SUGAR: A Sweeter Spot for Generative Unlearning of Many Identities

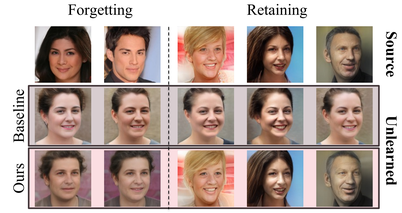
Dung Thuy Nguyen[1], Quang Nguyen[2], Preston K. Robinette[1], Eli Jiang[1], Taylor T. Johnson[1], Kevin Leach[1]
[1]Vanderbilt University, [2]Rutgers University

WACV 2026 TUCSON, AZ 3/6 – 3/10

Motivation

- As generative models continue to advance in their ability to synthesize and generate highly realistic images, concerns about ethical misuse have intensified: personal image misuses[1,2].
- Generative Identity Unlearning (GUI[3]): removing identity information from pretrained generative models without retraining from scratch, ensuring the-right-to-be-forgotten[5,6].
- No method was designed for (i) batch unlearning requests and (ii) sequential unlearning, i.e., often leading to retention corruption.
- Security remains critical, as mapping forgotten identities to fixed outputs (e.g., noise[4] or an average face[5]) may introduce detectable signatures, making the unlearning process vulnerable to erasure-detection or membership-inference attacks.

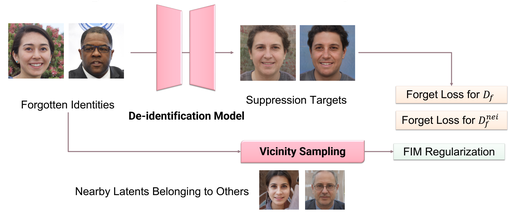
- ✓ Batch unlearning
- ✓ Sequential unlearning
- ✓ Controllable unlearning
- ✓ Addresses forgetting–retention trade-offs
- ✓ Avoids unintended lateral collapse and unlearning signatures



Multi-Identity Forgetting Results. Our method effectively removes target identities while preserving the quality and distinctiveness of retained ones.

High-level Idea:

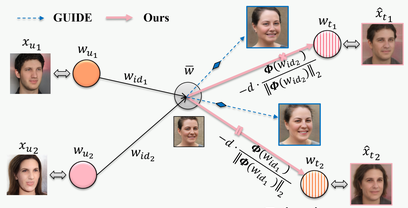
- Instead of their suppression mechanisms mapping to noise or an average face, can we make this process learnable?
- We explicitly protect the nearby retaining space, which is often close to the identities being removed and therefore more likely to be affected.



Methods and Materials

We propose a learnable and sample-specific procedure to determine where each forgotten sample should be mapped.

Key innovation: ID-specific De-identification $\mathcal{T}_\Theta$



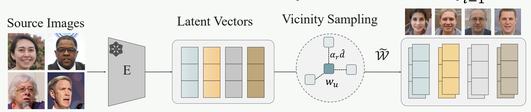
$$\Theta(\cdot, w_u) := w_t = \bar{w} - d \cdot \frac{\mathcal{T}_\Theta(w_{id})}{\|\mathcal{T}_\Theta(w_{id})\|}$$

$$\mathcal{L}_{de}(\Theta) = \frac{1}{|W_u|}\sum [\mathcal{L}_{mse}\left(\mathcal{F}(w_u), \mathcal{F}(\Theta(w_u))\right) - \mathcal{L}_{per}(x_u, \hat{x}_u) - \mathcal{L}_{id}(x_u, \hat{x}_u)]$$

$$\mathcal{L}_{\text{forget}}(w_u; \theta) = \mathcal{L}_{\text{map}}\left(w_u, \Theta(w_u)\right) + \lambda_{\text{nei}} \sum_{w_{u,a} \in W_u} \mathcal{L}_{\text{map}}\left(w_{u,a}, \Theta(w_{u,a})\right)$$

Maintaining Model Retention by Vicinity Sampling

We propose using FIM-based regularization to protect nearby samples from unintended disruption, i.e., $\widetilde{W} := \{\widetilde{w}^i \mid \widetilde{w}^i = w_u^i + \alpha_r \hat{d}^i\}_{i=1}^K$



