Supplemental Materials:
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

Table 1. Notation Table

Notation	Definition
$\mathcal{D}_u$	Unlearning set, containing images of unlearning identities
$\mathcal{D}_r$	Retaining set, containing images of retaining identities
W_u	Unlearning set, containing latent vectors of unlearning identities
W_r	Unlearning set, containing latent vectors of retaining identities
E_ψ	Encoder, inversion network
G_θ	Generator
G_θ^u	Unlearned Generator
$\mathcal{F}(w)$	Operator to extract tri-plane feature
$\mathcal{R}(w) := R(\mathcal{F}(w); c)$	Rendering operator to reconstruct image
w_{avg}, $\bar{w}$	Median face's latent vector
w_u	Unlearned identity's latent representation
w_{id}	Unlearned identity's ID vector, i.e., $w_{id} = w_u - w_{avg}$
w_{id}^t	Surrogate identity's latent vector
$\mathcal{I}^u$	Unlearned identity
$\mathcal{I}^{avg}$	Identity of the median face
$\Theta(\cdot, w_u)$	*de-identification operator* which determines the target latent code w_t for the forgotten identity w_u
$\mathcal{T}_\Theta$	De-identification Network, (i.e., UNet)
$\widetilde{W}$	Vicinity sampling set for retention

A. Implementation Details

We conduct all the experiments using PyTorch 2.1.0 [2]. All experiments are run on a computer with an Intel Xeon Gold 6330N CPU and an NVIDIA A6000 GPU.

A.1. Models and Hyper-parameters

We implemented the unlearning framework based on source code provided by Seo et al. [5]. We keep all training parameters to ensure a fair comparison. In specific, the generative model was built on a 3D generative adversarial network [1] pre-trained on FFHQ dataset [3]. We leveraged GOAE [7] as a GAN inversion network to obtain the latent code from an image and kept this component frozen during fine-tuning. The image resolution for all experiments is 512x512 with a rendering resolution of 128x128. We used Adam optimizer with a learning rate of 10^{-4}. The hyperparameters used in the experiments were: $d = 25$, $a_n = 15$, $\alpha_r = 30$, $\lambda_{nei} = 0.1$, $\lambda_{id} = 0.1$, $\lambda_{mse} = 0.01$, and $\lambda_{mse} = 50$.

A.2. Mapping Function in De-Identification Process

We implement a generating function model $\mathcal{T}_\Phi$ using a U-Net architecture [4], which takes as input a $\mathbb{R}^{14\times512}$ feature map from the encoder and outputs a transformed vector of the same dimension. We also experiment with alternative architectures, such as Transformers, but observe no significant performance difference.

Training is conducted using the Adam optimizer with a batch size of 2, a learning rate of 10^{-4}, and a maximum of 200 epochs. While more complex architectures like Transformers [6] can be used, our results indicate that a simple U-Net performs comparably when trained with the proposed algorithm.

Our de-identification model, $\mathcal{T}_\Phi$, can be implemented using alternative deep learning architectures, provided that the output preserves the required dimensions and maintains smooth transitions in target images, as illustrated in Figure 11.

A.3. Human Identity Recognition Study

We conducted a human recognition study comparing GUIDE and SUGAR. This study was IRB-approved, and participant details will be disclosed upon acceptance.

We recruited 59 participants from various universities and countries. Each participant completed the survey remotely without external pressure, and all responses were anonymous. Instructions and examples were provided in the welcome page prior to beginning the survey.

To construct the set of questions for the study, 20 unlearned identities (forgetting set) and 20 identities unseen during training (retaining set) were selected. GUIDE and SUGAReach generated images from these 40 identities, resulting in a total of 80 synthetic images. An additional 6 control identities were added, where the post-processing images were either identical to the original (positive control) or an entirely unrelated image (negative control).

These were sourced from the original CelebA dataset instead of synthetically generated.

Each participant evaluated five identities presented twice: once with the GUIDE-generated result, and once with the SUGAR-generated result. Each participant also answered three control questions. Of the 59 participants who began the study, 43 completed it in full. We received a total of 579 responses spread across 86 questions, for an average of 6-7 responses each question. For each question, participants were presented with the original identity alongside five synthetic candidate images, and asked to select which candidate image corresponded to the original image. If none matched, they were requested to select "None of the above." Among the candidate images, one was generated from the original image. The remaining four were generated from visually similar identities from CelebA, identified using DeepFace feature similarities to ensure demographic consistency in terms of race and sex.

This study design allowed us to measure whether human participants were able to recognize an identity before and after the unlearning process, testing the effectiveness of each model's ability to retain or forget target identities.

B. Additional Ablation Study

In this section, we expand the quantitative ablations from the main paper and include further experiments on single-identity and large-scale unlearning, a sequential-unlearning setting, and a potential application of our approach.

B.1. Ablation Result for Component Removal of Unlearning Loss

Table 2 presents the quantitative results for our forgetting ability under three considered scenarios including Random, In-distribution and Out-of-distribution. Figure 3 presents the numerical results for model performance given different values of λ_{mse}. To evaluate the role of each component in $\mathcal{L}_{unlearn}$, we perform ablation studies (Table 4) by progressively removing loss terms and reporting the corresponding forgetting and retention performance. Specifically, Ours-v1 excludes the neighbor loss $\mathcal{L}_{nei}$, Ours-v2 removes the EWC regularization $\mathcal{L}_{ewc}$, and Ours-v3 replaces the de-identification operator $\Theta(\cdot)$ with the median latent as in the baseline GUIDE. From the results, we observe that replacing $\Theta(\cdot)$ in Ours-v3 leads to the most significant degradation in forgetting effectiveness, confirming that the de-identification operator is the most critical component to enforce identity erasure. Eliminating $\mathcal{L}_{nei}$ in Ours-v1 weakens the consistency of forgetting across the latent neighborhood, while excluding $\mathcal{L}_{ewc}$ in Ours-v2 slightly reduces retention of unrelated identities. Overall, the complete method (Ours) achieves the best balance between forgetting and retention, with higher ID and lower FID on the retain set, demonstrating that each component contributes to stabilizing the trade-off.

