Unleashing Uncertainty: Efficient Machine Unlearning for Generative AI

Anonymous Authors[1]

Abstract

We introduce SAFEMax, a novel method for Machine Unlearning in diffusion models. Grounded in information-theoretic principles, SAFEMax maximizes the entropy in generated images, causing the model to generate Gaussian noise when conditioned on impermissible classes by ultimately halting its denoising process. Also, our method controls the balance between forgetting and retention by selectively focusing on the early diffusion steps, where class-specific information is prominent. Our results demonstrate the effectiveness of SAFEMax and highlight its substantial efficiency gains over state-of-the-art methods. We will share code upon acceptance.

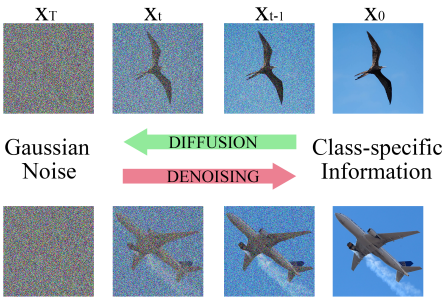


Figure 1. As the diffusion process progresses: (a) the entropy of the latent states increases due to the growing dominance of Gaussian noise, and (b) samples from different classes become increasingly similar. SAFEMax leverages both of these inherent properties of diffusion models to achieve effective and controlled unlearning.

1. Introduction

Machine Unlearning (MU) for generative AI aims to prevent the generation of impermissible content, such as samples from a specific class—referred to as the *forget samples* or *forget class*. The goal of MU is to correct pre-trained models efficiently, without retraining from scratch, thereby minimizing computational overhead. Efficiency is crucial to reducing both the cost and latency of model correction.

Most state-of-the-art methods in generative MU rely heavily on unlearning strategies originally developed for discriminative tasks, either by directly adapting existing methods to generative settings or by incorporating techniques, such as the Fisher Information Matrix (FIM) (Golatkar et al., 2020; Foster et al., 2024) and weight masking (Jia et al., 2023), originally introduced for discriminative MU.

For example, Selective Amnesia (Heng & Soh, 2023) uses FIM to guide unlearning through elastic weight consolidation in generative models. Saliency Unlearning (Fan et al., 2024) applies weight masks on top of discriminative MU methods—Random Labeling (Graves et al., 2021) and Gradient Ascent (Thudi et al., 2022)—which have been extended in generative settings. Other works (Wu & Harandi, 2024a;b; Ko et al., 2024; Patel & Qiu, 2025) adapt Gradient Ascent or NegGrad+ (Kurmanji et al., 2023) to generative models and apply multi-objective optimization techniques.

Despite progress in bridging discriminative and generative MU, insights from state-of-the-art approaches in discriminative tasks remain underexplored. Among recent discriminative MU methods, LoTUS (Spartalis et al., 2025) stands out for its strong unlearning performance and computational efficiency. It is an entropy-based method that increases the model's uncertainty on forget samples by smoothing the corresponding prediction probabilities up to an information-theoretic bound, thereby controlling the entropy increase. Motivated by this approach, we investigate the following research questions in the context of generative MU:

1. Can we efficiently train a generative model to forget specific samples by increasing the entropy on those samples?

2. Can we control this process to better balance the trade-off between unlearning and retention of useful knowledge?

To this end, we introduce a **S**imple **A**nd **F**ast **E**ntropy **Max**imization method for efficient MU in diffusion models. SAFEMax, for short, maximizes the entropy on the forget samples by training the model to generate Gaussian noise when conditioned on the forget class, ultimately halting the denoising process. Also, it leverages the progressive loss of information in the diffusion process, as shown in Figure 1, to control unlearning and balance forgetting and retention.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

Specifically, it applies stronger unlearning in early diffusion steps, where class-specific details are formed and refined, and weaker unlearning in the later steps, where latent states across different classes converge due to the dominant influence of Gaussian noise. SAFEMax significantly improves efficiency compared to existing unlearning methods while preserving state-of-the-art performance, establishing a scalable and cost-effective solution for MU in generative AI.

2. Simple and Fast Entropy Maximization for Machine Unlearning in Generative AI

The training of Denoising Diffusion Probabilistic Models (DDPMs) (Ho et al., 2020) consists of two processes. First, during the forward diffusion process, the original image $x_0 \sim q(x)$ is progressively corrupted with Gaussian noise over T steps, such that $x_T \sim \mathcal{N}(0, \mathbf{I})$. The noise level at each step is determined by a schedule α_t, and the latent state x_t remains tractable and can be sampled directly:

$$q(x_t \mid x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t} x_0, (1-\bar{\alpha}_t)\mathbf{I}), \quad \bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i \quad (1)$$

In the early diffusion steps, the Gaussian noise has lower variance (indicated by $\lim_{t\to 0} \bar{\alpha}_t = 1$), making the latent states x_t more similar to the original image x_0. These early states contain rich semantic information, which we refer to as *class-specific information*. As diffusion progresses, the latent states become more entropic (i.e., noisy) because of the dominance of cumulative Gaussian noise. In later diffusion steps, the variance of the injected Gaussian noise approximates its maximum (indicated by $\lim_{t\to T} \bar{\alpha}_t = 0$) and the distribution $q(x_t)$ becomes increasingly broad. Consequently, the latent states gradually lose the specific structure of the data and approach a state of maximum entropy. Specifically, for large T, the final latent state of the diffusion process, x_T, approximates the mean of the training data distribution for all inputs x_0 (Zhong et al., 2024).

