THE ART OF FORGETTING: ORTHOGONAL SUBSPACES AND LOSS FUNCTIONS FOR CLASS UNLEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Machine unlearning supports the "right to be forgotten" by removing the influence of designated classes without requiring full retraining. We introduce geometry-aware classifier heads that enforce intra-class alignment and inter-class orthogonality, embedding features as a union of one-dimensional orthogonal subspaces. Coupled with state-of-the-art unlearning methods and an error-maximizing noise scheme for data-independent updates, this structure enables selective suppression of the forgotten class while preserving classification accuracies for retained classes. To assess genuine forgetting rather than mere misclassification, we propose a spectral-angle test that certifies the removal of the forgotten subspace and complements standard metrics — unlearning/retention accuracy (UA/RA), their test-set counterparts (TUA/TRA), and a membership-inference measure (MIA). We further study loss-head pairings by contrasting cross-entropy (CE) and mean-squared error (MSE) under two operating regimes — *Quick* and *Optimum* — reflecting different compute budgets. On CIFAR-10 in a leave-one-class-out protocol (100 trials), the framework achieves near-perfect unlearning (UA $\leq$ 0.9%) with high retention (RA $\approx$ 95–96%) and consistent generalization to held-out data (low TUA, high TRA), often matching retraining baselines while reducing computational cost. These results show that enforcing subspace structure and choosing an appropriate loss yields robust and selective forgetting with strong retention and privacy. [1]

1 INTRODUCTION

Machine unlearning, the process of removing the influence of specified data from a trained model, has gained increasing importance in light of privacy regulations such as General Data Protection Regulation (GDPR) (Hoofnagle et al., 2019) and California Consumer Privacy Act (CCPA) (State of California, 2018), which grant individuals the "right to be forgotten." Assuming access to a sufficient retain dataset, unlearning strategies are conditioned on the availability of forget data. If the forget set is accessible, the process can closely approximate retraining on only the retained set through gradient updates or reweighting. In practice, however, privacy rules and storage limitations often prevent access to the forget set or even to the complete retained set, motivating stricter regimes such as zero-shot or zero-glance unlearning, which rely solely on model parameters and possibly a small subset of retained data (Chundawat et al., 2023).

Most existing unlearning methods (Izzo et al., 2021; Foster et al., 2024; Golatkar et al., 2020; Perifanis et al., 2024; Fan et al., 2023; Warnecke et al., 2021) follow a data-centric paradigm by applying gradient updates or reweighting, while only a few studies explore structural modifications to improve unlearning capabilities (Bourtoule et al., 2021). Moreover, almost all vision-based approaches employ cross-entropy (CE) loss, a common objective function in classification that penalizes incorrect predictions by comparing the predicted probability distribution with the true distribution. Studies have shown that squared error loss, also known as mean squared error (MSE) or squared error (SE), which measures the average of the squared differences between predicted and actual values, can be equally or more effective for classification (Golik et al., 2013; Hui & Belkin, 2020; Tyagi et al., 2024). Both CE and MSE converge to the true posterior under sufficient model capacity (Golik et al., 2013);

[1] The authors used large language models (LLMs) solely for editing and improving the clarity of the text. All technical content, experiments, and analyses are entirely the authors' own.

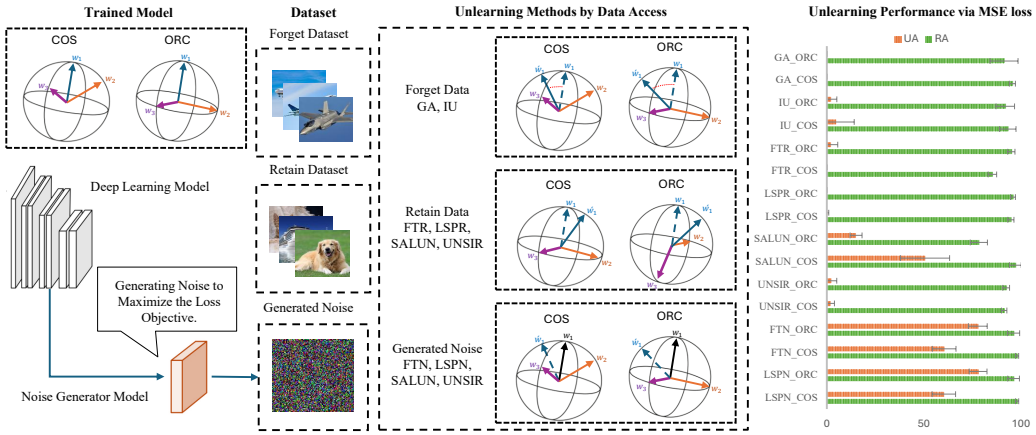


Figure 1: Overview of the proposed class unlearning framework with cosine (COS) and ortho-cosine (ORC) classifier heads under MSE loss. The left and right panels illustrate prototype representations before and after unlearning, respectively, where classes are modeled as unions of one-dimensional subspaces. Unlearning methods are grouped by their reliance on training data: (i) **forget-data based** approaches shift the target class to increase prediction loss on the forgotten set, (ii) **retain-data based** fine-tuning pushes retained classes away from the forgotten subspace, and (iii) **noise-based** strategies use synthetic inputs to overwrite the forgotten subspace. Results (right) demonstrate that our proposed classifier heads, combined with suitable loss functions, enable effective unlearning even in zero-glass settings where the forgotten data are inaccessible.

CE penalizes misclassification sharply, while MSE is bounded and often more robust to noise (Tyagi et al., 2024). Yet, the role of MSE in unlearning remains largely underexplored.

In this work, we address class unlearning by investigating whether embedding training samples such that each class occupies a distinct one-dimensional subspace enhances forgetting. This *union-of-subspaces* representation offers two advantages: (i) selective suppression of the forgotten subspace with minimal interference to retained classes, and (ii) an interpretable characterization via the dominant singular vector of each class, enabling evaluation metrics that distinguish genuine forgetting from incidental misclassification. Our approach builds on angular loss functions (Deng et al., 2019; Liu et al., 2017; Wang et al., 2018), hyper-spherical energy minimization (Liu et al., 2018), and OOD detection frameworks using subspace modeling (Zaeemzadeh et al., 2021). In addition, we adapt the error-maximization targeted noise of Tarun et al. (2023) to the zero-glance setting, using a universal noise to erase class traces rather than redirecting them into other classes. The overview of the proposed frame work illustrated in Figure 1.

Our contributions are:

- We propose enforcing a union-of-one-dimensional subspaces structure on class features, enabling precise suppression of forget classes during unlearning.
- We introduce the use of each class's leading singular vector as a robust signature for identifying and suppressing forgotten samples, providing a principled metric for true forgetting.
- We provide the first systematic comparison of CE and MSE losses across state-of-the-art unlearning methods under various classification heads.
- We validate our approach on benchmark datasets under data-available and zero-glance scenarios, demonstrating improved effectiveness, robustness, and efficiency.

2 RELATED WORK

2.1 CLASS UNLEARNING IN CLASSIFICATION

Class unlearning removes the influence of all samples from one or more target classes while preserving performance on the retained classes. Approaches differ in whether the forget set $\mathcal{D}_{\mathrm{FG}}$ is accessible. Retraining-based methods (Golatkar et al., 2020; Izzo et al., 2021; Warnecke et al., 2021; Fan et al., 2023) assume full availability of both forget and retain data, applying selective updates to approximate retraining. Although effective, these methods incur high computational cost and require full data access, which is often infeasible. Partition-based approaches such as SISA (Bourtoule et al., 2021) restructure training to enable localized retraining when data is removed. In contrast, data-free strategies (Chundawat et al., 2023; Tarun et al., 2023) update model parameters without storing the forget set, often using noise substitutes as replacements. However, most evaluations focus on increased misclassification of forgotten classes, which does not necessarily imply genuine removal of their representations. This gap motivates principled methods that disentangle forgotten information from the feature space while preserving accuracy on retained data.

2.2 SUBSPACE-CONSTRAINED FEATURE REPRESENTATIONS

Our motivation for constraining feature spaces to structured manifolds stems from the desire to enhance class separability, interpretability, and model robustness. For instance, approaches such as SphereFace (Liu et al., 2017) and ArcFace (Deng et al., 2019) introduced angular constraints to address the limitations of traditional embedding separability, aiming for compact intra-class and distinct inter-class features. Minimum Hyperspherical Energy (MHE) regularization (Liu et al., 2018) was motivated by the desire to maximize geometric diversity, leading to uniformly distributed neuron representations. More recently, Zaeemzadeh et al. (2021) modeled features as a union of one-dimensional subspaces, where a dominant singular vector characterizes each class, enabling robust out-of-distribution (OOD) detection. These studies demonstrate the benefits of explicit embedding constraints and inspire our unlearning approach, where a union-of-subspaces formulation allows targeted suppression of forget classes with minimal interference to retained ones.

2.3 LOSS FUNCTIONS: CROSS-ENTROPY VS. MEAN SQUARED ERROR

Finally, to implement these geometric constraints effectively within unlearning frameworks, it is important to examine the underlying loss functions. Cross-entropy (CE), derived as the negative log-likelihood of a multinomial model, remains the default for classification due to its strong penalization of incorrect predictions (Bishop & Nasrabadi, 2006). However, recent studies show that mean squared error (MSE) can match or surpass CE across NLP, speech, and vision tasks (Hui & Belkin, 2020; Liu et al., 2022; Tyagi et al., 2024). Unlike CE's unbounded gradients, which can amplify label noise and overfitting, MSE yields bounded gradients, reducing variance, enhancing stability, and improving robustness in over-parameterized or imbalanced regimes. Although both share the same theoretical optimum under sufficient capacity, empirical evidence highlights MSE's advantages in noise resistance and training stability (Golik et al., 2013). This motivates our systematic comparison of CE and MSE within state-of-the-art unlearning frameworks under both constrained and unconstrained feature representations.

3 PROBLEM FORMULATION

3.1 CLASSIFIER HEADS

Let $\mathcal{D} = \{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^{N}$ denote a training set of N image–label pairs, with $\mathbf{x}^{(i)} \in \mathbb{R}^{ch\times h\times w}$ and $y^{(i)} \in \mathcal{Y} = \{1, \ldots, C\}$. A classifier $\Pi_{\boldsymbol{\theta},\mathbf{W}}(\mathbf{x})$ consists of a feature extractor $f(\mathbf{x};\boldsymbol{\theta}) : \mathbb{R}^{ch\times h\times w} \to \mathbb{R}^d$ and final linear classifier weights $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_C] \in \mathbb{R}^{d\times C}$. The feature of sample i is $\mathbf{z}^{(i)} = f(\mathbf{x}^{(i)};\boldsymbol{\theta})$ and its normalized embedding $\mathbf{v}^{(i)} = \mathbf{z}^{(i)}/\|\mathbf{z}^{(i)}\|$. For class c, the normalized prototype is $\mathbf{u}_c = \mathbf{w}_c/\|\mathbf{w}_c\|$.

We define four classifier heads, cosine (COS), ortho-cosine (ORC), softmax (SFX), and MSE-logit (MSL), that differ in geometric constraints and loss objectives. We derive **COS** classifier head logit

$\tilde{\ell}_c^{(i)}$ as the absolute value of the cosine similarity logit

$$\ell_c^{(i)} = \mathbf{u}_c^\top \mathbf{v}^{(i)} = \frac{\mathbf{w}_c^\top \mathbf{z}^{(i)}}{\|\mathbf{w}_c\| \, \|\mathbf{z}^{(i)}\|}, \qquad \tilde{\ell}_c^{(i)} = |\ell_c^{(i)}|, \qquad c = 1, \ldots, C, \tag{1}$$

The COS classifier head enforces direction-invariant decisions that depend only on feature alignment with the class prototypes. This, in turn, encourages the formation of compact, axial clusters for each class. The **ORC** variant extends COS by additionally enforcing orthogonality among classifier weights, $\mathbf{W}^\top \mathbf{W} = \mathbf{I}_C$ by freezing prototypes with an orthonormal $\mathbf{W}$, so each class occupies a distinct one-dimensional subspace orthogonal to other subspaces.

Let $\mathbf{t}^{(i)} \in \{0, 1\}^C$ denote the one-hot target for sample i, which c-th component is $t_c^{(i)}$. Given that, the COS and ORC difference lies in the model's prototype layer structure, we can jointly define the COS and ORC objective function under CE and MSE losses as

$$\mathcal{L}_{\mathrm{COS-CE}} = \mathcal{L}_{\mathrm{ORC-CE}} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{c=1}^{C} t_c^{(i)} \log \tilde{\ell}_c^{(i)}, \tag{2}$$

$$\mathcal{L}_{\mathrm{COS-MSE}} = \mathcal{L}_{\mathrm{ORC-MSE}} = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{C} \sum_{c=1}^{C} \big(t_c^{(i)} - \tilde{\ell}_c^{(i)}\big)^2. \tag{3}$$

By applying a softmax activation function to the logits to produce class probabilities $p_c^{(i)}$, we define the **SFX** classifier head as the conventional softmax-based classifier. This head trains the model using CE and MSE loss functions

$$\mathcal{L}_{\mathrm{SFX-CE}} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{c=1}^{C} t_c^{(i)} \log p_c^{(i)}, \tag{4}$$

$$\mathcal{L}_{\mathrm{SFX-MSE}} = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{C} \sum_{c=1}^{C} \big(t_c^{(i)} - p_c^{(i)}\big)^2. \tag{5}$$

Finally, the **MSL** is trained with MSE directly on the raw inner-product logits, $\ell'^{(i)}_c = \mathbf{w}_c^\top \mathbf{z}^{(i)}, c = 1, \ldots, C$, without softmax normalization, defined as below:

$$\mathcal{L}_{\mathrm{MSL-MSE}} = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{C} \sum_{c=1}^{C} \big(t_c^{(i)} - \ell'^{(i)}_c\big)^2. \tag{6}$$

The MSL-MSE differs from SFX–MSE, as it supervises probability distributions and often yields smoother optimization by avoiding the effects of softmax normalization. Enforcing intra-class alignment (features aligned with their prototypes) together with inter-class orthogonality (orthogonal prototypes) yields compact and separable feature subspaces. Such structured embeddings allow forgotten classes to be suppressed by attenuating their subspaces rather than misclassifying them into retained categories (Zaeemzadeh et al., 2021). The detailed analysis of boundary losses for MSE and CE on classifier heads is presented in Appendix A.

