensitivity.** Test, retain, and forget accuracy are shown alongside Loss-MIA AUC for the five radius settings in Table 11. Points and error bars are means and standard deviations over five seeds. The dashed vertical line marks the canonical radius $r = 0.20$; the horizontal axis is logarithmic. Utility degrades at $r = 0.50$, and the model collapses at $r = 1.00$.

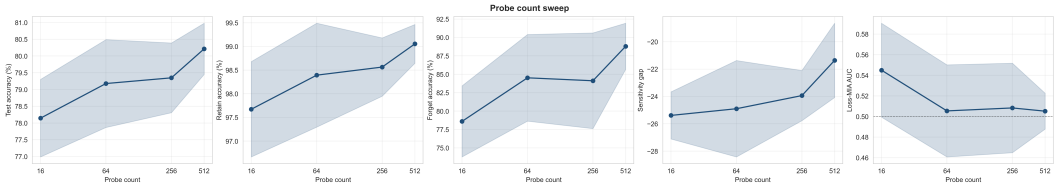


Figure 9: **Probe count sweep.** Four points at probe_count $\in \{16, 64, 256, 512\}$ on a log x-axis. Test and retain accuracy generally improve across these settings, while Loss-MIA AUC remains near 0.51 from 64 probes onward. Increasing the count from 256 to 512 raises runtime from 26 s to 71 s.

collapse: the geometric loss $\mathcal{L}_{\text{geo}}$ continues to decrease even after the classifier predicts a single class (Figure 11, second row).

Probe-count sensitivity (Fig. 9). Across the four swept settings (probe_count $= 16, 64, 256, 512$), increasing the probe count generally improves test and retain accuracy. Test accuracy increases as $78.14\% \rightarrow 79.18\% \rightarrow 79.35\% \rightarrow 80.21\%$, while Loss-MIA AUC remains near 0.51 from probe_count $= 64$ onward. Doubling the probe count from 256 to 512 raises test accuracy by 0.86 points and per-run wall-clock time from 26 s to 71 s. We use probe_count $= 128$ as the default.

Parameter-anchor regularization: flat for forgetting, load-bearing for utility (Fig. 10). Sweeping nonzero λ_{param} across three orders of magnitude (10^{-4}–10^{-1}), with an additional $\lambda_{\text{param}} = 0$ setting from the parameter-anchor-disabled ablation, shows that λ_{param} does not drive forgetting: $F_{\text{acc}} \in [77.2\%, 85.1\%]$ and Loss-MIA AUC $\in [0.51, 0.54]$ remain within seed noise across the sweep. It is, however, load-bearing for utility: the canonical $\lambda_{\text{param}} = 10^{-3}$ preserves 1.9–2.8 points more test accuracy than every other value ($T_{\text{acc}} = 81.1\%$ vs. $[78.3\%, 79.3\%]$). This pattern supports a trust-region interpretation of the parameter anchor. This refines the corresponding finding from Appendix P.3: at $\rho = 1\%$, the coupled geometry–margin objective drives the forgetting heuristic, while the parameter anchor protects utility.

P.7 Training Dynamics: Two Case Studies

To complement the aggregate-level ablation, Figure 11 plots the per-step training history of two notable runs (seed 42).



Figure 10: **Parameter-anchor regularization sweep.** Test, retain, and forget accuracy are shown alongside Loss-MIA AUC for four nonzero weights spanning three orders of magnitude (10^{-4}–10^{-1}) in Table 11. Points and error bars are means and standard deviations over five seeds. The dashed vertical line marks the canonical $\lambda_{\text{param}} = 10^{-3}$; the horizontal axis is logarithmic. Test accuracy favors this weight by 1.9–2.8 points.