λ's Value	Forget	Retain	
	$FID_{forget}(\uparrow)$	$FID_{FFHQ}(\downarrow)$	$FID_{random}(\downarrow)$
$\lambda = 0$	182.07	103.60	31.12
$\lambda = 5$	179.08	100.17	29.98
$\lambda = 50$	181.75	99.687	29.83
$\lambda = 100$	179.07	96.925	29.59
$\lambda = 200$	179.43	95.713	29.39
$\lambda = 400$	178.47	92.563	29.03

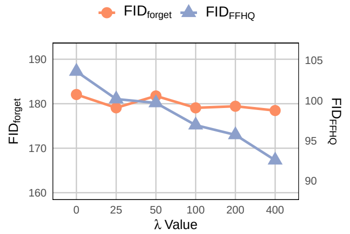

Table 3. Quantitative results in improving the image quality of retaining identities given different λ_{mse}. We use FID to measure the quality of the images generated by the unlearned model for forgetting, retaining, and random-noise images.

B.2. Ablation Result for De-Identification Loss

To better understand the design of our de-identification loss $\mathcal{L}_{de}$, Table 5 and the accompanying qualitative panel in Figure 1 examine the effect of each term in $\mathcal{L}_{de}$ by removing them in turn. Dropping $\mathcal{L}_{id}$ or $\mathcal{L}_{lpips}$ (v1, v2) still induces forgetting but increases retain-set FID, indicating degraded visual faithfulness; this aligns with the visuals, where v1 retains noticeable resemblance to the sources and v2 exhibits blur/loss of fine detail. Removing $\mathcal{L}_{mse}$ (v3) is clearly harmful: the forget-set ID score turns negative and FID spikes for both forget and retain, and the figure shows a collapse into unrealistic patterns—confirming that pixel-level anchoring is necessary for stable optimization. By contrast, the full objective $\mathcal{L}_{de}$ (all three terms) achieves the best trade-off: quantitatively, it lowers ID similarity on the forgotten set while keeping retain-set FID comparatively low; qualitatively, it produces realistic surrogate faces that conceal the original identities without visual artifacts. In short, $\mathcal{L}_{mse}$ stabilizes reconstruction, $\mathcal{L}_{id}$ drives identity separation, and $\mathcal{L}_{lpips}$ preserves perceptual realism; all three are complementary for balanced forgetting vs. retention.

B.3. Coefficients of Loss Functions

We further study the impact and stability of our method under different settings for the following hyper-parameters, including λ_{nei}, λ_{id}, and λ_{mse}, used in our loss functions in the main paper. The results are shown in Table 6, Table 7 and Table 8, respectively.

Table 2. Quantitative results of our method and the baseline (GUIDE) in the generative identity unlearning task with an increasing number of forgetting identities K.

#IDS	Methods	FFHQ (In-domain Distribution)		CelebAHQ (OOD Distribution)		Random	
		ID (↓)	FID (↑)	ID (↓)	FID (↑)	ID (↓)	FID (↑)
K=1	GUIDE	0.2773 ± 0.0099	133.9000 ± 6.9372	0.1180 ± 0.03	240.1000 ± 14.3486	0.0653 ± 0.0068	371.92 ± 0.2743
	Ours	0.3576 ± 0.0146	115.9752 ± 7.7846	0.3319 ± 0.04	200.0800 ± 10.2305	0.1980 ± 0.0103	377.76 ± 13.059
K=5	GUIDE	0.0040 ± 0.0077	177.2700 ± 2.2242	0.0472 ± 0.02	208.5533 ± 2.7544	0.0707 ± 0.0032	251.27 ± 1.3452
	Ours	0.2664 ± 0.0134	125.9833 ± 0.8796	0.3145 ± 0.01	180.3267 ± 2.1307	0.2421 ± 0.0008	227.88 ± 0.5831
K=10	GUIDE	0.0095 ± 0.0017	163.8333 ± 0.9667	0.0695 ± 0.00	253.8033 ± 4.5149	0.0454 ± 0.0009	209.84 ± 0.5237
	Ours	0.2559 ± 0.0078	125.9333 ± 0.3963	0.2382 ± 0.01	179.8500 ± 3.0294	0.1750 ± 0.0024	187.93 ± 0.8314
K=20	GUIDE	-0.0275 ± 0.0034	143.4433 ± 0.8456	0.0366 ± 0.00	196.5933 ± 3.6775	0.0538 ± 0.0016	176.84 ± 0.0924
	Ours	0.2917 ± 0.005	105.4033 ± 0.6074	0.2355 ± 0.01	154.1633 ± 3.3067	0.2224 ± 0.0016	140.69 ± 0.3707
K=50	GUIDE	-0.0431 ± 0.0018	146.4333 ± 0.6706	0.0291 ± 0.00	198.1010 ± 0.2452	0.0423 ± 0.0005	157.40 ± 0.3089
	Ours	0.2433 ± 0.0051	106.1167 ± 1.4283	0.2106 ± 0.00	150.5780 ± 0.8392	0.2535 ± 0.0007	118.44 ± 0.0636

Table 4. Ablation studies to demonstrate the differences in performance due to different loss functions by gradually removing our method components. We compare how good each version (Ours-v1, Ours-v2, Ours-v3, and Ours) balancing the trade-off between forgetting and retention ability via ID and FID.

Methods	Components			Forget		Retain	
	$\mathcal{L}_{nei}$	$\mathcal{L}_{ewc}$	$\Theta(\cdot)$	ID (↓)	FID (↑)	ID (↑)	FID (↓)
Ours-v1	✗	✓	✓	0.3639	184.89	0.5760	95.779
Ours-v2	✓	✗	✓	0.3487	182.89	0.5814	107.47
Ours-v3	✓	✓	✗	0.0495	203.31	0.3244	143.40
GUIDE	–	–	–	0.0380	211.66	0.3412	145.64
Ours	✓	✓	✓	0.3511	182.36	0.5897	101.42

Table 5. Ablation study for de-identification loss in $\mathcal{L}_{de}$. By gradually removing each loss term, we study the corresponding change in performance of balancing the trade-off between forgetting and retaining.

