On the Fragility of Out-of-Distribution Detectors Under Model Updates

A.Q.M. Sazzad Sayyed[†], Nathaniel D. Bastian[‡] and Francesco Restuccia[†]

[†]Northeastern University [‡]United States Military Academy

Abstract

While post-hoc Out-of-Distribution (OOD) detectors are typically evaluated assuming a static classifier, deployed models are often updated through techniques such as Machine Unlearning (MU) or Test-time Adaptation (TTA). Existing work has largely neglected this aspect; as such, we study the resulting *OOD fragility*, defined as the change in detection performance induced by a model update. Our study considers 16 update methods, 11 detectors, and 2 image benchmarks. Our key finding is that model updates can substantially degrade OOD detection up to 23.82% in terms of Area Under Receiver Operating Curve (AUROC) while preserving In-Distribution (ID) utility. Specifically, aggressive model update (increasing the size of the forget set for MU or increasing the corruption severity in TTA) causes larger degradation, degree of performance deterioration varies across OOD datasets and detector families, and feature-based detectors (ΔAUROC -0.0338 for RMDS) are generally more robust than confidence-based methods (ΔAUROC -0.1429 for GEN). These results expose a hidden reliability risk in dynamic deployment pipelines and motivate post-update reassessment and internal-signal-based prediction of OOD performance without access to OOD data.

1 Introduction

While Deep Neural Networks (DNNs) have achieved remarkable success, they usually exhibit highly confident predictions on inputs that do not belong to the training distribution (Sayyed et al. 2025). This limitation has motivated extensive research on Out-of-Distribution (OOD) detection, whose goal is to identify samples that do not follow the In-Distribution (ID) data while preserving predictive performance on ID examples (Hendrycks and Gimpel 2017; Lee et al. 2018; Liu et al. 2020). Reliable OOD detection is particularly important in safety-critical applications where overconfident predictions on anomalous inputs can lead to system failure (Nguyen, Yosinski, and Clune 2015).

Existing Issue. Most of the existing OOD detection methods are developed and evaluated *under the assumption that the model remains static throughout its lifetime*. Given a pretrained classifier, post-hoc OOD detectors exploit properties of its logits and feature representations to distinguish ID from OOD samples (Hendrycks and Gimpel 2017; Lee et al. 2018; Wang et al. 2022a). As such, benchmarks almost universally assume that the underlying DNN remains fixed after training. *However, this assumption is increasingly misaligned with real-world deployment, since modern DNNs are rarely static after deployment and are routinely modified to satisfy operational and regulatory requirements*.

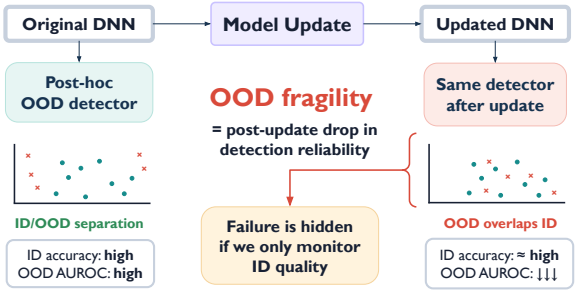

Figure 1: Fragility of OOD detection after model updates.

Two important examples of dynamic model updates are *Machine Unlearning (MU)* (Ginart et al. 2019) and *Test-time Adaptation (TTA)* (Wang et al. 2021). MU updates a trained model to remove the influence of designated training samples, classes, or subsets, often to satisfy data governance requirements, or user deletion requests (Bourtoule et al. 2021; Nguyen et al. 2022). Conversely, TTA updates a deployed model to improve robustness under distribution shift – usually without access to the original training process and/or data (Wang et al. 2021; Xiao and Snoek 2024). Although these paradigms are typically studied independently and serve different objectives, they share a common operational characteristic: *modifying the parameters or representations of a deployed model after training*.

Proposed Contribution. This observation raises a fundamental question that, to the best of our knowledge, has not been systematically investigated: ***what happens to OOD detectors when the underlying model is updated after deployment?*** Post-hoc OOD detectors rely on structural properties of the original model, such as class margins, feature concentration, prototype geometry, and representation subspaces. Unlearning or adaptation can potentially alter these properties – even after accounting for shift in representation – while preserving classification accuracy, causing a previously re-

liable detector to fail after a model update. In this work, we view MU and TTA as post-training modifications of a deployed model and abstract them as *Knowledge Updates (KU)*for this study. To this end, we define *OOD fragility* as the degradation in OOD detection performance induced by a KU and investigate the following questions:

> ❶ Do post-deployment model updates systematically degrade OOD detection performance? ❷ Are some update mechanisms more harmful than others? ❸ Are certain OOD detector families inherently more robust to model updates? ❹ What factors drive the observed degradation?

Technical Approach and Key Findings. To answer these questions, we conduct a comprehensive empirical study spanning 10 MU methods, 6 TTA algorithms, and 11 post-hoc OOD detectors across CIFAR10 and CIFAR100 image classification benchmarks. Our results reveal that OOD reliability can deteriorate substantially even when ID classification performance remains largely unchanged. Specifically, certain update mechanisms consistently induce up to 23.82% Area Under Receiver Operating Curve (AUROC) drop compared to the source DNN. Similarly, detector robustness varies significantly across detector families, with feature-based (average ΔAUROC of -0.0338 for RMDS(Ren et al. 2021)) approaches generally exhibiting greater resilience than confidence-based (average ΔAUROC of -.1429 for GEN(Liu, Lochman, and Christopher 2023)) methods. Finally, we show that changes in representation geometry provide a strong explanatory as well as monitoring signal for OOD performance prediction.

Summary of Novel Contributions

- We introduce the problem of OOD fragility under KU and provide the first systematic study of how post-deployment model modifications affect post-hoc OOD detection. We unify MU and TTA under a common KU framework and propose quantitative measures for evaluating OOD fragility.
- We perform a comprehensive empirical analysis across 10 MU and 7 TTA update mechanisms, 11 post-hoc OOD detectors, and CIFAR10 and CIFAR100 OOD benchmarks. We observe that KU can degrade average AUROC by upto 23.82% for the evaluated update mechanisms but it is largely dependent on the type of update, OOD detector, and OOD dataset. Interestingly, for MU methods, it can even improve OOD detection slightly when given access to retain dataset.
- We show that distortion in the feature-space and logit-structure can be reliable signals for the degradation in OOD detection performance. Specifically, alignment (measured with CKA) between the pre- and post-update features (and similarly logits) show > 0.9 correlation value for both MU and TTA.

In the rest of the work, we given an overview of the experimental setup in Section 2; impact of MU and TTA on OOD detection in Section 3 and 4 respectively; analyze the commonalities between MU and TTA in terms of post-update OOD detection performance; and finally, conclude with concise discussion of the limitations and future work.

2 Experimental Setup

Evaluation Pipeline

Our experimental pipeline consists of ❶ a source model trained on the in-distribution dataset, ❷ a post-deployment update method, ❸ a post-hoc OOD detector, and ❹ OOD benchmark datasets. Given a pretrained classifier f_θ, a knowledge update procedure $\mathcal{U}$ produces an updated model $f_{\theta'}$, where $\theta' = \mathcal{U}(\theta)$. We evaluate OOD detection performance before and after the update. For an OOD detector d and evaluation metric M, the change in OOD performance is computed as $\Delta M = M(d; f_{\theta'}) - M(d; f_\theta)$. For metric like AUROC, negative values of ΔM indicate degradation after the update. For metrics such as False Positive Rate (FPR), positive values indicate degradation. Throughout the paper, we use these performance differences to quantify OOD fragility.

Knowledge Update Methods

We consider two representative families of post-deployment model updates: MU and TTA.

MU. We evaluate SalUn (Fan et al. 2023) together with Fine-Tuning (FT) (Warnecke et al. 2021), Random Label (RL) (Golatkar, Achille, and Soatto 2020), Random Label Proximal (RL-Proximal) (Fan et al. 2023), Gradient Ascent (GA) (Thudi et al. 2022), and Gradient Ascent Prune (GA-Prune) (Jia et al. 2023). For class-unlearning experiments, we additionally consider Neggrad+ (Kurmanji et al. 2023), Boundary Shrinking (BS) and Boundary Expanding (BE), where the last two are part of Boundary Unlearning (BU) (Chen et al. 2023b). Retraining from scratch on the retained dataset is treated as the golden standard. Following prior MU evaluations, hyperparameters are selected using a seed-based model selection protocol and is detailed in Appendix 8.

TTA. We evaluate representative TTA methods: normalization-statistics adaptation (Norm) (Li et al. 2018), entropy minimization (TENT) (Wang et al. 2021), augmentation-based marginal entropy minimization (MEMO) (Zhang, Levine, and Finn 2022), efficient anti-forgetting adaptation (EATA) (Niu et al. 2022), sharpness-aware reliable adaptation (SAR) (Niu et al. 2023), and continual teacher–student adaptation (CoTTA) (Wang et al. 2022b). Detailed implementation settings are provided in the Appendix 9.

OOD Detectors

We evaluate a total of 11 of post-hoc OOD detectors that exploit different signals from the underlying model. Specifically, we evaluate Maximum Softmax Probability (MSP) (Hendrycks and Gimpel 2017), Generalized Entropy (GEN) (Liu, Lochman, and Christopher 2023), Energy (Liu et al. 2020), MaxLogit (Hendrycks et al. 2022), ReAct (Sun, Guo, and Li 2021), k-nearest neighbors (kNN) (Sun et al. 2022), relative Mahalanobis distance score (RMDS) (Ren et al. 2021), Residual (Wang et al. 2022a), ViM (Wang et al.

2022a), NECO (Ben Ammar et al. 2024), and fast decision-boundary distance (fDBD) (Liu and Qin 2024).

Datasets

For MU experiments, we use CIFAR-10 and CIFAR-100 as the in-distribution datasets. OOD performance is evaluated on MNIST, SVHN, Textures, Places365, and Tiny-ImageNet, following standard OOD detection protocols (Yang et al. 2022; Zhang et al. 2023). For TTA experiments, we adapt using standard distribution-shift benchmarks CIFAR10C and CIFAR100C (Hendrycks and Dietterich 2019) and afterwards assess OOD detection performance using the same detector suite.

3 Impact of MU on OOD Detection

Overall Impact of MU on OOD Detection

We first ask whether post-deployment unlearning degrades OOD detection at all, and if so, how severity scales with the type of the update. We consider both *instance unlearning* (removing 4,500 or 18,000 training examples) and *class unlearning* (removing an entire class), since the latter more explicitly reshapes the decision region for the retained classes.

Figure 2 shows that every evaluated CIFAR-10 unlearning method degrades OOD – detection, under both instance removal (mean ΔAUROC -0.1181, ΔFPR $+0.2069$ at 18,000 removed samples) and class removal (mean ΔAUROC -0.1033, ΔFPR $+0.1957$). Severity scales with the forget-set size for instance unlearning. For class unlearning, accessing the retain set generally proves helpful to preserve OOD detection performance as evidenced by BE being outperformed by other update methods. Gradient-ascent-based methods cause the largest, most consistent failures: GA-Prune reaches -0.3268 ΔAUROC / $+0.4444$ ΔFPR under instance unlearning and -0.3422 ΔAUROC under class unlearning. RL, RL-Proximal, and SalUn fall between these extremes for instance unlearning; for class unlearning, RL and RL-Proximal (along with WoodFisher) instead track the source baseline closely, and RL even slightly *improves* average AUROC ($+0.0013$). This indicates that degradation is not an unavoidable consequence of unlearning but depends on the specific unlearning mechanism. CIFAR-100 shows the same qualitative ordering of methods but with reduced gaps (e.g. mean ΔAUROC -0.0724 under instance unlearning), consistent with its weaker source operating point (0.7598 average AUROC vs. 0.9004 for CIFAR-10); we therefore treat CIFAR-10 as primary evidence and report full CIFAR-100 results in Appendix 11.

Our dataset-centric analysis shows that unlearning-induced OOD degradation is uneven across distribution shifts and depends on both the metric and the unlearning mechanism. Under instance unlearning, AUROC falls on all five OOD datasets, with Textures the most sensitive (-0.1384) and MNIST the most stable (-0.0842); under FPR, however, SVHN degrades most, so the sensitivity ranking is metric-dependent. Class unlearning shows a similar pattern, with SVHN most sensitive on average (-0.1357 AUROC), driven largely by an FT-Prune failure specific to that dataset. Aggressive pruning-based methods such as GA-Prune degrade every dataset broadly (least on MNIST), whereas conservative methods (Retrain, RL, RL-Proximal, WoodFisher) stay close to the source across shifts. Overall, OOD fragility emerges from the interaction of the unlearning mechanism, the removed data, and the evaluated distribution shift, and averaged performance alone can hide these failure modes.

Key Takeaway

These summary observations affirmatively answers the **first** research question: *MU degrades OOD performance in a structured way*. OOD detection performance degrades with number of samples removed. The observations also give insight about the **second** research question: *not all unlearning methods show same level of vulnerability*. Overall, OOD fragility is not an unavoidable consequence of unlearning, but depends on the interaction of the specific unlearning mechanism and kind of distribution shift being evaluated.

Robustness of Different OOD Detection Methods

We next ask whether different post-hoc OOD detectors are equally affected by MU, which reveals how the interaction between a detector's underlying signal and the update mechanism shapes fragility. We group detectors into three families: *output-based* (MSP, GEN, Energy, MaxLogit), which depend on classifier logits; *activation-based* (ReAct), which modifies penultimate-layer activations; and *representation-based* (Residual, ViM, NECO, RMDS, kNN, FDBD), which exploit feature-space geometry. For each detector we average performance over the five OOD datasets and measure its change relative to the source model.

Table 1: Average detector-family robustness under CIFAR-10 instance unlearning (18,000 removed samples), averaged over six unlearning methods, three seeds, and five OOD datasets. Per-detector values in Appendix 11.

Family	Avg. ΔAUROC	Avg. ΔFPR@95
Output-based (MSP, GEN, Energy, MaxLogit)	-0.1394	$+0.2589$
Activation-based (ReAct)	-0.1685	$+0.3234$
Representation-based (Residual, ViM, NECO, RMDS, kNN, FDBD)	-0.1175	$+0.2158$
RMDS *(most stable)*	-0.0338	$+0.1103$
kNN	-0.0745	$+0.1427$

Table 1 shows that detector robustness depends strongly on detector family. Averaged across six instance-unlearning methods, RMDS is the most stable detector (ΔAUROC -0.0338), followed by kNN (-0.0745) and ViM (-0.1146); output- and activation-based detectors degrade roughly 2–$5\times$ more on average. The same ordering holds under class unlearning: RMDS remains the most stable detector overall (ΔAUROC -0.0400), with ViM and kNN also comparatively stable, while NECO, Residual, FDBD, and ReAct show larger changes.