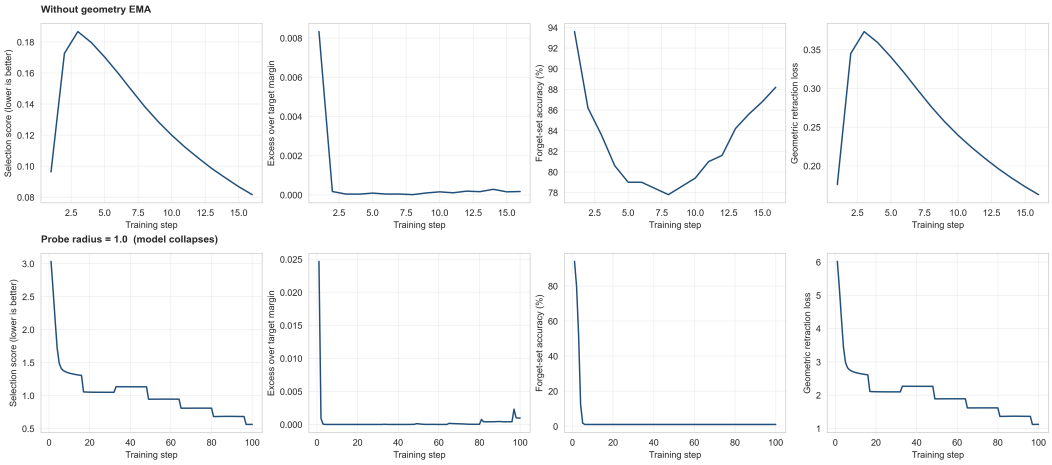


Figure 11: **Per-step training dynamics for two notable ablation configurations.** *Top row:* disabling the geometry EMA. The selection score peaks at steps 3–5 and then declines; checkpoint selection returns step 16 in all five seeds. *Bottom row:* probe_radius $= 1.0$, where the model collapses within ~ 5 steps as the forget set accuracy drops from $\sim 100\%$ to 1% and never recovers; the geometric retraction loss decays mechanically while the underlying classifier has lost its predictive structure entirely.

Why removing the EMA destabilizes selection. The top row of Figure 11 shows that without geometry EMA, the selection score is non-monotonic: it rises sharply in the first few steps as the noisy geometry estimate over-corrects, then decays as later steps partially recover. With **best_score** checkpointing, the algorithm is forced to commit to an early local minimum, which is why the no-EMA configuration's best_step is fixed at 16 across all five seeds (zero variance; Table 11). The geometry EMA serves precisely to suppress this oscillation, producing the smoother trajectories of the canonical baseline.

Anatomy of the probe-radius collapse. At $\mathrm{probe_radius} = 1.0$, the model predicts a single class and F_{acc} drops from $\sim 100\%$ to 1% within five steps (Figure 11, bottom row). The selection score plateaus near 1 while the geometric retraction loss continues to decrease after collapse. When on-manifold validation data are available, validation accuracy provides an additional candidate early-stopping criterion.

Q Limitations

We discuss the main limitations of HLMR and clarify the scope of the empirical claims.

First, we assess privacy empirically through membership-inference risk and alignment with matching retraining. These measurements neither certify deletion nor establish equivalence to retraining. Integrating certified removal or differential-privacy mechanisms remains future work.

Second, HLMR estimates its baseline and anchor from teacher-evaluated geodesic probes. Their reliability depends on the fidelity of the pretrained teacher's local directional margin field. Ambiguous class boundaries, sparse regions, and distribution shift warrant separate evaluation. Future work includes uncertainty-aware probes, calibration-aware margins, and adaptive probe filtering.

Third, HLMR requires a locally informative probe neighborhood. The geometric bound $r < \pi$ alone does not ensure locality. In the sensitivity sweep (Appendix P.6), the tested radii from 0.05 to 0.20 preserved utility, while the classifier collapsed at $r = 1.00$.

Fourth, our experiments focus on image classification across CNN and ViT models. Extending HLMR to generative models or large language models is future work, since defining appropriate hyperspherical representations, forget-instance margins, and local probes in those settings requires task-specific adaptation.

