SPECTRAL SALIENCY FOR MACHINE UNLEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Machine unlearning (MU) aims to remove the influence of specific training data while preserving model utility. As the name suggests, MU can be viewed as the inverse of learning, using gradient-based updates to reduce the influence of a forget-set by counteracting the previously learned behavior. Recently, Muon, a gradient descent variant, has been introduced. Muon applies spectral magnitude normalization to encourage exploration of rare directions and demonstrates promising performance. Inspired by Muon, we adopt the spectral view for unlearning and propose Spectral Saliency Unlearning (SSU). SSU thresholds weak singular components and updates only those directions supported by a confident unlearning signal. We further provide theoretical justification for this thresholding approach from the perspective of the forgetting-retention trade-off. Experiments across image classifiers, diffusion models, and LLMs demonstrate SSU's effectiveness.

1 INTRODUCTION

Machine unlearning (MU) studies how to remove the influence of specific training data from a trained model without resorting to full retraining (Cao & Yang, 2015). There is an increasing need for MU to satisfy data-governance requirements, such as the General Data Protection Regulation (GDPR) (European Union, 2016). Recent methods focus on approximate unlearning, which aims to efficiently reduce the influence of the forget set through a few gradient-based updates (Fan et al., 2024; Thudi et al., 2022; Izzo et al., 2021). These methods typically optimize a forgetting objective that increases the model's loss on the forget set or pushes a model's predictions toward misclassifications.

In this work, we conceptualize MU as the partial "reversal" of the optimization process. In the deep learning era, learning can be viewed as using stochastic gradient descent (SGD) to incrementally accumulate the influence of training examples into the model parameters. Correspondingly, approximate unlearning can be viewed as undoing the effect of gradient descent on the designated forget set. UnrollingSGD (Thudi et al., 2022) provides one concrete instantiation, analyzing unlearning through the lens of the SGD training dynamics. Motivated by the central role of SGD in deep learning, it is natural to ask how variants of SGD that alter the geometry and scaling of updates translate into the design of unlearning updates.

A canonical variant is SignSGD (Bernstein et al., 2018), which replaces the stochastic gradient $\boldsymbol{g}$ with its coordinate-wise sign, *i.e.*, $\operatorname{sign}(\boldsymbol{g})$, thereby equalizing update magnitudes across coordinates and emphasizing directional information over scale. For matrix-structured parameters (He et al., 2016; Vaswani et al., 2017), recent optimizers such as Muon (Jordan et al., 2024) extend this magnitude normalization beyond coordinates by applying the matrix sign in the spectral domain of layer gradients. This design is motivated in part by the idea that exploring "rare", weakly expressed directions can be beneficial for learning (Jordan et al., 2024). The approach is based on analyses of deep learning dynamics showing that training progresses along singular directions in order of singular value magnitude. Smaller singular values associated with weaker and more specific structure are learned later (Kleinman et al., 2024; Rahaman et al., 2019; Lampinen & Ganguli, 2019).

From this perspective, we view the recent unlearning method SalUn (Fan et al., 2024) as an analogue of SignSGD. Instead of uniformizing coordinate magnitudes, SalUn thresholds low-magnitude coordinates of the unlearning gradient. As SignSGD and Muon are both SGD-variants, where Muon promotes rare directions by equalizing magnitudes in the gradient spectrum, one might ask whether *unlearning could similarly benefit* from suppressing weak spectral components. We hypothesize that these weak components may encode interactions where forgetting and retention are entangled. This intuition is formalized through the concept of forgetting-retention conflict in Sec. 4.

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.

We propose Spectral Saliency Unlearning (SSU), a singular value thresholding method that suppresses weak spectral components of the unlearning gradient and can be plugged into general gradient-based unlearning methods. We justify SSU from the perspective of the forgetting-retention trade-off, characterizing how weak directions relate to forget-retain interference. We then explain why suppressing them can mitigate utility degradation, *i.e.*, the loss of model performance on data not in the forget set. We then extend the same perspective to coordinate-wise masking and offer a theoretical explanation for the efficacy of SalUn-style unlearning, which has largely been used as a heuristic.

Empirically, we evaluate SSU across diverse applications, including the unlearning of image classifiers, diffusion models for image generation (Fan et al., 2024), and large language models (Yuan et al., 2025). Results demonstrate consistent improvements when SSU is used with existing MU objectives. For image classifiers, SSU reduces the average gap by 9% and 18% at forgetting ratios of 10% and 50%, respectively, for CIFAR-10, and achieves a 27.7% reduction on Tiny-ImageNet, with improvements across different random seeds. For diffusion models, SSU achieves perfect unlearning efficacy with DDPM while improving generation quality on retained classes by 23.6%. For large language models, it consistently improves the utility-forgetting trade-off, yielding an average gain of 0.0125 on the TOFU benchmark. **Our main contributions are:**

- Inspired by recent advances in optimization, we introduce spectral saliency for matrix gradients and propose SSU, a novel singular value thresholding approach for gradient-based unlearning.
- We provide a justification for saliency-based thresholding from the forgetting-retention trade-off perspective, and extend the analysis to explain the efficacy of SalUn-style coordinate masking.
- Extensive experiments demonstrate the effectiveness of SSU across a range of models, including image classifiers, diffusion models, and LLM unlearning.

2 Background

Unlearning formulation. We consider a general gradient-based approximate unlearning framework (Thudi et al., 2022; Neel et al., 2021; Gandikota et al., 2023; Kumari et al., 2023; Gandikota et al., 2025). A pre-trained model is updated to reduce the influence of a given forget set $\mathcal{D}_f$ from the training data $\mathcal{D}$ while preserving utility on the complementary retain set $\mathcal{D}_r \coloneqq \mathcal{D} \setminus \mathcal{D}_f$. Let $\boldsymbol{W}$ denote model parameters. We define an unlearning objective

$$\mathcal{L}_u(\boldsymbol{W}) = \mathcal{L}_f(\boldsymbol{W}; \mathcal{D}_f) + \mathcal{L}_r(\boldsymbol{W}; \mathcal{D}_r) \tag{1}$$

composed of a forgetting loss $\mathcal{L}_f(\boldsymbol{W}; \mathcal{D}_f)$ and a retaining loss $\mathcal{L}_r(\boldsymbol{W}; \mathcal{D}_r)$, which are optimized jointly via gradient-based updates to minimize Eq. (1). We further denote the gradient update $\nabla \mathcal{L}_u(\boldsymbol{W})$ at each step as $\boldsymbol{G}_u = \boldsymbol{G}_f + \boldsymbol{G}_r$, where $\boldsymbol{G}_f = \nabla \mathcal{L}_f(\boldsymbol{W}; \mathcal{D}_f)$ and $\boldsymbol{G}_r = \nabla \mathcal{L}_r(\boldsymbol{W}; \mathcal{D}_r)$.

Muon and matrix sign. Muon (Jordan et al., 2024) is a recently proposed optimizer for matrix-parameterized functions. It orthogonalizes the layer gradient by the projection

$$\boldsymbol{G}_o = \arg\min_{\boldsymbol{O}} \left\{ \|\boldsymbol{G} - \boldsymbol{O}\|_F : \boldsymbol{O}^\top \boldsymbol{O} \text{ or } \boldsymbol{O}\boldsymbol{O}^\top = \boldsymbol{I} \right\}. \tag{2}$$

This operation is equivalent to applying the matrix sign in the gradient's spectral domain (Chen et al., 2025), *i.e.*, for $\boldsymbol{G} = \boldsymbol{U}\,\mathrm{diag}(\boldsymbol{\sigma})\,\boldsymbol{V}$, then we have $\mathrm{msign}(\boldsymbol{G}) \coloneqq \boldsymbol{U}\,\mathrm{diag}(\mathrm{sign}(\boldsymbol{\sigma}))\,\boldsymbol{V}$. We therefore view Muon as a matrix counterpart of SignSGD.

3 Method

For a gradient-based unlearning algorithm that minimizes the objective given in Eq. (1), our proposed SSU replaces the base unlearning gradient $\boldsymbol{G}_u$ with a singular value thresholded counterpart that suppresses weak spectral components and retains only the dominant singular directions for the update. The full procedure is summarized in Alg. 1. In the following, we elaborate on this procedure by first defining spectral saliency and then describing the singular value thresholding step.

3.1 Spectral Saliency

We define spectral saliency by measuring the magnitude of the unlearning update $\boldsymbol{G}_u$ along the singular directions of the layer gradient. For a matrix-structured parameter $\boldsymbol{W}$, let the unlearning gradient's SVD be denoted as

$$\boldsymbol{G}_u = \boldsymbol{U}\ \mathrm{diag}(\boldsymbol{\sigma})\,\boldsymbol{V}^\top \quad \text{with} \quad \boldsymbol{\sigma} = [\sigma_1, \cdots, \sigma_m]^\top, \tag{3}$$

where $\sigma_1 \geq \cdots \geq \sigma_m \geq 0$, so that each rank-one component $\sigma_i\, \boldsymbol{u}_i \boldsymbol{v}_i^\top$ represents the update along the singular direction pair $(\boldsymbol{u}_i, \boldsymbol{v}_i)$. Note, we omit layer and block indices for readability.

Algorithm 1 Spectral Saliency Unlearning (SSU)

input Pre-trained parameters $\boldsymbol{W}_0$; unlearning objective $\mathcal{L}_u(\boldsymbol{W}) = \mathcal{L}_f(\boldsymbol{W}) + \mathcal{L}_r(\boldsymbol{W})$, learning rate η, keep ratio $\gamma \in (0, 1]$

for $t = 0, 1, \ldots, T-1$ **do**
  $\boldsymbol{G}_{u,t} \leftarrow \nabla_{\boldsymbol{W}} \mathcal{L}_u(\boldsymbol{W}_t)$.
  for each matrix-structured gradient $\boldsymbol{G}_{u,t}^{(\ell)}$ **do**
    Compute SVD: $\boldsymbol{G}_{u,t}^{(\ell)} = \boldsymbol{U}\,\mathrm{diag}(\boldsymbol{\sigma})\boldsymbol{V}^\top$
    Set $k \leftarrow \lfloor \gamma m \rfloor$ for $\boldsymbol{\sigma} \in \mathbb{R}^m$.
    $\widetilde{\boldsymbol{G}}_{u,t}^{(\ell)} \leftarrow \boldsymbol{U}_{[:,1:k]}\ \mathrm{diag}(\boldsymbol{\sigma}_{[1:k]})\,\boldsymbol{V}_{[:,1:k]}^\top$
  end for
  $\boldsymbol{W}_{t+1} \leftarrow \boldsymbol{W}_t - \eta \widetilde{\boldsymbol{G}}_{u,t}$
end for

output $\boldsymbol{W}_T$

The spectral saliency of direction i is defined as its singular magnitude σ_i, which quantifies how strongly the unlearning objective drives updates along this direction. We view larger saliency as corresponding to a stronger and more reliable unlearning signal, whereas small singular magnitudes indicate weakly supported directions that are less confident for unlearning and may be susceptible to interference between forgetting and retention, an intuition we later formally justify in Sec. 4.

3.2 Singular Value Thresholding

To suppress these weak directions, we apply *singular value thresholding* (Cai et al., 2010) to $\boldsymbol{G}_u$. Given a threshold $\tau \geq 0$, we define the thresholded gradient

$$\widetilde{\boldsymbol{G}}_u \coloneqq \boldsymbol{U}\,\mathrm{diag}\left(\boldsymbol{\sigma} \odot \mathbb{I}_{[\boldsymbol{\sigma} \geq \tau]}\right)\boldsymbol{V}^\top, \quad \text{with} \quad \mathbb{I}_{[\boldsymbol{\sigma}\geq\tau]}(i) = \begin{cases} 1 & \text{if } \sigma_i \geq \tau \\ 0 & \text{otherwise} \end{cases}. \tag{4}$$

$\mathbb{I}_{[\boldsymbol{\sigma}\geq\tau]}$ is a vector-valued indicator whose i^{th} coordinate is 1 when $\sigma_i \geq \tau$ and 0 otherwise. Equivalently, we may retain the leading components and discard the tail as the singular values are sorted.

Different from coordinate-wise saliency methods, *e.g.*, SalUn (Fan et al., 2024), which constructs a coordinate selection mask using dataset-specific computations (*e.g.*, based on the forget set) either offline or online, SVD provides an intrinsic ordering of singular values. The most salient directions correspond naturally to the leading singular components. We use a fixed *keep ratio* $\gamma \in (0, 1]$ and set $k = \lfloor \gamma m \rfloor$, retaining the top γ fraction of singular components. We then reconstruct

$$\widetilde{\boldsymbol{G}}_u \coloneqq \sum_{i=1}^{k} \sigma_i\, \boldsymbol{u}_i \boldsymbol{v}_i^\top = \boldsymbol{U}_{[:,1:k]}\ \mathrm{diag}(\boldsymbol{\sigma}_{[1:k]})\,\boldsymbol{V}_{[:,1:k]}^\top \tag{5}$$

and proceed with standard gradient-based updates using this thresholded gradient. For vector- and scalar-structured parameters (*e.g.*, biases), we resort to coordinate-wise thresholding as in SalUn, mirroring Muon's practice of handling these parameters with vector-based optimizers rather than a spectral update.

4 Theoretical Justification

Our analysis adopts a directional view of the unlearning update, characterizing when and why suppressing weakly supported directions can improve unlearning from the perspective of the forgetting-retention trade-off. We first develop this justification for spectral thresholding on matrix-structured parameters, then derive an analogous interpretation for coordinate-wise masking.

