Unleashing Uncertainty: Efficient Machine Unlearning for Generative AI

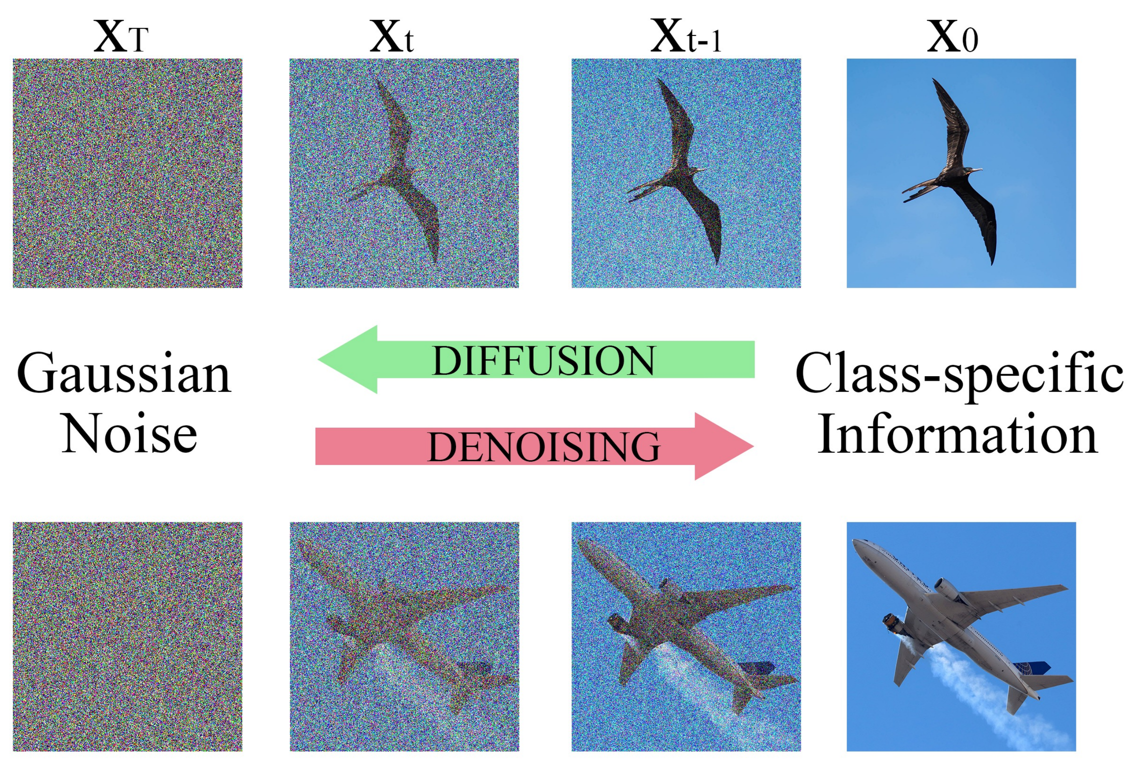
Christoforos Spartalis, Theodoros Semertzidis, Petros Daras, Efstratios Gavves

Motivations

- **Information-Theoretic** Unlearning for GenAI
- **Efficiency** ~ Cost & Latency of Model Correction

SAFEMax

- Leverages the inherent Gaussian noise of the diffusion process to maximize the entropy in generted images of impermissible classes ➔ **halts the denoising process**
- **Efficiently** balances forgetting and retention by focusing unlearning on the semantically rich steps.



Why Entropy Maximization?

$$Pr\{X \neq \hat{X}\} \geq \frac{H(X \mid \hat{X}) - 1}{log \mid \mathcal{X} \mid}$$

Semantics of an original image

Semantics of a generated image

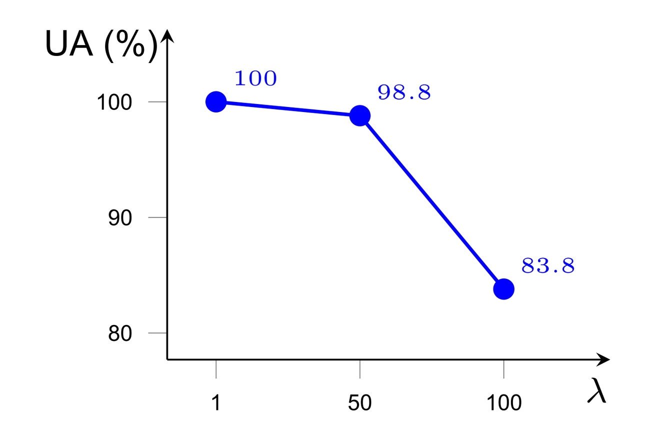
$$\mathcal{L}_f = \mathbb{E}_{t\in[1,T],\epsilon\sim\mathcal{N}(0,1)}[\psi(t) \mid\mid \overset{\epsilon_T}{\cancel{\epsilon_t}} - \epsilon_\theta(x_t, c_f, t) \mid\mid_2^2]$$

