ERASE, DON'T DISTORT: MACHINE UNLEARNING THROUGH REPRESENTATION REALIGNMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Machine unlearning is often carried out by deliberately increasing the loss on the forget set to suppress the model's response to targeted samples. While effective at inducing forgetting, this can leave a distinct statistical footprint that is vulnerable to membership inference attacks, a phenomenon known as the Streisand effect. The core issue is that loss maximization specifies what should be erased without controlling how the underlying representations evolve, often driving forgotten samples into abnormally high-loss regions that are easy to distinguish from naturally unseen data. We instead formulate unlearning as a controlled subspace realignment process that selectively modifies forget-specific dimensions while preserving the latent structure shared by retained data. To realize this formulation, we first isolate a low-dimensional subspace that captures the dominant discrepancy between forget and retain representations, while constraining its orthogonal complement to remain close to the original model. Within this subspace, we then realign the forget distribution toward a nearby retain distribution by matching both its centroid and its dispersion, thereby avoiding the abnormal concentration induced by destructive forgetting. This design confines forgetting to the representation directions where it is needed while preserving shared structure elsewhere. Comprehensive evaluations across CIFAR-10, CIFAR-100, SVHN, Tiny ImageNet, and ImageNet32-200 demonstrate a strong privacy–utility trade-off, driving membership inference performance close to the random-guess regime ($\sim 50\%$) while matching or surpassing representative baselines in retain utility. Our code is available here.

1 INTRODUCTION

Deep neural networks have a well-documented tendency to memorize their training data (Zhang et al., 2017), creating a fundamental tension with privacy requirements when sensitive records or proprietary information must be removed. To satisfy legal frameworks such as the "Right to be Forgotten" (Voigt & Von Dem Bussche, 2024; Villaronga et al., 2018), models should be able to eliminate the influence of selected training instances or entire classes (Bourtoule et al., 2019) without resorting to costly retraining from scratch. This requirement has motivated the development of *machine unlearning* (Cao & Yang, 2015), a paradigm designed to selectively remove targeted knowledge while preserving model utility on retained data. Ideally, a successful unlearning algorithm should render a model's outputs and latent representations on forgotten samples computationally and statistically indistinguishable from those of a model retrained from scratch (Xu et al., 2023).

Typical unlearning methods often pursue this goal by deliberately degrading the model's response to the forget set, with approaches like gradient ascent (Thudi et al., 2022), parameter-level dampening (Foster et al., 2023), or unconstrained entropy maximization (Kamberaj, 2020). Yet, such a process can be inherently destructive. These approaches frequently inflate forget-set losses to abnormally high values—for instance, driving prediction confidence to near-zero or forcing logit outputs into extreme, unconstrained regimes. Crucially, a naturally unseen sample does not require zero accuracy or an artificially maximized loss to be considered unlearned; it simply needs to reflect the model's natural generalization state. These abnormal high loss values also introduce detectable distributional anomalies in the latent and logit spaces, allowing Membership Inference Attacks (MIAs) (Yeom et al., 2018) to identify distributional discrepancies between unlearned samples and naturally unseen test data (Carlini et al., 2022b). Consequently, standard unlearning frameworks often fall

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

prey to a dual vulnerability: the *Streisand effect* (Shokri et al., 2017), where aggressive forgetting makes erased samples more statistically conspicuous, and *shared representation corruption*, where the same updates damage features needed by retained data.

To resolve these structural bottlenecks, we propose "QF-Entanglement", a representation-level unlearning framework that replaces destructive forgetting with controlled subspace realignment. The key idea is to separate forget-specific variation from representation structure shared with retained data. Specifically, we first construct a low-dimensional discriminative subspace from the discrepancy between forget and retain features, while constraining its orthogonal complement to remain close to the original model. Within this active subspace, we then redirect forgotten representations toward a nearby retain distribution by matching both their centroid and their dispersion. To control the latter, we represent the centered feature Gram matrix as a normalized density operator and regularize its linear entropy, which prevents the collapsed or abnormally concentrated feature distributions induced by aggressive loss maximization. Together, these components confine unlearning to the representation directions where it is needed while reducing the statistical artifacts associated with the Streisand effect and preserving retain utility.

We evaluate QF-Entanglement across five image-classification benchmarks: CIFAR-10, CIFAR-100, SVHN, Tiny ImageNet, and ImageNet32-200. Across these settings, QF-Entanglement consistently brings membership-inference performance close to the random-guess regime ($\sim 50\%$) while preserving strong retain accuracy. Compared with representative gradient-, parameter-, and representation-based unlearning baselines, it achieves a more favorable privacy–utility trade-off, with particularly strong gains on datasets where destructive forgetting otherwise causes substantial representation distortion.

Our main contributions are summarized as follows:

- **A Realignment Paradigm for Unlearning**: We shift the goal of machine unlearning from purely suppressing the forget set to statistically realigning forgotten representations with valid retained states, reducing the abnormal signatures that give rise to the Streisand effect.
- **Subspace-Isolated Distribution Realignment**: We formulate an unlearning objective that combines differential-PCA subspace isolation with orthogonal-complement preservation, together with first-order centroid anchoring and second-order linear-entropy alignment.
- **Comprehensive Empirical Evaluation**: Across five standard benchmarks, our method achieves a favorable privacy–utility trade-off, consistently bringing MIA performance close to the random-guess regime while maintaining strong target erasure and retain-set utility against representative unlearning baselines.

2 Related Works

Optimization-Based Unlearning. Early approaches formulate unlearning through direct optimization on the forget set, including Gradient Ascent (GA) (Thudi et al., 2022), Delete (Neel et al., 2020), and fine-tuning-based methods (Kurmanji et al., 2023; Cheng et al., 2026). To reduce the collateral forgetting caused by unconstrained ascent (Kirkpatrick et al., 2017), Gradient Difference Retraining (GDR) (Golatkar et al., 2020) jointly optimizes retain and forget objectives by combining their gradients in opposite directions. While effective at suppressing targeted behavior, these methods operate directly in parameter space, where forget and retain objectives can interfere through shared parameters (Pezeshki et al., 2021). Consequently, updates intended to remove forget-specific information may perturb representations of retained data, leading to degradation in retain utility.

Parameter Localization and Dampening. To reduce the collateral damage caused by broad parameter updates, later methods restrict unlearning to a subset of weights deemed most relevant to the forget set (Cha et al., 2024; Wuerkaixi et al., 2025; Patel & Qiu, 2025). Selective Synaptic Dampening (SSD) (Foster et al., 2023) uses diagonal Fisher information to identify and dampen parameters that are important for forgotten samples but less critical to retained data. Similarly, SALUN (Fan et al., 2024) employs gradient-based saliency to localize the parameters updated during unlearning. These methods improve retain utility by limiting parameter updates, but they do not control how forgotten representations move in feature space. Consequently, target suppression may still leave detectable deviations from naturally unseen data under membership inference attacks (MIAs).

Representation Distillation and Boundary Alignment. Recent methods also move beyond direct parameter manipulation toward representation matching and decision-boundary adjustment (Hinton et al., 2015). SCRUB (Kurmanji et al., 2023) adopts a teacher–student minimax objective that drives student predictions away from the original model on the forget set while preserving agreement on retained data. Geometric approaches such as MUNBA (Chen et al., 2023) instead modify decision boundaries relative to neighboring retain classes. While these methods provide finer control over the model's output behavior, they do not necessarily constrain how forget representations evolve throughout the feature hierarchy, leaving residual target information or abnormal latent geometry possible (Ganju et al., 2018). Inversely, recent entanglement-aware schemes (Cheng et al., 2026) primarily optimize unlearning through multi-stage constrained loss balancing, which can struggle to constrain geometric feature dispersion in intermediate latent spaces. Unlike existing paradigms that enforce destructive feature dispersion, isolating discriminative subspaces while statistically disguising erased representations remains an open challenge.

3 Problem Formulation

Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ denote the complete training dataset of N samples, where $x_i \in \mathcal{X}$ represents an input instance and $y_i \in \mathcal{Y} = \{1, \ldots, C\}$ is its corresponding ground-truth label. Consider a deep neural network $f_\theta : \mathcal{X} \to \mathcal{Y}$, parameterized by θ, which can be decomposed into a deep feature extractor $\Phi_\theta : \mathcal{X} \to \mathbb{R}^D$ and a linear classification head $g_\theta : \mathbb{R}^D \to \mathbb{R}^C$ such that $f_\theta(x) = g_\theta(\Phi_\theta(x))$. The model is initially trained on $\mathcal{D}$ with baseline parameters θ_0. Given a specific subset to be unlearned $\mathcal{D}_f \subset \mathcal{D}$ (the *forget set*) and the remaining data $\mathcal{D}_r := \mathcal{D} \setminus \mathcal{D}_f$ (the *retain set*), the fundamental objective of machine unlearning is to produce an updated model $f_{\tilde{\theta}}$ parameterized by $\tilde{\theta}$ that thoroughly eliminates the influence of $\mathcal{D}_f$ while preserving the predictive performance on $\mathcal{D}_r$.

Limitations of Classical Formulations. Standard unlearning paradigms typically evaluate forgetting efficacy merely through classification metrics on $\mathcal{D}_f$, predominantly optimizing for near-zero forget accuracy or maximized output cross-entropy. However, such destructive optimization regimes often suffer from two critical pitfalls:

1. **Superficial Erasure:** They tend to modify only the decision boundary of the classifier g_θ while leaving discriminative representations within the deep feature extractor Φ_θ intact.
2. **The Streisand Effect:** Forcibly suppressing predictions or indiscriminately scattering features introduces systematic distributional anomalies in the logit confidence space. Hence, state-of-the-art Membership Inference Attacks (MIAs), such as Likelihood Ratio Attacks (LiRA) (Carlini et al., 2022a), can reliably identify distinguish unlearned instances from genuinely unseen test data, paradoxically exacerbating privacy vulnerability.

