Machine Unlearning via Adaptive Gradient Reweighting and Multi-stage Objective Optimization

Supplementary Material

In this Appendix, We first provide the theoretical analysis and derivation of our update rule and show how it outperforms fixed-weight gradient combinations in Section A. Then, in Section B, We present comprehensive implementation details, including the experimental setup, baseline methods, and hyperparameter configurations to ensure full reproducibility. In Section C, we present additional experimental results, which include unlearning 10% on TinyImageNet and unlearning on CelebA-Top500 to further verify that our method performs reliably across different datasets and unlearning scenarios. Additionally, we provide an ablation study on hyper-parameter sensitivity to validate the robustness of our framework, as well as a quantitative comparison of execution time to highlight our method's computational efficiency. Finally, in Section D, we demonstrate how our framework can be seamlessly extended to generative models such as Diffusion Models.

A. Proof of the Superiority of Our Method

In this section, we theoretically demonstrate the superiority of our proposed method by systematically analyzing the trade-off between retention and forgetting objectives in machine unlearning. We first formalize the gradient conflict scenario using a 2D subspace framework, then introduce our Gradient Direction Rectification (GDR) mechanism to mitigate destructive interference between retention and forgetting gradients. We proceed to establish the pointwise dominance of symmetric GDR over fixed-weight baselines via analytical performance bounds, and further show that our adaptive multi-stage optimization strategy yields an optimal solution on the efficient trade-off front.

A.1. Preliminaries and Baseline

We analyze the unlearning update within the 2D subspace spanned by the retention and forgetting gradients. Let $\mathbf{g}_r, \mathbf{g}_f \in \mathbb{R}^d$ denote the unit-norm gradient directions corresponding to the retention and forgetting objectives, respectively. We assume a gradient conflict condition, where their cosine similarity is given by $c := \mathbf{g}_r^\top \mathbf{g}_f \in (0, 1)$.

We construct an orthonormal basis $\{\hat{\mathbf{u}}_r, \hat{\mathbf{u}}_o\}$ for this subspace, where $\hat{\mathbf{u}}_r := \mathbf{g}_r$ represents the retention axis, and $\hat{\mathbf{u}}_o$ denotes the orthogonal axis obtained via Gram–Schmidt orthogonalization of $\mathbf{g}_f$ with respect to $\mathbf{g}_r$.

To evaluate update effectiveness, we define two signed projection scores:

$$S_r(\mathbf{v}) := \mathbf{v}^\top \hat{\mathbf{u}}_r, \qquad S_f^\perp(\mathbf{v}) := -\mathbf{v}^\top \hat{\mathbf{u}}_o.$$

A higher S_r indicates stronger alignment with the retention direction (i.e., better preservation), while a higher $S_f^\perp$ indicates more effective progression along the forgetting direction orthogonal to $\mathbf{g}_r$ (i.e., less interference with retention).

As a baseline, we consider a fixed-weight combination of the two gradients:

$$\mathbf{g}_{\text{base}}(\lambda) := (1-\lambda)\mathbf{g}_r - \lambda \mathbf{g}_f,$$

where $\lambda \in [0, 1]$ is a manually tuned trade-off coefficient. The corresponding projection scores are given by:

$$S_r(\mathbf{g}_{\text{base}}) = (1-\lambda) - \lambda c, \qquad S_f^\perp(\mathbf{g}_{\text{base}}) = \lambda\sqrt{1-c^2}.$$

A.2. Gradient Direction Rectification

Our method adopts the *Gradient Direction Rectification (GDR)* mechanism. Its purpose is to directly mitigate the destructive interference between the retention and forgetting objectives by explicitly removing their opposing directional components, thereby enforcing mutual gradient orthogonality. In this analysis, we focus on its symmetric formulation, where both gradients are rectified in a mutually consistent manner.

Given positive rectification coefficients α and β, the rectified gradients are defined as follows. The rectified forgetting gradient $\mathbf{g}_f'$ removes the projection of $\mathbf{g}_f$ onto the retention direction $\mathbf{g}_r$:

$$\mathbf{g}_f' := \mathbf{g}_f - \alpha(\mathbf{g}_f \cdot \mathbf{g}_r)\mathbf{g}_r.$$

Similarly, the rectified retention gradient $\mathbf{g}_r'$ removes the projection of $\mathbf{g}_r$ onto the forgetting direction $\mathbf{g}_f$:

$$\mathbf{g}_r' := \mathbf{g}_r - \beta(\mathbf{g}_r \cdot \mathbf{g}_f)\mathbf{g}_f.$$

