UNDERSTANDING AND MITIGATING LATE-STAGE OOD DETECTION DEGRADATION IN EMA MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Exponential moving average (EMA) is widely used to improve model generalization, but its behavior in out-of-distribution (OOD) detection is not uniformly beneficial throughout training. We observe a consistent late-stage degradation phenomenon: EMA improves substantially during early and middle training, but its OOD performance later declines, while the plain model remains competitive and gradually closes the gap. Our analysis suggests that this behavior is associated with increasing redundancy in the late-stage plain-model trajectory, which reduces the complementary information available to EMA. We propose **ReDiv**, a simple **Re**wind-and-**Div**ersify framework for EMA-based OOD detection. ReDiv first detects the onset of late-stage degradation using an ID-only logit-margin signal and rewinds the complete training state to the corresponding checkpoint. It then diversifies the post-rewind trajectory by fixing the learning rate to avoid further scheduler-induced shrinkage and applying sharpness-aware minimization (SAM) to modify the effective update. Across multiple datasets and OOD benchmarks, ReDiv consistently improves over both plain training and standard EMA, requires no OOD validation data or architectural modification, and significantly complements existing training-time and test-time OOD methods. Through careful investigation, we show that the effectiveness of EMA depends not only on checkpoint quality, but also on the diversity of the model trajectory, which verifies the late-stage trajectory control in ReDiv as a powerful way to improve performance.

1 INTRODUCTION

Out-of-distribution (OOD) detection aims to identify inputs that deviate from the training distribution and is critical for the reliable deployment of modern neural networks (Nair et al., 2020; Corsini & Yang, 2023; Lu et al., 2025). Among the many techniques used to improve model generalization, exponential moving average (EMA) has become a widely adopted training practice (Tarvainen & Valpola, 2017; Shu et al., 2023), which maintains a running average of model parameters along the optimization trajectory. By smoothing stochastic parameter updates, EMA often yields a model that generalizes better than the final plain checkpoint (Morales-Brotons et al., 2024).

Prior work has shown that EMA can influence OOD detection performance (Xiong et al., 2025). However, how the OOD behavior of the EMA model itself evolves relative to its underlying non-averaged model, *i.e.,* the plain model, throughout training remains largely unexplored. To investigate this question, we track the complete training trajectories of both the plain and EMA models. As shown in Figure 1, we uncover a consistent temporal reversal: EMA gains a clear OOD advantage during early and middle training, but this advantage gradually vanishes in the late stage, whereas the plain model remains competitive and eventually reaches comparable OOD performance.

The late-stage degradation of EMA is not accompanied by a decline in the plain model; instead, the plain model continues to improve overall. Meanwhile, the plain-model trajectory becomes increasingly redundant, with successive checkpoints becoming progressively converged. Given prior evidence that diversity plays an important role in model averaging (Wang & Ji, 2023), this growing redundancy provides a plausible explanation for the diminishing OOD advantage of EMA. This perspective suggests two requirements for improving the late-stage behavior of EMA. First, the training procedure should identify when continued averaging ceases to provide additional benefit without re-

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

lying on OOD validation data. Second, once this transition is detected, the subsequent optimization trajectory should be modified to preserve useful diversity for averaging.

Based on these insights, we propose **ReDiv**, a simple **Re**wind-and-**Div**ersify framework for EMA-based OOD detection. To identify when continued averaging ceases to be beneficial, ReDiv monitors an ID-only signal, the logit margin, which is denoted by the median of the top-1/top-2 logit gap on the ID validation set. We find that when this margin persistently fails to improve, EMA degradation occurs. Thus, ReDiv rewinds the complete training state to the onset of the transition to prevent further improvement loss. Consequently, ReDiv modifies the post-rewind optimization trajectory through Rewind and Diversity steps: a fixed learning rate prevents further contraction of parameter movement, while sharpness-aware minimization (SAM) diversifies the intrinsic update geometry in the most vulnerable directions with variable magnitudes. Our theoretical analysis further provides an intuitive understanding to characterize the complementary roles of the fixed learning rate and SAM in preserving useful post-rewind trajectory diversity. Additionally, extensive experimental results across multiple datasets, OOD regimes, and scoring functions demonstrate the effectiveness and generalizability of ReDiv.

Our main contributions are summarized as follows:

- We identify an overlooked yet vital late-stage degradation phenomenon of EMA in OOD detection: while the plain model continues to improve overall, the benefit of EMA gradually vanishes as the underlying trajectory becomes increasingly redundant.
- We propose **ReDiv**, an OOD-free rewind-and-diversify framework that detects a persistent loss of improvement from the ID-validation logit margin and reshapes the post-rewind trajectory using a fixed learning rate and SAM.
- We provide theoretical analysis to characterize the complementary roles of fixed learning rate and SAM in preserving useful post-rewind trajectory diversity.
- Extensive experiments across multiple datasets, OOD regimes, and scoring functions demonstrate the effectiveness and generalizability of ReDiv.

2 Preliminaries and Observation

2.1 Learning Setup

We consider supervised learning on an in-distribution (ID) training set $\mathcal{D}_{\text{tr}}^{\text{ID}} = \{(x_i, y_i)\}_{i=1}^{N}$, with a held-out ID validation set $\mathcal{D}_{\text{val}}^{\text{ID}}$. A model f_θ maps an input x to logits $\mathbf{z} = f_\theta(x) \in \mathbb{R}^K$ over K ID classes. At iteration t, we sample a mini-batch $\mathcal{B}_t \subset \mathcal{D}_{\text{tr}}^{\text{ID}}$ and define the training objective as

$$\mathcal{L}_{\mathcal{B}_t}(\theta) = \frac{1}{|\mathcal{B}_t|} \sum_{(x,y)\in\mathcal{B}_t} \ell(f_\theta(x), y)\,, \tag{1}$$

where $\ell(\cdot,\cdot)$ is the cross-entropy loss. The parameters are then updated as

$$\theta_{t+1} = \theta_t - \eta_t \mu_t, \quad \mu_t = \nabla_\theta \mathcal{L}_{\mathcal{B}_t}(\theta_t), \tag{2}$$

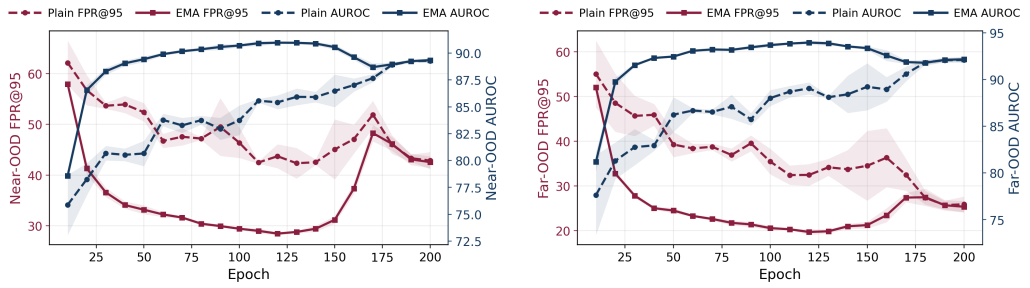
where θ_t and θ_{t+1} denote the model parameters before and after the t-th update, respectively, η_t is the learning rate, and μ_t is the stochastic gradient computed on $\mathcal{B}_t$. We refer to f_{θ_t} as the plain model, i.e., the online model whose parameters are updated directly by stochastic gradients without temporal averaging.

After each plain-model update, EMA model $f_{\bar{\theta}}$ updates its parameters as

$$\bar{\theta}_{t+1} = \alpha\bar{\theta}_t + (1-\alpha)\theta_{t+1}, \tag{3}$$

where $\alpha \in (0,1)$ is the EMA decay rate.

At test time, OOD detection is commonly performed using the EMA model $f_{\bar{\theta}}$. A post-hoc detector assigns an uncertainty score $S(x;\bar{\theta})$ to each test input x, where larger values indicate higher uncertainty and stronger evidence of being OOD.



(a) Near-OOD (b) Far-OOD

Figure 1: Checkpoint-wise OOD detection performance of plain and EMA models on CIFAR-10 under (a) near-OOD and (b) far-OOD scenarios.

2.2 Observation: EMA Exhibits Late-Stage OOD Degradation

We examine the temporal behavior of EMA for OOD detection by tracking the checkpoint-wise performance of plain and EMA models trained on CIFAR-10 with cross-entropy loss. Figure 1 reports MSP-based AUROC and FPR95 under the near-OOD and far-OOD settings.

As we can see, a consistent temporal reversal emerges in both near-OOD and far-OOD scenarios. EMA improves rapidly during early and middle training and reaches its best OOD performance around epochs 120–130. Beyond this point, however, its performance begins to decline despite continued training. In Figure 1 (a), EMA AUROC drops from approximately 91% to below 89%, while FPR95 increases from about 28–29$\%$ to nearly 48% around epoch 170. Figure 1 (b) shows the same trend: AUROC decreases from nearly 94% to about 92%, while FPR95 rises from around 20% to above 25%. Interestingly, the plain model does not exhibit the same late-stage decline. Its OOD performance continues to improve overall and progressively closes the gap to EMA, with the two models reaching nearly identical performance by epochs 180–200. This indicates that the late-stage failure of EMA is not caused by weaker plain checkpoints.

Theoretical advances in model averaging commonly decompose the ensemble error into individual error and diversity (Krogh & Vedelsby, 1994; Germain et al., 2015; Wood et al., 2023; Wortsman et al., 2022), therefore, we attribute the late-stage OOD degradation to a quality–diversity tradeoff along the plain-model trajectory. In the early stage, trajectory diversity is relatively high, while individual checkpoint quality remains limited. As training proceeds, checkpoint quality improves while sufficient diversity is retained, allowing EMA to benefit from averaging and reach its best performance. In the late stage, however, further gains in checkpoint quality are accompanied by a substantial reduction in trajectory diversity, leaving EMA to aggregate strong but increasingly redundant checkpoints.

These observations motivate two design principles for improving late-stage EMA. First, the training procedure should detect when the benefit of continued averaging begins to diminish, without relying on OOD validation data. Second, once this transition is identified, the subsequent optimization trajectory should maintain sufficiently strong checkpoints while avoiding excessive redundancy, so that EMA can continue to benefit from temporal averaging. Next, we further give a quality–diversity view under the EMA context.

2.3 A Quality–Diversity View of EMA

To explain why EMA may degrade even when later plain checkpoints remain strong, we analyze the parameter-averaged EMA model. Let

$$\mathcal{R}_{\mathbb{P}}(\theta) = \mathbb{E}_{(x,b)\sim\mathbb{P}}\left[\ell_{\text{det}}\big(S(x;\theta), b\big)\right] \tag{4}$$

denote the population surrogate risk under an unknown test distribution $\mathbb{P}$.

