PIU: Proximity-guided Identity Unlearning in ID-Conditioned Diffusion Models

Jose Edgar Hernandez Cancino Estrada[1,*], Mauro Díaz Lupone[1,*], Žiga Emeršič[1],
Vitomir Štruc[2], Peter Peer[1], and Darian Tomašević[1]
[1] University of Ljubljana, Faculty of Computer and Information Science, Ljubljana, Slovenia
[2] University of Ljubljana, Faculty of Electrical Engineering, Ljubljana, Slovenia
edgarcancinoe@exatec.tec.mx, maurodiazlupone@gmail.com, ziga.emersic@fri.uni-lj.si,
vitomir.struc@fe.uni-lj.si, {peter.peer, darian.tomasevic}@fri.uni-lj.si

Abstract

*Identity-conditioned diffusion models enable high-quality and identity-consistent face generation, but they also raise severe privacy concerns, as models may continue to synthesize individuals despite their right to be forgotten. While machine unlearning has been extensively studied for concept and data removal, identity unlearning remains largely unexplored, particularly in models conditioned directly on identity embeddings rather than text prompts. In this work, we study identity unlearning in Arc2Face, a state-of-the-art identity-conditioned latent diffusion model for face generation, and introduce **P**roximity-guided **I**dentity **U**nlearning (**PIU**), an anchor-guided framework for identity unlearning. Specifically, we formulate identity removal as an identity replacement objective that reassigns the source identity to a selected anchor identity in the learned identity space, and we complement it with a proximity-based anchor selection strategy motivated by the geometry of ArcFace representations. We further show that effective unlearning can be achieved through localized fine-tuning of a small subset of identity-sensitive cross-attention layers. Experiments across multiple target identities show that our framework effectively suppresses generation of the target identity while preserving realism and identity consistency for retained identities, as validated by improved performance on unlearning and image-quality metrics, together with qualitative evaluation. The source code for PIU is publicly available at* `https://github.com/edgarcancinoe/piu_unlearning`.

1. Introduction

Diffusion models are nowadays the dominant paradigm for high-quality image generation, serving as the foundation for many state-of-the-art conditional generative systems [33, 35, 29]. In biometrics, identity-driven synthesis has garnered particular attention for creating synthetic face images suitable for privacy-aware training and evaluation of deep models [4]. While text-based personalization methods can adapt pretrained generative backbones to specific subjects [35], their textual conditioning often compromises identity consistency across variations in pose, illumination, and expression [43]. In contrast, identity-conditioned models like Arc2Face [29] directly condition on rich identity features of pretrained recognition models, enabling identity-consistent yet diverse synthesis. Despite their value in producing large-scale synthetic datasets, such models also amplify privacy and misuse risks, including deepfake abuse [47], which is reinforced by evidence that diffusion models can memorize and reproduce training data [5, 40, 41]. These vulnerabilities, coupled with regulations like the CCPA [13] and GDPR [30] that formalize rights to erasure, necessitate the development of techniques to remove specific information from trained models without full retraining. This establishes a challenging objective: making ID-conditioned models forget a target individual.

Figure 1. **PIU unlearns a target identity from an identity-conditioned diffusion model [29] by mapping it to a selected anchor.** Afterwards, conditioning on the target identity produces images aligned with the anchor rather than the original subject.

*Jose Edgar Hernandez Cancino Estrada and Mauro Díaz Lupone are first authors with equal contributions.

To this end, machine unlearning [3] provides a formal framework for selective removal of data, concepts, or behaviours from trained models. In diffusion models, most prior work has focused on either *concept unlearning* [11, 12, 24], which suppresses high-level semantics like artistic styles or unsafe content, or on *data unlearning* [1, 16, 45], which aims to remove specific training images. Identity unlearning is related but distinct: in ID-conditioned models, identity is not controlled by discrete text prompts, but rather by continuous, structured feature representations, making selective forgetting substantially more delicate. Consequently, identity unlearning has only recently emerged as a distinct research direction. While GUIDE [37], LEGATO [6], and Forget-Me [31] study identity forgetting in different settings, the closest prior works exploring identity erasure in diffusion models include BIA [38] and WID [39]. However, BIA targets standard diffusion probabilistic models rather than ID-conditioned models like Arc2Face [29], while WID relies on a custom, unreleased base model, limiting its reproducibility. Thus, a critical gap remains between the conceptual formulation of identity unlearning in ID-conditioned diffusion models and its practical realization.