3.2 Machine Unlearning in the Class-Forgetting Setup

We focus on the *class forgetting* problem, where the objective is to remove the influence of entire classes or several classes from a pretrained model $\Pi_{\boldsymbol{\theta},\mathbf{W}}(\mathbf{x})$ while maintaining performance on the remaining classes. Given the complete dataset $\mathcal{D}$, the forget set $\mathcal{D}_{\mathrm{FG}}$ containing all samples of the forget classes and the retained set $\mathcal{D}_{\mathrm{RT}}$, the desired outcome is an unlearned model that approximates training on the retrain dataset $\mathcal{D}_{\mathrm{RT}} = \mathcal{D} \setminus \mathcal{D}_{\mathrm{FG}}$, without incurring the cost of full retraining. We analyze this setup under two practical conditions. In the *all-data available* case, both $\mathcal{D}_{\mathrm{FG}}$ and $\mathcal{D}_{\mathrm{RT}}$ are accessible, allowing direct removal of target forget classes. In contrast, the stricter *zero-glance privacy* regime (Tarun et al., 2023) assumes that $\mathcal{D}_{\mathrm{FG}}$ is entirely inaccessible due to privacy or deletion constraints, leaving only a limited subset of $\mathcal{D}_{\mathrm{RT}}$ for adaptation. Our proposed formulation introduces geometric constraints on the classifier space to enable principled suppression of forgotten classes, ensuring they are effectively erased from the model representation rather than misclassified into retained categories.

4 PROPOSED APPROACH

We present a unified class-forgetting framework that works in both all-data and zero-glance settings by combining three components: (i) an error-maximizing noise mechanism used to remove forget-class traces without access to the original samples; (ii) orthogonally constrained feature/classifier representations within which we adapt and extend state-of-the-art unlearning methods to evaluate performance in the class-forgetting regime; and (iii) a spectral-angle test that exploits the union-of-subspaces structure to verify genuine forgetting, offering an interpretable, geometry-aware complement to standard unlearning metrics.

4.1 ERROR–MAXIMIZING NOISE

Building on Tarun et al. (2023), we employ a differentiable generator $g_{\boldsymbol{\phi}}$ that maps Gaussian noise $\mathbf{n} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ to the initial synthetic input $\mathbf{X}_{\boldsymbol{\phi}}^{(0)} = g_{\boldsymbol{\phi}}(\mathbf{n}) \in \mathbb{R}^{B\times ch\times h\times w}$ for a frozen trained classifier $\Pi^*_{\boldsymbol{\theta},\mathbf{W}}$. After initialization, $\mathbf{X}_{\boldsymbol{\phi}}^{(k)} = g_{\boldsymbol{\phi}}(\mathbf{X}_{\boldsymbol{\phi}}^{(k-1)})$ would update the synthetic noise at each step k. For a target class $f \in \mathcal{Y}$, let $\mathbf{t}_f \in \{0,1\}^{|\mathcal{Y}|}$ denote the one–hot vector for f, where $|\mathcal{Y}|$ is the number of classes. Let $\mathbf{T}_f \in \{0,1\}^{B\times|\mathcal{Y}|}$ be the batchwise target obtained by repeating $\mathbf{t}_f$. We update the generator parameters $\boldsymbol{\phi}$ to maximize classification error while regularizing on the input noise of the previous step $\mathbf{X}_{\boldsymbol{\phi}}^{(k-1)}$ to prevent large noise generation:

$$\min_{\boldsymbol{\phi}} \mathbb{E}\Big[-\mathcal{L}\big(\Pi^*_{\boldsymbol{\theta},\mathbf{W}}(g_{\boldsymbol{\phi}}(\mathbf{X}_{\boldsymbol{\phi}}^{(k-1)})), \mathbf{T}_f\big)\Big] + \lambda\|\mathbf{X}_{\boldsymbol{\phi}}^{(k-1)}\|_2^2, \qquad \lambda > 0, \tag{7}$$

where $\mathcal{L}$ is the same loss used during classifier training, and $\mathbb{E}$ is the expectation over the noise batch. λ is a scalar hyperparameter that controls the magnitude of the synthetic input. With MSE, we set $\mathbf{t}_f = \mathbf{1}_C$ (an all-ones vector of dimension C), to maximize loss across all classes, yielding class-agnostic noise tailored to the model's geometry. In contrast, CE cannot directly enforce class-agnostic characteristics; we therefore alternate between minimizing error with random non-forget one-hot targets (95% of epochs) and maximizing error with forget one-hot target $\mathbf{t}_f$ for the remaining epochs. This strategy approximates class-agnostic behavior while still exploiting CE's discriminative structure. For cosine-based classifiers, the generated noise can further be interpreted via feature–prototype cosines, explicitly discouraging alignment with all prototypes. Optimization details are provided in Section B.1.

4.2 CLASS FORGETTING WITH PROTOTYPE CONSTRAINTS

To assess unlearning within our framework, we adapt multiple leading methods for class forgetting. Gradient Ascent (GA) (Thudi et al., 2022), which increases the loss of the model over $\mathcal{D}_{\mathrm{FG}}$ to mimic the reverse training process of model weights. We employ fine-tuning on the retain set (FTR) (Warnecke et al., 2021), which updates the model's weights using a portion of $\mathcal{D}_{\mathrm{RT}}$. We introduce FTR's data-independent variant, FTN, which trains on error-maximizing noise to replace $\mathcal{D}_{\mathrm{FG}}$. Influence Unlearning (IU) (Izzo et al., 2021) identifies dominant weights contributing to $\mathcal{D}_{\mathrm{FG}}$ predictions via influence function Koh & Liang (2017) and removes their effect. In addition, we employ sparsity-aware approaches (Jia et al., 2023) that impose an ℓ_1 − norm penalty ($\|.\|_1$) to promote weight pruning:

$$\min_{\boldsymbol{\theta},\mathbf{W}} \mathbb{E}_{\mathbf{x}^{(i)}\sim\mathcal{D}_{\mathrm{RT}}}\Big[-\mathcal{L}\big(\Pi_{\boldsymbol{\theta},\mathbf{W}}(\mathbf{x}^{(i)}), \mathbf{t}^{(i)}\big)\Big] + \gamma\|(\boldsymbol{\theta},\mathbf{W})\|_1, \qquad \gamma > 0. \tag{8}$$

Here, γ is a constant regularization term that controls the pruning of unimportant weights in the model. We extend sparsity-aware fine-tuning with $\mathcal{D}_{\mathrm{RT}}$ (LSPR) (Equation (8)) to its data-independent counterpart (LSPN) to fine-tune with the generated noise instead of $\mathcal{D}_{\mathrm{RT}}$. We also applied the saliency unlearning approach, SalUn (Fan et al., 2023), which identifies salient parameters that influence the prediction of the forget set via gradient-based masks. To support zero-glance privacy, we compute saliency maps using $\mathcal{D}_{\mathrm{RT}}$ instead of $\mathcal{D}_{\mathrm{FG}}$:

$$m_S{}' = \mathbb{0}\left(\left|\nabla_{\boldsymbol{\theta},\mathbf{W}}\mathcal{L}\big(\Pi_{\boldsymbol{\theta},\mathbf{W}}(\mathbf{x}^{(i)}), \mathbf{t}^{(i)}\big)\right|_{\mathbf{x}^{(i)}\in\mathcal{D}_{\mathrm{RT}}} \geq \xi\right), \tag{9}$$

where $\mathbb{0}$ is an element-wise indicator function which yields a value of 0 for the i-th element, gradient is less than gradient threshold ξ, and one otherwise. Parameters are then selectively updated:

$$(\boldsymbol{\theta}, \mathbf{W})_{unlearned} = (\mathbf{1} - m_S{}') \odot ((\boldsymbol{\theta}, \mathbf{W}) + \Delta(\boldsymbol{\theta}, \mathbf{W})) + m_S{}' \odot (\boldsymbol{\theta}, \mathbf{W}), \tag{10}$$

where $\odot$ denotes the element-wise product. The modified loss combines noise-based forgetting with retain fine-tuning:

$$\min_{\Delta(\boldsymbol{\theta},\mathbf{W})} \mathbb{E}_{(\mathbf{x}^{(i)},\mathbf{t}^{(i)})\sim(\mathbf{X}_{\boldsymbol{\phi}},\mathbf{t}_{\mathbf{f}})}\left[\mathcal{L}\big(\Pi_{\boldsymbol{\theta},\mathbf{W}}(\mathbf{x}^{(i)}), \mathbf{t}^{(i)}\big)\right] + \mathbb{E}_{\mathbf{x}^{(i)}\sim\mathcal{D}_{\mathrm{RT}}}\left[\mathcal{L}\big(\Pi_{\boldsymbol{\theta},\mathbf{W}}(\mathbf{x}^{(i)}, \mathbf{t}^{(i)})\big)\right]. \tag{11}$$

Finally, we adopt Unlearning by Selective Impair and Repair (UNSIR) (Tarun et al., 2023) on our framework. The *impair* step aggressively fine-tunes on noise and $\mathcal{D}_{\mathrm{RT}}$ with a high learning rate, corrupting influence of forget-class data on model's weights, followed by a *repair* step that stabilizes performance by retraining only on $\mathcal{D}_{\mathrm{RT}}$ at a lower rate.

4.3 Spectral Angle-Based Forgetting Test

We assess genuine class forgetting with a spectral-angle test inspired by OOD detection via unions of one-dimensional subspaces (Zaeemzadeh et al., 2021). Each class c is represented by the leading singular vector $\mathbf{v}_1^{(c)}$ of its feature subspace. For a test feature $\mathbf{z}$, we compute

$$\theta_c(\mathbf{z}) = \arccos\left(\frac{|\mathbf{z}^\top \mathbf{v}_1^{(c)}|}{\|\mathbf{z}\|}\right), \tag{12}$$

and use $\theta_{\min}(\mathbf{z}) = \min_{c\in\mathcal{C}} \theta_c(\mathbf{z})$ as the class assignment score. We select the threshold θ^* via Receiver Operating Characteristic (ROC) analysis to balance rejection of forgotten samples against retention accuracy. Therefore, if $\theta_{\min}(\mathbf{z}) \leq \theta^*$, $\mathbf{z}$'s corresponding sample is assigned to class c; otherwise it is labeled "I don't know" (IDK). This test directly probes the geometry of the embedding space, providing a measure of genuine forgetting against misclassification—a complement to standard metrics—and an indicator of the memory removal capability of unlearning methods.

5 Experiments

We study the impact of intra- and inter-class constraints together with the loss choice (CE vs. MSE) on the performance of introduced unlearning methods. Experiments are performed on CIFAR-10 with $C = 10$ classes. In each trial, one class is designated as the forget target and the remaining nine are retained; this is repeated for all classes, with 10 random seeds per class (100 trials total). Results are reported as mean $\pm$ standard deviation across trials and benchmarked against retrained baselines.

5.1 Dataset and Model Training

We use CIFAR-10 (Krizhevsky et al., 2009), which contains 50,000 training and 10,000 test RGB images of size 32×32 across 10 balanced classes. In the class-forgetting setup, the forget set $\mathcal{D}_{\mathrm{FG}}$ corresponds to all samples from one class, while the retained set $\mathcal{D}_{\mathrm{RT}}$ includes 1,000 samples from the remaining classes. We study both all-data and zero-glance conditions, where $\mathcal{D}_{\mathrm{FG}}$ is accessible or withheld, respectively. Standard CIFAR-10 preprocessing was applied, including 4-pixel reflective padding and per-channel normalization (mean $[125.3,\ 123.0,\ 113.9]/255$; standard deviation $[63.0,\ 62.1,\ 66.7]/255$).

We trained ResNet-18 models with introduced classifier heads from scratch for 200 epochs with a batch size of 64 on Stochastic Gradient Descent (SGD) with an initial learning rate of 0.1, momentum of 0.9, Nesterov acceleration enabled, and weight decay of 5×10^{-4} under both CE and MSE losses. The learning rate followed a StepLR schedule, decaying by a factor of 0.1 every 20 epochs.

5.2 Benchmarks and Evaluation Metrics

We evaluate unlearning across GA, FTR, FTN, LSPN, LSPR, modified SalUn, and UNSIR, under CE and MSE losses, as well as COS, ORC, SFX, and MSL classifier heads against the retrained model

trained only on $\mathcal{D}_{\mathrm{RT}}$. Training details and hyperparameter settings for each method are provided in the Appendix B.2.