Proposition 2.1 (Quality–diversity recursion)**.** *The EMA update satisfies*

$$\mathcal{R}_{\mathbb{P}}(\bar{\theta}_t) = \alpha\mathcal{R}_{\mathbb{P}}(\bar{\theta}_{t-1}) + (1-\alpha)\mathcal{R}_{\mathbb{P}}(\theta_t) - \mathcal{J}_t, \tag{5}$$

where $\mathcal{J}_t$ denotes the diversity credit contributed by the incoming plain checkpoint. Consequently, the EMA update improves the surrogate risk if and only if

$$\mathcal{R}_{\mathbb{P}}(\theta_t) < \mathcal{R}_{\mathbb{P}}(\bar{\theta}_{t-1}) + \frac{\mathcal{J}_t}{1-\alpha}. \tag{6}$$

Eq. 20 shows that the benefit of EMA depends jointly on checkpoint quality and useful variation relative to the current EMA model. Thus, a later plain checkpoint may improve over earlier checkpoints yet still fail to improve EMA when it contributes insufficient diversity.

Under a local quadratic approximation, the diversity credit takes the form

$$\mathcal{J}_t = \frac{\alpha(1-\alpha)}{2} \left\| \theta_t - \bar{\theta}_{t-1} \right\|_{H_t}^2, \tag{7}$$

where $\|v\|_{H_t}^2 = v^\top H_t v$. This shows that the relevant diversity is curvature-weighted rather than simple Euclidean separation. Further detailed definition of $\mathcal{J}_t$ and complete proofs are provided in the Appendix. Next, we carefully demonstrate ReDiv, which is designed to preserve useful post-rewind parameter variation while maintaining checkpoint quality.

3 METHOD

Given the quality–diversity view, ReDiv is designed via two complementary mechanisms, namely, margin-trigger rewinding and post-rewind diversity preservation. Finally, we theoretically validate that ReDiv can effectively fulfill quality and diversity.

3.1 MARGIN-TRIGGERED REWINDING

The first challenge is to determine when continued EMA training no longer provides additional benefit. To avoid requiring an OOD validation set, we use an ID-only statistic, logit margin, that remains informative after convergence and can be flexibly applied to various downstream tasks. Formally, for an ID validation sample x, let $\mathbf{z}_t(x) = f_{\bar{\theta}_t}(x) \in \mathbb{R}^K$ denote the EMA logit vector at checkpoint t, ordered such that $z_{t,1}(x) \geq z_{t,2}(x) \geq \cdots \geq z_{t,K}(x)$. We define the sample-wise logit margin as

$$m_t(x) = z_{t,1}(x) - z_{t,2}(x). \tag{8}$$

The logit margin is defined as the median gap between the largest and second-largest logits, which has been commonly used to define uncertainty from ID data while achieving consistent performance on OOD data. Thanks to this, we leverage it to indicate the OOD performance of EMA models.

Figure 2 compares the ID-validation logit margin with far-OOD FPR95 throughout training. We use the logit margin of the EMA model on ID validation data. Unlike ID accuracy, the logit margin remains sensitive to changes after predictions have stabilized. Notably, the EMA margin exhibits a clear late-stage decline that coincides exactly with the increase in FPR95. This correspondence makes it a perfect choice for detecting the late-stage transition. In practice, we use the ID-validation logit margin (P50) at t as the median of sample-wise margins to reduce the influence of outlier samples: $q_t = \text{Median}\left(\{m_t(x) : x \in \mathcal{D}_{\text{val}}^{\text{ID}}\}\right)$.

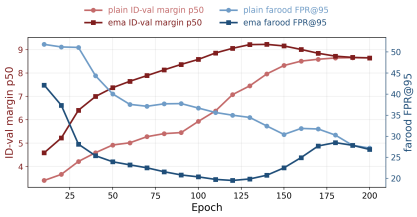


Figure 2: ID-validation logit margin and far-OOD FPR95 during training.

Particularly, ReDiv evaluates q_t periodically via the three most recent values. For checkpoints $t_1 < t_2 < t_3$, degradation is detected when neither of the two later evaluations exceeds the margin at t_1:

$$q_{t_2} \leq q_{t_1}, \qquad q_{t_3} \leq q_{t_1}. \tag{9}$$

Requiring two consecutive confirmations reduces sensitivity to a single downward fluctuation. Once Eq. 9 holds, we set the rewind point $t^\star$ to t_1.

Although the trigger is confirmed at t_3, ReDiv rewinds training to the detected point $t^\star$. We represent the restorable training state at checkpoint t as $\mathcal{R}_t = \left(t, \eta_t, \theta_t, \bar{\theta}_t, \xi_t\right)$, where η_t, θ_t, and $\bar{\theta}_t$ denote the learning rate, plain-model parameters, and EMA parameters, respectively. ξ_t denotes the optimizer state, including momentum buffers and other internal variables when applicable. ReDiv restores the training state to $\mathcal{R}_{t^\star}$.

3.2 POST-REWIND DIVERSITY PRESERVATION

Although ReDiv restores the training state to $\mathcal{R}_{t^\star}$, simply resuming training under the original optimization trajectory can eventually lead to the same late-stage EMA degradation. A natural alternative is to stop updating the EMA model once this transition is detected. While such an early-stopping strategy can prevent further degradation, it also prevents EMA from benefiting from later plain-model checkpoints, whose OOD performance may continue to improve. More importantly, early stopping does not address the underlying cause of the degradation. As discussed in Section 2.2, the late-stage decline of EMA is closely associated with reduced diversity among late-stage plain-model checkpoints. Based on this observation, ReDiv introduces post-rewind diversity preservation: instead of stopping EMA, it modifies the subsequent optimization trajectory to maintain useful variation among the plain-model checkpoints used for EMA averaging.

Consider the generic parameter update

$$\Delta_t = \theta_{t+1} - \theta_t = -\eta_t \mu_t, \tag{10}$$

where η_t denotes the learning rate and μ_t denotes the effective optimizer update at iteration t. The resulting parameter movement is jointly determined by the learning-rate scale η_t and the effective update μ_t, whose magnitude and direction characterize how the model moves in parameter space. Based on this decomposition, ReDiv preserves post-rewind diversity from two complementary aspects: it fixes the learning-rate scale to prevent continued learning-rate decay from suppressing parameter movement, and applies sharpness-aware minimization (SAM) to reshape the magnitude and direction of μ_t.

Post-rewind trajectory-scale preservation. Rewinding restores the model to an earlier training state, but the original learning-rate schedule continues to shrink the scale of subsequent updates. As training proceeds, this scheduler-induced reduction can drive the post-rewind trajectory back toward the same concentrated regime that limits EMA diversity.

To prevent this recurrence, ReDiv preserves the post-rewind update scale by anchoring the learning rate to its value at the rewind point: $\eta_t = \eta_{t^\star}$, where $t \geq t^\star$. From Eq. 10, $\|\Delta_t\|_2 = \eta_t \|\mu_t\|_2$, so fixing η_t removes the continued scheduler-driven reduction after rewinding. This design does not force the update magnitude to remain constant, since $\|\mu_t\|_2$ can still evolve during training.

Post-rewind gradient diversification. Although post-rewind trajectory-scale preservation prevents further shrinkage, it does not alter the effective update μ_t in Eq. 10. As a result, the updates may still become increasingly similar across iterations. To address this limitation, ReDiv applies SAM after rewinding to modify the magnitude and direction through local loss geometry.

For a mini-batch $\mathcal{B}_t$ at iteration t, let $\mu_t = \nabla_\theta \mathcal{L}_{\mathcal{B}_t}(\theta_t)$ denote the gradient of the plain model. SAM first computes a local perturbation by

$$\epsilon_t^\star \in \arg\max_{\|\epsilon\|_2 \leq \rho} \mathcal{L}_{\mathcal{B}_t}(\theta_t + \epsilon), \tag{11}$$

which is approximated by $\epsilon_t = \rho \frac{\mu_t}{\|\mu_t\|_2}$. The SAM gradient $\tilde{\mu}_t$ is then evaluated at the perturbed parameters by

$$\tilde{\mu}_t = \nabla_\theta \mathcal{L}_{\mathcal{B}_t}(\theta_t + \epsilon_t). \tag{12}$$

Therefore, the post-rewind update becomes

$$\theta_{t+1} = \theta_t - \eta_{t^\star} \cdot \tilde{\mu}_t. \tag{13}$$

To characterize how SAM reshapes the intrinsic update, let

$$H_t = \nabla^2 \mathcal{L}_{\mathcal{B}_t}(\theta_t), \qquad \hat{g}_t = \frac{g_t}{\|g_t\|_2}. \tag{14}$$

Assuming that the Hessian is locally Lipschitz, Taylor expansion around θ_t gives

$$\tilde{g}_t = g_t + \rho H_t \hat{g}_t + \mathcal{O}(\rho^2). \tag{15}$$

Relative to the vanilla gradient, SAM introduces the curvature-dependent correction $\rho H_t \hat{g}_t$.

Let $P_t^{\perp} = I - \hat{g}_t \hat{g}_t^{\top}$ denote the projection onto the subspace orthogonal to the plain-gradient direction. The first-order SAM correction in Eq.15 decomposes as

$$\rho H_t \hat{g}_t = \underbrace{\rho(\hat{g}_t^{\top} H_t \hat{g}_t)\hat{g}_t}_{\text{magnitude variation}} + \underbrace{\rho P_t^{\perp} H_t \hat{g}_t}_{\text{directional variation}} . \tag{16}$$

The first term in Eq. 16 changes the first-order coefficient along the plain-gradient direction, thereby modulating the intrinsic update magnitude. The second term deviates from the original gradient direction whenever it is nonzero. SAM can therefore induce first-order changes in both the magnitude and direction of the post-rewind effective update.

3.3 Theoretical Analysis

The quality–diversity analysis in Section 2.3 shows that, under a local quadratic approximation, the EMA diversity credit depends on the curvature-weighted separation between the incoming plain checkpoint and the current EMA model, as denoted by $d_t = \theta_t - \bar{\theta}_{t-1}$.

