SUGAR: A Sweeter Spot for Generative Unlearning of Many Identities

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

{dung.t.nguyen, preston.k.robinette, allison.z.jiang}@vanderbilt.edu
{taylor.johnson, kevin.leach}@vanderbilt.edu quang.ng@rutgers.edu

Abstract

Recent advances in 3D-aware generative models have enabled high-fidelity image synthesis of human identities. However, this progress raises urgent questions around user consent and the ability to remove specific individuals from a model's output space. We address this by introducing SUGAR*, a framework for scalable generative unlearning that enables the removal of many identities (simultaneously or sequentially) without retraining the entire model. Rather than projecting unwanted identities to unrealistic outputs or relying on static template faces,* SUGAR *learns a personalized surrogate latent for each identity, diverting reconstructions to visually coherent alternatives while preserving the model's quality and diversity. We further introduce a continual utility preservation objective that guards against degradation as more identities are forgotten.* SUGAR *achieves state-of-the-art performance in removing up to 200 identities, while delivering up to a 700% improvement in retention utility compared to existing baselines. Our code is publicly available at https://github.com/judydnguyen/SUGAR-Generative-Unlearn.*

1. Introduction

Despite the undeniable benefits of generative AI across various domains such as image synthesis and content creation [20, 23, 40, 44], growing concerns have emerged regarding their potential misuse and unintended consequences [7, 45, 46]. In particular, generative models can inadvertently memorize and reproduce identifiable faces or proprietary images from their training data, posing significant risks to intellectual property rights and personal privacy [5, 15, 32].

These issues have urged the *right to be forgotten* [29, 41], a movement supporting individuals' ability to request the removal of personal data (e.g., images of one's face) from digital systems like AI models. While possible to remove requested images from training sets, retraining models from scratch is computationally impractical, especially if there are multiple requests over time [22, 38].

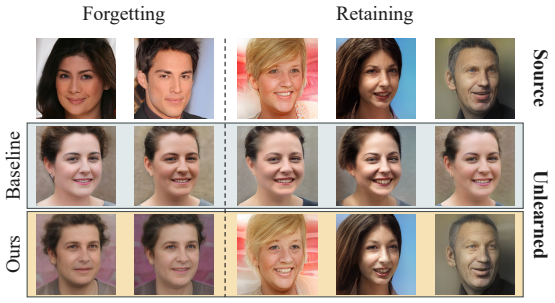


Figure 1. Results on forgetting multiple identities (left column). Our method effectively removes the specified identities from the generative model while maintaining the quality and distinctiveness of the retained identities. In contrast, the baseline (GUIDE) method causes noticeable degradation in the retained identities.

To address the growing need for post-hoc data removal, recent work has explored *generative unlearning*—the task of removing specific information from pretrained generative models without retraining from scratch [2, 13, 30, 42]. Approaches vary in how they suppress unwanted identities. Selective Amnesia [13] targets diffusion and VAE-based models, using continual learning techniques to push erased concepts into noise-like outputs. GUIDE [30], the first to address *Generative Identity Unlearning* (GUI) in GANs, instead manipulates the latent space by redirecting target embeddings toward a fixed mean identity representation. As illustrated in Figure 1, both methods demonstrate initial effectiveness for single-identity removal but face key limitations: (a) they lack support for forgetting multiple identities at once or over time; (b) their suppression mechanisms—mapping to noise or an average face—may leak information about the forgotten identity; and (c) they often degrade unrelated generations, causing semantic drift in retained identities. For example, in GUIDE, retained faces in

the Baseline row of Figure 1 begin to converge toward the median face, reducing output diversity and model utility.

Toward this end, we present SUGAR, which is capable of securely removing multiple identities while maintaining the usability of the model. Unlike previous approaches that introduce artifacts or distortions, SUGAR ensures that forgotten identities are effectively erased without compromising the fidelity, diversity, or usability of the model's outputs. The contributions of this work, therefore, are the following:

1. We introduce SUGAR, which—to the best of our knowledge—is the first GUI method to address *multiple-identity unlearning* and *sequential unlearning* within the context of generative AI.
2. We demonstrate the effectiveness of our approach over SOTA methods using Identity Similarity (ID) and Fréchet Inception Distance (FID) quantitative metrics, **achieving up to a 700% improvement over baselines in maintaining model utility**. Furthermore, our qualitative analysis and human judgment study reinforce these findings, showing that our method effectively unlearns multiple target identities while preserving the visual quality of retained identities.
3. We conduct extensive ablation studies to evaluate controllable unlearning, identity retention, utility preservation, and privacy enhancement, highlighting our method's advantages: (i) effectiveness in forgetting and retention; (ii) learnable and automatic determination of new identities for each forgotten identity; and (iii) privacy preservation throughout the unlearning process.

2. Related Works

Mitigating Misuse of Generative Models. As generative models continue to advance in their ability to synthesize and manipulate highly realistic images, concerns about ethical misuse have intensified. The integration of powerful image editing techniques—such as latent space manipulation [31], language-guided editing [26], and photorealistic face retouching [39]—has made it increasingly accessible for users to modify images of real individuals or apply stylistic transformations [33]. More alarmingly, recent studies have shown that these models can be exploited to extract memorized, copyrighted training data [3], often without the consent of the individuals or content owners involved. One of the most pressing risks is the unauthorized synthesis of a person's likeness in misleading or harmful contexts, such as deepfakes or defamatory media [1, 24, 28]. To counter this, Thanh et al. [36] developed an adversarial attack against UNet, effectively corrupting generated images. Other works in the concept erasure domain focus on eliminating certain visual concepts. For example, Gandikota et al. [8] were among the first to introduce removing specific concepts from diffusion models using negative guidance. This is still an ongoing research direction with follow-up works such as [8, 25, 27] to better improve the balancing between unlearning efficiency and maintaining model utility on remaining concepts. Previous methods are identity-specific, requiring tailoring for each individual, which makes them non-scalable for protecting multiple identities and introduces substantial computational overhead. Moreover, these approaches primarily focus on associations between concepts, such as the "Van Gogh style," and generated images, a focus more relevant to multi-modal learning, like text-to-image models. In contrast, our study centers on image-to-image models.

