Do Vision Models Truly Forget? New Findings from Representation-Level Certification of Visual Unlearning in Vertical Federated Learning

Zhenyu Yu[1,†], Yangchen Zeng[2,†], Chunlei Meng[3], Guangzhen Yao[4], and Shuigeng Zhou[1,*]

[1] College of Computer Science and Artificial Intelligence, Fudan University
[2] School of Cyber Science and Engineering, Southeast University
[3] College of Intelligent Robotics and Advanced Manufacturing, Fudan University
[4] School of Information Science and Technology, Northeast Normal University
zhenyuyu@fudan.edu.cn; zeng.yc@bytedance.com; clmeng23@m.fudan.edu.cn; guangzhenyao@163.com; sgzhou@fudan.edu.cn
[*] Corresponding author [†] Co-first author

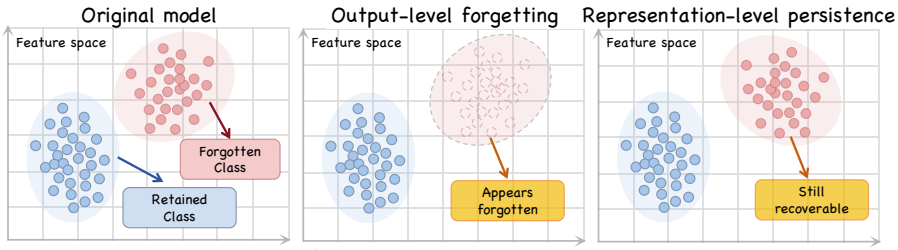

Fig. 1: Mirage reveals the illusion of forgetting. Suppressing classifier predictions may create the appearance of successful unlearning (middle), while the underlying feature geometry remains largely unchanged. Consequently, a linear probe can still recover forgotten-label information with high accuracy (right). This mismatch between behavioral suppression and representational persistence forms the *forgetting illusion*.

Abstract. Machine unlearning in Vertical Federated Learning (VFL) has attracted growing interest, yet existing methods certify forgetting solely using output-level metrics. We challenge these works by introducing **Mirage**, a representation-level auditing framework built from four complementary diagnostics. Mirage combines linear probe recovery (LPR), centered kernel alignment (CKA), feature separability scoring, and layer-wise recovery analysis to assess what a representation actually retains. Extensive experiments across seven datasets and seven baseline methods following recent VFL unlearning protocols reveal three key findings. *(1) Forgetting gap*: methods that pass output-level certification still retain substantial class structure in their representations, with LPR exceeding the retrained baseline by up to 15.4 points. CKA shows that these models remain structurally closer to the original than to the retrained reference, and separability scores indicate persistent geometric discrimination. *(2) Unlearning trilemma*: no existing method simultane-

ously achieves high utility, output-level forgetting, and representation-level forgetting. *(3) Class-sample asymmetry*: class-level forgetting leaves strong representational traces (LPR exceeding 96% on several datasets), whereas sample-level forgetting is indistinguishable from chance (LPR $\approx 50\%$). Layer-wise analysis further shows that residual class information persists across network depths. These findings call for representation-aware evaluation standards in federated unlearning research. Code is publicly available at https://github.com/YuZhenyuLindy/Mirage.

Keywords: Machine Unlearning · Vertical Federated Learning · Representation Auditing · Privacy

1 Introduction

Modern vision models are increasingly deployed in collaborative and regulated environments where selective forgetting is required [27, 28, 33, 41]. In medical imaging consortia, cross-organization feature pipelines, and large-scale model serving systems, intermediate representations are frequently shared, stored, or reused. Legal requirements such as the right to be forgotten have therefore motivated *visual unlearning*, which ensures that a trained model behaves as if designated classes or samples had never been observed [40]. Current evaluations focus almost exclusively on output behavior, including retained-data accuracy, forgotten-label accuracy, and computational overhead [7, 10, 19, 24, 40, 42]. If the model produces uninformative predictions on the forgotten target while maintaining utility, forgetting is deemed complete.

This evaluation protocol assumes that suppressing classifier predictions implies erasing the underlying information [33]. For vision models, this assumption is problematic. Deep networks encode class identity in high-dimensional feature geometries [28] where class-conditional subspaces remain separable across training variations. Modifying the classifier head does not guarantee that these internal structures collapse. A model may therefore satisfy behavioral forgetting criteria while preserving linearly recoverable class information in its intermediate embeddings [15, 17, 18, 20, 42].

The distinction becomes critical in settings such as Vertical Federated Learning (VFL) [35], where multiple passive parties compute and transmit intermediate embeddings to an active party holding labels. In this scenario, representations are explicitly exposed and can be analyzed independently of the classifier head. Yet existing vertical federated unlearning methods largely inherit output-level certification from federated learning literature [5, 34, 37], leaving the internal representation space effectively unaudited.

To examine whether visual models truly forget at the representational level, in this paper we introduce **Mirage**, a post-hoc auditing framework that probes recoverability, structural alignment, feature separability, and layer-wise information persistence. Mirage evaluates unlearning relative to a retrained-from-scratch baseline and formalizes the *forgetting gap*, which measures excess recoverability beyond what intrinsic class geometry would induce. This retraining-relative

perspective enables principled certification of representation-level erasure rather than mere output suppression. Our main **contributions** are:

- We introduce **Mirage**, a representation-level auditing framework that evaluates visual unlearning through complementary analyses of recoverability, structural alignment, and layer-wise information persistence.
- We formalize the **forgetting gap**, a retraining-relative criterion that quantifies residual representation structure beyond intrinsic class geometry.
- Through extensive empirical analysis using all four diagnostics, we reveal two systematic phenomena in visual unlearning: a misalignment between output-level forgetting and representation-level erasure, where Centered Kernel Alignment (CKA) and separability scores corroborate the Linear Probe Recovery (LPR)-based forgetting gap; and a **class-sample asymmetry**, in which class-level forgetting leaves strong representational traces, with positive gaps persisting across network depths, while sample-level forgetting is indistinguishable from random guessing.