Fifth, the structured forget-set evaluation in Appendix S.2 covers random, feature-clustered, and boundary-conditioned sets. Mislabeled, rare-subgroup, influence-selected, and adversarial constructions remain outside this evaluation.

R Broader Impact

Machine unlearning sits at the intersection of privacy regulation (e.g., the "right to be forgotten" under the GDPR), copyright-removal requests, and data-retention constraints, and zero-shot methods are particularly valuable for deployed systems where the retain corpus has been deleted, partitioned across silos, or moved beyond the unlearning operator's reach. HLMR offers a path toward fine-grained, instance-level data removal that does not require retraining from scratch and does not rely on storing or redistributing retain data, which can lower the operational cost of regulatory compliance and shrink the privacy surface of long-lived deployed models; the hyperspherical local-retraction perspective additionally provides a more interpretable target than ad-hoc relabeling or proxy-data construction, supporting downstream auditing of what has been forgotten. As with any unlearning method, operators should pair the procedure with logged, governed audit trails so that unlearning is not used to retroactively obscure training-set membership.

S Additional Analyses from the Review Discussion

This appendix reports additional CIFAR-100/ResNet-50 analyses carried out during the review discussion: the evaluation protocol and configuration (Appendix S.1), structured forget sets with local-retention controls (Appendix S.2), diagnostics of the probe-derived reference (Appendix S.3), and sensitivity to directional-field distortion (Appendix S.4).

S.1 Evaluation protocol and configuration

Forget sets. We evaluate three class-balanced 1% forget sets on CIFAR-100/ResNet-50, each containing 500 training images and five images per class. The random control follows class-balanced sampling: five images are drawn uniformly without replacement from each class. The feature-clustered set concentrates nearby same-class instances in unit-normalized penultimate representation space: for each class, a center image is drawn at random among the images that the pretrained

model classifies correctly under the directional score, and the set takes the five same-class images closest to that center in geodesic distance, the center included. The boundary-conditioned set uses a fixed low-margin/local-variation criterion: within each class, the 50 directionally correct images with the smallest positive directional margins form a candidate pool, and the five candidates with the highest weighted rank score are selected. The score combines low margin (weight 0.35), sign inconsistency of the original-class margin over 32 local probes at geodesic radius 0.1 (weight 0.25), the interquartile range of the probe margins (weight 0.20), and the entropy of the strongest competing class across the probes (weight 0.20). Both structured constructions use only the pretrained model's representations and directional margins. Each evaluation manifest has zero overlap with the union of the three ordinary development forget sets.

Local-retention controls. For the feature-clustered set, we additionally evaluate disjoint same-class retained controls nearest to and farthest from the forget region in normalized representation space. For each forget image, the Adjacent-Retain control takes the same-class retained image with the largest cosine similarity to it, and the Remote-Retain control takes the one with the smallest; each retained image is used at most once. This gives two disjoint control sets of 500 retained images (five per class), which are stored with the forget set so that every method is evaluated on the same images.

Matching retrained references. Let $\mathcal{D}_f$ denote the forget set. For each $\mathcal{D}_f$, a separate matching reference $\theta^\star$ is retrained with the paper's architecture and training protocol (Appendix M.2) after excluding exactly that $\mathcal{D}_f$. Each set needs its own reference because retraining depends on the deleted records. Each unlearning call receives only the pretrained model and $\mathcal{D}_f$; we denote its output by θ_u. Retain data and the matching retrained models are loaded only after each unlearning call, for evaluation.

Development stage and frozen configuration. The analyses in this appendix use a separate development stage. Three ordinary class-balanced 1% development forget sets and their matching retrained references are used to select one global HLMR configuration. This configuration, listed in Table 12, is frozen before evaluating the random, feature-clustered, and boundary-conditioned sets. The probe radius $r = 0.13$ is selected on the ordinary development splits and is used for all three evaluation sets and for every request. The references of the evaluation sets are used only for final evaluation, and no configuration is tuned on them. This design evaluates whether a configuration selected in ordinary development settings transfers to new structured forget sets. The main benchmarks keep the validation-based selection procedure of Appendix L.