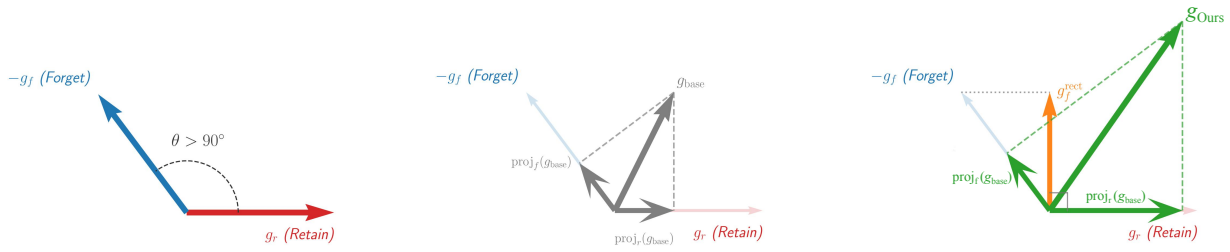
The final update direction under this symmetric rectification scheme is given by a linear combination of these two improved gradients:

$$\mathbf{g}_{\text{sym}}(\lambda) := (1-\lambda)\mathbf{g}_r' - \lambda \mathbf{g}_f'.$$

This formulation integrates GDR while retaining a fixed mixing coefficient λ, serving as a bridge between the baseline and our multi-stage objective optimization update to be introduced next.

A.3. Pointwise Dominance of Symmetric GDR

Having defined both the baseline update $\mathbf{g}_{\text{base}}(\lambda)$ and the GDR-based update $\mathbf{g}_{\text{sym}}(\lambda)$, we now proceed to quantitatively establish the superiority of the latter by showing that,



(a) The Problem: Gradient Conflict **(b) Baseline: Inefficient Compromise** **(c) Our Mtehod: Rectify and Optimize**

Figure 5. **Geometric illustration of gradient conflict resolution. (a)** Conflicting gradients ($\theta > 90°$); **(b)** Baseline: limited contributions (highlighted in gray); **(c)** Our Method: maximal contributions via rectification and optimization (highlighted in green).

under mild conditions, $\mathbf{g}_{\text{sym}}(\lambda)$ achieves a strictly better trade-off between the retention and forgetting objectives.

First, we compute the performance scores for the symmetric GDR update $\mathbf{g}_{\text{sym}}(\lambda)$. By substituting the definitions of the rectified gradients $(\mathbf{g}'_r, \mathbf{g}'_f)$ into the score functions $(S_r, S_f^{\perp})$, we obtain:

$$S_r(\mathbf{g}_{\text{sym}}) = (1-\lambda)(1-\beta c^2) - \lambda(1-\alpha)c, \tag{14}$$

$$S_f^{\perp}(\mathbf{g}_{\text{sym}}) = \sqrt{1-c^2}\,\big(\lambda + \beta c(1-\lambda)\big). \tag{15}$$

By direct algebraic manipulation, we obtain the following expressions for the performance improvements:

$$\Delta S_f^{\perp} = \sqrt{1-c^2}\,\beta c(1-\lambda), \tag{16}$$

$$\Delta S_r = c\big(\lambda\alpha - (1-\lambda)\beta c\big). \tag{17}$$

From Eq. (16), given that $\beta > 0$, $c \in (0,1)$, and $\lambda \in [0,1]$, it follows that $\Delta S_f^{\perp} \geq 0$. This is a crucial observation: symmetric GDR never deteriorates the harmless forgetting objective.

For the retention score, the improvement condition $\Delta S_r \geq 0$ holds if and only if $\lambda\alpha \geq (1-\lambda)\beta c$. Solving this inequality for λ gives the lower bound:

$$\lambda \geq \frac{\beta c}{\alpha + \beta c}.$$

We denote this critical threshold as $\lambda^\star := \frac{\beta c}{\alpha+\beta c}$. Let $\mathbf{S}(\mathbf{v}) := \big(S_r(\mathbf{v}), S_f^{\perp}(\mathbf{v})\big)$ represent the performance score vector of an update direction $\mathbf{v}$. These results directly lead to the following theorem, which formalizes the pointwise dominance of symmetric GDR over the baseline update.

Theorem 1 (Conditional Dominance of Symmetric GDR)**.** *Under the conflict setting defined above, for any mixing weight $\lambda \in [\lambda^\star, 1]$, the symmetric GDR update $\mathbf{g}_{sym}(\lambda)$ dominates the fixed-weight baseline $\mathbf{g}_{base}(\lambda)$ in the joint objective space:*

$$\mathbf{S}(\mathbf{g}_{sym}(\lambda)) \succeq \mathbf{S}(\mathbf{g}_{base}(\lambda)),$$

where $\succeq$ denotes component-wise dominance in the score space. The dominance becomes strict ($\succ$) if $\lambda > \lambda^\star$ (yielding a strictly higher S_r), or if $\beta > 0$ and $\lambda < 1$ (yielding a strictly higher $S_f^{\perp}$).