Proposition 3.1 (Post-rewind diversity dynamics)**.** *If the plain model is updated as $\theta_{t+1} = \theta_t - \eta_t u_t$, then the plain-to-EMA separation satisfies*

$$d_{t+1} = \alpha d_t - \eta_t u_t. \tag{17}$$

Under the same local quadratic approximation, the EMA diversity credit at step $t+1$ is

$$\mathcal{J}_{t+1} = \frac{\alpha(1-\alpha)}{2} \left\| \alpha d_t - \eta_t u_t \right\|^2_{H^{\mathcal{R}}_{t+1}} , \tag{18}$$

where $\mathcal{J}_{t+1}$ measures the curvature-weighted separation between the incoming plain checkpoint θ_{t+1} and the current EMA model $\bar{\theta}_t$.

Eq. 18 shows that the evolution of the EMA diversity credit depends jointly on the learning-rate scale η_t and the effective update u_t. ReDiv acts on these two factors separately through post-rewind learning-rate preservation and SAM, respectively. The full analysis is provided in the Appendix.

4 Experiments

4.1 Experimental Setup

CIFAR benchmarks. We use CIFAR-10 and CIFAR-100 (Krizhevsky et al., 2009) as ID datasets and follow the OpenOOD protocol (Zhang et al., 2023a) to organize OOD test sets into near- and far-OOD groups. For CIFAR-10, the near-OOD group contains CIFAR-100 and Tiny ImageNet (TIN) (Le & Yang, 2015); for CIFAR-100, it contains CIFAR-10 and TIN. The far-OOD group comprises MNIST, SVHN (Netzer et al., 2011), Texture (Cimpoi et al., 2014), and Places365 (Zhou et al., 2017) for both ID datasets.

ImageNet benchmarks. We additionally evaluate on ImageNet-50 and ImageNet-200, two class subsets of ImageNet (Deng et al., 2009). The near-OOD group contains SSB-hard (Vaze et al., 2021) and NINCO (Bitterwolf et al., 2023), while the far-OOD group contains iNaturalist (Van Horn et al., 2018), Texture (Cimpoi et al., 2014), and OpenImage-O (Wang et al., 2022).

Compared configurations. For each OOD score, *Plain* denotes the final online model and *EMA* denotes standard exponential averaging. *Rewind* uses the ID-margin trigger to locate $t^{\star}$, stops optimization, and directly evaluates the checkpoint at $t^{\star}$. *ReDiv* instead restores the complete training state at $t^{\star}$ and resumes optimization with a fixed learning rate and SAM. We evaluate MSP (Hendrycks & Gimpel, 2016), Energy (Liu et al., 2020), ReAct (Sun et al., 2021), and NCI (Liu & Qin, 2025) on all four ID benchmarks. We report FPR95 and AUROC as percentages.

4.2 Main Results

Consistent improvement over standard EMA. Tables 1 and 2 summarize the main OOD detection results on CIFAR-10, CIFAR-100, ImageNet-50, and ImageNet-200 under near-OOD and far-OOD

Table 1: OOD detection on CIFAR-10 and CIFAR-100. Results are mean ± standard deviation, with the best result in each scoring-function block shown in bold. Both Rewind and ReDiv locate $t^\star$ using only ID-validation margin P50; Rewind stops at $t^\star$, whereas ReDiv resumes training.

Method	CIFAR-10				CIFAR-100			
	Near-OOD		Far-OOD		Near-OOD		Far-OOD	
	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
MSP	42.85 ± 1.68	89.29 ± 0.22	25.87 ± 1.76	92.08 ± 0.29	53.01 ± 0.62	82.17 ± 0.10	55.47 ± 2.38	79.94 ± 1.14
MSP + EMA	42.55 ± 1.15	89.35 ± 0.17	25.32 ± 1.26	92.15 ± 0.25	52.20 ± 1.06	82.35 ± 0.21	54.77 ± 1.08	80.24 ± 0.59
MSP + Rewind	29.40 ± 0.60	90.89 ± 0.16	20.95 ± 0.54	93.53 ± 0.05	50.96 ± 0.88	81.73 ± 0.18	55.49 ± 1.30	79.01 ± 0.41
MSP + ReDiv	$\mathbf{26.46 \pm 0.50}$	$\mathbf{91.98 \pm 0.11}$	$\mathbf{18.51 \pm 0.26}$	$\mathbf{94.43 \pm 0.08}$	$\mathbf{49.60 \pm 0.54}$	$\mathbf{82.75 \pm 0.19}$	$\mathbf{52.91 \pm 0.79}$	$\mathbf{80.33 \pm 0.32}$
Energy	50.17 ± 1.85	89.50 ± 0.24	29.18 ± 2.40	92.73 ± 0.28	53.84 ± 0.21	82.52 ± 0.03	54.87 ± 2.69	81.07 ± 1.38
Energy + EMA	50.34 ± 1.67	89.58 ± 0.15	28.61 ± 1.72	92.82 ± 0.29	53.05 ± 0.72	82.75 ± 0.17	54.19 ± 1.26	81.51 ± 0.67
Energy + Rewind	28.95 ± 0.46	92.53 ± 0.12	17.64 ± 0.59	95.71 ± 0.10	51.55 ± 0.53	83.51 ± 0.18	49.16 ± 4.01	83.95 ± 1.62
Energy + ReDiv	$\mathbf{24.10 \pm 0.62}$	$\mathbf{93.76 \pm 0.15}$	$\mathbf{14.17 \pm 0.51}$	$\mathbf{96.53 \pm 0.17}$	$\mathbf{49.03 \pm 0.49}$	$\mathbf{84.25 \pm 0.06}$	$\mathbf{45.55 \pm 2.58}$	$\mathbf{85.38 \pm 0.79}$
ReAct	53.11 ± 3.46	89.09 ± 0.59	31.37 ± 5.25	92.40 ± 0.55	54.56 ± 2.41	82.06 ± 0.16	51.68 ± 3.78	81.85 ± 1.65
ReAct + EMA	53.38 ± 3.08	89.13 ± 0.55	31.37 ± 4.65	92.46 ± 0.51	53.80 ± 2.43	82.26 ± 0.22	51.21 ± 2.84	82.20 ± 1.09
ReAct + Rewind	29.96 ± 1.59	92.28 ± 0.32	18.09 ± 0.75	95.60 ± 0.12	54.17 ± 0.17	82.62 ± 0.15	43.60 ± 4.22	85.63 ± 1.99
ReAct + ReDiv	$\mathbf{24.52 \pm 0.54}$	$\mathbf{93.71 \pm 0.15}$	$\mathbf{14.45 \pm 0.43}$	$\mathbf{96.51 \pm 0.15}$	$\mathbf{50.94 \pm 0.75}$	$\mathbf{83.72 \pm 0.08}$	$\mathbf{41.16 \pm 3.42}$	$\mathbf{86.83 \pm 1.14}$
NCI	55.09 ± 2.78	88.48 ± 0.28	32.48 ± 6.02	91.62 ± 0.68	49.98 ± 0.08	82.56 ± 0.10	53.46 ± 0.93	81.27 ± 0.48
NCI + EMA	54.79 ± 2.40	88.49 ± 0.27	32.60 ± 5.88	91.63 ± 0.63	$\mathbf{49.54 \pm 0.34}$	82.66 ± 0.17	53.31 ± 0.47	81.40 ± 0.18
NCI + Rewind	41.92 ± 5.16	89.83 ± 0.63	25.16 ± 2.05	93.73 ± 0.25	50.03 ± 0.75	83.56 ± 0.16	48.88 ± 1.79	82.76 ± 0.71
NCI + ReDiv	$\mathbf{27.56 \pm 0.69}$	$\mathbf{92.19 \pm 0.14}$	$\mathbf{18.14 \pm 0.40}$	$\mathbf{95.25 \pm 0.06}$	50.03 ± 1.03	$\mathbf{83.99 \pm 0.18}$	$\mathbf{47.68 \pm 1.18}$	$\mathbf{83.34 \pm 0.38}$

Table 2: OOD detection on ImageNet-50 and ImageNet-200.

Method	ImageNet-50				ImageNet-200			
	Near-OOD		Far-OOD		Near-OOD		Far-OOD	
	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
MSP	34.15 ± 0.30	90.24 ± 0.08	24.10 ± 0.69	93.26 ± 0.09	54.57 ± 0.31	83.84 ± 0.02	34.70 ± 0.78	90.36 ± 0.12
MSP + EMA	33.76 ± 0.48	90.32 ± 0.07	24.00 ± 0.50	93.33 ± 0.06	54.57 ± 0.46	83.84 ± 0.02	34.26 ± 0.40	90.48 ± 0.05
MSP + Rewind	32.07 ± 0.46	90.68 ± 0.09	22.96 ± 0.68	94.09 ± 0.12	53.73 ± 0.76	83.89 ± 0.10	33.50 ± 0.21	90.83 ± 0.29
MSP + ReDiv	$\mathbf{30.17 \pm 0.27}$	$\mathbf{91.46 \pm 0.16}$	$\mathbf{19.84 \pm 0.24}$	$\mathbf{94.88 \pm 0.04}$	$\mathbf{52.40 \pm 0.62}$	$\mathbf{84.45 \pm 0.03}$	$\mathbf{30.66 \pm 0.42}$	$\mathbf{91.82 \pm 0.07}$
Energy	36.13 ± 0.33	90.70 ± 0.12	23.13 ± 0.92	94.22 ± 0.20	60.04 ± 1.20	82.87 ± 0.24	36.02 ± 0.88	90.47 ± 0.20
Energy + EMA	35.74 ± 0.61	90.82 ± 0.07	22.89 ± 0.76	94.30 ± 0.18	60.14 ± 1.34	82.83 ± 0.18	35.06 ± 0.37	90.69 ± 0.05
Energy + Rewind	30.84 ± 2.18	91.99 ± 0.33	16.95 ± 1.88	95.87 ± 0.42	59.32 ± 1.40	83.05 ± 0.18	32.79 ± 1.43	91.45 ± 0.54
Energy + ReDiv	$\mathbf{28.11 \pm 0.35}$	$\mathbf{92.70 \pm 0.05}$	$\mathbf{13.74 \pm 0.85}$	$\mathbf{96.53 \pm 0.10}$	$\mathbf{56.72 \pm 1.14}$	$\mathbf{83.71 \pm 0.22}$	$\mathbf{28.08 \pm 0.30}$	$\mathbf{92.87 \pm 0.05}$
ReAct	35.73 ± 1.33	90.91 ± 0.20	18.99 ± 1.65	95.04 ± 0.27	60.82 ± 0.76	82.65 ± 0.11	28.98 ± 0.53	91.94 ± 0.09
ReAct + EMA	35.39 ± 1.43	91.00 ± 0.17	18.71 ± 1.73	95.10 ± 0.31	60.62 ± 0.67	82.61 ± 0.10	28.75 ± 0.50	92.02 ± 0.11
ReAct + Rewind	31.51 ± 2.94	91.93 ± 0.39	14.74 ± 1.11	96.34 ± 0.26	60.47 ± 0.81	82.66 ± 0.09	27.36 ± 1.37	92.61 ± 0.58
ReAct + ReDiv	$\mathbf{27.36 \pm 1.17}$	$\mathbf{92.75 \pm 0.11}$	$\mathbf{12.61 \pm 1.11}$	$\mathbf{96.80 \pm 0.21}$	$\mathbf{57.91 \pm 0.73}$	$\mathbf{83.25 \pm 0.11}$	$\mathbf{23.32 \pm 0.54}$	$\mathbf{93.86 \pm 0.13}$
NCI	30.24 ± 0.73	91.69 ± 0.17	15.87 ± 0.53	95.88 ± 0.10	55.69 ± 0.20	84.04 ± 0.04	24.10 ± 0.28	93.48 ± 0.11
NCI + EMA	30.22 ± 0.56	91.73 ± 0.17	15.79 ± 0.54	95.91 ± 0.11	55.66 ± 0.37	84.01 ± 0.05	24.00 ± 0.30	93.52 ± 0.07
NCI + Rewind	33.73 ± 2.76	91.55 ± 0.46	14.19 ± 1.96	96.78 ± 0.28	54.94 ± 0.48	84.03 ± 0.05	23.41 ± 0.56	93.66 ± 0.15
NCI + ReDiv	$\mathbf{28.69 \pm 0.63}$	$\mathbf{92.53 \pm 0.10}$	$\mathbf{11.22 \pm 0.30}$	$\mathbf{97.45 \pm 0.06}$	$\mathbf{53.24 \pm 0.01}$	$\mathbf{84.46 \pm 0.06}$	$\mathbf{21.65 \pm 0.36}$	$\mathbf{94.30 \pm 0.05}$