Baselines. The baselines are NegaGrad [42], ADV+IMP [6], SULI [35], MU-MIS [10], and JiT [22]. Their hyperparameters are fixed together with the HLMR configuration before the structured-set evaluation: NegaGrad uses Adam with learning rate 1.1×10^{-5} for 6 epochs; ADV+IMP uses Adam with learning rate 10^{-5}, 15 PGD iterations, and at most 20 outer iterations; SULI uses Adam with learning rate 10^{-4} for 20 epochs; MU-MIS uses SGD with learning rate 2.5×10^{-5} for 10 epochs; and JiT uses Adam with learning rate 5×10^{-6} for 5 epochs. All methods share the pretrained checkpoint, forget/retain split, unlearning seeds, evaluation loaders, and membership-inference pipeline. No method is tuned on the references of the evaluation sets.

Seeds and reporting. Unlearning results are reported as mean $\pm$ standard deviation over unlearning seeds 42–46. The matching retrained references use retraining seed 42; the retraining-seed analysis in Appendix S.3 adds seeds 43–46. Probe-reference statistics are summarized over probe seeds 42–46. Accuracy and Loss-MIA AUC follow Appendix M.3.

S.2 Structured forget sets

The three forget sets, their matching retrained references, and the frozen configuration are described in Appendix S.1. The feature-clustered and boundary-conditioned constructions test concentrated local entanglement and a low-margin regime, respectively. One development-selected configuration is used across all three sets, with probe radius 0.13.

For each metric $M \in \{T, R, F, A\}$, define the absolute matching-retraining gap $g_M = |M(\theta_u) - M(\theta^\star)|$, where T, R, F denote Test, Retain, and Forget accuracy and A denotes Loss-MIA AUC. For

Table 12: Configuration used for the analyses in this appendix.

Setting	Symbol	Value
Optimizer	–	Adam
Learning rate	–	9×10^{-5}
Weight decay	–	0
Update steps	N_{step}	100
Gradient-norm clipping	–	5.0
Trainable scope	–	layer3 + layer4 + fc
Probe count	K	64
Near-probe fraction	K_{near}/K	0.5
Probe radius	r	0.13
Baseline statistic	$m_{f,t}^{\text{base}}$	lower quantile
Quantile level	q	0.15
Angular scale	σ	default $(0.75\, r)$
Margin temperature	T_m	0.1
Anchor construction	$u_{f,t}^{\text{anc}}$	weighted log-map
Base retraction step	τ_0	0.0325
Base margin coefficient	β_0	19.0
Margin weight	–	20.0
Criticality gain	κ	0
Criticality	$\omega_{f,t}$	adaptive, per instance, ReLU
Margin-target offset	–	0.075
Parameter-stabilizer weight	λ_{param}	3×10^{-4}
Geometry mode	–	hybrid
Refresh interval	Δ	8 steps
EMA decay	μ	0.8
Batch-normalization statistics	–	frozen
Dropout	–	off
Bad-update rejection	–	on
Early stopping	–	off
Checkpoint selection	–	best_score

this targeted evaluation, the compact summary is

$$J = \frac{g_T + g_R + g_F + 100 g_A}{4}.$$

Accuracy gaps are in percentage points and AUC gaps use the raw $[0, 1]$ scale. The factor 100 converts the AUC gap to the same numerical scale before taking the equally weighted average. Lower J indicates closer alignment to matching retraining. We interpret J as a compact summary and retain the component-wise gaps as the primary evidence. Table 15 reports the full cross-method comparison. Absolute metric gaps and J are computed for each unlearning seed, then summarized over seeds 42–46. The component gaps in Table 13 and local-retention results in Table 14 provide the corresponding utility and preservation evidence. For the latter, adjacent and remote controls are disjoint same-class retained examples nearest to and farthest from the feature-clustered forget region in normalized representation space. HLMR has the lowest mean J on the feature-clustered and boundary-conditioned sets and the highest mean Adjacent-Retain accuracy among the compared unlearning methods.