Version	Components			Forget		Retain	
	$\mathcal{L}_{mse}$	$\mathcal{L}_{id}$	$\mathcal{L}_{lpips}$	ID	FID	ID	FID
$\mathcal{L}_{de} - v1$	✓	✗	✓	0.4265	121.02	0.5762	103.12
$\mathcal{L}_{de} - v2$	✓	✓	✗	0.3746	164.88	0.5340	133.40
$\mathcal{L}_{de} - v3$	✗	✓	✓	-0.0154	392.19	-0.0219	410.78
$\mathcal{L}_{de}$	✓	✓	✓	0.3511	182.36	0.5897	104.23

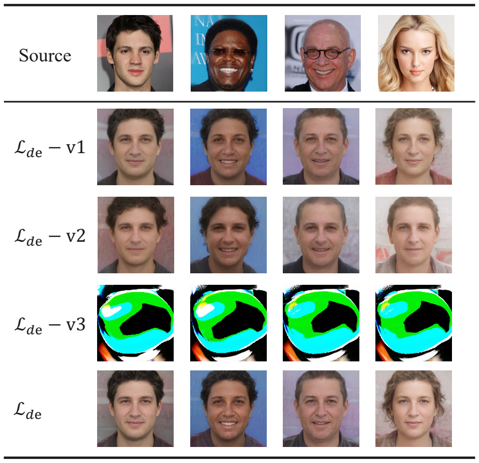
Source
$\mathcal{L}_{de} - \text{v1}$
$\mathcal{L}_{de} - \text{v2}$
$\mathcal{L}_{de} - \text{v3}$
$\mathcal{L}_{de}$

Figure 1. Qualitative results for surrogate identities produced by different variants of the de-identification loss in our ablation study for $\mathcal{L}_{de}$. When $\mathcal{L}_{id}$ is removed (v1), the generated faces still preserve noticeable resemblance to the source identities. Excluding $\mathcal{L}_{lpips}$ (v2) leads to identity shifts but introduces blurriness and loss of fine details. Without $\mathcal{L}_{mse}$ (v3), the generation collapses into unrealistic patterns, confirming the stabilizing role of pixel-level reconstruction. In contrast, the full $\mathcal{L}_{de}$ produces realistic surrogate faces that are visually natural while effectively concealing the original identities.

Coefficient of Neighbor Loss λ_{nei}. This parameter controls how much we should forget about the neighbors given a forgetting identity set. From results in Table 6, we can see that the method is quite stable with small values such as less than 0.1. Too high value of λ_{nei} can improve the retaining ability but can tamper with the unlearning effect. Our result indicates that selecting λ_{id} equal to 0.1 is the most effective.

Coefficient of Identity Loss λ_{id}. This parameter controls balancing between the human perceptual on forcing unlearning an identity versus distancing the corresponding vectors on the latent space in forgetting loss function. From results in Table 7, we select λ_{id} equal to 0.1 to ensure the stable performance of our method across various settings.

Coefficient of Feature Loss λ_{mse}This parameter also controls balancing between the human perceptual on forcing unlearning an identity versus distancing the corresponding vectors on the latent space in the forgetting loss, indicating

Table 6. Quantitative ablation study on different values of λ_{adj}.

Value	Forget		Retain	
	ID (↓)	FID (↑)	ID (↑)	FID (↓)
10^{-3}	0.3322	187.90	0.6768	86.671
10^{-2}	0.3385	186.64	0.6788	86.101
10^{-1}	**0.3457**	**185.258**	**0.6845**	**86.137**
1	0.3776	184.38	0.7040	85.002
10	0.3963	183.03	0.7131	85.351

Table 7. Quantitative ablation study on different values of λ_{id}.

Value	Forget		Retain	
	ID (↓)	FID (↑)	ID (↑)	FID (↓)
10^{-3}	0.3890	184.5962	**0.7058**	85.5134
10^{-2}	0.3890	184.6549	0.7057	85.3730
10^{-1}	**0.3585**	**185.0954**	0.7057	**85.3685**
1.0	0.3786	182.7092	0.7030	83.7749

how much we should focus on shifting feature spaces for forgetting identities. From results in Table 8, we select λ_{id} equal to 0.01 to achieve the highest performance of this dual task.

Table 8. Quantitative ablation study on different values of λ_{mse}.

Value	Forget		Retain	
	ID (↓)	FID (↑)	ID (↑)	FID (↓)
10^{-3}	0.3753	**186.80**	0.7056	87.165
10^{-2}	**0.3752**	185.39	**0.7058**	85.231
10^{-1}	0.3901	177.40	0.6971	**85.547**
1	0.4279	185.09	0.6933	87.284

B.4. Single-Identity and Large-Scale Unlearning

Forgetting Single Identity. To further demonstrate the superiority and consistency of our method over the baseline for single-identity unlearning, we conduct 10 independent experiments—each forgetting one randomly selected subject from Figure 2, which also underpins our human study summarized in the main paper. After unlearning, the model no longer reproduces the forgotten identity; instead, it generates a surrogate identity that preserves some high-level facial attributes while clearly depicting a different person. Moreover, as shown in Table 9, our method maintains strong utility on the retaining set, yielding higher ID and lower FID than the baseline. Taken together, these results indicate that our approach effectively enforces forgetting even when only a single identity is requested for removal, while better preserving non-target identities.

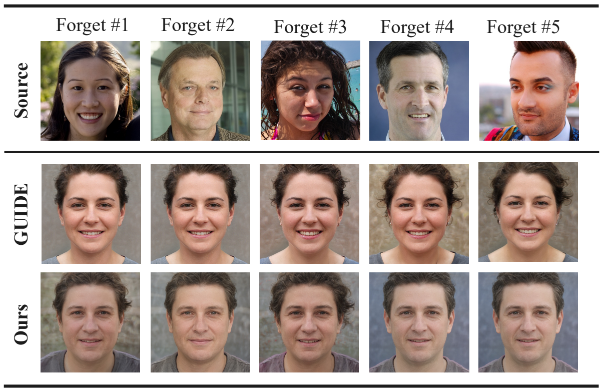

Figure 2. Quantitative results for forgetting a single identity. For each identity (ID), we independently apply the forgetting procedure and report ID and FID for both forgetting and retaining. The results show that our method more effectively removes resemblance to the target identity while better retaining non-target identities.

Table 9. Quantitative retaining performance after forgetting a single identity. In each experiment, the forgetting set $\mathcal{W}_f$ corresponds to one identity. For every identity, we report retaining scores for ID and FID, where lower values of FID indicate better performance.