Generative Unlearning. Machine unlearning is a technique that enables a machine learning model to "forget" or remove the influence of specific data points from its training data, without needing to be retrained from scratch [2, 35, 42]. The usage of approximate unlearning is more preferable in the context of GenAI because these models are often trained on large corpora and datasets, making the retraining approach infeasible [21, 37]. The interest in machine unlearning with generative models has grown recently, where the field was dominated by supervised unlearning before, i.e., classification tasks [4, 9–12, 34]. Seo et al. [30] were the first to address GIU for GAN-based models in an approach called GUIDE. However, their method does not account for practical scenarios involving **(i) simultaneous unlearning requests**—an essential requirement in real-world applications where multiple individuals may request data removal at once—or **(ii) sequential unlearning**, where requests arrive at different times [43]. Our empirical results also demonstrate that while GUIDE is effective for unlearning, it hampers the model's ability to maintain overall generation quality on retained identities.

Recent studies have also addressed generative unlearning by mapping forgotten identities to a Gaussian distribution $\mathcal{N}(0, \sigma)$ while preserving the original distribution for remaining data in an approach called Selective Amnesia (SA) [6, 13]. However, using Gaussian noise to replace forgotten identities is suboptimal, as it fails to accurately represent certain training data distributions, potentially corrupting other identities. Furthermore, SA, as well as GUIDE, **(iii) raise security concerns**, as the association between a forgotten identity and its output (noise or average face) can still be detected. To this end, we introduce SUGAR to address each of these concerns (**i**, **ii**, and **iii**).

3. SUGAR: Architecture and Components

3.1. Problem Formulation

Problem formulation (multi-identity unlearning). We build on the GUI setup from Seo et al. [30] and extend it to handle multiple identity unlearning. We consider a pre-trained EG3D-style pipeline (E_ψ, G_θ, R) and aim to obtain an *unlearned* generator G_θ^u that forgets a set of K tar-

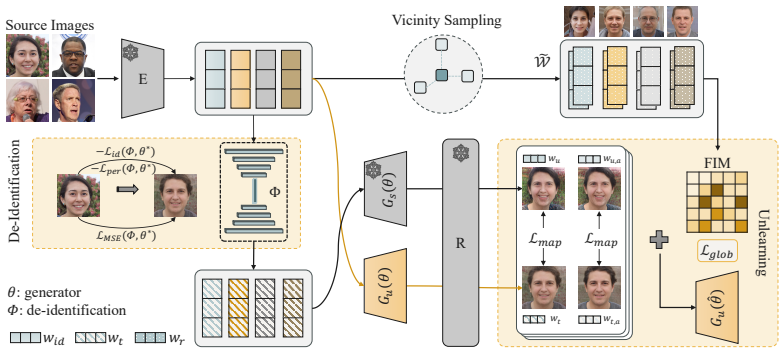


Figure 2. The overall pipeline of SUGAR consists of two phases: (i) ID-specific De-Identification and (ii) Forgetting in a Continual Learning Fashion. In the first phase, a mapping function, Φ, is trained to determine how a forgetting identity should be mapped using the forgetting set $\mathcal{D}_f$ and the pre-trained model G_s. In the second phase, the generator θ of the generative model is updated such that the forgotten identities are mapped to new, targeted identities generated by Φ.

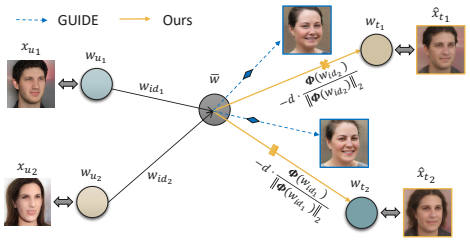


Figure 3. Our de-identification process, denoted as Φ, defines the target latent code w_t for unlearning by linearly subtracting the average identity vector $\bar{w}$ from the counterfactual identity $\Phi(w_{id})$ produced by the model. This approach contrasts with GUIDE, which sets the new target vector by extrapolating from the source latent code along the direction toward $\bar{w}$, maintaining a fixed distance along that direction.

get identities while retaining performance on others. Let $\mathcal{D} = \mathcal{D}_u \cup \mathcal{D}_r$ denote the full dataset, where $\mathcal{D}_u$ contains images of the K identities to remove and $\mathcal{D}_r$ the remainder ($|\mathcal{D}_r| \gg |\mathcal{D}_u|$). We operate in latent space: $W_u := \{\, w = E_\psi(x) \mid x \in \mathcal{D}_u \,\}$ and $W_r := \{\, w = E_\psi(x) \mid x \in \mathcal{D}_r \,\}$. We keep the mapping network and renderer R fixed and update only G.

For brevity, we define the feature operator $\mathcal{F}(w) := G(w)$, which returns the tri-plane feature, and the reconstruction operator $\mathcal{R}(w) := R(\mathcal{F}(w); c)$, which renders the final image. Accordingly, the objectives of multiple-identity unlearning can be formulated as follows.

- *Forget:* For every $w \in W_u$, the image $\hat{x}^u = \mathcal{R}_{G^u}(w)$ should not resemble the identity associated with $\hat{x}^* = \mathcal{R}_{G^*}(w)$.
- *Retain:* For every $w \in W_r$, $\hat{x}^u$ should remain close to $\hat{x}^*$ under standard perceptual/identity/photometric metrics.

Remapping function. To disassociate image generation from an identity $\mathcal{I}_u$ with latent vector w_u, a straightforward and effective strategy—used in **GUIDE** [30]—is to remap w_u to a surrogate target identity I^{avg} (e.g., the median face) with latent vector w_{avg}. With **GUI**, we adopt this idea and denote the GUIDE loss as the *remapping objective*.

$$\mathcal{L}_{map}(w_u, w_{avg}) = \lambda_{mse}\,\mathcal{L}_{mse}\big(\mathcal{F}(w_u), \mathcal{F}(w_{avg})\big) + \lambda_{per}\,\mathcal{L}_{per}\big(\mathcal{R}(w_u), \mathcal{R}(w_{avg})\big) + \lambda_{id}\,\mathcal{L}_{id}\big(\mathcal{R}(w_u), \mathcal{R}(w_{avg})\big). \tag{1}$$

This couples an MSE term in feature space with perceptual and identity terms in image space, aligning intermediate features and steering the generator toward the surrogate identity I^{avg} rather than the original $\mathcal{I}_u$.