No detector family, however, is uniformly robust across *all* update mechanisms. Under SalUn and RL-Proximal, the representation-based family degrades only mildly (-0.0421 and -0.0593 ΔAUROC, respectively) compared to output-based detectors (-0.1166 and -0.1095 respectively) and Re-

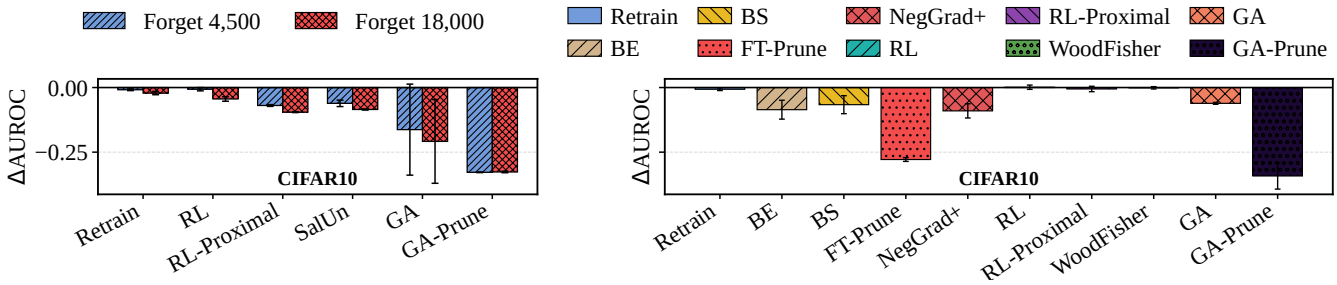

Figure 2: Overall Impact of MU on OOD Detection Performance. The left figure corresponds to instance unlearning and the right figure corresponds to class unlearning.

Act (-0.2150 and -0.2522 respectively); RMDS and kNN in fact remain close to or slightly exceed their source AUROC under these updates. RL inverts this ranking entirely: output-based detectors stay stable while Residual alone degrades badly (0.8896 to 0.6505), dragging down the representation-based family average for that specific update. GA-Prune is the one mechanism that degrades every family comparably severely (-0.3130 to -0.3423), though RMDS remains the single most resilient detector even here (0.9000 to 0.8331). Across MU mechanisms overall, RL, RL-Proximal, retraining, and WoodFisher preserve most detector scores, while FT-Prune and GA-Prune cause broad degradation across families.

Key Takeaway

While no detector family is universally robust, feature space distance based approaches (RMDS, kNN) show the least degradation in performance on average. As a rule of thumb, post-unlearning reliability depends on whether the update preserves the specific model property (logits, activations, or feature geometry) that a detector relies on to separate ID from OOD samples.

Factors Driving Post-Unlearning OOD Fragility

We next investigate which update-induced changes accompany OOD degradation, focusing on the penultimate feature and logit spaces that post-hoc detectors operate on. For source model f_θ and unlearned model $f_{\theta'}$, we evaluate both on the same retained examples and measure linear centered kernel alignment (CKA), paired feature/logit displacement, prediction agreement, confidence change, and k-nearest-neighbor preservation. These quantities capture changes to retained data induced by MU.

Representation alignment predicts fragility. Figure 20 shows a strong association between preserved source-model geometry and OOD reliability: feature CKA correlates with AUROC/FPR degradation at $-0.939/-0.915$ (class unlearning) and $-0.886/-0.943$ (instance unlearning), with logit CKA producing comparable associations. Relatively stable MU methods (Retrain, RL, RL-Proximal, WoodFisher) retain feature and logit CKA above 0.97 and stay close to source-model OOD performance, while the most damaging methods collapse it (GA-Prune: CKA 0.41–0.47, AUROC drop

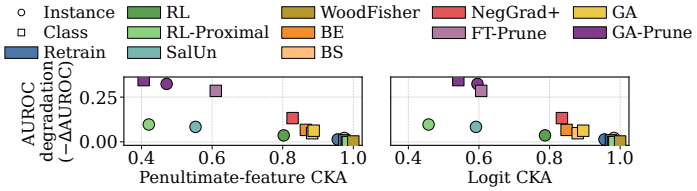

Figure 3: Association between feature-CKA alignment and OOD degradation in the sampled mechanistic panel. Circles: CIFAR-10 instance unlearning (18,000 removed); squares: CIFAR-10 class unlearning (class 0). Higher $-\Delta$AUROC indicates greater degradation.

0.32–0.34). Critically, retained classification accuracy is not a reliable predictor for this: Boundary Expanding/Shrinking preserve $> 99\%$ prediction agreement while feature CKA falls to $0.87/0.88$ (AUROC drops of $0.067/0.048$) – confirming that ID utility and OOD-relevant feature geometry can diverge sharply.

Logit-space distortion is the most consistent single predictor. Normalized logit drift (ℓ_2 distance between logit-vectors, normalized by source logit norm) correlates with AUROC degradation at $+0.943$ (instance) and $+0.879$ (class), and with FPR degradation at $+1.000$ and $+0.903$ – the strongest associations we observe. The most damaging methods also show the largest confidence collapse (GA-Prune: -0.79 instance, -0.72 class; FT-Prune: -0.45 class), versus < 0.004 for stable methods. This shows an update can preserve the predicted class while still shifting the logit magnitudes and margins that output-based detectors rely on – degrading OOD detection before classification accuracy visibly suffers, and connecting this analysis to the detector-family asymmetry in Section 11.

Local neighborhood structure corroborates this, but raw distance alone is insufficient. kNN overlap correlates with AUROC/FPR degradation at $-0.806/-0.842$ under class unlearning (e.g., WoodFisher preserves 96% of source neighbors vs. 3–4% for FT-Prune/GA-Prune), confirming that severe failures coincide with disrupted local geometry, not just shifted scores. However, raw coordinate displacement alone does not explain fragility: retraining shows high CKA alongside large paired feature displacement and low exact-neighbor overlap, since representations can rotate or rescale without losing their relational structure.

Table 2: Overall source-relative OOD degradation under TTA, averaged across architectures, severities, OOD datasets, and detectors (mean ± SD across three seeds). Full CIFAR-100 breakdown in Appendix 11.

Metric	Norm	TENT	MEMO	EATA	SAR	CoTTA
ΔAUROC	−0.1043 ±0.0000	−0.1081 ±0.0005	−0.1079 ±0.0023	−0.0938 ±0.0005	−0.2016 ±0.0090	−0.0000 ±0.0002
ΔFPR@95	+0.1995 ±0.0000	+0.1958 ±0.0011	+0.2008 ±0.0041	+0.1816 ±0.0016	+0.2673 ±0.0019	+0.0001 ±0.0000

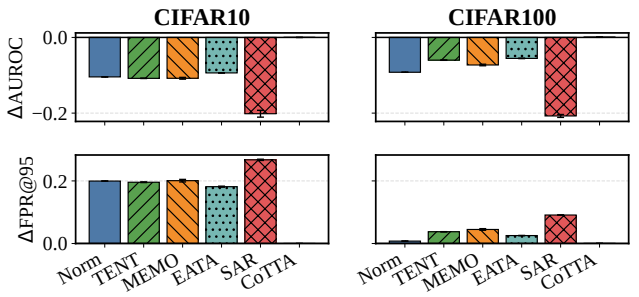

Figure 4: Overall source-relative AUROC and FPR@95 change under TTA, by method and ID dataset (single run per configuration).

Key Takeaway

Post-unlearning OOD fragility is associated with distortion of the retained-data representation and score space, rather than with retained-accuracy loss alone. These results establish mechanistic but not a causal association. They offer a possible internal, OOD-data-free signal for anticipating OOD fragility after a model update.

4 Impact of TTA on OOD Detection

Overall Effect and Dependence on Corruption Severity

Table 2 and Figure 4 show that, across the five methods that induce substantial degradation in OOD detection performance, adaptation consistently reduces mean OOD detection relative to the source model, averaged over three seeds. On CIFAR-10, mean AUROC falls by 0.0938–0.2016 and mean FPR rises by 0.1816–0.2673; EATA is the least damaging (ΔAUROC -0.0938 ± 0.0005) and SAR the most (-0.2016 ± 0.0090), with Norm, TENT, and MEMO in a narrower middle band (-0.1043 to -0.1081). The same ordering holds on CIFAR-100. CoTTA is qualitatively different: its OOD behavior is essentially unchanged on both CIFAR-10 (ΔAUROC$= -0.0000 \pm 0.0002$, ΔFPR@95$= +0.0001 \pm 0.0000$) and CIFAR-100 (ΔAUROC$= +0.0007 \pm 0.0005$, ΔFPR@95$= +0.0004 \pm 0.0002$). Improvement at the individual evaluation-cell level remains uncommon: on CIFAR-10, only 3.0–6.0% of cells improve in AUROC for EATA, MEMO, TENT, and SAR (Norm: 15.7%), confirming degradation is not driven by a small number of extreme detector failures.

As with MU, degradation generally scales with the aggressiveness, in this case, severity of the corruption the model is adapted to (Figure 5): for Norm, TENT, MEMO, and EATA, AUROC loss on CIFAR-10 roughly doubles from severity 1 to severity 5 (e.g., EATA: -0.0620 ± 0.0010 to -0.1229 ± 0.0008; TENT: -0.0777 to -0.1455), with the same direction on CIFAR-100 (e.g., EATA: -0.0190 to -0.0909; TENT: -0.0179 to -0.1074). SAR is the exception: it is uniformly fragile across severities on CIFAR-100 (-0.1982 to -0.2149), but on CIFAR-10 is highly non-monotonic (severity 1: -0.2455 ± 0.0065; severity 3: -0.3440 ± 0.0331; severity 5: -0.0154 ± 0.0005), reproduced across all three seeds and visible as the sharp reversal in Figure 5; the severity-3 estimate carries the largest cross-seed variability, so we treat this as configuration-specific rather than evidence that stronger corruption generally repairs SAR's OOD behavior. CoTTA remains near the source baseline at every severity on both benchmarks (CIFAR-10 ΔAUROC: $+0.0001$, -0.0001, -0.0001; CIFAR-100: -0.0003, $+0.0014$, $+0.0011$).

Key Takeaway

TTA degrades severly OOD detection except from CoTTA, mirroring the MU finding that task-directed model updates reduce OOD reliability in a structured way. The degradation scales with the severity of the corruption except for SAR's non-monotonic CIFAR-10 behavior (Figure 5)

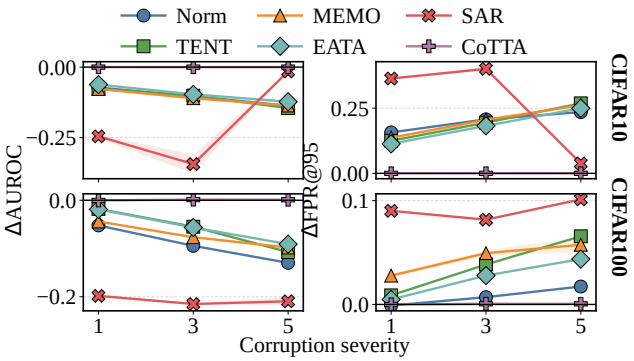

Figure 5: Severity dependence of TTA-induced OOD change, by method. Note SAR's non-monotonic CIFAR-10 trajectory (red), in contrast to the other four methods' roughly monotonic degradation.

Dependence on OOD Dataset and Detector

Fragility is not uniform across OOD distributions or scoring rules, similar to the dataset- and detector-dependence observed under MU.[1] For CIFAR-10 models, SVHN is the

[1] Architecture also modulates fragility: RepVGG-A2 is more fragile than ResNet-56 under every drifting method on CIFAR-10 (e.g., EATA: -0.1231 vs. -0.0656; SAR: -0.2692 vs. -0.1132), with a smaller but consistent gap on CIFAR-100; CoTTA remains near the source baseline for both architectures (Appendix 11).

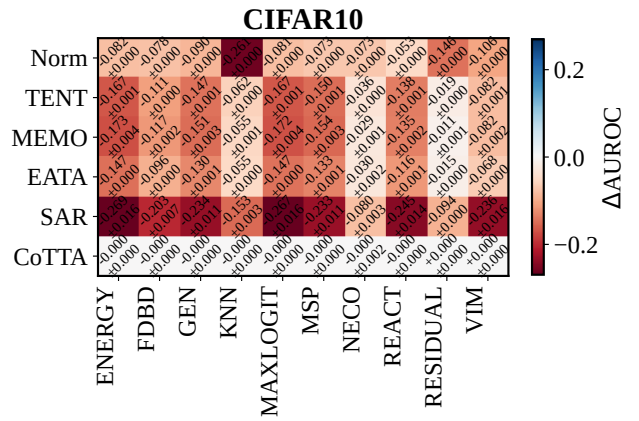

Figure 6: TTA-induced AUROC change by detector, for CIFAR10. Residual and NECO are comparatively stable under EATA/MEMO/TENT; energy, MaxLogit, and kNN (the latter specifically under Norm) degrade substantially more; SAR degrades all detectors.

most sensitive dataset for entropy-oriented methods (mean ΔAUROC: -0.1457 EATA, -0.1492 MEMO, -0.1655 TENT), roughly double their loss on TIN (-0.0719 to -0.0882); SAR causes large AUROC reductions on every OOD dataset, with its SVHN mean (-0.2309 ± 0.0328) showing the largest dataset-level seed variability we observe. For CIFAR-100 models, Textures is typically the most fragile OOD dataset (-0.0979 EATA to -0.2505 SAR), while MNIST is comparatively stable (e.g., $+0.0072$ for EATA, -0.0120 for TENT). CoTTA remains close to the source baseline across all evaluated OOD datasets and detectors, consistent with its near-zero aggregate change (Section 4).).

Detector choice interacts with the adaptation mechanism rather than failing uniformly (Figure 6). For CIFAR-10, residual detection is comparatively stable under EATA, MEMO, and TENT (ΔAUROC -0.0108 to -0.0195), and NECO is similarly stable (-0.0291 to -0.0363), while energy and MaxLogit degrade far more (roughly -0.147 to -0.173) – close to a $5\times$ gap within the same update mechanism. SAR degrades every detector we measured, including energy (-0.2695 ± 0.0156), MaxLogit (-0.2674 ± 0.0155), MSP (-0.2334 ± 0.0110), and ViM (-0.2362 ± 0.0156). Norm shows a distinct failure profile from the other four methods: kNN is by far its most fragile detector (ΔAUROC -0.2612 CIFAR-10, -0.2809 CIFAR-100), an order of magnitude larger than Norm's own aggregate AUROC loss (-0.1043), showing that method-level averages can conceal sharp method×detector interactions. This contrasts with the MU setting, where feature-space distance based kNN was one of the most broadly stable detectors across nearly all unlearning mechanisms (Section 11); under TTA, representation-based detectors are only *selectively* robust (residual, NECO), while kNN is stable under EATA/TENT/MEMO but among the single most fragile detectors under Norm. This indicates that detector robustness is not a fixed property of the representation-based family, but depends on which specific

Table 3: CIFAR-10 representation preservation (feature CKA, prediction agreement; single run) and source-relative OOD change (mean ± SD across three seeds) under TTA, per method. Full CIFAR-10/CIFAR-100 table with all measured quantities in Appendix 11.