Then, in the denoising process, a denoiser network ϵ_θ is trained to predict the noise ϵ_t at any arbitrary step t:

$$\mathcal{L} = \mathbb{E}_{t\in[1,T],\epsilon\sim\mathcal{N}(0,1)}[|| \epsilon_t - \epsilon_\theta(x_t, c, t) ||_2^2] \quad (2)$$

where c denotes the class of the original image x_0.

Unlearning with SAFEMax leverages the inherent Gaussian noise of the diffusion process, particularly the noise ϵ_T, which corresponds to the maximum entropy latent state x_T. To prevent the denoiser from reconstructing samples of an impermissible class c_f, we effectively halt the denoising process and maximize entropy in the model's output by fine-tuning the model with the following *forget loss*:

$$\mathcal{L}_f = \mathbb{E}_{t\in[1,T],\epsilon\sim\mathcal{N}(0,1)}[\psi(t) \, || \, \epsilon_T - \epsilon_\theta(x_t, c_f, t) ||_2^2] \quad (3)$$

where ϵ_T is the cumulative Gaussian noise in the final latent state, c_f is the *forget class*, and $\psi(t) = \exp(-t/T)$ is a monotonically decaying function within the range $[0, 1]$ that emphasizes unlearning in the early diffusion steps, where class-specific information is most prominent. This design encourages selective model updates, aiming to better balance unlearning and retention of useful prior knowledge.

Information-Theoretic Analysis of MU via Entropy Maximization. Consider a training sample x_0 and its reconstruction $\hat{x}$ generated by a diffusion model $g(x)$. This process can be viewed as a Markov chain: $x \mapsto g(x) \mapsto \hat{x}$. Let $P_e = \Pr\{x \neq \hat{x}\}$ denote the probability of reconstruction error. Within the context of diffusion models, we can define that the equality $x = \hat{x}$ applies when the reconstructed image $\hat{x}$ contains the same semantic information as the original image x (e.g., belonging to the same class). According to Fano's inequality (Cover & Thomas, 2006), we have:

$$P_e \geq \frac{H(x \mid \hat{x}) - 1}{log \mid \mathcal{X} \mid} \quad (4)$$

where $\mid \mathcal{X} \mid$ is the cardinality of the input space. Equation (4) indicates that as the reconstructed image $\hat{x}$ becomes less informative about the original image x (i.e., as the conditional entropy $H(x \mid \hat{x})$ increases), the lower bound of the error probability P_e increases, implying that the generated image $\hat{x}$ does not share the same semantic information as the original input x.

To prevent generating images from the *forget class*, we train the model to generate images similar to the ideal latent state x_T, which contains pure Gaussian noise (as $T \to \infty$ in the diffusion process) and thus no semantic information from x (denoted as x_0 in the diffusion process). Leveraging the entropy increase that is inherent in the diffusion trajectory, this design maximizes $H(x \mid \hat{x})$, thereby increasing the lower bound of reconstruction error P_e and guiding unlearning.

Balancing forgetting and retention with $\psi(t)$. Inspired by the strategies of delineating information for targeted MU in discriminative tasks (Spartalis et al., 2025), and loss scheduling for transfer learning in diffusion models (Zhong et al., 2024), we introduce a loss scheduling strategy for targeted MU in generative tasks. Our strategy is motivated by the observation that class-specific information is most prominent in the early diffusion steps and becomes increasingly obscured by Gaussian noise as the diffusion evolves, as shown in Figure 1. To leverage this, we propose a scheduling function that progressively reduces the influence of the forget loss throughout the diffusion process, thereby targeting unlearning in the semantically richer early stages:

$$\psi(t) = \exp(-\lambda \frac{t}{T}), \quad \text{for } t \in [0, T] \quad (5)$$

where λ is a hyperparameter to controls the rate of decay. A larger λ causes $\psi(t)$ to decay rapidly, concentrating the unlearning effect more narrowly on the early diffusion steps.

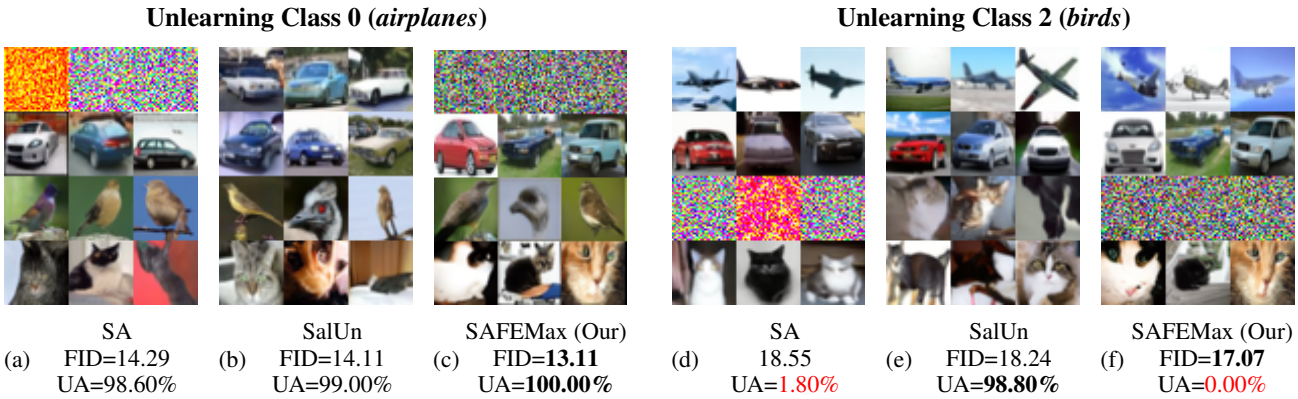


Figure 2. **Qualitative and quantitative results.** SAFEMax generates Gaussian noise for the forget class while preserving high-quality outputs for the retained classes, as reflected in FID. SAFEMax generates noise that is more consistent across forget samples compared to SA, whereas SalUn does not increase entropy and instead replaces forget samples with those of a fixed class. The UA drops for SAFEMax and SA on Class 2 (marked in red) not due to unlearning failure, but because the classifier erroneously identifies noise as *birds*.