2 Related Work

2.1 Machine Unlearning and Data Deletion

Machine unlearning aims to remove the influence of selected training data without full retraining [3,4]. Early work formalized the problem as efficient data deletion [11], while subsequent methods proposed partition-based retraining strategies such as SISA [3]. Approximate approaches include gradient-based removal [12], influence-function analysis [21], and distillation-based correction [9]. A parallel line of research pursues certified unlearning through differential-privacy-style guarantees or information-theoretic bounds on the residual influence of deleted data [36,39]. These certified approaches provide statistical guarantees on output distributions but do not directly examine the geometry of internal representations [15]. Mirage complements such guarantees by auditing whether representational structure, rather than output behavior, aligns with a retrained reference.

2.2 Federated and Distributed Unlearning

In federated learning, unlearning addresses client removal and data deletion in distributed settings [26]. While most existing works focus on horizontal federated learning, collaborative paradigms such as split learning explicitly transmit intermediate representations across parties [35]. In such settings, internal embeddings are shared and may be analyzed independently of the classifier head. However, evaluation protocols for federated unlearning largely inherit centralized output-level metrics, leaving the internal representation space unaudited.

2.3 Representation Probing and Geometric Analysis

A substantial body of work studies what neural networks encode in their intermediate layers. Linear probing has been widely used to assess recoverable information in learned representations [1, 2]. Representational similarity analysis, including Centered Kernel Alignment (CKA) [22], provides tools for comparing internal structures across models. These studies demonstrate that classifier performance does not fully characterize the geometry of embedding spaces. In vision models, class-conditional structure often remains linearly separable even under architectural or training variations. However, such representation-level analyses have rarely been integrated into unlearning evaluation [15, 18, 20].

2.4 Representation Leakage and Privacy Risks

Prior research has shown that trained models can retain sensitive information beyond their observable outputs. Membership inference attacks [30] and memorization analyses [31] reveal that internal representations may encode recoverable attributes even when output behavior appears constrained. In distributed learning frameworks where embeddings are transmitted or reused, this representational persistence raises additional risks. These findings suggest that certifying unlearning requires examining the recoverability and structural properties of internal representations rather than relying solely on output-level metrics.

3 Preliminaries

3.1 Collaborative Visual Learning Setting

We consider a collaborative visual learning setting in which feature extraction and classification are decoupled across parties. A model consists of K bottom encoders G_{θ_k} and a top classifier F_ω. For input sample $x = (x_1, \ldots, x_K)$, each party computes an embedding

$$H_k = G_{\theta_k}(x_k), \tag{1}$$

and the top model predicts

$$\hat{y} = F_\omega(H_1, \ldots, H_K). \tag{2}$$

The joint parameter set $\Theta = (\theta_1, \ldots, \theta_K, \omega)$ minimizes

$$\min_{\Theta} \frac{1}{n} \sum_{i=1}^{n} \ell\big(F_\omega(G_{\theta_1}(x_{1,i}), \ldots, G_{\theta_K}(x_{K,i})), y_i\big). \tag{3}$$

Vertical Federated Learning (VFL) is a representative instance of this formulation, where one active party holds labels and passive parties hold disjoint feature partitions. The key property of this setting is that intermediate representations H_k are explicitly exposed across parties.

3.2 Label Unlearning Objective

Given a trained model Θ^* and a target label set $\mathcal{Y}_u$ that are to be unlearned, label unlearning seeks to remove all information associated with $\mathcal{Y}_u$. Let

$$D_u = \{(x_i, y_i) \mid y_i \in \mathcal{Y}_u\}, \qquad D_r = D \setminus D_u. \tag{4}$$

An unlearned model Θ^u is considered ideal if it is functionally indistinguishable from a model Θ^r retrained from scratch on D_r. Existing evaluation protocols rely on output-level metrics. Retained accuracy (Acc_r) measures performance on D_r. Forgotten-label accuracy (y_u) measures prediction accuracy on D_u, where lower values indicate stronger behavioral forgetting. Runtime measures computational efficiency. These metrics assess classifier behavior but do not directly examine the internal representation space.

3.3 Representation-Level Criterion

Let $\phi_l(\cdot)$ denote the feature map of layer l. For a model Θ, define its linear probe recoverability on forgotten data as

$$\mathrm{LPR}(\Theta) = \max_{h \in \mathcal{H}_{\mathrm{lin}}} \mathbb{E}_{x \in D}\big[\mathbf{1}[h(\phi_l(x)) = \mathbf{1}[y \in \mathcal{Y}_u]]\big], \tag{5}$$

where $\mathcal{H}_{\mathrm{lin}}$ is the set of linear classifiers. We define the *forgetting gap* as

$$\Delta_{\mathrm{LPR}} = \mathrm{LPR}(\Theta^u) - \mathrm{LPR}(\Theta^r). \tag{6}$$

The retrained baseline Θ^r captures intrinsic class structure that naturally emerges in visual representations. A positive forgetting gap indicates residual class-specific structure retained by the unlearned model beyond what retraining would preserve. This criterion provides a representation-level perspective on whether unlearning achieves erasure rather than suppression of output behavior.

4 Methodology: The Mirage Framework

Mirage provides a principled framework for certifying representation-level forgetting in visual models (see Figure 2). The central question is whether unlearning removes class-specific structure from embeddings or suppresses the classifier head. Mirage answers this through complementary analyses of recoverability, geometric alignment, and spatial localization, all evaluated relative to a retrained-from-scratch baseline that defines the reference representation geometry.

4.1 Certification Objective

Let Θ^* denote the original trained model, Θ^u the unlearned model, and Θ^r a model retrained from scratch on retained data D_r. Representation-level certification aims to determine whether Θ^u is indistinguishable from Θ^r with respect to forgotten-label information in its internal embeddings.