Unlearning effectiveness is assessed using standard accuracy metrics. Unlearning Accuracy (UA) measures residual accuracy on the forget set $\mathcal{D}_{\mathrm{FG}}$, while Retained Accuracy (RA) evaluates fidelity on the retained set $\mathcal{D}_{\mathrm{RT}}$. Their test counterparts, TUA and TRA, are computed on unseen forget and retain samples. Privacy guarantees are quantified using the Membership Inference Attack (MIA) (Shokri et al., 2017), where a shadow model is trained to distinguish between member and non-member samples based on attack features. The MIA score is defined as $\frac{\mathrm{TN}}{|\mathcal{D}_{\mathrm{FG}}|}$. Where TN is the number of forget–set examples correctly identified as non–members (true negatives) and $|\mathcal{D}_{\mathrm{FG}}|$ denotes the cardinality of the forget set. Higher MIA values indicate stronger privacy protection.

Finally, for analyzing our IDK detection performance, we report four detection-based metrics defined as: FPR@TPR90, the false positive rate at 90% recall, Threshold@TPR90, the corresponding decision threshold reported in degrees, Detection Error@TPR90, representing minimum misclassification error at the decision threshold, and AUC, the area under the Receiver Operating Characteristic (ROC) curve, summarizing overall separability of the IDK detection.

5.3 Experiment Results

In this section, we present the results of unlearning under two experimental settings: Quick and Optimum unlearning. We impose Quick unlearning by limiting each method to a maximum of three training epochs, simulating fast but constrained updates. Moreover, we conduct Optimum unlearning by extending training up to 20 epochs, allowing methods to reach their best achievable performance. For each setting, hyperparameters are tuned to yield the strongest results under the respective setting as described in Appendix B.2.

Table 1 summarizes the best-performing unlearning methods in Quick and Optimum setups, across different classifier variants and loss functions. The results clearly indicate that COS, ORC, and MSL achieve unlearning stability and balance, characterized by low UA and high RA, when an appropriate loss function is chosen. In the Quick setup, GA with MSE loss on the COS head (GA-COS-MSE) offers the best trade-off, followed by LSPR-ORC-MSE and LSPR-SFX-CE, respectively. For MSL, UNSIR produces the best results, maintaining RA above 90%. In the Optimum setup, MSE's advantage is more apparent: GA on COS and MSL, and LSPR on ORC, all deliver near-zero UA while preserving RA above 95%, closely matching the retrain baselines.

These findings highlight that MSE-based approaches with structured classifiers (COS, ORC, MSL) consistently outperform CE-based ones, both in terms of stronger forgetting guarantees (lower UA) and higher retention (RA and TRA) in the Optimum setting. Moreover, the stability of results under Optimum settings indicates that prolonged fine-tuning epochs consolidate the gains from constrained feature geometries. The strong performance of LSPR and GA across multiple heads underscores their effectiveness as fast yet reliable unlearning strategies. By contrast, while UNSIR remains competitive in Quick settings, its improvements are less consistent once extended training epochs are allowed. It is worth mentioning that our proposed variant unlearning methods, FTN and LSPN, offer a possible solution for unlearning under total data unavailability (no access to $\mathcal{D}$) conditions; however, they generally exhibit weak trade-offs. These methods often have high UA or poor RA, making them less promising in practice. Therefore, we do not provide a detailed analysis of them in this paper. A more comprehensive comparison of all methods, along with additional ablations, is presented in the Appendix C.1, where the full scope of results is discussed.

Now, to examine the models based on their ability to reject a forget sample under the IDK detection notion, we identify the best-performing unlearning strategies under both Quick and Optimum setups by prioritizing high AUC and low FPR values (Table 2). Under the Quick setup, LSPR-SFX-MSE attains the strongest detection trade-off (AUC $= 0.921$, FPR $= 0.214$), followed by LSPR-COS-MSE (0.876, 0.268). For ORC, the GA-CE variant emerges as the most effective (0.890, 0.219), while in the MSL setting, the best achievable performance remains limited, with GA-MSE producing only moderate results (0.680, 0.736). Moving to the Optimum setup, performance improves consistently across models. GA-SFX-CE achieves the best overall trade-off (0.914, 0.255), with LSPR-COS-MSE and GA-ORC-CE also showing strong results (0.912, 0.218 and 0.910, 0.214, respectively). MSL again lags, with GA-MSE reaching only 0.863 AUC at 0.293 FPR.
Overall, these results indicate that GA and LSPR consistently yield the best detection trade-offs

Table 1: Best-performing unlearning methods under **Quick** and **Optimum** setups (FTN and LSPN excluded). For each model type (SFX, COS, ORC, MSL) and loss (CE, MSE), the method with the best UA–RA trade-off is reported, alongside Test UA/RA (TUA/TRA), Membership Inference Attack (MIA), and baseline retrain accuracies (Retrain TRA/TUA).

Method	Model	Loss	UA (↓)	RA (↑)	TUA (↓)	TRA (↑)	MIA (↑)	Retrain TRA (↑)	Retrain TUA (↓)
Quick									
LSPR	SFX	CE	2.65 ± 2.83	94.4 ± 0.81	2.55 ± 2.89	89.3 ± 1.03	1 ± 0.0004	92.5 ± 0.80	0 ± 0
GA	COS	MSE	0.14 ± 0.18	96.3 ± 1.26	0.15 ± 0.17	90.6 ± 1.55	1 ± 0.0007	94.6 ± 0.56	0 ± 0
LSPR	ORC	MSE	0.12 ± 0.25	94.3 ± 1.89	0.08 ± 0.23	88.9 ± 1.95	1 ± 0.0000	94.5 ± 0.63	0 ± 0
UNSIR	MSL	MSE	0.67 ± 1.00	90.9 ± 2.99	0.73 ± 1.11	85.9 ± 2.77	1 ± 0.0002	94.5 ± 0.56	0 ± 0
Optimum									
LSPR	SFX	CE	0.870 ± 1.270	96.4 ± 0.37	0.79 ± 1.21	90.8 ± 0.74	1 ± 0	92.5 ± 0.80	0 ± 0
GA	COS	MSE	0.005 ± 0.012	95.6 ± 1.55	0.02 ± 0.04	90.2 ± 1.98	1 ± 0	94.6 ± 0.56	0 ± 0
LSPR	ORC	MSE	0.001 ± 0.006	95.9 ± 1.26	0.00 ± 0.00	90.3 ± 1.44	1 ± 0	94.5 ± 0.63	0 ± 0
GA	MSL	MSE	0.044 ± 0.037	95.3 ± 1.30	0.04 ± 0.07	89.2 ± 1.35	1 ± 0	94.5 ± 0.56	0 ± 0

across different model types, particularly under the Optimum setup in terms of detecting forget set as IDK samples during testing.

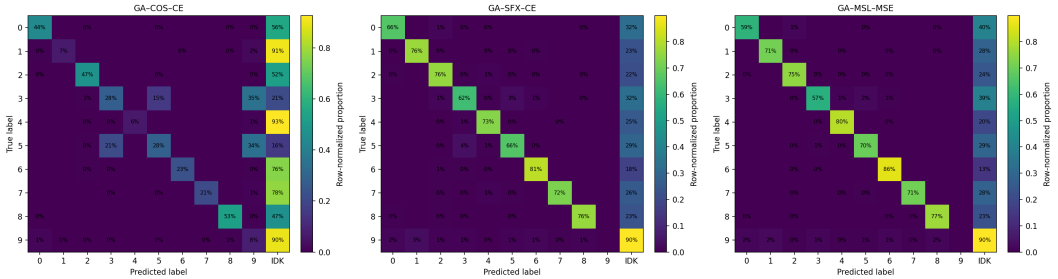
Table 2: Best IDK detection trade-offs under **Quick** and **Optimum** setups (FTN and LSPN excluded). For each model (SFX, COS, ORC, MSL), we select the method-loss pair that maximizes AUC (tied by the lowest FPR). The reported metrics are the area under the Receiver Operating Characteristic (ROC) curve (AUC), the false positive rate at 90% recall (FPR@TPR90), the corresponding decision threshold reported in degrees (Thr@TPR90), minimum misclassification error at the decision threshold (DetErr@TPR90).

Method	Model	Loss	AUC (↑)	FPR@TPR90 (↓)	Thr@TPR90	DetErr@TPR90 (↓)
Quick						
LSPR	SFX	MSE	0.921 ± 0.00817	0.214 ± 0.0164	13.2 ± 1.24	0.157 ± 0.00822
LSPR	COS	MSE	0.876 ± 0.12000	0.268 ± 0.1380	28.3 ± 6.38	0.184 ± 0.06900
GA	ORC	CE	0.890 ± 0.0212	0.219 ± 0.0678	10.1 ± 3.25	0.159 ± 0.03390
GA	MSL	MSE	0.680 ± 0.0597	0.736 ± 0.0869	9.41 ± 2.80	0.418 ± 0.04340
Optimum						
GA	SFX	CE	0.914 ± 0.0107	0.255 ± 0.0361	26.5 ± 0.981	0.178 ± 0.0181
LSPR	COS	MSE	0.912 ± 0.0378	0.218 ± 0.0555	40.7 ± 4.850	0.159 ± 0.0278
GA	ORC	CE	0.910 ± 0.0197	0.214 ± 0.0647	8.32 ± 2.680	0.157 ± 0.0323
GA	MSL	MSE	0.863 ± 0.0180	0.293 ± 0.0469	8.08 ± 0.476	0.197 ± 0.0234

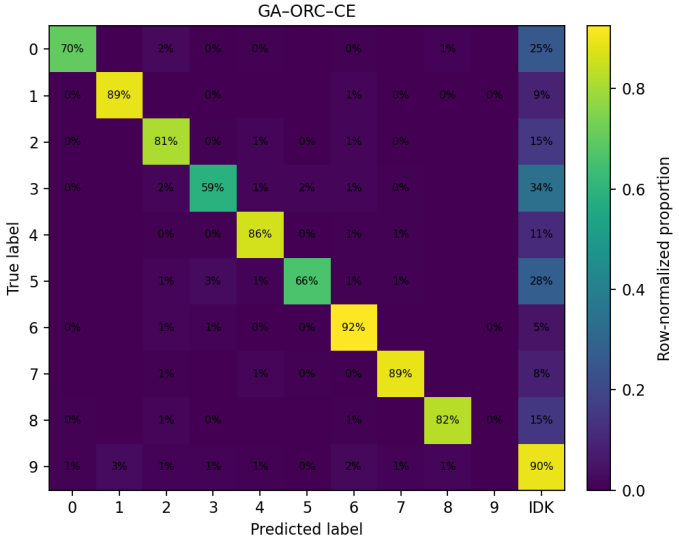
To represent a more detailed analysis of the effect of classifier heads and loss functions on IDN detection, we represent the confusion matrix after unlearning class 9 in the Optimum setting for all combinations of model-loss for GA method. In Figure 2, each row is the true class and each column is the predicted class; the last column (IDK) labels samples as *unseen* via comparison with a threshold ξ chosen to achieve 90% TPR (Thr@TPR90). Focusing on retained classes, GA–ORC–CE shows the cleanest diagonals ($\approx 70\%$–92%) with relatively low IDK spillover ($\approx 5\%$–34%), indicating selective rejection focused on the unlearned class. GA–SFX–CE is close but weaker (diagonals $\approx 66\%$–81%; IDK $\approx 18\%$–32%), while GA–COS–CE has poor diagonals and heavy IDK, suggesting weak separation. Under MSE, GA–ORC–MSE largely collapses to IDK for many retained classes (IDK $\gtrsim 80\%$), whereas GA–SFX–MSE and GA–MSL–MSE keep much higher in-class accuracy (diagonals $\approx$ 59%–86%) with moderate IDK ($\approx$ 13%–40%). Overall, in the Optimum setting, ORC+CE yields the most selective forgetting, while with MSE, the SFX/MSL heads are more robust, and ORC–MSE tends to over-reject.

6 Conclusion

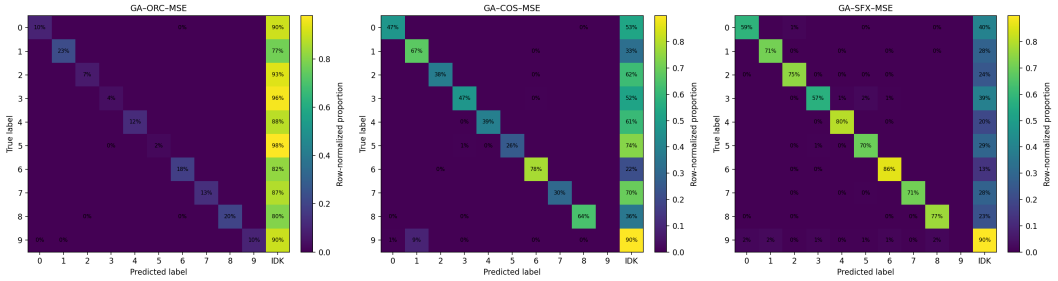
We introduced cosine (COS) and ortho-cosine (ORC) classifier heads that impose intra-class alignment and inter-class orthogonality, embedding features on a union of one-dimensional subspaces. This



(a) GA–COS–CE (b) GA–SFX–CE (c) GA–MSL–MSE



(d) GA–ORC–CE



(e) GA–ORC–MSE (f) GA–COS–MSE (g) GA–SFX–MSE

Figure 2: Confusion matrices after unlearning class 9 under the **Optimum** setup for GA with different heads and losses. Rows are true labels and columns are predicted labels (last column is IDK). Values are row-normalized proportions.

geometry enables the selective suppression of the forgotten class while preserving decision boundaries for retained classes, resulting in consistent gains in forget accuracy and retained accuracy across data-rich and zero-glance scenarios when paired with diverse unlearning methods. To assess whether a class is truly forgotten rather than merely misclassified, we proposed a geometry-aware spectral-angle criterion that certifies removal of the forgotten subspace and complements standard metrics (UA/RA, TUA/TRA) and privacy checks (MIA). Finally, our analysis of CE versus MSE across heads clarifies when each pairing is preferable: CE drives aggressive class separation and is most effective with ORC when selective forgetting is paramount, whereas MSE yields stable updates and, when applied to COS, reduces rival attraction and collateral rejection, supporting robust unlearning with strong retention.