S.3 Probe-reference diagnostics

The diagnostics separate the quality of the local margin reference from its relation to the instance-wise retraining response. For forget instance f, let $\mathcal{P}_f$ be its local geodesic probes and let $\mathcal{N}_f^R$ be the 128 nearest same-class retained embeddings. Neighbors are selected by cosine similarity in the pretrained normalized representation space, excluding the current forget set and all development forget records.

Table 13: **Frozen HLMR retains utility and tracks set-specific retraining on structured forget sets.** Absolute component gaps and J are mean $\pm$ standard deviation over unlearning seeds 42–46 (standard deviations appear on the second line). The last column gives the competing method with the lowest mean J in each set.

Forget set	g_T	g_R	g_F	g_A	HLMR J	Best competing J
Random	0.46 $\pm$0.18	0.46 $\pm$0.08	3.44 $\pm$1.22	0.0569 $\pm$0.0091	2.51 $\pm$0.25	ADV+IMP 2.68 ± 0.32
Clustered	0.80 $\pm$0.33	0.48 $\pm$0.09	2.88 $\pm$1.05	0.0515 $\pm$0.0160	2.33 $\pm$0.40	ADV+IMP 2.81 ± 0.18
Boundary	0.52 $\pm$0.15	0.41 $\pm$0.06	3.92 $\pm$0.81	0.0518 $\pm$0.0051	2.51 $\pm$0.31	SULI 5.84 ± 0.19

Table 14: **HLMR best preserves same-class retained examples adjacent to the forget cluster among the evaluated unlearning methods.** Accuracy is measured on disjoint retained controls nearest to and farthest from the feature-clustered forget region. Unlearning results are mean $\pm$ standard deviation over seeds 42–46; matching retraining is the set-specific reference.

Method	Adjacent-Retain accuracy	Remote-Retain accuracy
Matching retraining	100.00	98.00
HLMR	$\mathbf{98.72 \pm 0.32}$	95.04 ± 0.39
NegaGrad	85.96 ± 7.51	58.88 ± 13.74
ADV+IMP	95.44 ± 0.71	93.96 ± 0.74
SULI	95.48 ± 0.52	87.76 ± 1.86
MU-MIS	96.88 ± 0.10	74.68 ± 0.45
JiT	94.52 ± 0.30	60.60 ± 0.77

Table 15: **Full frozen comparison on structured forget sets.** Lower J indicates closer matching-retraining alignment. Values are mean $\pm$ standard deviation over unlearning seeds 42–46; accuracy gaps are percentage points and raw AUC gaps are multiplied by 100 in J.

Method	Random J	Feature-clustered J	Boundary-conditioned J
HLMR	$\mathbf{2.51 \pm 0.25}$	$\mathbf{2.33 \pm 0.40}$	$\mathbf{2.51 \pm 0.31}$
NegaGrad	16.45 ± 2.30	14.56 ± 6.57	39.25 ± 7.97
ADV+IMP	2.68 ± 0.32	2.81 ± 0.18	8.16 ± 0.54
SULI	6.74 ± 0.06	6.13 ± 0.62	5.84 ± 0.19
MU-MIS	10.91 ± 0.03	9.10 ± 0.03	17.05 ± 0.03
JiT	9.73 ± 0.11	9.72 ± 0.10	17.32 ± 0.11

Both references use the same lower empirical quantile Q_q:

$$b_f^P = Q_q\big(\{m_{y_f}(v;\theta^{\mathrm{pre}}) : v \in \mathcal{P}_f\}\big),$$
$$b_f^R = Q_q\big(\{m_{y_f}(u;\theta^{\mathrm{pre}}) : u \in \mathcal{N}_f^R\}\big).$$

The probe–retain Pearson and Spearman correlations and baseline MAE in Table 16 compare b_f^P with b_f^R. Each statistic is computed across 500 forget instances and reported as mean $\pm$ standard deviation over probe seeds 42–46. These standard deviations quantify variation across probe-sampling seeds with the pretrained model, forget set, and retained-neighbor reference held fixed. The positive association supports the baseline as a retain-free local reference at the level of this margin statistic.