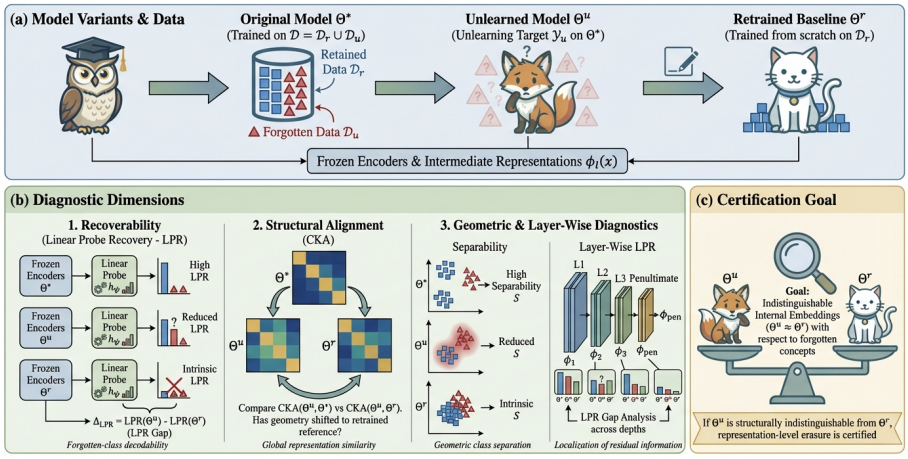

Fig. 2: Mirage: a representation-level certification framework. Mirage audits representation-level forgetting by comparing the original model Θ^*, the unlearned model Θ^u, and a retrained baseline Θ^r trained from scratch on retained data. The framework analyzes Linear Probe Recovery (LPR), Centered Kernel Alignment (CKA), feature separability, and layer-wise information persistence. Representation-level erasure is certified when the internal embeddings of Θ^u become structurally indistinguishable from the retrained reference.

Certification therefore requires that, for each diagnostic measure $\mathcal{D}$ in Mirage,

$$|\mathcal{D}(\Theta^u) - \mathcal{D}(\Theta^r)| \leq \epsilon, \tag{7}$$

under matched evaluation protocols. Any systematic deviation indicates residual representational traces beyond what retraining would preserve. Mirage evaluates this condition through three complementary dimensions: recoverability, structural alignment, and geometric localization.

4.2 Recoverability and Structural Analysis

This component examines whether forgotten-label information remains decodable and global embedding geometry has shifted toward the retrained reference.

Linear Probe Recovery. Linear Probe Recovery (LPR) evaluates whether forgotten-label information remains linearly decodable from frozen embeddings. The formal definition of LPR and the forgetting gap Δ_{LPR} are given in Eqs. (5)–(6). Concretely, given an unlearned model Θ^u, we freeze its parameters and extract features from layer l, denoted by $\phi_l(x) \in \mathbb{R}^{d_l}$. A linear classifier h_ψ is trained via cross-entropy to predict membership in the forgotten label set $\mathcal{Y}_u$:

$$\min_{\psi} \frac{1}{|D_{\text{probe}}|} \sum_{(x_i, y_i) \in D_{\text{probe}}} \ell_{\text{CE}}\big(h_\psi(\phi_l(x_i)),\ \mathbf{1}[y_i \in \mathcal{Y}_u]\big). \tag{8}$$

The probe thus performs binary classification by distinguishing samples belonging to $\mathcal{Y}_u$ from retained samples. Because visual representations naturally encode class structure, absolute LPR alone is insufficient. Mirage therefore evaluates Δ_{LPR} relative to the retrained baseline. A positive Δ_{LPR} indicates residual class-specific information beyond intrinsic representation geometry.

Recoverability Interpretation. If forgotten-class features remain separated from retained features in embedding space, linear classification accuracy exceeds chance. Under equal class priors and isotropic Gaussian class-conditional assumptions, let the signal-to-noise ratio be defined as

$$\mathrm{SNR} = \frac{\|\mu_u - \mu_r\|^2}{\sigma^2}, \tag{9}$$

the optimal linear classifier satisfies

$$\mathrm{Acc}_{\mathrm{LPR}} \geq \Phi\left(\frac{\sqrt{\mathrm{SNR}}}{2}\right), \tag{10}$$

where Φ is the standard normal cumulative distribution function. Under the isotropic assumption ($\Sigma_u = \Sigma_r = \sigma^2 I$), the feature separability score $\mathcal{F}$ defined in Section 4.3 satisfies $\mathcal{F} = \mathrm{SNR}/(2d)$, so $\mathcal{F}$ is a dimension-normalized proxy for SNR that preserves the same monotone relationship with probe accuracy. Empirically, methods with high separability in Table 2 (e.g., FT on COVID-19: $\mathcal{F} = 0.280$) indeed exhibit high LPR (96.2%), while methods with separability close to the retrained level (e.g., UNSIR on COVID-19: $\mathcal{F} = 0.040$ vs. Retrain $\mathcal{F} = 0.031$) show LPR near the baseline. This confirms that feature-space geometry governs linear recoverability, and $\mathcal{F}$ provides a principled, probe-independent measure of residual class structure.

CKA-Based Structural Alignment. While LPR measures decodability, it does not capture global geometric structure. Mirage therefore employs Centered Kernel Alignment (CKA) [22]. For any two models Θ_a and Θ_b, let $X \in \mathbb{R}^{n \times p}$ and $Y \in \mathbb{R}^{n \times q}$ denote their respective layer-l representations computed on the same n inputs. Linear CKA is defined as

$$\mathrm{CKA}(X, Y) = \frac{\|Y^\top X\|_F^2}{\|X^\top X\|_F \cdot \|Y^\top Y\|_F}. \tag{11}$$

We compare representations among Θ^*, Θ^u, and Θ^r. If Θ^u remains more aligned with Θ^* than with Θ^r, the unlearning procedure has not reshaped the embedding geometry toward the retrained reference. In such cases, behavioral forgetting may coexist with structural persistence.

4.3 Geometric and Layer-Wise Diagnostics

This component examines distribution-level separation and the spatial distribution of residual information across network depth.

Feature Separability. To quantify geometric class separation independent of probe optimization, Mirage computes a Fisher-inspired separability score. Let μ_u and Σ_u denote the mean and covariance of embeddings from the forgotten set D_u, and μ_r and Σ_r denote those from the retained set D_r. The score is:

$$\mathcal{F} = \frac{\|\mu_u - \mu_r\|^2}{\mathrm{tr}(\Sigma_u) + \mathrm{tr}(\Sigma_r)}. \tag{12}$$

A separability value exceeding that of the retrained baseline indicates residual structural discrimination of forgotten classes.