However, direct remapping can induce non-smooth behavior across multiple forgotten identities and trade off with retention. To enable scalable removal, we introduce **ID-specific De-Identification**, which assigns a personalized surrogate target to each identity in W_u and guides G^u_θ toward these substitutes while explicitly enforcing the retain constraint on W_r.

3.2. ID-specific De-indentification

Unlike existing unlearning approaches with predefined target concepts, we propose a **learnable** and **sample-specific** procedure to determine where each forgotten sample should be mapped. We call this *ID-specific De-identification*.

Following Seo et al. [30], we define the identity latent vector w_{id} for an individual as the difference between their latent representation w_u and the average latent vector $\bar{w}$, i.e., $w_{id} = w_u - \bar{w}$. In our method, we train a mapping model $\mathcal{T}_\Theta$ to produce a new identity vector w^t_{id} given any input latent vector $w_u \in \mathbb{R}^{d\times h\times w}$, such that:

$$w^t_{id} = \mathcal{T}_\Theta(w_{id}) \in \mathbb{R}^{d\times h\times w} \tag{2}$$

We then define the *de-identification operator* $\Theta(\cdot, w_u)$, which determines the target latent code w_t for the forgotten identity w_u, by shifting the reference latent $\bar{w}$ along a direction derived from w_{id}. Specifically, the operator subtracts a normalized transformation of w_{id}, scaled by a factor $d \in \mathbb{R}^+$, from $\bar{w}$:

$$\Theta(\cdot, w_u) := w_t = \bar{w} - d \cdot \frac{\mathcal{T}_\Theta(w_{id})}{\|\mathcal{T}_\Theta(w_{id})\|} \tag{3}$$

Here, d controls the displacement magnitude in latent space to ensure that the target code w_t removes identifiable traits of the original identity.

The intuition is that we learn a mapping $\mathcal{T}_\Theta$ that assigns each forgotten identity $w_u \in W_u$ a personalized *counterfactual* w_t: a different identity that (i) preserves coarse facial attributes of w_u while (ii) being recognizably distinct by facial recognition metrics such as ID or LPIPS [30]. In latent

space, w_t serves as the surrogate for w_u; after unlearning, G_θ^u synthesizes faces consistent with w_t rather than w_u. To learn this identity replacement, we train a transformation function Θ that maps each $w_u \in W_u$ to a replacement latent code $w_t = \Theta(\cdot, w_u)$ that meets the above criteria. The transformation is supervised using the following objective:

$$\mathcal{L}_{de}(\Theta) = \frac{1}{|W_u|} \sum_{w_u \in W_u} \Big[\lambda_{mse}\, \mathcal{L}_{mse}\big(\mathcal{F}(w_u),\, \mathcal{F}(\Theta(w_u))\big) - \lambda_{per}\, \mathcal{L}_{per}(x_u, \hat{x}_u) - \lambda_{id}\, \mathcal{L}_{id}(x_u, \hat{x}_u) \Big] \quad (4)$$

where the first term, $\mathcal{L}_{mse}$, encourages the replacement latent code to retain similar tri-plane features to w_u, ensuring that the substitute identity remains structurally plausible. The last two terms—$\mathcal{L}_{per}$ (perceptual loss) and $\mathcal{L}_{id}$ (identity loss)—enforce dissimilarity in image space between the reconstructed output $\hat{x}_u = \mathcal{R}(\Theta(w_u))$ and the original image x_u, promoting effective identity removal.

As shown in Figure 3, compared to the baseline GUIDE, our method introduces a more flexible mapping function, enabling the identification of a new target identity that better facilitates the subsequent unlearning process.

3.3. Forgetting in a Continual Learning Fashion

In unlearning, there are always two objectives tied together, which are enforcing forgetting and retention.

Enforcing Identity Forgetting. Building on the remapping objective of Seo et al. [30] in Equation 1, we replace the fixed median-face target with a *learnable* de-identification target produced by our operator $\Theta(\cdot)$; that is, the average face $\mathcal{I}^{avg}$ is replaced by $\Theta(w)$. As noted in the GUIDE baseline, successful unlearning must also suppress identity information in the local neighborhood of each forgotten latent. We therefore sample a perturbed neighbor $w_{u,a}$ for each forgotten code w_u:

$$w_{u,a} \;=\; w_u \;+\; \alpha_n \,\frac{w_{r,a} - w_u}{\big\|w_{r,a} - w_u\big\|_2}, \qquad \alpha_n \sim \mathcal{U}(0, \alpha_{\max}),$$

where $w_{r,a}$ is a randomly chosen reference latent that supplies a direction; this produces marginal perturbations in the vicinity of w_u to ensure robustness across local variations. We apply the de-identification operator Θ to both the anchor code w_u and its latent-space neighbor $w_{u,a}$, producing the learnable targets $\Theta(w_u)$ and $\Theta(w_{u,a})$. We then extend the remapping objective in Equation 1 to both targets, defining the overall forgetting loss:

$$\mathcal{L}_{forget}(w_u; \theta) = \mathcal{L}_{map}\big(w_u, \Theta(w_u)\big) + \lambda_{nei}\, \mathcal{L}_{map}\big(w_{u,a}, \Theta(w_{u,a})\big). \quad (5)$$

For brevity and further ablation study, we denote the second term as the neighbor-remapping loss, $\mathcal{L}_{nei} := \lambda_{nei}\, \mathcal{L}_{map}\big(w_{u,a}, \Theta(w_{u,a})\big)$.

Maintaining Model Utility. During unlearning, we observe that samples whose latent representations are close to those in the forgetting set W_u are more likely to experience performance degradation, whereas samples farther away are typically unaffected [1]. To preserve the model's utility, we propose a strategy to protect these nearby samples from unintended disruption.