To test its relationship with retraining, we compare

$$\widehat{\Delta}_f = [m_f^{\mathrm{pre}} - b_f^P]_+, \qquad \Delta_f^\star = m_f^{\mathrm{pre}} - m_f^{\mathrm{retrain}}.$$

Here $[x]_+ = \max(x, 0)$, m_f^{pre} is the frozen teacher's directional margin on the forget instance, and m_f^{retrain} uses the matching retrained model's own representation and normalized classifier. These frozen-teacher diagnostics evaluate the local reference construction, while unlearning recomputes local quantities around the current student center at each refresh, with scalar smoothing specified in Appendix I. HLMR uses the gated, normalized, and clipped excess for bounded adaptation; the raw contrast itself has near-zero correlation with the signed retraining response (Table 16).

Table 16: **The probe baseline tracks retained-neighbor margins, while raw excess exhibits near-zero correlations with the instance-wise retraining response.** Each statistic uses 500 forget instances. Cells give the mean, with ± standard deviation on the second line, over probe seeds 42–46. Clustered and Boundary denote the feature-clustered and boundary-conditioned sets.

	Probe–retain baseline			Excess–oracle response	
Forget set	Pearson	Spearman	MAE	Pearson	Spearman
Random	0.7960 ±0.0001	0.8116 ±0.0002	0.056327 ±0.000005	0.0232 ±0.0338	0.0112 ±0.0285
Clustered	0.7762 ±0.0001	0.7828 ±0.0002	0.066048 ±0.000011	−0.0128 ±0.0540	−0.0054 ±0.0377
Boundary	0.8794 ±0.0001	0.8997 ±0.0002	0.076667 ±0.000016	0.0241 ±0.0365	0.0230 ±0.0319

We test whether this finding depends on one retrained reference by training four additional matching oracles with seeds 43–46 for each forget set. All five oracles share the pinned ImageNet initialization and frozen retraining protocol; the seed controls classifier-head initialization and data order. Table 17 reports agreement among the five response vectors and the correlation between raw excess and their mean response. Pairwise Pearson summaries cover all ten retraining-seed pairs. ICC(3,1) is the single-measure consistency intraclass correlation coefficient. The excess–mean-oracle correlations are summarized over probe seeds 42–46. The response is repeatable across retraining seeds, while its mean remains nearly uncorrelated with the raw excess. Retained instances and retrained models enter only these evaluations. These diagnostics show that the probe baseline tracks a retained-neighbor margin statistic; the structured-set, local-retention, and field-perturbation experiments assess the resulting reference-guided unlearning behavior.

Table 17: **The near-zero excess–oracle association persists across retraining seeds.** Pairwise response correlations report median [minimum, maximum] over the ten pairs of retraining seeds. Excess–mean-oracle correlations are mean ± standard deviation over probe seeds 42–46.

Forget set	Pairwise $\Delta^\star$ Pearson median [min, max]	ICC(3,1)	Excess–mean-oracle	
			Pearson	Spearman
Random	0.847 [0.815, 0.864]	0.843	0.026 ±0.048	0.009 ±0.036
Feature-clustered	0.831 [0.809, 0.834]	0.828	−0.019 ±0.050	−0.021 ±0.033
Boundary-conditioned	0.820 [0.807, 0.829]	0.819	0.032 ±0.031	0.029 ±0.032