Unlearning Objectives via Statistical Indistinguishability. To mitigate these limitations, we argue that true unlearning should not be treated as empirical complete destruction, but rather as *representation-level statistical manifold realignment* (Guo et al., 2023)—erasing class-discriminative features while redirecting representations onto adjacent retain manifolds. Formally, the resulting model $f_{\tilde{\theta}}$ must simultaneously satisfy two synergistic goals:

- **Utility Preservation:** The updated model must retain its classification accuracy and generalization capacity on the retain distribution, satisfying:
$$\mathbb{E}_{(x,y)\sim\mathcal{D}_r}\left[\mathbb{I}(f_{\tilde{\theta}}(x) = y)\right] \approx \mathbb{E}_{(x,y)\sim\mathcal{D}_r}\left[\mathbb{I}(f_{\theta_0}(x) = y)\right]. \tag{1}$$
- **Feature-Space Indistinguishability (MIA Defense):** The representation distribution of the forget set $\mathcal{P}(\Phi_{\tilde{\theta}}(\mathcal{D}_f))$ must be reshaped to match the empirical moment statistics (Heusel et al., 2018) (e.g., mean centroid and variance dispersion) of a target retain class distribution $\mathcal{P}(\Phi_{\tilde{\theta}}(\mathcal{D}_{r,c^*}))$, such that any MIA adversary $\mathcal{A}_{\text{MIA}}$ achieves an attack accuracy indistinguishable from random guessing:
$$\text{Acc}\left(\mathcal{A}_{\text{MIA}}(f_{\tilde{\theta}}, \mathcal{D}_f)\right) \approx 50\%. \tag{2}$$

In this work, we operationalize this formulation by jointly aligning the first-order (centroid) and second-order (entropic dispersion) statistics of $\Phi_{\tilde{\theta}}(\mathcal{D}_f)$ to establish provable behavioral manifold realignment without corrupting shared semantic representations.

4 METHODS

4.1 DECOUPLING ENTANGLEMENT: DIFFERENTIAL SUBSPACE PROJECTION

Modifying representations across the entire ambient feature space $\mathbb{R}^D$ inevitably corrupts the shared generic features crucial for classifying $\mathcal{D}_r$. To decouple the forget-exclusive semantics from these domain-shared representations, we propose isolating the discriminative variance of $\mathcal{D}_f$ via differential principal component analysis (Shlens, 2014).

Let $\Phi_\theta : \mathcal{X} \to \mathbb{R}^D$ denote the deep feature extractor parameterized by θ. We first compute the retain distribution centroid $\mu_r = \frac{1}{|\mathcal{D}_r|} \sum_{x \in \mathcal{D}_r} \Phi_{\theta_0}(x)$ using the frozen pre-unlearning model θ_0. For a representative sample of forget representations $F_f \in \mathbb{R}^{M \times D}$ initially extracted with low computational cost using the baseline model θ_0, we construct the differential residual matrix $\Delta F = F_f - 1_M \mu_r$. By performing singular value decomposition (SVD) (Golub & Reinsch, 1970) on $\Delta F = U \Sigma V^\top$, where U and V contain the left and right singular vectors respectively, and Σ is the diagonal matrix of singular values, the right singular vectors $V \in \mathbb{R}^{D \times D}$ capture the orthogonal directions of maximum deviation. The initial projection matrix $P^{(0)}$ is extracted with complexity $\mathcal{O}(\min(MD^2, M^2D))$ (independent of total class count C). To accommodate representation drift during optimization, $P^{(e)}$ is dynamically tracked across epochs using current checkpoint representations. We determine the active subspace dimension k via the cumulative explained variance threshold τ:

$$k = \min \left\{ j \;\middle|\; \frac{\sum_{i=1}^{j} \sigma_i^2}{\sum_{i=1}^{\min(M,D)} \sigma_i^2} \geq \tau \right\} \tag{3}$$

where σ_i denotes the i-th singular value. Crucially, to prevent projecting onto isotropic noise once target concepts dissolve, we enforce a variance-gated freeze: subspace updates terminate when the leading singular value satisfies $\sigma_1^{(e)} < \gamma \sigma_1^{(0)}$ (with collapse threshold $\gamma = 0.5$), after which the projection operator $P = V_{:,1:k} \in \mathbb{R}^{D \times k}$ is permanently fixed. During training, any mini-batch of deep features $F \in \mathbb{R}^{B \times D}$ is uniquely decomposed into a discriminative component $F_{disc} = FP \in \mathbb{R}^{B \times k}$ and its orthogonal complement $F_{null} = F(I - PP^\top) \in \mathbb{R}^{B \times D}$. To preserve shared representation capacity, we enforce null-space invariance against the frozen reference model:

$$\mathcal{L}_{\text{preserve}}(\theta) = \frac{1}{BD} \left\| F_{\text{null}}(\theta) - F_{\text{null}}(\theta_0) \right\|_F^2 . \tag{4}$$

By restricting the main unlearning updates to the k-dimensional subspace while anchoring F_{null}, we formally bound the collateral utility degradation.

Assumption 1 (Retain-subspace stability)**.** *For the final parameter θ returned by unlearning, the retain-set representation perturbation in the orthogonal complement satisfies:*

$$\mathbb{E}_{x \sim \mathcal{D}_{retain}} \left[\left\| (I - PP^\top)(\Phi_\theta(x) - \Phi_{\theta_0}(x)) \right\|_2^2 \right] \leq \delta^2, \tag{5}$$

where $\delta^2 > 0$ denotes the upper bound on the energy of the representation perturbation.

Proposition 1 (Utility Preservation Bound)**.** *Let W_θ and W_{θ_0} denote the weight matrices of the linear classifier head for the post- and pre-unlearning models, respectively. Let $\mathcal{L}_{retain}(\theta) = \mathbb{E}_{(x,y) \sim \mathcal{D}_{retain}}[\ell(W_\theta \Phi_\theta(x), y)]$ be the loss on the retain set, where the classification loss $\ell(z, y)$ is L_ℓ-Lipschitz with respect to the logits z. If the perturbation in the k-dimensional subspace is bounded by ϵ_k (i.e., $\mathbb{E}\|P^\top(\Phi_\theta - \Phi_{\theta_0})\|_2 \leq \epsilon_k$), and the classifier head updates are bounded by $\|W_\theta - W_{\theta_0}\|_2 \leq \epsilon_W$, then the degradation of the retain loss is strictly upper bounded by:*

$$\mathcal{L}_{retain}(\theta) - \mathcal{L}_{retain}(\theta_0) \leq L_\ell \|W_\theta\|_2 (\epsilon_k + \delta) + L_\ell \epsilon_W \mathbb{E}_{x \sim \mathcal{D}_{retain}} \|\Phi_{\theta_0}(x)\|_2. \tag{6}$$

Detailed proofs for all theoretical results are provided in Appendix A.

4.2 DYNAMIC MULTIMODAL MANIFOLD REALIGNMENT VIA MONTE CARLO SAMPLING

To seamlessly disguise unlearned instances, we reshape forget representations to match the empirical distribution of semantically adjacent retain classes in both moments. To achieve scalable matching

on crowded manifolds while precluding distant interference, we maintain a stochastic candidate pool $\mathcal{C}_{\text{cand}} \subset \mathcal{C}_r$ via epoch-wise Monte Carlo sampling. We compute the forget subspace mean $\bar{f}_{\text{disc}}$ and assign it to the semantically nearest global anchor mode $c^* = \arg\min_{c \in \mathcal{C}_{\text{cand}}} \|\bar{f}_{\text{disc}} - \mu_c^{\text{sub}}\|_2$. This epoch-level anchoring directs forget representations toward semantically adjacent retain manifolds to mimic retrained models, while second-moment regularization preserves intra-class dispersion against point collapse (Ebrahimpour-Boroojeny et al., 2026; Chen et al., 2023; Shen et al., 2025), as corroborated by the preserved latent effective rank in Appendix B.6.

First-Moment Anchoring. During training, we pull the centroid of the mini-batch forget representation toward the selected retain mode:

$$\mathcal{L}_{\text{anchor}}(\theta) = \|\bar{f}_{\text{disc}}^{(b)}(\theta) - \mu_{c^*}^{\text{sub}}\|_2^2. \tag{7}$$

Second-Moment Entropy Matching. While $\mathcal{L}_{\text{anchor}}$ pulls the macroscopic centroid, relying solely on first-order alignment risks causing representations to hyper-concentrate around $\mu_{c^*}^{\text{sub}}$, leaving distinct density anomalies exploitable by membership audits. To enforce realistic intra-class dispersion, we regularize second-moment distribution via density-matrix linear entropy (Von Neumann, 2018).

Specifically, for centered subspace forget features $\tilde{F}_{\text{disc}} \in \mathbb{R}^{B \times k}$, we construct the normalized density operator $\rho = K/(\text{Tr}(K) + \epsilon)$, where $K = \tilde{F}_{\text{disc}} \tilde{F}_{\text{disc}}^\top$ and $\epsilon > 0$. We penalize its deviation from the target retain class linear entropy $S_{c^*}^{\text{sub}}$:

$$\mathcal{L}_{\text{entropy}}(\theta) = (S_L(\rho) - S_{c^*}^{\text{sub}})^2, \tag{8}$$

where $S_L(\rho) = 1 - \|\rho\|_F^2$. This constraint compels the microscopic spread of forget features to match the natural intra-class geometric width of the retain manifold, preventing single-point collapse.

4.3 Unified Optimization Framework and Theoretical Analysis

Combining the retain classification utility, localized subspace manifold realignment, and null-space preservation, the overall unlearning objective is formulated as:

$$\min_\theta \; \mathcal{L}_{\text{total}}(\theta) = \mathcal{L}_{\text{CE}}(\mathcal{D}_r; \theta) + \Big[\lambda_q \mathcal{L}_{\text{entropy}}(\theta) + \beta \mathcal{L}_{\text{anchor}}(\theta) + \mathcal{L}_{\text{preserve}}(\theta)\Big], \tag{9}$$

where $\beta > 0$ balances the first-moment anchoring, and λ_q scales the second-order entropic dispersion constraint. To balance the optimization scales between the cross-entropy loss and the quadratic entropy constraint, we set $\lambda_q = \ln 2 \approx 0.693$ as an empirical scaling factor, which consistently yields stable gradient updates across benchmarks without extensive tuning. For environments with exceptionally dense class manifolds (e.g., ImageNet32-200), λ_q can be proportionally scaled to match the broader target entropy capacity without requiring exhaustive hyperparameter sweeps. The two-phase training pipeline includes the lightweight subspace statistics initialization with low overhead and mini-batch manifold realignment optimization, is summarized in Algorithm 1.