REFERENCES

Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

Lucas Bourtoule, Varun Chandrasekaran, Christopher A Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. Machine unlearning. In *2021 IEEE symposium on security and privacy (SP)*, pp. 141–159. IEEE, 2021.

Vikram S Chundawat, Ayush K Tarun, Murari Mandal, and Mohan Kankanhalli. Zero-shot machine unlearning. *IEEE Transactions on Information Forensics and Security*, 18:2345–2354, 2023.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4690–4699, 2019.

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. *arXiv preprint arXiv:2310.12508*, 2023.

Jack Foster, Stefan Schoepf, and Alexandra Brintrup. Fast machine unlearning without retraining through selective synaptic dampening. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 12043–12051, 2024.

Aditya Golatkar, Alessandro Achille, and Stefano Soatto. Eternal sunshine of the spotless net: Selective forgetting in deep networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9304–9312, 2020.

Pavel Golik, Patrick Doetsch, and Hermann Ney. Cross-entropy vs. squared error training: a theoretical and experimental comparison. In *Interspeech*, volume 13, pp. 1756–1760, 2013.

Chris Jay Hoofnagle, Bart Van Der Sloot, and Frederik Zuiderveen Borgesius. The european union general data protection regulation: what it is and what it means. *Information & Communications Technology Law*, 28(1):65–98, 2019.

Like Hui and Mikhail Belkin. Evaluation of neural architectures trained with square loss vs cross-entropy in classification tasks. *arXiv preprint arXiv:2006.07322*, 2020.

Zachary Izzo, Mary Anne Smart, Kamalika Chaudhuri, and James Zou. Approximate data deletion from machine learning models. In *International conference on artificial intelligence and statistics*, pp. 2008–2016. PMLR, 2021.

Jinghan Jia, Jiancheng Liu, Parikshit Ram, Yuguang Yao, Gaowen Liu, Yang Liu, Pranay Sharma, and Sijia Liu. Model sparsity can simplify machine unlearning. *Advances in Neural Information Processing Systems*, 36:51584–51605, 2023.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *International conference on machine learning*, pp. 1885–1894. PMLR, 2017.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Sheng Liu, Zhihui Zhu, Qing Qu, and Chong You. Robust training under label noise by over-parameterization. In *International Conference on Machine Learning*, pp. 14153–14172. PMLR, 2022.

Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 212–220, 2017.

Weiyang Liu, Rongmei Lin, Zhen Liu, Lixin Liu, Zhiding Yu, Bo Dai, and Le Song. Learning towards minimum hyperspherical energy. *Advances in neural information processing systems*, 31, 2018.

Vasileios Perifanis, Efstathios Karypidis, Nikos Komodakis, and Pavlos Efraimidis. Sftc: Machine unlearning via selective fine-tuning and targeted confusion. In *Proceedings of the 2024 European Interdisciplinary Cybersecurity Conference*, pp. 29–36, 2024.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pp. 3–18. IEEE, 2017.

State of California. California consumer privacy act of 2018. `https://oag.ca.gov/privacy/ccpa`, 2018. California Civil Code Section 1798.100–1798.199.

Ayush K Tarun, Vikram S Chundawat, Murari Mandal, and Mohan Kankanhalli. Fast yet effective machine unlearning. *IEEE Transactions on Neural Networks and Learning Systems*, 35(9):13046–13055, 2023.

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

Kanishka Tyagi, Chinmay Rane, Ketaki Vaidya, Jeshwanth Challgundla, Soumitro Swapan Auddy, and Michael Manry. Making sigmoid-mse great again: Output reset challenges softmax cross-entropy in neural network classification. *arXiv preprint arXiv:2411.11213*, 2024.

Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5265–5274, 2018.

Alexander Warnecke, Lukas Pirch, Christian Wressnegger, and Konrad Rieck. Machine unlearning of features and labels. *arXiv preprint arXiv:2108.11577*, 2021.

Alireza Zaeemzadeh, Niccolo Bisagno, Zeno Sambugaro, Nicola Conci, Nazanin Rahnavard, and Mubarak Shah. Out-of-distribution detection using union of 1-dimensional subspaces. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pp. 9452–9461, 2021.

APPENDIX

A BOUNDARY-LOSS ANALYSIS FOR CE VS. MSE UNDER COS/ORC/SFX/MSL HEADS

For sample i, let $\mathbf{z}^{(i)} = f(\mathbf{x}^{(i)};\theta)$, $\mathbf{v}^{(i)} = \mathbf{z}^{(i)}/\|\mathbf{z}^{(i)}\|$, $\mathbf{u}_c = \mathbf{w}_c/\|\mathbf{w}_c\|$, and $\ell_c^{(i)} = \mathbf{u}_c^\top \mathbf{v}^{(i)}$. Given that the COS/ORC heads use the folded cosine score $\tilde{\ell}_c^{(i)} = |\ell_c^{(i)}|$ (direction-invariant alignment) which ORC further enforces $\mathbf{W}^\top\mathbf{W} = \mathbf{I}_C$ on the model structure. For SFX, the (dot-product) logits are $\ell'^{(i)}_c = \mathbf{w}_c^\top \mathbf{z}^{(i)}$ and $p_c^{(i)} = \mathrm{softmax}\big(\tilde{\ell}_c^{(i)}\big)$. The one-hot targets are defined as $\mathbf{t}^{(i)} \in \{0,1\}^C$ with $t_c^{(i)} = \mathbb{1}[y^{(i)} = c]$, where $\mathbb{1}$ is an element-wise indicator function which yields a value of 1 for the i-th element correspond to class c and zero otherwise.

AS mentioned, we pair each classifier head with the loss defined in its output space. The CE loss is applied to probabilities (COS/ORC/SFX-CE) where COS/ORC logits are converted to probabilities via a softmax and optimized with CE; SFX-CE is identical but uses the $\ell'^{(i)}_c$ logits. In addition, we employed MSE loss on probabilities (SFX-MSE) and MSE on folded cosine logits $\ell_c^{(i)}$ (COS/ORC-MSE) and dot-product logits $\ell'^{(i)}_c$ (MSL-MSE). To better understand the effect of these combinations on the unlearning process, we study their per-sample loss and derive their per-sample boundary-loss ranges.

MSE on probabilities (SFX-MSE) Let $y_c := p_c^{(i)}$ be a probability vector with true class c, and the rival probability $r = (p_j^{(i)})_{j\neq c}$ with $r_j \geq 0$, the MSE loss can be expressed in a decomposed form of an under-confidence term $S := 1 - y_c$ for the true class and an aggregate rival-mass term $\sum_{j\neq c} r_j = S$:

$$\mathcal{L}_{\mathrm{MSE}}(p;c) = \tfrac{1}{C}\left[(1-y_c)^2 + \sum_{j\neq c} p_j^2\right] = \tfrac{1}{C}\left[S^2 + \|r\|_2^2\right].$$

Therefore, the range of $\mathcal{L}_{\mathrm{MSE}}$ for fixed y_c is determined by the range of $\|r\|_2^2$ under the simplex constraint ($\|r\|_2^2 \in [\frac{S^2}{C-1}, S^2]$). The Lower bound is reached when the probability rival mass spread is distributed evenly among all classes ($r_i = S/(C-1)$). By Cauchy–Schwarz we have,

$$(\mathbf{1}^\top r)^2 \;\leq\; \|\mathbf{1}\|_2^2\,\|r\|_2^2 \;\Longrightarrow\; S^2 \;\leq\; (C-1)\sum_{i\neq c} r_i^2.$$

$$\|r\|_2^2 \;\geq\; \frac{(\sum_i r_i)^2}{C-1} \;=\; \frac{S^2}{C-1}.$$

Hence

$$\mathcal{L}_{\mathrm{MSE}}(y;c) \;\geq\; \frac{1}{C}(S^2 + \frac{S^2}{C-1}) \;=\; \frac{1}{C}(\frac{C}{C-1}\,S^2) \;=\; \frac{1}{C}(\frac{C}{C-1}(1-y_c)^2).$$

The Upper bound is reached when the probability rival mass spread is maximized in one rival class. Since a rival probability, $u \mapsto u^2$ is convex, $\|r\|_2^2$ is maximized at an extreme point of the simplex: take one rival $= S$, others $= 0$. Then $\|r\|_2^2 = S^2$, giving

$$\mathcal{L}_{\mathrm{MSE}}(y;c) \;\leq\; \frac{1}{C}(S^2 + S^2) \;=\; \frac{2}{C}(1-y_c)^2.$$

$$\boxed{\frac{1}{C}\,\frac{C}{C-1}(1-y_c)^2 \;\leq\; \mathcal{L}_{\mathrm{MSE}}(p;c) \;\leq\; \frac{1}{C}\,2(1-y_c)^2, \quad y_c \in [0,1].}$$

Thus $y_c \in [0,1]$, the global bound will be $0 \leq \mathcal{L}_{\mathrm{MSE}}(p;c) \leq 2/C$ per example.

CE on probabilities (SFX-CE). For the per-example CE loss

$$\mathcal{L}_{\mathrm{CE}}(y; c) = -\log y_c, \qquad y_c \in (0, 1].$$

Since $y_c \leq 1$, $-\log y_c \geq 0$, with equality under perfect confidence on the true class ($y_c = 1$) the CE lower bound will be 0. However, as $y_c \downarrow 0$, $-\log y_c \to \infty$ and no finite upper bound exists for CE. Thus, the per-sample CE loss is

$$\boxed{\mathcal{L}_{\mathrm{CE}}(p; c) = -\log y_c \;\geq\; 0, \qquad \text{unbounded above as } y_c \downarrow 0.}$$

CE can assign arbitrarily large penalties to hard errors; MSE on probabilities is bounded and caps per-sample influence.

MSE on folded-cosine logits (COS-MSE / ORC-MSE). All per-example error bounds we derived earlier depend on y_c. What changes is the mapping from geometry to y_c. Since $\tilde{\ell}_j^{(i)} \in [0, 1]$, $0 \leq (\tilde{\ell}_c - 1)^2 \leq 1, \quad \text{and} \quad 0 \leq (\tilde{\ell}_j)^2 \leq 1 \; (j \neq c)$

The Lower bound is 0 obtained at the ideal point $\tilde{\ell}_c = 1,\ \tilde{\ell}_{j \neq c} = 0$. However, the upper bound would require $\tilde{\ell}_c = 0$ and $\tilde{\ell}_{j \neq c} = 1 \; \forall j$, which is geometrically unattainable unless the feature vector v aligns with multiple classes. Therefore, the correct worst-case bound without any geometric assumptions is $\frac{1}{C}(1 + (C-1) \cdot 1) \;=\; 1$. The per-sample bound is

$$\boxed{0 \;\leq\; \mathcal{L}_{\mathrm{MSE}}(\tilde{\ell}; c) = \tfrac{1}{C} \sum_j \big(t_j - \tilde{\ell}_j\big)^2 \;\leq\; 1,}$$

where the upper per-example bound is tighter than the probability-space MSE bound.

CE on folded–cosine logits (COS–CE / ORC–CE). The per-example CE loss with folded cosine logit can be written as:

$$\mathcal{L}_{\mathrm{CE}}\big(p^{(i)}; c\big) = -\,\tilde{\ell}_c^{(i)} + \log\Big(\sum_{k=1}^{C} e^{\tilde{\ell}_k^{(i)}}\Big) = \log\Big(1 + \sum_{j \neq c} e^{\,\tilde{\ell}_j^{(i)} - \tilde{\ell}_c^{(i)}}\Big).$$

We defined a (direction-invariant) margin against the closest rival class:

$$\Delta^{(i)} \;\coloneqq\; \tilde{\ell}_c^{(i)} - \max_{j \neq c} \tilde{\ell}_j^{(i)} \;\in [-1, 1].$$

We can observe that, increasing the margin strictly decreases the loss as $\frac{\partial \mathcal{L}_{\mathrm{CE}}}{\partial \tilde{\ell}_c^{(i)}} = p_c^{(i)} - 1 < 0$ unless $p_c^{(i)} = 1$, while $\frac{\partial \mathcal{L}_{\mathrm{CE}}}{\partial \tilde{\ell}_j^{(i)}} = p_j^{(i)} \geq 0$ for each rival $j \neq c$. Equivalently,

$$\mathcal{L}_{\mathrm{CE}}\big(p^{(i)}; c\big) = \log\Big(1 + \sum_{j \neq c} e^{-(\tilde{\ell}_c^{(i)} - \tilde{\ell}_j^{(i)})}\Big),$$

and each summand decreases as the gap $\tilde{\ell}_c^{(i)} - \tilde{\ell}_j^{(i)}$ widens. This will give us the margin bounds:

$$\boxed{\log\big(1 + e^{-\Delta^{(i)}}\big) \;\leq\; \mathcal{L}_{\mathrm{CE}}\big(p^{(i)}; c\big) \;\leq\; \log\big(1 + (C-1)e^{-\Delta^{(i)}}\big),}$$

which are strictly decreasing in $\Delta^{(i)}$. Because $\tilde{\ell}_k^{(i)} \in [0, 1]$ implies $\Delta^{(i)} \in [-1, 1]$, plugging the endpoints gives a global range under this logit box:

$$\underbrace{\log\big(1 + e^{-1}\big)}_{\text{best separation } (\Delta^{(i)} = 1)} \;\leq\; \mathcal{L}_{\mathrm{CE}}\big(p^{(i)}; c\big) \;\leq\; \underbrace{\log\big(1 + (C-1)e^{\,1}\big)}_{\text{worst separation } (\Delta^{(i)} = -1)}.$$

The ORC head enforces $\mathbf{W}^\top \mathbf{W} = \mathbf{I}_C$, which tends to enlarge $\Delta^{(i)}$ by reducing rival alignments; however, the per-sample CE form and the bounds above remain unchanged for ORC-CE.