ID	Retaining (ID)		Retaining (FID)	
	GUIDE	Ours	GUIDE	Ours
1	0.4838	0.6504	113.27	83.80
2	0.4969	0.6858	107.72	79.71
3	0.5244	0.6235	102.09	88.40
4	0.4758	0.7125	109.67	56.76
5	0.5080	0.7553	102.36	61.08
6	0.4548	0.6250	86.64	81.63
7	0.4946	0.7146	106.65	66.05
8	0.5310	0.6507	112.04	93.45
9	0.5812	0.6745	115.00	76.29
10	0.4142	0.6447	113.18	83.01
Avg	0.4965 *± 0.0422*	0.6737 *± 0.0419*	106.96 *± 8.68*	77.42 *± 10.3*

Large-scale Unlearning. We further evaluate our method at scale by unlearning $N \in \{100, 200\}$ identities—an extreme and challenging setting. Compared to GUIDE, we sustain higher quality on the Retain set and obtain more realistic surrogates on the Forget set without mean-face collapse. Concretely, Retain ID improves by $\sim 270\%$ on average and Retain FID drops by $\sim 44\%$ (relative to GUIDE). However, at this scale, we also observe slight trait drift in some retained identities—subtle softening or shifts in facial cues—reflecting an inherent trade-off between enforcing coherent remapping for the forgotten cohort and perfectly preserving every retained detail. Overall, both the metrics and visuals indicate our approach scales more gracefully

Table 10. Large-scale identity unlearning with many identities (N=100, 200). When compared to GUIDE, our method keeps non-forgotten identities of higher quality.

#IDs	Methods	Forget		Retain	
		ID	FID	ID	FID
N=100	GUIDE	0.0661	176.03	0.1380	150.46
	Ours	0.3360	119.55	0.4695	80.940
N=200	GUIDE	0.0485	164.34	0.1042	138.41
	Ours	0.2793	114.01	0.4135	79.530

under large-cohort unlearning.

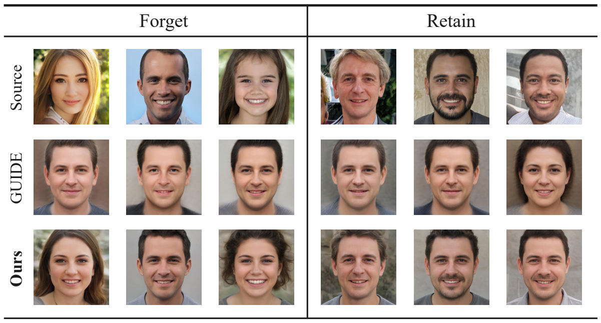

Figure 3. Qualitative results on large-scale identity unlearning with many identities (N=100, 200). Our method helps avoid the mean-face collapse observed compared with GUIDE.

B.5. Sequential Unlearning

In this experiment, we consider *sequential* unlearning, where unlearning requests arrive sequentially, reflecting a practical scenario. We define four unlearning stages, each associated with a forgetting identity set: $\mathcal{D}_f^1, \mathcal{D}_f^2, \mathcal{D}_f^3, \mathcal{D}_f^4$, with each set containing two identities. Starting with the pre-trained model G_s, we process each unlearning request d_f^i by applying our unlearning procedure and baseline methods for comparison, resulting in the unlearned models G_u^i. Quantitative results are presented in Table 11, while qualitative results are shown in the main paper. As observed, the GUIDE method experiences a gradual degradation in performance after each unlearning stage, leading to the loss of more facial features over time. In contrast, our approach demonstrates superior retention of model utility, consistently achieving high ID scores even after all unlearning requests. In the qualitative results, our method introduces only negligible changes in the reconstructed images, effectively preserving the identities while complying with unlearning requirements. Overall, our method outperforms GUIDE by effectively preserving model utility and identity retention while successfully fulfilling sequential unlearning requests.

Table 11. Quantitative results for sequential unlearning with $\mathcal{D}_f^1$, $\mathcal{D}_f^2$, $\mathcal{D}_f^3$ and $\mathcal{D}_f^4$, corresponding to four stages from 1 to 4 using GUIDE and our method for unlearning.

Stage	Methods	$\text{ID}_{\text{retain}}$ (↑)	$\text{FID}_{\text{retain}}$ (↓)
1	GUIDE	0.3262 ± 0.0116	140.44 ± 1.8840
	Ours	0.5702 ± 0.0068	68.089 ± 0.1155
2	GUIDE	0.2632 ± 0.0509	168.73 ± 0.0086
	Ours	0.5632 ± 0.0068	86.890 ± 3.1150
3	GUIDE	0.2310 ± 0.0069	128.37 ± 2.5130
	Ours	0.5774 ± 0.0027	96.885 ± 1.7595
4	GUIDE	0.2243 ± 0.0001	171.45 ± 0.0357
	Ours	0.6187 ± 0.0003	100.19 ± 2.0856

Table 12. Quantitative results of evaluating forgetting results on forgotten identities with edited images.

Settings	ID		FID	
	GUIDE	Ours	GUIDE	Ours
w/o. modification	-0.0458 ± 0.0007	0.2983 ± 0.0008	171.0607 ± 0.0543	120.3780 ± 2.3486
w.glasses	-0.0388 ± 0.0010	0.2768 ± 0.0020	207.0635 ± 0.6468	141.3038 ± 0.3380
w.tattoo	-0.0136 ± 0.0003	0.3353 ± 0.0002	189.9124 ± 0.0014	132.9188 ± 0.0009

B.6. Robustness Analysis

We further study the robustness of our forgetting scheme from two angles: (i) whether the identity is genuinely forgotten, or if the model simply forgets external features such as "glasses," and (ii) whether the unlearned model remains robust against edited images of forgotten identities.

To conduct this study, we consider three image groups for the same set of 5 identities: (i) unmodified images, (ii) modified images with glasses, and (iii) modified images with tattoos. For the first question, we investigate the forgetting of a set of 5 identities on the (ii) modified images with glasses, denoted as $\mathcal{D}_f$. Afterward, we generate images for these forgotten identities using all three image groups. The results, shown in Figure 5, reveal that modifying the forgotten identities' images (e.g., with glasses) has only a slight effect on the generated images for these identities. Importantly, both our method and the baseline are able to generate images of different identities, rather than the forgotten ones.