Definition 1 (Moment-matched representation unlearning criterion)**.** *Given a mini-batch of forget representations $\mathcal{B}_f \sim \mathcal{D}_f$ assigned to a globally selected target retain class $c^* \in \mathcal{C}_r$, a model achieves conditional moment-matched unlearning if the gaps in the first moment and linear entropy between the batch forget distribution $P_f^{(b)}$ and the target retain distribution P_{r,c^*} are bounded by ϵ_u, while the overall retain classification degradation satisfies $\Delta\mathcal{L}_{\text{retain}} \leq \epsilon_r$.*

Theoretical Guarantees: Mitigating the Streisand Effect. Our theoretical analysis provides a rigorous representation-level guarantee demonstrating how matching linear entropy prevents the catastrophic manifold collapse that typically triggers membership inference vulnerabilities. Unlike standard parameter-level dampening which drives forget variances to zero, our objective explicitly preserves the geometric dimensionality of the feature space.

Proposition 2 (Linear Convergence of Moment Matching)**.** *Under the Polyak-Łojasiewicz (PL) condition (detailed in Appendix A.2), gradient descent guarantees linear convergence for the forget moment-matching sub-objective $\mathcal{J}_{\text{forget}}(\theta) := \lambda_q \mathcal{L}_{\text{entropy}}(\theta) + \beta \mathcal{L}_{\text{anchor}}(\theta)$. Consequently, the empirical optimization gaps ϵ_a and ϵ_e can be tightly bounded.*

Theorem 1 (Manifold Collapse Prevention via Entropy Matching). *Let Σ_f and Σ_r be the covariance matrices of the forget and retain distributions in the k-dimensional subspace. Let $R_{\mathit{eff}}(\Sigma) = (Tr\Sigma)^2/Tr(\Sigma^2)$ denote the effective representation rank. If the linear entropy gap is bounded by ϵ_e (i.e., $|S_L(\rho_f) - S_L(\rho_r)| \leq \epsilon_e$), the discrepancy in their effective dimensionalities is strictly bounded:*

$$|R_{\mathit{eff}}(\Sigma_f) - R_{\mathit{eff}}(\Sigma_r)| \leq \frac{\epsilon_e}{(1 - S_L(\rho_r))(1 - S_L(\rho_f))} \tag{10}$$

Theorem 1 proves that entropy matching structurally prevents point-collapse. If the target retain class is diverse ($S_L(\rho_r) \gg 0$), the unlearned forget representations are mathematically guaranteed to maintain an equivalent macroscopic dispersion, averting the density singularities exploited by likelihood-based attacks.

Corollary 1 (Likelihood Anomaly Neutralization). *Assuming convergence of the proposed gradient updates (Kingma & Ba, 2017), the forget features align with the target retain manifold in both centroid and effective rank. Consequently, for any likelihood-based membership auditor, the expected log-density of the forget samples remains bounded relative to the retain distribution, neutralizing the variance-collapse anomaly (the Streisand effect) while leaving only higher-order non-Gaussian residuals exploitable.*

5 Experiment

5.1 Experimental Setup

We evaluate QF-Entanglement on CIFAR-10, CIFAR-100 (Krizhevsky, 2009), SVHN (Netzer et al., 2011), Tiny ImageNet (Le & Yang, 2015), and ImageNet32-200 (Chrabaszcz et al., 2017). To ensure a completely fair sandbox environment and maintain the purity of the initial weight space, we employ a ResNet-18 (He et al., 2015) backbone trained strictly from scratch on the full datasets, explicitly avoiding pre-trained weight initialization.

Our primary unlearning protocol focuses on class unlearning Kurmanji et al. (2023), where Class 0 is designated as the primary target class following convention (other classes/superclasses in Appendix B.8). The unlearning phase is executed for 10 epochs on SVHN, CIFAR-10, and CIFAR-100, and 3 epochs for the denser Tiny ImageNet and ImageNet32-200 distributions. Baseline implementations (Fine-Tune, Gradient Ascent, GDR, SSD, SALUN, SCRUB, Munba, Delete and RF-Entanglement) default to the optimal hyperparameters reported in their original literature, supplemented by empirical fine-tuning to guarantee competitive baselines (configurations are deferred to the supplementary).

To evaluate algorithmic stability, we conduct 5-seed trials (seeds: 1, 42, 123, 999, 2027) on CIFAR-100 (detailed benchmark results and multiseed analysis are deferred to Appendix B.3). Given the negligible variance observed for our method across independent runs, and considering the computational constraints, we standardize on seed 42 for remaining benchmark evaluations following common practice. We evaluate unlearning performance via retain accuracy, forget accuracy, and the LiRA-MIA true-positive rate (MIA_{true}, with an ideal target of $\approx 50\%$). By estimating sample-specific likelihood ratios through shadow models, LiRA provides a strong likelihood-based privacy audit. Extended evaluations under strict low false-positive rates (TPR@Low FPR) are detailed in Appendix B.7.

5.2 Main Results And Efficacy

QF-Entanglement significantly mitigates membership inference threats, consistently centering MIA_{true} at random-guess levels across all tested distributions. In contrast, many baseline methods hover in abnormal MIA ranges or suffer massive utility penalties to achieve privacy. Utility preservation is remarkably stable even in dense classification tasks where high-dimensional feature corruption usually collapses models. In our manifold realignment paradigm, forget representations have their class-discriminative features erased and redirected onto adjacent retain manifolds rather than being heuristically suppressed. While heuristic baselines often prioritize zero forget accuracy via direct logit suppression at the expense of retention utility or MIA stability (introducing detectable statistical anomalies), QF-Entanglement balances this trade-off: by focusing on representation

manifold realignment rather than strict logit nullification, it achieves substantial target suppression while preserving feature integrity and maintaining MIA close to the nominal 50% random-guess baseline, where minor residual forget accuracy naturally reflects shared manifold support.

Table 1: Unlearning Performance on CIFAR-10

Method	Retain Acc	Forget Acc	Retain Test	MIA
Retrain	99.46	0.00	93.44	50.00
Delete	99.82	0.00	93.83	51.10
Fine-tune	98.60	1.28	92.87	51.50
GDR	93.84	0.00	87.59	48.80
Gradient Ascent	98.79	3.98	92.22	49.60
Munba	99.58	57.92	93.53	51.00
SalUn	98.95	0.80	92.54	50.60
SCRUB	99.48	66.32	93.38	51.90
SSD	50.30	0.00	49.78	48.60
RF-Entanglement	99.38	0.00	93.28	49.20
Ours	**99.92**	**0.00**	**94.18**	**50.30**

Unlearning Performance on CIFAR-10. As shown in Table 1, SCRUB and Munba exhibit insufficient erasure on the forget set (66.32% and 57.92% forget accuracy, respectively). While baselines such as Delete, SalUn, and RF-Entanglement achieve substantially lower forget accuracy, methods such as FT and GDR incur noticeable drops in retain test accuracy (92.87% and 87.59%), and SSD suffers severe utility degradation (49.78%). In comparison, our method achieves complete target erasure (0.00% forget accuracy) while preserving the highest retain test accuracy among the evaluated methods (94.18%), comparable to the retrained reference (93.44%). More importantly, strong forgetting does not come at the cost of increased membership exposure: our method yields an MIA score of 50.30%, closely matching the random-guess regime. This is particularly important because low forget accuracy alone does not guarantee successful unlearning; aggressive suppression can leave forgotten samples statistically distinguishable from naturally unseen data. The results therefore show that our method simultaneously removes the target behavior, preserves retained utility, and avoids introducing a detectable membership signature.

Table 2: Unlearning Performance on SVHN

Method	Retain Acc	Forget Acc	Retain Test	MIA
Retrain	99.99	0.00	96.18	50.00
Delete	**100.00**	**0.00**	**96.26**	51.15
Fine-tune	99.71	**0.00**	94.78	50.97
GDR	98.31	**0.00**	93.25	40.71
Gradient Ascent	99.99	**0.00**	95.48	45.58
Munba	98.50	10.41	95.56	48.91
SalUn	99.99	**0.00**	95.69	45.07
SCRUB	**100.00**	31.59	96.20	51.83
SSD	**100.00**	**0.00**	95.93	51.72
RF-Entanglement	99.98	**0.00**	95.48	38.65
Ours	**100.00**	0.49	96.01	**49.71**

Results on Structurally Simple Domains (SVHN). On SVHN (Table 2), optimization- and gradient-based baselines (GDR, SalUn, RF-Entanglement) achieve complete forgetting (0.00%) but leave distributional distortions, driving MIA to abnormal regimes (38.65%–45.58%). Conversely, MUNBA and SCRUB exhibit insufficient erasure. SSD serves as a strong benchmark (0.00% forget, 95.93% retain, 51.72% MIA). In comparison, our method achieves near-complete erasure (0.49% forget) while slightly improving retain generalization (96.01% retain) and tightly aligning with the ideal MIA threshold (49.71%). We note that for intrinsically compact manifolds like SVHN, the energy of ΔF concentrates sharply in the leading singular spectrum. Therefore, the adaptive explained variance criterion in Eq.(3) naturally resolves to a minimal subspace dimension ($k \ll D$) without requiring manual rank intervention or compromising orthogonal complement invariance.

Table 3: Unlearning Performance on ImageNet32-200

Method	Retain Acc	Forget Acc	Retain Test	MIA
Retrain	99.82	0.00	65.63	50.00
Delete	69.08	3.08	45.61	**50.00**
Fine-tune	77.71	0.38	58.74	52.00
GDR	87.73	34.92	56.19	54.00
Gradient Ascent	92.37	1.08	58.63	**50.00**
Munba	93.99	47.92	59.44	56.00
SalUn	93.58	0.69	59.13	48.00
SCRUB	93.97	28.15	59.35	54.00
SSD	**98.39**	**0.00**	**64.11**	46.00
RF-Entanglement	96.10	27.38	61.52	60.00
Ours	97.78	**0.00**	62.87	**50.00**

Results on Dense-Class Feature Spaces (ImageNet32-200). Under high class density (Table 3), representation-level entanglement becomes severe: Munba, GDR, SCRUB, and RF-Entanglement exhibit substantial information leakage, while FT suffers from degraded retain fitting (77.71%). Delete suffers catastrophic utility collapse (45.61% retain). Among robust baselines, SSD preserves retain accuracy (64.11%) but introduces an MIA shift (46.00%). In contrast, our method eliminates target information (0.00% forget accuracy) while preserving 62.87% retain test accuracy and maintaining the ideal 50.00% MIA rate. This indicates that subspace preservation prevents severe feature space distortion across 200 entangled classes.