MSE on dot-logits (MSL–MSE). For the dot head with unbounded logits ℓ'_j, no finite global upper bound exists unless norms are controlled. Considering the MSE-MSL objective (MSL) as:

$$\tilde{\mathcal{L}}_{\mathrm{MSE}}(\boldsymbol{\ell}'; c) = \frac{1}{C}\Big[(\ell'_c - 1)^2 + \sum_{j \neq c} (\ell'_j)^2\Big], \qquad \ell'_j = \mathbf{w}_j^\top \mathbf{z} \in \mathbb{R}.$$

If the logits ℓ'_j are unconstrained (no bounds on $\|\mathbf{w}_j\|$ and $\|\mathbf{z}\|$), then:

$$\boxed{0 \;\le\; \tilde{\mathcal{L}}_{\mathrm{MSE}}(\boldsymbol{\ell}'; c) \;<\; \infty}$$

with the minimum 0 achieved at $\ell'_c = 1$ and $\ell'_{j \neq c} = 0$, with no finite upper bound.

Assume logits are bounded coordinate-wise:

$$|\ell'_j| \;\le\; B \quad \text{for all } j \in \{1, \ldots, C\},$$

where $\|\mathbf{w}_j\| \le R$ and $\|\mathbf{z}\| \le H$, giving $B \le R \times H$ by Cauchy–Schwarz. Therefore, MSL-MSE is optimized over the hyper-rectangle $[-B, B]^C$. The lower bound is achieved as the rival terms are nonnegative and minimized at $\ell'_{j \neq c} = 0$, and the target error $(\ell'_c - 1)^2$ is minimized by:

$$\ell'^{\star}_c = \arg\min_{\ell'_c \in [-B, B]} (\ell'_c - 1)^2 = \begin{cases} 1, & B \ge 1, \\ B, & B < 1, \end{cases}$$

hence we derive the lower bound as:

$$\tilde{\mathcal{L}}_{\mathrm{MSE}}(\boldsymbol{\ell}'; c) \;\ge\; \frac{1}{C}\left(\max\{0,\, 1 - B\}\right)^2 \;=\; \begin{cases} 0, & B \ge 1, \\ \dfrac{(1-B)^2}{C}, & B < 1. \end{cases}$$

On the other hand, the rival sum is maximized at $|\ell'_{j \neq c}| = B$, giving $\sum_{j \neq c} (\ell'_j)^2 = (C-1)B^2$ and the target error term $(\ell'_c - 1)^2$ is maximized at the endpoint farthest from 1, namely $\ell'_c = -B$, giving target error $= (1 + B)^2$. Therefore, the upper bound is achieved as:

$$\tilde{\mathcal{L}}_{\mathrm{MSE}}(\boldsymbol{\ell}'; c) \;\le\; \frac{1}{C}\Big[(1 + B)^2 + (C - 1)B^2\Big] = \frac{1 + 2B + CB^2}{C}.$$

We can show the MSL-MSE per-sample error under the box constraint as:

$$\boxed{\frac{(\max\{0,\, 1 - B\})^2}{C} \;\le\; \tilde{\mathcal{L}}_{\mathrm{MSE}}(\boldsymbol{\ell}'; c) \;\le\; \frac{1 + 2B + CB^2}{C}.}$$

The lower bound is 0 whenever the target logit can reach 1 ($B \ge 1$), and the upper bound grows quadratically with the logit radius B and linearly with C. Thus, MSL has bounded per-sample loss if and only if logits are bounded (via norm control or explicit clipping); otherwise, the loss is unbounded above.

B Implementation details

B.1 Noise Optimizations

For each forget class $f \in \mathcal{Y}$ we independently initialize n, freeze $\Pi_{\boldsymbol{\theta}, \mathbf{W}}$. We optimize $\boldsymbol{\phi}$ using the Adam optimizer, with a decreasing learning rate (lr) via the StepLR schedule, set to a step size of 5 and a decay factor of 0.1, initialized at $lr = 0.01$. The energy regularization weight is λ=0.1. The Algorithm 1 will summarize the unlearning noise training with CE and MSE loss. Figure 3 represents the average angles between $\mathbf{v}$ (normalized noise features $\mathbf{z}$ to forget class 0) and $\mathbf{u}$ of all classes, including forget class, over noise training epochs.

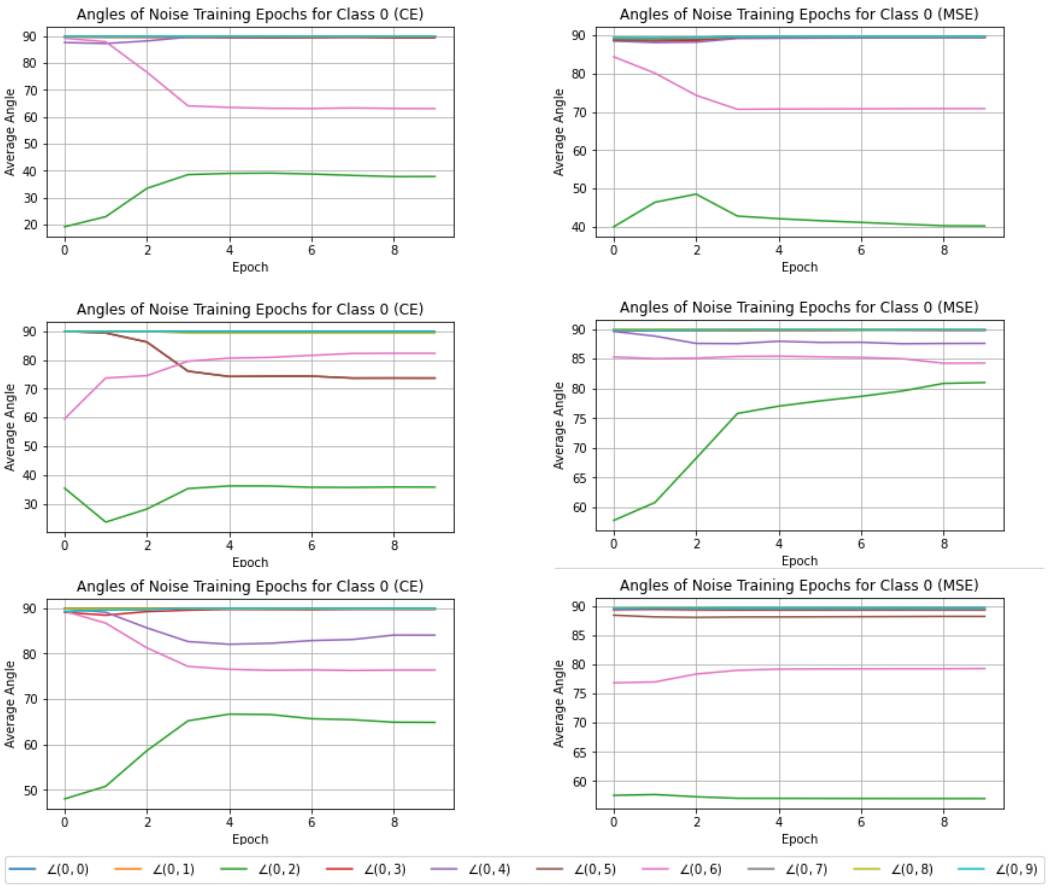


Figure 3: Training unlearning noise results with *(left)* CE loss and *(right)* MSE loss. The $\angle(i, j)$ represents the angle between $\mathbf{v}_i$, the normalized feature vector of class i, and $\mathbf{u}_j$, the normalized prototype weight of class j. Top: regular prototype logit. Middle: cosine similarity logit with trainable prototype layer. Bottom: cosine similarity logit with fixed orthonormal prototype layer.

B.2 Benchmark Methods

For the Retrain method, training is conducted over 150 epochs using the SGD optimizer with a cosine-scheduled learning rate initialized at 0.1. For all the benchmark methods, we do the learning-rate search within $[10^{-5}, 10^{-1}]$ and epoch search within $[1, 20]$. We used the SGD optimizer with a batch size of 128 for all methods for fair comparison. For IU, we explore the parameter α associated with the WoodFisher Hessian inverse approximation within the range $[0.1, 20]$. For ℓ_1-sparse, a search for the parameter γ is executed within $[10^{-6}, 10^{-2}]$. For SalUn, we are searching for sparsity ratios (S) in the $[0.1, 0.9]$ range. The summary of the exact hyperparameter for reproducibility of the results is presented in table Table 3 for Quick unlearning and in table Table 4 for Optimum unlearning.

C Complementary Experiment Results

C.1 Extensive Unlearning Performance Analysis

We evaluate four model constraints and two loss functions across eight baseline methods via wide-ranging experiments in two settings: Quick unlearning and optimum unlearning. The Quick-unlearning results in Table 5 compare classification heads under CE loss, whereas Table 6 reports the corresponding results under MSE loss. As shown in Table 5, ORC consistently outperforms other classification heads in balancing forgetting effectiveness (low UA) and task retention (high

Algorithm 1 Training Unlearning Noise

Input: Pretrained model Π^*, classes to forget $\mathcal{C}_f$, batch size $B = 128$, epochs $E = 10$, steps per epoch $S = 35$, learning rate $\alpha = 0.01$, number of classes C, input channels ch, Loss MSE or CE
Noise Training:
Set model to evaluation mode and Loss to cross-entropy
for each class $c \in \mathcal{C}_{forget}$ **do**
  Initialize $noise_c \leftarrow Noise(B, ch, 32, 32)$ on device
  Initialize optimizer $opt \leftarrow Adam(noise_c,\ \alpha)$
  Initialize scheduler $sched \leftarrow StepLR(opt,\ step\_size = 5,\ \gamma = 0.1)$
  for epoch $e = 1$ to E **do**
    for step $s = 1$ to S **do**
      if Loss CE **then**
        if $s \leq 0.95S$ **then**
          $y \leftarrow \text{RANDINT}([0, C-1], B)$ on device
        else
          $y \leftarrow c \cdot \mathbf{1}_B$
        end if
      else
        $y \leftarrow \mathbf{1}_B$
      end if
      $logits \leftarrow model(x)$
      Compute loss: $\mathcal{L} \leftarrow -CE(logits, y) + 0.1 \cdot \text{mean}(\sum x^2)$
      Update ϕ
    end for
    Update α with $sched.step()$
  end for
  $noises[c] \leftarrow noise_c$
end for
return $noises$

Table 3: Quick unlearning – compact hyperparameter configurations under MSE and CE loss.

	ORC					COS					SFX / MSL				
Method	Unl.	lr	γ	α	S	Unl.	lr	γ	α	S	Unl.	lr	γ	α	S
MSE Loss															
GA	1	0.005	-	-	-	1	5e-4	-	-	-	2	5e-4	-	-	-
FTR	2	0.05	-	-	-	3	0.05	-	-	-	3	0.05	-	-	-
FTN	1	1e-5	-	-	-	1	0.001	-	-	-	1	1e-5	-	-	-
LSPN	1	1e-5	1e-5	-	-	1	1e-5	0.001	-	-	1	1e-5	1e-4	-	-
LSPR	3	0.01	1e-4	-	-	3	0.01	1e-4	-	-	3	5e-5	0.01	-	-
IU	1	-	-	2	-	1	-	-	2	-	1	-	-	2	-
SALUN	3	0.1	-	-	0.9	3	0.005	-	-	0.2	3	0.05	-	-	0.6
	Imp.	Rep.	lr_{imp}	lr_{rep}		Imp.	Rep.	lr_{imp}	lr_{rep}		Imp.	Rep.	lr_{imp}	lr_{rep}	
UNSIR	1	2	0.08	0.02		1	2	0.08	0.01		1	2	1e-4	0.01	
CE Loss															
GA	3	1e-4	-	-	-	1	5e-4	-	-	-	3	5e-5	-	-	-
FTR	3	0.05	-	-	-	2	0.05	-	-	-	2	0.1	-	-	-
FTN	1	1e-5	-	-	-	1	1e-5	-	-	-	1	1e-5	-	-	-
LSPN	1	1e-5	0.01	-	-	1	1e-5	5e-6	-	-	1	1e-4	0.01	-	-
LSPR	3	0.005	0.001	-	-	3	0.001	0.005	-	-	3	0.001	0.01	-	-
IU	1	-	-	1	-	1	0.1	-	1	-	1	-	-	0.2	-
SALUN	1	0.05	-	-	0.5	3	0.005	-	-	0.7	1	0.01	-	-	0.5
	Imp.	Rep.	lr_{imp}	lr_{rep}		Imp.	Rep.	lr_{imp}	lr_{rep}		Imp.	Rep.	lr_{imp}	lr_{rep}	
UNSIR	1	2	0.08	0.01		1	2	0.01	0.005		1	2	0.05	0.02	

RA). GA-ORC-CE achieves the best trade-off with a UA of only 0.28% while maintaining an RA of 95.3%.