Table 1. **Evaluation of Unlearning** using a ResNet34 classifier on images generated by a DDPM for a forgotten CIFAR-10 class. SAFEMax achieves the highest scores in Unlearning Accuracy (UA) and average entropy (H). The class highlighted in red indicates a case where the classifier failed to evaluate correctly.

Class	SA (Heng & Soh, 2023)		SalUn (Fan et al., 2024)		SAFEMax (Our)	
	$H\uparrow$	UA (%)$\uparrow$	$H\uparrow$	UA (%)$\uparrow$	$H\uparrow$	UA (%)$\uparrow$
0	1.062	98.60	0.051	99.00	**1.132**	**100.00**
1	0.987	99.60	0.032	**100.00**	**1.156**	**100.00**
2	0.948	1.80	0.084	**98.80**	**1.156**	0.00
3	1.006	**100.00**	0.068	99.60	**1.122**	**100.00**
4	0.926	**100.00**	0.085	99.60	**1.128**	**100.00**
5	0.908	**100.00**	0.040	99.60	**1.118**	**100.00**
6	0.993	**100.00**	0.045	**100.00**	**1.144**	**100.00**
7	1.007	**100.00**	0.027	**100.00**	**1.136**	**100.00**
8	0.900	**100.00**	0.045	99.20	**1.152**	**100.00**
9	0.998	**100.00**	0.057	99.20	**1.124**	**100.00**

Table 2. **Evaluation of Retention & Efficiency.** We report the mean (μ) and standard deviation (σ) across all CIFAR-10 classes. SAFEMax achieves the best score in Fréchet Inception Distance (FID) on generated images from the non-forgotten classes, Runtime Estimation (RTE) in minutes, and GPU memory usage in GB.

Class	SA (Heng & Soh, 2023)			SalUn (Fan et al., 2024)			SAFEMax (Our)		
	FID$\downarrow$	RTE$\downarrow$	GPU$\downarrow$	FID$\downarrow$	RTE$\downarrow$	GPU$\downarrow$	FID$\downarrow$	RTE$\downarrow$	GPU$\downarrow$
0	14.29	174.32	17.29	14.11	11.56	23.22	**13.11**	**5.82**	**9.50**
1	18.72	174.37	17.29	**16.85**	11.96	23.23	18.01	**5.79**	**9.50**
2	18.55	174.38	17.29	18.24	11.97	23.24	**17.07**	**5.80**	**9.50**
3	17.66	174.76	17.29	16.84	12.03	23.23	**15.64**	**5.89**	**9.50**
4	17.67	174.87	17.29	**16.64**	12.03	23.24	16.89	**5.80**	**9.50**
5	17.31	174.62	17.29	**16.95**	11.29	23.23	17.07	**5.79**	**9.50**
6	17.71	173.75	17.29	**16.78**	12.00	23.23	16.80	**5.79**	**9.50**
7	18.37	173.76	17.29	**16.93**	12.00	23.23	17.93	**5.90**	**9.50**
8	18.56	174.26	17.29	18.72	11.99	23.24	**18.20**	**5.80**	**9.50**
9	18.28	174.65	17.29	**15.55**	11.98	23.24	16.66	**5.85**	**9.50**
μ	17.71	174.37	17.29	16.76	11.81	23.23	**16.74**	**5.83**	**9.50**
σ	1.29	0.38	0.00	1.27	0.25	0.01	**0.32**	**0.04**	**0.00**

3. Experiments & Discussion

We follow the experimental setup and evaluation framework of Selective Amnesia (SA) and Saliency Unlearning (SalUn). We aim to forget a specific class within a DDPM trained on CIFAR-10 (Krizhevsky, 2009). We apply SalUn on top of the Random Labeling strategy using saliency masks that update 50% of the weights, as proposed by the authors for DDPMs. For SA, we follow the author's configuration and perform unlearning for 20,000 iterations. For SAFEMax, we adopt the same hyperparameter values as SalUn and perform unlearning for 1,000 iterations, consistent with SalUn.

To evaluate **unlearning**, we use a ResNet34 classifier, pre-trained on ImageNet (Deng et al., 2009) and fine-tuned on CIFAR-10 for 20 epochs. We report two key metrics: (i) *Unlearning Accuracy (UA)*, defined as 100% minus the accuracy of the classifier on the forget class, and (ii) the average *entropy (H)* of the classifier's prediction for the forget class, which captures the uncertainty introduced by the unlearning method. To evaluate **retention** (i.e., the model's performance on the remaining classes), we compute the *Fréchet Inception Distance* on generated images for retained classes. Finally, we measure **efficiency** by reporting each method's *Runtime Estimation (RTE)* and *peak GPU memory usage*.

Effectiveness. SAFEMax demonstrates strong unlearning performance by excelling in both forgetting and retention of the respective information. Overall, our method achieves better unlearning accuracy (UA) and retention quality (FID) than state-of-the-art approaches, as shown in Tables 1 and 2.