Table 4: Unlearning Performance on Tiny-ImageNet

Method	Retain Acc	Forget Acc	Retain Test	MIA
Retrain	99.97	0.00	64.53	50.00
Delete	98.68	19.20	58.35	56.00
Fine-tune	87.32	1.20	55.66	**50.00**
GDR	97.22	20.40	58.07	52.00
Gradient Ascent	99.42	**0.00**	60.10	52.00
Munba	99.68	56.00	60.75	56.00
SalUn	99.45	**0.00**	60.11	**50.00**
SCRUB	99.65	11.20	60.52	52.00
SSD	**99.89**	**0.00**	**62.98**	52.00
RF-Entanglement	99.77	17.40	60.97	56.00
Ours	99.73	**0.00**	60.94	**50.00**

Results on Structured Superclass Domains (Tiny-ImageNet). Across structured semantic groups (Table 4), boundary adjustments (Munba) fail to sufficiently eliminate target concepts (56.00% forget), while FT incurs noticeable retain utility degradation (55.66%) and degraded retain fitting (87.32%). Partial suppression is observed in Delete, GDR, and RF-Entanglement, accompanied by elevated MIA rates (52.00%–56.00%). Baselines such as SSD and GA enforce 0.00% forget accuracy but deviate from the random-guess MIA baseline (52.00%). In comparison, our method achieves complete target erasure (0.00% forget) and solid retain generalization (60.94%) while strictly aligning with the optimal 50.00% MIA rate. This confirms that our framework effectively isolates target representations even within complex superclass hierarchies.

5.3 Ablation Studies

To evaluate the contribution of each component, we conduct ablation studies across CIFAR-100 and ImageNet32-200 (Table 5). Removing subspace projection substantially degrades retain accuracy and leads to incomplete unlearning. Without centroid anchoring, target erasure collapses entirely, leaving forget accuracy above 89%. Furthermore, omitting entropy matching increases MIA vulnerability and induces target concept leakage (yielding 0.31% forget accuracy on ImageNet32-200), confirming that second-order dispersion regularization is necessary to eliminate residual density

Table 5: Ablation Studies on CIFAR-100 and ImageNet32-200

Method	CIFAR-100				ImageNet32-200			
	Ret. Acc	Forget Acc	Ret. Test	MIA	Ret. Acc	Forget Acc	Ret. Test	MIA
w/o Subspace	98.53	50.00	71.53	52.00	90.88	13.23	57.33	56.00
w/o Entropy	99.96	**0.00**	76.45	51.00	**98.15**	0.31	63.24	52.00
w/o Anchor	**99.97**	99.80	**76.85**	53.00	**98.15**	89.92	**63.36**	52.00
Full Model	99.96	**0.00**	76.56	**50.00**	97.78	**0.00**	62.87	**50.00**

anomalies on complex manifolds. These results demonstrate that each component is indispensable for balancing retain utility, complete erasure, and privacy defense.

5.4 Sensitivity Analysis

We examine the sensitivity of QF-Entanglement to the anchoring coefficient β on CIFAR-100. As detailed in Appendix B.4, evaluating $\beta \in [0.01, 300]$ yields consistently high retain accuracy ($\geq 99.89\%$) and marginal variations in retain test accuracy (75.72%–76.64%). For all tested $\beta \geq 0.05$, the forget accuracy is effectively reduced to 0.00% (0.40% and 0.60% at $\beta = 0.01$ and 0.03), while the true MIA score stays close to the 50% random-guess baseline (47.00%–52.00%). These observations suggest that our method is relatively insensitive to the choice of β across several orders of magnitude, preserving the utility-privacy trade-off without tedious hyperparameter tuning.

5.5 White-Box Representation Auditing and Relearning Dynamics

To verify that QF-Entanglement achieves intrinsic representation-level unlearning rather than superficial output-logit masking, we conduct white-box audits on CIFAR-100 (Class 0): binary linear probing on the frozen backbone, relearning dynamics upon re-introducing $\mathcal{D}_f$, and layer-wise CKA alongside latent geometry analysis.

Representation Invariance via Linear Probing. We freeze the backbone feature extractor $\Phi_{\tilde{\theta}}$ and train linear and non-linear probes over 20 epochs. As reported in Table 9, QF-Entanglement yields probe accuracies of 89.40% (linear) and 89.60% (non-linear), closely following the Retrain reference baseline (89.20% and 88.30%, with deviations of $+0.20\%$ and $+1.30\%$). These results demonstrate that our method suppresses class-specific feature retention while preserving generic visual representations.

Representation Alignment and Latent Geometry. As explained in Appendix B.6, layer-wise CKA with Retrain remains high across Layers 1–3 (≥ 0.9837) and selectively declines at Layer 4 (0.8308), confirming that representation adjustments concentrate within high-level semantic boundaries. Concurrently, latent geometry analysis shows that the effective rank is largely preserved ($21.91 \rightarrow 20.38$) despite moderate reductions in point dispersion ($6.90 \rightarrow 5.52$) and total variance ($26.25 \rightarrow 18.53$), verifying non-degeneracy without manifold collapse.

6 Conclusion

In this work, we present QF-Entanglement, reframing machine unlearning not as destructive parameter erasure, but as representation-level statistical manifold realignment. Specifically, our framework isolates forget-specific variations into a compact subspace via differential PCA while simultaneously preserving shared foundational representations in the orthogonal complement. Within the active subspace, joint first-order centroid anchoring and second-order density-matrix linear entropy matching smoothly disperse forgotten representations across adjacent retain manifolds, eliminating the statistical anomalies that trigger the Streisand effect. Comprehensive evaluations and white-box audits demonstrate that QF-Entanglement consistently suppresses membership inference attacks to the theoretical random-guess baseline ($\approx 50\%$) while preserving model utility and representation plasticity, providing a principled foundation for privacy-compliant deep learning.

AI USE STATEMENT

Generative AI tools were used to assist with conversion of the manuscript to the ICLR 2027 LaTeX format. The authors reviewed the resulting source and take responsibility for the final content of the work, including all text, claims, and artifacts.

REPRODUCIBILITY STATEMENT

Codes are provided in an anonymous website to guarantee reproducibility. The algorithm and its hyperparameters are specified in the experiment part; assumptions and complete proofs are provided in the appendix. Experimental settings, additional datasets and ablations are also documented in the appendix and supplementary material.

REFERENCES

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes, November 2018. URL http://arxiv.org/abs/1610.01644. arXiv:1610.01644 [stat.ML].

Lucas Bourtoule, Varun Chandrasekaran, Christopher Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. Machine Unlearning, December 2019. URL http://arxiv.org/abs/1912.03817. arXiv:1912.03817 [cs.CR] version: 1.

Yinzhi Cao and Junfeng Yang. Towards Making Systems Forget with Machine Unlearning. In *2015 IEEE Symposium on Security and Privacy*, pp. 463–480, May 2015. doi: 10.1109/SP.2015.35. URL https://ieeexplore.ieee.org/document/7163042. ISSN: 2375-1207.

Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramer. Membership Inference Attacks From First Principles, April 2022a. URL http://arxiv.org/abs/2112.03570. arXiv:2112.03570 [cs.CR].

Nicholas Carlini, Matthew Jagielski, Chiyuan Zhang, Nicolas Papernot, Andreas Terzis, and Florian Tramer. The Privacy Onion Effect: Memorization is Relative, June 2022b. URL http://arxiv.org/abs/2206.10469. arXiv:2206.10469 [cs.LG].

Sungmin Cha, Sungjun Cho, Dasol Hwang, Honglak Lee, Taesup Moon, and Moontae Lee. Learning to unlearn: Instance-wise unlearning for pre-trained classifiers. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 11186–11194, 2024.

Min Chen, Weizhuo Gao, Gaoyang Liu, Kai Peng, and Chen Wang. Boundary Unlearning, March 2023. URL http://arxiv.org/abs/2303.11570. arXiv:2303.11570 [cs.CV].

Jingpu Cheng, Ping Liu, Qianxiao Li, and Chi Zhang. Machine Unlearning under Retain-Forget Entanglement, March 2026. URL http://arxiv.org/abs/2603.26569. arXiv:2603.26569 [cs.LG].

Patryk Chrabaszcz, Ilya Loshchilov, and Frank Hutter. A Downsampled Variant of ImageNet as an Alternative to the CIFAR datasets, August 2017. URL http://arxiv.org/abs/1707.08819. arXiv:1707.08819 [cs.CV].

Ali Ebrahimpour-Boroojeny, Yian Wang, and Hari Sundaram. On the Necessity of Output Distribution Reweighting for Effective Class Unlearning, July 2026. URL http://arxiv.org/abs/2506.20893. arXiv:2506.20893 [cs.LG] version: 5.

Chongyu Fan, Jiancheng Liu, Yihua Zhang, Eric Wong, Dennis Wei, and Sijia Liu. SalUn: Empowering Machine Unlearning via Gradient-based Weight Saliency in Both Image Classification and Generation, April 2024. URL http://arxiv.org/abs/2310.12508. arXiv:2310.12508 [cs.LG].

Jack Foster, Stefan Schoepf, and Alexandra Brintrup. Fast Machine Unlearning Without Retraining Through Selective Synaptic Dampening, December 2023. URL http://arxiv.org/abs/2308.07707. arXiv:2308.07707 [cs.LG].

Karan Ganju, Qi Wang, Wei Yang, Carl A. Gunter, and Nikita Borisov. Property Inference Attacks on Fully Connected Neural Networks using Permutation Invariant Representations. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, CCS '18, pp. 619–633, New York, NY, USA, October 2018. Association for Computing Machinery. ISBN 978-1-4503-5693-0. doi: 10.1145/3243734.3243834. URL https://dl.acm.org/doi/10.1145/3243734.3243834.

Quentin Garrido, Randall Balestriero, Laurent Najman, and Yann Lecun. RankMe: Assessing the downstream performance of pretrained self-supervised representations by their rank, June 2023. URL http://arxiv.org/abs/2210.02885. arXiv:2210.02885 [cs.LG].

Aditya Golatkar, Alessandro Achille, and Stefano Soatto. Eternal Sunshine of the Spotless Net: Selective Forgetting in Deep Networks, March 2020. URL http://arxiv.org/abs/1911.04933. arXiv:1911.04933 [cs.LG].