To address this gap, we present **P**roximity-guided **I**dentity **U**nlearning (**PIU**), a framework designed for surgical identity erasure in ID-conditioned diffusion models like Arc2Face [29]. Unlike methods that attempt to naively suppress a conditioning signal, PIU formulates forgetting as an explicit replacement and reassigns the target identity to a safe anchor identity within the learned identity space, as shown in Figure 1. To achieve this, PIU combines a proximity-based anchor selection strategy, grounded in the geometry of ArcFace [7] identity embeddings, with an anchor-guided unlearning objective tailored for diffusion architectures. Furthermore, we provide an empirical analysis of identity-sensitive layers in Arc2Face, enabling a more surgical and efficient fine-tuning procedure that maximizes identity erasure while minimizing negative impact on the remaining identity manifold. Extensive experiments on the state-of-the-art Arc2Face model [29] demonstrate that PIU effectively removes target identities from CelebA-HQ [20] while preserving the quality and identity consistency of images for retained, non-target subjects. Through a reproducible benchmark, both quantitative and qualitative results show that identity unlearning in ID-conditioned models is feasible and distinct from prior concept or data unlearning paradigms. Our contributions are summarized as follows:

- We propose PIU, to our knowledge the first reproducible framework for identity unlearning in ID-conditioned diffusion models, namely Arc2Face [29].
- We formulate unlearning as explicit geometric replacement, erasing the target identity by reassigning it to an anchor selected via a novel proximity-based strategy within the continuous identity space.
- We provide a layer-level analysis of identity-sensitive components in Arc2Face, enabling surgical and efficient fine-tuning that targets the most relevant weights.
- We compare PIU to the state-of-the-art through extensive experiments and demonstrate that it delivers effective identity unlearning while avoiding degradation for non-target identities observed in prior works.

2. Related work

Image generation. Image synthesis has advanced rapidly with the development of deep generative models, progressing from Generative Adversarial Networks (GANs) [14, 21, 22] to diffusion models [18]. In particular, Latent Diffusion Models (LDMs) [33] have nowadays established a robust foundation for efficient and high-quality conditional synthesis by operating in a lower-dimensional latent space. Building on this framework, many works addressed identity-driven face generation. Early methods personalized generation from a few subject images through embedding optimization [10] and model fine-tuning [35], with recent advancements incorporating identity-based training objectives to enhance identity consistency [43]. To bypass the need for per-subject optimization, other approaches introduced adapter modules [48, 25, 46], which extract subject-aware visual features using pretrained image encoders and inject them via cross-attention layers. In contrast, Arc2Face [29] enables identity-consistent generation by injecting identity embeddings from a pretrained recognition model [7] into the text encoder by replacing token embeddings of a template prompt. This approach achieves exceptional identity preservation, establishing Arc2Face as a core framework for generating high-quality, diverse, and identity-consistent face images suitable for training recognition models.

Identity unlearning. Modern diffusion models have raised privacy and copyright concerns due to their capacity to reproduce training samples [40, 5, 41]. This memorization risk, coupled with the requirements of legal frameworks such as the General Data Protection Regulation (GDPR) [30], has motivated the development of machine unlearning techniques. The most widely explored direction is *concept unlearning* for text-to-image models. Early iterative fine-tuning methods like Erased Stable Diffusion (ESD) [11] suppress concepts by steering the conditional noise prediction toward its unconditional counterpart. Subsequent works explore minimalist unlearning through sparse weight masking [49], while closed-form approaches like Unified Concept Editing (UCE) [12] compute analytic updates to targeted model weights, typically in cross-attention layers. Another prominent direction is *data un-*