In terms of **unlearning**, SAFEMax consistently achieves the maximum UA score (100%) in all-but-one case. The only exception is Class 2 (*birds*), where the UA drops to 0%, indicating that the classifier predicts all generated images for the unlearned class as *birds*. However, this anomaly is not due a failure of our method. Instead, the classifier mistakenly identifies Gaussian noise as birds, as evidenced in Figure 2. This observation highlights that **UA scores can be misleading in isolation** and underscores the importance

of examining both quantitative and qualitative results.

Notably, SAFEMax consistently achieves the highest increase in entropy, directly aligning with its design goal of maximizing uncertainty for the forget class. As shown in Figure 2, it generates high-entropy noise more reliably than SA, leading to greater classifier uncertainty, as reflected in the entropy scores (H) in Table 1. In contrast, SalUn undermines the entropy-increase objective of Random Labeling in discriminative tasks. While Random Labeling originally assigned random incorrect labels to forget samples at each unlearning iteration, its adaptation in SalUn maps the forget class to a fixed alternative class (as shown in Figure 2). This leads the evaluation model to make incorrect yet high-confidence predictions, as shown in Table 1.

In terms of **retention**—preserving high image quality for the remaining classes as measured by FID—SAFEMax generally achieves state-of-the-art results as shown in Table 2. Overall, SAFEMax delivers effective forgetting and retention while requiring significantly less time and memory.

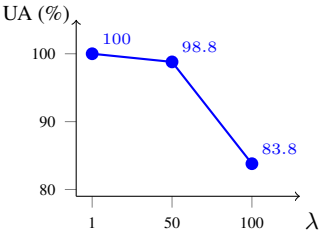
Efficiency. As shown in Table 2, a key advantage of SAFEMax is achieving a strong balance between forgetting and retention without relying on computationally intensive techniques (e.g., the use of FIM, weight masks, or regularization terms for multi-objective optimization) that introduce significant overhead before or during unlearning. Compared to SalUn, which runs for the same number of iterations, SAFEMax is $2\times$ faster. Against SA, our method achieves a $30\times$ speed-up due to its ability to unlearn effectively in far fewer iterations. When also accounting for the time to compute the FIM for SA (1226.98 minutes, not included in the RTE metric), SAFEMax offers a total $230\times$ improvement in runtime.

In terms of GPU memory usage, SAFEMax is 59% more efficient than SalUn and 45% more efficient than SA. The high memory demands of SalUn and SA result from their reliance on storing the saliency masks and FIM, respectively. In contrast, SAFEMax does not require complex auxiliary structures to balance forgetting and retention, and instead uses a simple, predefined scheduler. This simple yet effective design makes SAFEMax not only faster but also more practical for large-scale or resource-constrained applications.

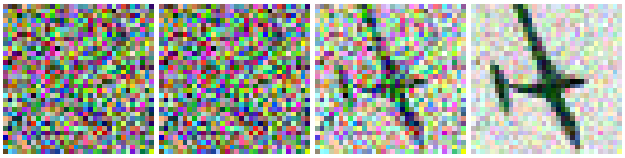
Ablation Study. To assess the role of our scheduling function $\psi(t)$, we conducted an ablation study, presented in Figure 3. The results show that SAFEMax achieves strong unlearning performance even without the scheduler, underscoring the robustness of our core method. However, introducing the decaying scheduler $\psi(t)$ improves the trade-off between unlearning and retention, as evidenced by improved image quality in the generated outputs. We further analyze the impact of the decay parameter λ and verify that a faster decay (i.e., larger λ) improves retention—even for the forget class. In our experiments, we used a moderate value of $\lambda = 1$, without additional tuning. Tuning λ may yield better results.

(a) **Effect of decaying scheduler ($\lambda = 1$) vs. no scheduler ($\lambda = 0$).** SAFEMax improves the image quality for retained classes (see **5.62%** improvement in FID), while still unlearning perfectly.

λ	UA (%) ↑	FID ↓
0	100.00	13.89
1	100.00	13.11



(b) **As λ increases, more information is retained**—even for the forget class, as show by the drop in UA.



(c) $\lambda = 0$ (d) $\lambda = 1$ (e) $\lambda = 50$ (f) $\lambda = 100$

Figure 3. **Ablation study: Increasing λ enhances information retention.** Quantitative and qualitative results from unlearning Class 0 (*airplane*) support our hypothesis behind the decaying scheduler $\psi(t)$: Faster decay (i.e., larger λ) results in greater information retention, as shown in (b), (e), and (f) and $\psi(t)$ can enhance the quality of generated images from the remaining classes, while still enabling perfect unlearning, as shown in (a), (c), and (d).

4. Conclusion

In this paper, we introduced SAFEMax, an effective and significantly more efficient Machine Unlearning strategy for diffusion models. Motivated by the information-theoretic analysis of LoTUS (Spartalis et al., 2025) in discriminative tasks, we maximized the entropy of the forget samples by leveraging the natural entropy increase inherent to diffusion models. We further proposed a simple scheduling mechanism that targets class-specific information, enhancing the balance between unlearning and the retention of useful knowledge. We compared SAFEMax against the most prominent state-of-the-art approaches for DDPM unlearning and showed that our method not only achieves strong unlearning performance but also offers substantial improvements in computational efficiency. These findings suggest that SAFEMax is a promising, scalable, and cost-effective unlearning strategy.