G. H. Golub and C. Reinsch. Singular value decomposition and least squares solutions. *Numerische Mathematik*, 14(5):403–420, April 1970. ISSN 0945-3245. doi: 10.1007/BF02163027. URL https://doi.org/10.1007/BF02163027.

Chuan Guo, Tom Goldstein, Awni Hannun, and Laurens van der Maaten. Certified Data Removal from Machine Learning Models, November 2023. URL http://arxiv.org/abs/1911.03030. arXiv:1911.03030 [cs.LG].

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition, December 2015. URL http://arxiv.org/abs/1512.03385. arXiv:1512.03385 [cs.CV].

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium, January 2018. URL http://arxiv.org/abs/1706.08500. arXiv:1706.08500 [cs.LG].

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the Knowledge in a Neural Network, March 2015. URL http://arxiv.org/abs/1503.02531. arXiv:1503.02531 [stat.ML].

Hiqmet Kamberaj. Information Theory and Statistical Mechanics. In Hiqmet Kamberaj (ed.), *Molecular Dynamics Simulations in Statistical Physics: Theory and Applications*, pp. 343–369. Springer International Publishing, Cham, 2020. ISBN 978-3-030-35702-3. doi: 10.1007/978-3-030-35702-3_9. URL https://doi.org/10.1007/978-3-030-35702-3_9.

Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear Convergence of Gradient and Proximal-Gradient Methods Under the Polyak-Łojasiewicz Condition, September 2020. URL http://arxiv.org/abs/1608.04636. arXiv:1608.04636 [cs.LG].

Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization, January 2017. URL http://arxiv.org/abs/1412.6980. arXiv:1412.6980 [cs.LG].

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13): 3521–3526, March 2017. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1611835114. URL http://arxiv.org/abs/1612.00796. arXiv:1612.00796 [cs.LG].

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of Neural Network Representations Revisited, July 2019. URL http://arxiv.org/abs/1905.00414. arXiv:1905.00414 [cs.LG].

A. Krizhevsky. Learning multiple layers of features from tiny images. 2009. URL https://www.semanticscholar.org/paper/Learning-Multiple-Layers-of-Features-from-Tiny-Krizhevsky/5d90f06bb70a0a3dced62413346235c02b1aa086.

Meghdad Kurmanji, Peter Triantafillou, Jamie Hayes, and Eleni Triantafillou. Towards Unbounded Machine Unlearning, October 2023. URL http://arxiv.org/abs/2302.09880. arXiv:2302.09880 [cs.LG].

Ya Le and Xuan Yang. Tiny ImageNet Visual Recognition Challenge. 2015. URL https://www.semanticscholar.org/paper/Tiny-ImageNet-Visual-Recognition-Challenge-Le-Yang/384ce792cf2b2afbe001f2168bfe7d5e7804c736.

Seth Neel, Aaron Roth, and Saeed Sharifi-Malvajerdi. Descent-to-Delete: Gradient-Based Methods for Machine Unlearning, July 2020. URL http://arxiv.org/abs/2007.02923. arXiv:2007.02923 [stat.ML].

Yu. Nesterov. Gradient methods for minimizing composite functions. *Mathematical Programming*, 140(1):125–161, August 2013. ISSN 1436-4646. doi: 10.1007/s10107-012-0629-5. URL https://doi.org/10.1007/s10107-012-0629-5.

Yuval Netzer, Tao Wang, A. Coates, A. Bissacco, Bo Wu, and A. Ng. Reading Digits in Natural Images with Unsupervised Feature Learning. 2011. URL https://www.semanticscholar.org/paper/Reading-Digits-in-Natural-Images-with-Unsupervised-Netzer-Wang/02227c94dd41fe0b439e050d377b0beb5d427cda.

Vardan Papyan, X. Y. Han, and David L. Donoho. Prevalence of Neural Collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40): 24652–24663, October 2020. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.2015509117. URL http://arxiv.org/abs/2008.08186. arXiv:2008.08186 [cs.LG].

Gaurav Patel and Qiang Qiu. Learning to unlearn while retaining: Combating gradient conflicts in machine unlearning. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4211–4221. IEEE, 2025.

Mohammad Pezeshki, Sékou-Oumar Kaba, Yoshua Bengio, Aaron Courville, Doina Precup, and Guillaume Lajoie. Gradient Starvation: A Learning Proclivity in Neural Networks, November 2021. URL http://arxiv.org/abs/2011.09468. arXiv:2011.09468 [cs.LG].

William F. Shen, Xinchi Qiu, Meghdad Kurmanji, Alex Iacob, Lorenzo Sani, Yihong Chen, Nicola Cancedda, and Nicholas D. Lane. LLM Unlearning via Neural Activation Redirection, October 2025. URL http://arxiv.org/abs/2502.07218. arXiv:2502.07218 [cs.LG].

Jonathon Shlens. A Tutorial on Principal Component Analysis, April 2014. URL http://arxiv.org/abs/1404.1100. arXiv:1404.1100 [cs.LG].

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership Inference Attacks against Machine Learning Models, March 2017. URL http://arxiv.org/abs/1610.05820. arXiv:1610.05820 [cs.CR].

Anvith Thudi, Gabriel Deza, Varun Chandrasekaran, and Nicolas Papernot. Unrolling SGD: Understanding Factors Influencing Machine Unlearning, March 2022. URL http://arxiv.org/abs/2109.13398. arXiv:2109.13398 [cs.LG].

Eduard Fosch Villaronga, Peter Kieseberg, and Tiffany Li. Humans forget, machines remember: Artificial intelligence and the Right to Be Forgotten. *Computer Law & Security Review*, 34(2): 304–313, April 2018. ISSN 2212-473X. doi: 10.1016/j.clsr.2017.08.007. URL https://www.sciencedirect.com/science/article/pii/S0267364917302091.

Paul Voigt and Axel Von Dem Bussche. *The EU General Data Protection Regulation (GDPR): A Practical Guide*. Springer Nature Switzerland, Cham, 2024. ISBN 978-3-031-62327-1 978-3-031-62328-8. doi: 10.1007/978-3-031-62328-8. URL https://link.springer.com/10.1007/978-3-031-62328-8.

John Von Neumann. *Mathematical foundations of quantum mechanics*. Princeton University Press, Princeton, new edition edition, 2018. ISBN 978-0-691-17856-1 978-1-4008-8992-1 978-0-691-17857-8.

Abudukelimu Wuerkaixi, Qizhou Wang, Sen Cui, Wutong Xu, Bo Han, Gang Niu, Masashi Sugiyama, and Changshui Zhang. Adaptive localization of knowledge negation for continual llm unlearning. In *Forty-second International Conference on Machine Learning*, 2025.

Heng Xu, Tianqing Zhu, Lefeng Zhang, Wanlei Zhou, and Philip S. Yu. Machine Unlearning: A Survey, June 2023. URL http://arxiv.org/abs/2306.03558. arXiv:2306.03558 [cs.CR].

Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy Risk in Machine Learning: Analyzing the Connection to Overfitting, May 2018. URL http://arxiv.org/abs/1709.01604. arXiv:1709.01604 [cs.CR].

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization, February 2017. URL http://arxiv.org/abs/1611.03530. arXiv:1611.03530 [cs.LG].

A Detailed Proofs

A.1 Proof of Proposition 1

The change in logits for any input x is given by $z_\theta - z_{\theta_0} = W_\theta \Phi_\theta - W_{\theta_0}\Phi_{\theta_0} = W_\theta(\Phi_\theta - \Phi_{\theta_0}) + (W_\theta - W_{\theta_0})\Phi_{\theta_0}$. By the triangle inequality and sub-multiplicativity of the spectral norm:

$$\|z_\theta - z_{\theta_0}\|_2 \le \|W_\theta\|_2 \|\Phi_\theta - \Phi_{\theta_0}\|_2 + \|W_\theta - W_{\theta_0}\|_2 \|\Phi_{\theta_0}\|_2 \tag{11}$$

Given the L_ℓ-Lipschitz continuity of the cross-entropy loss w.r.t logits:

$$\ell(z_\theta, y) - \ell(z_{\theta_0}, y) \le L_\ell \|z_\theta - z_{\theta_0}\|_2 \tag{12}$$

For the feature perturbation, we decompose it into the subspace and its orthogonal complement:

$$\Phi_\theta - \Phi_{\theta_0} = PP^\top(\Phi_\theta - \Phi_{\theta_0}) + (I - PP^\top)(\Phi_\theta - \Phi_{\theta_0}) \tag{13}$$

Taking the L_2 norm and applying the triangle inequality:

$$\|\Phi_\theta - \Phi_{\theta_0}\|_2 \le \|P^\top(\Phi_\theta - \Phi_{\theta_0})\|_2 + \|(I - PP^\top)(\Phi_\theta - \Phi_{\theta_0})\|_2 \tag{14}$$

Taking the expectation over $\mathcal{D}_{\text{retain}}$, applying Jensen's inequality to the null-space term ($\mathbb{E}[\|X\|_2] \le \sqrt{\mathbb{E}[\|X\|_2^2]} \le \delta$), and substituting the assumptions yields the final bound.

Remark A.1 (Scope of Utility Preservation). Proposition 1 is a conditional perturbation bound: it does not claim that the retain representation is strictly invariant, but quantifies how utility degradation scales with active-subspace perturbation, bounded null-space drift, and classifier-head change. The $\mathcal{L}_{\text{preserve}}$ objective is a soft invariance constraint; therefore, the theory quantifies bounded deviation rather than absolute preservation of null-space representations.