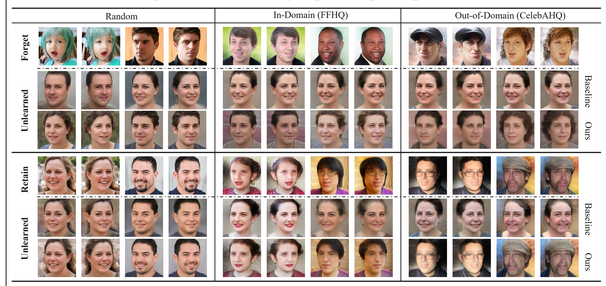
This regularizer encourages parameters most critical for preserving behavior on the vicinity set to remain near their pre-unlearning values

$$FIM_i \approx \frac{1}{|\widetilde{W}|}\sum_{w \in \widetilde{W}} \left(\frac{\partial \ell_{\text{ref}}(w;\theta)}{\partial \theta_i}\bigg|_{\theta=\theta^*}\right)^2$$

$$\mathcal{L}_{\text{unlearn}} = \mathbb{E}_{w_u \sim W_u}\left[\mathcal{L}_{\text{forget}}(w_u;\theta)\right] + \lambda_{\text{ewc}}\mathcal{L}_{\text{ewc}}(\theta;\theta^*)$$

Results

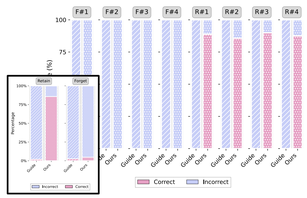
Qualitative comparison of model utility in generating unforgotten identities



Quantitative comparison of model utility in generating unforgotten identities

#IDS	Methods	FFHQ (In-domain Distribution) ID (↑)	FFHQ (In-domain Distribution) FID (↓)	CelebAHQ (OOD Distribution) ID (↑)	CelebAHQ (OOD Distribution) FID (↓)	Random ID (↑)	Random FID (↓)
N=1	GUIDE	0.3301 ±0.0073	112.06 ±5.1507	0.5118 ±0.0027	133.16 ±4.9765	0.6061 ±0.0061	66.531 ±2.4249
	Ours	0.5730 ±0.0103	94.873 ±0.3855	0.7057 ±0.0055	66.280 ±4.3198	0.6457 ±0.0030	87.257 ±6.9018
N=5	GUIDE	0.2233 ±0.0040	143.23 ±1.4068	0.3244 ±0.0292	116.47 ±2.0877	0.4346 ±0.0019	144.40 ±0.1790
	Ours	0.5500 ±0.0022	113.10 ±0.9406	0.5766 ±0.0070	101.78 ±2.6185	0.6538 ±0.0017	115.53 ±0.5600
N=10	GUIDE	0.2132 ±0.0063	181.66 ±2.5615	0.2893 ±0.0024	165.37 ±0.8487	0.2798 ±0.0012	138.19 ±0.6920
	Ours	0.5001 ±0.0030	140.86 ±2.3990	0.4791 ±0.0060	118.32 ±0.9936	0.4392 ±0.0010	98.023 ±0.4619
N=20	GUIDE	0.1150 ±0.0036	191.50 ±1.5908	0.1748 ±0.0116	168.34 ±2.5693	0.2281 ±0.0007	155.08 ±0.6582
	Ours	0.4515 ±0.0051	149.56 ±1.1429	0.4182 ±0.0051	124.31 ±5.2634	0.4774 ±0.0024	90.080 ±0.1015
N=50	GUIDE	0.0115 ±0.0112	197.91 ±0.5813	0.1339 ±0.0054	174.96 ±3.2256	0.1569 ±0.0005	141.61 ±0.3855
	Ours	0.3583 ±0.0043	164.07 ±1.8977	0.3504 ±0.0043	137.41 ±1.9314	0.4612 ±0.0009	88.306 ±0.0092
Average Impr. (%)		*+ 96.424*	*+ 19.568*	*+ 732.18*	*+ 27.780*	*+ 83.586*	*+ 18.928*

Human judgment result on selected forgetting (F#) and retaining (R#) identities.