regimes. ReDiv consistently outperforms standard EMA across datasets and OOD regimes. Notably, the final EMA model is often close to the corresponding plain model. In contrast, ReDiv yields substantial improvements. For example, on CIFAR-10 with ReAct, ReDiv reduces near-OOD FPR95 from 53.38% to 24.52% and far-OOD FPR95 from 31.37% to 14.45%, while improving the corresponding AUROC from 89.13% to 93.71% and from 92.46% to 96.51%. The improvement also extends to larger-scale benchmarks: on ImageNet-200 with Energy, ReDiv reduces far-OOD FPR95 from 35.06% to 28.08% and improves AUROC from 90.69% to 92.87%. Overall, these results show that ReDiv consistently strengthens EMA-based OOD detection.

Generalization across scoring functions. ReDiv consistently improves over standard EMA across MSP, Energy, ReAct, and NCI, indicating that its effectiveness is not tied to a specific OOD scoring function. For example, on CIFAR-100 far-OOD, ReDiv reduces FPR95 from 54.77% to 52.91% with MSP, from 54.19% to 45.55% with Energy, from 51.21% to 41.16% with ReAct, and from 53.31% to 47.68% with NCI. A similar pattern is observed on ImageNet-50 near-OOD, where FPR95 decreases from 33.76% to 30.17% with MSP, from 35.74% to 28.11% with Energy, from 35.39% to 27.36% with ReAct, and from 30.22% to 28.69% with NCI. These consistent gains across scoring rules show that ReDiv is complementary to different post-hoc OOD scoring functions rather than relying on a particular score formulation.

Compatibility with Training-Time Methods. To evaluate whether ReDiv complements training-time OOD detection methods, we combine it with LogitNorm Wei et al. (2022) and use MSP for scoring. Table 3 compares plain LogitNorm, LogitNorm with EMA, Rewind, and the complete ReDiv method. EMA provides only modest gains over LogitNorm, reducing Near-Avg and Far-Avg FPR95 from 26.86% and 13.40% to 26.66% and 13.26%, respectively. Rewind mainly benefits far-OOD detection, whereas ReDiv improves both regimes, achieving 25.41% Near-Avg and 12.45%

Table 3: Compatibility of ReDiv with LogitNorm on CIFAR-10 (FPR95↓).

Method	CIFAR-100	TIN	Near-Avg	MNIST	SVHN	Texture	Places365	Far-Avg
LogitNorm	31.99 ± 0.57	21.73 ± 0.19	26.86 ± 0.35	3.45 ± 0.56	10.41 ± 1.85	20.74 ± 1.16	18.99 ± 1.41	13.40 ± 0.47
LogitNorm + EMA	31.77 ± 0.56	21.56 ± 0.56	26.66 ± 0.46	3.58 ± 0.63	10.30 ± 1.97	20.35 ± 1.30	18.80 ± 0.93	13.26 ± 0.46
LogitNorm + Rewind	32.27 ± 0.58	21.30 ± 0.67	26.79 ± 0.61	4.68 ± 0.82	10.40 ± 3.40	$\mathbf{16.89 \pm 1.34}$	18.67 ± 1.84	12.66 ± 0.51
LogitNorm + ReDiv	$\mathbf{30.55 \pm 0.32}$	$\mathbf{20.26 \pm 0.78}$	$\mathbf{25.41 \pm 0.53}$	$\mathbf{3.20 \pm 0.43}$	$\mathbf{8.99 \pm 0.70}$	19.61 ± 1.50	$\mathbf{18.01 \pm 0.43}$	$\mathbf{12.45 \pm 0.17}$

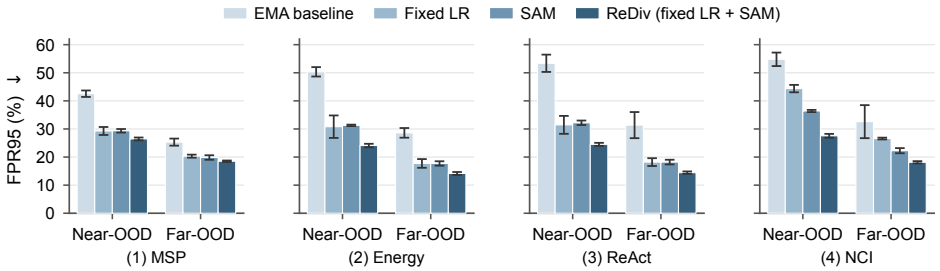


Figure 3: Ablation of fixed learning rate and SAM on CIFAR-10. EMA denotes the standard-training baseline; the other variants continue from the same rewound state. Fixed LR and SAM activate one diversification component at a time, whereas ReDiv combines both. Bars show means and error bars denote one standard deviation. Lower FPR95 is better.

Far-Avg. Compared with LogitNorm + EMA, ReDiv reduces Near-Avg and Far-Avg FPR95 by 1.25% and 0.81%, respectively. These results indicate that ReDiv remains effective on top of LogitNorm, supporting its complementarity with training-time OOD detection methods.

4.3 Ablation Studies

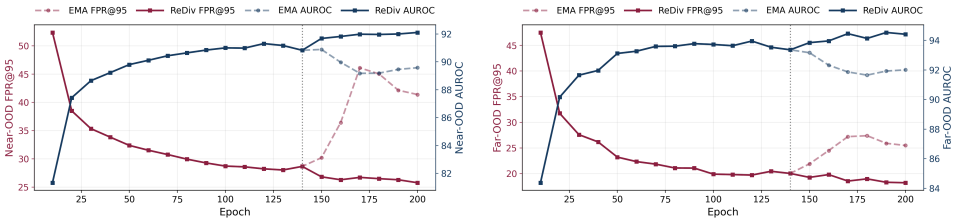
Rewinding is effective but not sufficient. Tables 1 and 2 disentangle the effect of margin-based checkpoint selection from that of post-rewind optimization. Compared with the final standard EMA model, Rewind improves OOD detection in most results, showing that the margin trigger can identify a more favorable checkpoint before late-stage degradation. Starting from the same detected checkpoint, ReDiv resumes training with post-rewind diversity preservation and further improves over Rewind in 31 of 32 FPR95 comparisons and all 32 AUROC comparisons. These results show that rewinding itself is effective and continued post-rewind optimization is important.

To isolate the two post-rewind components, we compare ReDiv with variants using only the fixed learning rate or SAM, together with the standard EMA baseline. All post-rewind variants start from the same restored state and use the same update budget. Figure 3 reports FPR95 on CIFAR-10 across four OOD scoring functions under near- and far-OOD settings.

A fixed learning rate consistently improves OOD detection. Fixing the post-rewind learning rate consistently improves FPR95 over standard EMA across MSP, Energy, ReAct, and NCI under both near- and far-OOD scenarios. The improvement is particularly clear with ReAct: near-OOD FPR95 decreases from 53.38% to 31.46%, while far-OOD FPR95 decreases from 31.37% to 18.20%. These results empirically show that preventing continued learning-rate decay is beneficial.

SAM independently improves OOD detection. Applying SAM alone exhibits the same consistent advantage over standard EMA across the four scoring functions and both OOD scenarios. Its distinct contribution is clearest with NCI: SAM lowers near- and far-OOD FPR95 from 54.79% and 32.60% to 36.40% and 22.27%, outperforming the corresponding fixed-LR-only results of 44.37% and 26.59%. This empirical advantage is consistent with the update-geometry mechanism.

A fixed learning rate and SAM are complementary. Combining the two components consistently yields lower FPR95 than either component alone across all four scoring functions and both OOD scenarios. The complementarity is most pronounced with NCI: ReDiv reduces near-OOD FPR95 to 27.56%, compared with 44.37% for fixed LR and 36.40% for SAM, and reduces far-OOD FPR95 to 18.14%, compared with 26.59% and 22.27%, respectively. These results empirically support Proposition 3.1 in Section 3.3: the fixed learning rate acts on the learning-rate scale η_t, while SAM modifies the effective update u_t, and their combination yields the strongest performance.



(a) Near-OOD (b) Far-OOD

Figure 5: Checkpoint-wise OOD performance of standard EMA and ReDiv on CIFAR-10 in a representative run. The vertical dotted line marks the intervention at epoch 140; both trajectories are identical before this point. Each panel reports FPR95 on the left axis and AUROC on the right axis. After the intervention, standard EMA exhibits late-stage degradation, whereas ReDiv continues to improve under both (a) near-OOD and (b) far-OOD settings.