4.1 Why Spectral Thresholding Helps

Motivated by the Muon perspective of operating in the gradient's spectral domain, and by analyses linking smaller singular values to weaker, more specific structure (Kleinman et al., 2024; Rahaman et al., 2019; Lampinen & Ganguli, 2019), we take the reverse stance for unlearning. Weak spectral

components tend to be the least reliable directions to update for unlearning, as they may reflect forget-retain entanglement and thereby worsen the forgetting-retention trade-off. Accordingly, we use the singular basis of the unlearning gradient to explain why suppressing these components via spectral thresholding can improve unlearning. We start by examining how the forgetting gradient and the retaining gradient contribute to and interact within the unlearning gradient. We consider the compact SVD of the unlearning gradient, retaining only non-zero singular values:

$$\boldsymbol{G}_u = \boldsymbol{U}_e \operatorname{diag}(\boldsymbol{\sigma}_e) \boldsymbol{V}_e^\top = \sum_{i=1}^{r_e} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^\top, \sigma_1 \geq \cdots \geq \sigma_{r_e} > 0 \tag{6}$$

where r_e denotes the effective rank: the exact rank in theory and the number of significantly nonzero singular values in practice. We refer to $\mathcal{E} = \operatorname{span}(\boldsymbol{V}_e)$ as the effective subspace of update directions induced by $\boldsymbol{G}_u$. In this subspace, each singular direction $\boldsymbol{v}_i$ defines a canonical one-dimensional update mode with a magnitude σ_i.

To understand how the forgetting and retention gradients interact within the effective subspace, we introduce the notion of *forget-retain alignment*, which measures the directional alignment between the projected forgetting and retention gradients.

Definition 4.1. (Forget-Retain Alignment) We define the alignment between the forgetting and retention directions of $\boldsymbol{v}_i$ as

$$a(\boldsymbol{v}_i) = \frac{\langle \boldsymbol{G}_f \boldsymbol{v}_i, \boldsymbol{G}_r \boldsymbol{v}_i \rangle}{\|\boldsymbol{G}_f \boldsymbol{v}_i\|_2 \|\boldsymbol{G}_r \boldsymbol{v}_i\|_2}. \tag{7}$$

Using this notion, Prop. 4.2 shows that when the forgetting and retaining gradients exhibit strong negative alignment in the effective subspace, this indicates that there is significant conflict between the forgetting and retention objectives, which therefore cannot be simultaneously improved along the same update direction, yielding an unavoidable trade-off.

Proposition 4.2. *(Informal) Assume $\mathcal{L}_f$ and $\mathcal{L}_r$ are smooth. For unlearning gradient $\boldsymbol{G}_u = \eta \sum_{i=1}^{r_e} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^\top$, consider a single-direction update along the ith singular component $\boldsymbol{v}_i$: $\boldsymbol{W}' = \boldsymbol{W} - \eta \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^\top$. There exist constants $a_0 < 0$ and $\eta_0 > 0$ such that if the alignment $a(\boldsymbol{v}_i) < a_0$, and the step size $\eta < \eta_0$, then the forgetting progress $\Delta\mathcal{L}_f = \mathcal{L}_f(\boldsymbol{W}') - \mathcal{L}_f(\boldsymbol{W})$ and the retaining progress $\Delta\mathcal{L}_r = \mathcal{L}_r(\boldsymbol{W}') - \mathcal{L}_r(\boldsymbol{W})$ satisfy $\Delta\mathcal{L}_f \Delta\mathcal{L}_r < 0$. That is, along any singular direction $\boldsymbol{v}_i$ where the retain and forget gradients are in significant conflict, any update that improves one objective must necessarily worsen the other.*

We refer the readers to Appx. A.1 for the formal statement, assumptions, and proof of this proposition. Next, we show how this conflict is reflected in the spectrum of the unlearning gradient.

Proposition 4.3. **(a)** *If the direction carries a nontrivial forget/retain signal,* i.e., $\|\boldsymbol{G}_f \boldsymbol{v}_i\| + \|\boldsymbol{G}_r \boldsymbol{v}_i\| \geq \xi$ *for $\xi > 0$ and its singular value is small relative to this signal,* i.e., $\sigma_i \leq \rho\xi$ *for $\rho \in \left(0, \frac{1}{\sqrt{2}}\right)$, then its forget-retain alignment satisfies $a(\boldsymbol{v}_i) \leq 2\rho^2 - 1 < 0$.*
(b) *(Informal) Assume the projections of $\boldsymbol{G}_f$ and $\boldsymbol{G}_r$ onto the effective subspace of $\boldsymbol{G}_u$ have bounded spectral disparity. For two directions $\boldsymbol{v}_i, \boldsymbol{v}_j$ with alignment scores $a(\boldsymbol{v}_i) \leq -\delta_i$ and $a(\boldsymbol{v}_j) \geq \delta_j$ where $\delta_i, \delta_j \in (0, 1]$, if the separation $(\delta_i + \delta_j)$ exceeds a gap threshold, then $\sigma_i < \sigma_j$.*

We refer the readers to Appx. A.2 for the assumption of spectral disparity, the formal statement, and proof. Prop. 4.3 (a) shows that as long as there is a nontrivial forgetting or retention signal, small singular values correspond to directions in which the forgetting and retaining gradients are negatively aligned. If a direction carries neither forgetting nor retention signal, *i.e.*, $\|\boldsymbol{G}_f \boldsymbol{v}_i\| = \|\boldsymbol{G}_r \boldsymbol{v}_i\| = 0$, then $\boldsymbol{v}_i$ lies in the null space of $\boldsymbol{G}_u$, which is not affected by singular value thresholding. Prop. 4.3 (b) complementarily shows that, between sufficiently positively aligned directions and negatively aligned ones, the former attain larger singular values. Together with Prop. 4.2, this analysis shows that by suppressing small singular components of $\boldsymbol{G}_u$, SSU preferentially removes directions that are more likely to exhibit strong forget-retain conflict, thereby mitigating unnecessary retention degradation while preserving the dominant unlearning signal.

4.2 Justification for SalUn-Style Masking

We further extend the same directional analysis to SalUn-style coordinate thresholding, which operates on the unlearning gradient in the standard coordinate basis. Each coordinate $i \in [d]$ defines

a one-dimensional update direction, and the unlearning gradient decomposes as $g_{u,i} = g_{f,i} + g_{r,i}$. A key regime of interest is when a coordinate carries a nontrivial forgetting or retention signal (otherwise updates along that coordinate are negligible), yet the magnitude $|g_{u,i}|$ is small, suggesting cancellation between forgetting and retention effects along that coordinate. This explains why SalUn's heuristic of suppressing small-magnitude coordinates can be beneficial: updating along such weakly supported coordinates is likely to couple forgetting progress with retention degradation, as characterized by the following proposition whose proof can be found in Appx. A.3.

Proposition 4.4. *Assume $\mathcal{L}_f$ and $\mathcal{L}_r$ are β_f- and β_r-smooth. Consider any coordinate $i \in [d]$ such that the forgetting/retention signal is nontrivial,* i.e., *$\forall i \in [d]$, $|g_{f,i}| + |g_{r,i}| \geq \xi$ where $\xi > 0$, and the combined unlearning gradient is small,* i.e., *$|g_{u,i}| \leq \rho\xi$ for some $\rho \in (0, 1]$. For the coordinate-wise update $\boldsymbol{w}' = \boldsymbol{w} - \eta g_{u,i}$ with step size $\eta < \min\left(\frac{2|g_{f,i}|}{\beta_f |g_{u,i}|}, \frac{2|g_{r,i}|}{\beta_r |g_{u,i}|}\right)$, the induced progress $\Delta\mathcal{L}_f = \mathcal{L}_f(\boldsymbol{w}') - \mathcal{L}_f(\boldsymbol{w})$ and $\Delta\mathcal{L}_r = \mathcal{L}_r(\boldsymbol{w}') - \mathcal{L}_r(\boldsymbol{w})$ satisfy $\Delta\mathcal{L}_f \Delta\mathcal{L}_r < 0$.*

That is to say, the unlearning update necessarily improves one objective while worsening the other, between forgetting and retention. To our knowledge, existing saliency-masking methods are primarily motivated empirically. The proposition above offers a nontrivial theoretical justification from the forgetting-retention trade-off perspective.

Limitations and discussion. Our results justify SSU as a principled mechanism for mitigating forget-retain conflict by suppressing weak spectral components. At the same time, the theory is intended as a characterization of the forgetting-retention trade-off rather than a complete quantitative prediction of end-to-end unlearning performance. Extending the analysis to weaker assumptions and deriving tighter performance-level guarantees remain important directions for future work. A natural variant of SSU would be to threshold directions by forget-retain alignment. However, this requires separately computing $\boldsymbol{G}_f$ and $\boldsymbol{G}_r$, projecting both onto the singular directions of $\boldsymbol{G}_u$, and then evaluating their directional inner products, introducing nontrivial computation overhead. Singular values, in contrast, are obtained directly from $\boldsymbol{G}_u$ and, supported by Prop. 4.3, serve as a principled proxy for forget-retain conflict. While a more fine-grained, alignment-aware criterion may improve thresholding, we focus on the simple, broadly applicable choice of singular value thresholding.

5 Experiments

We conduct experiments across three applications spanning over image classification (Sec. 5.1), image generation (Sec. 5.2), and language modeling (Sec. 5.3) following existing MU works.

5.1 Random Subset Unlearning in Image Classification

Setup and evaluation. Following Fan et al. (2024), we focus on random subset unlearning in image classification tasks using the CIFAR-10 (Krizhevsky, 2009) and Tiny-ImageNet datasets (Le & Yang, 2015). We employ ResNet-18 (He et al., 2016) as our architecture and compare our method against three simple baselines: fine-tuning (FT) (Warnecke et al., 2021), gradient ascent (GA) (Thudi et al., 2022), influence unlearning (IU) (Izzo et al., 2021), and six competitive baselines: ℓ_1-sparse (Jia et al., 2023), SCRUB (Kurmanji et al., 2023), SSD (Foster et al., 2024), SEMU (Sendera et al., 2025), SFRON (Huang et al., 2024), and SalUn (Fan et al., 2024).

Following SalUn's setup, we regard the retrained oracle as the gold standard of unlearning. To compare each method with this reference, we report four metrics: forgetting set accuracy (FA) to measure unlearning efficacy, remaining set accuracy (RA) and test set accuracy (TA) to assess preserved generalization, and the membership inference attack (MIA) (Fan et al., 2024) success rate on the forgetting set as a privacy metric. We measure each metric's distance from its corresponding oracle value and report the average gap across the four metrics as the overall performance. Implementation details are provided in Appx. B.1.

Results. In Tab. 1, we summarize the results for random subset unlearning on CIFAR-10 with ResNet-18, where 10% of the training data is forgotten. Most baseline methods face a fundamental trade-off: techniques such as FT, GA, IU, SCRUB, SSD, SEMU, and SFRON yield high FA, indicating insufficient forgetting, while methods like ℓ_1-sparse that achieve lower FA suffer large drops in RA and TA, compromising model utility.

Among the baselines, SalUn achieves the best overall performance with an average gap of 1.21. Our method improves upon SalUn across all five different seeds, achieving the lowest average gap of 1.10 with a paired t-test p-value of 0.034, while maintaining comparable performance across all metrics. Although SSU employs SVD, the actual computational overhead is negligible during unlearning due to the low-rankness in model weights.

Table 1: CIFAR-10, 10% random data forgetting (Mean ± std over five seeds).

Methods	FA	RA	TA	MIA	Avg. Gap ↓	Time (min) ↓
Retrain	94.57±0.57	100.00±0.00	94.25±0.10	13.35±0.57	0.00	43.29
FT	99.91±0.04	99.95±0.04	94.11±0.15	1.29±0.15	4.42±0.27	2.37
GA	100.00±0.00	100.00±0.00	94.66±0.06	0.49±0.10	4.68±0.27	0.13
IU	100.00±0.00	100.00±0.00	94.63±0.10	0.39±0.10	4.70±0.26	3.22
ℓ_1-sparse	94.81±0.54	97.16±0.43	90.76±0.65	10.95±0.91	2.38±0.23	2.36
SCRUB	100.00±0.00	100.00±0.00	94.69±0.11	0.40±0.11	4.71±0.25	1.88
SSD	99.48±0.70	99.46±0.63	93.31±1.28	2.47±1.77	4.31±0.14	2.78
SEMU	100.00±0.00	100.00±0.00	94.61±0.06	1.00±0.32	4.54±0.32	1.05
SFRON	97.26±0.30	99.79±0.03	94.13±0.08	8.24±0.71	2.04±0.12	1.90
SalUn	97.03±0.55	99.28±0.26	93.16±0.45	13.78±0.86	1.21±0.25	2.61
+ SSU	97.06±0.56	99.38±0.21	93.26±0.38	13.38±0.67	**1.10±0.20**	2.62

Table 2: CIFAR-10, 50% random data forgetting (Mean ± std over five seeds).

Methods	FA	RA	TA	MIA	Avg. Gap ↓
Retrain	91.07±0.24	100.00±0.00	90.70±0.20	20.77±0.60	0.00
FT	99.91±0.04	99.97±0.02	94.02±0.33	1.54±0.18	7.85±0.32
GA	99.93±0.09	99.94±0.08	94.31±0.52	0.64±0.20	8.19±0.12
IU	100.00±0.00	100.00±0.00	94.64±0.10	0.41±0.09	8.31±0.25
ℓ_1-sparse	98.82±0.49	99.34±0.39	92.54±0.55	4.58±1.07	6.61±0.50
SCRUB	100.00±0.00	100.00±0.00	94.67±0.16	0.47±0.06	8.28±0.27
SSD	99.99±0.02	99.99±0.01	94.48±0.12	0.66±0.13	8.20±0.25
SEMU	99.79±0.29	99.76±0.31	93.36±0.89	3.90±0.59	7.12±0.31
SFRON	95.87±0.21	99.95±0.01	93.04±0.13	11.68±0.35	4.07±0.18
SalUn	96.04±1.84	97.37±1.66	91.34±1.61	14.09±1.79	3.91±0.76
+ SSU	93.07±2.06	94.86±1.93	89.43±1.56	16.86±2.60	**3.21±0.68**

Table 3: Tiny ImageNet, 10% random data forgetting (Mean ± std over three seeds).