Limitations & Future Work. A more comprehensive evaluation could further strengthen the validity and applicability of our approach, even though SAFEMax has already demonstrated strong performance across key benchmarks. We plan to compare against more recent DDPM unlearning methods, such as (Patel & Qiu, 2025), once their code is publicly released, and to extend our evaluation to additional datasets and Stable Diffusion models.

References

Cover, T. and Thomas, J. *Elements of information theory*. Wiley-Interscience, 2nd edition, 2006.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Fan, C., Liu, J., Zhang, Y., Wei, D., Wong, E., and Liu, S. Salun: Empowering machine unlearning via gradient-based weight saliency in both image classification and generation. In *International Conference on Learning Representations*, 2024.

Foster, J., Schoepf, S., and Brintrup, A. Fast machine unlearning without retraining through selective synaptic dampening. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 12043–12051, 2024.

Golatkar, A., Achille, A., and Soatto, S. Eternal sunshine of the spotless net: Selective forgetting in deep networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9304–9312, 2020.

Graves, L., Nagisetty, V., and Ganesh, V. Amnesiac machine learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 11516–11524, 2021.

Heng, A. and Soh, H. Selective amnesia: A continual learning approach to forgetting in deep generative models. *Advances in Neural Information Processing Systems*, 36: 17170–17194, 2023.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Jia, J., Liu, J., Ram, P., Yao, Y., Liu, G., Liu, Y., Sharma, P., and Liu, S. Model sparsity can simplify machine unlearning. *Advances in Neural Information Processing Systems*, 36:51584–51605, 2023.

Ko, M., Li, H., Wang, Z., Patsenker, J., Wang, J. T., Li, Q., Jin, M., Song, D., and Jia, R. Boosting alignment for post-unlearning text-to-image generative models. *Advances in Neural Information Processing Systems*, 37:85131–85154, 2024.

Krizhevsky, A. Learning multiple layers of features from tiny images. 2009.

Kurmanji, M., Triantafillou, P., Hayes, J., and Triantafillou, E. Towards unbounded machine unlearning. *Advances in neural information processing systems*, 36:1957–1987, 2023.

Patel, G. and Qiu, Q. Learning to unlearn while retaining: Combating gradient conflicts in machine unlearning, 2025.

Spartalis, C. N., Semertzidis, T., Gavves, S., and Daras, P. Lotus: Large-scale machine unlearning with a taste of uncertainty. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2025.

Thudi, A., Deza, G., Chandrasekaran, V., and Papernot, N. Unrolling sgd: Understanding factors influencing machine unlearning. In *2022 IEEE 7th European Symposium on Security and Privacy (EuroS&P)*, pp. 303–319. IEEE, 2022.

Wu, J. and Harandi, M. Munba: Machine unlearning via nash bargaining. *arXiv preprint arXiv:2411.15537*, 2024a.

Wu, J. and Harandi, M. Scissorhands: Scrub data influence via connection sensitivity in networks. In *European Conference on Computer Vision*, pp. 367–384. Springer, 2024b.

Zhong, J., Guo, X., Dong, J., and Long, M. Diffusion tuning: Transferring diffusion models via chain of forgetting. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

A. Broader Social Impact

Methods like SAFEMax align machine learning models with privacy regulations and ethical standards, by particularly preventing the generation of impermissible or sensitive content. However, they can also be misused by adversaries to deliberately degrade the performance of otherwise well-functioning models, suggesting careful consideration of deployment practices.

B. More Visualizations

Figure 4. Unlearning Class 0 (*airplanes*).

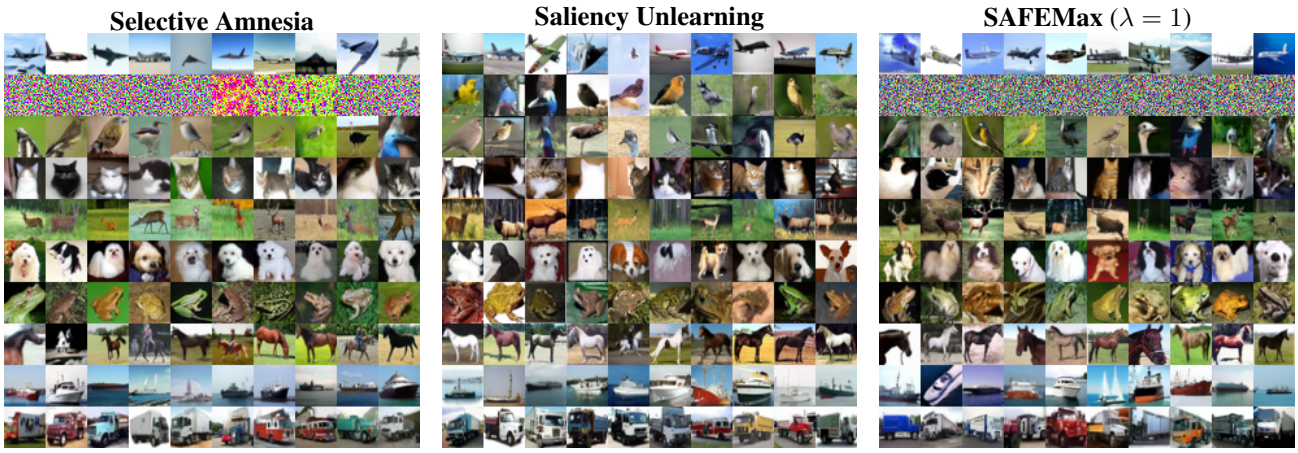


Figure 5. Unlearning Class 1 (*cars*).