A.2 Proof of Proposition 2

Let $\mathcal{M}(\theta) \in \mathbb{R}^{k+1}$ denote the vector of empirical forget-subspace moments and M_{target} be the cached retain target moments, with feature Jacobian $J(\theta) = \nabla_\theta \mathcal{M}(\theta)$. The gradient is:

$$\nabla_\theta \mathcal{J}_{\text{forget}}(\theta) = J(\theta)^\top(\mathcal{M}(\theta) - M_{\text{target}}) \tag{15}$$

Analyzing the squared L_2 norm:

$$\|\nabla_\theta \mathcal{J}_{\text{forget}}(\theta)\|_2^2 = (\mathcal{M}(\theta) - M_{\text{target}})^T J(\theta) J(\theta)^T (\mathcal{M}(\theta) - M_{\text{target}}) \tag{16}$$

Assuming the feature Jacobian satisfies local full-rank conditioning such that the minimum eigenvalue of $J(\theta)J(\theta)^T$ is bounded below by $\mu^2 > 0$, by the Rayleigh quotient property:

$$\|\nabla_\theta \mathcal{J}_{\text{forget}}(\theta)\|_2^2 \ge \mu^2 \|\mathcal{M}(\theta) - M_{\text{target}}\|_2^2 = 2\mu^2 \mathcal{J}_{\text{forget}}(\theta) \tag{17}$$

This establishes the Polyak-Lojasiewicz (PL) condition (Nesterov, 2013):

$$\frac{1}{2}\|\nabla_\theta \mathcal{J}_{\text{forget}}(\theta)\|_2^2 \ge \mu^2 \mathcal{J}_{\text{forget}}(\theta) \tag{18}$$

Given the PL condition and the L_J-smoothness of the objective in region Ω, Gradient Descent with step size $\eta \leq 1/L_J$ guarantees linear convergence to the optimal value (Karimi et al., 2020) $\mathcal{J}^\star$:

$$\mathcal{J}_{\text{forget}}(\theta_t) - \mathcal{J}^\star \leq \left(1 - \eta\mu^2\right)^t \left(\mathcal{J}_{\text{forget}}(\theta_0) - \mathcal{J}^\star\right) \tag{19}$$

Assuming moment realizability on the retain manifold such that $\mathcal{J}^\star = 0$, we obtain the final bound:

$$\mathcal{J}_{\text{forget}}(\theta_t) \leq \left(1 - \eta\mu^2\right)^t \mathcal{J}_{\text{forget}}(\theta_0) \tag{20}$$

This completes the proof.

Remark A.2 (Optimization-Side Result). Proposition 2 is an optimization result conditional on local Jacobian conditioning and moment realizability. It characterizes the local convergence of the moment-matching subproblem and is *not* required for the validity of the distributional Wasserstein bound (Theorem 1) once the achieved optimization gaps ϵ_a, ϵ_e are empirically specified. The theoretical analysis adopts the squared entropy deviation for analytical tractability. Implementation details follow the benchmark-specific configurations reported in the released code.

A.3 Proof of Theorem 1

Geometric Interpretation of Linear Entropy. The density-matrix linear entropy for a normalized covariance matrix $\rho = \Sigma/\text{Tr}(\Sigma)$ is defined as $S_L(\rho) = 1 - \text{Tr}(\rho^2)$. The effective rank of a covariance matrix Σ, which quantifies the continuous dimensionality of the representation manifold, is defined as $R_{\text{eff}}(\Sigma) = (\text{Tr}\Sigma)^2/\text{Tr}(\Sigma^2)$. By substituting the definition of ρ, we establish an exact algebraic equivalence between linear entropy and effective rank:

$$R_{\text{eff}}(\Sigma) = \frac{(\text{Tr}\Sigma)^2}{\text{Tr}(\Sigma^2)} = \frac{1}{\text{Tr}(\rho^2)} = \frac{1}{1 - S_L(\rho)} \tag{21}$$

Proof of Theorem 1. Given the linear entropy matching objective $\mathcal{L}_{\text{entropy}}$, assume the optimization gap is bounded by ϵ_e, yielding $|S_L(\rho_f) - S_L(\rho_r)| \leq \epsilon_e$. We seek to bound the discrepancy in the effective rank between the forget distribution Σ_f and the retain target distribution Σ_r.

Using the relation $R_{\text{eff}}(\Sigma) = (1 - S_L(\rho))^{-1}$, we evaluate the absolute difference:

$$\begin{aligned} |R_{\text{eff}}(\Sigma_f) - R_{\text{eff}}(\Sigma_r)| &= \left| \frac{1}{1 - S_L(\rho_f)} - \frac{1}{1 - S_L(\rho_r)} \right| \\ &= \frac{|S_L(\rho_f) - S_L(\rho_r)|}{(1 - S_L(\rho_r))(1 - S_L(\rho_f))} \\ &\leq \frac{\epsilon_e}{(1 - S_L(\rho_r))(1 - S_L(\rho_f))} \end{aligned} \tag{22}$$

This completes the proof.

Significance for Membership Inference Attacks (MIAs). Standard parameter dampening or gradient ascent techniques force target features toward zero variance, causing $\text{Tr}(\Sigma_f^2) \to \text{Tr}(\Sigma_f)^2$ and collapsing the effective rank $R_{\text{eff}} \to 1$. In a multi-dimensional feature space, such degenerate manifolds produce likelihood singularities (near-infinite log-densities on the target subspace or near-zero densities elsewhere), which LiRA and other variance-sensitive MIAs easily detect as the Streisand effect. By proving that $|R_{\text{eff}}(\Sigma_f) - R_{\text{eff}}(\Sigma_r)|$ is tightly bounded, Theorem 1 guarantees that the unlearned distribution preserves a macroscopic volume proportional to natural retain classes. This mathematically precludes point-collapse and ensures that the Gaussian log-likelihoods of the unlearned samples remain statistically structurally comparable to genuine unseen data, thereby neutralizing variance-based detection.

A.4 Proof of Corollary 1

The Gaussian log-likelihood for a sample x is given by $\ln p(x) \propto -\frac{1}{2}(x - \mu)^T\Sigma^{-1}(x - \mu) - \frac{1}{2}\ln|\Sigma|$. When the effective rank R_{eff} collapses to 1, the condition number of Σ diverges, causing the precision matrix Σ^{-1} and $\ln|\Sigma|$ to explode. This singularity forces the likelihood ratio (e.g.,

in LiRA) to exhibit highly anomalous values. By Theorem 1, matching linear entropy bounds the effective rank discrepancy, strictly bounding Σ_f away from singularity (assuming the retain target Σ_r is non-degenerate). Consequently, the expected log-likelihoods are structurally stabilized, and the likelihood ratio avoids the extreme anomalies characteristic of the Streisand effect.

A.5 Target Interpretation

Remark A.3 (Conditional Target Interpretation). The theoretical analysis (Theorem 1) is conditioned on a fixed target retain mode c^*. The global anchor assignment mechanism can therefore be viewed as instantiating this conditional guarantee for the selected retain manifold. Furthermore, restricting target assignment to a Monte Carlo subsampled candidate pool $\mathcal{C}_{\text{cand}} \subset \mathcal{C}_r$ preserves this theoretical guarantee for the selected mode while mitigating optimization variance.

Lemma A.1 (Conditional-to-Empirical manifold realignment). Suppose for each batch b, the conditional Wasserstein bound satisfies $W_2(P_f^{(b)}, P_{r,c^*}) \leq \epsilon_b$. Then the batch-conditioned manifold realignment error's empirical average satisfies:

$$\frac{1}{B}\sum_{b=1}^{B} W_2(P_f^{(b)}, P_{r,c^*}) \leq \frac{1}{B}\sum_{b=1}^{B} \epsilon_b. \tag{23}$$

This aligns the mini-batch optimization theoretical guarantee with the overall global anchor alignment executed in our experiments.

B Additional Algorithm and Experiments

B.1 Training Pipeline and Pseudocode

To provide a self-contained operational reference, Algorithm 1 outlines the complete optimization workflow of QF-Entanglement. The procedure comprises two sequential phases: (i) a lightweight initialization stage that extracts the initial differential PCA projection matrix P with low computational overhead and caches empirical retain-manifold statistics, followed by (ii) a dual-moment manifold realignment optimization loop driven by the unified objective $\mathcal{L}_{total}(\theta)$.

B.2 Computational Complexity and Sequential Unlearning Scalability

To evaluate scalability across large label spaces (e.g., ImageNet-1k/21k) and continuous unlearning, we analyze the computational and storage footprints across both phases.

Lightweight Subspace Extraction (SVD): For an unlearning request with cohort size $M = |\mathcal{D}_f|$ and feature dimension D, computing the thin SVD on the differential residual $\Delta F \in \mathbb{R}^{M\times D}$ requires $\mathcal{O}(min(MD^2, M^2D))$ operations with low computational overhead. This cost depends solely on the local request size M (e.g., $M \approx 1300$ for ImageNet-1k) and feature dimension D ($D = 512$ for ResNet-18), decoupling completely from the overall class vocabulary C. The subspace rank k is dynamically determined via the cumulative explained variance threshold τ (Eq. 3). To establish an empirical upper bound on computational overhead, we benchmark Phase 1 on an NVIDIA GeForce RTX 5070 Ti Laptop GPU over 100 trials using synthetic isotropic Gaussian representations ($\tau = 0.95$, Table 6). Because isotropic features lack semantic clustering and possess a flat eigenspectrum, they yield substantially larger dynamic ranks k (262–481) than structured natural manifolds. Even under this conservative setting, end-to-end extraction executes in 20–23 ms for ResNet-18 (42 ms for ViT-Base) with peak dynamic memory below 38 MB.

Class Statistics Caching and Nearest-Neighbor Assignment: Pre-computing retain subspace centroids $\{\mu_c^{\text{sub}}\}_{c\in\mathcal{C}_r} \subset \mathbb{R}^k$ and scalar entropies $\{S_c^{\text{sub}}\}_{c\in\mathcal{C}_r} \subset \mathbb{R}$ requires a single forward inference pass over $\mathcal{D}_r$. The storage footprint for caching C classes is $C \times (k+1) \times 4$ bytes. On natural manifolds where the dynamic rank satisfies $k \ll D$, this cache footprint is negligible (e.g., < 1 MB across typical benchmarks; even under the conservative isotropic rank $k = 364$ on ImageNet-1k, storage is under 1.9 MB). During Phase 2, identifying the target global class $c^* = \arg\min_c \|\overline{f}_{\text{disc}} - \mu_c^{\text{sub}}\|_2$ requires only $\mathcal{O}(Ck)$ FLOPs in total, introducing sub-millisecond latency.

Algorithm 1 QF-ENTANGLEMENT with Variance-Gated Tracking

Require: Forget set $\mathcal{D}_f$, retain set $\mathcal{D}_r$, learning rate η, collapse threshold $\gamma = 0.5$, candidate anchor pool size K_{cand} (default ≥ 5). Initialize $\theta = \theta_0$.