A.4. Multi-stage Objective Optimization

The previous section established that the GDR mechanism enlarges the feasible set of update directions. Beyond this, our method further adaptively identifies the optimal trade-off between retention and forgetting at each unlearning step, without relying on a fixed, manually-tuned mixing coefficient λ.

This adaptive selection can be naturally formulated as a multi-objective optimization problem, which we address via a standard *scalarization* technique. We introduce a scalarization function $\Phi : \mathbb{R}^2 \to \mathbb{R}$ that maps the two-dimensional performance vector $\mathbf{S}(\mathbf{v}) = (S_r(\mathbf{v}), S_f^{\perp}(\mathbf{v}))$ to a single real-valued score. A crucial requirement for Φ is to be *strictly monotonic* in each coordinate, ensuring that improvements in either objective strictly increase the scalarized score. A simple yet effective choice is a weighted linear sum, $\Phi(s_1, s_2) = w_1 s_1 + w_2 s_2$ with $w_1, w_2 > 0$.

At each step t, our method determines the optimal mixing weight λ_t by maximizing this scalarized objective over $\lambda \in [0,1]$:

$$\lambda_t = \arg\max_{\lambda \in [0,1]} \Phi(\mathbf{S}(\mathbf{g}_{\text{sym}}(\lambda))).$$

The final update direction of our method is then given by $\mathbf{g}_{\text{Ours}} = \mathbf{g}_{\text{sym}}(\lambda_t)$. This construction directly leads to two desirable theoretical properties. First, by standard results in multi-objective optimization, maximizing any strictly monotonic scalarization function yields an efficient solution. Consequently, $\mathbf{g}_{\text{Ours}}$ always resides on the efficient front of the trade-off curve $\mathbf{g}_{\text{sym}}(\lambda)$. Second, by definition of optimality, the scalarized performance of $\mathbf{g}_{\text{Ours}}$ is guaranteed to be no worse than that of any fixed-weight update on the curve:

$$\Phi(\mathbf{S}(\mathbf{g}_{\text{Ours}})) \geq \Phi(\mathbf{S}(\mathbf{g}_{\text{sym}}(\lambda_0))), \quad \forall \lambda_0 \in [0,1].$$

This inequality completes the argument, establishing a formal connection between the adaptively weighted update in our method and the entire family of fixed-weight GDR variants.

A.5. Main Result and Conclusion

We now consolidate the results from the preceding sections to formally establish the dominance of our method over the fixed-weight baseline.

The proof proceeds in two steps. First, by Theorem 1 in Section A.3, for any baseline policy with a fixed weight λ_0 satisfying the dominance condition ($\lambda_0 \geq \lambda^\star$), the GDR mechanism achieves an improvement over the baseline:

$$\mathbf{S}(\mathbf{g}_{\text{sym}}(\lambda_0)) \succeq \mathbf{S}(\mathbf{g}_{\text{base}}(\lambda_0)).$$

Because Φ is strictly increasing in each coordinate, this dominance directly translates to the scalarized objective:

$$\Phi(\mathbf{S}(\mathbf{g}_{\text{sym}}(\lambda_0))) \geq \Phi(\mathbf{S}(\mathbf{g}_{\text{base}}(\lambda_0))).$$

Second, as established in Section A.4, the adaptive weight selection in our method guaranties scalarized optimality over the entire family of fixed-weight GDR updates.

$$\Phi(\mathbf{S}(\mathbf{g}_{\text{Ours}})) \geq \Phi(\mathbf{S}(\mathbf{g}_{\text{sym}}(\lambda_0))), \quad \forall \lambda_0 \in [0, 1].$$

Combining the two dominance relations yields our main result. For any strictly increasing scalarization function Φ and any baseline weight $\lambda_0 \geq \lambda^\star$, the following hierarchy holds:

$$\underbrace{\Phi(\mathbf{S}(\mathbf{g}_{\text{Ours}}))}_{\text{Our Method}} \geq \underbrace{\Phi(\mathbf{S}(\mathbf{g}_{\text{sym}}(\lambda_0)))}_{\text{Symmetric GDR}} \geq \underbrace{\Phi(\mathbf{S}(\mathbf{g}_{\text{base}}(\lambda_0)))}_{\text{Fixed-Weight Baseline}}.$$

This chain of inequalities formally establishes that our method consistently attains a solution that is never worse and typically superior to any fixed-weight policy.

Remark on One-sided GDR. The simpler one-sided variant of GDR, in which only the forgetting gradient $\mathbf{g}_f$ is rectified, corresponds to the special case $\beta = 0$ in our framework. In this case, the dominance threshold reduces to $\lambda^\star = 0$, implying that one-sided GDR yields a strict improvement over the baseline for all $\lambda \in (0, 1]$ (see Fig. 5).