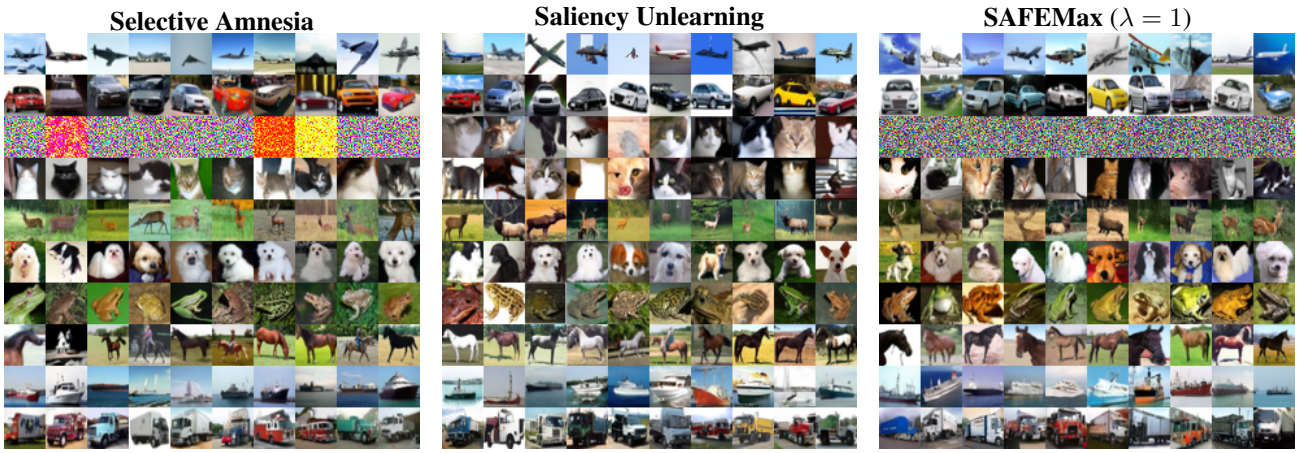


Figure 6. Unlearning Class 2 (*birds*).

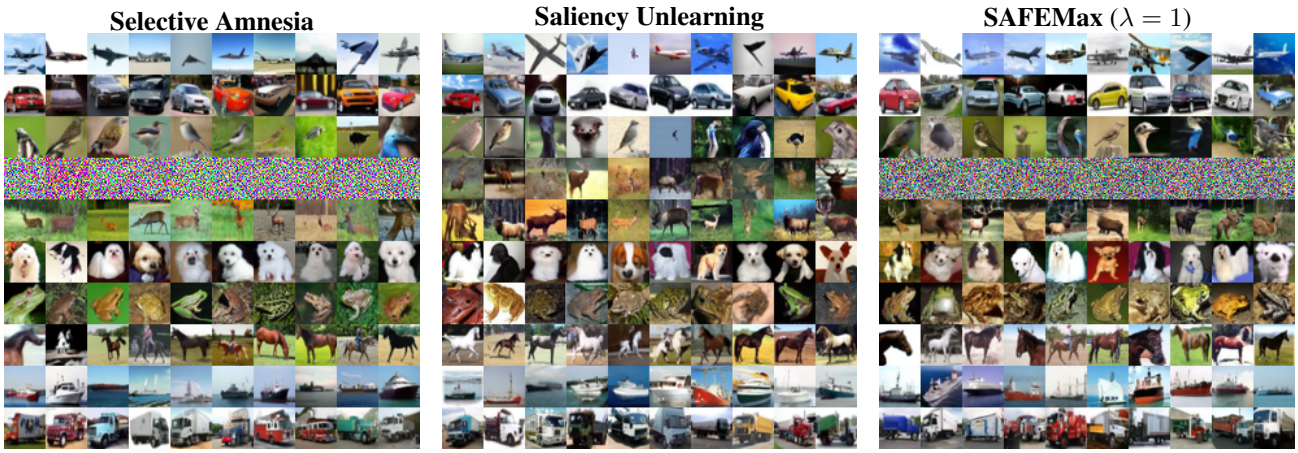


Figure 7. Unlearning Class 3 (*cats*).

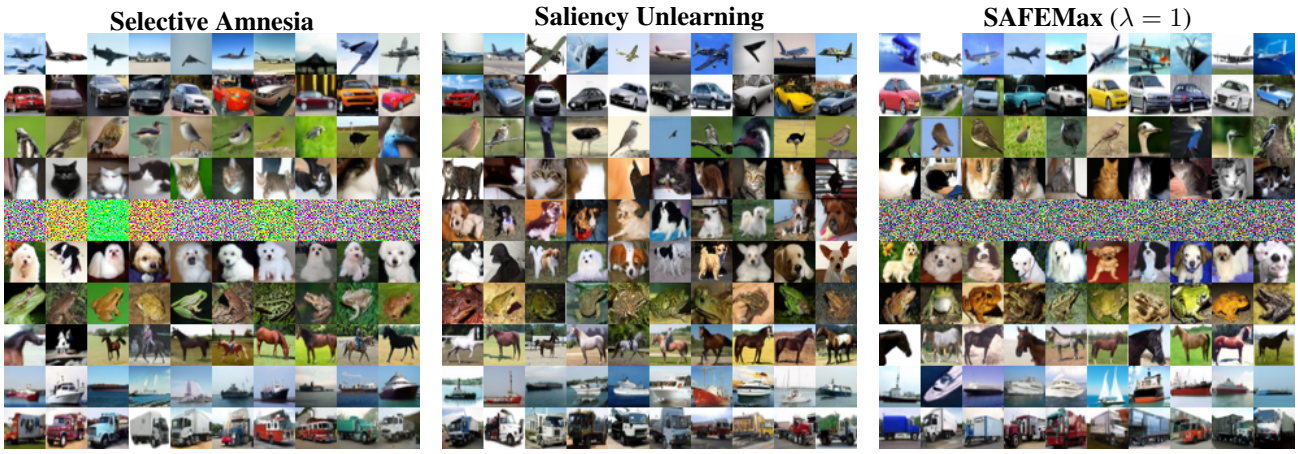


Figure 8. Unlearning Class 4 (*deer*).