Vicinity Sampling. Unlearning a set of identities can disrupt nearby latents belonging to other identities; in particular, identities whose codes lie close to the forgotten set are most prone to collateral degradation (cf. Appendix D.3). To proactively guard against this effect, we construct a *vicinity set* $\tilde{W}$ of perturbation probes around each forgotten latent and use them as the retention objective. Formally, let $d^i \sim \mathcal{N}(0, I)$ be a random direction and $\hat{d}^i := d^i / \|d^i\|_2$ the corresponding unit vector. For a step radius $\alpha_r > 0$, we define the probe set:

$$\widetilde{W} \;:=\; \Big\{ \widetilde{w}^{\,i} \;\Big|\; \widetilde{w}^{\,i} \;=\; w_u^i \;+\; \alpha_r\, \hat{d}^i \Big\}_{i=1}^{K}. \quad (6)$$

Intuitively, $\widetilde{W}$ samples points on a small spherical shell centered at w_u^i. These probes are incorporated into the retention objective to prevent unlearning W_f from inadvertently distorting neighboring latents and, in turn, degrading identities not intended to be forgotten.

While one might preserve nearby samples by adding explicit reconstruction or retain losses (e.g., enforcing the same reconstruction quality via Equation 5), this introduces two drawbacks: (1) the computational cost grows linearly with the number of protected samples, becoming infeasible at scale; and (2) directly constraining generation can interfere with effective unlearning.

Elastic Weight Consolidation (EWC). To efficiently preserve behavior in regions most susceptible to collateral damage, we adopt Elastic Weight Consolidation (EWC) [19]. We treat the vicinity probe set $\widetilde{W}$ as a proxy for prior behavior to be retained and regularize parameters that are important for agreement with the source model around forgotten latents.

Let θ^* denote the pre-unlearning generator parameters. We define a reference (retain) loss that measures alignment with the source model on a probe $w \in \widetilde{W}$: $\ell_{\text{ref}}(w; \theta) := \mathcal{L}_{map}(\widetilde{w}^{\,i}, \widetilde{w}^{\,i})$, $\forall\, \widetilde{w}^{\,i} \in \widetilde{W}$. We approximate the diagonal Fisher information over $\widetilde{W}$ at $\theta = \theta^*$ by the average squared gradient:

$$FIM_i \;\approx\; \frac{1}{|\widetilde{W}|} \sum_{w \in \widetilde{W}} \left(\frac{\partial\, \ell_{\text{ref}}(w; \theta)}{\partial \theta_i}\Big|_{\theta=\theta^*} \right)^2, \quad (7)$$

computed with a mini-batch estimator in practice. Given these importances, the EWC penalty constrains drift from

[1] See Appendix D.3 for further analysis.

θ^* proportionally to F_i:

$$\mathcal{L}_{ewc} = \frac{1}{2}\sum_i F_i \left(\theta_i - \theta_i^*\right)^2. \quad (8)$$

This regularizer encourages parameters most critical for preserving behavior on the *vicinity* set to remain near their pre-unlearning values, thereby avoiding per-sample retain losses during unlearning and reducing collateral distortion around forgotten identities.

Final Objective. Combining the forgetting loss over the identities in W_u with the EWC regularization, the final unlearning objective is:

$$\mathcal{L}_{unlearn} = \mathbb{E}_{w_u \sim W_u}\left[\, \mathcal{L}_{forget}(w_u;\theta)\,\right] + \lambda_{ewc}\,\mathcal{L}_{ewc}(\theta;\theta^*). \quad (9)$$

This formulation enables us to mitigate side effects on samples near the forgetting set while preserving overall model utility, without incurring the high cost of sample-specific loss terms.

Overall, SUGAR effectively forgets identities in a pre-trained EG3D generator by (1) leveraging a new ID-specific DeIdentification process that maps a target identity to forget to a new identity in the latent space, and (2) combining forgetting and retaining loss function. This approach ensures identities can be effectively removed while ensuring model utility on other identities.

4. Experiments

4.1. Experimental Setup

Datasets. Following the settings of Seo et al. [30], we evaluated GUIDE in three forgetting scenarios: Random, InD (in-domain), and OOD (out-of-domain). These scenarios represent different sources of identities to unlearn. Specifically, identities are sampled randomly from the latent space (Random), from the FFHQ dataset [18] used for pre-training (InD), or from the CelebA-HQ dataset [17] as unlearning targets (OOD). For both in-distribution (InD) and out-of-distribution (OOD) settings, we use a GAN inversion network $\mathcal{I}$ to map each forgetting image $x_u^i \in \mathcal{D}_u$ to a latent code $w_u^i = E_\psi(x_u^i)$, forming the forget set $W_u = \{w_u^i\}_{i=1}^N$; the retaining set W_r is sampled from the same datasets and inverted analogously.

Baselines. We select GUIDE [30] as our baseline for comparison, as it aligns with our focus on identity-specific unlearning. GUIDE effectively removes identity-related features while preserving the overall structure and quality of the generated images. In contrast, other methods, such as Selective Amnesia (SA) [13] and Feng et al. [6], map forgotten identities to Gaussian noise. This approach often results in unnatural artifacts and disrupts facial reconstruction, making it unsuitable for our task. Specifically, mapping to Gaussian noise prevents the generation of coherent and realistic reconstructions after unlearning, which is a critical requirement in our setting.

Evaluation Metrics. We use Identity Similarity (ID), computed with the face recognition network CurricularFace [16], to measure the similarity between images of faces generated from the same latent codes before and after unlearning. Additionally, we use the Fréchet Inception Distance (FID) score [14] to quantify the overall visual quality and distribution alignment of the generated images. These metrics are calculated on both the forgetting set $\mathcal{D}_u$ and the retaining set $\mathcal{D}_r$, where $\mathcal{D}_r$ is sampled from the same datasets to which $\mathcal{D}_u$ belongs.

4.2. Experimental Results

4.2.1. Balancing Forgetting and Retaining

To evaluate the ability of SUGAR to balance the trade-off between forgetting and retaining in unlearning tasks, we conducted experiments with an increasing number of forgotten identities, where $K \in \{1, 5, 10, 20, 50\}$, across three categories: *In-domain Distribution*, *OOD Distribution*, and *Random*. We then compare our method with the baseline both quantitatively and qualitatively, and incorporate human judgment.