4.4 Mechanism Analysis

Does ReDiv preserve trajectory diversity? Improved OOD performance alone does not establish the proposed mechanism. We therefore examine the late-stage plain-model trajectory that supplies checkpoints to EMA. For a 10-epoch interval, we define $\Delta\theta_t = \theta_t - \theta_{t-10}$. A small relative displacement indicates reduced parameter movement, while high directional similarity indicates increasingly aligned updates. As shown in Figure 4, the vanilla plain trajectory contracts substantially in late training: its relative displacement drops from about 0.31 at epoch 160 to nearly zero at epoch 200, while directional similarity increases from about -0.13 to 0.71. In contrast, ReDiv maintains a relative displacement of about 0.38–0.40 and low directional similarity. These results indicate that ReDiv preserves larger and less aligned late-stage updates, reducing trajectory redundancy.

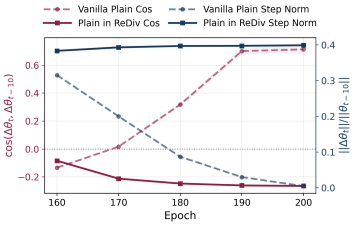


Figure 4: Late-stage statistics of the plain-model trajectory.

Does ReDiv prevent late-stage EMA degradation? Figure 5 compares the checkpoint-wise OOD performance of standard EMA and ReDiv on CIFAR-10. The two methods share the same training history before the intervention at epoch 140 and diverge only afterward, isolating the effect of post-rewind optimization. After the intervention, ReDiv continues to improve under both near- and far-OOD settings, whereas standard EMA exhibits the late-stage reversal observed previously. For near-OOD, ReDiv reduces FPR95 from about 29% to 26% and raises AUROC above 92%, while standard EMA ends above 41% FPR95 and below 90% AUROC. A similar trend appears for far-OOD, where ReDiv reaches about 18% FPR95 and above 94% AUROC, compared with roughly 25% and 92% for standard EMA. These results show that post-rewind optimization effectively prevents the late-stage EMA degradation and continues to improve the EMA model.

5 Conclusion

We revisited exponential moving average in OOD detection and identified a consistent late-stage degradation phenomenon: EMA initially improves OOD performance but gradually loses its advantage as training proceeds. Our analysis links this behavior to increasing redundancy in the late-stage optimization trajectory. To address this issue, we proposed **ReDiv**, a rewind-and-diversify framework that detects a persistent loss of improvement using only the ID-validation logit margin and reshapes the post-rewind trajectory with a fixed learning rate and SAM, while leaving the EMA update rule unchanged. Experiments across multiple datasets, OOD regimes, and scoring functions show that ReDiv consistently improves over both plain training and standard EMA without requiring OOD validation data or architectural modifications.

AI USE STATEMENT

We used generative AI tools to assist with language editing. The authors developed and verified all scientific ideas, theoretical results, experimental design, analysis, and conclusions. The authors take full responsibility for the content of this paper.

ETHICS STATEMENT

This work studies OOD detection using publicly available models and datasets. It does not involve human subjects or private personal data.

REPRODUCIBILITY STATEMENT

We have provided detailed proofs of the theoretical results, as well as detailed experimental setups, model configurations, training protocols, and evaluation benchmarks in the experiments and appendix sections to facilitate reproduction of our study.

REFERENCES

Julian Bitterwolf, Maximilian Mueller, and Matthias Hein. In or out? fixing imagenet out-of-distribution detection evaluation. *arXiv preprint arXiv:2306.00826*, 2023.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3606–3613, 2014.

Andrea Corsini and Shanchieh Jay Yang. Are existing out-of-distribution techniques suitable for network intrusion detection? *arXiv preprint arXiv:2308.14376*, 2023.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Xuefeng Du, Zhaoning Wang, Mu Cai, and Yixuan Li. Vos: Learning what you don't know by virtual outlier synthesis. *arXiv preprint arXiv:2202.01197*, 2022.

Pascal Germain, Alexandre Lacasse, Francois Laviolette, Mario Marchand, and Jean-Francis Roy. Risk bounds for the majority vote: From a pac-bayesian analysis to a learning algorithm. *arXiv preprint arXiv:1503.08329*, 2015.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Rundong He, Yue Yuan, Zhongyi Han, Fan Wang, Wan Su, Yilong Yin, Tongliang Liu, and Yongshun Gong. Exploring channel-aware typical features for out-of-distribution detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 12402–12410, 2024.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.

Rui Huang, Andrew Geng, and Yixuan Li. On the importance of gradients for detecting distributional shifts in the wild. *Advances in Neural Information Processing Systems*, 34:677–689, 2021.

Julian Katz-Samuels, Julia B Nakhleh, Robert Nowak, and Yixuan Li. Training ood detectors in their natural habitats. In *ICML*, pp. 10848–10865. PMLR, 2022.

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

Anders Krogh and Jesper Vedelsby. Neural network ensembles, cross validation, and active learning. *Advances in neural information processing systems*, 7, 1994.

Yann Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 31, 2018.

Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv preprint arXiv:1706.02690*, 2017.

Ziqian Lin, Sreya Dutta Roy, and Yixuan Li. Mood: Multi-level out-of-distribution detection. In *CVPR*, pp. 15313–15323, 2021.

Litian Liu and Yao Qin. Detecting out-of-distribution through the lens of neural collapse. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 15424–15433, 2025.

Weitang Liu, Xiaoyun Wang, John D Owens, and Yixuan Li. Energy-based out-of-distribution detection. *arXiv preprint arXiv:2010.03759*, 2020.

Shuo Lu, Yingsheng Wang, Lijun Sheng, Lingxiao He, Aihua Zheng, and Jian Liang. Out-of-distribution detection: A task-oriented survey of recent advances. *ACM Computing Surveys*, 58 (2):1–39, 2025.

Yifei Ming, Ying Fan, and Yixuan Li. Poem: Out-of-distribution detection with posterior sampling. In *ICML*, pp. 15650–15665. PMLR, 2022.

Daniel Morales-Brotons, Thijs Vogels, and Hadrien Hendrikx. Exponential moving average of weights in deep learning: Dynamics and benefits. *Trans. Mach. Learn. Res.*, 2024, 2024.

Tanya Nair, Doina Precup, Douglas L Arnold, and Tal Arbel. Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation. *Medical image analysis*, 59:101557, 2020.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011.

Chandramouli Shama Sastry and Sageev Oore. Detecting out-of-distribution examples with gram matrices. In *International Conference on Machine Learning*, pp. 8491–8501. PMLR, 2020.

Yang Shu, Xingzhuo Guo, Jialong Wu, Ximei Wang, Jianmin Wang, and Mingsheng Long. Clipood: Generalizing clip to out-of-distributions. In *International conference on machine learning*, pp. 31716–31731. PMLR, 2023.

Yiyou Sun and Yixuan Li. Dice: Leveraging sparsification for out-of-distribution detection. In *European Conference on Computer Vision*, pp. 691–708. Springer, 2022.

Yiyou Sun, Chuan Guo, and Yixuan Li. React: Out-of-distribution detection with rectified activations. *Advances in Neural Information Processing Systems*, 34, 2021.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017.

Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8769–8778, 2018.

Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Open-set recognition: A good closed-set classifier is all you need? 2021.

Hanjing Wang and Qiang Ji. Diversity-enhanced probabilistic ensemble for uncertainty estimation. In *Uncertainty in Artificial Intelligence*, pp. 2214–2225. PMLR, 2023.

Haoqi Wang, Zhizhong Li, Litong Feng, and Wayne Zhang. Vim: Out-of-distribution with virtual-logit matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4921–4930, 2022.

Hongxin Wei, Renchunzi Xie, Hao Cheng, Lei Feng, Bo An, and Yixuan Li. Mitigating neural network overconfidence with logit normalization. *arXiv preprint arXiv:2205.09310*, 2022.

Danny Wood, Tingting Mu, Andrew M Webb, Henry WJ Reeve, Mikel Lujan, and Gavin Brown. A unified theory of diversity in ensemble learning. *Journal of machine learning research*, 24(359): 1–49, 2023.

Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, et al. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International conference on machine learning*, pp. 23965–23998. Pmlr, 2022.

Haipeng Xiong, Kai Xu, and Angela Yao. Diagnosing pretrained models for out-of-distribution detection. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1836–1845. IEEE, 2025.

Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *arXiv preprint arXiv:2110.11334*, 2021.

Jingyang Zhang, Jingkang Yang, Pengyun Wang, Haoqi Wang, Yueqian Lin, Haoran Zhang, Yiyou Sun, Xuefeng Du, Yixuan Li, Ziwei Liu, et al. Openood v1. 5: Enhanced benchmark for out-of-distribution detection. *arXiv preprint arXiv:2306.09301*, 2023a.

Jinsong Zhang, Qiang Fu, Xu Chen, Lun Du, Zelin Li, Gang Wang, Xiaoguang Liu, Shi Han, and Dongmei Zhang. Out-of-distribution detection based on in-distribution data patterns memorization with modern hopfield energy. In *International Conference on Learning Representations (ICLR'23)*, February 2023b.

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017.

Yao Zhu, YueFeng Chen, Chuanlong Xie, Xiaodan Li, Rong Zhang, Hui Xue, Xiang Tian, bolun zheng, and Yaowu Chen. Boosting out-of-distribution detection with typical features, 2022.

A Related Works

A.1 Out-of-Distribution Detection

Test-time OOD detection methods. Test-time OOD detection methods have the advantage of being easy to use without modifying the training procedure and objective (Yang et al., 2021). The test-time approaches do not require retraining the model, perform well, and are easy to implement in the real world. Test-time OOD detection methods can be categorized into confidence-based, feature-based, distance-based, gradient-based, pruning-based, and activation-based methods. Confidence-based methods use the confidence score of a pre-trained classifier to detect OOD data. The underlying assumption is that the ID data should receive a high confidence score, while the OOD data should receive a low confidence score. MSP (Hendrycks & Gimpel, 2016) directly uses the maximum Softmax score to determine whether the test sample is an ID or OOD. ODIN (Liang et al., 2017) improves the SoftMax score by perturbing the input and applying temperature scaling to the logits. Energy (Liu et al., 2020) demonstrates that the Energy score (i.e., logsumexp of logits) outperforms the SoftMax score in distinguishing between ID and OOD data. Feature-based methods include GRAM (Sastry & Oore, 2020) and SHE (Zhang et al., 2023b). GRAM computes the Gram matrix within the hidden layers. SHE uses the energy function defined in modern Hopfield networks. Distance-based methods consider OOD data to be farther away from the training set than ID data. Mahalanobis (Lee et al., 2018) calculates the minimum Mahalanobis distance between the test data and the class centroids of the training set as an OOD score. Gradient-based approaches (Huang et al.,

2021) use gradient statistics to calculate OOD scores. Pruning-based methods prune the weights of the model to address overconfident predictions of OOD data. DICE (Sun & Li, 2022) prunes the weights of the classification layer to address overconfidence in the model's prediction of OOD data. Activation-based methods attempt to maximize the gap between ID and OOD data by truncating abnormally low or high activations. ReAct (Sun et al., 2021) observes that OOD inputs trigger abnormally high activations. BATS (Zhu et al., 2022; He et al., 2024) exhibits efficacy by truncating both abnormally low and abnormally high activations of each channel.