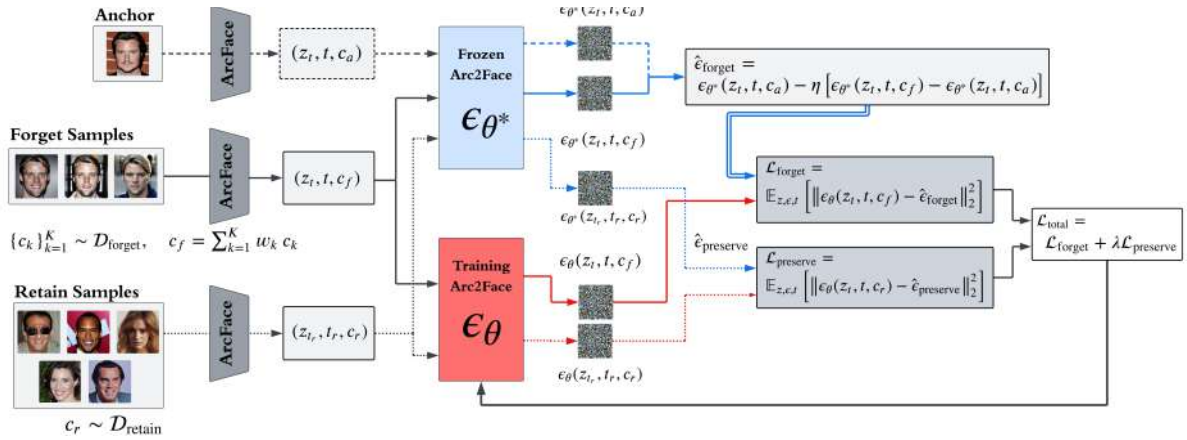

Figure 2. **Overview of the proposed Proximity-guided Identity Unlearning (PIU) framework.** During training, target (forget) and retain identity embeddings are processed by frozen and trainable versions of Arc2Face [29]. The forget loss $\mathcal{L}_{\text{forget}}$ steers the trainable prediction toward a proximity-based anchor trajectory, while the preservation loss $\mathcal{L}_{\text{preserve}}$ enforces alignment with the frozen model for non-target identities. The joint objective is optimized to ensure efficient identity removal while preserving the generative prior of the model.

learning, which aims to remove specific training samples so they cannot be reproduced [9, 16]. Within this paradigm, Subtracted Importance Sampled Scores (SISS) [1] introduces an importance-sampled loss function providing theoretical guarantees and empirical evidence for erasing specific identities. Despite this progress, *identity unlearning* remains largely underexplored, especially for identity-conditioned diffusion models. Unlike standard text-to-image models where concepts are tied to discrete text tokens, these architectures condition image generation on dense identity embeddings [29], making unlearning highly sensitive to how the target is redirected within the identity space. While GUIDE [37] introduced anchor-based identity unlearning, the mechanisms remained limited to GAN-based generators. For diffusion models, Forget-Me [31] studies identity forgetting in a federated setting, but its reliance on decentralized client updates renders it unsuitable for centralized models like Arc2Face [29]. The closest prior works are Black Hole-Driven Identity Absorbing (BIA) [38] and the Unlearn and Protect variant With ID-guidance (WID) [39]. While BIA [38] transfers anchor-guided unlearning to the h-space of standard diffusion models, it is not designed for ID-conditioned architectures and lacks publicly available code. To our knowledge, WID [39] is the only prior work explicitly addressing identity unlearning in ID-conditioned diffusion models. However, it relies on a simple identity-guided objective lacking both a dedicated preservation strategy to protect non-target identities and a principled mechanism for anchor selection. Furthermore, because it utilizes a custom, unreleased base model, direct reproduction is severely hindered. Our method addresses these gaps by introducing an explicit preservation term and a proximity-based anchor selection strategy on top of a publicly available framework, establishing a robust and reproducible identity unlearning pipeline.

3. Methodology

This section presents our Proximity-guided Identity Unlearning (PIU) approach for the identity-conditioned Arc2Face [29] diffusion model, which steers the target identity toward a proximity-based anchor rather than an unconditional prior, as depicted in Figure 2.

3.1. Identity-conditioned Latent Diffusion

Latent Diffusion Models (LDMs) [33] improve the efficiency of standard denoising diffusion probabilistic models [18] by operating in a compressed latent space of a pretrained variational autoencoder. Given an image x_0, the encoder first maps it to a lower dimensional latent representation $z_0 = \mathcal{E}(x_0)$. The forward diffusion process then gradually corrupts z_0 over timesteps $t \in \{0, \dots, T\}$ by adding Gaussian noise according to a predefined schedule, so that $z_t = \sqrt{\bar{\alpha}_t}\, z_0 + \sqrt{1-\bar{\alpha}_t}\, \epsilon$, where $\epsilon \sim \mathcal{N}(0, I)$. The reverse process can be learned by a denoising network, typically a U-Net [34], that is trained to predict the added noise ϵ at each timestep t as $\epsilon_\theta(z_t, t, c)$ with c as the conditioning information. The denoised sample $\hat{z}_0$ can then be mapped back to the image space through the decoder as $\hat{x}_0 = \mathcal{D}(\hat{z}_0)$. At inference time, conditional generation can even be strengthened with classifier-free guidance (CFG) [19], which combines unconditional and conditional predictions. In text-to-image implementations of LDMs like Stable Diffusion [33], the condition c is derived from textual prompts processed by a CLIP text encoder [32].