Table 4: Optimum unlearning – compact hyperparameter configurations under MSE and CE loss.

	ORC					COS					SFX / MSL				
Method	Unl.	lr	γ	α	S	Unl.	lr	γ	α	S	Unl.	lr	γ	α	S
MSE Loss															
GA	1	0.005	-	-	-	15	1e-4	-	-	-	18	5e-5	-	-	-
FTR	20	0.01	-	-	-	18	0.05	-	-	-	12	0.05	-	-	-
FTN	1	1e-5	-	-	-	4	1e-5	-	-	-	1	1e-5	-	-	-
LSPN	1	1e-5	1e-5	-	-	1	1e-5	0.001	-	-	1	1e-5	1e-4	-	-
LSP	20	0.01	1e-5	-	-	10	0.001	5e-4	-	-	20	1e-5	0.01	-	-
IU	1	-	-	2	-	1	-	-	2	-	1	-	-	2	-
SALUN	8	0.1	-	-	0.4	12	0.005	-	-	0.9	18	0.05	-	-	0.8
	Imp.	Rep.	lr_{imp}	lr_{rep}		Imp.	Rep.	lr_{imp}	lr_{rep}		Imp.	Rep.	lr_{imp}	lr_{rep}	
UNSIR	3	10	5e-4	0.01		1	10	0.08	0.01		1	2	1e-4	0.01	
CE Loss															
GA	20	1e-5	-	-	-	1	5e-4	-	-	-	3	5e-5	-	-	-
FTR	20	0.05	-	-	-	18	0.05	-	-	-	20	0.05	-	-	-
FTN	1	1e-5	-	-	-	1	1e-5	-	-	-	1	1e-5	-	-	-
LSPN	1	1e-5	0.01	-	-	1	1e-5	5e-6	-	-	1	1e-4	0.01	-	-
LSP	8	5e-4	0.005	-	-	12	0.01	1e-4	-	-	15	5e-4	0.005	-	-
IU	1	-	-	1	-	1	0.1	-	1	-	1	-	-	0.2	-
SALUN	15	0.05	-	-	0.5	20	0.01	-	-	0.4	8	0.05	-	-	0.5
	Imp.	Rep.	lr_{imp}	lr_{rep}		Imp.	Rep.	lr_{imp}	lr_{rep}		Imp.	Rep.	lr_{imp}	lr_{rep}	
UNSIR	1	10	0.08	0.01		1	8	0.02	0.01		2	10	0.02	0.02	

Table 6 results indicate that, using MSE loss, COS paired with GA or ORC paired with LSPR maximize retention while maintaining a low forgetting error. Methods such as FTN-MSL-MSE and LSPN-MSL-MSE achieve RA of 96.9% and 97.0%, respectively, but with extremely high UA of 28.4% in both cases, reflecting ineffective unlearning.

Under the longer training schedule of optimum unlearning (Tables 7 and 8), we observe retention gains (RA, TRA) for both loss functions, accompanied by reduced residual memorization (UA, TUA), in many cases. However, LSPR-COS-CE shows signs of over-forgetting (UA $\approx 0\%$) at the cost of markedly reduced RA 57.1% (± 5.99), whereas LSPR-COS-MSE balances the trade-off better (RA 94.9%, UA 0.28%). SALUN exhibits high RA under both losses for COS but with large UA. Under optimum unlearning, the best CE-loss configurations are GA–ORC, LSPR–SFX, and FTR–SFX, achieving low UA and high RA. For MSE-loss, GA–COS and LSPR–ORC likewise rank among the top performers, with low UA and high RA.

Tables 5 to 8 show that moving from Quick to optimum unlearning generally increases retention (RA/TRA) for both losses. Under CE, ORC (and sometimes SFX) tends to raise RA but often increases UA; therefore, when minimizing UA is paramount in CE, some ORC/SFX configurations are better left at the Quick schedule. The gains are most consistent for COS–MSE; ORC–MSE typically boosts RA with a mixed effect on UA, e.g., LSPR–ORC–MSE reaches near-zero UA with RA $> 95\%$ under the optimum schedule. By methods, GA and LSPR frequently benefit from the longer schedule and are often among the strongest performers.

Across the four tables, MIA is near ceiling in the Quick setting for most head–loss–method combinations ($\approx 0.98-1.00$) and typically reaches ≈ 1.00 with reduced variance under optimum unlearning setting. The only consistent outlier is COS-CE in the Quick regime, which shows lower and unstable MIA relative to ORC-CE and SFX-CE; the optimum schedule largely closes this gap. Under MSE, all classification heads exhibit uniformly high MIA in both regimes, with optimum further tightening variability.

Table 9 shows that, after retraining, MSE yields the most stable outcomes: across models, retained accuracy is $\approx 94.5\%$, and the forget set is perfectly excluded (TUA $= 0$). Under CE, ORC (93.9 ± 0.6) and SFX (92.5 ± 0.8) are comparable and also attain TUA $= 0$, whereas COS is lower and unstable (TRA $= 64.9 \pm 37.2$, TUA $= 0.01 \pm 0.03$). Thus, post-retraining differences are driven mainly by the loss: MSE is slightly higher and markedly more stable than CE, while the head choice is largely immaterial under MSE and under CE except for the COS case.

Table 5: Quick unlearning performance results across methods and models, with CE.

Method	Model	Loss	UA(↓)	RA(↑)	TUA(↓)	TRA(↑)	MIA(↑)
FTN	COS	CE	48.1 ± 30.2	47.3 ± 3.62	46.6 ± 30	45.7 ± 3.55	0.657 ± 0.295
FTR	COS	CE	1.05 ± 2.32	70.6 ± 7.61	0.929 ± 2.17	68.5 ± 7.29	0.992 ± 0.019
GA	COS	CE	0.0546 ± 0.078	76.3 ± 3.83	0.029 ± 0.0624	73 ± 3.54	0.996 ± 0.00663
IU	COS	CE	0.018 ± 0.0534	39.4 ± 21	0.013 ± 0.0464	38.2 ± 19.7	0.986 ± 0.063
LSPN	COS	CE	48.2 ± 30.4	47.3 ± 3.62	46.7 ± 30.1	45.7 ± 3.54	0.663 ± 0.287
LSPR	COS	CE	7.96 ± 12.8	84.2 ± 4.63	7.15 ± 11.2	80.6 ± 4.28	0.997 ± 0.00916
SALUN	COS	CE	47.6 ± 38	87.5 ± 7.39	43.3 ± 35	83.6 ± 7.01	0.691 ± 0.346
UNSIR	COS	CE	5.37 ± 5.24	77.4 ± 5.3	4.85 ± 4.74	73.9 ± 5.1	0.981 ± 0.0284
FTN	ORC	CE	27.8 ± 24.9	27.2 ± 3.81	27.1 ± 24.3	26.5 ± 3.79	0.8 ± 0.141
FTR	ORC	CE	0.0698 ± 0.264	75.5 ± 5.82	0.068 ± 0.244	73 ± 5.62	1 ± 0
GA	ORC	CE	0.277 ± 0.771	95.3 ± 1.53	0.193 ± 0.493	89.1 ± 1.73	0.999 ± 0.00289
IU	ORC	CE	0.0638 ± 0.122	46.6 ± 27.6	0.058 ± 0.0945	44.7 ± 25.9	0.929 ± 0.256
LSPN	ORC	CE	28 ± 25	27.4 ± 3.83	27.4 ± 24.4	26.8 ± 3.81	0.796 ± 0.139
LSPR	ORC	CE	0.0858 ± 0.246	88.2 ± 3.25	0.082 ± 0.262	84.2 ± 3.18	1 ± 0
SALUN	ORC	CE	1.71 ± 3.12	68.3 ± 8.57	1.65 ± 3.02	66.4 ± 8.3	0.989 ± 0.0233
UNSIR	ORC	CE	0.607 ± 0.694	89.2 ± 1.73	0.578 ± 0.672	84.7 ± 1.81	1 ± 0.000811
FTN	SFX	CE	20 ± 28.7	18.2 ± 3.47	19.7 ± 28.5	17.9 ± 3.42	0.786 ± 0.178
FTR	SFX	CE	0 ± 0	76.3 ± 5.05	0 ± 0	73.4 ± 4.83	1 ± 0
GA	SFX	CE	12.6 ± 1.65	92 ± 2.78	12 ± 2.1	85.4 ± 2.67	0.882 ± 0.0207
IU	SFX	CE	11.6 ± 14.4	88.1 ± 12.8	11.4 ± 13.5	81.9 ± 11	0.912 ± 0.112
LSPN	SFX	CE	20.2 ± 28.7	18.5 ± 3.49	19.8 ± 28.5	18.1 ± 3.44	0.787 ± 0.178
LSPR	SFX	CE	2.65 ± 2.83	94.4 ± 0.805	2.55 ± 2.89	89.3 ± 1.03	1 ± 0.000374
SALUN	SFX	CE	25.6 ± 23.9	98.8 ± 0.456	22.7 ± 19.6	92.2 ± 0.938	0.923 ± 0.145
UNSIR	SFX	CE	0.46 ± 0.967	89 ± 1.83	0.435 ± 0.85	84 ± 1.87	0.999 ± 0.00358

Table 6: Quick unlearning performance results across methods and models, with MSE.

Method	Model	Loss	UA(↓)	RA(↑)	TUA(↓)	TRA(↑)	MIA(↑)
FTN	COS	MSE	63 ± 7.15	97.9 ± 1.2	57.4 ± 8.26	91.9 ± 1.64	0.669 ± 0.0851
FTR	COS	MSE	1.33 ± 1.17	84.9 ± 2.84	1.34 ± 1.12	81.2 ± 2.84	0.997 ± 0.00416
GA	COS	MSE	0.141 ± 0.18	96.3 ± 1.26	0.15 ± 0.173	90.6 ± 1.55	1 ± 0.000706
IU	COS	MSE	4.37 ± 8.82	93.5 ± 4.16	4.31 ± 8.83	87.8 ± 3.34	0.987 ± 0.0314
LSPN	COS	MSE	60.6 ± 5.32	98.1 ± 0.753	55.2 ± 6.72	92.1 ± 1.22	0.681 ± 0.085
LSPR	COS	MSE	0.904 ± 1.42	94.3 ± 2.57	1.01 ± 1.66	88.8 ± 2.6	1 ± 0
SALUN	COS	MSE	50.4 ± 10.4	97.3 ± 1.23	46 ± 8.74	91.2 ± 1.64	0.752 ± 0.22
UNSIR	COS	MSE	1.07 ± 1.07	89.8 ± 1.79	1.05 ± 1.1	85.8 ± 1.82	0.999 ± 0.00269
FTN	ORC	MSE	98.5 ± 0.88	73.3 ± 15.6	96.1 ± 1.98	66.5 ± 13.9	0.807 ± 0.136
FTR	ORC	MSE	1.05 ± 0.833	85.2 ± 2.45	1.02 ± 0.947	81.7 ± 2.48	0.998 ± 0.00191
GA	ORC	MSE	0 ± 0	81.6 ± 7.85	0 ± 0	75.2 ± 7.42	1 ± 0
IU	ORC	MSE	2.23 ± 2.89	91.9 ± 4.87	1.96 ± 2.46	86.6 ± 4.14	0.998 ± 0.00709
LSPN	ORC	MSE	78.4 ± 4.79	96.2 ± 3.2	71.7 ± 6.91	90 ± 3.57	0.857 ± 0.0782
LSPR	ORC	MSE	0.121 ± 0.25	94.3 ± 1.89	0.076 ± 0.232	88.9 ± 1.95	1 ± 0
SALUN	ORC	MSE	15.6 ± 3.96	77 ± 4.4	15.6 ± 3.99	74.3 ± 4.26	0.883 ± 0.108
UNSIR	ORC	MSE	0.857 ± 1.06	91 ± 1.79	0.897 ± 1.17	86.7 ± 1.93	0.999 ± 0.00137
FTN	SFX	MSE	28.4 ± 2.35	96.9 ± 0.767	26.6 ± 2.88	90.6 ± 1.03	0.393 ± 0.0198
FTR	SFX	MSE	0 ± 0	60.3 ± 9.23	0 ± 0	59.2 ± 8.68	1 ± 0
GA	SFX	MSE	3.96 ± 1.82	30.5 ± 5.09	4.05 ± 2.11	30.3 ± 4.99	0.404 ± 0.0273
IU	SFX	MSE	39 ± 4.13	7.84 ± 2.04	39.1 ± 4.51	7.85 ± 2.04	0.765 ± 0.312
LSPN	SFX	MSE	28.4 ± 2.26	97 ± 0.793	26.7 ± 2.88	90.6 ± 1.03	0.393 ± 0.0196
LSPR	SFX	MSE	10.3 ± 8.74	94.8 ± 2.2	9.49 ± 7.37	88.4 ± 2.42	1 ± 0.000369
SALUN	SFX	MSE	24 ± 4.15	35.7 ± 9.98	23.9 ± 4.14	35.5 ± 9.83	0.609 ± 0.204
UNSIR	SFX	MSE	0.67 ± 1	90.9 ± 2.99	0.727 ± 1.11	85.9 ± 2.77	1 ± 0.000237
FTN	MSL	MSE	28.4 ± 2.35	96.9 ± 0.767	26.6 ± 2.88	90.6 ± 1.03	0.393 ± 0.0198
FTR	MSL	MSE	0 ± 0	60.3 ± 9.23	0 ± 0	59.2 ± 8.68	1 ± 0
GA	MSL	MSE	3.96 ± 1.82	30.5 ± 5.09	4.05 ± 2.11	30.3 ± 4.99	0.404 ± 0.0273
IU	MSL	MSE	39 ± 4.13	7.84 ± 2.04	39.1 ± 4.51	7.85 ± 2.04	0.765 ± 0.312
LSPN	MSL	MSE	28.4 ± 2.26	97 ± 0.793	26.7 ± 2.88	90.6 ± 1.03	0.393 ± 0.0196
LSPR	MSL	MSE	10.3 ± 8.74	94.8 ± 2.2	9.49 ± 7.37	88.4 ± 2.42	1 ± 0.000369
SALUN	MSL	MSE	24 ± 4.15	35.7 ± 9.98	23.9 ± 4.14	35.5 ± 9.83	0.609 ± 0.204
UNSIR	MSL	MSE	0.67 ± 1	90.9 ± 2.99	0.727 ± 1.11	85.9 ± 2.77	1 ± 0.000237

C.2 IDK Detection

Tables 10 to 13 present IDK-detection performance across models and methods for (i) Quick–CE, (ii) Quick–MSE, (iii) optimum–CE, and (iv) optimum–MSE, respectively.