S.4 Sensitivity to directional-field distortion

The controlled intervention changes only the frozen pretrained teacher classifier directions used for probe-margin evaluation on the CIFAR-100/ResNet-50 boundary-conditioned set. For a rotation angle ϑ, each unit-normalized class direction $\hat{w}_c$ is replaced by $\cos(\vartheta)\,\hat{w}_c + \sin(\vartheta)\,q_c$, where q_c is a random unit vector orthogonal to $\hat{w}_c$ drawn from a field seed. The student starts from the unchanged pretrained model; the stabilizer's parameter reference, forget set, matching retraining oracle, and configuration also remain fixed. Rotating the normalized classifier directions preserves continuity and Lipschitz regularity while degrading quantitative directional fidelity. Table 18 gives the complete results for rotation angles 0, 0.4, 0.8, 1.2, and $\pi/2$, and Figure 12 plots them. Accuracy and J are mean ± standard deviation over five clean runs and 15 runs per nonzero angle (three field seeds, each with unlearning seeds 42–46). Field Pearson correlation and sign agreement compare clean and corrupted probe margins and are means across field seeds. The matching retraining Test/Retain/Forget accuracy is $81.46/99.94/58.80$.

High sign agreement coexists with deteriorating retraining alignment, demonstrating the role of quantitative directional fidelity. This requirement differs from global probability calibration: global temperature scaling can alter softmax calibration while leaving normalized representations, classifier directions, and probe targets unchanged. The controlled rotations identify a directional-field failure

mode; distribution shifts that invalidate semantic locality require separate, matching-retraining evaluations.

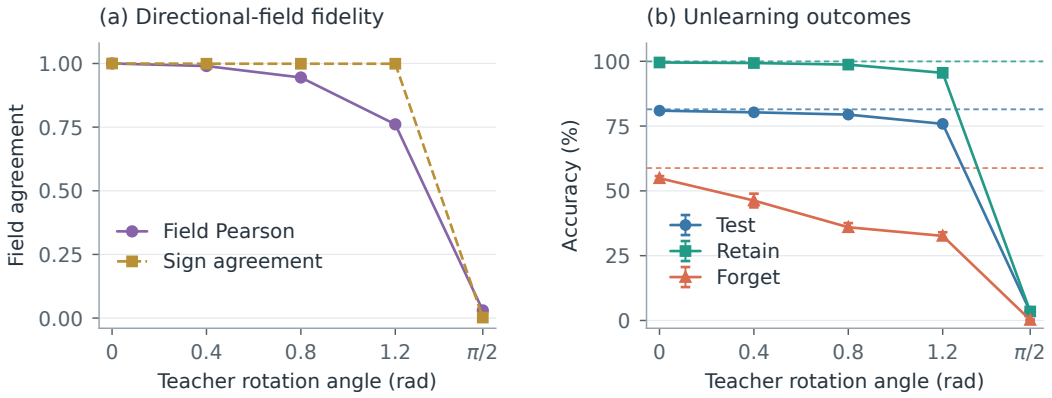


Figure 12: **Directional-field distortion degrades retraining alignment.** (a) Field Pearson correlation falls while sign agreement remains high through rotation 1.2. (b) Accuracy departs from matching retraining (dashed lines). Field diagnostics are means across field seeds; accuracy error bars show one standard deviation over five clean runs or 15 runs per nonzero angle.

Table 18: **Directional-field distortion exposes a quantitative reliability requirement.** Field diagnostics compare clean and rotated probe margins. Accuracy and J are mean $\pm$ standard deviation over five clean runs or 15 runs per nonzero angle; field diagnostics are means across field seeds. The matching retraining Test/Retain/Forget target is $81.46/99.94/58.80$.