To address the second question, we conduct a similar study, but this time we focus on forgetting a set of 5 identities using (i) unmodified images with glasses, denoted as $\mathcal{D}_f$. As before, we generate images for the forgotten identities using all three image groups. The results, shown in Figure 4, reveal a similar trend. From this, we conclude that the

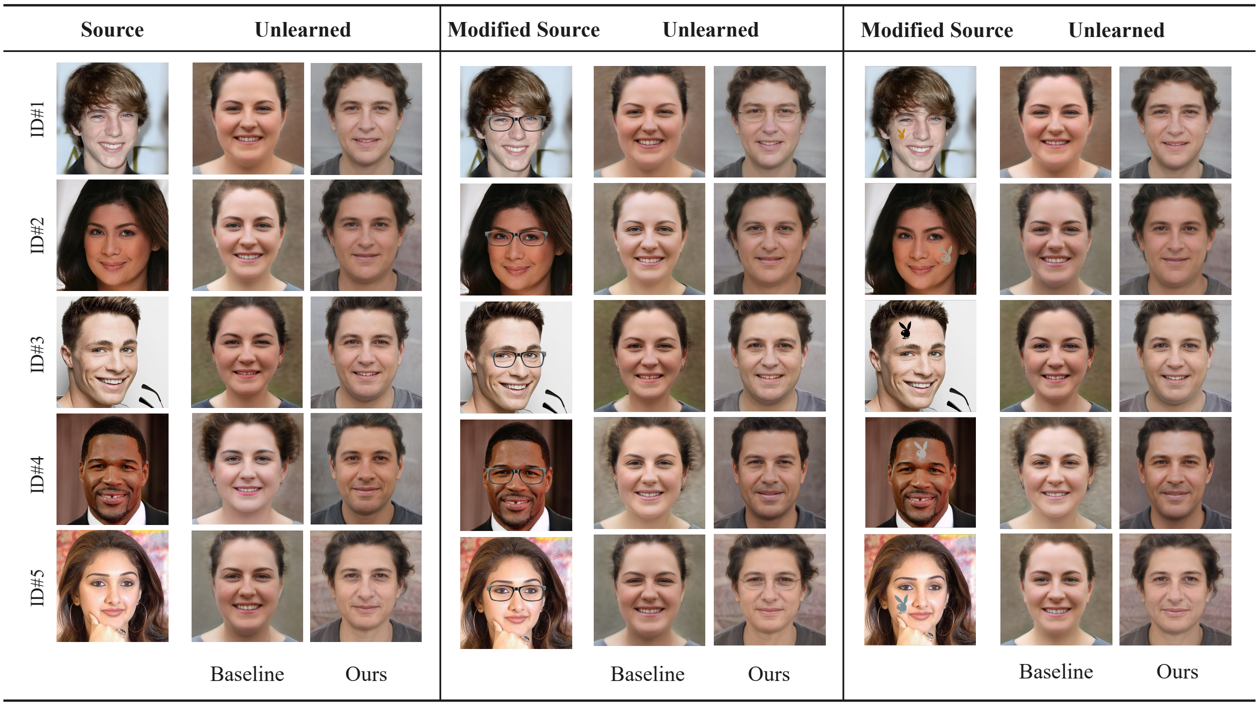

Figure 4. Qualitative results for forgetting identities and testing with the modified images of those same identities, i.e., adding wearing glasses or tattoos.

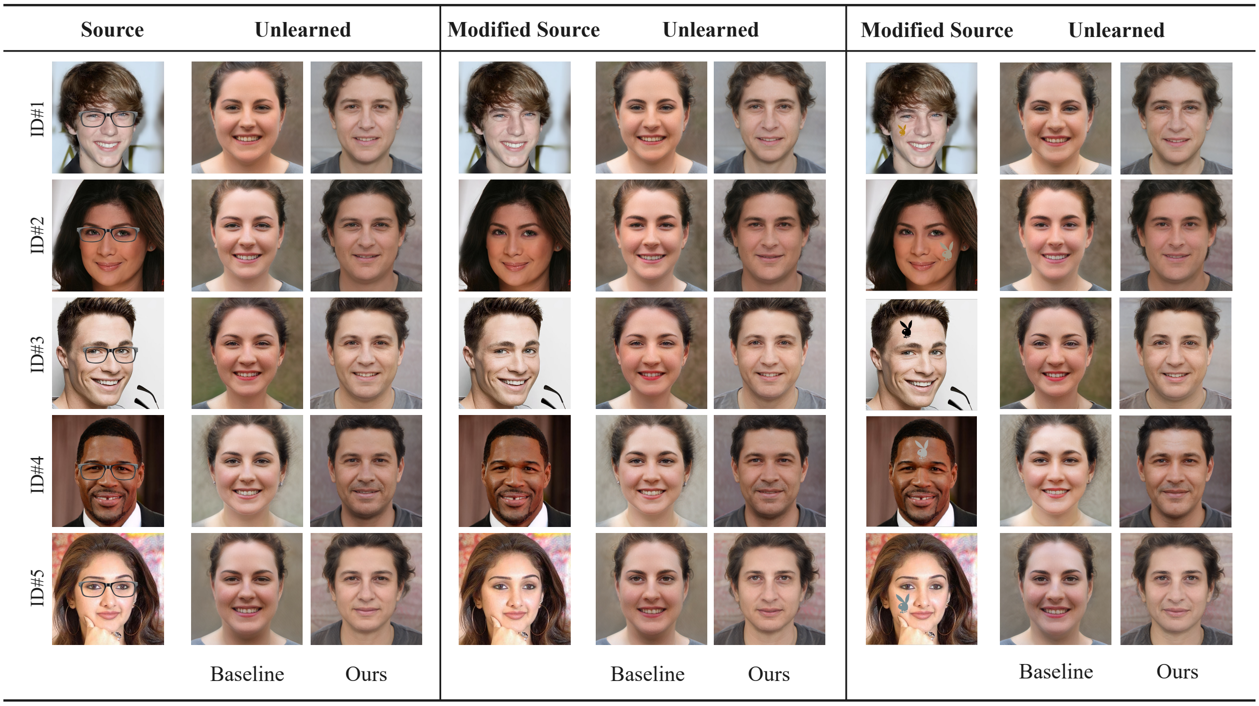

Figure 5. Qualitative results for forgetting identities with specific features (i.e., wearing glasses) and testing with the same identities but without those features, i.e., not wearing glasses.

forgetting loss used in both our method and GUIDE is indeed effective at unlearning the identities themselves, rather than merely removing external features such as glasses or tattoos.