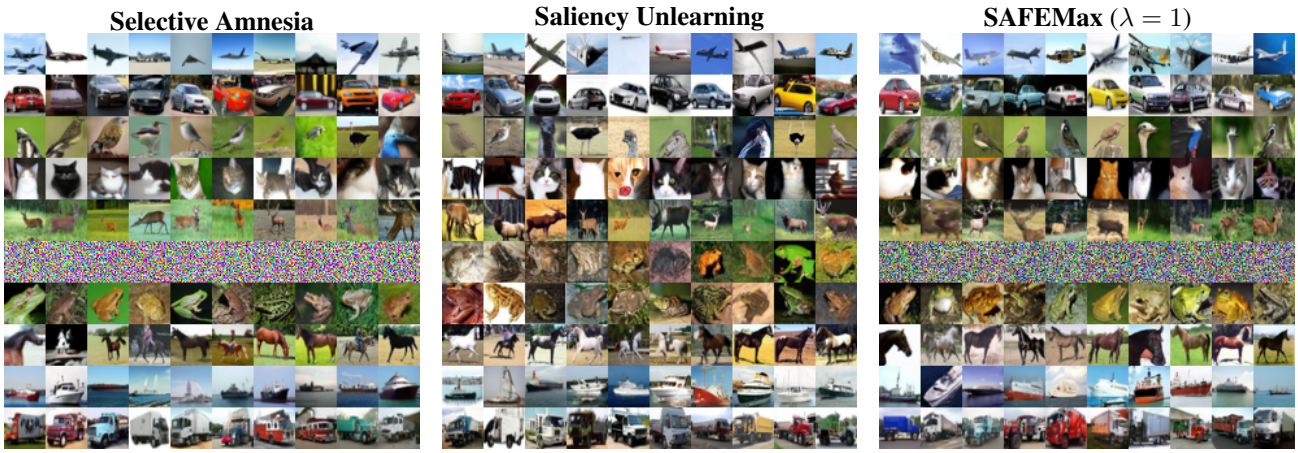


Figure 9. Unlearning Class 5 (*dogs*).

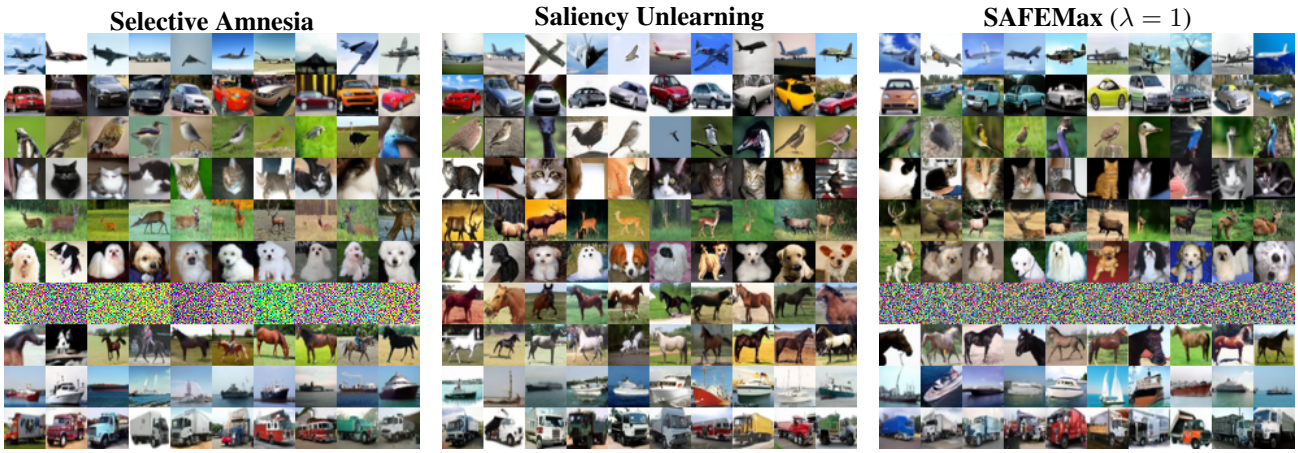


Figure 10. Unlearning Class 6 (*frogs*).