Method	Feat. CKA	Pred. Agree.	ΔAUROC	ΔFPR@95
CoTTA	1.0000	1.0000	-0.0000 ± 0.0002	$+0.0001 \pm 0.0000$
EATA	0.9097	0.9155	-0.0938 ± 0.0005	$+0.1816 \pm 0.0016$
TENT	0.9078	0.9142	-0.1081 ± 0.0005	$+0.1958 \pm 0.0011$
MEMO	0.8997	0.9049	-0.1079 ± 0.0023	$+0.2008 \pm 0.0041$
Norm	0.8242	0.8528	-0.1043 ± 0.0000	$+0.1995 \pm 0.0000$
SAR	0.7438	0.7726	-0.2016 ± 0.0090	$+0.2673 \pm 0.0019$

mechanism the update perturbs. Full dataset-level breakdowns for both benchmarks are in Appendix 11.

Key Takeaway

TTA-induced OOD fragility is highly dataset- and detector-dependent, so aggregate averages can hide sharp failures. Unlike MU, robustness here is mechanism-specific: kNN is Norm's most fragile detector yet stable under EATA/TENT/MEMO, so no detector family is reliable by default.

Factors Driving Post-Adaptation OOD Fragility

Paralleling the MU analysis (Section 11), we examine whether the representation changes induced by TTA explain its OOD degradation. For each of TTA method and configuration, we compare the final adapted checkpoint against its source model on a stratified sample of 5,000 clean ID test examples, measuring linear CKA in feature and logit space, prediction agreement, principal-subspace, label purity, class-prototype displacement, and kNN overlap.

Adaptation methods induce distinct degrees of geometric drift. Table 3 shows preservation of source geometry varies sharply by method. CoTTA is the clear preservation extreme on both benchmarks: feature/logit CKA and prediction agreement round to 1.0000 on CIFAR-10 and $\geq$ 0.9975 on CIFAR-100, with negligible OOD change (ΔAUROC$= -0.0000/+0.0007$, mean across three seeds). Among the drifting methods, EATA and TENT preserve geometry most (feature CKA 0.91, prediction agreement $\sim$ 0.915), followed by MEMO (0.90); all three still degrade mean AUROC (-0.09 to -0.11), showing even $\sim$90% representation preservation does not guarantee OOD separation. Norm shifts internals more sharply (CKA 0.8242), and SAR is least aligned (CKA 0.7438) and most damaging (ΔAUROC -0.2016 ± 0.0090). CIFAR-100 shows sharper separation: EATA/TENT/MEMO retain CKA $\sim$ 0.74, Norm falls to 0.6337, and SAR collapses to 0.2577 (18.6% prediction agreement) with its largest AUROC drop (-0.2075 ± 0.0038). Full CIFAR-100 values are in Appendix 11.

Representation preservation correlates with OOD preservation. On CIFAR-10, principal-subspace distance correlates most strongly with ΔAUROC ($\rho = -0.956$); feature/logit CKA, prediction agreement, and kNN overlap are

also strongly associated ($|\rho| \geq 0.81$). On CIFAR-100, local and class-conditional geometry are the strongest predictors ($|\rho| \geq 0.966$ for kNN label-purity change, logit/feature CKA, kNN overlap, and class-prototype displacement), converging global, local, and class-level evidence on the same conclusion. Full correlation tables are in Appendix 11.

Severity compounds geometric drift, mirroring the MU pattern. On CIFAR-100, feature CKA falls monotonically with severity (0.76 to 0.68 to 0.62), aligned with the monotonic AUROC decline in Section 4. CIFAR-10 mirrors SAR's non-monotonic pattern: CKA is 0.90, 0.85, 0.89 at all three severities (severity level 1,3, and 5), recovering alongside SAR's partial AUROC recovery at severity 5. This severity-coupling reinforces that geometric drift is not merely a static byproduct of a given adaptation method, but scales with the magnitude of distribution shift the method is forced to accommodate – strengthening the case that geometric disruption, rather than adaptation, is the operative mechanism behind OOD degradation under TTA.

Key Takeaway

TTA-induced OOD fragility aligns with disruption of source representation geometry: CoTTA preserves both and shows no OOD change; EATA/TENT retain the most alignment among drifting methods and degrade least; SAR is least aligned and most damaging. This relationship strengthens with severity, though associations remain descriptive (single run per configuration).

Adaptation Accuracy Does Not Predict OOD Preservation

As with MU, where retained-set accuracy failed to expose OOD-relevant representation drift, accuracy on corrupted data is not a reliable proxy for OOD preservation under TTA. The Pearson correlation between mean corruption accuracy across seeds and mean ΔAUROC across all configurations is weak on CIFAR-10 ($r = -0.191$) and weak-to-moderate on CIFAR-100 ($r = +0.243$); both relationships are weak and inconsistent in sign across datasets. Method rankings reinforce this: on CIFAR-10, MEMO attains the highest average corruption accuracy (0.8758 ± 0.0005) yet loses 0.1079 AUROC, while EATA achieves similar accuracy (0.8715 ± 0.0002) with smaller OOD loss (0.0938); SAR attains reasonable accuracy (0.8533 ± 0.0001) but the largest OOD loss (0.2016). CoTTA matches Norm's CIFAR-10 corruption accuracy (0.8224 ± 0.0000) while losing essentially no OOD performance, underscoring that corruption accuracy alone does not predict OOD preservation even between methods with comparable target-task utility. Selecting a TTA method by corrupted-ID accuracy alone can therefore leave a substantial, unmeasured OOD-detection cost.

5 Overall Impact of KU on OOD Detection

Despite operating on entirely different objectives, MU and TTA produce a common pattern: post-deployment updates directed at a target task (forgetting or corruption robustness) can substantially degrade OOD detection without a corresponding loss in that target task's own metric (retained accuracy for MU; corruption accuracy for TTA). This parallels our finding that update aggressiveness (number of samples removed or severity of distribution shift) generally aligns with degradation, though not monotonically for every method (GA under MU; SAR under TTA), and that the specific OOD dataset and detector most affected shifts across conditions, so aggregate metrics can conceal important failure modes. The two settings diverge, however, in which detectors remain robust: representation-distance based detectors (RMDS, kNN) are broadly the most stable family under MU, whereas under TTA robustness is more selective (residual, NECO stable; kNN sometimes the most fragile detector), suggesting that no single detector family is universally safe against post-deployment updates. Both settings also converge mechanistically. Our analyses of representations for retained set (for MU) and adapted (for TTA) show that degradation is not a property of the update paradigm, but of how much a given method perturbs the source model's feature and logit geometry: methods that preserve this geometry with high fidelity – Retrain, RL, RL-Proximal, and WoodFisher under MU; CoTTA under TTA – leave OOD detection close to source-model performance, while methods that substantially reorganize this geometry – GA-Prune under MU; SAR under TTA – produce the largest OOD failures. This convergence is notable given that MU and TTA optimize entirely different objectives using entirely different data (retained ID data vs. unlabeled corrupted data).

6 Conclusion

In this work, we have shown that KU affects the OOD detection performance drastically through comprehensive study. The results of our study motivate treating post-deployment KU as inherently multi-objective intervention: not only achieving its stated goal (e.g., forgetting, adaptation accuracy) but also preserving the representational structure that downstream OOD detectors rely on. Reporting only the target-task metric, as is standard practice in both the MU and TTA literature, can obscure a substantial and otherwise invisible degradation in OOD reliability. Despite a comprehensive experimental campaign, we believe there are still many interesting questions left for investigation. An immediate task is extension of the study to include KU mechanisms as well as OOD detectors that fundamentally differ from the ones presented here. Apart from that, it would be interesting to study how the KU and OOD detectors interact in high-resolution natural image setting (e.g., Imagenet). Finally, in this era of foundation models, it is necessary to investigate how stable the vision language models are across different scales in terms of OOD detection performance under KU. We leave these question for future work.

References

Ben Ammar, M.; Belkhir, N.; Popescu, S.; Manzanera, A.; and Franchi, G. 2024. Neco: Neural collapse based out-of-distribution detection. In *International Conference on Learning Representations*, volume 2024, 1182–1211.

Bourtoule, L.; Chandrasekaran, V.; Choquette-Choo, C. A.;

Jia, H.; Travers, A.; Zhang, B.; Lie, D.; and Papernot, N. 2021. Machine Unlearning. In *Proceedings of IEEE Symposium on Security and Privacy (S&P)*, 141–159. IEEE.

Chen, L.; Zhang, Y.; Song, Y.; Shan, Y.; and Liu, L. 2023a. Improved Test-Time Adaptation for Domain Generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24172–24182.

Chen, M.; Gao, W.; Liu, G.; Peng, K.; and Wang, C. 2023b. 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 (CVPR)*, 7766–7775.

Döbler, M.; Marsden, R. A.; and Yang, B. 2023. Robust Mean Teacher for Continual and Gradual Test-Time Adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7704–7714.

Fan, C.; Liu, J.; Zhang, Y.; Wong, E.; Wei, D.; and Liu, S. 2023. SalUn: Empowering Machine Unlearning via Gradient-based Weight Saliency in Both Image Classification and Generation. *arXiv preprint arXiv:2310.12508*.

Fan, C.; Liu, J.; Zhang, Y.; Wong, E.; Wei, D.; and Liu, S. 2024. SalUn: Empowering Machine Unlearning via Gradient-based Weight Saliency in Both Image Classification and Generation. In *The Twelfth International Conference on Learning Representations*.

Ginart, A.; Guan, M.; Valiant, G.; and Zou, J. Y. 2019. Making AI Forget You: Data Deletion in Machine Learning. In Wallach, H.; Larochelle, H.; Beygelzimer, A.; d'Alché-Buc, F.; Fox, E.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Golatkar, A.; Achille, A.; and Soatto, S. 2020. Eternal Sunshine of the Spotless Net: Selective Forgetting in Deep Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9304–9312.

Hendrycks, D.; Basart, S.; Mazeika, M.; Zou, A.; Kwon, J.; Mostajabi, M.; Steinhardt, J.; and Song, D. 2022. Scaling Out-of-Distribution Detection for Real-World Settings. In Chaudhuri, K.; Jegelka, S.; Song, L.; Szepesvari, C.; Niu, G.; and Sabato, S., eds., *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, 8759–8773. PMLR.

Hendrycks, D.; and Dietterich, T. 2019. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*.

Hendrycks, D.; and Gimpel, K. 2017. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. *ICLR*.

Jia, J.; Liu, J.; Ram, P.; Yao, Y.; Liu, G.; Liu, Y.; Sharma, P.; and Liu, S. 2023. Model Sparsity Can Simplify Machine Unlearning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Kodge, S.; Saha, G.; and Roy, K. 2024. Deep Unlearning: Fast and Efficient Gradient-free Class Forgetting. *Transactions on Machine Learning Research*.

Kurmanji, M.; Triantafillou, P.; Hayes, J.; and Triantafillou, E. 2023. Towards Unbounded Machine Unlearning. In Oh, A.; Naumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; and Levine, S., eds., *Advances in Neural Information Processing Systems*, volume 36, 1957–1987. Curran Associates, Inc.

Lee, D.; Yoon, J.; and Hwang, S. J. 2024. BECoTTA: Input-Dependent Online Blending of Experts for Continual Test-Time Adaptation. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, 27072–27093. PMLR.

Lee, K.; Lee, K.; Lee, H.; and Shin, J. 2018. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 31.

Li, Y.; Wang, N.; Shi, J.; Hou, X.; and Liu, J. 2018. Adaptive Batch Normalization for Practical Domain Adaptation. *Pattern Recognition*, 80: 109–117.

Liu, L.; and Qin, Y. 2024. Fast Decision Boundary based Out-of-Distribution Detector. In *Forty-first International Conference on Machine Learning*.

Liu, W.; Wang, X.; Owens, J.; and Li, Y. 2020. Energy-based Out-of-distribution Detection. In *Thirty-fourth Conference on Neural Information Processing Systems*, 21464–21475.

Liu, X.; Lochman, Y.; and Christopher, Z. 2023. GEN: Pushing the Limits of Softmax-Based Out-of-Distribution Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Nguyen, A.; Yosinski, J.; and Clune, J. 2015. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 427–436.

Nguyen, T. T.; Huynh, T. T.; Nguyen, P. L.; Liew, A. W.-C.; Yin, H.; and Nguyen, Q. V. H. 2022. A Survey of Machine Unlearning. *arXiv preprint arXiv:2209.02299*.

Niu, S.; Wu, J.; Zhang, Y.; Chen, Y.; Zheng, S.; Zhao, P.; and Tan, M. 2022. Efficient Test-Time Model Adaptation without Forgetting. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, 16888–16905. PMLR.

Niu, S.; Wu, J.; Zhang, Y.; Wen, Z.; Chen, Y.; Zhao, P.; and Tan, M. 2023. Towards Stable Test-Time Adaptation in Dynamic Wild World. In *International Conference on Learning Representations*.

Papyan, V.; Han, X.; and Donoho, D. L. 2020. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40): 24652–24663.

Ren, J.; Fort, S.; Liu, J.; Roy, A. G.; Padhy, S.; and Lakshminarayanan, B. 2021. A simple fix to mahalanobis distance for improving near-ood detection. *arXiv preprint arXiv:2106.09022*.

Sayyed, A. S.; Shahriar, R.; Zhang, M.; Swami, A.; De Lucia, M.; Bastian, N.; and Restuccia, F. 2025. Out-of-distribution detection in computer vision: A comprehensive survey and research challenges. *ACM Computing Surveys*.

Sun, Y.; Guo, C.; and Li, Y. 2021. React: Out-of-distribution detection with rectified activations. *Advances in neural information processing systems*, 34: 144–157.

Sun, Y.; Ming, Y.; Zhu, X.; and Li, Y. 2022. Out-of-distribution Detection with Deep Nearest Neighbors. *ICML*.

Thudi, A.; Deza, G.; Chandrasekaran, V.; and Papernot, N. 2022. Unrolling SGD: Understanding Factors Influencing Machine Unlearning . In *2022 IEEE 7th European Symposium on Security and Privacy (EuroS&P)*, 303–319.

Wang, D.; Shelhamer, E.; Liu, S.; Olshausen, B.; and Darrell, T. 2021. Tent: Fully Test-Time Adaptation by Entropy Minimization. In *International Conference on Learning Representations*.