B.7. Computational Cost Analysis

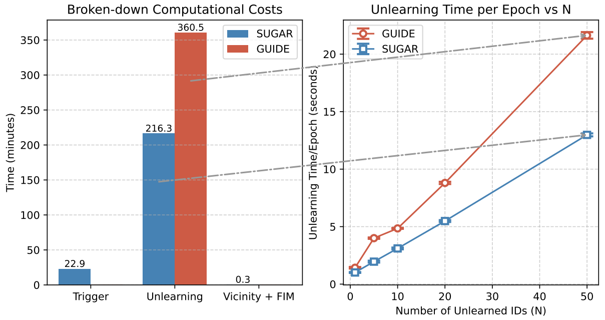

Figure 6. **Computational cost analysis.** *Left:* Breakdown of runtime under identical hardware for our method vs. GUIDE. *Right:* Per–epoch unlearning time vs. number of identities N (mean ± s.d.). Our method scales with a lower slope (∼0.24 s/ID) than GUIDE (∼0.41 s/ID); arrows link the *Unlearning* bars to the N=50 points, where ours is 12.98 ± 0.13 s vs. GUIDE's 21.63 ± 0.29 s.

As shown in Figure 6, we measured runtime on identical hardware and decomposed our pipeline into three components: trigger training, Vicinity+FIM, and the unlearning loop. The unlearning stage dominates wall–clock time, whereas trigger training is minor and Vicinity+FIM adds only negligible overhead. Overall, our end-to-end runtime is substantially lower than GUIDE. In the right panel, per-epoch unlearning time grows approximately linearly with the number of identities N, but our curve exhibits a clearly smaller slope than GUIDE, so the time savings widen as N increases. Crucially, we do *not* train a separate model per identity: a single lightweight generator produces de-identification targets for many identities, avoiding per-identity memory/compute overhead.

Our generator–based design eliminates per–identity models and uses far fewer parameters than a full StyleGAN2, leading to markedly faster training and inference. Under identical hardware, our method consistently finishes sooner than GUIDE and maintains lower per-epoch unlearning time across the tested range of N. Overall, we achieve state-of-the-art unlearning effectiveness while reducing computational burden and avoiding additional overheads.

C. Additional Baseline and Potential Applications

Additional baseline (Selective Amnesia). Beyond GUIDE, we compare against Selective Amnesia (SA) in Figure 7 and Table 13. Qualitatively, SA collapses the *forget* outputs to noisy/mean-like templates and even corrupts the *retain* images, while our method produces realistic, diverse surrogates for the forget set and leaves the retain set visually faithful. Quantitatively, SA's collapse manifests as an undefined ID score on the forget set (NaN, due to non-face outputs) and extremely poor fidelity (FID 520.67 on forget and 418.36 on retain). In contrast, ours achieves effective forgetting with valid identity measurements (ID 0.3511 on forget) and dramatically better realism—about 3× lower FID on the forget set (182.36 vs. 520.67) and about 4× lower on the retain set (104.23 vs. 418.36). Crucially, we also preserve the retained identities (retain ID 0.5897 vs. SA's 0.0519), avoiding SA's utility collapse. These results reinforce our security analysis: SA's low-entropy collapse creates a tell-tale fingerprint that is easy to detect, whereas our method avoids mode collapse, yielding natural surrogates that intermix with the retained distribution and thus better resist erasure-detection and membership-inference attacks.

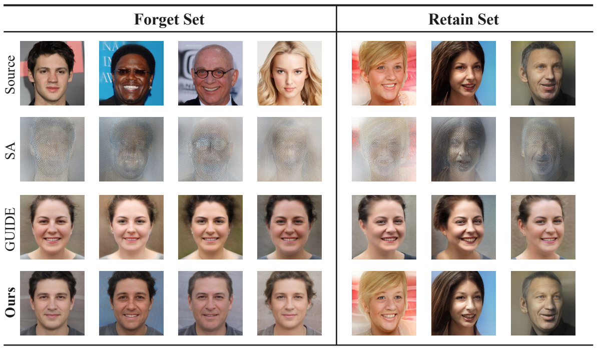

Figure 7. Qualitative comparison on identities to *forget* (left) and to *retain* (right). Rows show the source images, Selective Amnesia (SA), GUIDE, and our method. SA collapses forgotten identities to noise/mean-like patterns and also corrupts retained faces. GUIDE maps many forgotten subjects to a near-identical surrogate and slightly distorts retained samples, yielding a separable cluster. Our method produces realistic, diverse surrogates for the forget set while preserving the appearance of the retain set, avoiding mode collapse.

Table 13. Comparison with Selective Amnesia (SA) on *forget*/*retain* sets. SA collapses to non-face outputs (NaN ID) with very high FID, while ours achieves effective forgetting and preserves retain-set fidelity.

Method	Forget		Retain	
	ID	FID	ID	FID
SA	NaN	520.67	0.0519	418.36
Ours	0.3511	182.36	0.5897	104.23

Potential Application on Diffusion Model. As stated in our problem formulation, we work in an image-to-image (I2I) setting where a frozen encoder serves as a surrogate to train our de-identification model: each identity image is mapped to a latent vector, and we directly reuse this de-identification model to learn a *surrogate set* for a forgotten identity (e.g., “Angelina Jolie”) in Stable Diffusion 1.4; experiments are run by replacing the forgotten concept with this targeted surrogate set, with images generated using our trained generator and, for comparison, Selective Amnesia, which is presented in Figure 8. Conceptually, this setup also clarifies how our approach interfaces with diffusion-based unlearning: our surrogate generator can supply the *replacement target* for a diffusion model, but deploying it natively in text-conditioned diffusion requires additional tailoring because unlearning in DMs is fundamentally about *text–image association*. Identity information in diffusion models is distributed and context-dependent (token embeddings, CLIP/image encoders, and cross-attention features), so a suitable encoder must be chosen or adapted to the specific DM to capture prompt-conditioned identity semantics; in practice, this means tailoring the encoder (and possibly the training objective) to bind the learned surrogate set to the relevant text tokens without leakage. In short, our method provides a principled “what to replace with,” while diffusion-model unlearning must handle “how to bind it to text,” which is feasible but demands model-specific encoder tailoring.