Enhancing Retained Model Utility. First, we report the improvement of our method in maintaining model utility in Table 1 in the generative identity unlearning task across three categories. Our method consistently outperforms the baseline (GUIDE) in maintaining higher ID scores and lower FID values for retained identities, even in scenarios where only one identity is forgotten. As the number of forgotten identities (N) increases, unlearning becomes more challenging, leading to lower identity similarity and higher image distortion. However, our method still achieves better performance, with ID scores improving by 50–150% over GUIDE, especially in the CelebAHQ dataset, where ID scores are roughly 2x higher for $N = 1$ to $N = 20$. Additionally, our method demonstrates a substantial reduction in FID values, highlighting its ability to preserve image quality while unlearning more identities. **On average, our method can improve the retainability of the model by up to 700% across these scenarios.** This observation is represented clearly in the qualitative results in Figure 4. The images in the second row exhibit visually apparent distortions, suggesting that the baseline approach inadvertently removes crucial facial attributes beyond the intended identity modification. In contrast, our method, shown in the third row, maintains a higher level of visual consistency, preserving defining characteristics such as facial structure, expression, and texture. This indicates that the proposed approach is more effective in unlearning forgotten identities without dramatically affecting the retained identities.

Ensuring Identity Forgetting. To evaluate the model's performance in effectively unlearning identities, we present

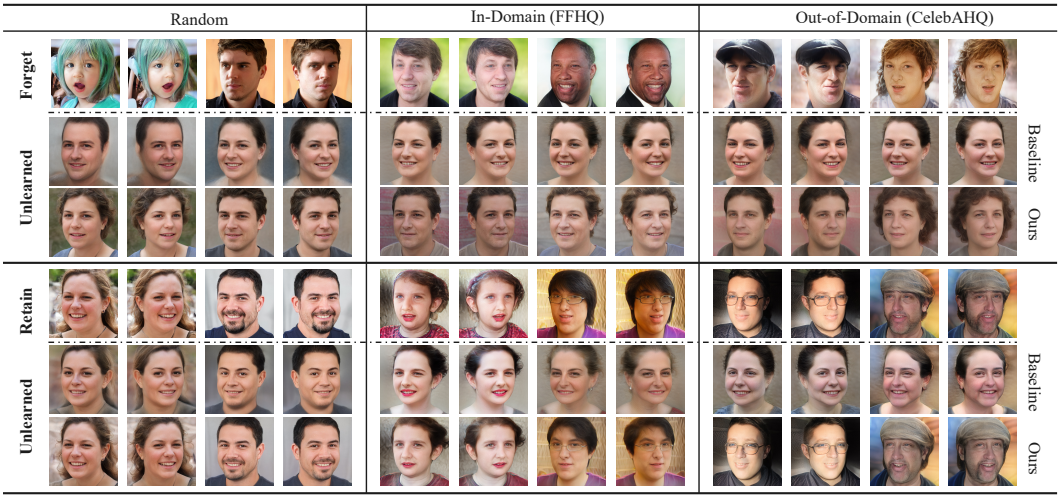


Figure 4. Qualitative results comparing our method with the baseline for the unlearning multiple identities task. The first block illustrates identities targeted for forgetting, while the second block shows identities to be retained. Within each block, the first row contains the original images, while the second and third rows present the generated images corresponding to these identities.

Table 1. Quantitative comparison of model utility in generating **unforgotten** identities between our method and the baseline (GUIDE) in the generative identity unlearning task, as the number of forgotten identities (N) increases.

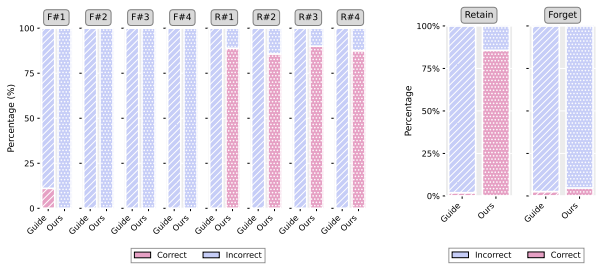
#IDS	Methods	FFHQ (In-domain Distribution)		CelebAHQ (OOD Distribution)		Random	
		ID (↑)	FID (↓)	ID (↑)	FID (↓)	ID (↑)	FID (↓)
N=1	GUIDE	0.3301 ±0.0073	112.06 ±5.1507	0.5118 ±0.0027	133.16 ±4.9765	0.6061 ±0.0061	66.531 ±2.4249
	Ours	0.5730 ±0.0103	94.873 ±0.3855	0.7057 ±0.0055	66.280 ±4.3198	0.6457 ±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*

quantitative results in Table 2. On average, ID scores for forgotten identities are 54.7% lower than those for retained identities, indicating that SUGAR effectively prevents the generation of images from forgotten identities. We argue that forgetting an identity does not require aggressively minimizing metrics like ID, as this could harm utility, resulting in the degraded performance demonstrated in Figure 4. This raises the question of when an identity can be considered sufficiently forgotten. Since there is no clear threshold for ID scores at which humans perceive an unlearned image as distinct rather than identical, determining sufficiency is inherently subjective. To address this, we conducted a human recognition study to assess (i) whether unlearned identities are no longer identifiable by humans and (ii) whether retained identities remain recognizable. This evaluation provides a more robust, quantitatively grounded assessment of each model's effectiveness in identity unlearning beyond relying solely on ID scores.

Table 2. Quantitative comparison of the forgetting ability of the unlearned model, measured on **forgotten** identities, between our method and the baseline (GUIDE) in the generative identity unlearning task, as the number of forgotten identities (N) increases.

Metric	Model	Number of Forgetting IDs				
		N=1	N=5	N=10	N=20	N=50
ID	GUIDE	0.2773 *±0.0099*	0.0040 *±0.0077*	0.0095 *±0.0017*	-0.0275 *±0.0034*	-0.0431 *±0.0018*
	Ours	0.3576 *±0.0146*	0.2664 *±0.0134*	0.2559 *±0.0078*	0.2917 *±0.0050*	0.2433 *±0.0051*
FID	GUIDE	133.90 *±6.9372*	177.27 *±2.2242*	163.83 *±0.9667*	143.44 *±0.8456*	146.43 *±0.6706*
	Ours	115.98 *±7.7846*	125.98 *±0.8796*	125.93 *±0.3963*	105.40 *±0.6074*	106.12 *±1.4283*

Human Judgments' Results. In this study, we sample the generated images for 20 forgotten identities and 20



(a) Accuracy of Selected Identities (b) Overall Accuracy

Figure 5. Human judgment result on selected forgetting (F#) and retaining (R#) identities.

retaining identities produced by our unlearned model and GUIDE's unlearned model. We recruited 60 participants and received 579 responses. Each participant was asked to select the image among five sampled images that they believe depicts the same person given a set of five random identities' images. Interestingly, participants were unable to consistently identify the original identity among the unlearned images for either GUIDE or SUGAR, demonstrating comparable effectiveness in identity forgetting. As shown in Figure 5, performance on the forgetting set is indistinguishable between the two methods. The individual breakdown in Figure 5a reveals that 80–100% of participants failed to match original images to any unlearned outputs. However, for retained identities, Specifically, participants correctly identified SUGAR-generated identities 86% of the time, whereas for GUIDE, this rate is only 1.9%. These results highlight that **SUGAR effectively balances identity forgetting while preserving overall model utility.**

5. Ablation Study

In this section, we conduct ablation studies to evaluate controllable unlearning, identity retention, global utility preservation, and sequential unlearning.