Wang, H.; Li, Z.; Feng, L.; and Zhang, W. 2022a. Vim: Out-of-distribution with virtual-logit matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4921–4930.

Wang, Q.; Fink, O.; Van Gool, L.; and Dai, D. 2022b. Continual Test-Time Domain Adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7201–7211.

Warnecke, A.; Pirch, L.; Wressnegger, C.; and Rieck, K. 2021. Machine Unlearning of Features and Labels. *arXiv preprint arXiv:2108.11577*.

Xiao, Z.; and Snoek, C. G. 2024. Beyond model adaptation at test time: A survey. *arXiv preprint arXiv:2411.03687*.

Yang, J.; Wang, P.; Zou, D.; Zhou, Z.; Ding, K.; Peng, W.; Wang, H.; Chen, G.; Li, B.; Sun, Y.; Du, X.; Zhou, K.; Zhang, W.; Hendrycks, D.; Li, Y.; and Liu, Z. 2022. OpenOOD: Benchmarking Generalized Out-of-Distribution Detection.

Yuan, L.; Xie, B.; and Li, S. 2023. Robust Test-Time Adaptation in Dynamic Scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15922–15932.

Zhang, J.; Yang, J.; Wang, P.; Wang, H.; Lin, Y.; Zhang, H.; Sun, Y.; Du, X.; Li, Y.; Liu, Z.; Chen, Y.; and Li, H. 2023. OpenOOD v1.5: Enhanced Benchmark for Out-of-Distribution Detection. *arXiv preprint arXiv:2306.09301*.

Zhang, M.; Levine, S.; and Finn, C. 2022. MEMO: Test Time Robustness via Adaptation and Augmentation. In *Advances in Neural Information Processing Systems*, volume 35.

Supplementary Material

7 Related Work

This section reviews the prior work most closely related to our study, first covering post-hoc OOD detection and then machine unlearning and test-time adaptation.

Out-of-Distribution Detection

OOD detection aims to identify inputs that do not belong to the training distribution of a classifier. Early approaches relied on prediction confidence, with maximum softmax probability (MSP) serving as a simple yet effective baseline (Hendrycks and Gimpel 2017). Subsequent work improved confidence-based detection through alternative scoring functions such as energy-based scores (Liu et al. 2020) and logit transformations like GEN (Liu, Lochman, and Christopher 2023) or Maxlogit (Hendrycks et al. 2022). Another major line of research exploits feature representations learned by the classifier. Representative examples include Mahalanobis-distance methods (Lee et al. 2018), nearest-neighbor approaches (Sun et al. 2022), and subspace-based techniques such as ViM (Wang et al. 2022a). Recent studies have further explored geometric and representation-aware detectors that leverage structural properties of the learned feature space. NECO(Ben Ammar et al. 2024) falls into this category utilizing the neural collapse theory (Papyan, Han, and Donoho 2020) for OOD detection. *Nearly all post-hoc OOD detectors are developed and evaluated under the assumption that the underlying classifier remains fixed after training*.

Machine Unlearning and Test-Time Adaptation

MU seeks to remove the influence of designated training data from a trained model without requiring complete retraining (Bourtoule et al. 2021). Existing approaches include gradient-ascent-based methods (Thudi et al. 2022; Kurmanji et al. 2023; Kodge, Saha, and Roy 2024), fine-tuning strategies(Golatkar, Achille, and Soatto 2020), parameter pruning techniques(Jia et al. 2023), and more recent saliency-guided approaches such as SalUn (Fan et al. 2024). Other methods focus on class-level forgetting through explicit decision-boundary manipulation (Chen et al. 2023b). Prior work primarily evaluates unlearning through forgetting efficacy, utility preservation, privacy guarantees, or membership-inference resistance (Nguyen et al. 2022).

TTA adapts a pretrained model to an unlabeled target distribution during inference without labeled source data. AdaBN (Li et al. 2018) corrects distribution shifts using target-domain batch-normalization statistics. TENT (Wang et al. 2021) minimizes prediction entropy by updating only normalization-layer affine parameters, while MEMO (Zhang, Levine, and Finn 2022) performs instance-wise adaptation by minimizing the entropy of predictions aggregated across augmented views. EATA (Niu et al. 2022) selects reliable, non-redundant samples and constrains important parameters through Fisher-based regularization. SAR (Niu et al. 2023) combines reliable-sample selection with sharpness-aware optimization to stabilize adaptation under small batches, mixed shifts, and label imbalance. ITTA (Chen et al. 2023a) optimizes lightweight auxiliary parameters using a learnable consistency objective while keeping the source network fixed. Continual TTA addresses nonstationary streams with sequential domain shifts. CoTTA (Wang et al. 2022b) uses a teacher model, augmentation-averaged predictions, and stochastic source-parameter restoration to limit errors. RoTTA (Yuan, Xie, and Li 2023) combines robust normalization, a category-balanced memory bank, and time-aware teacher–student learning. RMT (Döbler, Marsden, and Yang 2023) uses robust mean-teacher learning with symmetric cross-entropy and contrastive alignment, while BECoTTA (Lee, Yoon, and Hwang 2024) routes inputs through low-rank domain experts for parameter-efficient continual adaptation.

Although OOD detection, MU, and TTA are well studied, they have largely developed independently. OOD methods typically assume a static model, while MU and TTA primarily assess post-update classification performance. Consequently, it remains unclear whether model updates preserve OOD reliability, even though they may alter the feature geometry, calibration, and representation statistics on which many detectors depend. To the best of our knowledge, this is the first systematic study of OOD fragility under both MU and TTA.

8 Hyperparameter Selection

We selected hyperparameters without using any of the out-of-distribution (OOD) evaluation datasets or OOD-detection scores. Selection was performed on seed 1, separately for each experimental setting, and the selected configuration was then held fixed for seeds 2 and 3. Thus, the OOD results measure the consequences of configurations selected for their primary objective—accurate test-time adaptation or effective unlearning—rather than configurations tuned to improve OOD detection.

Test-time adaptation

For test-time adaptation (TTA), we swept each method independently for every dataset, architecture, and corruption severity. The adaptation datasets were CIFAR-10-C and CIFAR-100-C; the architectures were ResNet-56 and RepVGG-A2; and the evaluated severities were $\{1, 3, 5\}$. For each dataset–architecture–method–severity tuple, we selected the checkpoint with the highest classification accuracy on the corresponding corrupted test stream. A strict greater-than comparison was used, so an exact tie retained the first configuration encountered in the sweep. Only after this selection were the chosen checkpoints evaluated on the clean in-distribution test set and the five OOD datasets. Hyperparameters were therefore allowed to differ between architectures and severities, but never between OOD datasets or OOD scoring rules.

Table 4 gives the complete TTA search space. All gradient-based methods used SGD with momentum 0.9. SAR used

SAM with zero weight decay, one update step, and $\rho = 0.05$. The SAR entropy margin was $0.8 \log(1000) \approx 5.526$. Norm has no tuned hyperparameter. Parameters shown as singletons were fixed rather than selected.

We also swept ITTA over learning rates $\{10^{-2}, 10^{-3}\}$ and regularization weights $\{0.1, 0.01\}$. ITTA was excluded from the reported comparison after the corrected implementation exhibited prediction collapse; its sweep is stated here for completeness but was not used to support the paper's TTA conclusions.

Machine unlearning

For machine unlearning, hyperparameters were selected by agreement with the matched retraining baseline, which represents training from scratch after removing the forget set. Selection was performed separately for each dataset and forget setting (forget-set size for instance unlearning, or forgotten class for class unlearning). Let h denote a candidate configuration and let r denote its matched seed-1 retraining run. For class unlearning, the selection score was

$$D(h,r) = \frac{1}{5}\Big(\big|A^h_{\text{retain}} - A^r_{\text{retain}}\big| + \big|A^h_{\text{forget}} - A^r_{\text{forget}}\big| + \big|A^h_{\text{val}} - A^r_{\text{val}}\big| + \big|A^h_{\text{test}} - A^r_{\text{test}}\big| + \big|M^h_{\text{forget}} - M^r_{\text{forget}}\big|\Big), \tag{1}$$

where A denotes accuracy on the indicated split and M_{forget} is the correctness-based membership-inference metric on the forget set. We selected the candidate minimizing D. The instance unlearning selectors used the same retrain-matching principle.

Table 5 reports the candidate grids used for the final seed-1 class-unlearning sweeps. Epoch and learning-rate combinations were evaluated as Cartesian products. For RL and RL-Proximal, the saliency mask was generated for the corresponding forget setting and then held fixed within the hyperparameter sweep. NegGrad+ used equal retain and forget weights. The chosen seed-1 hyperparameters were rerun unchanged with both the data seed and training seed set to 2 and 3.

The instance-unlearning sweeps used forget-set sizes 4,500 and 18,000 and the same grids shown above except where the regime required a different RL-Proximal range: its learning-rate candidates were $\{10^{-3}, 5 \times 10^{-4}, 10^{-4}, 5 \times 10^{-5}, 10^{-5}, 5 \times 10^{-6}, 10^{-6}\}$, with mask ratios $\{0.05, 0.1\}$. Boundary expanding, boundary shrink, and NegGrad+ were evaluated as class-unlearning baselines and therefore do not define an instance-unlearning grid in this study.

Retraining is a reference procedure rather than an approximate-unlearning configuration and was trained with its prescribed training schedule. Because the selected approximate-unlearning configuration can vary by dataset and forget setting, the machine-readable selection CSVs provide the exact chosen epoch, learning rate, method-specific coefficient, and checkpoint for every reported run. This design avoids choosing one globally favorable setting and then attributing differences caused by poor primary-task tuning to the unlearning mechanism.

9 Detailed Update Methods

This appendix briefly summarizes the test-time adaptation (TTA) and machine- unlearning baselines evaluated in this work. Hyperparameters were selected using only the primary-task criteria in Section 8; OOD performance was not used for model selection.

Test-time adaptation baselines

Table 6 summarizes the source baseline and the six test-time adaptation methods, including which model components or statistics each method updates at inference time.

Table 8 reports the corruption-set classification accuracy of each adapted method for every seed. Each adapted entry averages the two architectures and corruption severities 1, 3, and 5. The Source row instead reports clean-test accuracy averaged over the two source architectures and is therefore labeled "Source (clean)"; it is repeated only to make the source utility explicit and is not directly comparable to corruption accuracy.

Machine-unlearning baselines

Table 7 summarizes the source and retraining references together with the approximate machine-unlearning methods, highlighting the update used by each method to remove the influence of the forget set.

Tables 9 and 10 report retain, forget, and test accuracy, together with membership-inference attack (MIA) correctness. For class unlearning, each seed entry is the mean $\pm$ sample standard deviation across the common forgotten-class panel (CIFAR-10 classes 0, 1, and 2; CIFAR-100 classes 0 and 15). These error terms quantify class-to-class variation within a seed, not uncertainty across seeds. A dash denotes a metric that was not evaluated. In particular, the source checkpoints store clean-test accuracy but not source retain/forget subset accuracy or source MIA. Missing WoodFisher seed entries similarly indicate unexecuted seed-level evaluations and are not imputed.

10 Performance of the KU Methods

The tables in this section summarize the primary-task performance of the evaluated update methods. Table 8 reports corruption accuracy for test-time adaptation, while Tables 9 and 10 report retain, forget, test, and membership-inference metrics for instance and class unlearning.

11 Detailed Results

This section provides detailed views of post-update OOD behavior under both test-time adaptation and machine unlearning. The tables and figures below compare overall degradation, sensitivity across OOD datasets and detectors, and the representation- and score-space changes associated with that degradation.

Detailed Test-Time Adaptation Results

This subsection complements the aggregate TTA results in the main text with breakdowns across OOD datasets, detectors, corruption severities, and internal representation

Table 4: TTA hyperparameter search space. Cartesian products were used when a method has more than one listed parameter.

Method	Hyperparameter	Candidate values
Norm	–	–
TENT	learning rate	$\{10^{-3}, 10^{-4}\}$
MEMO	learning rate; augmentations; noise std.	10^{-3}; 4; 0.01
EATA	learning rate; entropy margin	10^{-3}; 0.4
	redundancy threshold; Fisher weight	0.95; 0
SAR	learning rate; reset threshold	$\{2.5\times10^{-2}, 2.5\times10^{-3}\}$; $\{0.2, 0.05\}$
	margin; ρ; steps	$0.8\log(1000)$; 0.05; 1
CoTTA	learning rate; teacher momentum	10^{-3}; $\{0.35, 0.9\}$
	restoration probability; augmentation probability; steps	$\{0.01, 0.05\}$; 0.9; 1

Table 5: Machine-unlearning hyperparameter search space. "Epochs" refers to unlearning epochs and η to the unlearning learning rate.

Method	Hyperparameter	Candidate values
RL	epochs; η	$\{3, 5, 7, 10\}$; $\{10^{-2}, 5\times10^{-3}, 10^{-3}, 5\times10^{-4}, 10^{-4}, 5\times10^{-5}, 10^{-5}\}$
RL-Proximal	epochs; η; mask ratio	$\{3, 5, 7, 10\}$; $\{10^{-2}, 5\times10^{-3}, 10^{-3}, 5\times10^{-4}, 10^{-4}, 5\times10^{-5}, 10^{-5}\}$; $\{0.05, 0.1\}$
FT	epochs; η	$\{3, 5, 7, 10\}$; $\{10^{-3}, 5\times10^{-4}, 10^{-4}, 5\times10^{-5}, 10^{-5}, 5\times10^{-6}, 10^{-6}\}$
GA	epochs; η	same as FT
GA-Prune	epochs; η	same as FT
FT-Prune	epochs; η; pruning α	same as FT; $\{0.1, 0.2\}$
WoodFisher	noise scale α	$\{0.05, 0.2\}$
NegGrad+	epochs; η; retain/forget weights	same as FT; $1/1$
Boundary expanding/shrink	epochs; η	$\{3, 5, 7, 10\}$; $\{10^{-3}, 5\times10^{-4}, 10^{-4}, 5\times10^{-5}, 10^{-5}\}$

changes. Figures 7 and 8 show where adaptation-induced fragility is concentrated, while Figures 9, 10, and 11 relate that fragility to changes in the adapted model. Figure 12 compares adaptation accuracy with OOD preservation.

Dependence on OOD Dataset and Detector Figure 7 reports source-relative AUROC changes for each OOD dataset and adaptation method on both CIFAR benchmarks. Figure 8 provides the corresponding detector-level breakdown, exposing method–detector interactions that can be hidden by the overall averages reported in the main text.

Representation Changes and Corruption Severity Figure 9 compares representation preservation with OOD degradation across adapted checkpoints. Figure 10 summarizes the associations between the measured internal changes and OOD metrics, and Figure 11 shows how these changes vary with corruption severity.