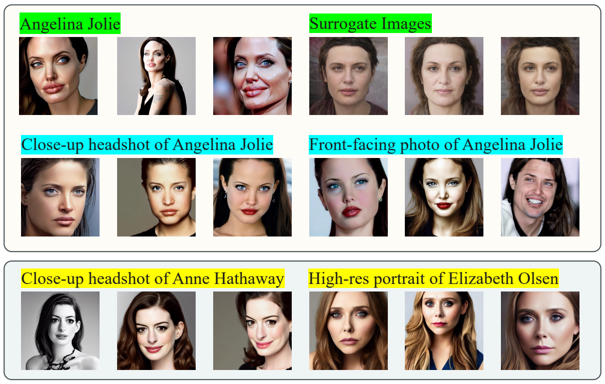
Angelina Jolie
Surrogate Images
Close-up headshot of Angelina Jolie
Front-facing photo of Angelina Jolie
Close-up headshot of Anne Hathaway
High-res portrait of Elizabeth Olsen

Figure 8. Surrogate–concept replacement in an I2I setting (Stable Diffusion 1.4). We use a frozen encoder to represent identity images as latent vectors and train a de-identification generator that learns a *surrogate set* for the forgotten concept (“Angelina Jolie”). At inference, the forgotten token is replaced by the learned surrogate. Top: exemplar images of the forgotten concept and the corresponding surrogate images learned by our generator. Middle: prompts that explicitly invoke the forgotten identity (e.g., “Close-up headshot of Angelina Jolie”, “Front-facing photo of Angelina Jolie”) produce realistic, diverse surrogates rather than reconstructing the original identity. Bottom: control prompts for other identities (Anne Hathaway, Elizabeth Olsen) remain faithful, indicating that the replacement is targeted and does not degrade unrelated concepts.

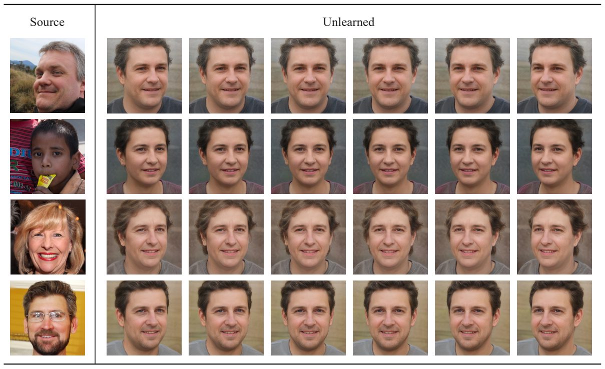

Figure 9. Generated images after unlearning multiple identities under different camera poses using the FFHQ dataset.

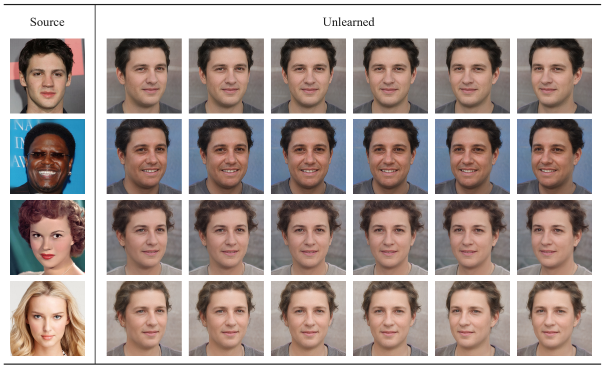

Figure 10. Generated images after unlearning multiple identities under different camera poses using the CelebA-HQ dataset.

D. Additional Qualitative Results

D.1. Multi-View Synthesized Images

We generate images after unlearning multiple identities while maintaining variations in camera poses using the CelebA-HQ and FFHQ datasets to evaluate the model’s ability to forget specific identities while preserving pose diversity (Figure 10 and Figure 9). As shown, SUGAR can forget multiple identities from multiple angles, successfully generating a new identity for each forgotten identity.

D.2. Targets Generated By De-Identification Process

We plot the evolution of generated target faces over training iterations in Figure 11. Initially, the generated faces exhibit substantial variation, but over time, they begin to resemble a consistent synthesized identity, appearing as an averaged or transformed representation of the source. As training progresses, the generated faces gradually stabilize, aligning with certain facial features. Eventually, the targeted images for different source identities converge toward faces that share some features with the forgotten IDs but ultimately belong to different identities. This suggests a model adaptation process where the generated outputs retain partial resemblance, i.e., glasses and skin colors, hairs, etc., to the original sources while shifting toward distinct identities. One advantage of this process is that the model vendor can completely control this process and the forgetting ability that they think is reasonable for their settings.

D.3. Retain ability with different closeness

In Figure 12, we take the latent vector w_f of a forgotten identity and sample new vectors with distance Δ from the original identity. Identities sampled closer to w_f experience a larger forgetting effect, as seen most prominently in the $\Delta = 35$ or $\Delta = 40$ cases, where the generated images show the typical characteristics of convergence toward the identities synthesized by GUIDE and SUGAR. As the distance from w_f increases, the generated results exhibit less prominent changes, and begin to match the source image almost exactly.