Phase 1: Subspace & Empirical Anchor Initialization
Extract features $F_f \in \mathbb{R}^{M \times D}$ via θ_0; compute initial projection $P^{(0)}$ and leading singular value $\sigma_1^{(0)}$ via thin SVD on $\Delta F = F_f - \mathbf{1}_M \mu_r^\top$.
Monte Carlo sample a candidate subset of retain classes $\mathcal{C}_{\text{cand}} \subset \mathcal{C}_r$ ($|\mathcal{C}_{\text{cand}}| = K_{\text{cand}}$); cache projected centroids $\{\mu_c^{\text{sub}}\}$ and entropies $\{S_c^{\text{sub}}\}_{c \in \mathcal{C}_{\text{cand}}}$.
Phase 2: Dual-Moment manifold realignment Optimization
for epoch $e = 1$ to E **do**
  for each mini-batch **do**
    Sample mini-batches $(X_f, y_f) \sim \mathcal{D}_f$ and $(X_r, y_r) \sim \mathcal{D}_r$;
    Use the nearest global retain disguise $c^* = \arg\min_{c \in \mathcal{C}_{\text{cand}}} \|\bar{f}_{\text{disc}} - \mu_c^{\text{sub}}\|_2$;
    Compute $\mathcal{L}_{\text{total}}(\theta) = \mathcal{L}_{\text{CE}}(\theta) + \lambda_q \mathcal{L}_{\text{entropy}}(\theta) + \beta \mathcal{L}_{\text{anchor}}(\theta) + \mathcal{L}_{\text{preserve}}(\theta)$;
    Update model parameters: $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}_{\text{total}}(\theta)$.
  end for
  if subspace updates active **then**
    Re-extract $\Delta F^{(e)}$ via current checkpoint θ to obtain updated $\sigma_1^{(e)}$;
    if $\sigma_1^{(e)} < \gamma \sigma_1^{(0)}$ **then**
      Freeze subspace updates permanently (set $P \leftarrow P^{(e)}$);
    else
      Refresh projection matrix $P^{(e)}$ and update cache on resampled $\mathcal{C}_{\text{cand}}$.
    end if
  end if
end for
return Unlearned model parameters $\tilde{\theta}$

Table 6: Empirical computational and memory overhead of Phase 1 differential subspace extraction ($\tau = 0.95$, averaged over 100 runs on synthetic isotropic representations). Both SVD and Full Phase 1 times are reported in ms, and peak memory is in MB.

Dataset	M	D	k	SVD Time	Full Phase 1	Peak Mem
CIFAR-100	500	512	262	$19.57_{\pm 0.44}$	$20.01_{\pm 0.52}$	7.84
Tiny-ImageNet	500	512	262	$19.60_{\pm 0.49}$	$20.08_{\pm 0.56}$	7.84
ImageNet-1k	1300	512	364	$20.12_{\pm 0.47}$	$20.54_{\pm 0.53}$	12.14
CIFAR-10	5000	512	425	$22.83_{\pm 0.41}$	$23.26_{\pm 0.47}$	37.91
ImageNet-1k (ViT-Base)	1300	768	481	$41.57_{\pm 1.02}$	$41.96_{\pm 1.06}$	19.47

Scalability under Sequential Unlearning: Handling sequential requests $\{\mathcal{D}_f^{(1)}, \ldots, \mathcal{D}_f^{(T)}\}$ requires neither retraining nor global subspace re-computation:

- *Independent Local Decomposition:* Each task computes an instantaneous local projection $P^{(t)}$ from current checkpoint representations in constant $\mathcal{O}(MD^2)$ time (≈ 20 ms).
- *Incremental Subspace Augmentation:* When an aggregated null space is required, the basis is updated via Gram-Schmidt orthogonalization: $P^{(t)} \leftarrow \text{orth}([P^{(t-1)}, V_f^{(t)}])$, maintaining representation invariance while $\sum_t k_t < D$.

Therefore, QF-ENTANGLEMENT scales efficiently to massive label spaces and continuous forgetting regimes.

B.3 UNLEARNING PERFORMANCE AND MULTISEED STABILITY ON CIFAR-100

To evaluate consistency, we conduct five independent trials on CIFAR-100 under Class 0 unlearning (Table 7). As further visualized in Figure 1, the error-bar plots illustrate the distributional variance across seeds, highlighting how closely our method aligns with the ideal Retrain bounds. Stan-

Table 7: Multiseed Performance on CIFAR-100

Method	Retain Acc	Forget Acc	Retain Test	MIA
Retrain	99.97	0.00	77.24	50.00
Delete	$85.59_{\pm 0.17}$	$2.70_{\pm 0.59}$	$61.65_{\pm 0.11}$	$47.00_{\pm 0.00}$
Fine-tune	$96.32_{\pm 0.61}$	$0.12_{\pm 0.16}$	$70.26_{\pm 0.65}$	$52.60_{\pm 3.50}$
GDR	$94.30_{\pm 0.17}$	$1.20_{\pm 0.28}$	$68.37_{\pm 0.24}$	$43.40_{\pm 0.80}$
Gradient Ascent	$96.85_{\pm 0.20}$	$0.00_{\pm 0.00}$	$70.13_{\pm 0.19}$	$46.00_{\pm 0.00}$
Munba	$86.17_{\pm 2.33}$	$34.04_{\pm 15.71}$	$62.26_{\pm 1.32}$	$49.00_{\pm 1.26}$
SalUn	$97.59_{\pm 0.10}$	$0.00_{\pm 0.00}$	$70.84_{\pm 0.11}$	$46.00_{\pm 0.00}$
SSD	$94.85_{\pm 8.70}$	$0.00_{\pm 0.00}$	$71.95_{\pm 7.51}$	$47.40_{\pm 1.02}$
Ours	$\mathbf{99.97}_{\pm 0.01}$	$\mathbf{0.00}_{\pm 0.00}$	$\mathbf{76.54}_{\pm 0.05}$	$\mathbf{51.00}_{\pm 0.89}$

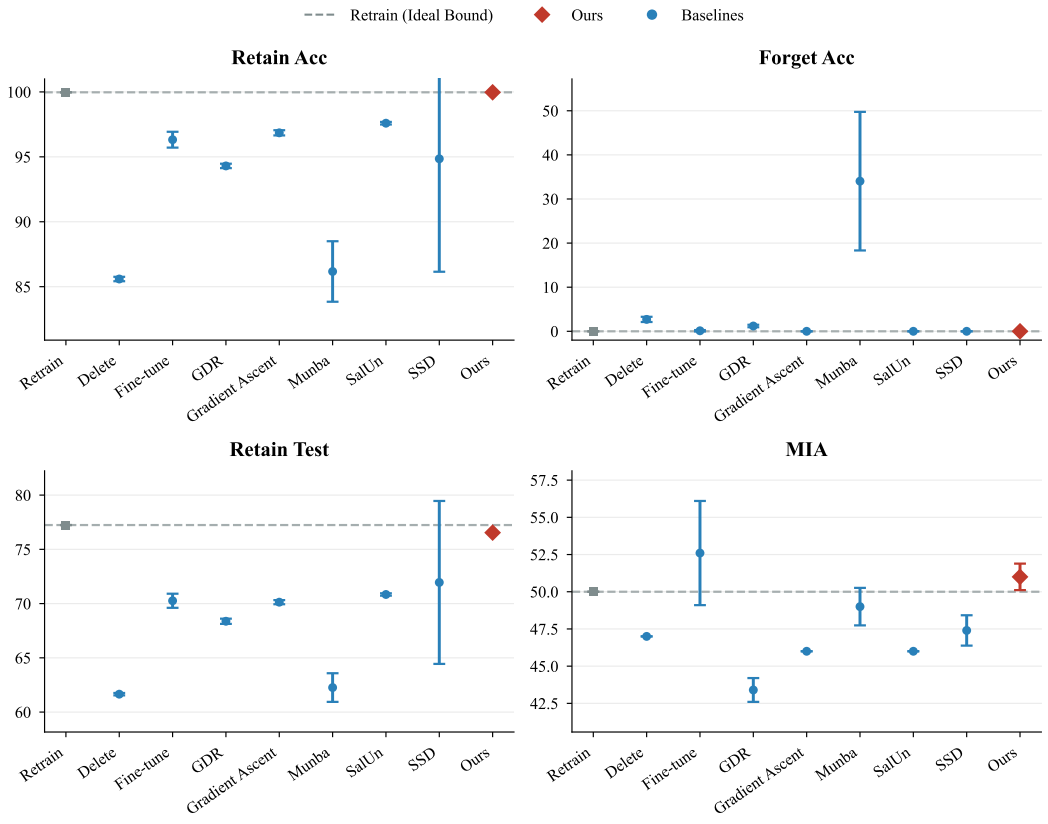

Figure 1: Performance comparison on CIFAR-100.

dard baselines exhibit distinct failure modes: Delete incurs a severe utility penalty ($61.65 \pm 0.11\%$ retain test), while FT exhibits notable MIA auditing variance ($52.60 \pm 3.50\%$). Boundary- and regularization-based methods (Munba) leave substantial forget-set residuals. Optimization and dampening baselines (GA, SalUn, GDR, SSD) achieve near-zero forget accuracy but introduce statistical anomalies that shift MIA rates away from the random-guess baseline or exhibit high variance. In contrast, our method achieves complete target erasure (0.00 forget) and obviously highest retain utility ($76.54 \pm 0.05\%$). In particular, our approach maintains an MIA rate close to the ideal random-guess threshold ($51.00 \pm 0.89\%$) with minimal variance, confirming the empirical stability of our subspace-isolated manifold realignment framework.

Table 8: Sensitivity of QF-Entanglement to the anchoring coefficient β on CIFAR-100.

Method	Retain Acc	Forget Acc	Retain Test	MIA_true
Ours ($\beta = 0.01$)	99.97	0.40	76.56	52.00
Ours ($\beta = 0.03$)	99.96	0.60	76.53	49.00
Ours ($\beta = 0.05$)	99.96	0.00	76.56	50.00
Ours ($\beta = 0.1$)	99.96	0.00	76.64	49.00
Ours ($\beta = 20$)	99.93	0.00	75.87	48.00
Ours ($\beta = 35$)	99.93	0.00	75.94	48.00
Ours ($\beta = 50$)	99.92	0.00	76.03	48.00
Ours ($\beta = 100$)	99.93	0.00	75.73	48.00
Ours ($\beta = 200$)	99.92	0.00	75.72	47.00
Ours ($\beta = 300$)	99.89	0.00	75.76	47.00

B.4 Extended Sensitivity Analysis

We examine the sensitivity of our method to the anchoring coefficient β on CIFAR-100. Table 8 reports the results across a wide range of $\beta \in [0.01, 300]$. Across these values, the retain accuracy remains consistently high ($\geq 99.89\%$), and the retain test accuracy shows minimal variation, fluctuating within a narrow range between 75.72% and 76.64%. Regarding unlearning efficacy, the forget accuracy is reduced to 0.00% for all $\beta \geq 0.05$, with only slight residual accuracy observed at smaller values (0.40% and 0.60% for $\beta = 0.01$ and 0.03, respectively). Concurrently, the true MIA score stays close to the 50% random-guess baseline across all settings (47.00%–52.00%). These observations suggest that our method is relatively insensitive to the choice of β across several orders of magnitude, provided $\beta \geq 0.05$, and does not require fine-grained tuning.