Adaptation Accuracy and OOD Preservation Figure 12 compares corruption accuracy against source-relative OOD performance. It shows whether methods selected for strong adaptation accuracy also preserve the reliability of post-hoc OOD detectors.

Overall Impact of MU on OOD Detection

Tables 11 and 12 quantify the average AUROC and FPR@95 changes under instance unlearning. Figures 13 and 14 summarize overall fragility for instance and class unlearning, while Figure 15, Table 13, and Figure 16 show how the effects vary across individual OOD datasets.

Robustness of Different OOD Detection Methods

Figures 17 and 19 compare detector-level AUROC changes under instance and class unlearning. Table 14 reports the average robustness of each detector, and Figure 18 aggregates these results by output-, activation-, and representation-based detector families.

Factors Driving Post-Unlearning OOD Fragility

The figures in this subsection relate post-unlearning OOD degradation to changes measured on retained data. Figure 20 examines feature and logit alignment, Figure 21 examines normalized score-space displacement, and Figure 22 summarizes alignment, drift, confidence, and local-neighborhood preservation across the sampled checkpoints.

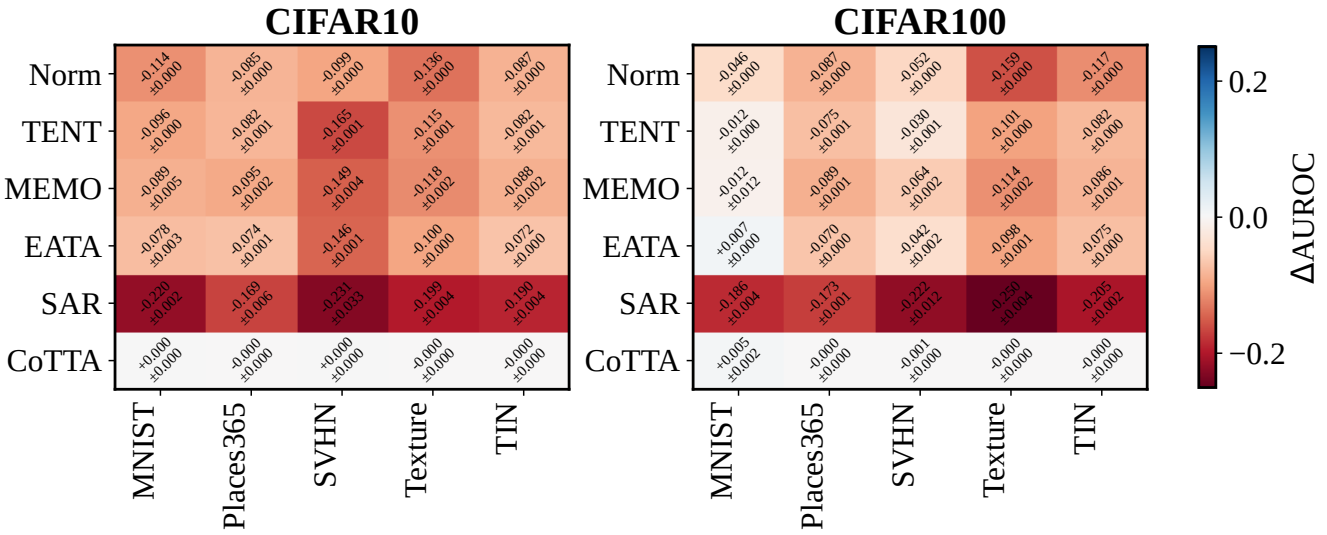

Figure 7: Dataset-level source-relative AUROC change under TTA for CIFAR-10 and CIFAR-100. Results are grouped by adaptation method and OOD evaluation dataset; more negative values indicate greater degradation.

Table 6: Summary of the TTA baselines.

Method	Brief description
Source	The pretrained model is evaluated without any test-time parameter or normalization-statistic update.
Norm	Re-estimates batch-normalization statistics from the unlabeled test stream while leaving learned weights fixed (**?**).
TENT	Minimizes the entropy of test predictions and updates only the trainable affine parameters in normalization layers (Wang et al. 2021).
MEMO	Augments each test input and minimizes the entropy of the marginal prediction averaged across its augmented views (Zhang, Levine, and Finn 2022).
EATA	Applies entropy minimization only to reliable, non-redundant test samples and regularizes important parameters to reduce forgetting (Niu et al. 2022).
SAR	Selects low-entropy samples and performs a sharpness-aware entropy update, with recovery mechanisms intended to prevent model collapse (Niu et al. 2023).
CoTTA	Uses an exponential-moving-average teacher, augmentation-averaged predictions, and stochastic restoration of source weights for continual adaptation (Wang et al. 2022b).

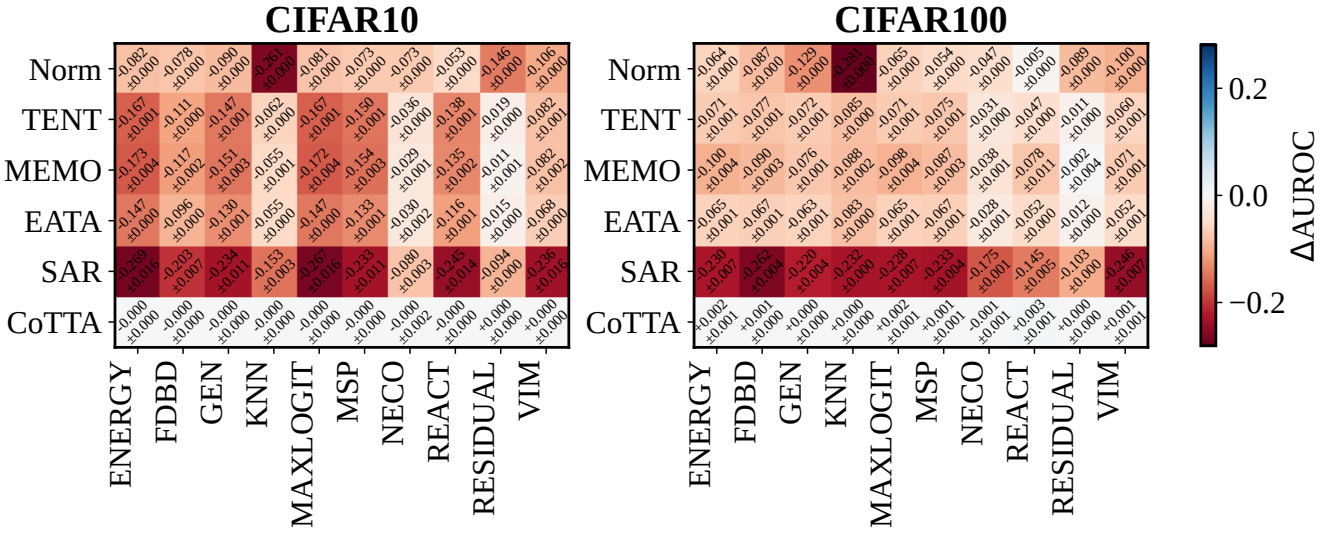

Figure 8: Detector-level source-relative AUROC change under TTA for CIFAR-10 and CIFAR-100. The figure shows how detector robustness depends on the adaptation mechanism rather than only on detector family.

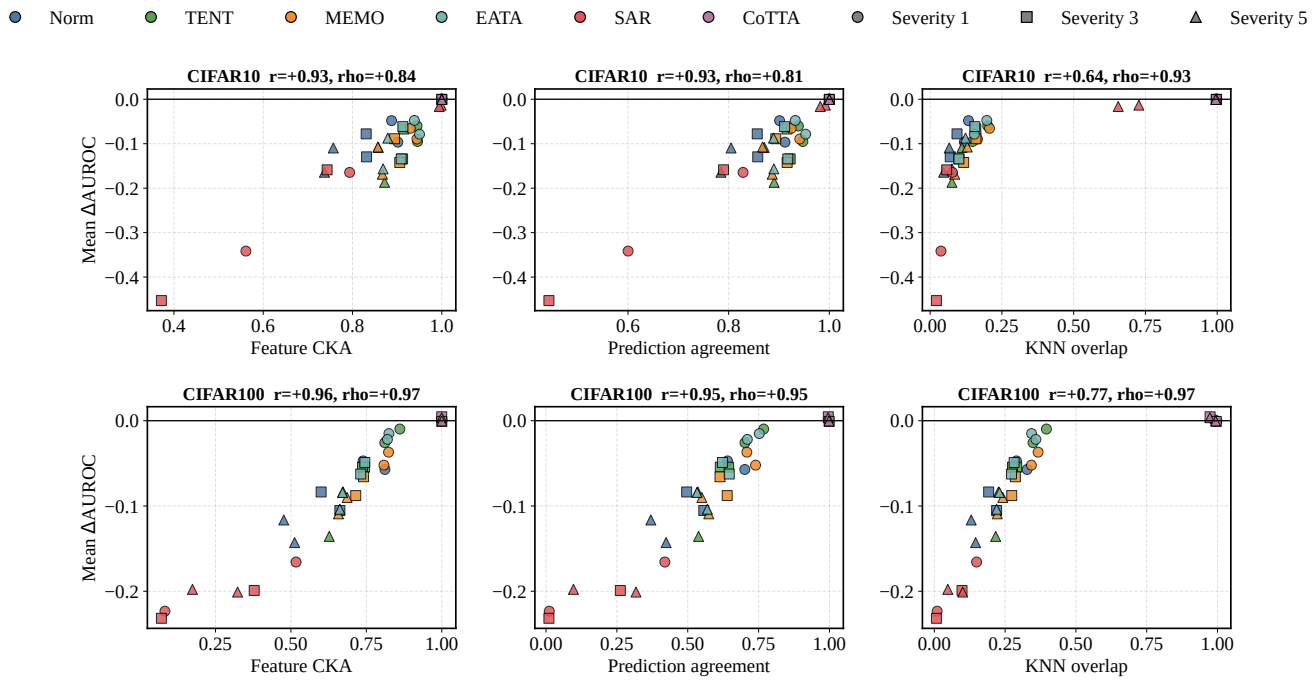

Figure 9: Association between source-model representation alignment and OOD degradation after test-time adaptation. Greater preservation of the source feature and logit geometry generally accompanies smaller source-relative degradation.

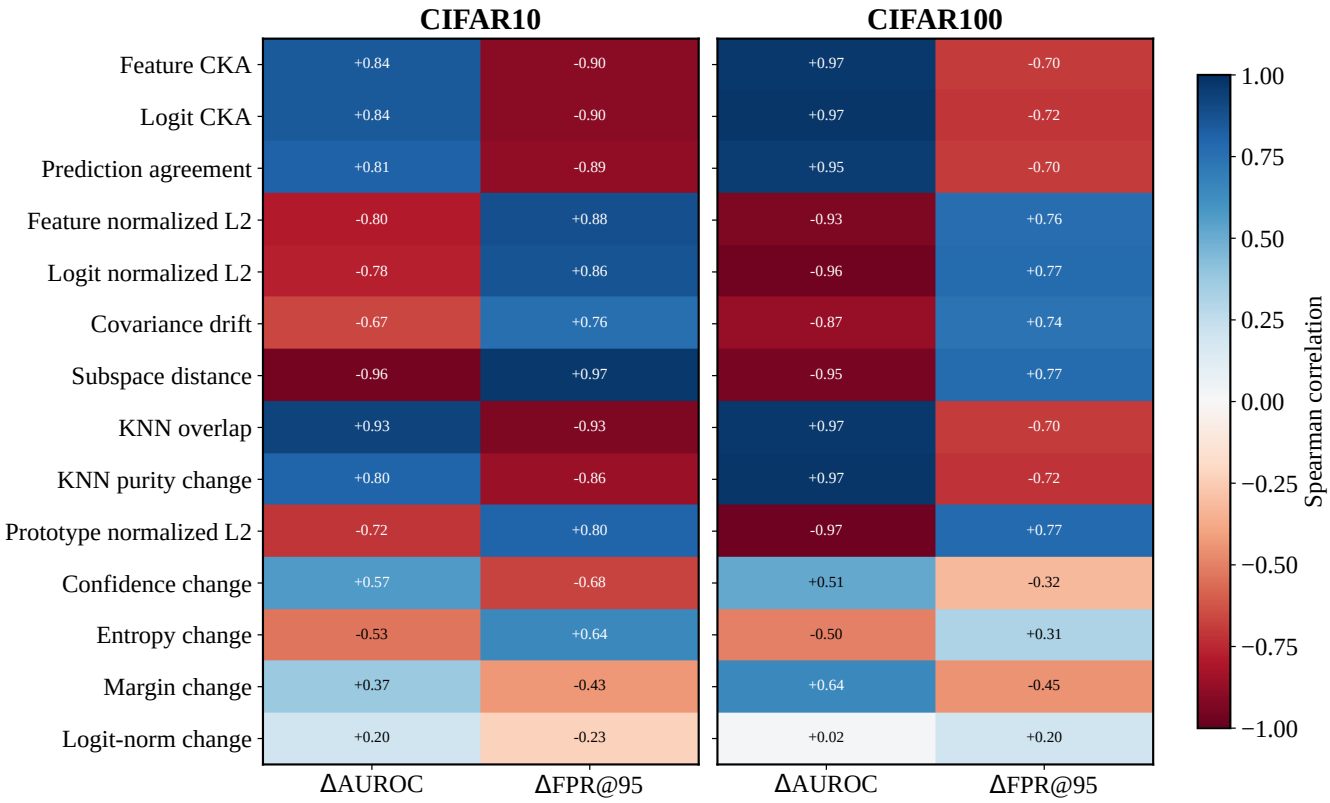

Figure 10: Correlations between adaptation-induced internal changes and source-relative OOD performance. The heatmap summarizes which representation, prediction, and neighborhood quantities most closely track AUROC and FPR@95 changes.