Methods	FA	RA	TA	MIA	Avg. Gap ↓
Retrain	59.42±0.51	99.98±0.00	59.90±0.04	73.97±0.69	0.00
FT	99.96±0.01	99.98±0.00	61.08±0.59	1.87±0.34	28.46±0.38
GA	99.98±0.01	99.98±0.01	60.79±0.54	1.39±0.45	28.51±0.44
IU	99.98±0.01	99.98±0.00	61.39±0.57	1.12±0.26	28.73±0.41
ℓ_1-sparse	99.96±0.02	99.98±0.00	60.42±0.40	3.36±0.86	27.92±0.28
SCRUB	99.97±0.02	99.98±0.00	61.47±0.49	1.54±0.21	28.64±0.38
SSD	99.97±0.00	99.98±0.00	61.43±0.52	1.04±0.20	28.76±0.43
SEMU	5.73±0.82	5.74±0.69	5.86±0.53	2.57±1.66	68.34±0.94
SFRON	98.83±0.33	99.98±0.00	60.67±0.22	23.50±4.11	22.66±0.83
SalUn	74.08±3.57	99.76±0.09	53.30±0.34	82.06±1.71	7.40±0.57
+ SSU	56.51±0.60	99.76±0.04	51.30±1.00	83.64±2.51	**5.35±0.22**

For more challenging settings, Tab. 2 shows the results for unlearning 50% of CIFAR-10 with the same baselines and setup. We see that SSU remains the best method, reducing the average gap to 3.21±0.68, improving that of SalUn by 17.9%. Tab. 3 moves to Tiny-ImageNet with 200 classes, where 10% of the training data is forgotten. The average gaps of all non-saliency baselines stay above 22, while SSU reaches 5.35 ± 0.22, improving the result of SalUn by 27.7%.

5.2 Class-wise Unlearning in Image Generation

Setup and evaluation. Following the setup in (Fan et al., 2024), we evaluate our method on class-wise unlearning for image generation using DDPM (Ho et al., 2020) on CIFAR-10 and Stable Diffusion V1.4 (SD) (Rombach et al., 2022) on Imagenette (Howard, 2019). We compare against four baselines: SA (Heng & Soh, 2023), ESD (Gandikota et al., 2023), SFRON (Huang et al., 2024), and SalUn (Fan et al., 2024). For DDPM, SVD is applied only to the convolution layers of the denoising U-Net, while for Stable Diffusion, SVD is applied only to the cross-attention layers of the U-Net. We evaluate unlearning quality using the FID (Seitzer, 2020) score to measure generation quality on retained classes, and forgetting set accuracy (FA) of unlearned classes via a pre-trained classifier. Implementation details are provided in Appx. B.2.

Results on DDPM. Tab. 4 presents class-wise unlearning results for DDPM on CIFAR-10 across five classes. While all methods achieve near-perfect forgetting, they differ significantly in generation quality. Baseline methods yield FID scores ranging from 18.44 to 29.04, indicating moderate image quality on retained classes. Our method achieves substantial improvements, with FID scores between 15.19 and 16.32 across all classes, which is a 25-30% reduction compared to baselines. Notably, SSU maintains perfect forgetting on all classes while delivering the best generation quality, with the most significant improvements on Automobile, Cat, and Truck. These results show that our approach can effectively unlearn target data while preserving and even improving the generative capabilities.

Table 4: Class-wise unlearning of image generation on CIFAR-10 with DDPM.

Method	CIFAR-10 Class-wise Unlearning									
	Automobile		Cat		Dog		Horse		Truck	
	FA↓	**FID↓**	**FA↓**	**FID↓**	**FA↓**	**FID↓**	**FA↓**	**FID↓**	**FA↓**	**FID↓**
SA	**0.00**	23.56	14.20	21.34	8.60	21.19	**0.00**	21.13	**0.00**	29.04
SFRON	**0.00**	20.70	7.40	18.44	0.20	18.89	**0.00**	19.93	**0.00**	20.61
SalUn	0.20	21.23	1.40	20.29	**0.00**	20.18	0.60	20.70	0.80	20.45
+ SSU	**0.00**	**15.46**	**1.00**	**16.01**	**0.00**	**15.19**	**0.00**	**15.60**	**0.00**	**16.32**

While the method requires SVD computation during training, we note that the method has minimal overhead. For U-Net architectures, gradients are reshaped to moderate-sized 2D matrices (typically $\leq 512 \times 512$) for SVD, which can be efficiently decomposed using optimized GPU implementations.



Figure 1: Qualitative results of unlearning and retaining images across Imagenette classes using Stable Diffusion. The top row shows generated samples for each class when it is targeted for unlearning, while the bottom row displays samples for the same classes when they are retained.

In practice, our method takes approximately 40 seconds per 100 training steps on one NVIDIA L40S, nearly identical to SalUn, while providing better unlearning through spectral filtering.

Results on SD. Tab. 5 shows class-wise unlearning performance on Stable Diffusion (Rombach et al., 2022) with the Imagenette dataset (Howard, 2019). We generate 300 images per class for the computation of both FA and FID. All methods achieve strong forgetting performance with average FA below 0.30%, confirming effective removal of target concepts. In terms of generation quality, ESD achieves an average FID of 1.71, while SalUn improves this to 1.55. Our method further advances the state-of-the-art with an average FID of 1.24, achieving the best overall performance. SSU demonstrates consistent improvements across most classes, with larger gains on Tench, French Horn, Gas Pump, and Golf Ball. Moreover, SSU achieves slightly better forgetting with an average FA of 0.20% compared to 0.30% for ESD, matching SalUn's 0.20%. These results validate the generalizability of our approach across different diffusion architectures.

Fig. 1 shows qualitative results for class-wise unlearning on Stable Diffusion with the Imagenette dataset. We observe that SSU effectively removes the model's ability to generate images corresponding to the unlearned classes, as evidenced by the lack of recognizable features in those classes, *e.g.*, a tench turns into a dog. At the same time, the generations of retained classes show high fidelity and remain semantically accurate. This demonstrates that our approach preserves generative quality for non-targeted concepts. These results visually verify the effectiveness of our method in selectively unlearning specific classes while maintaining overall model utility.

Table 5: Class-wise unlearning of image generation on Imagenette with SD.

Forget Class	**ESD**		**SalUn**		**+ SSU**	
	FA (↓)	FID (↓)	FA (↓)	FID (↓)	FA (↓)	FID (↓)
Tench	0.00	2.37	0.00	0.91	0.00	0.76
English Springer	0.00	1.70	0.00	0.92	0.00	1.17
Cassette Player	0.00	1.34	0.67	1.30	0.67	1.16
Chain Saw	0.00	1.48	0.00	1.38	0.00	1.55
Church	2.33	2.56	0.00	1.73	0.00	1.65
French Horn	0.00	1.61	0.00	1.36	0.00	1.14
Garbage Truck	0.00	1.57	0.00	1.01	0.00	1.00
Gas Pump	0.00	1.34	0.00	2.78	0.00	1.48
Golf Ball	0.67	1.35	1.33	2.44	1.33	1.12
Parachute	0.00	1.75	0.00	1.68	0.00	1.35
Average	0.30	1.71	**0.20**	1.55	**0.20**	**1.24**

5.3 Random Subset Unlearning in LLMs

Experimental setup and evaluation. Following the setup by Yuan et al. (2025), we evaluate SSU on LLM unlearning using the TOFU benchmark (Maini et al., 2024) with Llama2-7B (Touvron et al., 2023). The TOFU benchmark consists of three scenarios: forget01, forget05, and forget10, corresponding to forgetting 1%, 5%, and 10% of the training data, respectively. We consider two types of unlearning tasks: untargeted unlearning, including gradient-based methods (GA+GD, GA+KL), preference-based methods (NPO+GD, NPO+KL), and knowledge manipulation methods (ME+GD); and targeted unlearning, where we evaluate on preference-based methods (DPO+GD, DPO+KL) and knowledge manipulation methods (IDK+GD, IDK+AP). The baselines are adopted from Yuan et al. (2025). SSU is built on top of the best-performing baseline for each task.

As in prior work (Yuan et al., 2025), we evaluate performance using two aggregated metrics: Model Utility (MU) and Forget Efficacy (FE). These metrics combine multiple complementary signals, including ROUGE, prediction probability, truth ratio, token entropy, semantic similarity, and entailment score, to jointly assess lexical accuracy, semantic consistency, generation quality, and factual correctness. MU is computed on the retain set using the harmonic mean to measure overall utility preservation, while FE is computed on the forget set as one minus the arithmetic mean (excluding

Table 6: Results of different unlearning methods on the TOFU benchmark with Llama2-7B. MU and FE represent Model Utility and Forget Efficacy, respectively, and we indicate the best results in bold.

Dataset	Metric	Method										
		GA+GD	GA+KL	NPO+GD	NPO+KL	ME+GD	+ SSU	DPO+GD	DPO+KL	IDK+GD	IDK+AP	+ SSU
forget01	MU	0.6671	0.6385	0.6402	0.7404	0.7245	**0.7477**	0.7554	0.7601	0.6704	0.7579	**0.7769**
	FE	0.5935	0.6028	0.6137	0.4943	0.9156	**0.9599**	0.5260	0.3013	0.7700	0.7625	**0.7943**
	Avg.	0.6303	0.6206	0.6269	0.6174	0.8201	**0.8538**	0.6407	0.5307	0.7202	0.7602	**0.7856**
forget05	MU	0.2913	0.0000	0.5718	0.5470	**0.7521**	0.7277	0.0000	0.4536	0.0000	**0.7522**	0.7515
	FE	0.9135	0.8927	0.6967	0.6192	0.9262	**0.9407**	**0.8242**	0.7831	0.7948	0.7480	0.7608
	Avg.	0.6024	0.4463	0.6342	0.5831	**0.8391**	0.8342	0.4121	0.6183	0.3974	0.7501	**0.7562**
forget10	MU	0.5001	0.0000	0.5660	0.4904	0.7312	**0.7480**	0.0000	0.0000	0.0527	0.7444	**0.7463**
	FE	0.9449	0.9484	0.7530	0.7449	0.9505	**0.9557**	0.8043	**0.8346**	0.7602	0.7432	0.7487
	Avg.	0.7225	0.4742	0.6595	0.6177	0.8409	**0.8519**	0.4022	0.4173	0.4065	0.7438	**0.7475**

Table 7: Ablating the top singular value ratio when randomly unlearning 10% of the CIFAR-10 data.

Keep ratio	FA	RA	TA	MIA	Avg. Gap ↓
0.1	98.27±0.64	99.53±0.31	93.22±0.45	11.10±0.93	1.86±0.11
0.3	97.45±0.57	99.34±0.25	93.04±0.43	12.81±0.37	1.32±0.24
0.5	97.00±0.61	99.29±0.26	93.08±0.38	13.34±0.99	1.23±0.25
0.7	97.06±0.56	99.38±0.21	93.26±0.38	13.38±0.67	**1.10±0.20**
0.9	97.21±0.61	99.54±0.15	93.29±0.40	13.81±0.68	1.15±0.21

Table 8: Ablating the modules to which SVD is applied for class-wise unlearning with Stable Diffusion.

	Conv	MLP	Full	XAttn
FA ↓	0.17	0.23	0.23	0.20
FID ↓	1.29	1.29	1.28	1.24

token entropy) to quantify the strength of forgetting. We further report their average as a single summary score. Implementation details are provided in Appx. B.3.

Results. Tab. 6 presents random subset unlearning on Llama2-7B across three forgetting scenarios. We apply SSU on top of the best baseline methods for each unlearning type. For untargeted unlearning, ME+GD emerges as the strongest baseline, significantly outperforming gradient-based (GA) and preference-based (NPO) alternatives. Applying SSU on top of ME+GD further enhances performance. SSU improves model utility for forget01 from 0.7245 to 0.7477 and forget10 from 0.7312 to 0.7480 while simultaneously improving forget efficacy across all scenarios, demonstrating superior balance between preserving model capabilities and effective unlearning.

For targeted unlearning, IDK+AP serves as the best baseline with consistent performance. Building upon IDK+AP, SSU achieves the best results with average scores of 0.7856, 0.7562, and 0.7475 for the three scenarios. The improvements are particularly noticeable for forget01, where SSU enhances both model utility and forget efficacy. These results demonstrate that SSU generalizes effectively to large language models, consistently improving state-of-the-art unlearning methods for both untargeted and targeted scenarios across different forgetting ratios.

5.4 Ablation Studies

Effect of the keep ratio. To investigate the impact of the keep ratio on unlearning performance, we conduct an ablation study by varying the ratio parameter in our method for random subset unlearning on CIFAR-10. As in Tab. 7, the method is relatively robust with respect to the keep ratio except in extreme cases such as 0.1. Our default setting of ratio $= 0.7$ achieves the best balance across different metrics.

Impact of SVD-applied weight selection. We ablate the effect of applying SVD to different subsets of weights in the U-Net architecture. Tab. 8 reports results on Stable Diffusion with Imagenette, comparing SVD applied to all layers, only cross-attention layers, only convolutional layers, and only MLP layers. We observe that restricting SVD to cross-attention layers achieves the best trade-off between forgetting and generation quality. In contrast, applying SVD to all layers slightly degrades generation quality, likely because it perturbs low-level visual features and global denoising dynamics that are less directly related to semantic concepts. This suggests that effective unlearning requires targeted modification of concept-carrying components. These results motivate our design choice of selectively applying SVD to cross-attention layers, which maximizes unlearning efficacy while preserving overall generative fidelity.