Layer-Wise Recovery. Residual information may localize to specific network depths. Mirage evaluates LPR across early, intermediate, and penultimate layers:

$$\Delta_{\mathrm{LPR}}^{(l)} = \mathrm{LPR}_l(\Theta^u) - \mathrm{LPR}_l(\Theta^r), \tag{13}$$

where LPR_l denotes recoverability measured at layer l. Persistent positive gaps across multiple depths indicate that unlearning has not altered representation hierarchies beyond superficial classifier adjustments (see Appendix A2).

5 Experiments

5.1 Experimental Setup

Datasets. We evaluate on seven datasets: **MNIST** [25], **CIFAR-10** [23], **CIFAR-100** [23], **ModelNet** [38], **Brain Tumor MRI** [37], **COVID-19 Radiography** [8, 29], and **Yahoo Answers** [6]. The first six datasets are image classification benchmarks with increasing visual complexity, while Yahoo Answers is a text classification dataset included to verify that the proposed evaluation protocol is not limited to purely visual inputs.

Baseline Methods. We compare the following unlearning methods: **Retrain** (retraining from scratch on D_r), **Fine-Tuning** (FT) [12], **Fisher Forgetting** [12], **Amnesiac Unlearning** [13], **UNSIR** [32], **Boundary Unlearning** (BU) [7], **SSD** [10], and the **Target Method** [14]. All implementations follow the original specifications and hyperparameters provided by the respective authors.

Evaluation Metrics. **Output-Level Metrics.** We report retained accuracy Acc_r and forgotten-label accuracy y_u as defined in Section 3.2. Effective output-level forgetting requires low y_u while maintaining high Acc_r. **Representation-Level Metrics.** We apply the Mirage diagnostics described in Linear Probe Recovery (LPR) and its gap Δ_{LPR} (Section 4.2), CKA-based structural alignment (Section 4.2), and feature separability (Section 4.3). For LPR, we use a logistic regression probe with ℓ_2 regularization ($C = 1.0$), trained for 1000 iterations on features extracted from the penultimate layer. CKA is computed using linear CKA on 5000 randomly sampled data points. Feature separability uses all available data from D_u and a balanced sample from D_r.

Table 1: Output-level metrics for single-label unlearning across seven datasets. Acc_r (%) is retained accuracy (↑) and y_u (%) is forgotten-label accuracy (↓). **Bold** entries indicate apparent output-level success.

Method	MNIST		CIFAR-10		CIFAR-100		Brain Tumor		COVID-19		ModelNet		Yahoo Answers	
	Acc_r	y_u	Acc_r	y_u	Acc_r	y_u	Acc_r	y_u	Acc_r	y_u	Acc_r	y_u	Acc_r	y_u
Retrain	99.4	0.0	89.7	0.0	62.2	0.0	99.1	0.0	92.8	0.0	82.3	0.0	56.7	0.0
FT	99.3	99.4	89.7	48.5	61.9	31.7	98.5	77.4	92.0	91.8	81.7	52.7	56.4	5.0
Fisher	6.8	33.3	9.0	14.6	0.9	0.0	13.7	58.9	4.6	64.4	11.1	0.0	11.4	0.1
Amnesiac	75.0	12.6	70.1	13.5	30.1	0.0	69.2	19.6	57.5	24.1	75.3	29.3	42.5	2.7
UNSIR	48.4	0.0	23.3	0.0	2.4	0.0	51.4	0.0	**70.9**	**0.0**	14.9	5.3	32.7	0.0
BU	44.7	0.0	47.4	0.0	33.4	0.0	**72.4**	**0.0**	**77.9**	**0.0**	73.3	8.7	51.7	0.1
SSD	99.3	97.0	89.6	88.2	62.7	72.7	98.8	74.7	92.1	90.9	81.7	39.3	55.2	37.6
Target	14.9	0.0	24.1	0.0	1.0	0.0	33.3	0.0	34.3	0.0	11.8	0.0	49.7	0.0

Implementation Details. We follow the experimental protocol [14], including identical data splits, model architectures, optimization settings, and unlearning procedures to ensure fair comparison. For all vision datasets, we adopt **ResNet18** [16] as the backbone model, while for Yahoo Answers we use a multilayer perceptron (MLP) consistent with the baseline configuration. In the vertical federated learning (VFL) setting, input features are partitioned equally between two passive parties, while the active party holds the labels and the top classifier. All models are first trained on the full dataset D to obtain the original model Θ^*. For single-label unlearning, one class $\mathcal{Y}_u$ is designated as the forgotten class and the remaining data form the retained set D_r. For sample-level unlearning, 5% or 10% of samples are randomly selected for removal. The retrained baseline Θ^r is obtained by training from scratch on D_r with the same architecture and hyperparameters. Results are averaged over three random seeds.

5.2 Comparison

Quantitative Analysis We begin with quantitative comparison across all seven datasets using both output-level and representation-level metrics.

Output-Level Behavior. As shown in Table 1, no method consistently achieves both high Acc_r and low y_u. FT and SSD maintain Acc_r close to Retrain but exhibit high y_u, which indicates ineffective forgetting. Target and Fisher reduce y_u to zero on most datasets, but this reduction is accompanied by severe utility degradation. Amnesiac shows unstable behavior across datasets. BU is the only method that achieves near-zero y_u while preserving reasonable Acc_r on selected datasets such as Brain Tumor, COVID-19, and Yahoo Answers.

Representation-Level Diagnostics. Mirage reveals a starkly different picture. As shown in Table 2, substantial LPR values (54 to 83% across datasets) are observed even for the Retrain model, confirming that class structure is inherently encoded in visual representations. Absolute LPR values are therefore uninformative. The principled criterion is instead the forgetting gap Δ_{LPR} relative to Retrain.

BU exhibits consistently positive Δ_{LPR} on datasets where it appears successful at the output level. On COVID-19, BU achieves $y_u = 0\%$ while producing an LPR of 94.7%, which is 15.4 points above the Retrain baseline. This indicates that forgotten-class information remains linearly recoverable in feature space.

FT and SSD show even larger positive Δ_{LPR}, which aligns with their high CKA similarity to the original model. Target shows negative Δ_{LPR} on several datasets. However, this coincides with catastrophic utility collapse and low CKA values, suggesting that representational structure is destroyed rather than selectively modified. The only exception is Yahoo Answers, where Target retains high CKA (0.957) and a positive Δ_{LPR} (+7.8), behaving like a suppression-style method rather than collapsing.