$d-$Controllable Unlearning. Our method can be used as a controllable unlearning solution in I2I generative models, where d is the control coefficient. In Equation 3, with a well-trained model Φ, the targeted identity for each identity w_{id} is adjusted with a different coefficient d. As shown in Figure 6, the higher the value of d, the more features of the forgetting ID w_{id} are retained in the new mapping identity w_t. With our method, the model vendor can select this controllable parameter to balance the trade-off and the forgetting level that reaches a consensus between the vendor and the requested unlearning party. However, there is a trade-off between forgetting and retaining utility, as discussed in previous work [6, 42]. This relationship is presented in Table 3. While reducing the distance can make the model forget the forgetting set $\mathcal{D}_f$ more drastically (i.e., lower ID and higher FID for the forgetting set), this leads to a greater impact on the retaining set (i.e., lower ID). It is worth noting that even

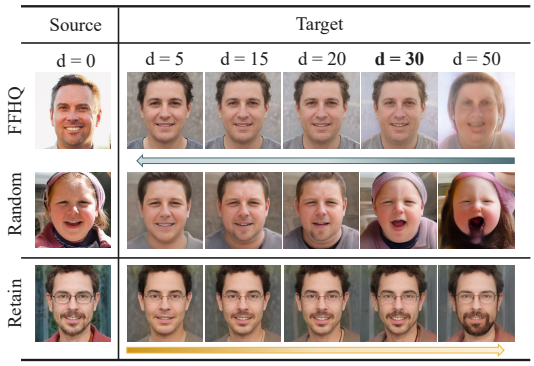


Figure 6. Visualization of our controllable unlearning. The effect of unlearning in I2I generative models is controlled by d (Equation 2) for the vectors generated by our de-identification models. With our approach, the model vendor can selectively control the forgetting level. The higher the value of d (i.e., up to 30), the more original features of forgetting people are retained.

Table 3. Ablation study on the effect of d-controllable value in balancing forgetting and retaining capabilities. We used FFHQ and Random datasets in this experiment.

Metric	Dataset	Value of d						
		d=-5	d=5	d=10	d=15	d=20	d=30	d=50
ID	Forget	0.1815	0.2408	0.2713	0.3194	0.3904	0.4472	0.5124
	Retain	0.4501	0.4960	0.5107	0.5542	0.5880	0.7073	0.5949

with $d = 30$, the generated images are much different from the original identities.

Dual Objective: Identity Removal and Global Utility Preservation. We study the impact of our two loss terms: forgetting neighbor loss $\mathcal{L}_{nei}$ for thorough identity forgetting and the global preservation loss $\mathcal{L}_{\text{ewc}}$ for utility retention. To evaluate identity removal, we conduct a multi-image test on CelebAHQ identities, assessing whether our method forgets other images of the same identity that are not explicitly included in W_u. Figure 7 summarizes the results. "Ours-v1" denotes the variant *without* the neighbor loss $\mathcal{L}_{nei}$ in Equation 5, so the forgetting objective applies only to the anchor images of each target identity. Without this term, residual attributes of the forgotten identity can persist, even when the generated images depict distinct individuals. Adding $\mathcal{L}_{nei}$ enforces consistency across the local neighborhood in latent space, which is crucial for the ultimate goal of removing an entire identity rather than only a single image. To analyze $\mathcal{L}_{ewc}$, Figure 8 compares reconstructions with (Ours) and without (Ours-v2) the global preservation loss. The first two retained identities come from the same distribution as the forgotten ones, while the last three (IDs 3–5) are randomly sampled latents. Random identities remain largely unaffected by unlearning (even under GUIDE), whereas closer identities (IDs 1–2) are more susceptible to collateral forgetting. Ours-v2 already preserves these better than GUIDE due to our de-identification, and adding $\mathcal{L}_{ewc}$ further improves retention by preserving fine details (e.g.,

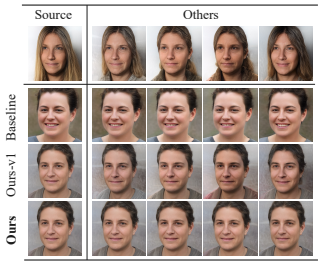


Figure 7. Reconstructed images for the forgotten identity and their unseen neighbor samples (Others) from the same person to demonstrate the thoroughness of identity removal. "Ours-v1" denotes the variant without the neighbor (vicinity) loss from Equation 5.

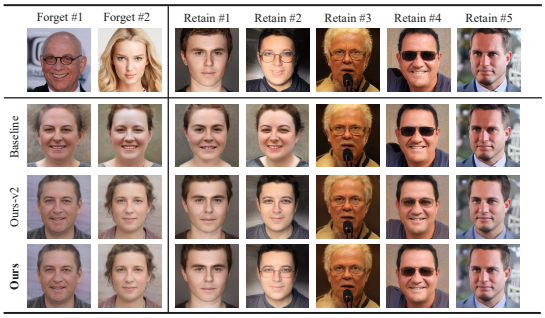


Figure 8. Qualitative results when removing the $\mathcal{L}_{ewc}$ term from the full objective (Eq. 9); we denote this variant as *Ours-v2*. The full model (with $\mathcal{L}_{ewc}$) better preserves retained identities, maintaining fine-grained details (e.g., glasses, hair, wrinkles).