Table 7: Optimum unlearning performance results across methods and models, with CE.

Method	Model	Loss	UA(↓)	RA(↑)	TUA(↓)	TRA(↑)	MIA(↑)
FTN	COS	CE	48.2 ± 30.1	47.2 ± 3.59	46.7 ± 29.8	45.7 ± 3.51	0.663 ± 0.286
FTR	COS	CE	0.131 ± 0.365	76.9 ± 4.5	0.127 ± 0.338	73.9 ± 4.33	1 ± 0.00179
GA	COS	CE	0.0546 ± 0.078	76.3 ± 3.83	0.029 ± 0.0624	73 ± 3.54	0.996 ± 0.00663
IU	COS	CE	0.018 ± 0.0534	39.4 ± 21	0.013 ± 0.0464	38.2 ± 19.7	0.986 ± 0.063
LSPN	COS	CE	48 ± 30.2	47.3 ± 3.61	46.6 ± 29.9	45.7 ± 3.53	0.654 ± 0.293
LSPR	COS	CE	0.001 ± 0.00595	57.1 ± 5.99	0.001 ± 0.01	55.7 ± 5.8	1 ± 0
SALUN	COS	CE	34.7 ± 31.1	85.5 ± 6.23	32.1 ± 29	81.8 ± 5.97	0.763 ± 0.287
UNSIR	COS	CE	2.44 ± 3.03	76.7 ± 4.6	2.17 ± 2.71	73.3 ± 4.32	0.994 ± 0.013
FTN	ORC	CE	29.3 ± 25.3	28.6 ± 4.59	28.7 ± 24.8	28.1 ± 4.56	0.786 ± 0.158
FTR	ORC	CE	0.0014 ± 0.00513	85.8 ± 2.71	0 ± 0	81.8 ± 2.8	1 ± 0
GA	ORC	CE	0.647 ± 1.71	95.6 ± 1.35	0.443 ± 1.15	89.5 ± 1.61	0.998 ± 0.00663
IU	ORC	CE	0.0502 ± 0.0991	46.5 ± 27.7	0.047 ± 0.0731	44.6 ± 26	0.986 ± 0.0448
LSPN	ORC	CE	29.4 ± 25.4	28.8 ± 4.6	28.8 ± 24.9	28.2 ± 4.58	0.786 ± 0.141
LSPR	ORC	CE	1.41 ± 2.52	95.8 ± 0.493	1.3 ± 2.51	90.6 ± 0.791	1 ± 0
SALUN	ORC	CE	7.03 ± 8.11	81.6 ± 4.19	6.89 ± 8.05	78.4 ± 3.97	0.971 ± 0.0478
UNSIR	ORC	CE	0.607 ± 0.694	89.2 ± 1.73	0.578 ± 0.672	84.7 ± 1.81	1 ± 0.000811
FTN	SFX	CE	18.9 ± 27.5	18.5 ± 3.2	18.6 ± 27.2	18.1 ± 3.15	0.793 ± 0.162
FTR	SFX	CE	0.446 ± 0.531	94.6 ± 0.732	0.44 ± 0.557	90.3 ± 1.05	1 ± 0.000148
GA	SFX	CE	12.6 ± 1.65	92 ± 2.78	12 ± 2.1	85.4 ± 2.67	0.882 ± 0.0207
IU	SFX	CE	11.6 ± 14.4	88.1 ± 12.8	11.4 ± 13.5	81.9 ± 11	0.912 ± 0.112
LSPN	SFX	CE	22 ± 28.9	20.1 ± 3.59	21.5 ± 28.6	19.6 ± 3.54	0.782 ± 0.179
LSPR	SFX	CE	0.865 ± 1.27	96.4 ± 0.367	0.794 ± 1.21	90.8 ± 0.738	1 ± 0
SALUN	SFX	CE	1.64 ± 3.95	83.3 ± 3.68	1.57 ± 3.7	79.7 ± 3.52	0.99 ± 0.0268
UNSIR	SFX	CE	1.44 ± 1.83	93.1 ± 1.47	1.43 ± 1.72	86.8 ± 1.59	0.998 ± 0.00439

Table 8: Optimum unlearning performance results across methods and models, with MSE.

Method	Model	Loss	UA(↓)	RA(↑)	TUA(↓)	TRA(↑)	MIA(↑)
FTN	COS	MSE	60.5 ± 6.1	98.1 ± 0.797	54.8 ± 7.34	92.1 ± 1.25	0.678 ± 0.089
FTR	COS	MSE	0.0301 ± 0.0567	85.3 ± 2.19	0.0209 ± 0.0723	81.4 ± 2.23	$1 \pm 2.1e-05$
GA	COS	MSE	0.00527 ± 0.0119	95.6 ± 1.55	0.0198 ± 0.0401	90.2 ± 1.98	$1 \pm 4.12e-05$
IU	COS	MSE	4.87 ± 9.25	93.3 ± 4.23	4.83 ± 9.28	87.7 ± 3.4	0.985 ± 0.0329
LSPN	COS	MSE	60.3 ± 6.04	98.1 ± 0.828	54.7 ± 7.27	92.1 ± 1.28	0.679 ± 0.0896
LSPR	COS	MSE	0.28 ± 0.7	94.9 ± 1.57	0.327 ± 0.832	89.6 ± 1.75	1 ± 0
SALUN	COS	MSE	50.6 ± 12.7	97.2 ± 3.01	45.7 ± 10.8	91.2 ± 3.12	0.728 ± 0.214
UNSIR	COS	MSE	2.01 ± 1.92	91.4 ± 1.39	2.05 ± 2	86.9 ± 1.55	0.998 ± 0.00415
FTN	ORC	MSE	77.8 ± 4.85	96.2 ± 3.15	71.2 ± 7.03	90.1 ± 3.51	0.854 ± 0.0785
FTR	ORC	MSE	2.21 ± 3.41	95.1 ± 1.87	2.3 ± 3.5	89.1 ± 2.03	1 ± 0
GA	ORC	MSE	0 ± 0	91.3 ± 7.21	0 ± 0	85.2 ± 7.23	1 ± 0
IU	ORC	MSE	2.25 ± 2.9	92 ± 4.77	2 ± 2.48	86.6 ± 4.06	0.998 ± 0.00702
LSPN	ORC	MSE	78 ± 4.58	96.3 ± 3.01	71.4 ± 6.8	90.1 ± 3.38	0.855 ± 0.0786
LSPR	ORC	MSE	0.001 ± 0.00659	95.9 ± 1.26	0 ± 0	90.3 ± 1.44	1 ± 0
SALUN	ORC	MSE	15.1 ± 3.08	78.4 ± 4.42	15.1 ± 3.4	75.1 ± 4.34	0.895 ± 0.0317
UNSIR	ORC	MSE	2.4 ± 2.71	92.4 ± 1.74	2.42 ± 2.85	87.4 ± 1.91	1 ± 0.000117
FTN	SFX	MSE	28 ± 2.16	96.9 ± 0.817	26.3 ± 2.85	90.5 ± 1.04	0.998 ± 0.00208
FTR	SFX	MSE	0 ± 0	86.7 ± 2.36	0 ± 0	82.4 ± 2.46	1 ± 0
GA	SFX	MSE	0.044 ± 0.0374	95.3 ± 1.3	0.0386 ± 0.0721	89.2 ± 1.35	1 ± 0
IU	SFX	MSE	1.73 ± 1.17	91.8 ± 3.76	1.57 ± 1.08	86.3 ± 2.82	$1 \pm 9.31e-05$
LSPN	SFX	MSE	28 ± 2.43	96.9 ± 0.793	26.2 ± 3.05	90.5 ± 1.03	0.998 ± 0.00206
LSPR	SFX	MSE	13.9 ± 4.28	95.8 ± 1.55	13.2 ± 4.3	89.9 ± 1.53	1 ± 0
SALUN	SFX	MSE	12.8 ± 2	86.3 ± 2.6	12.7 ± 2.29	82 ± 2.66	0.933 ± 0.0153
UNSIR	SFX	MSE	0.161 ± 0.366	92.8 ± 1.73	0.184 ± 0.407	87.2 ± 1.89	1 ± 0
FTR	MSL	MSE	0 ± 0	86.7 ± 2.36	0 ± 0	82.4 ± 2.46	1 ± 0
GA	MSL	MSE	0.044 ± 0.0374	95.3 ± 1.3	0.0386 ± 0.0721	89.2 ± 1.35	1 ± 0
IU	MSL	MSE	1.73 ± 1.17	91.8 ± 3.76	1.57 ± 1.08	86.3 ± 2.82	$1 \pm 9.31e-05$
LSPN	MSL	MSE	28 ± 2.43	96.9 ± 0.793	26.2 ± 3.05	90.5 ± 1.03	0.998 ± 0.00206
LSPR	MSL	MSE	13.9 ± 4.28	95.8 ± 1.55	13.2 ± 4.3	89.9 ± 1.53	1 ± 0
SALUN	MSL	MSE	12.8 ± 2	86.3 ± 2.6	12.7 ± 2.29	82 ± 2.66	0.933 ± 0.0153
UNSIR	MSL	MSE	0.161 ± 0.366	92.8 ± 1.73	0.184 ± 0.407	87.2 ± 1.89	1 ± 0

IDK detection is loss-driven: MSE consistently delivers higher AUC and lower FPR@TPR90 and DetErr@TPR90 than CE for all models and methods. Switching from Quick to optimum yields modest, consistent gains mainly under MSE (higher AUC, lower FPR/DetErr); CE changes little and optimization does not close its gap to MSE. COS and ORC benefit similarly from MSE, with

Table 9: Retrain performance across models, grouped by loss functions.

Loss	Model	Retrain TRA(↑)	Retrain TUA(↓)
CE	COS	64.9 ± 37.200	0.01 ± 0.0316
	ORC	93.9 ± 0.627	0 ± 0
	SFX	92.5 ± 0.804	0 ± 0
MSE	COS	94.6 ± 0.561	0 ± 0
	ORC	94.5 ± 0.634	0 ± 0
	SFX	94.5 ± 0.613	0 ± 0
	MSL	94.5 ± 0.557	0 ± 0

Table 10: IDK detection performance under Quick unlearning across methods and models, with CE.

Method	Model	Loss	AUC(↑)	FPR@TPR90(↓)	Thr@TPR90	DetErr@TPR90(↓)
FTR	COS	CE	0.625 ± 0.136	0.687 ± 0.17	1.77 ± 0.698	0.393 ± 0.085
GA	COS	CE	0.769 ± 0.0449	0.523 ± 0.0807	2.35 ± 0.432	0.312 ± 0.0404
IU	COS	CE	0.479 ± 0.207	0.789 ± 0.207	3.05 ± 2.99	0.444 ± 0.104
LSPR	COS	CE	0.783 ± 0.171	0.435 ± 0.267	9.25 ± 4.78	0.268 ± 0.134
SALUN	COS	CE	0.728 ± 0.206	0.485 ± 0.294	2.22 ± 1.11	0.292 ± 0.147
UNSIR	COS	CE	0.818 ± 0.0566	0.404 ± 0.0944	3.22 ± 2.66	0.252 ± 0.0472
FTR	ORC	CE	0.654 ± 0.091	0.65 ± 0.112	2.22 ± 0.607	0.375 ± 0.0558
GA	ORC	CE	0.89 ± 0.0212	0.219 ± 0.0678	10.1 ± 3.25	0.159 ± 0.0339
IU	ORC	CE	0.416 ± 0.167	0.869 ± 0.124	2.4 ± 1.81	0.485 ± 0.0619
LSPR	ORC	CE	0.773 ± 0.064	0.465 ± 0.113	4.24 ± 1.16	0.282 ± 0.0565
SALUN	ORC	CE	0.619 ± 0.115	0.685 ± 0.123	4.04 ± 1.37	0.393 ± 0.0613
UNSIR	ORC	CE	0.813 ± 0.0243	0.397 ± 0.0658	5.41 ± 1.96	0.248 ± 0.0328
FTR	SFX	CE	0.573 ± 0.107	0.74 ± 0.111	15.9 ± 2.69	0.42 ± 0.0557
GA	SFX	CE	0.914 ± 0.0107	0.255 ± 0.0361	26.5 ± 0.981	0.178 ± 0.0181
IU	SFX	CE	0.654 ± 0.189	0.651 ± 0.176	15.9 ± 4.16	0.376 ± 0.088
LSPR	SFX	CE	0.694 ± 0.0956	0.543 ± 0.136	18.1 ± 2.51	0.322 ± 0.068
SALUN	SFX	CE	0.755 ± 0.0511	0.465 ± 0.0847	18 ± 1.65	0.282 ± 0.0423
UNSIR	SFX	CE	0.862 ± 0.0117	0.341 ± 0.0253	25.7 ± 1.13	0.221 ± 0.0127

Table 11: IDK detection performance under Quick unlearning across methods and models, with MSE.