Feature Separability. The separability scores in Table 2 provide a complementary perspective independent of probe optimization. FT and SSD exhibit the highest separability values (e.g., 0.280 and 0.340 on COVID-19), reflecting their near-complete preservation of feature geometry. BU shows moderate separability, lower than FT and SSD but consistently above Retrain. This confirms that the forgotten class remains geometrically distinguishable. Target produces heterogeneous separability. It is near-zero on datasets where it collapses utility (e.g., 0.038 on CIFAR-100), but occasionally high (e.g., 0.354 on Brain Tumor) where residual class structure persists in a distorted feature space. Separability and LPR are correlated but not redundant. High separability generally accompanies high LPR, yet cases like UNSIR on CIFAR-10 (Sep = 0.097, LPR = 69.5%) show that moderate linear decodability can occur even with low geometric separation, suggesting the probe exploits subtle directional structure beyond centroid differences. These patterns are shown in Figure 3, where BU consistently lies in the forgetting-illusion region ($y_u \approx 0$, $\Delta_{\mathrm{LPR}} > 0$) while Target falls into a collapse regime with degraded utility on the vision datasets. Output-level forgetting therefore does not imply representation-level erasure, and Mirage distinguishes these two failure modes, revealing cases where apparent forgetting masks substantial residual class structure.

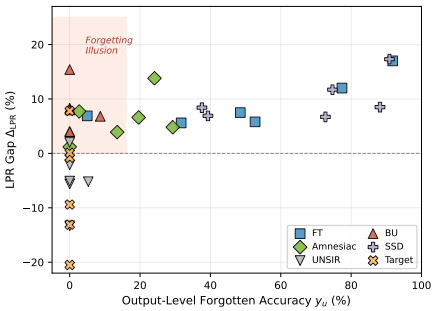

Fig. 3: Forgetting gap across methods and datasets. Each point represents a method–dataset pair with coordinates $(y_u, \Delta_{\mathrm{LPR}})$. The red region ($y_u \approx 0$, $\Delta_{\mathrm{LPR}} > 0$) is the *forgetting illusion*. BU (triangles) consistently falls here.

Qualitative Analysis We examine the geometry of learned representations through feature-space visualization and structural comparison.

t-SNE Visualization. t-SNE embeddings of bottom-model features on the COVID-19 dataset are presented in Figure 4. In the Retrain model, the forgotten class

Table 2: Mirage representation-level diagnostics for single-label unlearning. LPR denotes Linear Probe Recovery (%) and Δ denotes the LPR gap relative to Retrain. CKA_O measures similarity to the original model and Sep denotes feature separability. Red values indicate $\Delta_{\text{LPR}} > 5\%$. **MNIST** is omitted because all methods achieve near-ceiling LPR (> 97%) with negligible Δ_{LPR} (< 1.5 pp), making gap analysis uninformative (output-level results remain in Table 1). **Fisher** is omitted because it causes catastrophic utility collapse ($\text{Acc}_r < 14\%$ on all vision datasets, see Table 1), rendering representation diagnostics uninformative.

Method	CIFAR-10				COVID-19				Brain Tumor			
	LPR	Δ	CKA_O	Sep	LPR	Δ	CKA_O	Sep	LPR	Δ	CKA_O	Sep
Retrain	82.8	—	.989	.093	79.2	—	.981	.031	79.8	—	.978	.176
FT	90.4	+7.5	.999	.160	96.2	+17.0	.998	.280	91.8	+12.0	.986	.549
Amnesiac	86.7	+3.9	.974	.103	93.0	+13.8	.971	.091	86.4	+6.6	.943	.234
UNSIR	69.5	−13.3	.947	.097	77.2	−2.1	.954	.040	74.2	−5.6	.936	.118
BU	86.8	+4.0	.956	.149	94.7	**+15.4**	.957	.115	88.1	**+8.3**	.909	.289
SSD	91.3	+8.5	1.00	.246	96.5	+17.3	1.00	.340	91.5	+11.7	1.00	.641
Target	82.0	−0.8	.928	.104	69.9	−9.4	.841	.089	79.9	+0.1	.816	.354

Method	CIFAR-100				ModelNet				Yahoo Answers			
	LPR	Δ	CKA_O	Sep	LPR	Δ	CKA_O	Sep	LPR	Δ	CKA_O	Sep
Retrain	73.2	—	.992	.453	72.4	—	.990	.132	53.7	—	.926	.008
FT	78.8	+5.6	.999	.641	78.2	+5.8	1.00	.241	60.6	+6.9	.992	.020
Amnesiac	74.4	+1.2	.986	.284	77.2	+4.8	.983	.191	61.4	+7.7	.927	.020
UNSIR	68.1	−5.1	.960	.277	67.2	−5.2	.921	.030	55.7	+2.0	.836	.003
BU	77.1	+4.0	.985	.381	79.2	+6.8	.969	.165	61.7	+8.0	.939	.018
SSD	79.9	+6.7	1.00	.686	79.3	+6.9	1.00	.240	62.1	+8.4	.999	.023
Target	52.7	−20.5	.622	.038	59.4	−13.1	.778	.021	61.5	+7.8	.957	.020

forms a clearly separable cluster despite the absence of its labels during training. This confirms that general-purpose representations naturally encode class-level structure. For BU, the forgotten-class cluster remains compact and well-separated from retained classes, and its spatial configuration closely resembles that of Retrain. Although output-level accuracy on the forgotten class drops to zero, the geometric structure persists. This visual evidence is consistent with the large positive Δ_{LPR} observed quantitatively. In contrast, Target produces a scattered feature distribution with no coherent clustering. The collapse of structure aligns with the sharp drop in retained accuracy and the low CKA similarity. The apparent reduction in LPR is therefore attributable to representational destruction rather than targeted forgetting. Analogous patterns are observed across all remaining datasets (see Figures A1 to A6 in the Appendix).