6 Related Work

Forgetting-retention trade-off and spectral structure. The forgetting-retention trade-off is a central consideration in machine unlearning (Sekhari et al., 2021). As updates that remove forget-set influence can degrade retained performance, unlearning inherits the same interference phenomenon

studied as catastrophic forgetting in continual learning (Li et al., 2019). Many methods, therefore, use SVD-based projection to project the unlearning update onto a subspace orthogonal to an estimated retain subspace (Wang et al., 2025b; Lin et al., 2024; Biswas et al., 2025; Fang et al., 2025; Chen et al., 2024), following parameter-isolation strategies from continual learning (Farajtabar et al., 2020; Bennani et al., 2020; Saha et al., 2021).

SSU differs from these approaches in both object and mechanism. It operates directly on the singular basis of the unlearning gradient and suppresses weak spectral components. This yields an optimizer-agnostic mechanism that targets forget-retain interaction directly, rather than constraining updates solely through an estimated retain subspace. Related work also studies forget-retain conflict via Euclidean gradient alignment between the forgetting and retaining gradients (Patel & Qiu, 2025; Wang et al., 2025a; Asif & Amiri, 2026), whereas SSU analyzes their interaction in the spectral basis of the unlearning gradient and uses the resulting spectrum as a structured proxy for interference. In addition, Sendera et al. (2025) applies SVD-based low-rank parameterization for efficient updates, while SSU uses singular values as a saliency signal for thresholding.

Saliency-based MU. Existing saliency-based unlearning selectively updates a subset of parameters deemed most responsible for forget-set behavior. Motivated by evidence that model sparsity improves unlearning (Savani et al., 2025), SalUn (Fan et al., 2024) computes a coordinate-wise saliency map from the forget-set gradient evaluated with the pretrained model, and thresholds low-magnitude coordinates to sparsify unlearning updates. Building on SalUn, Ding et al. (2025) further argues via a fine-tuning analysis that saliency should instead be derived from retained data to better preserve overlapping features and mitigate the forgetting-retention trade-off.

Unlike coordinate masking, SSU defines spectral saliency using the singular values of the unlearning gradient, selecting dominant matrix-valued update directions that reflect signals from both forgetting and retention. Huang et al. (2024) develops a unified view of saliency by decomposing the unlearning objective into a forgetting term, a retention term, and an explicit weight-saliency matrix that modulates the unlearning direction. Overall, while saliency-driven sparsification is empirically effective, it remains largely a heuristic. Our analysis offers a complementary theoretical explanation for why saliency-based masking can improve the forgetting-retention trade-off.

General MU has been explored through a range of optimization-based approaches, including gradient ascent methods that increase the forget-set loss (Thudi et al., 2022; Neel et al., 2021), influence-function-based approximations to leave-one-out retraining (Izzo et al., 2021; Koh & Liang, 2017), and Fisher-information-based selective updates (Golatkar et al., 2020; Becker & Liebig, 2022). More recently, practical approximate unlearning has increasingly relied on fine-tuning (Warnecke et al., 2021; Golatkar et al., 2020; Kurmanji et al., 2023), particularly for large models where full retraining is prohibitive. These techniques have been developed across modalities, including image classification models (Jia et al., 2023; Fan et al., 2024), generative image models such as diffusion models (Gandikota et al., 2023; Zhang et al., 2024a; Heng & Soh, 2023; Wu et al., 2025), and LLMs (Zhang et al., 2024b; Yuan et al., 2025; Liu et al., 2025).

Despite their differences, many existing methods can be expressed under a common template that jointly optimizes a forgetting loss and a retention or utility-preserving loss (Fan et al., 2024; Yuan et al., 2025; Zhong et al., 2025). SSU can be used as a plug-in enhancement to methods of this structure, as we have demonstrated in Sec. 5.

7 Conclusion

Motivated by a spectral view of gradient-based optimization, we proposed Spectral Saliency Unlearning (SSU), which applies singular-value thresholding to suppress weak spectral components of the unlearning gradient and can be used as a drop-in enhancement for gradient-based unlearning pipelines. We provided a theoretical justification for SSU from the perspective of the forgetting-retention trade-off, characterizing how weak directions relate to forget-retain interference and how suppressing them can mitigate utility degradation. We further extended this perspective to coordinate-wise thresholding, offering a nontrivial explanation for the empirical effectiveness of SalUn-style masking. Empirically, we evaluated SSU across diverse unlearning settings, including image classification, diffusion-model unlearning, and large language model unlearning. Across these settings, SSU consistently demonstrates its effectiveness in unlearning and improves the forgetting-utility trade-off.

AI USE STATEMENT

In this work, we used generative AI tools to edit a research paper to improve readability. We have not used generative AI tools to help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, design or provide feedback on research methodology or experiments, implement methods, support qualitative and thematic data analysis, or interpret results. Generating synthetic data sets, proposing or refining hypotheses, assisting with translation, cleaning and reformatting datasets are not applicable to this work. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Sadia Asif and Mohammad Mohammadi Amiri. OFMU: Optimization-driven framework for machine unlearning. In *Proc. ICLR*, 2026. 9

Alexander Becker and Thomas Liebig. Evaluating machine unlearning via epistemic uncertainty. *arXiv preprint arXiv:2208.10836*, 2022. 9

Mehdi Abbana Bennani, Thang Doan, and Masashi Sugiyama. Generalisation guarantees for continual learning with orthogonal gradient descent. *arXiv preprint arXiv:2006.11942*, 2020. 9

Jeremy Bernstein, Yu-Xiang Wang, Kamyar Azizzadenesheli, and Animashree Anandkumar. SignSGD: Compressed optimisation for non-convex problems. In *Proc. ICML*, 2018. 1

Shristi Das Biswas, Arani Roy, and Kaushik Roy. CURE: Concept unlearning via orthogonal representation editing in diffusion models. In *Proc. NeurIPS*, 2025. 9

Jian-Feng Cai, Emmanuel J Candès, and Zuowei Shen. A singular value thresholding algorithm for matrix completion. *SIOPT*, 2010. 3

Yinzhi Cao and Junfeng Yang. Towards making systems forget with machine unlearning. In *IEEE symposium on security and privacy*, 2015. 1

Huiqiang Chen, Tianqing Zhu, Xin Yu, and Wanlei Zhou. Machine unlearning via null space calibration. In *Proc. IJCAI*, 2024. 9

Lizhang Chen, Jonathan Li, and Qiang Liu. Muon optimizes under spectral norm constraints. *arXiv preprint arXiv:2506.15054*, 2025. 2

Meng Ding, Rohan Sharma, Changyou Chen, Jinhui Xu, and Kaiyi Ji. Understanding fine-tuning in approximate unlearning: A theoretical perspective. *TMLR*, 2025. 9

European Union. Regulation (EU) 2016/679 of the european parliament and of the council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (general data protection regulation), 2016. URL https://eur-lex.europa.eu/eli/reg/2016/679/oj. OJ L 119, 4.5.2016, p. 1–88. 1

Chongyu Fan, Jiancheng Liu, Yihua Zhang, Eric Wong, Dennis Wei, and Sijia Liu. Salun: Empowering machine unlearning via gradient-based weight saliency in both image classification and generation. In *Proc. ICLR*, 2024. 1, 2, 3, 5, 6, 9

Junfeng Fang, Houcheng Jiang, Kun Wang, Yunshan Ma, Jie Shi, Xiang Wang, Xiangnan He, and Tat-Seng Chua. AlphaEdit: Null-space constrained model editing for language models. In *Proc. ICLR*, 2025. 9

Mehrdad Farajtabar, Navid Azizan, Alex Mott, and Ang Li. Orthogonal gradient descent for continual learning. In *Proc. AISTATS*, 2020. 9

Jack Foster, Stefan Schoepf, and Alexandra Brintrup. Fast machine unlearning without retraining through selective synaptic dampening. In *Proc. AAAI*, 2024. 5

Rohit Gandikota, Joanna Materzynska, Jaden Fiotto-Kaufman, and David Bau. Erasing concepts from diffusion models. In *Proc. ICCV*, 2023. 2, 6, 9

Rohit Gandikota, Sheridan Feucht, Samuel Marks, and David Bau. Erasing conceptual knowledge from language models. 2025. 2

Aditya Golatkar, Alessandro Achille, and Stefano Soatto. Eternal sunshine of the spotless net: Selective forgetting in deep networks. In *Proc. CVPR*, 2020. 9

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. CVPR*, 2016. 1, 5

Alvin Heng and Harold Soh. Selective amnesia: A continual learning approach to forgetting in deep generative models. In *Proc. NeurIPS*, 2023. 6, 9

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Proc. NeurIPS*, 2020. 6

Jeremy Howard. Imagenette: A smaller subset of 10 easily classified classes from imagenet, March 2019. URL https://github.com/fastai/imagenette. 6, 7

Zhehao Huang, Xinwen Cheng, JingHao Zheng, Haoran Wang, Zhengbao He, Tao Li, and Xiaolin Huang. Unified gradient-based machine unlearning with remain geometry enhancement. In *Proc. NeurIPS*, 2024. 5, 6, 9

Zachary Izzo, Mary Anne Smart, Kamalika Chaudhuri, and James Zou. Approximate data deletion from machine learning models. In *Proc. AISTATS*, 2021. 1, 5, 9

Jinghan Jia, Jiancheng Liu, Parikshit Ram, Yuguang Yao, Gaowen Liu, Yang Liu, Pranay Sharma, and Sijia Liu. Model sparsification can simplify machine unlearning. In *Proc. NeurIPS*, 2023. 5, 9

Keller Jordan, Yuchen Jin, Vlado Boza, You Jiacheng, Franz Cecista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks. *URL https://kellerjordan.github.io/posts/muon/*, 2024. 1, 2

Michael Kleinman, Alessandro Achille, and Stefano Soatto. Critical learning periods emerge even in deep linear networks. In *Proc. ICLR*, 2024. 1, 3

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *Proc. ICML*, 2017. 9

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf. 5

Nupur Kumari, Bingliang Zhang, Sheng-Yu Wang, Eli Shechtman, Richard Zhang, and Jun-Yan Zhu. Ablating concepts in text-to-image diffusion models. In *Proc. ICCV*, 2023. 2

Meghdad Kurmanji, Peter Triantafillou, Jamie Hayes, and Eleni Triantafillou. Towards unbounded machine unlearning. In *Proc. NeurIPS*, 2023. 5, 9

Andrew K. Lampinen and Surya Ganguli. An analytic theory of generalization dynamics and transfer learning in deep linear networks. In *Proc. ICLR*, 2019. 1, 3

Ya Le and Xuan Yang. Tiny ImageNet visual recognition challenge. CS231n course project report, Stanford University, 2015. URL https://cs231n.stanford.edu/reports/2015/pdfs/yle_project.pdf. 5

Xilai Li, Yingbo Zhou, Tianfu Wu, Richard Socher, and Caiming Xiong. Learn to grow: A continual structure learning framework for overcoming catastrophic forgetting. In *Proc. ICML*, 2019. 9

Shen Lin, Xiaoyu Zhang, Willy Susilo, Xiaofeng Chen, and Jun Liu. GDR-GMA: Machine unlearning via direction-rectified and magnitude-adjusted gradients. In *ACM Multimedia*, 2024. 9

Sijia Liu, Yuanshun Yao, Jinghan Jia, Stephen Casper, Nathalie Baracaldo, Peter Hase, Yuguang Yao, Chris Yuhao Liu, Xiaojun Xu, Hang Li, et al. Rethinking machine unlearning for large language models. *Nature Machine Intelligence*, 2025. 9

Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary C Lipton, and J Zico Kolter. Tofu: A task of fictitious unlearning for llms. *arXiv preprint arXiv:2401.06121*, 2024. 7

Seth Neel, Aaron Roth, and Saeed Sharifi-Malvajerdi. Descent-to-delete: Gradient-based methods for machine unlearning. In *Proc. ALT*, 2021. 2, 9

Gaurav Patel and Qiang Qiu. Learning to unlearn while retaining: Combating gradient conflicts in machine unlearning. In *Proc. CVPR*, 2025. 9

Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In *Proc. ICML*, 2019. 1, 3

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proc. CVPR*, 2022. 6, 7

Gobinda Saha, Isha Garg, and Kaushik Roy. Gradient projection memory for continual learning. In *Proc. ICLR*, 2021. 9

Yash Savani, Asher Trockman, Zhili Feng, Yixuan Even Xu, Avi Schwarzschild, Alexander Robey, Marc Finzi, and J Zico Kolter. Antidistillation sampling. *Proc. NeurIPS*, 2025. 9

Maximilian Seitzer. pytorch-fid: FID Score for PyTorch. https://github.com/mseitzer/pytorch-fid, August 2020. Version 0.3.0. 6

Ayush Sekhari, Jayadev Acharya, Gautam Kamath, and Ananda Theertha Suresh. Remember what you want to forget: Algorithms for machine unlearning. In *Proc. NeurIPS*, 2021. 8

Marcin Sendera, Łukasz Struski, Kamil Książek, Kryspin Musiol, Jacek Tabor, and Dawid Damian Rymarczyk. SEMU: Singular value decomposition for efficient machine unlearning. In *Proc. ICML*, 2025. 5, 9

Anvith Thudi, Gabriel Deza, Varun Chandrasekaran, and Nicolas Papernot. Unrolling SGD: Understanding factors influencing machine unlearning. In *Proc. EuroS&P*, 2022. 1, 2, 5, 9

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 7

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proc. NeurIPS*, 2017. 1

Qizhou Wang, Jin Peng Zhou, Zhanke Zhou, Saebyeol Shin, Bo Han, and Kilian Q Weinberger. Rethinking LLM unlearning objectives: A gradient perspective and go beyond. In *Proc. ICLR*, 2025a. 9

Yue Wang, Qizhou Wang, Feng Liu, Wei Huang, Yali Du, Xiaojiang Du, and Bo Han. GRU: Mitigating the trade-off between unlearning and retention for LLMs. In *Proc. ICML*, 2025b. 9

Alexander Warnecke, Lukas Pirch, Christian Wressnegger, and Konrad Rieck. Machine unlearning of features and labels. In *Proc. NDSS*, 2021. 5, 9

Jing Wu, Trung Le, Munawar Hayat, and Mehrtash Harandi. Erasing undesirable influence in diffusion models. In *Proc. CVPR*, 2025. 9

Xiaojian Yuan, Tianyu Pang, Chao Du, Kejiang Chen, Weiming Zhang, and Min Lin. A closer look at machine unlearning for large language models. In *Proc. ICLR*, 2025. 2, 7, 9

Gong Zhang, Kai Wang, Xingqian Xu, Zhangyang Wang, and Humphrey Shi. Forget-me-not: Learning to forget in text-to-image diffusion models. In *Proc. CVPR*, 2024a. 9

Ruiqi Zhang, Licong Lin, Yu Bai, and Song Mei. Negative preference optimization: From catastrophic collapse to effective unlearning. In *Proc. COLM*, 2024b. 9

Xuyang Zhong, Haochen Luo, and Chen Liu. Dualoptim: Enhancing efficacy and stability in machine unlearning with dual optimizers. In *Proc. NeurIPS*, 2025. 9

Appendix

The appendix is organized as follows:

- In Sec. A, we provide the formal assumptions, statements, and complete proofs for the propositions stated in the main paper.
- In Sec. B, we provide additional experiment details. The code will be open-sourced upon the acceptance of this paper.
- In Sec. C, we provide additional image generation results after unlearning with SSU.