B. Experimental Details

B.1. Experimental Setup

Our implementation is based on Python 3.8 and PyTorch 2.4. All experiments are conducted on a workstation with an NVIDIA GeForce RTX 4090 GPU and an Intel Xeon Platinum 8383C CPU. We evaluate our method under two standard unlearning scenarios: **(1) Random Sample Unlearning**, where 10% or 50% of the training samples are randomly selected across all classes to form the forget set (D_f); and **(2) Class Unlearning**, where all samples from one or more specific classes constitute D_f. The remaining data form the retain set (D_r).

For CIFAR-10/100, we train ResNet-18 and Vision Transformer (ViT-B/16) from scratch. For larger datasets, including Tiny ImageNet, CelebA-Top500, and VGGFace2 subsets, we use ResNet-50 initialized with ImageNet-1k pre-trained weights and fine-tuned on the corresponding training sets. The retrained model follows the same training procedure but is trained only on the retain set D_r.

B.2. Baseline Details

We evaluate our method against various unlearning methods covering multiple paradigms:

- **Finetune**: The pre-trained model undergoes further training exclusively on the retain set (D_r), with no exposure to forget samples during this phase.
- **NegGrad+** [35]: The model continues to train on the full dataset but reverses the gradient sign for forget samples during backpropagation to mitigate their influence.
- **Random Label** [24]: During continued training, forget samples are assigned random incorrect labels to disrupt the model's memorization of their original patterns.
- **Bad Teacher** [10]: A distillation-based method where the student model aligns with a pre-trained teacher on retain samples while following a randomly initialized teacher (providing uninformative guidance) on forget samples.
- **SCRUB** [35]: A distillation approach that minimizes the KL divergence between the model and the pre-trained teacher on retain samples, while maximizing this divergence on forget samples.
- **SSD** [21]: Identifies parameters most influenced by forget samples via the Fisher Information Matrix, then suppresses these parameters to reduce their impact.
- **UNSIR** [51]: Introduces a noise matrix derived from forget samples, which is designed to maximize the model's prediction error on these samples.
- **SalUn** [19]: A gradient-based approach that distinguishes between weights associated with forget samples (removable) and those critical to retain samples (preserved), followed by targeted unlearning of the former.
- **LoTUS** [50]: Smooths the model's outputs on the forget set up to an information-theoretic bound, aiming to reduce overconfidence arising from data memorization.

For class unlearning, we additionally compare against the baselines below:

- **Boundary Unlearning** [8]: This method directly manipulates the decision boundary, either by shrinking it away from the forget class data or expanding the boundaries of retain classes.
- **DELETE** [61]: The unlearning objective is decoupled into forgetting and retention terms, using masked distillation to simultaneously erase knowledge of the forget class

Table 7. Hyperparameter configurations of baseline methods on the CIFAR-10 random unlearning task (10% forget ratio).

Method	Learning Rate	Weight Decay	Optimizer
Finetune	1×10^{-3}	5×10^{-4}	SGD
NegGrad+	1×10^{-3}	5×10^{-4}	SGD
RndLbl	1×10^{-3}	5×10^{-4}	SGD
BadT	1×10^{-4}	0	Adam
SCRUB	1×10^{-4}	0.1	Adam
SSD	0.1	0	SGD
UNSIR	1×10^{-3}	1×10^{-3}	SGD
SalUn	1×10^{-3}	5×10^{-4}	SGD
LoTUS	1×10^{-5}	5×10^{-4}	AdamW

Table 8. Hyperparameter configurations of baseline methods on the CIFAR-10 class-level unlearning task (1 forget class), where DELETE adopts layered learning rates with 1×10^{-5} for the backbone and 5×10^{-3} for the head.

Method	Learning Rate	Weight Decay	Optimizer
Finetune	5×10^{-2}	5×10^{-4}	SGD
NegGrad+	1×10^{-3}	1×10^{-4}	SGD
RndLbl	5×10^{-2}	5×10^{-4}	SGD
BShrink	2×10^{-4}	0	SGD
BExpand	5×10^{-5}	0	SGD
BadT	1×10^{-4}	1×10^{-4}	Adam
SalUn	1×10^{-3}	5×10^{-4}	SGD
DELETE	5×10^{-3}	5×10^{-4}	SGD

while preserving knowledge for the retain classes.

Finetune, NegGrad+, and Random Label are considered simple yet widely used unlearning baselines, whereas the latter methods represent state-of-the-art approaches with more sophisticated mechanisms. The hyperparameters for the unlearning phase of each method are detailed in Tables 7 and 8.