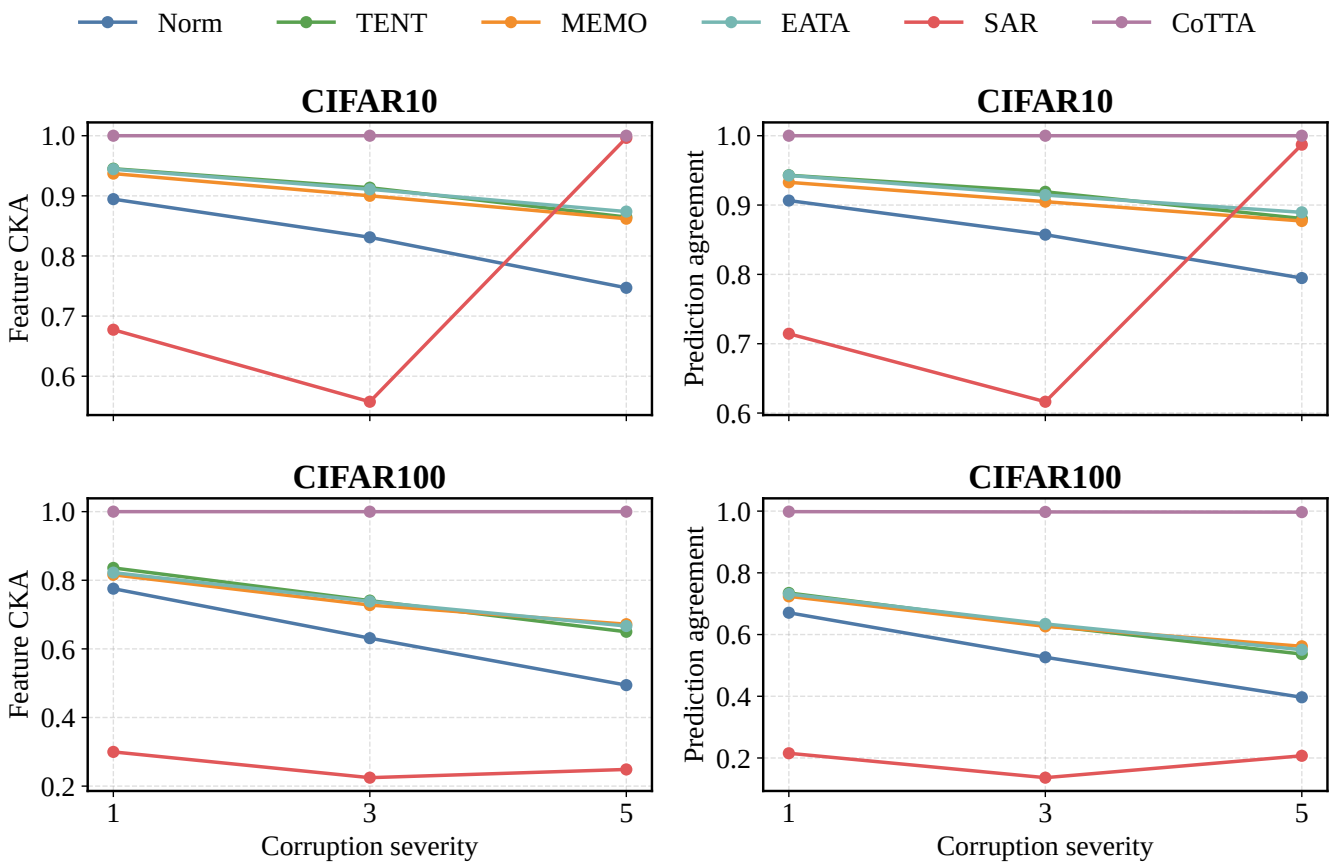

Figure 11: Variation of the measured representation changes with corruption severity under TTA. The figure complements the severity-dependent OOD degradation reported in the main text.

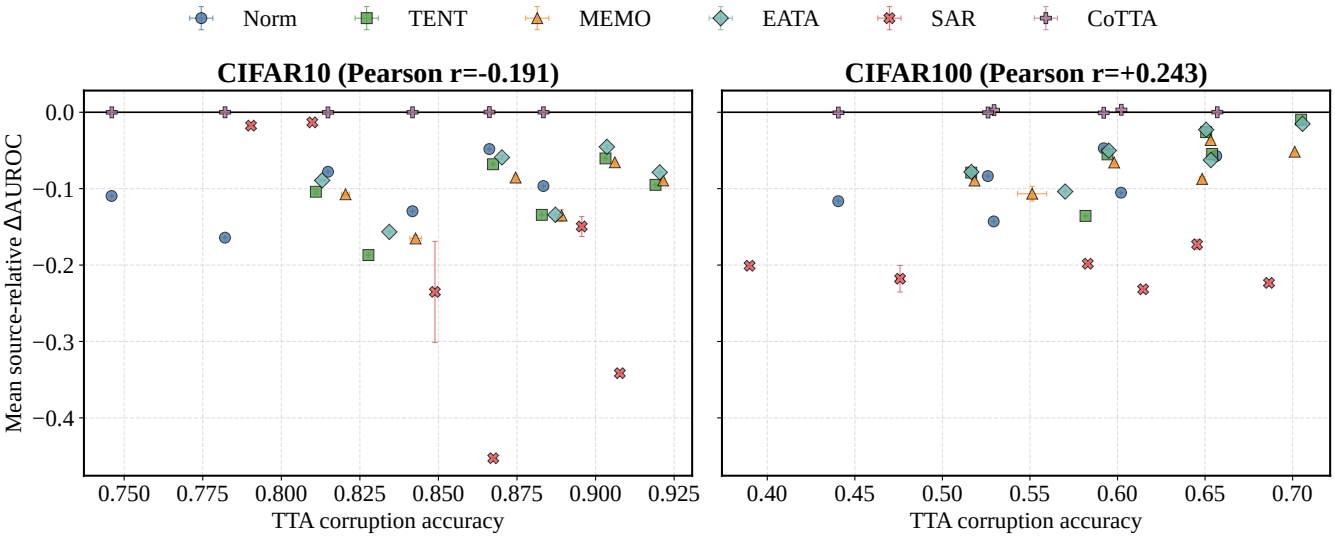

Figure 12: Relationship between corruption-set accuracy and source-relative OOD change across TTA configurations. Similar adaptation accuracy can coincide with substantially different levels of OOD preservation.

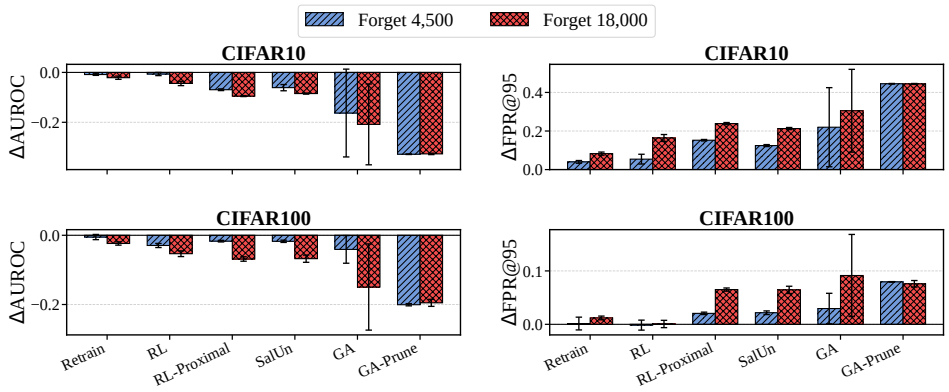

Figure 13: Overall OOD degradation under machine instance unlearning. Each bar reports the change relative to the corresponding source model, averaged over five OOD datasets and 11 detectors. Error bars denote standard deviation across available seeds. Negative ΔAUROC and positive ΔFPR indicate degradation.

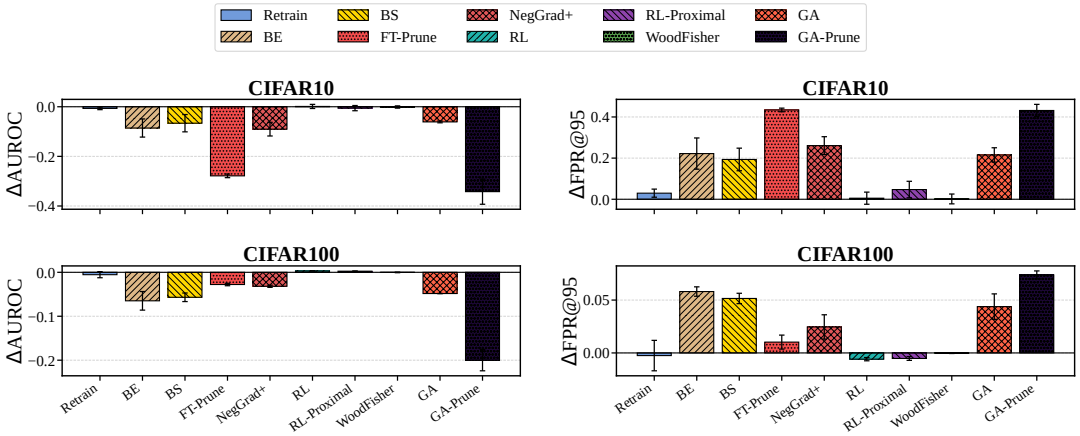

Figure 14: Overall OOD degradation under machine class unlearning. Results are averaged over five OOD datasets and 11 detectors.

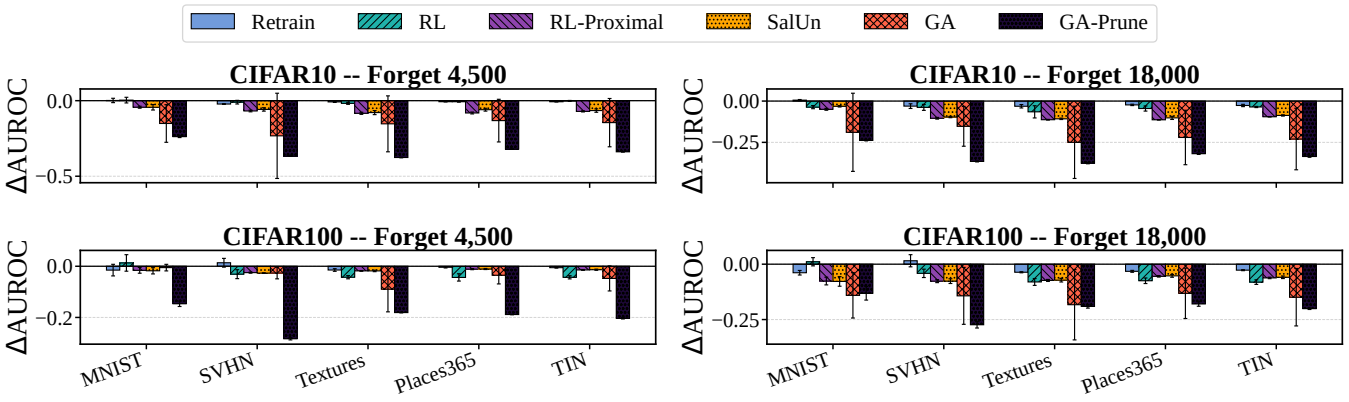

Figure 15: Dataset-centric AUROC degradation under machine instance unlearning. Each bar reports the change relative to the source model, averaged over 11 OOD detectors. Error bars denote standard deviation across available seeds. More negative values indicate greater degradation.

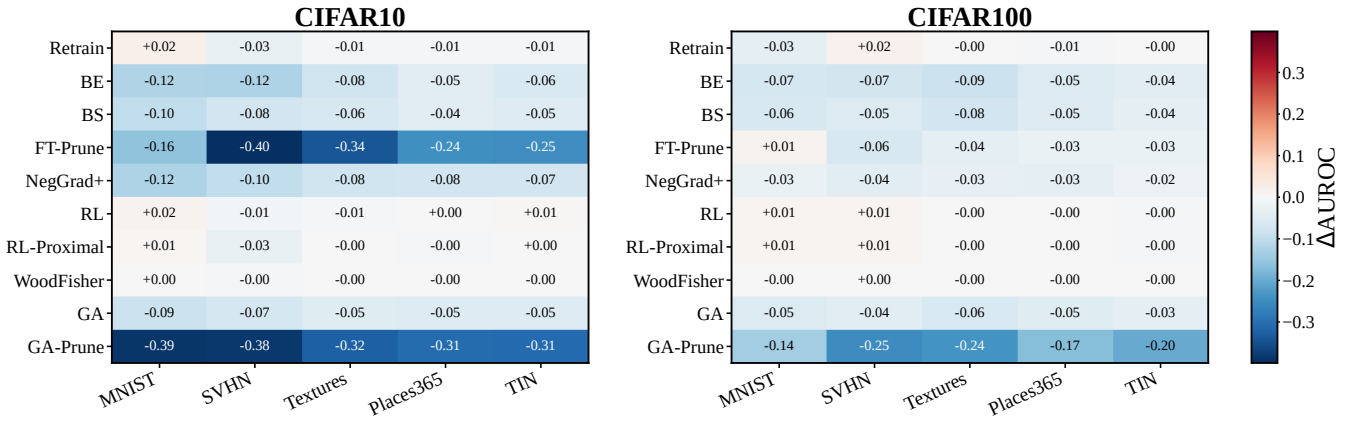

Figure 16: Dataset-centric AUROC degradation under machine class unlearning. Each cell reports the mean change relative to the source model, averaged over available class-and-seed runs and 11 detectors. CIFAR-10 uses forgotten classes 0–2, while CIFAR-100 is restricted to the common evaluated classes 0 and 15.

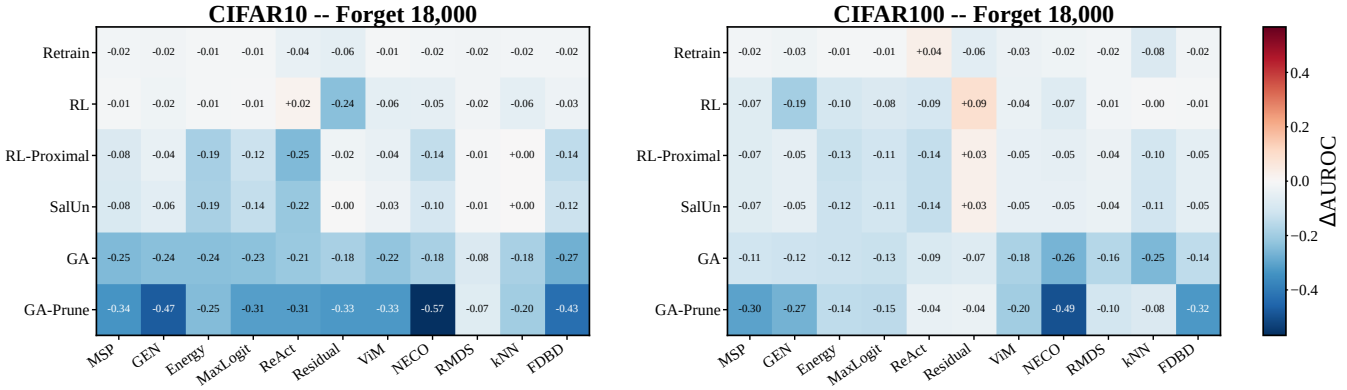

Figure 17: Detector-level AUROC change under machine instance unlearning with 18,000 removed samples. Each cell reports the mean change relative to that detector's source-model performance, averaged over five OOD datasets and available seeds. Negative values indicate degradation.