Representation Similarity Patterns. We further compare CKA similarity across methods. FT and SSD exhibit near-unity CKA with the original model, indicating minimal structural change. BU shows moderately high similarity, which suggests that most of the representation space remains intact. Target shows substantially lower similarity on several datasets, reflecting large-scale alteration of feature geometry. These qualitative patterns reinforce the quantitative findings by demonstrating that methods appearing successful under output-level evalu-

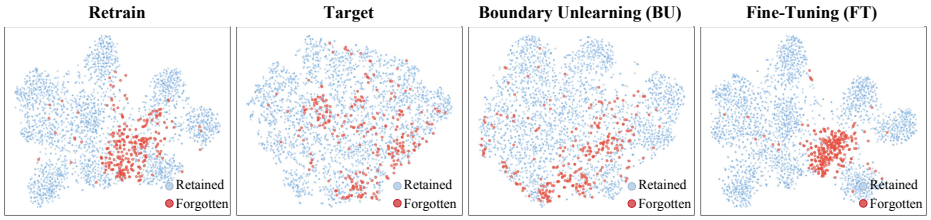

Fig. 4: t-SNE visualization of bottom-model features on COVID-19. Blue: retained classes; red: forgotten class. **Retrain** (left) shows the forgotten class forming a separable cluster even without training on its labels. **Target** scatters all points, reflecting model collapse ($\mathrm{Acc}_r = 34.3\%$). **BU** preserves the forgotten-class cluster almost identically to Retrain, visually confirming the forgetting illusion ($\Delta_{\mathrm{LPR}} = +15.4\%$). **FT** likewise preserves the full representational structure.

Table 3: Sample-level unlearning results across seven datasets. LPR (%) for forgetting 5% and 10% of random samples (mean over 3 rounds). All methods, including Retrain, remain near chance level ($\approx 50\%$), in contrast to class-level recovery (Table 2).

Method	MNIST		CIFAR-10		CIFAR-100		Brain Tumor		COVID-19		ModelNet		Yahoo Answers	
	5%	10%	5%	10%	5%	10%	5%	10%	5%	10%	5%	10%	5%	10%
Retrain	48.6	49.6	50.8	50.9	49.0	50.0	49.6	51.8	51.4	49.8	51.0	50.4	52.0	52.4
FT	49.9	49.2	49.2	49.6	50.8	50.2	50.4	49.0	49.6	50.4	53.3	50.0	49.6	50.4
Amnesiac	50.4	49.6	49.5	50.4	51.0	49.9	48.8	50.1	52.5	51.4	51.8	46.8	49.0	51.1
UNSIR	49.1	50.2	49.6	49.5	50.7	50.4	52.0	49.6	51.9	50.2	44.6	49.6	51.0	49.2
BU	50.0	49.7	49.2	48.5	50.5	49.4	50.4	49.6	50.0	51.7	49.6	49.6	49.2	51.1
Target	48.5	50.2	49.8	49.8	49.0	50.2	54.2	47.4	49.5	49.1	53.8	50.3	48.8	50.6

ation can fully preserve the geometric signature of the forgotten class. Mirage makes this residual structure directly visible and quantifiable.

Class-Sample Asymmetry We extend the analysis to sample-level unlearning, where a fixed proportion (5% or 10%) of randomly selected samples is designated for forgetting rather than entire classes. The LPR results are shown in Table 3. A striking *class-sample asymmetry* emerges across the two forgetting settings. Class-level LPR reaches ∼99% on MNIST and exceeds 96% on COVID-19 (Table 2, where the Retrain baseline is 54–83% on the six reported datasets), whereas sample-level LPR $\approx 50\%$ for *all* methods including Retrain. This dichotomy has a clear geometric explanation rooted in how representations encode information. Class identity is captured by stable feature-space structures, such as cluster centers and inter-class margins, that persist across training runs. Individual sample membership, by contrast, does not manifest as a linearly separable signal in penultimate-layer embeddings. The implication for auditing is therefore twofold. Linear probes are effective and necessary for class-level auditing, but fundamentally insufficient for sample-level verification. The two forgetting settings demand categorically different diagnostic tools.

5.3 Ablation Study

Sensitivity to Strength. We first analyze how increasing unlearning strength affects both utility and representation-level residuals. We vary the number of BU epochs while keeping all other hyperparameters fixed, asking: *can the forgetting illusion be eliminated by simply training longer?* As shown in Table 4, the answer is no. Even at 20 epochs, Δ_{LPR} remains positive on all three datasets. On COVID-19, LPR stays above 94% regardless of epoch count, which is 15–16 points above the Retrain baseline. Meanwhile, Acc_r degrades steadily, dropping on CIFAR-10 from 89.7% (Retrain) to 21.8% at 20 epochs. No operating point exists where representation-level residuals vanish while utility remains intact. This reveals a fundamental limitation that BU operates at the decision boundary without modifying the underlying feature geometry. More epochs progressively destroy utility, but the forgotten class remains linearly recoverable because the representational structure responsible for class separability is never altered.

Table 4: Effect of unlearning epochs for BU (averaged over 3 rounds). Δ_{LPR} remains positive across epochs on all datasets.

	CIFAR-10			CIFAR-100			COVID-19		
Epochs	Acc_r	LPR	Δ	Acc_r	LPR	Δ	Acc_r	LPR	Δ
0 (Ret.)	89.7	83.3	–	62.2	74.3	–	92.7	79.5	–
1	79.2	87.7	+4.4	57.4	80.5	+6.2	89.3	95.4	+15.9
3	58.9	87.8	+4.5	42.3	81.6	+7.3	85.6	94.6	+15.1
5	43.3	86.2	+2.9	32.9	77.7	+3.4	81.7	94.6	+15.1
10	32.7	85.3	+2.0	25.5	77.7	+3.4	78.8	94.4	+14.9
20	21.8	83.9	+0.6	22.0	75.4	+1.1	68.2	95.4	+15.9

Multi-Party Scaling. We next examine how representation-level residuals scale with the number of passive parties across which features are partitioned. We vary $K \in 2, 4, 8$ on CIFAR-10, forgetting a single class, and measure LPR and Δ_{LPR}. As shown in Table 5, BU maintains consistently positive Δ_{LPR} across all K. This indicates that residual information persists even as features are split across more parties. FT exhibits even larger positive gaps, consistent with its minimal representational change. In contrast, Target becomes increasingly negative as K increases. This pattern reflects progressive destruction of the feature space rather than selective removal of specific class structure. The scaling behavior therefore distinguishes persistent leakage from global collapse.