60 participants; 579 responses

Forgotten Identities: neither GUIDE nor SUGAR allowed consistent identification of the original identity after unlearning.
Retained Identities: Participants correctly identified SUGAR-generated identities **86%** of the time, compared to only **1.9%** for GUIDE.

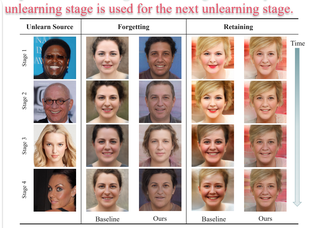
Forgetting an identity does not require aggressively minimizing metrics like ID

Baseline: Causes visible distortions and unintentionally removes key facial attributes.
Ours: Preserves facial structure, expression, and texture; achieving effective identity unlearning while maintaining visual consistency of retained identities.

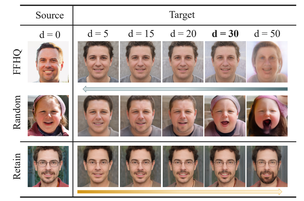
On average, SUGAR can improve the retainability of the model by up to 700%.
Maps each erased identity to a diverse counterfactual on the data manifold—improving retention and limiting adaptive probe-and-collapse attacks.

Appendix: security analysis, ablation study, unlearning robustness, extension with diffusion models, etc.

Sequential Unlearning. The resulting model of previous unlearning stage is used for the next unlearning stage.



Controllable Unlearning. Adapting forgetting strength controlled by parameter d (Eq. 2).



Controllable Unlearn. Larger d values retain more original features, enabling adjustable control over the privacy–utility trade-off.
Continual Unlearn. *GUIDE*: Accumulates collateral damage, progressively degrading retained identities after each request; *SUGAR*: Continues forgetting target identities while consistently preserving retained ones.

Conclusions

- SUGAR, an approach for effectively unlearning multiple identities from a generative model, either simultaneously or sequentially.
- Outperforms SOTA methods on ID similarity and FID, achieving up to 700% improvement in utility preservation (quantitatively and qualitatively)
- TL;DR: (i) effective forgetting with better retention, (ii) learnable and automatic counterfactual identity selection (iii) privacy enhancement.

Github

Contact

Dung (Judy) Nguyen
Vanderbilt University
Email: dung.t.nguyen@vanderbilt.edu
Website: judydnguyen.github.io

Acknowledgements

We acknowledge partial support from the National Security Agency under the Science of Security program, from the Defense Advanced Research Projects Agency under the CASTLE program, the Advanced Research Projects Agency for Health, and from an Amazon Research Award. The contents of this paper do not necessarily reflect the views of the US Government.

References

1. Plamena Zlateva, Liudmila Steshina, Igor Petukhov, and Dimiter Velev. A conceptual framework for solving ethical issues in generative artificial intelligence. In Electronics, Communications and Networks, pages 110–119. IOS Press, 2024.
2. Maria Vittoria Zucca and Gaia Fiorinelli. Regulating ai as a cybersecurity defense: Fighting the misuse of generative ai for cyber attacks and cybercrime. Technology and Regulation, 2025:247–262, 2025.
3. Juwon Seo, Sung-Hoon Lee, Tae-Young Lee, Seungjun Moon, and Gyeong-Moon Park. Generative unlearning for any identity. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 9151–9161, 2024.
4. Alvin Heng and Harold Soh. Selective amnesia: A continual learning approach to forgetting in deep generative models. Advances in Neural Information Processing Systems, 36, 2024.
5. Jeffrey Rosen. The right to be forgotten. Stan. L. Rev. Online, 64:88, 2011.
6. Dawen Zhang, Pamela Finckenberg-Broman, Thong Hoang, Shidong Pan, Zhenchang Xing, Mark Staples, and Xiwei Xu. Right to be forgotten in the era of large language models: Implications, challenges, and solutions. AI and Ethics, pages 1–10, 2024.