A Proof of the Theoretical Justification

A.1 Formal Statement and Proof of Proposition 4.2

For the unlearning objective $\mathcal{L}_u = \mathcal{L}_f + \mathcal{L}_r$ composed of a forgetting loss $\mathcal{L}_f$ and a retaining loss $\mathcal{L}_r$, we have for their gradients $\boldsymbol{G}_u = \boldsymbol{G}_f + \boldsymbol{G}_r$. The SVD of $\boldsymbol{G}_u$ is given by $\boldsymbol{G}_u = \boldsymbol{U}\,\mathrm{diag}\,(\boldsymbol{\sigma})\,\boldsymbol{V} = \sum_{i=1}^{m} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^\top$. Let $r_e = \mathrm{rank}\,(G_u)$. Then we have $\forall i \in [r_e]$, $\sigma_i > 0$ and $\forall i \in \{r_e + 1, \cdots, m\}$, $\sigma_i = 0$. In practice, in the context of effective rank, we let the former represent significant non-zero singular values and the latter include those that are approximately zero. We have compact SVD $\boldsymbol{G}_u = \boldsymbol{U}_e \mathrm{diag}\,(\boldsymbol{\sigma}_e)\,\boldsymbol{V}_e^\top$ where $\boldsymbol{U}_e \in \mathbb{R}^{m \times r_e}$ and $\boldsymbol{V}_e \in \mathbb{R}^{n \times r_e}$.

Assumption A.1 (Smoothness)**.** For constants $\beta_f, \beta_r > 0$, the forgetting loss $\mathcal{L}_f$ and the retaining loss $\mathcal{L}_r$ are smooth, i.e., for $\boldsymbol{W}$, $\boldsymbol{W}'$,

$$\left|\mathcal{L}_f\left(\boldsymbol{W}'\right) - \mathcal{L}_f\left(\boldsymbol{W}\right) - \left\langle \boldsymbol{G}_f\left(\boldsymbol{W}\right), \boldsymbol{W}' - \boldsymbol{W}\right\rangle\right| \leq \frac{\beta_f}{2}\left\|\boldsymbol{W}' - \boldsymbol{W}\right\|^2,$$

$$\left|\mathcal{L}_r\left(\boldsymbol{W}'\right) - \mathcal{L}_r\left(\boldsymbol{W}\right) - \left\langle \boldsymbol{G}_r\left(\boldsymbol{W}\right), \boldsymbol{W}' - \boldsymbol{W}\right\rangle\right| \leq \frac{\beta_r}{2}\left\|\boldsymbol{W}' - \boldsymbol{W}\right\|^2.$$

Definition 4.1. (Forget-Retain Alignment) We define the alignment between the forgetting and retention directions of $\boldsymbol{v}_i$ as

$$a\left(\boldsymbol{v}_i\right) = \frac{\left\langle \boldsymbol{G}_f \boldsymbol{v}_i, \boldsymbol{G}_r \boldsymbol{v}_i\right\rangle}{\left\|\boldsymbol{G}_f \boldsymbol{v}_i\right\|_2 \left\|\boldsymbol{G}_r \boldsymbol{v}_i\right\|_2}. \tag{7}$$

Lemma A.2. *If* $a\left(\boldsymbol{v}_i\right) < -\frac{\min\{\|\boldsymbol{G}_f \boldsymbol{v}_i\|, \|\boldsymbol{G}_r \boldsymbol{v}_i\|\}}{\max\{\|\boldsymbol{G}_f \boldsymbol{v}_i\|, \|\boldsymbol{G}_r \boldsymbol{v}_i\|\}}$*, then for* $f_i = \left\langle \boldsymbol{G}_f, \boldsymbol{u}_i \boldsymbol{v}_i^\top \right\rangle$, $r_i = \left\langle \boldsymbol{G}_r, \boldsymbol{u}_i \boldsymbol{v}_i^\top \right\rangle$*, we have* $f_i r_i < 0$.

Proof. By definition,

$$f_i = \left\langle \boldsymbol{G}_f, \boldsymbol{u}_i \boldsymbol{v}_i^\top \right\rangle = \boldsymbol{u}_i^\top \boldsymbol{G}_f \boldsymbol{v}_i, \qquad r_i = \left\langle \boldsymbol{G}_r, \boldsymbol{u}_i \boldsymbol{v}_i^\top \right\rangle = \boldsymbol{u}_i^\top \boldsymbol{G}_r \boldsymbol{v}_i$$

Since $\boldsymbol{G}_f \boldsymbol{v}_i + \boldsymbol{G}_r \boldsymbol{v}_i = \boldsymbol{G}_u \boldsymbol{v}_i = \sum_{j=1}^{r_e} \sigma_j \boldsymbol{u}_j \boldsymbol{v}_j^\top \boldsymbol{v}_i = \sigma_i \boldsymbol{u}_i$, we know $\boldsymbol{u}_i = \frac{\boldsymbol{G}_f \boldsymbol{v}_i + \boldsymbol{G}_r \boldsymbol{v}_i}{\sigma_i}$. Therefore,

$$f_i = \boldsymbol{u}_i^\top \boldsymbol{G}_f \boldsymbol{v}_i = \frac{\left(\boldsymbol{G}_f \boldsymbol{v}_i + \boldsymbol{G}_r \boldsymbol{v}_i\right)^\top \boldsymbol{G}_f \boldsymbol{v}_i}{\sigma_i} = \frac{\left\|\boldsymbol{G}_f \boldsymbol{v}_i\right\|_2^2 + \left\langle \boldsymbol{G}_r \boldsymbol{v}_i, \boldsymbol{G}_f \boldsymbol{v}_i\right\rangle}{\sigma_i}.$$

Similarly, $r_i = \frac{\|\boldsymbol{G}_r \boldsymbol{v}_i\|_2^2 + \langle \boldsymbol{G}_r \boldsymbol{v}_i, \boldsymbol{G}_f \boldsymbol{v}_i\rangle}{\sigma_i}$.

If $a\left(\boldsymbol{v}_i\right) = \frac{\langle \boldsymbol{G}_f \boldsymbol{v}_i, \boldsymbol{G}_r \boldsymbol{v}_i\rangle}{\|\boldsymbol{G}_f \boldsymbol{v}_i\|_2 \|\boldsymbol{G}_r \boldsymbol{v}_i\|_2} < -\frac{\min\{\|\boldsymbol{G}_f \boldsymbol{v}_i\|, \|\boldsymbol{G}_r \boldsymbol{v}_i\|\}}{\max\{\|\boldsymbol{G}_f \boldsymbol{v}_i\|, \|\boldsymbol{G}_r \boldsymbol{v}_i\|\}}$, without loss of generality, we assume $\|\boldsymbol{G}_f \boldsymbol{v}_i\| > \|\boldsymbol{G}_r \boldsymbol{v}_i\|$, then $\frac{\langle \boldsymbol{G}_f \boldsymbol{v}_i, \boldsymbol{G}_r \boldsymbol{v}_i\rangle}{\|\boldsymbol{G}_f \boldsymbol{v}_i\|_2 \|\boldsymbol{G}_r \boldsymbol{v}_i\|_2} < -\frac{\|\boldsymbol{G}_r \boldsymbol{v}_i\|}{\|\boldsymbol{G}_f \boldsymbol{v}_i\|}$ yields $\langle \boldsymbol{G}_f \boldsymbol{v}_i, \boldsymbol{G}_r \boldsymbol{v}_i\rangle < -\|\boldsymbol{G}_r \boldsymbol{v}_i\|^2$. Also, we have by Cauchy-Schwarz inequality,

$$\left\langle \boldsymbol{G}_f \boldsymbol{v}_i, \boldsymbol{G}_r \boldsymbol{v}_i\right\rangle \geq -\left\|\boldsymbol{G}_f \boldsymbol{v}_i\right\|_2 \left\|\boldsymbol{G}_r \boldsymbol{v}_i\right\|_2 > -\left\|\boldsymbol{G}_f \boldsymbol{v}_i\right\|_2^2.$$

Therefore, we know $\left\langle \boldsymbol{G}_f \boldsymbol{v}_i, \boldsymbol{G}_r \boldsymbol{v}_i\right\rangle + \left\|\boldsymbol{G}_r \boldsymbol{v}_i\right\|^2 < 0$ and $\left\langle \boldsymbol{G}_f \boldsymbol{v}_i, \boldsymbol{G}_r \boldsymbol{v}_i\right\rangle + \left\|\boldsymbol{G}_f \boldsymbol{v}_i\right\|_2^2 > 0$. As a result, we have

$$f_i r_i = \frac{\left(\left\|\boldsymbol{G}_f \boldsymbol{v}_i\right\|_2^2 + \left\langle \boldsymbol{G}_r \boldsymbol{v}_i, \boldsymbol{G}_f \boldsymbol{v}_i\right\rangle\right)\left(\left\|\boldsymbol{G}_r \boldsymbol{v}_i\right\|_2^2 + \left\langle \boldsymbol{G}_r \boldsymbol{v}_i, \boldsymbol{G}_f \boldsymbol{v}_i\right\rangle\right)}{\sigma_i^2} < 0.$$

□

Lemma A.3. *For gradient $\boldsymbol{G}_u = \eta \sum_{i=1}^{r_e} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^\top$, the update in direction $\boldsymbol{v}_i$ is given by $\boldsymbol{W}' = \boldsymbol{W} - \eta\sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^\top$. For $f_i = \langle \boldsymbol{G}_f, \boldsymbol{u}_i \boldsymbol{v}_i^\top \rangle$, $r_i = \langle \boldsymbol{G}_r, \boldsymbol{u}_i \boldsymbol{v}_i^\top \rangle$, $\eta < \min_i \left\{ \frac{2|f_i|}{\beta_f \sigma_i}, \frac{2|r_i|}{\beta_r \sigma_i} \right\}$, forgetting loss $\mathcal{L}_f$, and retaining loss $\mathcal{L}_r$, we have*

(a) *$\mathcal{L}_f(\boldsymbol{W}') - \mathcal{L}_f(\boldsymbol{W}) < 0$ implies $f_i > 0$, and $f_i < 0$ implies $\mathcal{L}_f(\boldsymbol{W}') - \mathcal{L}_f(\boldsymbol{W}) > 0$.*

(b) *$\mathcal{L}_r(\boldsymbol{W}') - \mathcal{L}_r(\boldsymbol{W}) < 0$ implies $r_i > 0$, and $r_i < 0$ implies $\mathcal{L}_r(\boldsymbol{W}') - \mathcal{L}_r(\boldsymbol{W}) > 0$.*

Proof. (a) We prove by contradiction. Suppose when $\mathcal{L}_f(\boldsymbol{W}') - \mathcal{L}_f(\boldsymbol{W}) < 0$, $f_i \le 0$. For the update $\boldsymbol{W}' = \boldsymbol{W} + \Delta\boldsymbol{W}$ where $\Delta\boldsymbol{W} = -\eta\sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^\top$, we have $\|\Delta\boldsymbol{W}\| = \eta\sigma_i$. By Assumption A.1, we have

$$\begin{aligned}
\mathcal{L}_f(\boldsymbol{W}') - \mathcal{L}_f(\boldsymbol{W}) &\ge \langle \boldsymbol{G}_f, \Delta\boldsymbol{W} \rangle - \frac{\beta_f}{2} \|\Delta\boldsymbol{W}\|^2 \\
&= -\eta\sigma_i \langle \boldsymbol{G}_f, \boldsymbol{u}_i \boldsymbol{v}_i^\top \rangle - \frac{\beta_f}{2} \eta^2 \sigma_i^2 \\
&= -\eta\sigma_i f_i - \frac{\beta_f}{2} \eta^2 \sigma_i^2 \\
&> 0,
\end{aligned}$$

where the last inequality follows from $\eta < \frac{2|f_i|}{\beta_f \sigma_i}$. This contradicts the condition $\mathcal{L}_f(\boldsymbol{W}') - \mathcal{L}_f(\boldsymbol{W}) < 0$. Therefore, when $\mathcal{L}_f(\boldsymbol{W}') - \mathcal{L}_f(\boldsymbol{W}) < 0$, we must have $f_i > 0$. In addition, from the derivation above, we know $f_i = \langle \boldsymbol{G}_f, \boldsymbol{u}_i \boldsymbol{v}_i^\top \rangle \le 0$ implies $\mathcal{L}_f(\boldsymbol{W}') - \mathcal{L}_f(\boldsymbol{W}) > 0$.