Training-time OOD detection methods. Unlike testing-time methods, training-time methods aim to mitigate overconfident predictions for OOD data during the training period. According to whether the OOD-supervised signals are used in the training process, training-time methods can be categorized into OOD-free and OOD-needed methods. The representatives of OOD-free methods are (Wei et al., 2022; Lin et al., 2021). Wei et al. (2022) decoupled the influence of logits' norm from the training procedure by incorporating LogitNorm into the cross-entropy loss. Lin et al. (2021) exploited intermediate classifier outputs for dynamic and efficient OOD inference. The OOD-needed methods aim to calibrate the model by OOD-supervised signals (Ming et al., 2022; Katz-Samuels et al., 2022; Du et al., 2022).

Our work is orthogonal to both lines of research. Rather than designing a new test-time score or introducing additional OOD supervision during training, we study the training dynamics of EMA itself. ReDiv improves EMA-based OOD detection by locating its saturation point with ID validation data and diversifying the subsequent optimization trajectory. As a result, ReDiv can be inserted into existing training pipelines and combined with both training-time and test-time OOD methods.

B Experiments

B.1 Experimental Setup

Evaluation metrics. We report FPR95 and AUROC as percentages. FPR95 is the false-positive rate on OOD inputs when 95% of ID inputs are accepted; lower is better. AUROC measures threshold-independent ranking quality; higher is better. We also report ID classification accuracy to assess whether improved OOD detection preserves ID utility. All results are presented as mean $\pm$ standard deviation over three independent runs. Unless stated otherwise, all reported improvements are absolute differences between these percentage-valued metrics.

Training details. Following OpenOOD (Zhang et al., 2023a), we use ResNet-18 (He et al., 2016) as the backbone and optimize the cross-entropy objective with SGD for 200 epochs. The initial learning rate is 0.1 and follows cosine decay, with momentum 0.9 and weight decay 5×10^{-4}. We use batch sizes of 128 for CIFAR-10/100 and 256 for ImageNet-50/200. EMA uses decay $\alpha = 0.999$. ReDiv evaluates the ID-validation margin every 10 epochs, restores the complete training state when the three-checkpoint condition is met, fixes the learning rate at its restored value, and applies SAM with $\rho = 0.05$ for the remaining trajectory. ReDiv accesses only the standard ID validation split; no OOD samples, labels, or test performance are used for training, rewinding, or model selection.

B.2 More Results

Sensitivity to the SAM radius. The perturbation radius ρ controls the strength of the SAM update. We vary ρ while keeping the rewind point, fixed learning rate, and all other training settings unchanged. As shown in Table 4, increasing ρ from 0.01 to 0.05 improves both near- and far-OOD performance, reducing Near-Avg FPR95 from 28.36% to 26.46% and Far-Avg FPR95 from 20.57% to 18.51%. Larger values exhibit a trade-off: $\rho = 0.20$ further improves Far-Avg to 17.14% but slightly worsens Near-Avg, while $\rho = 0.30$ degrades the aggregate performance. This suggests that ρ balances post-rewind diversification against checkpoint quality. ReDiv remains effective over the intermediate range $\rho \in [0.05, 0.20]$, indicating limited sensitivity to a single tuned value. We use $\rho = 0.05$ as the fixed default in all main experiments.

Table 4: Sensitivity to the SAM radius ρ on CIFAR-10 (FPR95↓).

ρ	CIFAR-100	TIN	Near-Avg	MNIST	SVHN	Texture	Places365	Far-Avg
0.01	30.65 ± 0.44	26.07 ± 0.22	28.36 ± 0.17	18.61 ± 0.38	15.21 ± 0.75	20.64 ± 1.13	27.84 ± 0.82	20.57 ± 0.59
0.02	29.73 ± 0.84	25.00 ± 0.51	27.36 ± 0.57	15.83 ± 2.20	14.27 ± 0.28	21.17 ± 1.95	28.26 ± 1.56	19.88 ± 1.11
0.05	$\mathbf{28.50 \pm 0.87}$	24.43 ± 0.36	$\mathbf{26.46 \pm 0.61}$	$\mathbf{15.13 \pm 0.44}$	12.49 ± 1.12	19.54 ± 0.48	$\mathbf{26.90 \pm 0.61}$	18.51 ± 0.32
0.20	29.61 ± 1.19	$\mathbf{23.84 \pm 0.92}$	26.72 ± 1.04	15.66 ± 2.21	9.55 ± 0.33	15.48 ± 0.49	27.88 ± 0.52	$\mathbf{17.14 \pm 0.47}$
0.30	32.26 ± 0.29	25.97 ± 0.36	29.11 ± 0.26	16.67 ± 0.89	$\mathbf{9.30 \pm 1.11}$	$\mathbf{15.20 \pm 0.45}$	29.27 ± 0.23	17.61 ± 0.39

Table 5: Comparison (FPR95↓) on CIFAR-10 benchmark. All values are percentages.

Method	CIFAR-100	TIN	Near-Avg	MNIST	SVHN	Texture	Places365	Far-Avg
MSP	49.59±2.94	36.11±0.54	42.85±1.68	25.21±7.04	17.34±1.11	24.89±0.27	36.02±0.85	25.87±1.76
MSP+EMA	49.22±1.88	35.88±0.74	42.55±1.15	25.04±5.12	17.08±1.09	24.28±0.31	34.91±0.33	25.32±1.26
MSP+Rewind	31.46±0.84	27.33±0.49	29.40±0.60	18.55±1.77	14.95±0.57	20.67±0.16	29.64±0.69	20.95±0.54
MSP+ReDiv	28.50±0.71	24.43±0.29	26.46±0.50	15.13±0.36	12.49±0.92	19.54±0.39	26.90±0.49	18.51±0.26
Energy	58.12±2.75	42.23±1.28	50.17±1.85	28.24±10.09	17.08±1.60	29.67±1.29	41.71±2.03	29.18±2.40
Energy+EMA	58.10±1.75	42.57±1.81	50.34±1.67	28.34±7.63	16.92±1.64	28.69±1.45	40.51±0.62	28.61±1.72
Energy+Rewind	32.58±0.34	25.33±0.60	28.95±0.46	11.51±1.50	13.50±0.26	18.83±0.34	26.70±0.87	17.64±0.59
Energy+ReDiv	27.50±0.51	20.69±0.73	24.10±0.62	7.36±0.47	10.19±1.59	16.94±0.51	22.17±0.60	14.17±0.51
ReAct	59.76±2.72	46.45±4.20	53.11±3.46	34.87±17.37	19.56±1.29	33.87±6.12	37.16±4.23	31.37±5.25
ReAct+EMA	60.14±2.48	46.62±3.73	53.38±3.08	35.78±14.64	19.57±1.82	33.04±5.46	37.07±4.28	31.37±4.65
ReAct+Rewind	33.51±1.82	26.41±1.38	29.96±1.59	12.83±1.76	14.94±0.95	19.52±1.05	25.07±1.72	18.09±0.75
ReAct+ReDiv	27.82±0.49	21.21±0.59	24.52±0.54	7.66±0.58	10.67±1.68	17.31±0.81	22.17±0.75	14.45±0.43
Nci	61.29±2.65	48.89±2.95	55.09±2.78	47.27±19.74	19.48±0.32	24.94±3.36	38.22±0.85	32.48±6.02
Nci+EMA	60.91±2.02	48.67±2.82	54.79±2.40	48.34±18.76	19.52±0.32	24.49±3.64	38.06±0.81	32.60±5.88
Nci+Rewind	43.49±5.49	40.34±4.84	41.92±5.16	23.12±5.32	17.27±2.10	25.51±4.50	34.72±3.21	25.16±2.05
Nci+ReDiv	29.33±0.90	25.80±0.73	27.56±0.69	12.98±0.60	12.36±0.22	19.03±1.25	28.20±3.10	18.14±0.40

B.3 Detailed Results on CIFAR-10

Tables 5 and 6 provide per-dataset FPR95 and AUROC results on CIFAR-10, supplementing the aggregate results in the main text. ReDiv consistently improves upon the corresponding Plain, EMA, and Rewind variants across all evaluated detection scores and OOD datasets.

B.4 Detailed Results on CIFAR-100

Table 7 provides per-dataset FPR95 results on CIFAR-100, supplementing the aggregate results in the main text. ReDiv generally improves detection performance, although the gains vary across detection scores and OOD datasets.

B.5 Detailed Results on ImageNet-50

Table 8 provides per-dataset FPR95 results on ImageNet-50, supplementing the aggregate results in the main text. ReDiv achieves the lowest Near-Avg and Far-Avg FPR95 within each detection score, with improvements on most individual OOD datasets.

B.6 Detailed Results on ImageNet-200

Table 9 provides per-dataset FPR95 results on ImageNet-200, supplementing the aggregate results in the main text. ReDiv consistently reduces FPR95 relative to the corresponding baseline, EMA, and Rewind variants across all evaluated detection scores and OOD datasets.

AUROC comparison of ablation analysis. We complement the ablation in Figure 6 with the corresponding AUROC comparison on CIFAR-10.

Averaged over four scores and two OOD regimes, fixed LR and SAM improve AUROC over standard EMA by 2.19% and 2.31%, respectively, while ReDiv improves it by 3.59%. ReDiv further exceeds the fixed-LR-only and SAM-only variants by 1.40% and 1.29% on average and achieves the highest AUROC in all eight comparisons. Relative to the stronger single-component variant in each comparison, ReDiv improves AUROC by 1.35% on near-OOD data and 1.00% on far-OOD data on average. The largest gains again occur with NCI: ReDiv improves near-OOD AUROC by 1.66%

Table 6: Comparison (AUROC↑) on CIFAR-10 benchmark. All values are percentages.