Method	Model	Loss	AUC(↑)	FPR@TPR90(↓)	Thr@TPR90	DetErr@TPR90(↓)
FTR	COS	MSE	0.824 ± 0.0224	0.402 ± 0.0471	10.7 ± 1.26	0.251 ± 0.0235
GA	COS	MSE	0.822 ± 0.0314	0.392 ± 0.0943	6.85 ± 0.733	0.246 ± 0.0472
IU	COS	MSE	0.794 ± 0.0862	0.513 ± 0.208	13.7 ± 5.75	0.306 ± 0.104
LSPR	COS	MSE	0.876 ± 0.12	0.268 ± 0.138	28.3 ± 6.38	0.184 ± 0.069
SALUN	COS	MSE	0.826 ± 0.0604	0.436 ± 0.192	10.8 ± 3.45	0.268 ± 0.096
UNSIR	COS	MSE	0.837 ± 0.0572	0.345 ± 0.0753	10.9 ± 3.72	0.222 ± 0.0376
FTR	ORC	MSE	0.82 ± 0.0174	0.386 ± 0.0377	5.24 ± 0.576	0.243 ± 0.0188
GA	ORC	MSE	0.2 ± 0.0858	0.961 ± 0.0599	4.17 ± 0.486	0.531 ± 0.03
IU	ORC	MSE	0.734 ± 0.0754	0.604 ± 0.139	8.55 ± 2.89	0.352 ± 0.0696
LSPR	ORC	MSE	0.875 ± 0.0296	0.218 ± 0.0323	21.3 ± 1.82	0.159 ± 0.0162
SALUN	ORC	MSE	0.774 ± 0.0295	0.481 ± 0.0506	7.25 ± 1.71	0.291 ± 0.0253
UNSIR	ORC	MSE	0.855 ± 0.0139	0.305 ± 0.0258	5.71 ± 1.51	0.203 ± 0.0129
FTR	SFX	MSE	0.806 ± 0.0124	0.424 ± 0.0301	13.5 ± 1.24	0.262 ± 0.0151
GA	SFX	MSE	0.912 ± 0.0098	0.23 ± 0.024	11.3 ± 1.29	0.165 ± 0.012
IU	SFX	MSE	0.814 ± 0.0343	0.413 ± 0.0832	9.98 ± 1.29	0.257 ± 0.0416
LSPR	SFX	MSE	0.921 ± 0.00817	0.214 ± 0.0164	13.2 ± 1.24	0.157 ± 0.00822
SALUN	SFX	MSE	0.832 ± 0.0196	0.4 ± 0.0322	15.1 ± 1.57	0.25 ± 0.0161
UNSIR	SFX	MSE	0.845 ± 0.111	0.337 ± 0.132	12.3 ± 5.9	0.219 ± 0.0662
FTR	MSL	MSE	0.671 ± 0.059	0.644 ± 0.089	14 ± 2.65	0.372 ± 0.0445
GA	MSL	MSE	0.68 ± 0.0597	0.736 ± 0.0869	9.41 ± 2.8	0.418 ± 0.0434
IU	MSL	MSE	0.561 ± 0.0977	0.861 ± 0.0759	5.06 ± 2.64	0.48 ± 0.038
LSPR	MSL	MSE	0.513 ± 0.0661	0.806 ± 0.0742	4.51 ± 0.658	0.453 ± 0.0371
SALUN	MSL	MSE	0.641 ± 0.0622	0.722 ± 0.0822	11.4 ± 2.38	0.411 ± 0.0411
UNSIR	MSL	MSE	0.667 ± 0.0763	0.738 ± 0.0714	9.21 ± 9.3	0.419 ± 0.0357

ORC sometimes gaining a small extra FPR drop in the optimum step. Within CE, SFX is usually strongest, whereas with MSE the MSL variant typically matches or exceeds SFX after optimization. In the Quick regime, SFX–MSE clearly outperforms MSL–MSE for most methods (notably GA, IU, SALUN, UNSIR), whereas under Optimum they are largely comparable. Method-wise, UNSIR and

Table 12: IDK detection performance under optimum unlearning across methods and models, with CE.

Method	Model	Loss	AUC(↑)	FPR@TPR90(↓)	Thr@TPR90	DetErr@TPR90(↓)
FTR	COS	CE	0.685 ± 0.0986	0.63 ± 0.159	1.45 ± 0.444	0.365 ± 0.0794
GA	COS	CE	0.769 ± 0.0449	0.523 ± 0.0807	2.35 ± 0.432	0.312 ± 0.0404
IU	COS	CE	0.479 ± 0.207	0.789 ± 0.207	3.05 ± 2.99	0.444 ± 0.104
LSPR	COS	CE	0.574 ± 0.0956	0.77 ± 0.105	2.31 ± 1.2	0.435 ± 0.0524
SALUN	COS	CE	0.762 ± 0.197	0.425 ± 0.299	2.56 ± 1.21	0.262 ± 0.15
UNSIR	COS	CE	0.806 ± 0.0516	0.436 ± 0.0993	2.91 ± 1.3	0.268 ± 0.0497
FTR	ORC	CE	0.723 ± 0.0694	0.569 ± 0.119	1.88 ± 0.346	0.335 ± 0.0593
GA	ORC	CE	0.91 ± 0.0197	0.214 ± 0.0647	8.32 ± 2.68	0.157 ± 0.0323
IU	ORC	CE	0.415 ± 0.167	0.869 ± 0.124	2.41 ± 1.83	0.485 ± 0.0619
LSPR	ORC	CE	0.864 ± 0.0778	0.309 ± 0.15	8.36 ± 2.83	0.205 ± 0.0748
SALUN	ORC	CE	0.73 ± 0.0909	0.566 ± 0.135	4.38 ± 1.55	0.333 ± 0.0677
UNSIR	ORC	CE	0.808 ± 0.0321	0.422 ± 0.0904	3.78 ± 1.55	0.261 ± 0.0452
FTR	SFX	CE	0.738 ± 0.0554	0.474 ± 0.0917	17.2 ± 1.84	0.287 ± 0.0458
GA	SFX	CE	0.914 ± 0.0107	0.255 ± 0.0361	26.5 ± 0.981	0.178 ± 0.0181
IU	SFX	CE	0.654 ± 0.189	0.651 ± 0.176	15.9 ± 4.16	0.376 ± 0.088
LSPR	SFX	CE	0.765 ± 0.0574	0.433 ± 0.103	18.9 ± 2.4	0.267 ± 0.0516
SALUN	SFX	CE	0.628 ± 0.11	0.674 ± 0.137	16.1 ± 2.76	0.387 ± 0.0683
UNSIR	SFX	CE	0.875 ± 0.0124	0.304 ± 0.0309	26.2 ± 1.29	0.202 ± 0.0154

Table 13: IDK detection performance under optimum unlearning across methods and models, with MSE.

Method	Model	Loss	AUC(↑)	FPR@TPR90(↓)	Thr@TPR90	DetErr@TPR90(↓)
FTR	COS	MSE	0.82 ± 0.0169	0.391 ± 0.0336	7.47 ± 0.647	0.246 ± 0.0168
GA	COS	MSE	0.81 ± 0.0372	0.384 ± 0.0449	6.44 ± 0.617	0.242 ± 0.0224
IU	COS	MSE	0.807 ± 0.0813	0.475 ± 0.186	14.5 ± 5.57	0.288 ± 0.093
LSPR	COS	MSE	0.912 ± 0.0378	0.218 ± 0.0555	40.7 ± 4.85	0.159 ± 0.0278
SALUN	COS	MSE	0.836 ± 0.0655	0.408 ± 0.191	11.5 ± 3.98	0.254 ± 0.0953
UNSIR	COS	MSE	0.861 ± 0.0226	0.314 ± 0.0408	8.69 ± 2.65	0.207 ± 0.0204
FTR	ORC	MSE	0.815 ± 0.0647	0.247 ± 0.0757	16.2 ± 1.76	0.174 ± 0.0379
GA	ORC	MSE	0.528 ± 0.141	0.687 ± 0.132	4.49 ± 1.17	0.394 ± 0.0662
IU	ORC	MSE	0.735 ± 0.0741	0.603 ± 0.138	8.56 ± 2.86	0.351 ± 0.0688
LSPR	ORC	MSE	0.853 ± 0.0343	0.217 ± 0.0384	15.5 ± 1.2	0.158 ± 0.0192
SALUN	ORC	MSE	0.775 ± 0.0345	0.479 ± 0.0613	5.84 ± 1.15	0.289 ± 0.0307
UNSIR	ORC	MSE	0.893 ± 0.0228	0.248 ± 0.0368	11.6 ± 2.12	0.174 ± 0.0184
FTR	SFX	MSE	0.813 ± 0.0155	0.406 ± 0.033	11.5 ± 1.21	0.253 ± 0.0165
GA	SFX	MSE	0.865 ± 0.0141	0.289 ± 0.0327	8.08 ± 0.473	0.195 ± 0.0164
IU	SFX	MSE	0.814 ± 0.0341	0.412 ± 0.0831	9.98 ± 1.32	0.256 ± 0.0416
LSPR	SFX	MSE	0.839 ± 0.0149	0.274 ± 0.0216	14.4 ± 0.66	0.187 ± 0.0108
SALUN	SFX	MSE	0.836 ± 0.0169	0.377 ± 0.0322	12.4 ± 1.63	0.238 ± 0.0161
UNSIR	SFX	MSE	0.864 ± 0.0117	0.304 ± 0.0247	9.54 ± 1.06	0.202 ± 0.0124
FTR	MSL	MSE	0.813 ± 0.0155	0.406 ± 0.033	11.5 ± 1.21	0.253 ± 0.0165
GA	MSL	MSE	0.863 ± 0.018	0.293 ± 0.0469	8.08 ± 0.476	0.197 ± 0.0234
IU	MSL	MSE	0.814 ± 0.0341	0.412 ± 0.0831	9.98 ± 1.32	0.256 ± 0.0416
LSPR	MSL	MSE	0.839 ± 0.0149	0.274 ± 0.0216	14.4 ± 0.66	0.187 ± 0.0108
SALUN	MSL	MSE	0.836 ± 0.0169	0.377 ± 0.0322	12.4 ± 1.63	0.238 ± 0.0161
UNSIR	MSL	MSE	0.715 ± 0.0813	0.644 ± 0.102	7.76 ± 1.4	0.372 ± 0.0511

LSPR are top-tier under MSE in both regimes. Thr@TPR90 reflects score scaling and should be compared only within the same loss. Overall, optimum unlearning modestly improves robustness relative to Quick unlearning, but the decisive factor is the loss: MSE outperforms CE in both accuracy and stability. Also, comparing the performance of MSL and SFX models reveals that adding a softmax layer generally improves stability and detection, particularly in the Quick setup, with accompanying reductions in FPR and detection error.

Figure 4 visualizes the unlearning–retention trade-off by plotting TUA versus TRA with AUC encoded as bubble size; color denotes method, marker denotes model variant, and fill denotes loss. SALUN achieves high TUA and TRA having a large bubble. COS tends to preserve retention better than MSL, and ORC is more balanced. GA–COS–MSE offers the best balance between TRA and TUA. Overall, the figure summarizes how method, classifier head, and loss jointly shape the unlearning–retention trade-off.

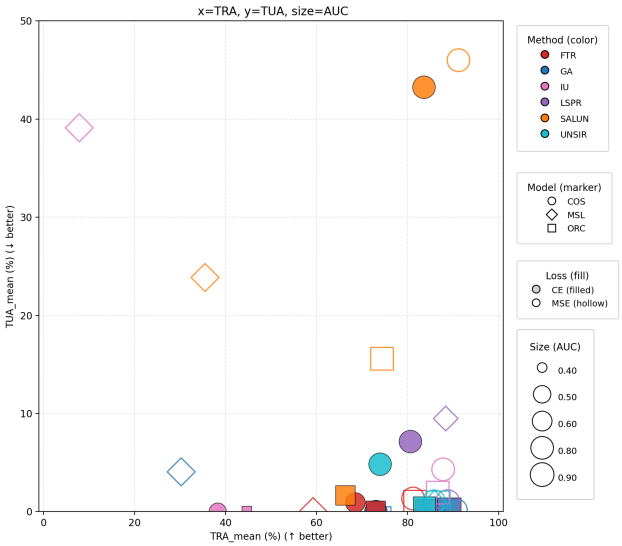


Figure 4: Merged view of IDK detection and test unlearning/retention under the optimum setting, illustrating forgetting effectiveness versus retained-data performance (the trade-off between unlearning accuracy and retention fidelity).