C. Experimental Results

C.1. The results of unlearning 10% on TinyIN

In the main paper, we primarily present comprehensive experimental results for the 10% forgetting task on the widely used CIFAR-10 and CIFAR-100 datasets, which serve as standard benchmarks for evaluating machine unlearning performance. To further validate the generalizability and robustness of the proposed method across diverse data distributions and more complex image domains, we extend our evaluations to the Tiny ImageNet dataset, which is a larger-scale benchmark with richer semantic categories and finer-grained visual variations. Specifically, Tables 9 and 10 provide detailed performance comparisons across all competing unlearning methods (including Finetune, NegGrad+, SCRUB, and others) under the 10% forgetting setting on Tiny ImageNet. These supplementary results confirm that our method achieves the best retain-forget trade-off despite higher-dimensional inputs and a more extensive fine-grained label space, validating its effectiveness in complex data scenarios.

Table 9. **Performance Summary of unlearning 10% of TinyIN on RN18.** Avg Gap denotes overall performance, with the best results highlighted in **bold**.

Method	$\mathrm{Acc_r}$	$\mathrm{Acc_f}$	$\mathrm{Acc_t}$	$\mathrm{Acc_{MIA}}$	Avg Gap
Gold Std	99.99	62.26	63.36	30.30	–
Finetune	99.99	99.94	65.64	93.31	25.74
NegGrad+	99.75	95.02	63.78	70.91	18.51
RndLbl	72.11	45.80	38.44	21.75	19.45
BadT	92.13	90.80	59.32	62.62	18.19
SCRUB	92.27	90.68	62.68	84.27	22.70
SSD	99.49	99.48	65.52	95.45	26.26
UNSIR	99.39	99.36	58.72	83.96	24.00
SalUn	99.97	99.90	63.20	87.21	23.68
LoTUS	99.44	94.62	61.18	52.55	14.34
Ours	93.62	79.67	51.74	30.38	**8.87**

Table 10. **Performance Summary of unlearning 10% of TinyIN on ViT.** Avg Gap denotes overall performance, with the best results highlighted in **bold**.

Method	Acc_r	Acc_f	Acc_t	$\mathrm{Acc_{MIA}}$	Avg Gap
Gold Std	96.26	89.77	90.04	74.98	–
Finetune	97.78	92.71	89.14	76.86	1.81
NegGrad+	93.35	92.76	88.80	71.28	2.71
RndLbl	91.16	90.34	88.06	69.86	3.19
BadT	86.87	84.57	83.22	67.43	7.24
SCRUB	89.61	89.32	86.06	69.21	4.21
SSD	94.85	94.85	89.74	76.56	2.09
UNSIR	93.51	93.21	88.54	77.10	2.45
SalUn	96.32	82.48	82.02	69.05	5.33
LoTUS	94.79	94.67	89.92	75.32	1.71
Ours	95.03	92.75	89.78	74.57	**1.22**

Table 11. **Performance comparison of unlearning methods on the CelebA-Top500 dataset.** All methods achieve complete forgetting ($\mathrm{Acc_{ft}}$ = 0%). Our method achieves the best trade-off between retention and forgetting, yielding an H-Mean closest to the retraining baseline.

Method	$\mathrm{Acc}_r \uparrow$	$\mathrm{Acc}_{rt} \uparrow$	H-Mean ↑	Gap Score ↓
Retrain	99.99	73.25	84.56	0.00
RndLbl	87.49	56.41	72.13	7.34
NegGrad+	93.72	58.85	74.10	5.17
BShrink	89.72	52.64	68.97	7.72
BExpand	95.93	58.52	73.83	5.74
SCRUB	94.03	60.83	75.65	4.60
BadT	95.22	61.43	76.10	4.15
SalUn	95.02	61.69	76.31	4.13
DELETE	96.50	59.58	74.67	7.42
Ours	**99.89**	**71.00**	**83.04**	**1.63**

Table 12. **Effect of unlearning on similar identities in CelebA-Top500 dataset.** Performance is evaluated on nine similar identities ranked by cosine similarity to ID 107 (the forgotten identity). Our method attains the highest average accuracy on both training and test sets, while ensuring complete forgetting of the target identity.