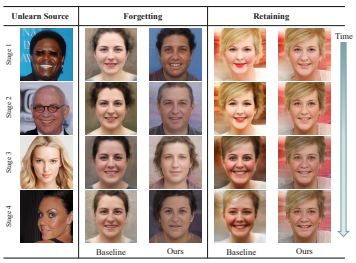


Figure 9. Qualitative results for sequential unlearning setting from stage 1 to stage 4: in each stage, we compare the reconstructed images for both forgetting and retaining identities. *The resulting model of previous unlearning stage is used for the next unlearning stage.*

glasses, hair, wrinkles). We provide a detailed quantitative result for our ablation study in Appendices B.1. and B.2.

Sequential unlearning. We consider the setting where unlearning requests arrive *sequentially* rather than in a single batch. Let $G_\theta^{(0)} = G_\theta^*$ be the pretrained model and let $\mathcal{D}_f^{(k)}$ denote the (new) identities to forget at stage $k{=}1, \ldots, 4$. At each stage, we initialize from the previously unlearned model and run the *same* unlearning procedure (architecture, losses, schedule, and hyperparameters unchanged): $G_\theta^{(k)} \leftarrow \textsc{Unlearn}\big(G_\theta^{(k-1)}, \mathcal{D}_f^{(k)}\big)$. Figure 9 shows qualitative results after four stages for GUIDE and SUGAR. While GUIDE accumulates collateral damage—progressively eroding features of the *retained* identities across stages—SUGAR continues to forget the requested identities while preserving the appearance of the retained set. Detailed setup and additional analysis are provided in Appendix B.5.

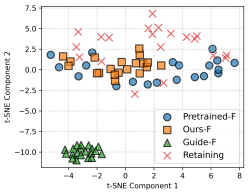


Table 4. Quantitative Results

	Methods	Metrics			
		MSE	PSNR	LPIPS	SSIM
Forgetting	GUIDE	254.31 ± 376.36	25.4262 ± 0.2026	0.2048 ± 0.0000	25.4262 ± 0.2026
	Ours	1022.81 ± 55301.1	18.5459 ± 0.8418	0.2936 ± 0.0002	0.7156 ± 0.0002

Figure 10. Privacy analysis of images generated by GUIDE and SUGAR. The left figure is a t-SNE visualization of intermediate tri-plane representation from forgetting (−F) and retaining identities; from the pre-trained model (before unlearning), the unlearned model produced by SUGAR, and the baseline GUIDE.

Privacy Discussion In this section, we assess the security of SUGAR in concealing forgotten identities. Prior unlearning methods (e.g., GUIDE, Selective Amnesia) collapse erased identities to an average face or noise, creating a recognizable *signature* that enables simple erasure-detection and membership-inference attacks (e.g., nearest-centroid tests, one-class density checks) by pushing outputs into low-density, out-of-manifold regions. In contrast, SUGAR maps each erased identity to a *close, diverse counterfactual* that remains on the data manifold, eliminating a single surrogate template and reducing separability from natural variation. Empirically, Table 4 shows larger divergence from each identity's pretrained counterpart yet higher *intra-forgotten* variability (higher MSE/LPIPS, lower PSNR/SSIM) than GUIDE, which makes unsupervised identification harder; Figure 10 likewise shows GUIDE's forgotten samples forming a distinct cluster, while SUGAR intermixes with retained samples. This distributional indistinguishability limits adaptive "probe-and-collapse" and template-matching attacks, preserves fidelity for retained identities (avoiding collateral degradation seen with noise/mean collapse), and scales securely since a single counterfactual generator yields varied surrogates without a deterministic average-face cue. To this end, these properties make SUGAR more secure and privacy-preserving than average/noise-based unlearning, challenging membership inference attacks.

6. Conclusion

In this work, we introduce SUGAR, an approach for effectively unlearning multiple identities from a generative model, either simultaneously or sequentially. Unlike existing methods, SUGAR autonomously determines optimal target representations for each forgotten identity, ensuring a structured and efficient continual unlearning process. Through qualitative and quantitative evaluations, we demonstrate that SUGAR outperforms baselines in preserving model utility while achieving effective identity removal. Furthermore, our analysis provides deeper insights into how unlearning impacts different regions of the data distribution, offering a foundation for future research on adaptive target determination in unlearning generative models.

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] Fakhar Abbas and Araz Taeihagh. Unmasking deepfakes: A systematic review of deepfake detection and generation techniques using artificial intelligence. *Expert Systems with Applications*, page 124260, 2024. 2

[2] Lucas Bourtoule, Varun Chandrasekaran, Christopher A Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. Machine unlearning. In *2021 IEEE symposium on security and privacy (SP)*, pages 141–159. IEEE, 2021. 1, 2

[3] Nicolas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramer, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. In *Proceedings of the 32nd USENIX Security Symposium (USENIX Security 23)*, pages 5253–5270, 2023. 2

[4] Vikram S Chundawat, Ayush K Tarun, Murari Mandal, and Mohan Kankanhalli. Zero-shot machine unlearning. *IEEE Transactions on Information Forensics and Security*, 2023. 2

[5] Daniel DeAlcala, Gonzalo Mancera, Aythami Morales, Julian Fierrez, Ruben Tolosana, and Javier Ortega-Garcia. A comprehensive analysis of factors impacting membership inference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3585–3593, 2024. 1

[6] XiaoHua Feng, Yuyuan Li, Chaochao Chen, Li Zhang, Longfei Li, JUN ZHOU, and Xiaolin Zheng. Controllable unlearning for image-to-image generative models via ϵ-constrained optimization. In *The Thirteenth International Conference on Learning Representations*, 2025. 2, 5, 7

[7] Emilio Ferrara. Genai against humanity: Nefarious applications of generative artificial intelligence and large language models. *Journal of Computational Social Science*, 7(1):549–569, 2024. 1

[8] Rohit Gandikota, Joanna Materzynska, Jaden Fiotto-Kaufman, and David Bau. Erasing concepts from diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2426–2436, 2023. 2

[9] Aditya Golatkar, Alessandro Achille, and Stefano Soatto. Eternal sunshine of the spotless net: Selective forgetting in deep networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9304–9312, 2020. 2

[10] Aditya Golatkar, Alessandro Achille, and Stefano Soatto. Forgetting outside the box: Scrubbing deep networks of information accessible from input-output observations. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 383–398. Springer, 2020.