(b) Suppose when $\mathcal{L}_r(\boldsymbol{W}') - \mathcal{L}_r(\boldsymbol{W}) < 0$, $r_i \le 0$. By Assumption A.1, we have

$$\begin{aligned}
\mathcal{L}_r(\boldsymbol{W}') - \mathcal{L}_r(\boldsymbol{W}) &\ge \langle \boldsymbol{G}_r, \Delta\boldsymbol{W} \rangle - \frac{\beta_r}{2} \|\Delta\boldsymbol{W}\|^2 \\
&= -\eta\sigma_i \langle \boldsymbol{G}_r, \boldsymbol{u}_i \boldsymbol{v}_i^\top \rangle - \frac{\beta_r}{2} \eta^2 \sigma_i^2 \\
&= -\eta\sigma_i r_i - \frac{\beta_r}{2} \eta^2 \sigma_i^2 \\
&> 0,
\end{aligned}$$

where the last inequality follows from $\eta < \frac{2|r_i|}{\beta_r \sigma_i}$. This contradicts the condition $\mathcal{L}_r(\boldsymbol{W}') - \mathcal{L}_r(\boldsymbol{W}) < 0$. Therefore, when $\mathcal{L}_r(\boldsymbol{W}') - \mathcal{L}_r(\boldsymbol{W}) < 0$, we must have $r_i > 0$. In addition, from the derivation above, we know $r_i = \langle \boldsymbol{G}_r, \boldsymbol{u}_i \boldsymbol{v}_i^\top \rangle \le 0$ implies $\mathcal{L}_r(\boldsymbol{W}') - \mathcal{L}_r(\boldsymbol{W}) > 0$. □

Proposition 4.2. *For gradient of the unlearning objective $\boldsymbol{G}_u = \eta \sum_{i=1}^{r_e} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^\top$, the update in direction $\boldsymbol{v}_i$ is given by $\boldsymbol{W}' = \boldsymbol{W} - \eta\sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^\top$. Under Assumption A.1, if $a(\boldsymbol{v}_i) < -\frac{\min\{\|\boldsymbol{G}_f \boldsymbol{v}_i\|, \|\boldsymbol{G}_r \boldsymbol{v}_i\|\}}{\max\{\|\boldsymbol{G}_f \boldsymbol{v}_i\|, \|\boldsymbol{G}_r \boldsymbol{v}_i\|\}}$, and $\eta < \min_i \left\{ \frac{2|f_i|}{\beta_f \sigma_i}, \frac{2|r_i|}{\beta_r \sigma_i} \right\}$ for $f_i = \langle \boldsymbol{G}_f, \boldsymbol{u}_i \boldsymbol{v}_i^\top \rangle$ and $r_i = \langle \boldsymbol{G}_r, \boldsymbol{u}_i \boldsymbol{v}_i^\top \rangle$, then we have* **(a)** *if $\Delta\mathcal{L}_f = \mathcal{L}_f(\boldsymbol{W}') - \mathcal{L}_f(\boldsymbol{W}) < 0$ then $\Delta\mathcal{L}_r = \mathcal{L}_r(\boldsymbol{W}') - \mathcal{L}_r(\boldsymbol{W}) > 0$;* **(b)** *if $\Delta\mathcal{L}_r < 0$, then $\Delta\mathcal{L}_f > 0$. That is, along any singular direction $\boldsymbol{v}_i$ where the retain and forget gradients are in significant conflict, any update that improves one objective must necessarily worsen the other.*

Proof. (a) If the update makes progress in forgetting, that is, $\Delta\mathcal{L}_f = \mathcal{L}_f(\boldsymbol{W}') - \mathcal{L}_f(\boldsymbol{W}) < 0$, then by Lemma A.3 (a), we know $f_i > 0$. Also, from Lemma A.2, we know that if the retain and forget gradients are in significant conflict in the effective space, i.e., $a(\boldsymbol{v}_i) < -\frac{\min\{\|\boldsymbol{G}_f \boldsymbol{v}_i\|, \|\boldsymbol{G}_r \boldsymbol{v}_i\|\}}{\max\{\|\boldsymbol{G}_f \boldsymbol{v}_i\|, \|\boldsymbol{G}_r \boldsymbol{v}_i\|\}}$, we have $f_i r_i < 0$. Since $f_i > 0$, we must have $r_i < 0$. Then by Lemma A.3 (b), we have $\mathcal{L}_r(\boldsymbol{W}') - \mathcal{L}_r(\boldsymbol{W}) > 0$, meaning that the effort for retention is damaged.

(b) If the update makes progress in retention, that is, $\Delta\mathcal{L}_r = \mathcal{L}_r(\boldsymbol{W}') - \mathcal{L}_r(\boldsymbol{W}) < 0$, then by Lemma A.3 (b), we know $r_i > 0$. Again, by Lemma A.2, we know if $a(\boldsymbol{v}_i) < -\frac{\min\{\|\boldsymbol{G}_f \boldsymbol{v}_i\|, \|\boldsymbol{G}_r \boldsymbol{v}_i\|\}}{\max\{\|\boldsymbol{G}_f \boldsymbol{v}_i\|, \|\boldsymbol{G}_r \boldsymbol{v}_i\|\}}$, we have $f_i r_i < 0$. Since $r_i > 0$, we must have $f_i < 0$. Then by Lemma A.3 (a), we have $\mathcal{L}_f(\boldsymbol{W}') - \mathcal{L}_f(\boldsymbol{W}) > 0$, meaning that unlearning is not making progress in forgetting. □

A.2 Formal Statement and Proof of Proposition 4.3

For compact SVD $\boldsymbol{G}_u = \boldsymbol{U}_e \mathrm{diag}(\boldsymbol{\sigma}_e) \boldsymbol{V}_e^\top$ where $\boldsymbol{U}_e \in \mathbb{R}^{m \times r_e}$ and $\boldsymbol{V}_e \in \mathbb{R}^{n \times r_e}$, we call the space spanned by $\boldsymbol{V}_e$ the effective subspace, $\mathcal{E} = \mathrm{span}(\boldsymbol{V}_e)$, and the corresponding projection $\boldsymbol{P} = \boldsymbol{V}_e \boldsymbol{V}_e^\top$. Consider the Gram matrix of $\boldsymbol{G}_r$ and $\boldsymbol{G}_f$: $\boldsymbol{M}_r = \boldsymbol{P}\boldsymbol{G}_r^\top \boldsymbol{G}_r \boldsymbol{P}$ and $\boldsymbol{M}_f = \boldsymbol{P}\boldsymbol{G}_f^\top \boldsymbol{G}_f \boldsymbol{P}$. Throughout this analysis, the eigenvalues of $\boldsymbol{M}_f$ and $\boldsymbol{M}_r$ refer to their restriction to the effective subspace $\mathcal{E}$, or equivalently, their eigen-decompositions are defined as $\boldsymbol{M}_f = \sum_{j=1}^{r_e} \lambda_{f,j} \boldsymbol{e}_{f,j} \boldsymbol{e}_{f,j}^\top$, $\boldsymbol{M}_r = \sum_{j=1}^{r_e} \lambda_{r,j} \boldsymbol{e}_{r,j} \boldsymbol{e}_{r,j}^\top$.

Assumption A.3 (Bounded Spectral Disparity)**.** $\boldsymbol{M}_r$ and $\boldsymbol{M}_f$ exhibit bounded spectral disparity. Specifically, there exist constants $\alpha_f^- < 1 < \alpha_f^+, \alpha_r^- < 1 < \alpha_r^+$ but close enough to 1 such that for the maximum, minimum, and average eigenvalues of $\boldsymbol{M}_r, \boldsymbol{M}_f$,

$$\alpha_f^- \lambda_f^{\texttt{max}} \le \lambda_f^{\texttt{avg}} \le \alpha_f^+ \lambda_f^{\texttt{min}}, \qquad \alpha_r^- \lambda_r^{\texttt{max}} \le \lambda_r^{\texttt{avg}} \le \alpha_r^+ \lambda_r^{\texttt{min}}.$$

Bounded spectral disparity formalizes that the projected retain and forget signals are bulk-distributed on the effective space $\mathcal{E}$, and the spectrum within $\mathcal{E}$ is not so heavy-tailed that $\boldsymbol{v}^\top \boldsymbol{M}_r \boldsymbol{v}$ or $\boldsymbol{v}^\top \boldsymbol{M}_f \boldsymbol{v}$ varies by orders of magnitude across admissible directions. This is a natural regime for subspace-based unlearning, as $\mathcal{E}$ is defined from the joint gradient $\boldsymbol{G}_u$, so it filters out directions where the update signal is negligible and where task-specific curvature can be extremely ill-conditioned. Empirically, the projected spectra within $\mathcal{E}$ are typically far less heavy-tailed than in the full parameter space.

Lemma A.4. *Under Assumption A.3, for* $\epsilon = \max\left\{\frac{1}{\alpha_f^-} - 1, 1 - \frac{1}{\alpha_f^+}, \frac{1}{\alpha_r^-} - 1, 1 - \frac{1}{\alpha_r^+}\right\}$*, we have* $\forall i \in [r_e]$,

$$(1-\epsilon)\lambda_f^{\texttt{avg}} \le \boldsymbol{v}_i^\top \boldsymbol{M}_f \boldsymbol{v}_i \le (1+\epsilon)\lambda_f^{\texttt{avg}}, \qquad (1-\epsilon)\lambda_r^{\texttt{avg}} \le \boldsymbol{v}_i^\top \boldsymbol{M}_r \boldsymbol{v}_i \le (1+\epsilon)\lambda_r^{\texttt{avg}}.$$

Proof. Given that $\boldsymbol{M}_f = \sum_{j=1}^{r_e} \lambda_{f,j} \boldsymbol{e}_{f,j} \boldsymbol{e}_{f,j}^\top$, we have

$$\boldsymbol{v}_i^\top \boldsymbol{M}_f \boldsymbol{v}_i = \sum_{j=1}^{r_e} \lambda_{f,j} \left(\langle \boldsymbol{e}_{f,j}, \boldsymbol{v}_i \rangle\right)^2 \le \lambda_f^{\texttt{max}} \sum_{j=1}^{r_e} \left(\langle \boldsymbol{e}_{f,j}, \boldsymbol{v}_i \rangle\right)^2 \le \lambda_f^{\texttt{max}}$$

as $\boldsymbol{e}_{f,j}$ forms an orthonormal basis and $\boldsymbol{v}_i$ is a unit vector so that $\sum_{j=1}^{r_e} \left(\langle \boldsymbol{e}_{f,j}, \boldsymbol{v}_i \rangle\right)^2 = 1$. Similarly, we have

$$\lambda_f^{\texttt{min}} \le \boldsymbol{v}_i^\top \boldsymbol{M}_f \boldsymbol{v}_i \le \lambda_f^{\texttt{max}}, \qquad \lambda_r^{\texttt{min}} \le \boldsymbol{v}_i^\top \boldsymbol{M}_r \boldsymbol{v}_i \le \lambda_r^{\texttt{max}}.$$

Given Assumption A.3, we know $\lambda_f^{\texttt{max}} \le \frac{1}{\alpha_f^-} \lambda_f^{\texttt{avg}}$ and by definition, $\epsilon \ge \frac{1}{\alpha_f^-} - 1$, therefore

$$\boldsymbol{v}_i^\top \boldsymbol{M}_f \boldsymbol{v}_i \le \lambda_f^{\texttt{max}} \le \frac{1}{\alpha_f^-} \lambda_f^{\texttt{avg}} \le (1+\epsilon)\lambda_f^{\texttt{avg}}.$$

Similarly, we can show $\boldsymbol{v}_i^\top \boldsymbol{M}_f \boldsymbol{v}_i \ge (1-\epsilon)\lambda_f^{\texttt{avg}}$, and $(1-\epsilon)\lambda_r^{\texttt{avg}} \le \boldsymbol{v}_i^\top \boldsymbol{M}_r \boldsymbol{v}_i \le (1+\epsilon)\lambda_r^{\texttt{avg}}$. □

Proposition 4.3. **(a)** *If the direction carries a nontrivial forget/retain signal, i.e.,* $\|\boldsymbol{G}_f \boldsymbol{v}_i\| + \|\boldsymbol{G}_r \boldsymbol{v}_i\| \ge \xi$ *for* $\xi > 0$ *and its singular value is small relative to this signal, i.e.,* $\sigma_i \le \rho\xi$ *for* $\rho \in \left(0, \frac{1}{\sqrt{2}}\right)$*, then its forget-retain alignment satisfies* $a(\boldsymbol{v}_i) \le 2\rho^2 - 1 < 0$.
(b) *Under Assumption A.3, for two directions* $\boldsymbol{v}_i, \boldsymbol{v}_j$*,* $i, j \in [r_e]$*, with alignment scores* $a(\boldsymbol{v}_i) \le -\delta_i$ *and* $a(\boldsymbol{v}_j) \ge \delta_j$ *where* $\delta_i, \delta_j \in (0, 1]$*, if the separation* $(\delta_i + \delta_j) > \frac{\epsilon}{1-\epsilon} \cdot \frac{s^2+1}{s}$ *for* $\epsilon = \max\left\{\frac{1}{\alpha_f^-} - 1, 1 - \frac{1}{\alpha_f^+}, \frac{1}{\alpha_r^-} - 1, 1 - \frac{1}{\alpha_r^+}\right\}$ *where* $s = \frac{\|\boldsymbol{G}_r \boldsymbol{P}\|_F}{\|\boldsymbol{G}_f \boldsymbol{P}\|_F}$*, then* $\sigma_i < \sigma_j$.