Method	ID 258 (0.711)		ID 164 (0.563)		ID 452 (0.557)		ID 132 (0.519)		ID 334 (0.508)		ID 229 (0.478)		ID 426 (0.468)		ID 165 (0.451)		ID 436 (0.380)		Average	
	$Acc_r \uparrow$	$Acc_{rt} \uparrow$	$Acc_r \uparrow$	$Acc_{rt} \uparrow$	$Acc_r \uparrow$	$Acc_{rt} \uparrow$	$Acc_r \uparrow$	$Acc_{rt} \uparrow$	$Acc_r \uparrow$	$Acc_{rt} \uparrow$	$Acc_r \uparrow$	$Acc_{rt} \uparrow$	$Acc_r \uparrow$	$Acc_{rt} \uparrow$	$Acc_r \uparrow$	$Acc_{rt} \uparrow$	$Acc_r \uparrow$	$Acc_{rt} \uparrow$	$Acc_r \uparrow$	$Acc_{rt} \uparrow$
Retrain	100	66.7	100	66.7	100	66.7	100	100	100	0.0	100	100	100	0.0	100	33.3	100	66.7	100	55.6
RndLbl	66.7	33.3	100	100	62.5	33.3	62.5	66.7	91.7	33.3	83.3	33.3	100	0.0	70.8	33.3	83.3	33.3	80.1	40.7
NegGrad+	100	100	66.7	0.0	95.8	66.7	87.5	66.7	75.0	0.0	91.7	33.3	95.8	0.0	100	0.0	100	66.7	90.3	37.0
BShrink	4.2	0.0	87.5	33.3	20.8	0.0	70.8	66.7	79.2	0.0	87.5	0.0	58.3	0.0	87.5	0.0	91.7	66.7	65.3	18.5
BExpand	29.2	0.0	95.8	66.7	50.0	33.3	91.7	100	87.5	0.0	91.7	0.0	62.5	0.0	95.8	33.3	95.8	100	77.8	37.0
SCRUB	75.0	33.3	95.8	66.7	75.0	33.3	100	100	91.7	33.3	100	33.3	100	0.0	83.3	33.3	100	100	91.2	48.1
BadT	100	33.3	100	100	83.3	33.3	100	66.7	83.3	0.0	100	66.7	87.5	0.0	95.8	33.3	95.8	66.7	94.0	44.4
SalUn	100	100	100	66.7	87.5	33.3	91.7	66.7	100	0.0	100	66.7	95.8	0.0	95.8	0.0	95.8	33.3	96.3	40.7
DELETE	100	33.3	100	66.7	100	100	95.8	100	87.5	0.0	95.8	66.7	87.5	0.0	100	0.0	100	66.7	96.3	48.1
Ours	**100**	**100**	**100**	66.7	**100**	66.7	95.8	**100**	**100**	**33.3**	**100**	**66.7**	95.8	0.0	**100**	**33.3**	**100**	**100**	**99.1**	**59.3**

Figure 6. **Visualization of similar identities on the CelebA-Top500 dataset.** Ten highly similar identities are selected for unlearning evaluation. ID 107 (highlighted in red) represents the target identity to be forgotten, while the remaining nine serve as similar retain identities. Numbers below each image denote cosine similarity to ID 107, ranging from 0.711 (most similar) to 0.380 (least similar). The average similarity across the group is 0.515, reflecting strong visual resemblance.

C.2. The results of unlearning on CelebA-Top500

To further verify that our method can generalize to the task of face unlearning, we conduct additional experiments on the CelebA-Top500 dataset. Following the same setup as in Section 5.2, we designate ID 107 as the forgotten identity and treat the remaining identities as retained classes. Table 11 reports the main results on CelebA-Top500 class unlearning. Our method achieves a Gap Score of 1.63, outperforming all baselines by more than 60% (next best: SalUn at 4.13). In terms of overall performance, measured by H-Mean, our method achieves 83.04%, closely approaching the retraining upper bound of 84.56% while maintaining the highest retention accuracy (Acc_r = 99.89%). These results indicate that our method delivers consistently better retain-forget trade-offs.

To additionally evaluate how effectively our method preserves identities that are visually most confusable with the forgotten identity, we identify the nine remaining identities with the highest cosine similarity to ID 107. Figure 6 visualizes the ten identities involved in this evaluation. The similarity scores span from 0.711 to 0.380, reflecting a continuous spectrum of visual resemblance and highlighting the fine-grained difficulty of disentangling the forgotten identity from its most similar counterparts.

Our method achieves the highest average accuracy on both the training set (99.1%) and the test set (59.3%), exceeding the strongest baseline by approximately 23%. This strong performance on identities most confusable with the forgotten class demonstrates our method's ability to preserve fine-grained distinctions without compromising un-