References

[1] Eric R. Chan, Connor Z. Lin, Matthew A. Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas Guibas, Jonathan Tremblay, Sameh Khamis, Tero Karras, and Gordon Wetzstein. Efficient geometry-aware 3D generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1

[2] Sagar Imambi, Kolla Bhanu Prakash, and GR Kanagachidambaresan. Pytorch. *Programming with TensorFlow: Solution for Edge Computing Applications*, pages 87–104, 2021. 1

[3] 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. 1

[4] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 1

[5] 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

[6] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 1

[7] Ziyang Yuan, Yiming Zhu, Yu Li, Hongyu Liu, and Chun Yuan. Make encoder great again in 3d gan inversion through geometry and occlusion-aware encoding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2437–2447, 2023. 1

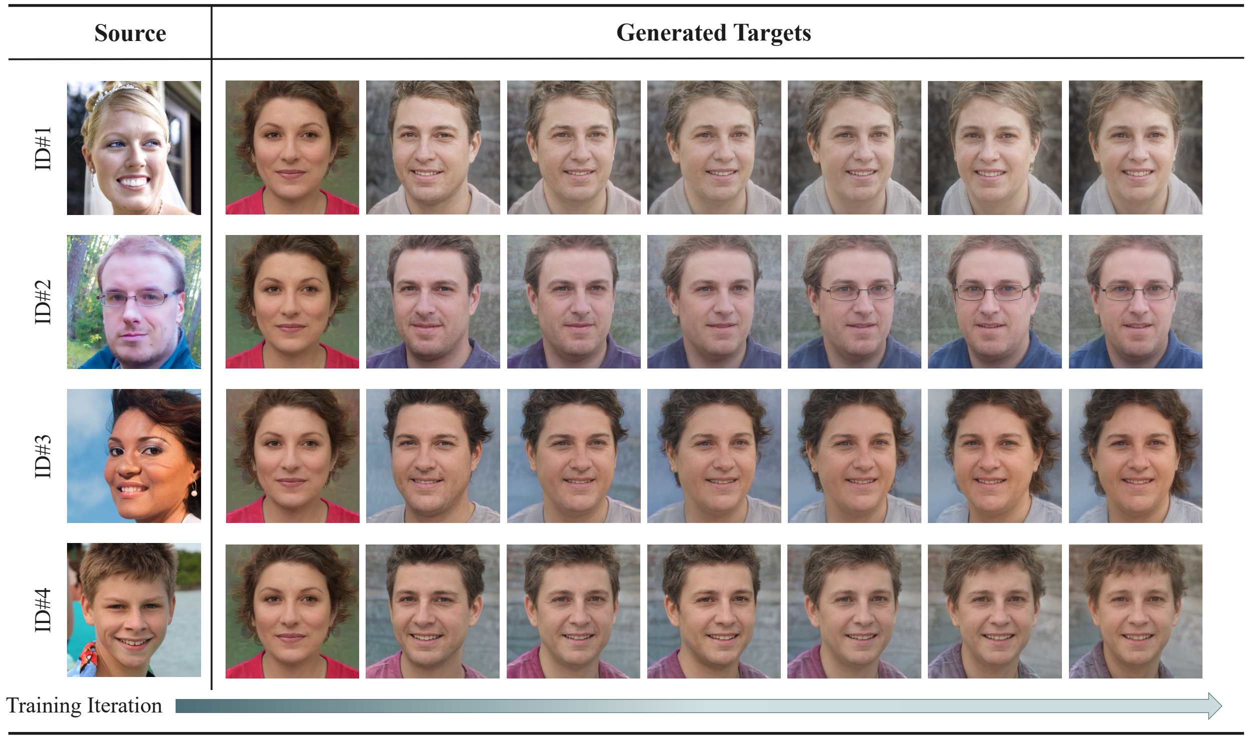

Figure 11. Generated triggers by time using our de-identification process Θ. As training progresses, the generated faces gradually stabilize, aligning with certain facial features.

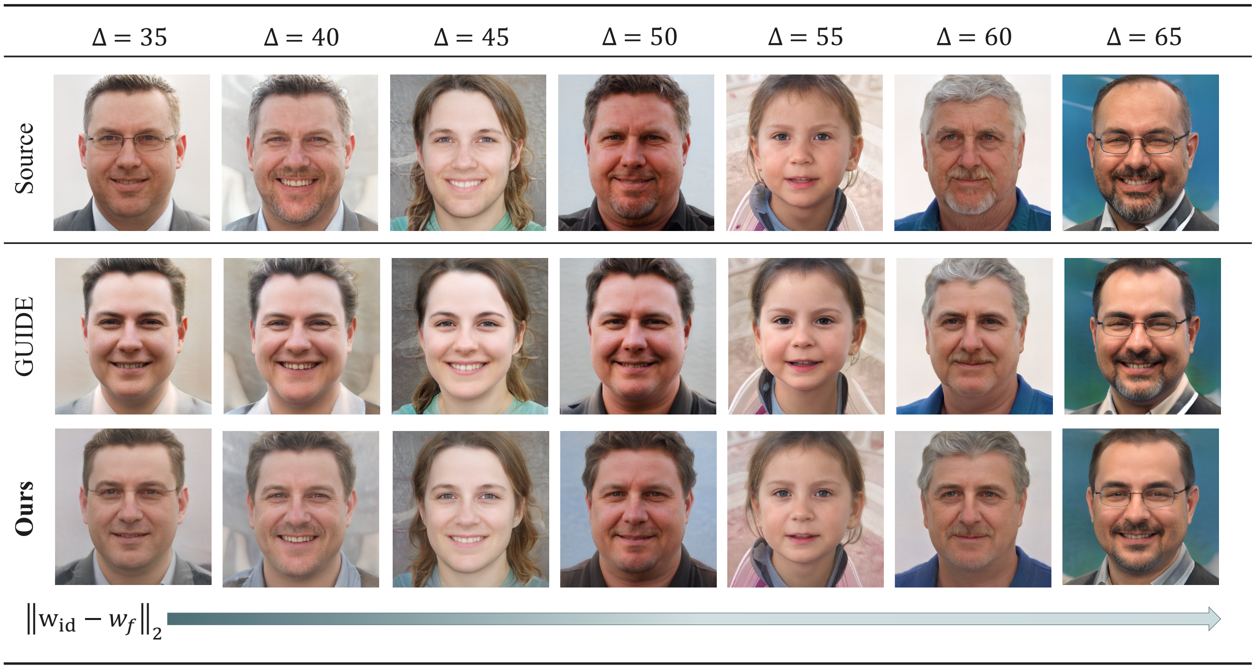

Figure 12. Qualitative results for model performance on retaining identities. The higher the distance Δ is, the further the identity is from the forgetting identity in the latent space. As shown, the unlearning process tends to affect the performance on the closer identities, i.e., when Δ equal to 35.