Method	CIFAR-100	TIN	Near-Avg	MNIST	SVHN	Texture	Places365	Far-Avg
MSP	88.27±0.28	90.31±0.16	89.29±0.22	92.41±1.43	93.86±0.50	91.84±0.12	90.21±0.13	92.08±0.29
MSP+EMA	88.33±0.21	90.36±0.14	89.35±0.17	92.34±1.15	93.91±0.48	92.02±0.15	90.35±0.07	92.15±0.25
MSP+Rewind	90.16±0.14	91.62±0.21	90.89±0.16	94.37±0.42	95.24±0.16	93.51±0.12	91.03±0.25	93.53±0.05
MSP+ReDiv	91.23±0.10	92.74±0.11	91.98±0.11	95.61±0.23	95.95±0.18	94.04±0.11	92.12±0.14	94.43±0.08
Energy	88.25±0.30	90.74±0.18	89.50±0.24	93.25±1.61	94.81±0.55	92.09±0.25	90.79±0.16	92.73±0.28
Energy+EMA	88.33±0.19	90.82±0.12	89.58±0.15	93.14±1.36	94.87±0.53	92.29±0.31	90.99±0.05	92.82±0.29
Energy+Rewind	91.45±0.15	93.61±0.15	92.53±0.12	97.33±0.37	96.86±0.21	95.10±0.05	93.53±0.20	95.71±0.10
Energy+ReDiv	92.71±0.15	94.81±0.15	93.76±0.15	98.32±0.13	97.45±0.51	95.58±0.21	94.78±0.15	96.53±0.17
ReAct	87.86±0.70	90.33±0.49	89.09±0.59	92.34±2.36	94.13±0.25	91.72±0.44	91.41±0.65	92.40±0.55
ReAct+EMA	87.90±0.63	90.37±0.47	89.13±0.55	92.24±2.00	94.19±0.35	91.91±0.39	91.50±0.70	92.46±0.51
ReAct+Rewind	91.25±0.35	93.31±0.33	92.28±0.32	97.09±0.38	96.55±0.08	94.96±0.30	93.79±0.27	95.60±0.12
ReAct+ReDiv	92.70±0.15	94.72±0.15	93.71±0.15	98.27±0.13	97.36±0.51	95.60±0.24	94.80±0.19	96.51±0.15
Nci	87.37±0.36	89.60±0.20	88.48±0.28	90.07±2.30	93.26±0.20	92.53±0.53	90.61±0.06	91.62±0.68
Nci+EMA	87.38±0.32	89.60±0.22	88.49±0.27	89.98±2.11	93.29±0.11	92.58±0.52	90.67±0.09	91.63±0.63
Nci+Rewind	89.18±0.75	90.48±0.53	89.83±0.63	94.65±0.92	95.59±0.33	93.57±0.64	91.12±0.67	93.73±0.25
Nci+ReDiv	91.46±0.16	92.91±0.16	92.19±0.14	97.04±0.18	96.81±0.07	94.91±0.39	92.25±0.71	95.25±0.06

Table 7: Comparison (FPR95↓) on CIFAR-100 benchmark. All values are percentages.

Method	CIFAR-10	TIN	Near-Avg	MNIST	SVHN	Texture	Places365	Far-Avg
MSP	58.53±0.52	47.49±0.77	53.01±0.62	49.06±2.16	54.69±6.00	62.37±2.16	55.75±1.03	55.47±2.38
MSP+EMA	57.43±0.82	46.98±1.30	52.20±1.06	49.25±1.29	53.85±4.27	61.57±1.34	54.43±0.21	54.77±1.08
MSP+Rewind	55.49±1.49	46.43±0.43	50.96±0.88	47.72±1.69	61.36±5.25	60.20±1.69	52.68±0.31	55.49±1.30
MSP+ReDiv	54.36±0.90	44.85±0.19	49.60±0.54	45.00±2.24	58.25±5.93	57.57±1.83	50.83±0.11	52.91±0.79
Energy	60.13±0.94	47.56±0.68	53.84±0.21	46.68±2.41	53.31±7.90	62.29±2.35	57.21±1.13	54.87±2.69
Energy+EMA	58.98±0.47	47.12±1.03	53.05±0.72	46.95±0.98	52.44±5.67	61.60±1.90	55.77±0.36	54.19±1.26
Energy+Rewind	57.54±1.46	45.57±0.53	51.55±0.53	35.49±2.92	50.73±8.99	55.90±5.60	54.50±1.46	49.16±4.01
Energy+ReDiv	55.27±0.88	42.79±0.35	49.03±0.49	30.41±1.62	47.55±10.09	52.07±3.11	52.19±0.67	45.55±2.58
ReAct	60.97±4.05	48.14±0.77	54.56±2.41	52.39±7.63	44.35±7.22	51.14±1.75	58.83±2.02	51.68±3.78
ReAct+EMA	59.99±3.89	47.61±0.99	53.80±2.43	52.30±6.69	44.16±6.24	50.81±2.31	57.58±1.02	51.21±2.84
ReAct+Rewind	61.34±0.68	47.00±0.77	54.17±0.17	41.42±2.91	34.06±10.33	44.89±5.43	54.03±0.70	43.60±4.22
ReAct+ReDiv	58.02±1.87	43.87±0.42	50.94±0.75	34.61±0.88	36.07±11.07	42.91±4.83	51.06±1.53	41.16±3.42
Nci	53.60±0.72	46.35±0.60	49.98±0.08	49.88±2.10	56.61±2.93	53.55±2.11	53.79±1.50	53.46±0.93
Nci+EMA	53.28±0.78	45.81±0.64	49.54±0.34	49.94±1.93	56.87±3.87	53.15±1.57	53.30±0.88	53.31±0.47
Nci+Rewind	57.03±1.62	43.02±0.22	50.03±0.75	44.12±2.05	54.97±5.38	46.29±2.33	50.16±0.33	48.88±1.79
Nci+ReDiv	58.06±1.67	42.00±0.50	50.03±1.03	41.67±0.93	53.29±4.48	45.90±2.06	49.85±0.29	47.68±1.18

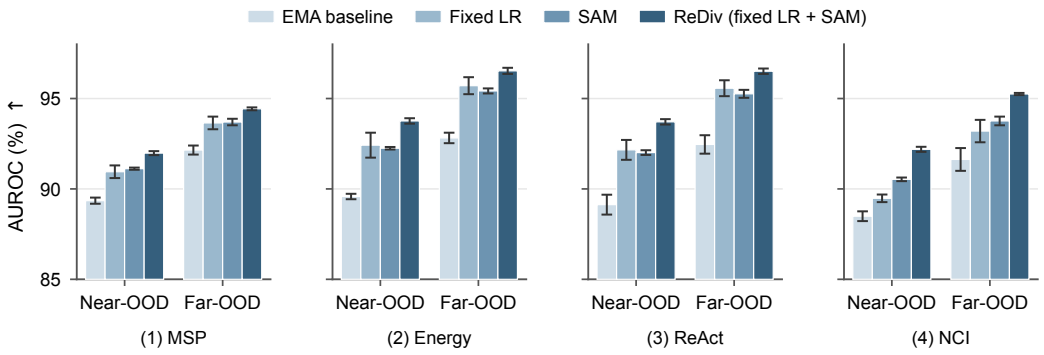


Figure 6: AUROC ablation of fixed learning rate and SAM on CIFAR-10. EMA denotes the standard-training baseline; the other variants continue from the same rewound state. Fixed LR and SAM activate one diversification component at a time, whereas ReDiv combines both. Bars show means and error bars denote one standard deviation. The y-axis begins at 85 to resolve the differences between configurations. Higher AUROC is better.

(92.19 vs. 90.53) and far-OOD AUROC by 1.49% (95.25 vs. 93.76). Together with the FPR95 results, this consistent ranking shows that fixed LR and SAM provide complementary gains.

Table 8: Comparison (FPR95↓) on ImageNet-50 benchmark. All values are percentages.

Method	Ssb_hard	Ninco	Near-Avg	iNaturalist	Texture	Openimage-o	Far-Avg
MSP	36.00±0.51	32.30±0.87	34.15±0.30	21.78±0.74	25.04±0.86	25.50±1.04	24.10±0.69
MSP+EMA	35.75±0.34	31.78±0.92	33.76±0.48	21.50±0.60	25.20±0.16	25.29±0.89	24.00±0.50
MSP+Rewind	34.30±0.26	29.84±0.77	32.07±0.46	20.30±0.32	23.94±0.98	24.64±1.26	22.96±0.68
MSP+ReDiv	32.50±0.42	27.84±0.13	30.17±0.27	17.64±0.34	19.38±0.50	22.49±0.55	19.84±0.24
Energy	37.81±0.60	34.44±0.16	36.13±0.33	19.67±0.79	24.90±0.90	24.80±1.16	23.13±0.92
Energy+EMA	37.48±0.92	34.00±0.31	35.74±0.61	19.36±1.06	24.73±0.56	24.58±0.98	22.89±0.76
Energy+Rewind	33.02±1.84	28.65±2.53	30.84±2.18	14.10±2.19	17.99±2.21	18.76±1.55	16.95±1.88
Energy+ReDiv	30.68±0.57	25.54±0.75	28.11±0.35	11.78±1.20	12.76±1.12	16.68±0.76	13.74±0.85
ReAct	38.10±1.12	33.35±1.61	35.73±1.33	16.27±1.47	20.12±1.69	20.59±1.84	18.99±1.65
ReAct+EMA	37.73±1.30	33.04±1.59	35.39±1.43	15.76±1.55	20.01±1.68	20.34±2.03	18.71±1.73
ReAct+Rewind	34.52±2.71	28.50±3.19	31.51±2.94	11.81±0.41	15.47±2.21	16.95±0.89	14.74±1.11
ReAct+ReDiv	30.55±1.02	24.16±1.73	27.36±1.17	10.41±0.99	12.15±1.40	15.27±1.07	12.61±1.11
NCI	33.04±0.44	27.45±1.07	30.24±0.73	16.19±0.42	13.08±0.40	18.33±1.06	15.87±0.53
NCI+EMA	32.96±0.43	27.48±0.69	30.22±0.56	16.15±0.48	13.04±0.33	18.18±1.11	15.79±0.54
NCI+Rewind	37.66±1.97	29.81±3.55	33.73±2.76	12.18±1.58	12.13±2.31	18.27±2.12	14.19±1.96
NCI+ReDiv	33.05±0.14	24.33±1.16	28.69±0.63	9.99±0.60	8.16±0.45	15.53±0.29	11.22±0.30

Table 9: Comparison (FPR95↓) on ImageNet-200 benchmark. All values are percentages.