Table 5: Effect of the number of passive parties (K) on CIFAR-10. Δ_{LPR} remains positive for FT and BU but becomes negative for Target as K increases.

	LPR (%)			Δ_{LPR} (%)		
Method	K=2	K=4	K=8	K=2	K=4	K=8
Retrain	82.8	81.6	80.9	–	–	–
FT	90.7	87.6	87.1	+7.9	+6.0	+6.2
BU	87.6	86.1	84.6	+4.8	+4.6	+3.7
Target	83.9	76.7	69.0	+1.2	-4.9	-11.9

Sensitivity to Forgotten Class. We investigate whether the forgetting gap depends on which class is selected for removal. On CIFAR-10, we test each of the 10 classes individually. The Retrain LPR varies across classes, which reflects inherent differences in class separability within the feature space. However, the relative behavior of each method remains stable. BU consistently produces positive Δ_{LPR}, while Target produces values near or below zero that coincide with degraded utility. Similar patterns are observed on CIFAR-100. Although absolute LPR varies across classes, the qualitative relationship between Δ_{LPR} and retained accuracy remains unchanged. This confirms that the forgetting gap is

not an artifact of a particular class choice but a structural property of the unlearning mechanism. Detailed per-class results are provided in Appendix A3.

6 Discussion

Representation-Level Erasure as a Structural Constraint. Our results suggest that the gap between output-level forgetting and representation-level erasure is structural. In deep models, class identity is encoded in feature geometry through cluster structure and inter-class margins. Adjusting classifier weights or suppressing logits can change decision boundaries without altering this geometry, leaving forgotten-class information linearly recoverable. We therefore adopt a retraining-relative criterion. Because a retrained model reflects the geometry induced by retained data, a positive forgetting gap ($\Delta_{\mathrm{LPR}} > 0$) indicates residual class structure beyond retraining. Representation-level forgetting thus requires geometric alignment with the retrained reference.

An Empirical Unlearning Trilemma. Across datasets and methods, we observe a consistent tension among three desiderata: (1) **utility**, (2) **output forgetting**, and (3) **representation forgetting** (small Δ_{LPR}). No evaluated method achieves all three simultaneously. Methods that preserve utility often retain recoverable class structure, whereas methods that strongly suppress outputs tend to degrade retained performance. Although not a formal impossibility result, this pattern suggests a structural tension between preserving feature geometry and eliminating class separability. Additional discussion appears in Appendix A6.

7 Conclusion

We presented Mirage, a representation-level certification framework for visual unlearning. Across seven datasets, we show that suppressing forgotten-label predictions does not imply erasure of underlying feature geometry. Methods that appear successful at the output level often retain recoverable class information. By defining the forgetting gap relative to a retrained baseline, Mirage reframes unlearning evaluation as a geometric certification problem and exposes a tension among utility, behavioral forgetting, and representation-level erasure.

Protocol Alignment Statement

We strictly follow the original datasets, architectures, unlearning scenarios, and evaluation metrics of the target baseline. Mirage introduces only representation-level diagnostics applied to the same trained models, ensuring complete experimental consistency and fair comparison.

Limitations

Mirage is designed as an auditing framework rather than a new unlearning algorithm. Our experiments follow a VFL protocol consistent with prior work, so behavior in heterogeneous real-world deployments may differ. The linear probe provides a conservative lower bound on recoverable information. Nonlinear probes may reveal additional residual structure beyond linear separability (see Appendix A4). Extending certification to stronger adversaries and to horizontal federated settings remains an important open direction.

References

1. Alain, G., Bengio, Y.: Understanding intermediate layers using linear classifier probes. arXiv preprint arXiv:1610.01644 (2016)
2. Belinkov, Y.: Probing classifiers: Promises, shortcomings, and advances. Computational Linguistics **48**(1), 207–219 (2022)
3. Bourtoule, L., Chandrasekaran, V., Choquette-Choo, C.A., Jia, H., Travers, A., Zhang, B., Lie, D., Papernot, N.: Machine unlearning. In: 2021 IEEE symposium on security and privacy (SP). pp. 141–159. IEEE (2021)
4. Cao, Y., Yang, J.: Towards making systems forget with machine unlearning. In: 2015 IEEE symposium on security and privacy. pp. 463–480. IEEE (2015)
5. Che, T., Zhou, Y., Zhang, Z., Lyu, L., Liu, J., Yan, D., Dou, D., Huan, J.: Fast federated machine unlearning with nonlinear functional theory. In: International conference on machine learning. pp. 4241–4268. PMLR (2023)
6. Chen, J., Yang, Z., Yang, D.: Mixtext: Linguistically-informed interpolation of hidden space for semi-supervised text classification. In: Proceedings of the 58th annual meeting of the association for computational linguistics. pp. 2147–2157 (2020)
7. Chen, M., Gao, W., Liu, G., Peng, K., Wang, C.: Boundary unlearning: Rapid forgetting of deep networks via shifting the decision boundary. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7766–7775 (2023)
8. Chowdhury, M.E., Rahman, T., Khandakar, A., Mazhar, R., Kadir, M.A., Mahbub, Z.B., Islam, K.R., Khan, M.S., Iqbal, A., Al Emadi, N., et al.: Can ai help in screening viral and covid-19 pneumonia? Ieee Access **8**, 132665–132676 (2020)
9. Chundawat, V.S., Tarun, A.K., Mandal, M., Kankanhalli, M.: Zero-shot machine unlearning. IEEE Transactions on Information Forensics and Security **18**, 2345–2354 (2023)
10. Foster, J., Schoepf, S., Brintrup, A.: Fast machine unlearning without retraining through selective synaptic dampening. In: Proceedings of the AAAI conference on artificial intelligence. vol. 38, pp. 12043–12051 (2024)
11. Ginart, A., Guan, M., Valiant, G., Zou, J.Y.: Making ai forget you: Data deletion in machine learning. Advances in neural information processing systems **32** (2019)
12. Golatkar, A., Achille, A., Soatto, S.: 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)
13. Graves, L., Nagisetty, V., Ganesh, V.: Amnesiac machine learning. In: Proceedings of the AAAI conference on artificial intelligence. vol. 35, pp. 11516–11524 (2021)