Proof. (a) Given that $\boldsymbol{G}_u = \sum_{i=1}^{r_e} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^\top$, we have for the singular value

$$\begin{aligned}
\sigma_i^2 &= \|\boldsymbol{G}_u \boldsymbol{v}_i\|_2^2 \\
&= \|(\boldsymbol{G}_f + \boldsymbol{G}_r)\boldsymbol{v}_i\|_2^2 \\
&= \|\boldsymbol{G}_f \boldsymbol{v}_i\|_2^2 + \|\boldsymbol{G}_r \boldsymbol{v}_i\|_2^2 + 2\langle \boldsymbol{G}_f \boldsymbol{v}_i, \boldsymbol{G}_r \boldsymbol{v}_i \rangle \\
&= \|\boldsymbol{G}_f \boldsymbol{v}_i\|_2^2 + \|\boldsymbol{G}_r \boldsymbol{v}_i\|_2^2 + 2\|\boldsymbol{G}_f \boldsymbol{v}_i\|_2 \|\boldsymbol{G}_r \boldsymbol{v}_i\|_2 a(\boldsymbol{v}_i)
\end{aligned}$$

where the last equality follows from Definition 4.1. Also, we know from the conditions that

$$\sigma_i \leq \rho\xi \leq \rho\left(\|\boldsymbol{G}_f\boldsymbol{v}_i\| + \|\boldsymbol{G}_r\boldsymbol{v}_i\|\right).$$

As a result, we have

$$\|\boldsymbol{G}_f\boldsymbol{v}_i\|_2^2 + \|\boldsymbol{G}_r\boldsymbol{v}_i\|_2^2 + 2\|\boldsymbol{G}_f\boldsymbol{v}_i\|_2\|\boldsymbol{G}_r\boldsymbol{v}_i\|_2 a\left(\boldsymbol{v}_i\right) = \sigma_i^2 \leq \rho^2\left(\|\boldsymbol{G}_f\boldsymbol{v}_i\| + \|\boldsymbol{G}_r\boldsymbol{v}_i\|\right)^2.$$

Rearranging the terms yields

$$2\|\boldsymbol{G}_f\boldsymbol{v}_i\|_2\|\boldsymbol{G}_r\boldsymbol{v}_i\|_2 a\left(\boldsymbol{v}_i\right) \leq \left(\rho^2-1\right)\left(\|\boldsymbol{G}_f\boldsymbol{v}_i\|^2 + \|\boldsymbol{G}_r\boldsymbol{v}_i\|^2\right) + 2\rho^2\|\boldsymbol{G}_f\boldsymbol{v}_i\|_2\|\boldsymbol{G}_r\boldsymbol{v}_i\|_2.$$

Dividing both sides by $2\|\boldsymbol{G}_f\boldsymbol{v}_i\|_2\|\boldsymbol{G}_r\boldsymbol{v}_i\|_2$,

$$a\left(\boldsymbol{v}_i\right) \leq \rho^2 + \frac{\rho^2-1}{2}\left(\frac{\|\boldsymbol{G}_f\boldsymbol{v}_i\|_2}{\|\boldsymbol{G}_r\boldsymbol{v}_i\|} + \frac{\|\boldsymbol{G}_r\boldsymbol{v}_i\|}{\|\boldsymbol{G}_f\boldsymbol{v}_i\|_2}\right) \leq 2\rho^2 - 1,$$

where the last inequality follows from the facts that $\left(\frac{\|\boldsymbol{G}_f\boldsymbol{v}_i\|_2}{\|\boldsymbol{G}_r\boldsymbol{v}_i\|} + \frac{\|\boldsymbol{G}_r\boldsymbol{v}_i\|}{\|\boldsymbol{G}_f\boldsymbol{v}_i\|_2}\right) \geq 2$ and $\rho^2 - 1 \leq 0$. This completes the proof.

(b) From the proof of (a), we know

$$\sigma_i^2 = \|\boldsymbol{G}_f\boldsymbol{v}_i\|_2^2 + \|\boldsymbol{G}_r\boldsymbol{v}_i\|_2^2 + 2\|\boldsymbol{G}_f\boldsymbol{v}_i\|_2\|\boldsymbol{G}_r\boldsymbol{v}_i\|_2 a\left(\boldsymbol{v}_i\right).$$

Furthermore,

$$\begin{aligned}\|\boldsymbol{G}_f\boldsymbol{v}_i\|_2^2 &= \boldsymbol{v}_i^\top\boldsymbol{G}_f^\top\boldsymbol{G}_f\boldsymbol{v}_i \\ &= \boldsymbol{v}_i^\top\boldsymbol{P}\boldsymbol{G}_f^\top\boldsymbol{G}_f\boldsymbol{P}\boldsymbol{v}_i \\ &= \boldsymbol{v}_i^\top\boldsymbol{M}_f\boldsymbol{v}_i\end{aligned}$$

where we used the fact that $\boldsymbol{P}\boldsymbol{v}_i = \boldsymbol{v}_i$ as $\boldsymbol{P} = \boldsymbol{V}_e\boldsymbol{V}_e^\top$ and $\boldsymbol{v}_i$ is one column of $\boldsymbol{V}_e$. Therefore, by Lemma A.4 we know for $\|\boldsymbol{G}_f\boldsymbol{v}_i\|_2^2$ and similarly for $\|\boldsymbol{G}_r\boldsymbol{v}_i\|_2^2$

$$\left(1-\epsilon\right)\lambda_f^{\texttt{avg}} \leq \|\boldsymbol{G}_f\boldsymbol{v}_i\|_2^2 \leq \left(1+\epsilon\right)\lambda_f^{\texttt{avg}}, \qquad \left(1-\epsilon\right)\lambda_r^{\texttt{avg}} \leq \|\boldsymbol{G}_r\boldsymbol{v}_i\|_2^2 \leq \left(1+\epsilon\right)\lambda_r^{\texttt{avg}}.$$

As a result, we can have for direction $\boldsymbol{v}_i$ with singular value σ_i and conflict $a\left(\boldsymbol{v}_i\right) \leq -\delta_i < 0$,

$$\begin{aligned}\sigma_i^2 &= \|\boldsymbol{G}_f\boldsymbol{v}_i\|_2^2 + \|\boldsymbol{G}_r\boldsymbol{v}_i\|_2^2 + 2\|\boldsymbol{G}_f\boldsymbol{v}_i\|_2\|\boldsymbol{G}_r\boldsymbol{v}_i\|_2 a\left(\boldsymbol{v}_i\right) \\ &\leq \left(1+\epsilon\right)\lambda_f^{\texttt{avg}} + \left(1+\epsilon\right)\lambda_r^{\texttt{avg}} + 2\sqrt{\left(1-\epsilon\right)\lambda_f^{\texttt{avg}}\left(1-\epsilon\right)\lambda_r^{\texttt{avg}}}a\left(\boldsymbol{v}_i\right) \\ &\leq \left(1+\epsilon\right)\left(\lambda_f^{\texttt{avg}} + \lambda_r^{\texttt{avg}}\right) - 2\left(1-\epsilon\right)\delta_i\sqrt{\lambda_f^{\texttt{avg}}\lambda_r^{\texttt{avg}}}.\end{aligned}$$

And for direction $\boldsymbol{v}_j$ with singular value σ_j and conflict $a\left(\boldsymbol{v}_j\right) \geq \delta_j > 0$,

$$\begin{aligned}\sigma_j^2 &= \|\boldsymbol{G}_f\boldsymbol{v}_j\|_2^2 + \|\boldsymbol{G}_r\boldsymbol{v}_j\|_2^2 + 2\|\boldsymbol{G}_f\boldsymbol{v}_j\|_2\|\boldsymbol{G}_r\boldsymbol{v}_j\|_2 a\left(\boldsymbol{v}_j\right) \\ &\geq \left(1-\epsilon\right)\lambda_f^{\texttt{avg}} + \left(1-\epsilon\right)\lambda_r^{\texttt{avg}} + 2\sqrt{\left(1-\epsilon\right)\lambda_f^{\texttt{avg}}\left(1-\epsilon\right)\lambda_r^{\texttt{avg}}}a\left(\boldsymbol{v}_j\right) \\ &\geq \left(1-\epsilon\right)\left(\lambda_f^{\texttt{avg}} + \lambda_r^{\texttt{avg}} + 2\delta_j\sqrt{\lambda_f^{\texttt{avg}}\lambda_r^{\texttt{avg}}}\right).\end{aligned}$$

Given the condition of a sufficient gap in conflict $(\delta_i + \delta_j) > \frac{\epsilon}{1-\epsilon}\cdot\frac{s^2+1}{s}$ for $s = \frac{\|\boldsymbol{G}_r\boldsymbol{P}\|_F}{\|\boldsymbol{G}_f\boldsymbol{P}\|_F} = \sqrt{\frac{\lambda_r^{\texttt{avg}}}{\lambda_f^{\texttt{avg}}}}$, multiplying both sides by $2\left(1-\epsilon\right)\sqrt{\lambda_f^{\texttt{avg}}\lambda_r^{\texttt{avg}}}$ yields

$$2\left(1-\epsilon\right)\left(\delta_i+\delta_j\right)\sqrt{\lambda_f^{\texttt{avg}}\lambda_r^{\texttt{avg}}} > 2\epsilon\left(\lambda_f^{\texttt{avg}} + \lambda_r^{\texttt{avg}}\right),$$

which, by rearranging the terms, is equivalent to

$$\left(1+\epsilon\right)\left(\lambda_f^{\texttt{avg}} + \lambda_r^{\texttt{avg}}\right) - 2\left(1-\epsilon\right)\delta_i\sqrt{\lambda_f^{\texttt{avg}}\lambda_r^{\texttt{avg}}} < \left(1-\epsilon\right)\left(\lambda_f^{\texttt{avg}} + \lambda_r^{\texttt{avg}} + 2\delta_j\sqrt{\lambda_f^{\texttt{avg}}\lambda_r^{\texttt{avg}}}\right).$$

Therefore, we have

$$\begin{aligned}\sigma_i^2 &\leq \left(1+\epsilon\right)\left(\lambda_f^{\texttt{avg}} + \lambda_r^{\texttt{avg}}\right) - 2\left(1-\epsilon\right)\delta_i\sqrt{\lambda_f^{\texttt{avg}}\lambda_r^{\texttt{avg}}} \\ &< \left(1-\epsilon\right)\left(\lambda_f^{\texttt{avg}} + \lambda_r^{\texttt{avg}} + 2\delta_j\sqrt{\lambda_f^{\texttt{avg}}\lambda_r^{\texttt{avg}}}\right) \\ &\leq \sigma_j^2.\end{aligned}$$

□

A.3 Theoretical Justification for SalUn

Assumption A.4 (Smoothness)**.** For constants $\beta_f, \beta_r > 0$, the forgetting loss $\mathcal{L}_f$ and the retaining loss $\mathcal{L}_r$ are smooth, i.e., for $\boldsymbol{w}$, $\boldsymbol{w}'$,

$$\left|\mathcal{L}_f\left(\boldsymbol{w}'\right) - \mathcal{L}_f\left(\boldsymbol{w}\right) - \left\langle \boldsymbol{g}_f\left(\boldsymbol{w}\right), \boldsymbol{w}' - \boldsymbol{w}\right\rangle\right| \leq \frac{\beta_f}{2}\left\|\boldsymbol{w}' - \boldsymbol{w}\right\|^2,$$

$$\left|\mathcal{L}_r\left(\boldsymbol{w}'\right) - \mathcal{L}_r\left(\boldsymbol{w}\right) - \left\langle \boldsymbol{g}_r\left(\boldsymbol{w}\right), \boldsymbol{w}' - \boldsymbol{w}\right\rangle\right| \leq \frac{\beta_r}{2}\left\|\boldsymbol{w}' - \boldsymbol{w}\right\|^2.$$

Proposition 4.4. *Assume $\mathcal{L}_f$ and $\mathcal{L}_r$ are β_f- and β_r-smooth. Consider any coordinate $i \in [d]$ such that the forgetting/retention signal is nontrivial,* i.e.*, $\forall i \in [d]$, $|g_{f,i}| + |g_{r,i}| \geq \xi$ where $\xi > 0$, and the combined unlearning gradient is small,* i.e.*, $|g_{u,i}| \leq \rho\xi$ for some $\rho \in (0, 1]$. For the coordinate-wise update $\boldsymbol{w}' = \boldsymbol{w} - \eta g_{u,i}$ with step size $\eta < \min\left(\frac{2|g_{f,i}|}{\beta_f |g_{u,i}|}, \frac{2|g_{r,i}|}{\beta_r |g_{u,i}|}\right)$, the induced progress $\Delta\mathcal{L}_f = \mathcal{L}_f\left(\boldsymbol{w}'\right) - \mathcal{L}_f\left(\boldsymbol{w}\right)$ and $\Delta\mathcal{L}_r = \mathcal{L}_r\left(\boldsymbol{w}'\right) - \mathcal{L}_r\left(\boldsymbol{w}\right)$ satisfy $\Delta\mathcal{L}_f \Delta\mathcal{L}_r < 0$.*

Proof. We first show that $g_{f,i} g_{r,i} < 0$. Assume that $g_{f,i} g_{r,i} \geq 0$, then we have $|g_{u,i}| = |g_{f,i}| + |g_{r,i}| \geq \xi$, which contradicts the condition $|g_{u,i}| \leq \rho\xi$ for $\rho < 1$. Therefore, $g_{f,i} g_{r,i} < 0$.