B.5 Extended White-Box Auditing: Probing and Relearning Dynamics

This section provides the full experimental setup, numerical logs, and mechanistic discussions for the Binary Probing and Relearning evaluations conducted on CIFAR-100 (Class 0 unlearning).

B.5.1 Binary Linear Probing Protocol and Results

Protocol. We freeze the backbone feature extractor $\Phi_{\tilde{\theta}}$ of each unlearned model and train classification heads (Alain & Bengio, 2018) on top of the frozen representations for 20 epochs using SGD. Table 9 reports the mean probe test accuracy over the final 5 epochs.

Table 9: Mean test accuracy of binary linear and non-linear probing on frozen representations over the final 5 epochs (trained for 20 epochs) on CIFAR-100. Δ denotes the performance deviation relative to the gold standard (RETRAIN); a smaller $|\Delta|$ indicates that the unlearned representation aligns more closely with retraining from scratch.

Method	Linear Probing		Non-linear Probing	
	Acc. (%)	Δ (%)	Acc. (%)	Δ (%)
RETRAIN (Gold Standard)	89.20	0.00	88.30	0.00
QF-ENTANGLEMENT (Ours)	**89.40**	**+0.20**	**89.60**	**+1.30**
RF-Entanglement	89.40	+0.20	90.40	+2.10
SSD	88.10	-1.10	86.20	-2.10
SCRUB	91.20	+2.00	91.60	+3.30
MUNBA	91.80	+2.60	90.80	+2.50
SALUN	92.00	+2.80	92.50	+4.20
Gradient Ascent (GA)	92.10	+2.90	92.90	+4.60

Empirical Analysis. Linear and non-linear probes trained on the frozen backbone of RETRAIN establish the empirical baseline (89.20% and 88.30%, respectively), reflecting the inherent separability of shared visual primitives on unseen data. Thus, the objective of representation-level unlearning is not to eliminate feature separability entirely, but to match the reference behavior of RETRAIN.

As shown in Table 9, QF-ENTANGLEMENT closely follows the reference model, yielding probe accuracies of 89.40% ($\Delta = +0.20\%$) under linear probing and 89.60% ($\Delta = +1.30\%$) under non-linear probing, achieving the smallest overall deviation $|\Delta|$. Although RF-Entanglement matches our linear probe accuracy (89.40%), it exhibits a higher gap under non-linear probing (90.40%, $\Delta = +2.10\%$). In contrast, baselines such as SALUN (92.00% / 92.50%) and GA (92.10% / 92.90%) show consistently elevated probe accuracies across both regimes (Δ up to $+4.60\%$), indicating residual class-specific information in their feature spaces. On the other end, SSD yields probe accuracies below RETRAIN (88.10% / 86.20%, $\Delta = -1.10\%/-2.10\%$), suggesting collateral degradation of general feature representations. Together, these results demonstrate that QF-ENTANGLEMENT preserves general visual representations while substantially reducing class-specific feature retention.

B.6 REPRESENTATION ALIGNMENT AND LATENT GEOMETRY ANALYSIS

To verify that the unlearning process selectively eliminates target-class discriminative features without distorting foundational representations or causing latent manifold collapse, we conduct a two-fold representation audit: (i) layer-wise Centered Kernel Alignment (CKA) (Kornblith et al., 2019) against the retrained reference model, and (ii) a geometric dispersion analysis of the latent embedding space (Garrido et al., 2023; Papyan et al., 2020) ($d = 512$).

Table 10: Layer-wise linear CKA similarity between QF-ENTANGLEMENT and the retrained baseline (CIFAR-100), alongside geometric characteristics of the forget cohort (Class 0) embeddings.

(a) Layer-wise CKA vs. RETRAIN

Network Stage	CKA Similarity
Layer 1	0.9969
Layer 2	0.9948
Layer 3	0.9837
Layer 4	0.8308

(b) Latent Space Geometry ($d = 512$)

Metric	Original	Ours	Δ
L_2	6.9014	5.5230	-1.38
$\mathrm{Tr}(\mathbf{\Sigma})$	26.2486	18.5290	-7.72
R_{eff}	21.91	20.38	-1.53

Quantitative Findings and Structural Observations. As reported in Table 10:

- **Layer-wise Representation Fidelity:** As shown in Table 10a, the intermediate feature representations of QF-ENTANGLEMENT across Layers 1 to 3 maintain high alignment with the genuinely retrained model (CKA ≥ 0.9837). This indicates that low- and mid-level feature extractors remain structurally intact, preventing unnecessary perturbation to general visual primitives. At the final representation stage (Layer 4), the CKA score moderately declines to 0.8308. This localized divergence confirms that representation adjustments are concentrated within the high-level semantic space where class-discriminative boundaries are formed, aligning with the expected behavioral divergence of target-concept unlearning.
- **Latent Manifold Preservation and Non-Degeneracy:** Table 10b characterizes the global geometry of the forget cohort representations via mean pairwise Euclidean distance, total variance ($\mathrm{Tr}(\mathbf{\Sigma})$), and effective rank ($R_{\mathrm{eff}} = \exp(-\sum_i \tilde{\lambda}_i \ln \tilde{\lambda}_i)$, where $\tilde{\lambda}_i$ denotes normalized eigenvalues of the covariance matrix). While the unlearning objective moderately reduces point dispersion ($6.9014 \rightarrow 5.5230$) and total variance ($26.2486 \rightarrow 18.5290$) due to the attenuation of class-specific discriminative axes, the effective rank remains largely preserved ($21.91 \rightarrow 20.38$). Rather than collapsing into an anomalous low-rank subspace or a singular cluster, the unlearned samples continue to occupy a 20-dimensional orthogonal subspace supported by the preservation objective. This structural consistency ensures that the unlearned cohort retains natural geometric properties, mitigating detectable statistical artifacts in representation auditing.

B.7 EXTENDED MEMBERSHIP INFERENCE AUDITING AT LOW FALSE POSITIVE RATES

To further verify the privacy-preserving properties of QF-ENTANGLEMENT under strict auditing conditions, we evaluate the True Positive Rate (TPR) at low False Positive Rate (FPR) thresholds

Table 11: Membership Inference Attack True Positive Rate (TPR) Evaluation on CIFAR100

Method	TPR@1% FPR	TPR@0.1% FPR
Retrain	1.00	0.10
Delete	1.40	1.40
GDR	2.40	2.00
Gradient Ascent	2.00	1.80
Munba	1.20	0.60
SalUn	2.60	2.60
SCRUB	2.60	0.60
SSD	2.20	0.00
RF-Entanglement	2.80	1.00
Ours	**0.80**	**0.20**

(1% and 0.1%). Following the Likelihood Ratio Attack (LiRA) protocol, this evaluation incorporates a retrained shadow model to compute sample-specific confidence calibrations. This calibration isolates target memorization residuals from shared semantic generalizations, mitigating the distributional artifacts that can distort uncalibrated MIA metrics.

As presented in Table 11, QF-ENTANGLEMENT consistently suppresses information leakage under stringent FPR constraints. At 1% FPR, our method yields a TPR of 0.80%, marginally lower than the empirical Retrain gold standard (1.00%). When evaluated at the stricter 0.1% FPR threshold, the TPR remains bounded at 0.20%, closely aligning with the Retrain baseline (0.10%).

In contrast, several baseline methods exhibit residual structural anomalies in these low-FPR regimes. Methods such as SalUn, SCRUB, and RF-Entanglement record elevated TPRs at 1% FPR (2.60%, 2.60%, and 2.80%, respectively). SalUn maintains a 2.60% TPR even under the 0.1% FPR constraint, suggesting the persistence of class-discriminative traces. These observations indicate that while such baselines may perform adequately under median-threshold MIA evaluations, they can still expose unlearned samples to highly selective threshold attacks. The findings corroborate that our subspace-isolated manifold realignment effectively mitigates long-tail privacy risks, achieving an unlearned state statistically comparable to retraining from scratch.

B.8 Extended Evaluation Across Additional Classes and Superclasses

While the main evaluation designates fine class 0 as the default forget target, real-world unlearning scenarios may involve diverse semantic categories or coarser superclass-level concepts. To assess generalizability across varying semantic granularities, we conduct extended experiments on CIFAR-100 covering both fine-class and superclass unlearning tasks. We evaluate individual fine classes from distinct semantic domains (classes 6, 42, and 78) as well as the complete superclass 0, which comprises five constituent fine classes (classes 4, 30, 55, 72, and 95). Due to computational resource constraints, the MIA evaluation in this subsection adopts a standard z-score metric-based MIA.

Table 12: Unlearning performance on CIFAR-100 across superclass and fine class target settings.

Type	Target	Retain Acc	Forget Acc	Retain Test Acc	MIA
Superclass	0	99.74	0.00	76.13	50.60
Fine class	6	99.94	0.00	76.31	50.00
	42	99.97	0.00	76.08	50.00
	78	99.96	0.00	76.29	51.00

As summarized in Table 12, the framework exhibits consistent performance across all evaluated targets. For individual fine-class forgetting (classes 6, 42, and 78), the target forget accuracy is reduced to 0.00% while retain test accuracy remains stable around 76.08%–76.31%. Simultaneously, the corresponding MIA detection rates center tightly around the nominal 50.00%–51.00% random-guess baseline. When scaling to superclass 0, which requires the simultaneous removal of all five fine classes, the method retains 76.13% test accuracy on the remaining data with 0.00% forget accu-

racy and an MIA rate of 50.60%. These empirical results demonstrate that our representation-level unlearning maintains a steady utility-privacy trade-off under different semantic targets and forget cohort scales.