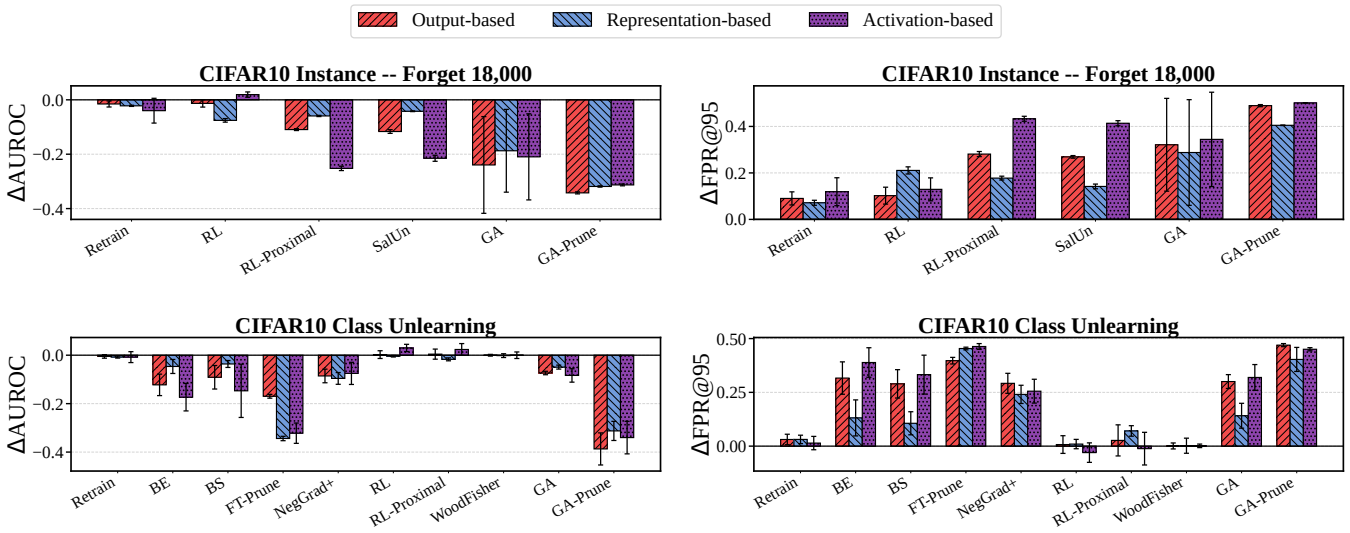

Figure 18: Average degradation of different detector families under CIFAR-10 instance and class unlearning. Output-based detectors include MSP, GEN, Energy, and MaxLogit; representation-based detectors include Residual, ViM, NECO, RMDS, kNN, and FDBD; ReAct forms the activation-based category. Error bars denote standard deviation across available checkpoint-level runs.

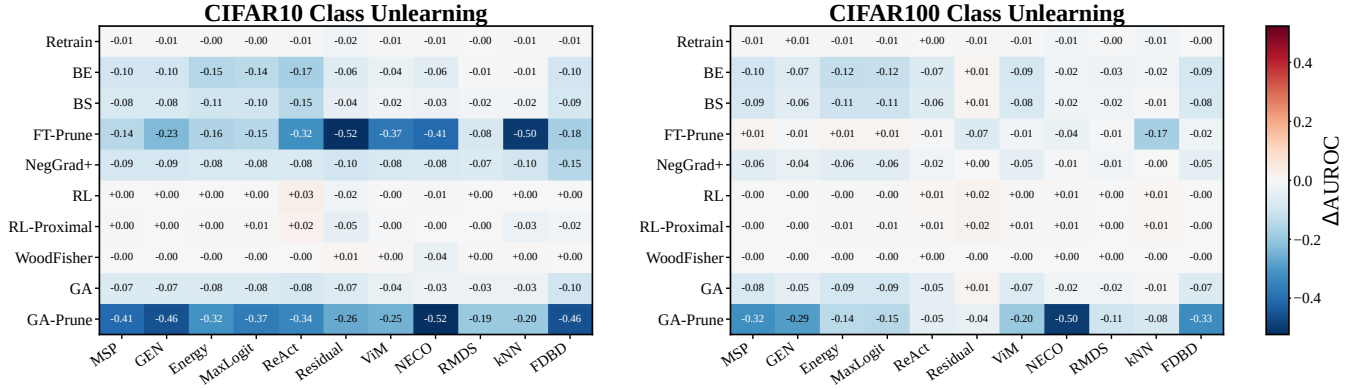

Figure 19: Detector-level AUROC change under machine class unlearning. Results are averaged over five OOD datasets and available class-and-seed runs. CIFAR-10 uses forgotten classes 0–2, while CIFAR-100 is restricted to the common evaluated classes 0 and 15.

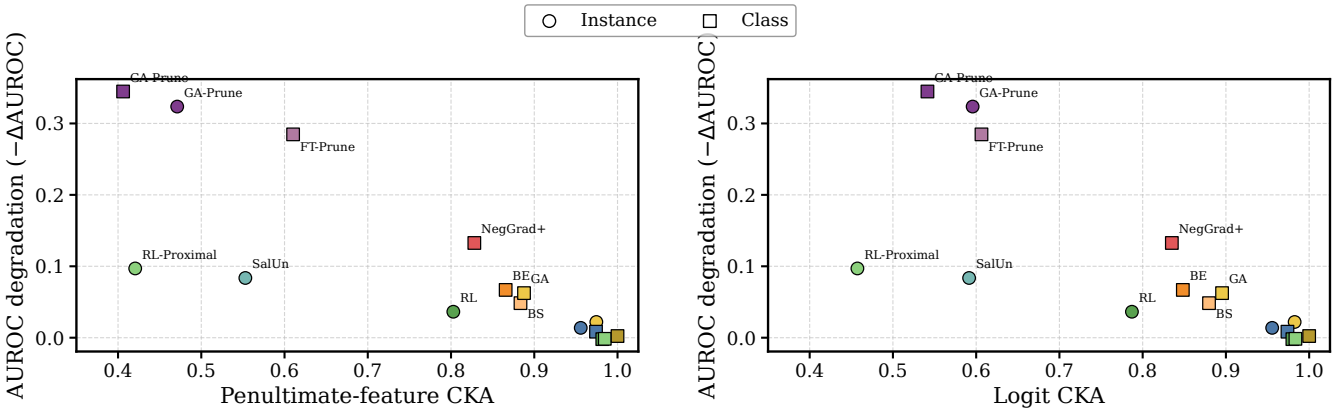

Figure 20: Association between retained-data alignment and OOD degradation in the sampled mechanistic panel. Circles denote CIFAR-10 instance unlearning with 18,000 removed examples; squares denote CIFAR-10 class unlearning for forgotten class 0. Higher $-\Delta$AUROC indicates greater degradation. CKA is computed on paired retained examples.

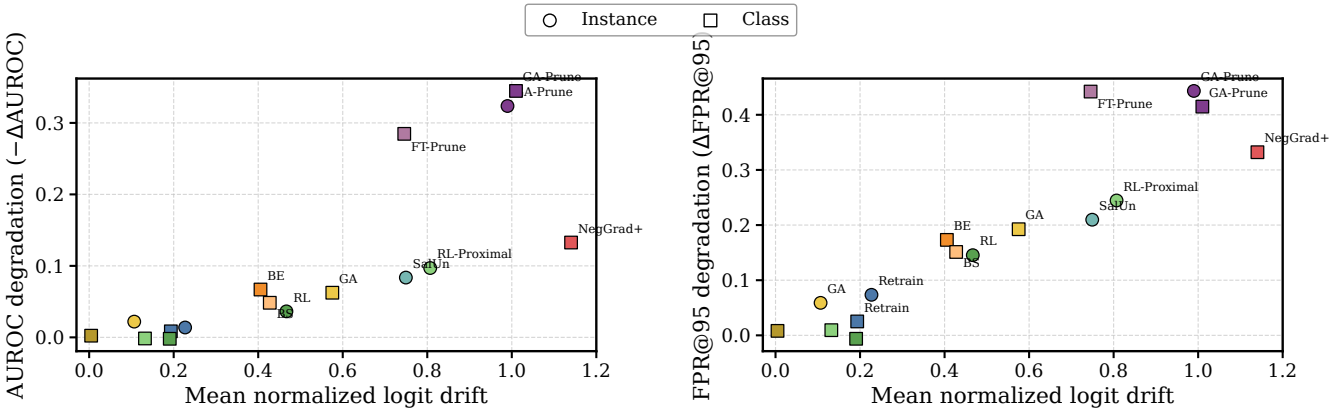

Figure 21: Association between paired retained-example logit drift and OOD degradation. Logit drift is the ℓ_2 distance between source and updated logits normalized by the source logit norm, then averaged over retained examples. Larger values indicate greater score-space distortion.

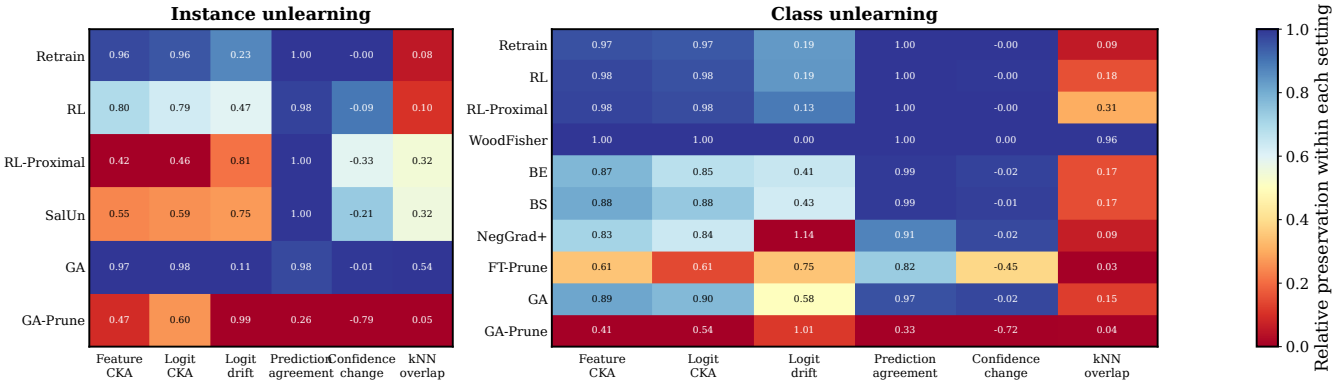

Figure 22: Summary of retained-data changes in the sampled mechanistic panel. Cell annotations report raw metric values. Colors are normalized separately within each column and unlearning setting so that blue denotes greater preservation and red denotes greater distortion. For logit drift and confidence change, smaller displacement magnitude denotes greater preservation. Colors should therefore be compared within, not across, columns.

Table 7: Summary of the machine-unlearning baselines.

Method	Brief description
Source	The original model trained on the complete training set; no forgetting update is applied.
Retrain	Exact-unlearning reference trained from scratch after removing the forget set.
RL	Randomly relabels forget-set examples and fine-tunes the model on the resulting incorrect labels, weakening the original input–label mapping (Fan et al. 2024).
RL-Proximal	Our RL variant with a proximal penalty that limits parameter movement away from the source model and helps preserve retained knowledge (Fan et al. 2023).
SalUn	Computes a gradient-based weight-saliency mask and applies the random-label unlearning update only to the selected parameters (Fan et al. 2024).
GA	Performs gradient ascent on the forget-set classification loss so that the model becomes worse at predicting the forgotten examples (Fan et al. 2024).
GA-Prune	Combines forget-set gradient ascent with parameter pruning, restricting or removing weights identified by the pruning mask (Fan et al. 2024).
BE	Boundary Expanding introduces a shadow class for forgotten examples and then removes it, shifting the decision region away from the forgotten class (Chen et al. 2023b).
BS	Boundary Shrinking relabels forgotten examples toward nearby non-forgotten classes and fine-tunes the classifier to contract the forgotten decision region (Chen et al. 2023b).
FT-Prune	Fine-tunes on retained data while pruning selected parameters, aiming to preserve utility while removing parameters associated with the forgotten data (Jia et al. 2023).
NegGrad+	Jointly minimizes retain-set loss and maximizes forget-set loss; our configuration assigns equal weight to the retain and forget objectives (Kurmanji et al. 2023).
WoodFisher	Uses an inverse-Fisher approximation to estimate a second-order parameter update that removes the contribution of forgotten data (Jia et al. 2023).

Table 8: TTA corruption accuracy grouped by random seed. Each entry averages two architectures and severities 1, 3, and 5 (six configurations).

Dataset	Method	Seed 1	Seed 2	Seed 3
CIFAR10	NORM	0.8224	0.8224	0.8224
CIFAR10	TENT	0.8686	0.8685	0.8685
CIFAR10	MEMO	0.8763	0.8757	0.8752
CIFAR10	EATA	0.8717	0.8714	0.8714
CIFAR10	SAR	0.8534	0.8533	0.8533
CIFAR10	COTTA	0.8224	0.8224	0.8224
CIFAR100	NORM	0.5577	0.5577	0.5577
CIFAR100	TENT	0.6171	0.6170	0.6170
CIFAR100	MEMO	0.6109	0.6112	0.6131
CIFAR100	EATA	0.6151	0.6153	0.6153
CIFAR100	SAR	0.5655	0.5661	0.5661
CIFAR100	COTTA	0.5580	0.5578	0.5579

Table 9: Instance-unlearning task performance grouped by random seed. Accuracy values are percentages; MIA is attack correctness.