14. Gu, H., Tae, H.X., Fan, L., Chan, C.S.: Towards privacy-guaranteed label unlearning in vertical federated learning: Few-shot forgetting without disclosure. In: The Fourteenth International Conference on Learning Representations (2026)
15. Hayes, J., Shumailov, I., Triantafillou, E., Khalifa, A., Papernot, N.: Inexact unlearning needs more careful evaluations to avoid a false sense of privacy. In: 2025 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML). pp. 497–519. IEEE (2025)
16. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
17. Jang, Y., Lee, J., Kim, D., Jo, J., Woo, S.S.: Suppression or deletion: A restoration-based representation-level analysis of machine unlearning (2026), https://arxiv.org/abs/2602.18505, to appear in Proc. ACM Web Conference (WWW), 2026
18. Jeon, D., Jeung, W., Kim, T., No, A., Choi, J.: An information theoretic evaluation metric for strong unlearning. Proceedings of the AAAI Conference on Artificial Intelligence **40**(26), 22173–22181 (2026). https://doi.org/10.1609/aaai.v40i26.39373, https://arxiv.org/abs/2405.17878
19. Jia, J., Liu, J., Ram, P., Yao, Y., Liu, G., Liu, Y., Sharma, P., Liu, S.: Model sparsity can simplify machine unlearning. Advances in Neural Information Processing Systems **36**, 51584–51605 (2023)
20. Kim, Y., Cha, S., Kim, D.: Are we truly forgetting? a critical re-examination of machine unlearning evaluation protocols. Engineering Applications of Artificial Intelligence **167**, 113785 (Mar 2026). https://doi.org/10.1016/j.engappai.2026.113785, https://doi.org/10.1016/j.engappai.2026.113785
21. Koh, P.W., Liang, P.: Understanding black-box predictions via influence functions. In: International conference on machine learning. pp. 1885–1894. PMLR (2017)
22. Kornblith, S., Norouzi, M., Lee, H., Hinton, G.: Similarity of neural network representations revisited. In: International conference on machine learning. pp. 3519–3529. PMlR (2019)
23. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009)
24. Kurmanji, M., Triantafillou, P., Hayes, J., Triantafillou, E.: Towards unbounded machine unlearning. Advances in neural information processing systems **36**, 1957–1987 (2023)
25. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE **86**(11), 2278–2324 (1998)
26. Liu, G., Ma, X., Yang, Y., Wang, C., Liu, J.: Federaser: Enabling efficient client-level data removal from federated learning models. In: 2021 IEEE/ACM 29th international symposium on quality of service (IWQOS). pp. 1–10. IEEE (2021)
27. Meng, C., Luo, J., Yan, Z., Yu, Z., Fu, R., Gan, Z., Ouyang, C.: Tri-subspaces disentanglement for multimodal sentiment analysis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8791–8800 (2026)
28. Papyan, V., Han, X., Donoho, D.L.: Prevalence of neural collapse during the terminal phase of deep learning training. Proceedings of the National Academy of Sciences **117**(40), 24652–24663 (2020)
29. Rahman, T., Khandakar, A., Qiblawey, Y., Tahir, A., Kiranyaz, S., Kashem, S.B.A., Islam, M.T., Al Maadeed, S., Zughaier, S.M., Khan, M.S., et al.: Exploring the effect of image enhancement techniques on covid-19 detection using chest x-ray images. Computers in biology and medicine **132**, 104319 (2021)

30. Shokri, R., Stronati, M., Song, C., Shmatikov, V.: Membership inference attacks against machine learning models. In: 2017 IEEE symposium on security and privacy (SP). pp. 3–18. IEEE (2017)
31. Song, C., Ristenpart, T., Shmatikov, V.: Machine learning models that remember too much. In: Proceedings of the 2017 ACM SIGSAC Conference on computer and communications security. pp. 587–601 (2017)
32. Tarun, A.K., Chundawat, V.S., Mandal, M., Kankanhalli, M.: Fast yet effective machine unlearning. IEEE transactions on neural networks and learning systems **35**(9), 13046–13055 (2023)
33. Thudi, A., Jia, H., Shumailov, I., Papernot, N.: On the necessity of auditable algorithmic definitions for machine unlearning. In: 31st USENIX security symposium (USENIX Security 22). pp. 4007–4022 (2022)
34. Varshney, A.K., Vandikas, K., Torra, V.: Unlearning clients, features and samples in vertical federated learning. arXiv preprint arXiv:2501.13683 (2025)
35. Vepakomma, P., Gupta, O., Swedish, T., Raskar, R.: Split learning for health: Distributed deep learning without sharing raw patient data. arXiv preprint arXiv:1812.00564 (2018)
36. Wang, J., Lin, Y., Niyato, D., Gao, Z., Du, H., Zhang, T., Tang, X., Fan, J., Han, Z.: A zero-shot federated unlearning framework with stability verification. IEEE Transactions on Cognitive Communications and Networking (2025)
37. Wang, Z., Gao, X., Wang, C., Cheng, P., Chen, J.: Efficient vertical federated unlearning via fast retraining. ACM Transactions on Internet Technology **24**(2), 1–22 (2024)
38. Wu, Z., Song, S., Khosla, A., Yu, F., Zhang, L., Tang, X., Xiao, J.: 3d shapenets: A deep representation for volumetric shapes. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1912–1920 (2015)
39. Yang, W., Al-Masri, E., Kotevska, O.: Mic-dp: A scalable correlation-aware differential privacy framework for high-dimensional data. IEEE Transactions on Privacy **2**, 131–143 (2025)
40. Yu, Z., Han, L., Wang, P., IDRIS, M.Y.I., Xiang, Y.: Instantforget: Training-free functional feature unlearning via subspace projection and inference-time smoothing (2025)
41. Yu, Z., Idris, M.Y.I., Wang, P., Xia, Y., Xiang, Y.: Forgetme: Benchmarking the selective forgetting capabilities of generative models. Engineering Applications of Artificial Intelligence **161**, 112087 (2025)
42. Zhang, B., Chen, Z., Shen, C., Li, J.: Verification of machine unlearning is fragile. In: Proceedings of the 41st International Conference on Machine Learning. pp. 58717–58738 (2024)