Rotation angle	Field Pearson / sign agreement	Test	Retain	Forget	J
0	$1.000/1.000$	80.94 ± 0.15	99.53 ± 0.06	54.88 ± 0.81	2.51 ± 0.31
0.4	$0.990/0.999$	80.30 ± 0.18	99.32 ± 0.08	46.27 ± 2.65	5.27 ± 0.83
0.8	$0.945/0.999$	79.43 ± 0.23	98.71 ± 0.10	35.97 ± 1.60	8.29 ± 0.48
1.2	$0.761/0.999$	75.85 ± 0.46	95.53 ± 0.38	32.63 ± 1.40	9.93 ± 0.53
$\pi/2$	$0.030/0.002$	3.40 ± 1.20	3.47 ± 1.33	0.20 ± 0.16	64.01 ± 1.22

Table 19: Information budget and deletion task of related method families. "None" means no additional resource beyond the pretrained model and forget set at request time. The empirical ranking concerns the matched instance-wise deep-classifier panel.

Method family	Additional information	Task/model scope	Role in comparison
HLMR and the five zero-shot baselines	None; internal probes or proxies are permitted	Instance-wise; deep classifiers	Main matched panel
Ahmed et al. [1]	None	Instance-wise; linear/mixed-linear models	Information-matched; different model formulation
SCAR [3]	External OOD data and stored training statistics	Instance-wise; deep models	Auxiliary-data protocol
LoTUS [40]	Retain subset and unseen/calibration set	Instance-wise; deep models	Auxiliary-data protocol
Min-Max [13]	None	Class-wise; deep models	Different deletion unit
UNSIR [41]	Retain samples	Class-wise; deep models	Different deletion unit and information budget

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state that the paper studies zero-shot instance-wise machine unlearning and proposes HLMR as a hyperspherical local margin retraction method using teacher-evaluated geodesic probes. The empirical claims are limited to the evaluated vision classification benchmarks and architectures, and the conclusion explicitly frames the results relative to the retraining oracle and the evaluated zero-shot baselines.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Appendix Q discusses the empirical scope of the privacy evaluation, dependence on the pretrained teacher's directional margin field, the locality assumption on probe radius, and the vision benchmarks and forget-set constructions covered by the evaluation.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The regularity bound on the directional-margin field (Eq. (11)) is stated with its assumptions and proved in full in Appendix B. The remaining equations define the hyperspherical representation space, geodesic probes, margin baseline, retraction anchor, and optimization objective.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper describes the datasets, preprocessing, forget-retain split construction, original and oracle model training, evaluation metrics, baselines, and ablations in the appendices.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We release the HLMR implementation and demo as supplementary material.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Appendix specifies the datasets, preprocessing, forget ratios, class-stratified forget-set construction, original/retrained model training recipes, and evaluation metrics. We also specify baseline implementation details, hyperparameters, ablation study, and hyperparameter sensitivity. The protocol and configuration of the additional analyses are given in Appendix S.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

Justification: The paper reports mean and standard deviation over five seeds for the main experimental results and ablations. The variability captures stochasticity from the unlearning procedure under fixed dataset splits and matched evaluation protocols. The additional analyses in Appendix S likewise report mean and standard deviation over seeds.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: In appendix, we report per-method unlearning runtime and complexity, together with the hardware and software environment used for the experiments, including GPU type, GPU memory, CPU resources, and software versions (Appendix M). Runtime is measured as wall-clock time for the core unlearning computation and excludes original training, retraining, attack evaluation, plotting, logging, and checkpoint input/output.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work uses publicly available benchmark datasets and does not involve human-subject experiments, private personal data collection, or deployment on real users. The research is aimed at improving privacy-preserving and accountable model maintenance through machine unlearning.

   Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Appendix discusses positive impacts for privacy regulation, copyright-removal requests, data-retention constraints, and fine-grained model maintenance. It also notes the need for governed audit trails so that unlearning is not used to obscure training-set membership or overstate deletion guarantees.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release high-risk pretrained language models, image generators, scraped datasets, or models intended for direct deployment.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.

- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The paper credits the existing assets, and the supplementary material identifies the software dependencies and their license references.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The HLMR implementation is documented through the accompanying demo.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing, human-subject experiments, user studies, or participant compensation.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve human subjects, crowdsourcing, user studies, or participant compensation.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs are not used as a component of the proposed method, experiments, evaluation protocol, or scientific claims.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.