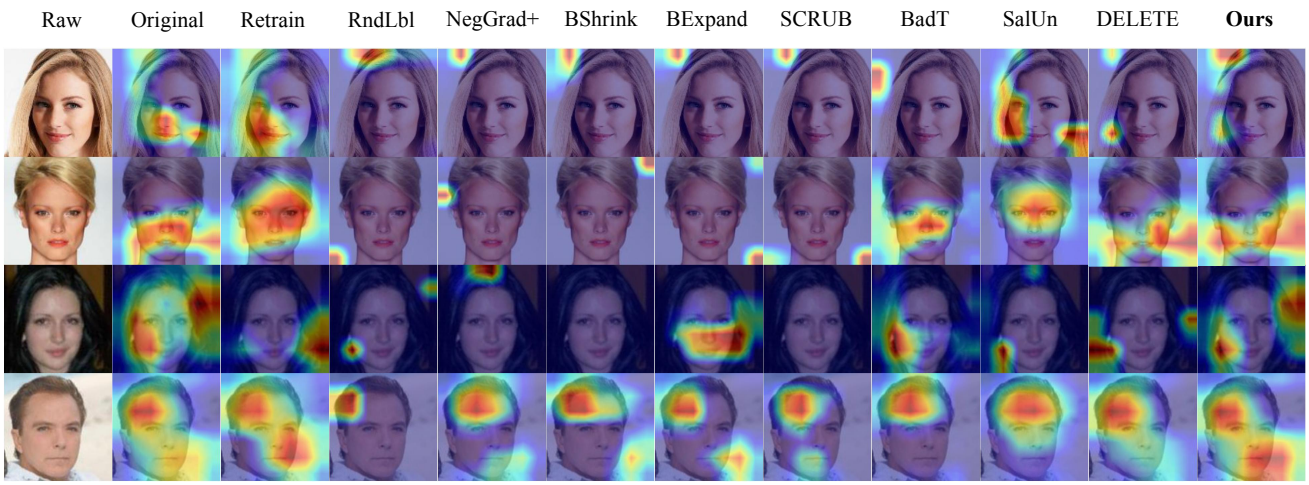


Figure 7. **Grad-CAM Activation Heatmaps for Retained Identities on the CelebA-Top500 dataset.** Our method preserves stable and concentrated activation in key facial regions for retained identities, while other approaches often exhibit weakened or scattered responses. This demonstrates that our method maintains the discriminative information of retained identities more effectively.

learning quality, as summarized in Table 12.

C.3. Visualization Results on CelebA-Top500

Figure 7 shows the Grad-CAM activation maps for several retained identities. Our method produces stable and concentrated activations in key facial regions, indicating that the model continues to extract reliable and discriminative cues after unlearning. In comparison, many baseline methods display weaker or scattered responses, suggesting disruptions to the underlying feature representations. These results provide intuitive evidence that our method better preserves identity-specific information on the challenging CelebA-Top500 dataset.

C.4. Ablation study on hyper-parameter sensitivity

In this section, we investigate the sensitivity of our proposed method to key hyperparameters across its core components: the learning rate η_a in Adaptive Gradient Reweighting (AGR), the moving average momentum μ in Temporal Stabilization (TS), and the geometric rectification strength α in Direction Rectification (DR). The results on Random (10%) and Class (1 class) unlearning are summarized in Table 14. Throughout the experiments, we adopt a control variable approach, maintaining our default configuration of η_a = 5e-3, μ = 0.9, and α = 1.0, which yields the best overall performance.

C.5. Quantitative comparison of execution time

As shown in Table 13, we compare the execution time (in minutes) for unlearning 10% of CIFAR-10 and CIFAR-100. Overall, our method achieves highly competitive efficiency across both ResNet-18 and ViT architectures. Particularly on the more complex ViT model, our approach consistently requires the lowest execution time across both datasets. This confirms that our proposed framework is highly efficient and introduces minimal computational overhead.

Figure 8. Visual results of erasing "Church" and preserving the other concepts.

D. Extension to Generative Models

D.1. Methodology for Diffusion Models

Our approach is general and can be readily extended to large-scale generative models. In diffusion models, we minimize the forgetting loss by pulling the erase concept set E toward a neutral/null concept c_n (e.g., "a photo" or " "):

$$\mathcal{L}_f(\theta; c_e) = \mathbb{E}_{x,t,\epsilon}\left[\|\epsilon_\theta(x_t, c_e) - -\epsilon_\theta(x_t, c_n)\|_2^2\right], c_e \in E. \tag{18}$$

To preserve the remaining concepts, we define the retention loss with respect to a frozen teacher θ_0:

$$\mathcal{L}_r(\theta; c) = \mathbb{E}_{x,t,\epsilon}\left[\|\epsilon_\theta(x_t, c) - \epsilon_{\theta_0}(x_t, c)\|_2^2\right], c \in R. \tag{19}$$

Table 13. Efficiency comparison for unlearning 10% of CIFAR-10 and CIFAR-100. The best results are shown in **bold**.

Model	Dataset	Finetuning	NegGrad+	RndLbl	BadT	SCRUB	SSD	UNSIR	SalUn	LoTUS	Ours
ViT	C-10	11.69	12.49	12.75	8.97	16.50	13.48	10.23	35.08	7.85	**6.19**
	C-100	11.35	12.57	12.72	9.86	16.55	13.23	10.76	38.15	7.15	**7.00**
RN18	C-10	0.46	0.51	0.73	0.43	0.58	0.49	0.57	1.38	**0.33**	0.47
	C-100	0.83	0.89	0.97	0.74	0.98	0.93	0.85	1.99	0.68	**0.60**

Table 14. Hyperparameter sensitivity of η_a, μ, and α on Random (10%) and Class (1 class) unlearning.