Balancing Forgetting & Retention

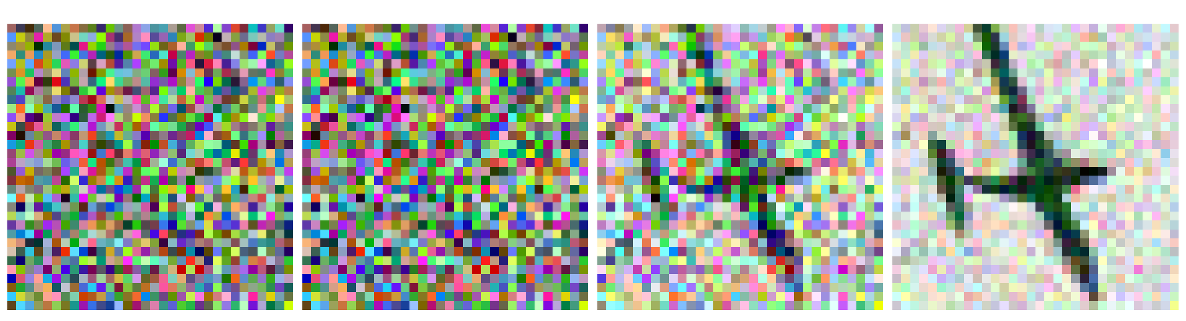
$$\psi(t) = \exp(-\lambda\frac{t}{T}), \quad \text{for } t \in [0, T]$$

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



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



$\lambda = 0$ $\lambda = 1$ $\lambda = 50$ $\lambda = 100$

Results

Class	SA (Heng & Soh, 2023)			SalUn (Fan et al., 2024)			SAFEMax (Ours)		
	FID ↓	RTE ↓	GPU ↓	FID ↓	RTE ↓	GPU ↓	FID ↓	RTE ↓	GPU ↓
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**

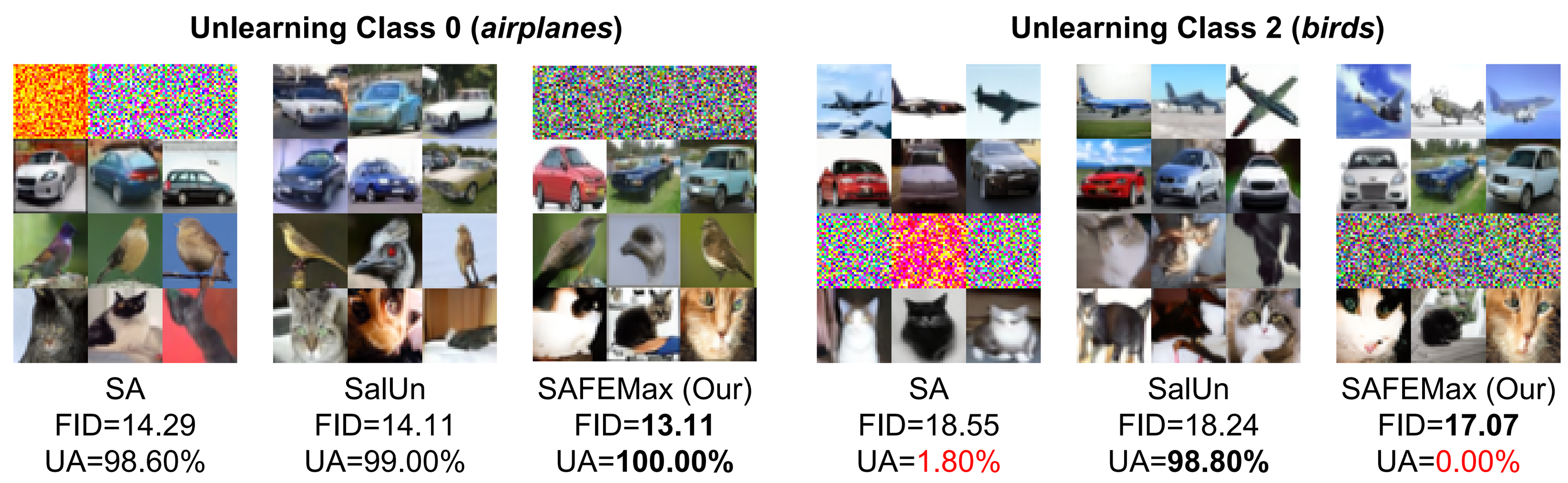
Run Time Estimation

- **30x faster** than Selective Amnesia (SA)
  - **230x faster** including the priors used by SA
- **2x faster** than Saliency Unlearning (SalUn)

GPU Memory Usage

- **45% more efficient** than Selective Amnesia
- **59% more efficient** than Saliency Unlearning

Class	SA (Heng & Soh, 2023)		SalUn (Fan et al., 2024)		SAFEMax (Ours)	
	H ↑	UA (%) ↑	H ↑	UA (%) ↑	H ↑	UA (%) ↑
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**



Future Work

- Alternative Evaluation of Unlearning Accuracy
- Resilience to Unlearning Attacks
- Concept Unlearning

Reference

Spartalis et al. "LoTUS: Large-Scale Machine Unlearning with a Taste of Uncertainty". *Proceedings of the Computer Vision and Pattern Recognition Conference* 2025.