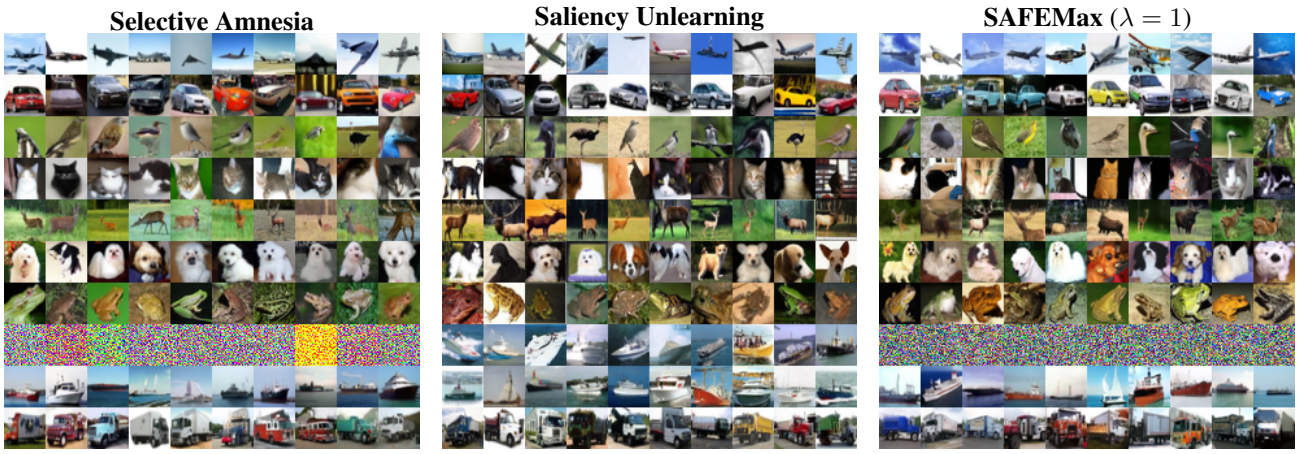


Figure 11. Unlearning Class 7 (*horses*).

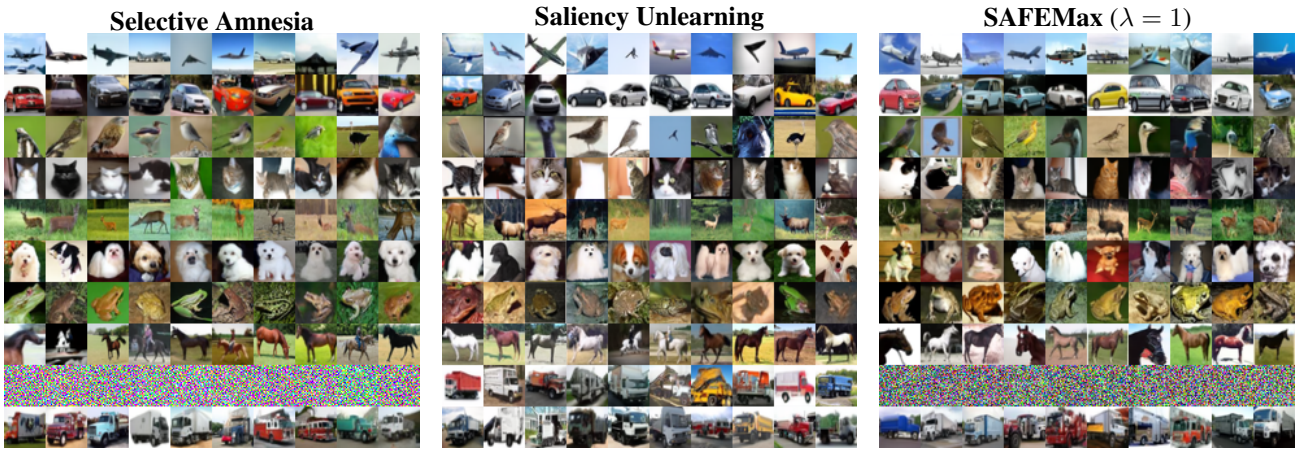


Figure 12. Unlearning Class 8 (*ships*).

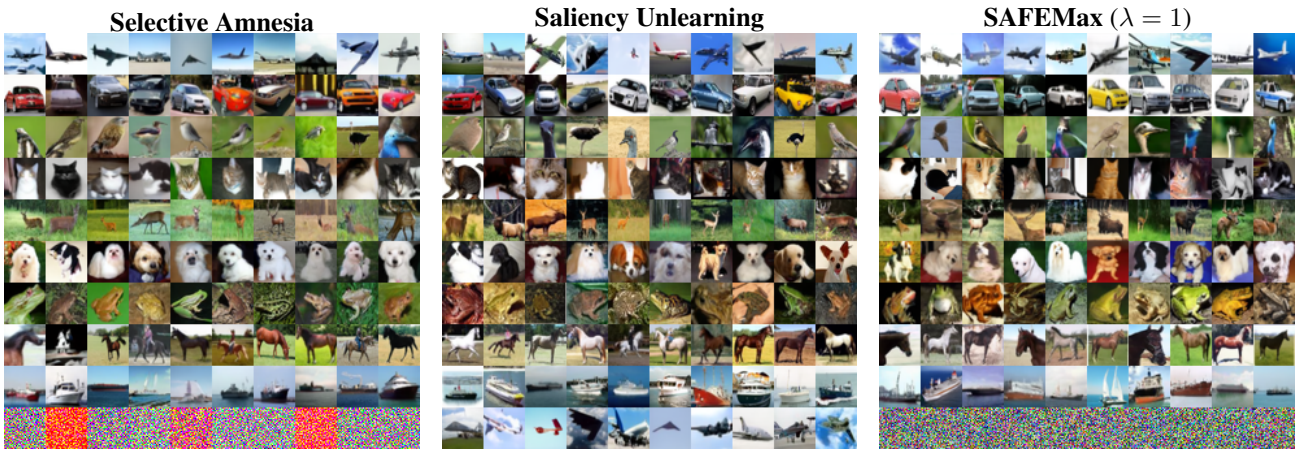


Figure 13. Unlearning Class 9 (*trucks*).