[11] Aditya Golatkar, Alessandro Achille, Avinash Ravichandran, Marzia Polito, and Stefano Soatto. Mixed-privacy forgetting in deep networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 792–801, 2021.

[12] Varun Gupta, Christopher Jung, Seth Neel, Aaron Roth, Saeed Sharifi-Malvajerdi, and Chris Waites. Adaptive machine unlearning. *Advances in Neural Information Processing Systems*, 34:16319–16330, 2021. 2

[13] 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. 1, 2, 5

[14] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in Neural Information Processing Systems*, 30, 2017. 5

[15] Hongsheng Hu, Zoran Salcic, Lichao Sun, Gillian Dobbie, Philip S Yu, and Xuyun Zhang. Membership inference attacks on machine learning: A survey. *ACM Computing Surveys (CSUR)*, 54(11s):1–37, 2022. 1

[16] Yuge Huang, Yuhan Wang, Ying Tai, Xiaoming Liu, Pengcheng Shen, Shaoxin Li, Jilin Li, and Feiyue Huang. Curricularface: adaptive curriculum learning loss for deep face recognition. In *proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5901–5910, 2020. 5

[17] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018. 5

[18] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 5

[19] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017. 4

[20] Yeonkyeong Lee, Taeho Choi, Hyunsung Go, Hyunjoon Lee, Sunghyun Cho, and Junho Kim. Exp-gan: 3d-aware facial image generation with expression control. In *Proceedings of the Asian Conference on Computer Vision*, pages 3812–3827, 2022. 1

[21] Jiaqi Li, Qianshan Wei, Chuanyi Zhang, Guilin Qi, Miaozeng Du, Yongrui Chen, Sheng Bi, and Fan Liu. Single image unlearning: Efficient machine unlearning in multimodal large language models. *Advances in Neural Information Processing Systems*, 37:35414–35453, 2025. 2

[22] Na Li, Chunyi Zhou, Yansong Gao, Hui Chen, Zhi Zhang, Boyu Kuang, and Anmin Fu. Machine unlearning: Taxonomy, metrics, applications, challenges, and prospects. *IEEE Transactions on Neural Networks and Learning Systems*, 2025. 1

[23] Wentong Liao, Kai Hu, Michael Ying Yang, and Bodo Rosenhahn. Text to image generation with semantic-spatial aware gan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18187–18196, 2022. 1

[24] Yixin Liu, Chenrui Fan, Yutong Dai, Xun Chen, Pan Zhou, and Lichao Sun. Metacloak: Preventing unauthorized subject-driven text-to-image diffusion-based synthesis via meta-learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24219–24228, 2024. 2

[25] Shilin Lu, Zilan Wang, Leyang Li, Yanzhu Liu, and Adams Wai-Kin Kong. Mace: Mass concept erasure in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6430–6440, 2024. 2

[26] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2085–2094, 2021. 2

[27] Vitali Petsiuk and Kate Saenko. Concept arithmetics for circumventing concept inhibition in diffusion models. In *European Conference on Computer Vision*, pages 309–325. Springer, 2024. 2

[28] Felipe Romero Moreno. Generative ai and deepfakes: a human rights approach to tackling harmful content. *International Review of Law, Computers & Technology*, 38(3):297–326, 2024. 2

[29] Jeffrey Rosen. The right to be forgotten. *Stan. L. Rev. Online*, 64:88, 2011. 1

[30] 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. 1, 2, 3, 4, 5

[31] Yujun Shen, Ceyuan Yang, Xiaoou Tang, and Bolei Zhou. Interfacegan: Interpreting the disentangled face representation learned by gans. *IEEE transactions on pattern analysis and machine intelligence*, 44(4):2004–2018, 2020. 2

[32] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pages 3–18. IEEE, 2017. 1

[33] Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Diffusion art or digital forgery? investigating data replication in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6048–6058, 2023. 2

[34] Ayush K Tarun, Vikram S Chundawat, Murari Mandal, and Mohan Kankanhalli. Fast yet effective machine unlearning. *IEEE Transactions on Neural Networks and Learning Systems*, 2023. 2

[35] Piyush Tiwary, Atri Guha, Subhodip Panda, et al. Adapt then unlearn: Exploiting parameter space semantics for unlearning in generative adversarial networks. *arXiv preprint arXiv:2309.14054*, 2023. 2

[36] Thanh Van Le, Hao Phung, Thuan Hoang Nguyen, Quan Dao, Ngoc N Tran, and Anh Tran. Anti-dreambooth: Protecting users from personalized text-to-image synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2116–2127, 2023. 2

[37] Jie Xu, Zihan Wu, Cong Wang, and Xiaohua Jia. Machine unlearning: Solutions and challenges. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2024. 2

[38] Yuanshun Yao, Xiaojun Xu, and Yang Liu. Large language model unlearning. *Advances in Neural Information Processing Systems*, 37:105425–105475, 2025. 1

[39] Jaejun Yoo, Youngjung Uh, Sanghyuk Chun, Byeongkyu Kang, and Jung-Woo Ha. Photorealistic style transfer via wavelet transforms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9036–9045, 2019. 2

[40] Bowen Zhang, Shuyang Gu, Bo Zhang, Jianmin Bao, Dong Chen, Fang Wen, Yong Wang, and Baining Guo. Styleswin: Transformer-based gan for high-resolution image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11304–11314, 2022. 1

[41] 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. 1

[42] Haibo Zhang, Toru Nakamura, Takamasa Isohara, and Kouichi Sakurai. A review on machine unlearning. *SN Computer Science*, 4(4):337, 2023. 1, 2, 7

[43] Yihua Zhang, Chongyu Fan, Yimeng Zhang, Yuguang Yao, Jinghan Jia, Jiancheng Liu, Gaoyuan Zhang, Gaowen Liu, Ramana Rao Kompella, Xiaoming Liu, and Sijia Liu. Unlearncanvas: Stylized image dataset for enhanced machine unlearning evaluation in diffusion models. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. 2

[44] Xiaoming Zhao, Fangchang Ma, David Güera, Zhile Ren, Alexander G Schwing, and Alex Colburn. Generative multiplane images: Making a 2d gan 3d-aware. In *European conference on computer vision*, pages 18–35. Springer, 2022. 1

[45] 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. 1

[46] 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. 1