Next, by Assumption A.4,

$$-\eta g_{f,i} g_{u,i} - \frac{\beta_f}{2}\eta^2 g_{u,i}^2 \leq \mathcal{L}_f\left(\boldsymbol{w}'\right) - \mathcal{L}_f\left(\boldsymbol{w}\right) \leq -\eta g_{f,i} g_{u,i} + \frac{\beta_f}{2}\eta^2 g_{u,i}^2,$$

If $g_{f,i} g_{u,i} > 0$, then for $\Delta\mathcal{L}_f = \mathcal{L}_f\left(\boldsymbol{w}'\right) - \mathcal{L}_f\left(\boldsymbol{w}\right)$,

$$\Delta\mathcal{L}_f \leq -\eta g_{f,i} g_{u,i} + \frac{\beta_f}{2}\eta^2 g_{u,i}^2 = \eta\left(\eta\frac{\beta_f}{2} g_{u,i}^2 - g_{f,i} g_{u,i}\right) < 0$$

where the last line follows from $\eta < \frac{2|g_{f,i}|}{\beta_f |g_{u,i}|}$. Also, if $g_{f,i} g_{u,i} < 0$, then

$$\Delta\mathcal{L}_f \geq -\eta g_{f,i} g_{u,i} - \frac{\beta_f}{2}\eta^2 g_{u,i}^2 = \eta\left(|g_{f,i}|\,|g_{u,i}| - \eta\frac{\beta_f}{2}|g_{u,i}|^2\right) > 0$$

where the last line again follows from $\eta < \frac{2|g_{f,i}|}{\beta_f |g_{u,i}|}$. As a result, we conclude that $\operatorname{sign}\left(\Delta\mathcal{L}_f\right) = -\operatorname{sign}\left(g_{f,i} g_{u,i}\right)$.

Similarly, we can show for $\Delta\mathcal{L}_r = \mathcal{L}_r\left(\boldsymbol{w}'\right) - \mathcal{L}_r\left(\boldsymbol{w}\right)$ that $\operatorname{sign}\left(\Delta\mathcal{L}_r\right) = -\operatorname{sign}\left(g_{r,i} g_{u,i}\right)$. Therefore,

$$\operatorname{sign}\left(\Delta\mathcal{L}_f \Delta\mathcal{L}_r\right) = \operatorname{sign}\left(g_{f,i} g_{u,i}\right)\operatorname{sign}\left(g_{r,i} g_{u,i}\right) = \operatorname{sign}\left(g_{f,i} g_{r,i} g_{u,i}^2\right) = \operatorname{sign}\left(g_{f,i} g_{r,i}\right),$$

which completes the proof. □

B Detailed Experiment Setup

B.1 Details of unlearning on image classification.

CIFAR-10. In the setting of Tab. 1, we randomly select 10% of the training data to forget with the model choice of ResNet-18. This corresponds to 5,000 samples that should be unlearned. We use seeds 1-5 for selecting forget samples and initialization to ensure reproducibility. Each experiment is evaluated on four metrics: retain set accuracy with the 40,000 training samples (excluding 5,000 in the validation set), forget set accuracy with the 5,000 samples to be unlearned, test set accuracy (TA) with all 10,000 test samples, and the membership inference attack (MIA) with the forget set and test set.

During unlearning, we apply SVD on-the-fly. The unlearning process fine-tunes the model on the 40,000-sample retain set for 10 epochs with the learning rate of 0.0113 while applying spectral filtering according to the keep ratio $\gamma = 0.70$, which controls the proportion of singular values retained after reconstruction and was chosen by grid search, selecting the value that yields the best validation performance.

In the setting of Tab. 2, we randomly select 50% of the training data to forget, corresponding to 25,000 samples to be unlearned. The keep ratio is set to be $\gamma = 0.70$ in consistency and learning rate 0.013. The seeds, models, and other settings follow the 10% setting.

Tiny-ImageNet. We use the same ResNet-18 model on Tiny-ImageNet, which contains 200 classes. Images are upsampled from 64×64 to 224×224 before standard crop-and-flip augmentation. We randomly select 10,000 of the 100,000 training images as the forget set and report the test set accuracy on the 10,000 validation images. The keep ratio is set to $\gamma = 0.5$ and the learning rate to 0.003. Results are averaged over three random seeds.

B.2 Details of unlearning on image generation.

DDPM on CIFAR-10. We evaluate class-conditional unlearning on CIFAR-10 where we aim to remove the model's ability to generate images from a specific class. The diffusion model uses a U-Net architecture with 128 base channels, channel multipliers of $[1, 2, 2, 2]$, 2 residual blocks per resolution, and self-attention at resolution 16. The diffusion process uses a linear noise schedule with 1000 timesteps, $\beta_{\text{start}} = 0.0001$ and $\beta_{\text{end}} = 0.02$.

During unlearning, we fine-tune the diffusion model for 1000 iterations with batch size 128 using gradient ascent-based unlearning loss. For computational efficiency, we apply on-the-fly SVD reconstruction only to convolutional layers, while using pre-computed saliency coordinate masks for non-convolutional layers. For the gradient block of each convolutional layer, we compute its SVD $\mathbf{G} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ during training and apply eigenvalue-based filtering according to the keep ratio $\gamma = 0.50$, which controls the proportion of singular values retained. The unlearning uses Adam optimizer with learning rate 1×10^{-5}, remain alpha $\alpha_{\text{remain}} = 1 \times 10^{-3}$, and forget alpha $\alpha_{\text{forget}} = 1.0$. We obtained the hyperparameters by grid search, selecting the value that yields the best validation performance. For evaluation, we generate 5,000 images for each class to compute FID and FA.

Stable Diffusion on Imagenette. Stable Diffusion uses a latent diffusion model with a U-Net denoising network operating in the latent space of a pre-trained autoencoder. The U-Net has 320 base channels, channel multipliers of $[1, 2, 4, 4]$, 2 residual blocks per resolution, 8 attention heads, and transformer depth of 1 with context dimension 768 for CLIP text conditioning. The diffusion process uses a linear noise schedule with 1000 timesteps, $\beta_{\text{start}} = 0.00085$ and $\beta_{\text{end}} = 0.012$.

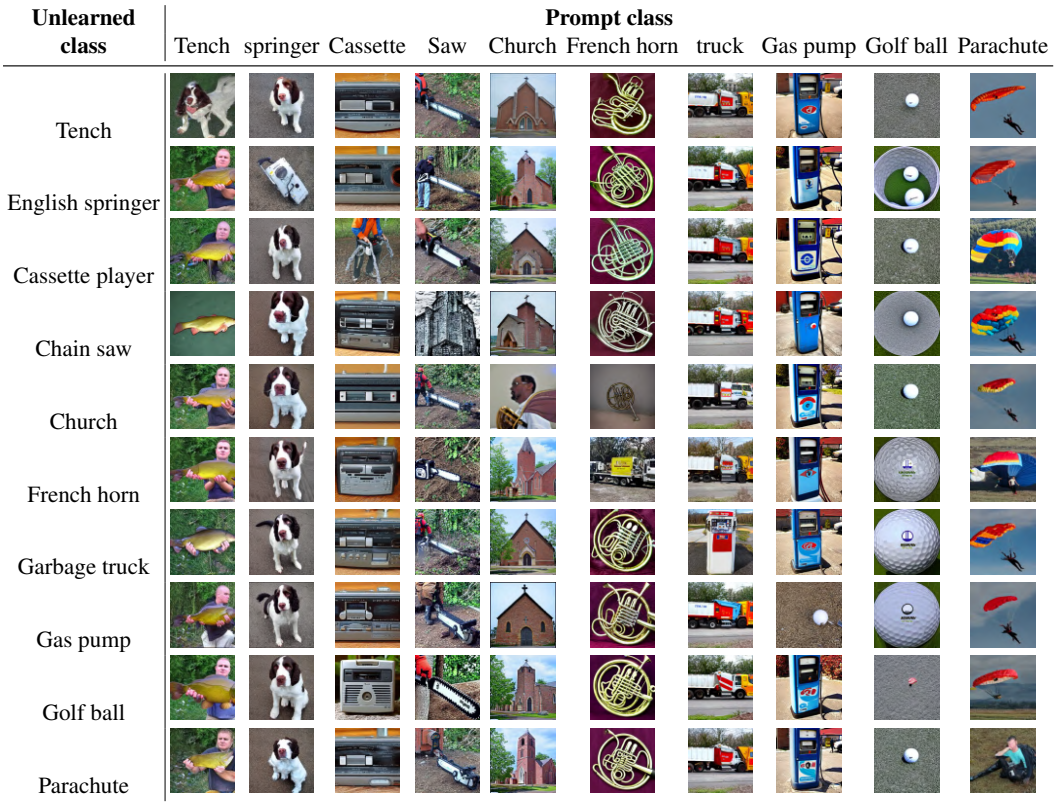
During unlearning, we fine-tune the model for 5 epochs with batch size 8 using gradient ascent-based unlearning loss with randomly assigned labels for the forget class. For computational efficiency, we apply on-the-fly SVD reconstruction only to cross-attention layers, while using pre-computed saliency coordinate masks for the other layers. For each convolutional layer, we compute its SVD $\mathbf{W} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ during training and apply eigenvalue-based filtering according to the keep ratio $\gamma = 0.50$, which controls the proportion of eigenvalues retained and was chosen by grid search, selecting the value that yields the best validation performance. The unlearning uses Adam optimizer with learning rate 1×10^{-5} and classifier-free guidance scale of 7.5. For evaluation, we generate 300 images for each class to compute FID and FA.

B.3 Details of unlearning on large language models.

During unlearning, we fine-tune the model for 5 epochs with batch size 8 and gradient accumulation steps of 4, resulting in an effective batch size of 32. We experiment with two unlearning objectives: IDK+AP (I-Don't-Know with answer preservation) and ME+GD (mismatch entropy with gradient difference). For computational efficiency, we apply on-the-fly SVD reconstruction only to attention projection layers (Q, K, V), while using pre-computed saliency coordinate masks for feed-forward and other layers. For attention layers, we compute SVD during training and apply eigenvalue-based filtering according to the QKV mask ratio $\alpha_{\text{qkv}} = 0.20$, which controls the proportion of eigenvalues retained. For non-attention layers, we use coordinate masks with mask ratio of 0.10.

For IDK+AP, we use a learning rate 3×10^{-5} with forget coefficient 1.0 and regularization coefficient 1.0. For ME+GD, we use a learning rate 2×10^{-5} with forget coefficient 0.1 and regularization coefficient 1.0. The top eigenvalue ratio for both methods is 0.2, which was chosen by grid search, selecting the value that yields the best validation performance. All experiments use the AdamW optimizer with weight decay 0.01 and are trained with distributed data parallel across 2 GPUs.

Figure 2: Examples of generated images using SSU. From the rows below, diagonal images represent the forgetting class, while non-diagonal images represent the remaining class.



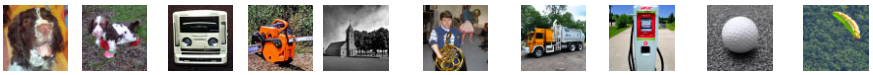
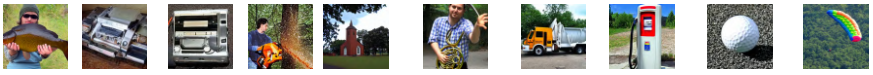
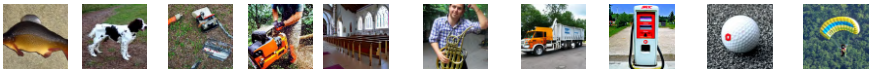
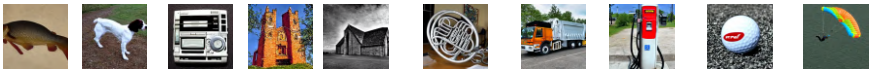
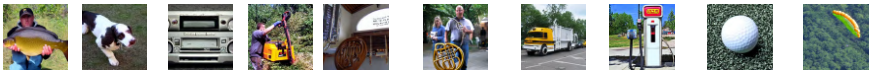
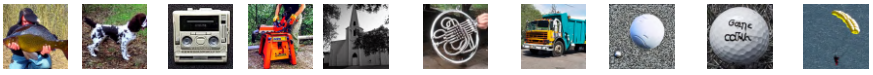
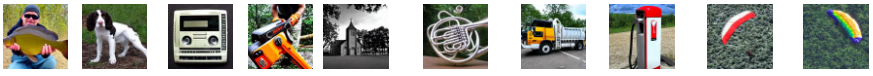
C ADDITIONAL GENERATION RESULTS

Fig. 2, Fig. 3, and Fig. 4 present class-wise unlearning results on the Imagenette dataset using the SSU method, evaluated under different random seeds. Each figure is organized as a matrix, where rows indicate the "Unlearned class" and columns indicate the "Prompt class," clearly separating the intended unlearning target from the generated outputs. Diagonal images correspond to the class being unlearned, highlighting the effectiveness of SSU in removing specific concepts. Off-diagonal images show generations for other classes, demonstrating the model's ability to generalize and distinguish among the remaining categories.

Figure 3: Examples of generated images using SSU. From the rows below, diagonal images represent the forgetting class, while non-diagonal images represent the remaining class.

Unlearned class	**Prompt class**									
	Tench	springer	Cassette	Saw	Church	French horn	truck	Gas pump	Golf ball	Parachute
Tench										
English springer										
Cassette player										
Chain saw										
Church										
French horn										
Garbage truck										
Gas pump										
Golf ball										
Parachute										

Figure 4: Examples of generated images using SSU. From the rows below, diagonal images represent the forgetting class, while non-diagonal images represent the remaining class.

Unlearned class	**Prompt class**									
	Tench	springer	Cassette	Saw	Church	French horn	truck	Gas pump	Golf ball	Parachute
Tench										
English springer										
Cassette player										
Chain saw										
Church										
French horn										
Garbage truck										
Gas pump										
Golf ball										
Parachute	