Method	Ssb_hard	Ninco	Near-Avg	iNaturalist	Texture	Openimage-o	Far-Avg
MSP	65.85±0.96	43.29±0.35	54.57±0.31	25.37±0.79	44.01±1.40	34.72±0.29	34.70±0.78
MSP+EMA	65.73±0.95	43.40±0.23	54.57±0.46	25.09±0.49	43.26±0.90	34.43±0.27	34.26±0.40
MSP+Rewind	65.19±0.80	42.27±0.76	53.73±0.76	24.26±1.16	42.34±0.86	33.90±0.40	33.50±0.21
MSP+ReDiv	64.27±0.71	40.52±0.70	52.40±0.62	22.16±0.46	38.15±0.83	31.68±0.12	30.66±0.42
Energy	69.31±0.96	50.96±1.73	60.14±1.34	27.11±0.36	40.84±0.88	37.23±0.77	35.06±0.37
Energy+EMA	69.31±0.96	50.96±1.73	60.14±1.34	27.11±0.36	40.84±0.88	37.23±0.77	35.06±0.37
Energy+Rewind	68.80±1.01	49.85±1.80	59.32±1.40	25.41±2.28	37.69±1.05	35.28±1.10	32.79±1.43
Energy+ReDiv	66.81±0.82	46.63±1.60	56.72±1.14	22.21±0.19	29.89±1.00	32.14±0.19	28.08±0.30
ReAct	69.86±0.04	51.79±1.56	60.82±0.76	23.23±1.18	31.13±0.59	32.59±0.28	28.98±0.53
ReAct+EMA	69.65±0.29	51.59±1.61	60.62±0.67	23.03±1.35	30.84±0.44	32.37±0.34	28.75±0.50
ReAct+Rewind	69.61±0.33	51.33±1.95	60.47±0.81	22.01±1.88	28.71±1.13	31.36±1.18	27.36±1.37
ReAct+ReDiv	68.28±0.55	47.55±2.01	57.91±0.73	18.90±1.09	22.74±0.61	28.32±0.24	23.32±0.54
NCI	69.26±0.78	42.11±0.98	55.69±0.20	17.43±0.40	27.87±0.88	26.99±0.30	24.10±0.28
NCI+EMA	69.35±0.73	41.97±1.00	55.66±0.37	17.36±0.20	27.73±1.05	26.91±0.48	24.00±0.30
NCI+Rewind	68.72±0.33	41.17±1.03	54.94±0.48	16.69±0.92	26.76±0.31	26.77±0.68	23.41±0.56
NCI+ReDiv	67.29±0.69	39.19±0.69	53.24±0.01	15.37±0.46	24.42±1.01	25.16±0.15	21.65±0.36

Does margin locate EMA saturation? We compare margin with five alternative statistics computed on the ID validation set: MSP, predictive entropy, logit norm, mean loss, and maximum logit. Each statistic uses the same three-checkpoint trigger rule, while the post-rewind optimization and evaluation protocol remain fixed. No OOD data are used to select the rewind point. Table 10 reports the resulting CIFAR-10 performance. Margin achieves the lowest FPR95 under both OOD regimes, reaching 26.46% on near-OOD data and 18.51% on far-OOD data. We therefore use margin as the default trigger in ReDiv.

Table 10: FPR95↓ of signals used to trigger rewinding on CIFAR-10.

Trigger signal	Near-Avg	Far-Avg
MSP	27.14 ± 1.19	19.05 ± 0.91
Entropy	27.28 ± 0.34	19.61 ± 0.33
LogitNorm	28.47 ± 0.68	20.21 ± 0.53
Loss	26.70 ± 0.64	18.76 ± 0.79
MaxLogit	27.56 ± 0.14	19.26 ± 0.35
Margin	$\mathbf{26.46 \pm 0.50}$	$\mathbf{18.51 \pm 0.26}$

C A Quality–Diversity View of EMA

To explain why EMA may degrade even when later plain checkpoints remain strong, we analyze the parameter-averaged EMA model. Let

$$\mathcal{R}_{\mathbb{P}}(\theta) = \mathbb{E}_{(x,b)\sim\mathbb{P}}\left[\ell_{\mathrm{det}}\big(S(x;\theta), b\big)\right] \tag{19}$$

denote the population surrogate risk under an unknown test distribution $\mathbb{P}$.

Proposition C.1 (Quality–diversity recursion)**.** *The EMA update satisfies*

$$\mathcal{R}_{\mathbb{P}}(\bar{\theta}_t) = \alpha\mathcal{R}_{\mathbb{P}}(\bar{\theta}_{t-1}) + (1-\alpha)\mathcal{R}_{\mathbb{P}}(\theta_t) - \mathcal{J}_t, \tag{20}$$

where $\mathcal{J}_t$ denotes the diversity credit contributed by the incoming plain checkpoint. Consequently, the EMA update improves the surrogate risk if and only if

$$\mathcal{R}_{\mathbb{P}}(\theta_t) < \mathcal{R}_{\mathbb{P}}(\bar{\theta}_{t-1}) + \frac{\mathcal{J}_t}{1-\alpha}. \tag{21}$$

Equation 20 shows that the benefit of EMA depends jointly on checkpoint quality and useful variation relative to the current EMA model. Thus, a later plain checkpoint may improve over earlier checkpoints yet still fail to improve EMA when it contributes insufficient diversity credit.

Under a local quadratic approximation, the diversity credit takes the form

$$\mathcal{J}_t = \frac{\alpha(1-\alpha)}{2}\left\|\theta_t - \bar{\theta}_{t-1}\right\|_{H_t}^2, \tag{22}$$

where $\|v\|_{H_t}^2 = v^\top H_t v$. This shows that the relevant diversity is curvature-weighted rather than simple Euclidean separation.

ReDiv is therefore designed to preserve useful post-rewind parameter variation while maintaining checkpoint quality.

D Additional Analysis of Post-Rewind Diversity

D.1 Plain-to-EMA Diversity Dynamics

Recall

$$d_t = \theta_t - \bar{\theta}_{t-1}. \tag{23}$$

The plain model follows

$$\theta_{t+1} = \theta_t - \eta_t u_t, \tag{24}$$

while EMA is updated by

$$\bar{\theta}_t = \alpha\bar{\theta}_{t-1} + (1-\alpha)\theta_t. \tag{25}$$

Proposition D.1 (Plain-to-EMA diversity dynamics)**.** *The plain-to-EMA separation satisfies*

$$d_{t+1} = \alpha d_t - \eta_t u_t. \tag{26}$$

Proof. By definition,

$$\begin{aligned} d_{t+1} &= \theta_{t+1} - \bar{\theta}_t && (27)\\ &= \theta_t - \eta_t u_t - \left[\alpha\bar{\theta}_{t-1} + (1-\alpha)\theta_t\right] && (28)\\ &= \alpha\left(\theta_t - \bar{\theta}_{t-1}\right) - \eta_t u_t && (29)\\ &= \alpha d_t - \eta_t u_t. && (30) \end{aligned}$$

□

D.2 Effect of the Learning-Rate Scale

For a fixed positive-semidefinite matrix H, define

$$F_t(\eta) = \|\alpha d_t - \eta u_t\|_H^2, \qquad \|v\|_H^2 = v^\top H v. \tag{31}$$

For two learning rates $\eta_a > \eta_b \geq 0$,

$$F_t(\eta_a) - F_t(\eta_b) = (\eta_a - \eta_b)\left[(\eta_a + \eta_b)\|u_t\|_H^2 - 2\alpha\langle d_t, u_t\rangle_H\right], \tag{32}$$

where

$$\langle a, b\rangle_H = a^\top H b. \tag{33}$$

Therefore, whenever

$$(\eta_a + \eta_b)\|u_t\|_H^2 > 2\alpha\langle d_t, u_t\rangle_H, \tag{34}$$

a smaller learning rate produces a smaller curvature-weighted separation. This formalizes how continued learning-rate decay can further contract the post-rewind trajectory. Fixing the learning rate does not guarantee increased diversity at every step; it removes this additional scheduler-induced source of shrinkage.

D.3 How SAM Modifies the Effective Update

For a mini-batch loss $\mathcal{L}_{\mathcal{B}_t}$, let

$$g_t = \nabla_\theta \mathcal{L}_{\mathcal{B}_t}(\theta_t), \qquad \hat{g}_t = \frac{g_t}{\|g_t\|_2}. \tag{35}$$

SAM evaluates the gradient at locally perturbed parameters

$$\theta_t + \epsilon_t, \qquad \epsilon_t = \rho\hat{g}_t, \tag{36}$$

which gives

$$\tilde{g}_t = \nabla_\theta \mathcal{L}_{\mathcal{B}_t}(\theta_t + \epsilon_t). \tag{37}$$

Let

$$H_t^{\mathcal{L}} = \nabla_\theta^2 \mathcal{L}_{\mathcal{B}_t}(\theta_t). \tag{38}$$

Assuming that the Hessian is locally Lipschitz,

$$\tilde{g}_t = g_t + \rho H_t^{\mathcal{L}}\hat{g}_t + \mathcal{O}(\rho^2). \tag{39}$$

Define

$$P_t^\perp = I - \hat{g}_t\hat{g}_t^\top. \tag{40}$$

Then

$$\rho H_t^{\mathcal{L}}\hat{g}_t = \underbrace{\rho(\hat{g}_t^\top H_t^{\mathcal{L}}\hat{g}_t)\hat{g}_t}_{\text{parallel: magnitude change}} + \underbrace{\rho P_t^\perp H_t^{\mathcal{L}}\hat{g}_t}_{\text{orthogonal: direction change}}. \tag{41}$$

If

$$P_t^\perp H_t^{\mathcal{L}}\hat{g}_t \neq 0, \tag{42}$$

SAM introduces a nonzero first-order component orthogonal to the vanilla-gradient direction. Thus, SAM can modify both the magnitude and direction of the effective update at first order.

D.4 Complementarity

Combining the previous results,

$$d_{t+1} = \alpha d_t - \eta_t u_t \tag{43}$$

shows that the learning rate and effective update affect different parts of the post-rewind diversity dynamics. The fixed learning rate acts on the scale η_t, whereas SAM modifies the magnitude and direction of u_t.

Under the local quadratic approximation introduced in Section 2.3,

$$\mathcal{J}_{t+1} = \frac{\alpha(1-\alpha)}{2} \left\| \alpha d_t - \eta_t u_t \right\|^2_{H^{\mathcal{R}}_{t+1}} . \tag{44}$$

Hence, the two components of ReDiv act on complementary factors governing the evolution of the EMA diversity credit.

These results characterize an update-level mechanism rather than a universal monotonicity guarantee. They do not imply that either component increases $\mathcal{J}_t$ at every iteration, nor do they directly guarantee improvements in AUROC or FPR95.