Dataset	Forget Size	Method	Seed 1				Seed 2				Seed 3			
			Retain	Forget	Test	MIA	Retain	Forget	Test	MIA	Retain	Forget	Test	MIA
CIFAR10	4,500	Retrain	100.00	94.47	94.14	0.055	100.00	94.69	94.43	0.053	100.00	95.56	94.33	0.044
CIFAR10	4,500	RL	99.69	97.18	94.08	0.028	99.68	97.40	93.97	0.026	99.60	97.33	94.19	0.027
CIFAR10	4,500	RL-Proximal	99.97	99.93	94.15	0.001	99.45	99.38	94.20	0.006	99.49	99.47	94.38	0.005
CIFAR10	4,500	SalUn	99.97	99.93	94.31	0.001	99.45	99.38	94.37	0.006	99.48	99.47	94.41	0.005
CIFAR10	4,500	GA	100.00	99.98	94.65	0.000	27.70	28.24	27.11	0.718	88.70	88.27	83.45	0.117
CIFAR10	4,500	GA-Prune	22.88	23.40	23.17	0.234	22.00	22.38	22.76	0.224	19.91	20.02	20.42	0.200
CIFAR10	18,000	Retrain	100.00	93.13	92.93	0.069	100.00	93.21	92.89	0.068	100.00	93.32	92.71	0.067
CIFAR10	18,000	RL	98.41	96.95	92.63	0.030	98.23	96.86	92.60	0.031	98.40	96.69	92.51	0.033
CIFAR10	18,000	RL-Proximal	99.80	99.69	92.75	0.003	99.19	99.23	93.10	0.008	99.29	99.19	93.25	0.008
CIFAR10	18,000	SalUn	99.84	99.76	92.87	0.002	99.16	99.13	92.99	0.009	99.28	99.19	92.93	0.008
CIFAR10	18,000	GA	97.73	97.69	92.12	0.023	3.09	2.80	2.99	0.972	2.05	1.54	2.79	0.015
CIFAR10	18,000	GA-Prune	25.76	25.44	25.88	0.254	19.76	20.39	20.45	0.204	19.70	19.92	20.18	0.199
CIFAR100	4,500	Retrain	99.98	74.60	74.71	0.254	99.98	76.36	74.39	0.236	99.97	74.40	74.28	0.256
CIFAR100	4,500	RL	99.89	72.33	68.38	0.277	99.18	71.60	69.63	0.284	99.19	73.91	69.83	0.261
CIFAR100	4,500	RL-Proximal	99.97	99.98	74.62	0.000	97.51	97.40	74.95	0.026	97.44	97.84	74.89	0.022
CIFAR100	4,500	SalUn	99.97	99.98	74.59	0.000	97.49	97.36	75.02	0.026	97.43	97.82	75.03	0.022
CIFAR100	4,500	GA	99.98	99.96	75.72	0.000	83.68	80.98	62.13	0.190	89.56	88.42	67.68	0.116
CIFAR100	4,500	GA-Prune	1.82	1.67	1.75	0.017	1.80	1.73	1.75	0.983	1.81	2.22	1.78	0.022
CIFAR100	18,000	Retrain	99.99	69.67	69.71	0.303	99.97	69.06	69.18	0.309	99.99	69.63	69.27	0.304
CIFAR100	18,000	RL	99.10	66.34	54.97	0.337	98.11	66.84	57.11	0.332	97.59	64.03	55.71	0.360
CIFAR100	18,000	RL-Proximal	99.29	99.25	69.60	0.007	96.53	96.29	70.36	0.037	96.45	96.34	70.17	0.037
CIFAR100	18,000	SalUn	99.31	99.27	69.94	0.007	96.79	96.53	71.16	0.035	96.61	96.57	70.96	0.034
CIFAR100	18,000	GA	98.20	98.17	74.26	0.018	1.81	1.78	1.75	0.018	1.08	1.15	1.13	0.011
CIFAR100	18,000	GA-Prune	2.23	2.13	2.11	0.979	1.95	1.90	1.86	0.981	2.29	2.18	2.09	0.978

Table 10: Class-unlearning task performance grouped by random seed, averaged over the common forgotten-class panel within each seed. Accuracy values are percentages; MIA is attack correctness.

Dataset	Method	Seed 1				Seed 2				Seed 3			
		Retain	Forget	Test	MIA	Retain	Forget	Test	MIA	Retain	Forget	Test	MIA
CIFAR10	Source	–	–	94.84	–	–	–	94.84	–	–	–	94.84	–
CIFAR10	Retrain	100.00 ± 0.00	0.00 ± 0.00	94.94 ± 0.10	1.000 ± 0.000	100.00 ± 0.00	0.00 ± 0.00	94.56 ± 0.31	1.000 ± 0.000	100.00 ± 0.00	0.00 ± 0.00	94.83 ± 0.25	1.000 ± 0.000
CIFAR10	BE	97.74 ± 2.19	10.16 ± 8.61	91.26 ± 2.85	0.898 ± 0.086	96.91 ± 2.21	10.93 ± 8.71	91.05 ± 2.75	0.891 ± 0.087	96.93 ± 2.27	10.57 ± 8.24	91.16 ± 2.94	0.894 ± 0.082
CIFAR10	BS	98.48 ± 1.21	16.67 ± 9.26	91.71 ± 1.98	0.833 ± 0.093	97.67 ± 1.42	17.07 ± 9.11	91.55 ± 2.12	0.829 ± 0.091	97.58 ± 1.61	17.29 ± 9.15	91.47 ± 2.36	0.827 ± 0.092
CIFAR10	FT-Prune	80.78 ± 1.03	1.58 ± 1.75	77.70 ± 0.94	0.984 ± 0.017	81.54 ± 1.30	2.11 ± 2.05	78.64 ± 1.66	0.979 ± 0.021	81.00 ± 1.48	0.68 ± 0.50	78.19 ± 1.47	0.993 ± 0.005
CIFAR10	NegGrad+	94.19 ± 2.52	2.43 ± 1.98	87.94 ± 2.29	0.976 ± 0.020	94.02 ± 2.06	2.24 ± 2.08	88.33 ± 1.80	0.978 ± 0.021	94.21 ± 1.88	2.24 ± 1.87	88.58 ± 1.75	0.978 ± 0.019
CIFAR10	RL	99.96 ± 0.04	0.08 ± 0.06	94.89 ± 0.32	0.999 ± 0.001	99.77 ± 0.15	0.17 ± 0.30	94.58 ± 0.43	0.998 ± 0.003	99.81 ± 0.13	0.09 ± 0.08	94.67 ± 0.35	0.999 ± 0.001
CIFAR10	RL-Proximal	99.96 ± 0.03	1.26 ± 1.92	94.61 ± 0.49	0.987 ± 0.019	99.73 ± 0.12	12.44 ± 13.89	94.15 ± 0.53	0.876 ± 0.139	99.73 ± 0.17	18.05 ± 30.20	94.24 ± 0.54	0.819 ± 0.302
CIFAR10	WoodFisher	100.00 ± 0.00	99.99 ± 0.01	94.62 ± 0.25	0.000 ± 0.000	–	–	–	–	–	–	–	–
CIFAR10	GA	96.79 ± 0.53	10.51 ± 6.66	90.01 ± 0.27	0.895 ± 0.067	96.22 ± 0.52	10.91 ± 7.04	90.01 ± 0.29	0.891 ± 0.070	96.27 ± 0.48	11.02 ± 7.21	90.11 ± 0.33	0.890 ± 0.072
CIFAR10	GA-Prune	34.16 ± 12.82	0.00 ± 0.00	33.90 ± 12.51	0.667 ± 0.577	32.93 ± 14.09	0.00 ± 0.00	32.73 ± 13.77	0.333 ± 0.577	34.47 ± 12.10	0.00 ± 0.00	34.17 ± 12.20	0.333 ± 0.577
CIFAR100	Source	–	–	75.54	–	–	–	75.54	–	–	–	75.54	–
CIFAR100	Retrain	99.97 ± 0.00	0.00 ± 0.00	74.94 ± 0.01	1.000 ± 0.000	99.97 ± 0.00	0.00 ± 0.00	74.37 ± 0.36	1.000 ± 0.000	99.97 ± 0.01	0.00 ± 0.00	75.00 ± 0.40	1.000 ± 0.000
CIFAR100	BE	93.90 ± 6.96	4.22 ± 0.94	66.47 ± 5.94	0.958 ± 0.009	91.19 ± 6.75	3.33 ± 1.57	66.43 ± 5.89	0.967 ± 0.016	90.84 ± 7.35	4.00 ± 0.63	66.26 ± 6.16	0.960 ± 0.006
CIFAR100	BS	94.95 ± 5.49	12.00 ± 10.06	67.40 ± 4.67	0.880 ± 0.101	92.15 ± 5.40	11.44 ± 9.90	67.41 ± 4.57	0.886 ± 0.099	92.14 ± 5.53	12.00 ± 10.69	67.42 ± 4.52	0.880 ± 0.107
CIFAR100	FT-Prune	99.71 ± 0.02	100.00 ± 0.00	74.31 ± 0.06	0.000 ± 0.000	97.13 ± 0.01	98.67 ± 0.94	74.28 ± 0.09	0.013 ± 0.009	97.17 ± 0.03	98.78 ± 1.10	74.25 ± 0.16	0.012 ± 0.011
CIFAR100	NegGrad+	98.13 ± 2.24	1.78 ± 0.63	70.97 ± 2.99	0.982 ± 0.006	95.03 ± 2.80	1.56 ± 0.31	70.69 ± 3.63	0.984 ± 0.003	95.26 ± 2.62	1.67 ± 0.79	70.75 ± 3.35	0.983 ± 0.008
CIFAR100	RL	99.97 ± 0.00	0.44 ± 0.31	75.54 ± 0.17	0.996 ± 0.003	99.71 ± 0.00	9.56 ± 13.51	75.44 ± 0.03	0.904 ± 0.135	99.70 ± 0.03	2.11 ± 2.99	75.62 ± 0.15	0.979 ± 0.030
CIFAR100	RL-Proximal	99.97 ± 0.00	0.44 ± 0.31	75.47 ± 0.06	0.996 ± 0.003	99.58 ± 0.00	9.00 ± 12.41	75.43 ± 0.09	0.910 ± 0.124	99.57 ± 0.03	2.22 ± 3.14	75.71 ± 0.13	0.978 ± 0.031
CIFAR100	WoodFisher	99.98 ± 0.00	99.89 ± 0.16	75.75 ± 0.05	0.001 ± 0.002	–	–	–	–	–	–	–	–
CIFAR100	GA	95.97 ± 3.81	5.44 ± 6.76	68.14 ± 3.54	0.946 ± 0.068	93.07 ± 3.94	6.00 ± 7.23	68.09 ± 3.67	0.940 ± 0.072	93.23 ± 4.02	5.33 ± 6.60	68.28 ± 3.77	0.947 ± 0.066
CIFAR100	GA-Prune	2.40 ± 0.27	0.00 ± 0.00	2.31 ± 0.41	1.000 ± 0.000	2.34 ± 0.34	0.00 ± 0.00	2.28 ± 0.44	1.000 ± 0.000	2.35 ± 0.29	0.00 ± 0.00	2.31 ± 0.42	0.500 ± 0.707

Table 11: Overall OOD degradation under CIFAR-10 machine instance unlearning. Results are mean ± standard deviation across three seeds, averaged over all OOD datasets and detectors.

	Forget 4,500		Forget 18,000	
Method	**ΔAUROC**	**ΔFPR@95**	**ΔAUROC**	**ΔFPR@95**
Retrain	-0.0086 ± 0.0032	$+0.0400 \pm 0.0070$	-0.0213 ± 0.0065	$+0.0827 \pm 0.0081$
RL	-0.0065 ± 0.0064	$+0.0542 \pm 0.0253$	-0.0442 ± 0.0087	$+0.1639 \pm 0.0179$
RL-Proximal	-0.0695 ± 0.0034	$+0.1523 \pm 0.0038$	-0.0951 ± 0.0017	$+0.2385 \pm 0.0055$
SalUn	-0.0611 ± 0.0125	$+0.1245 \pm 0.0051$	-0.0849 ± 0.0022	$+0.2128 \pm 0.0054$
GA	-0.1629 ± 0.1761	$+0.2198 \pm 0.2051$	-0.2084 ± 0.1621	$+0.3050 \pm 0.2146$
GA-Prune	-0.3280 ± 0.0011	$+0.4449 \pm 0.0016$	-0.3268 ± 0.0028	$+0.4444 \pm 0.0021$

Table 12: Overall OOD degradation under CIFAR-100 machine instance unlearning. Results are mean ± standard deviation across three seeds.

	Forget 4,500		Forget 18,000	
Method	**ΔAUROC**	**ΔFPR@95**	**ΔAUROC**	**ΔFPR@95**
Retrain	-0.0051 ± 0.0075	$+0.0015 \pm 0.0120$	-0.0237 ± 0.0048	$+0.0119 \pm 0.0035$
RL	-0.0295 ± 0.0059	-0.0015 ± 0.0094	-0.0533 ± 0.0082	$+0.0006 \pm 0.0068$
RL-Proximal	-0.0169 ± 0.0023	$+0.0205 \pm 0.0023$	-0.0686 ± 0.0064	$+0.0650 \pm 0.0030$
SalUn	-0.0174 ± 0.0033	$+0.0217 \pm 0.0034$	-0.0677 ± 0.0104	$+0.0645 \pm 0.0068$
GA	-0.0408 ± 0.0400	$+0.0296 \pm 0.0285$	-0.1495 ± 0.1245	$+0.0914 \pm 0.0768$
GA-Prune	-0.2007 ± 0.0032	$+0.0795 \pm 0.0005$	-0.1952 ± 0.0102	$+0.0759 \pm 0.0061$

Table 13: Dataset-centric OOD performance under CIFAR-10 machine instance unlearning with 18,000 removed samples. Each cell reports AUROC mean ± standard deviation across three seeds, averaged over all 11 OOD detectors.

Method	**MNIST ↑**	**SVHN ↑**	**Textures ↑**	**Places365 ↑**	**TIN ↑**
Source	0.9055 ± 0.0000	0.9287 ± 0.0000	0.9145 ± 0.0000	0.8834 ± 0.0000	0.8702 ± 0.0000
Retrain	0.9111 ± 0.0037	0.8980 ± 0.0142	0.8830 ± 0.0102	0.8602 ± 0.0041	0.8435 ± 0.0069
RL	0.8687 ± 0.0095	0.8905 ± 0.0180	0.8493 ± 0.0370	0.8383 ± 0.0155	0.8343 ± 0.0036
RL-Proximal	0.8556 ± 0.0056	0.8246 ± 0.0051	0.8015 ± 0.0021	0.7699 ± 0.0015	0.7750 ± 0.0006
SalUn	0.8738 ± 0.0042	0.8321 ± 0.0040	0.8069 ± 0.0042	0.7816 ± 0.0095	0.7832 ± 0.0038
GA	0.7164 ± 0.2366	0.7758 ± 0.1206	0.6646 ± 0.2187	0.6629 ± 0.1649	0.6403 ± 0.1861
GA-Prune	0.6673 ± 0.0029	0.5631 ± 0.0029	0.5388 ± 0.0040	0.5647 ± 0.0044	0.5344 ± 0.0046

Table 14: Average detector robustness under CIFAR-10 instance unlearning with 18,000 removed samples. Deltas are averaged over six unlearning methods, three seeds, and five OOD datasets.

Detector	**Family**	**Source AUROC**	**Avg. ΔAUROC**	**Avg. ΔFPR@95**
MSP	Output	0.8929	−0.1303	+0.2036
GEN	Output	0.8953	−0.1429	+0.1820
Energy	Output	0.8959	−0.1465	+0.3407
MaxLogit	Output	0.8950	−0.1379	+0.3092
ReAct	Activation	0.8448	−0.1685	+0.3234
Residual	Representation	0.8896	−0.1381	+0.1389
ViM	Representation	0.9244	−0.1146	+0.2163
NECO	Representation	0.9300	−0.1760	+0.3592
RMDS	Representation	0.9000	−0.0338	+0.1103
kNN	Representation	0.9238	−0.0745	+0.1427
FDBD	Representation	0.9131	−0.1682	+0.3272