Hyperparameter			Random (10%)	Class (1 class)
η_a	μ	α	Avg Gap ↓	H-Mean ↑
5e-2	0.9	1.0	1.74	94.13
5e-3	0.9	1.0	1.57	94.67
5e-4	0.9	1.0	2.75	91.18
5e-3	0.6	1.0	2.85	93.74
5e-3	0.9	1.0	1.57	94.67
5e-3	1.0	1.0	3.57	93.02
5e-3	0.9	0.8	2.36	92.52
5e-3	0.9	1.0	1.57	94.67
5e-3	0.9	1.2	1.91	93.10

Table 15. **Erasing object-related concepts.** Higher values indicate better performance.

Method	ESR-1↑	ESR-5↑	PSR-1↑	PSR-5↑
SD	15.2	2.3	83.7	98.1
ESD [22]	94.8	88.2	47.1	53.2
UCE [23]	**99.9**	**99.9**	24.0	51.6
CA [34]	97.3	92.7	53.4	70.1
Ours	98.1	93.5	**67.8**	**81.3**

We learn an adversarial concept attribution policy $\pi \in \Delta^{|R|-1}$ and formulate a worst-case retention objective:

$$\min_{\theta} \max_{\pi \in \Delta^{|R|-1}} \mathbb{E}_{z \sim G(\pi;\tau)} [\mathcal{L}_r(\theta;\, c(z))]\,, c(z) = z^\top \mathbf{R}. \tag{20}$$

where R is instantiated as a finite vocabulary (e.g., Oxford 3000 or the CLIP token vocabulary) without E the erase concept removed, $\mathbf{R}$ denotes the corresponding concept embedding matrix, and $G(\cdot)$ denotes the Gumbel-Softmax operator. By Eqs. (18) and (20), we compute the forgetting and retention gradients, g_f and g_r, and then obtain the final update direction by our Multi-stage Objective Optimization.

D.2. Experimental Evaluation

In this experiment, following prior work on concept erasure [2, 3, 22, 23, 34], we evaluate the effectiveness of our method in removing different types of concepts using a pre-trained Stable Diffusion v1.4 model as the foundation model for all experimental trials. Specifically, we consider two representative settings: object-wise forgetting and concept-wise forgetting. For object-wise forgetting, we adopt the Imagenette dataset, a 10-class subset of ImageNet [36] designed for simple object recognition, with Stable Diffusion (SD), while for concept-wise forgetting, we use the I2P dataset [46] with SD.

Table 16. **Evaluation on the nudity erasure setting.** Lower values indicate better performance.

Method	NER-0.3↓	NER-0.5↓	NER-0.7↓	FID↓
SD	18.1	9.3	1.9	-
ESD [22]	6.8	3.7	1.9	23.1
UCE [23]	8.1	4.2	1.4	19.5
CA [34]	13.3	7.3	4.1	27.8
Ours	**4.1**	**1.5**	**0.8**	**17.7**

Object-wise Unlearning. In class-wise unlearning, a subset of object categories is selected as the target concepts to be removed, while the remaining categories are used to evaluate the preservation of unrelated concepts. For each target category, we generate 200 images using prompts that describe the corresponding object. A pretrained ResNet-50 classifier is used to identify object categories in the generated images.

To evaluate the erasure effectiveness, we adopt the Erasing Success Rate (ESR-k), defined as the percentage of generated images in which the removed object does not appear within the top-k predictions of the classifier. To measure the preservation ability, we use the Preserving Success Rate (PSR-k), defined as the percentage of generated images where non-target objects remain correctly recognized within the top-k predictions.

The quantitative results are summarized in Table 15. Our method achieves competitive erasure performance while preserving significantly more non-target concepts compared with existing approaches. In particular, it maintains higher PSR scores while achieving comparable ESR values, indicating a better balance between concept removal and knowl-

edge preservation. Figure 8 shows qualitative examples, where the target concept ("Church") is effectively removed while other semantic elements remain intact.

Concept-wise Unlearning. For concept-wise unlearning, we evaluate the removal of unsafe content using the I2P dataset [46]. Images are generated from all prompts using the edited diffusion model. The generated images are then analyzed using an automatic nudity detector to determine whether explicit body parts appear in the outputs.

We evaluate Not-Safe-For-Work (NSFW) removal using the Nudity Exposure Rate (NER), defined as the percentage of generated images with detected nude body parts, where lower values indicate better suppression of unsafe content. We further report the Fréchet Inception Distance (FID) against the COCO-30K validation set to measure the generation quality after concept removal.

The quantitative comparison is reported in Table 16. Our method consistently achieves lower NER scores across different detection thresholds, indicating stronger suppression of unsafe content. At the same time, it obtains the best FID score, suggesting that our approach preserves the overall image generation quality more effectively than existing methods.