Arc2Face [29] adapts this text-conditioned framework for identity-driven generation of face images. Rather than relying on open text prompts, the conditioning space is restricted to facial recognition features. To this end, Arc2Face extracts an identity embedding $v \in \mathbb{R}^{512}$ with a pretrained ArcFace [7] recognition network. This vector is zero-padded to $\hat{v} \in \mathbb{R}^{768}$ and substituted for the `<id>` token

within a template prompt *"photo of a <id> person"*. The CLIP [32] text encoder then processes this sequence to produce the final identity-encoded condition c_{id}. Arc2Face trains the denoising network to predict $\epsilon_\theta(z_t, t, c_{\text{id}})$, with c_{id} encoding the target identity, thus replacing the flexibility of textual conditioning with identity conditioning.

3.2. Anchor-Guided Identity Unlearning

The goal of PIU is to erase a target identity from an identity-conditioned diffusion model, in our case Arc2Face [29]. Methods for prompt-based concept unlearning are not directly applicable in this setting, due to the fixed prompt constraints of the architecture [29]. For this reason, we build on the core ideas of concept unlearning, namely Erased Stable Diffusion (ESD) [11], and fundamentally extend this paradigm for geometrically-informed unlearning tailored to the specific structure of the identity space.

Let θ^* denote the parameters of the frozen original model, and θ be the parameters of the trainable model to be unlearned. Given a noisy latent $z_t = \mathcal{E}(x_t)$ at timestep t, let c_f be the condition embedding of the forget identity. In standard ESD [11], concept unlearning is achieved by pushing the conditional prediction away from the target concept and toward the unconditional prior $\varnothing$. In our setting, however, steering the forget identity toward the unconditional prediction is suboptimal, since the unconditional prior still maps to a generic identity. Consequently, using it as a target can unpredictably entangle identity unlearning with non-identity features, leading to attribute changes or even severe quality degradation. Instead, we propose to steer the forget identity c_f toward an anchor identity condition c_a. We define the prediction for the forget target as:

$$\hat{\epsilon}_{\text{forget}} = \epsilon_{\theta^*}(z_t, t, c_a) - \eta \left[\epsilon_{\theta^*}(z_t, t, c_f) - \epsilon_{\theta^*}(z_t, t, c_a)\right], \quad (1)$$

where η is the negative guidance scale. The forget loss then forces the trainable model, conditioned on the forget identity, to match this modified trajectory:

$$\mathcal{L}_{\text{forget}} = \mathbb{E}_{z,\epsilon,t}\left[\|\epsilon_\theta(z_t, t, c_f) - \hat{\epsilon}_{\text{forget}}\|_2^2\right]. \quad (2)$$

Since identities are highly interconnected within the U-Net [34] weights, solely optimizing for the forget objective may degrade the generation of non-target identities. To alleviate this, we introduce an identity preservation loss, inspired by recent fine-tuning methods [35]. For retain identities c_r, we enforce the trainable model to remain aligned with the predictions of the frozen model $\hat{\epsilon}_{\text{preserve}} = \epsilon_{\theta^*}(z_t, t, c_r)$ through the identity preservation objective as:

$$\mathcal{L}_{\text{preserve}} = \mathbb{E}_{z,\epsilon,t}\left[\|\epsilon_\theta(z_t, t, c_r) - \hat{\epsilon}_{\text{preserve}}\|_2^2\right]. \quad (3)$$

The final training objective is then the weighted sum:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{forget}} + \lambda \mathcal{L}_{\text{preserve}}, \quad (4)$$

where the preservation weight λ balances forgetting and model preservation: if too large, unlearning weakens; if too small, non-target identities may drift. We use $\lambda = 10$ by default and defer its analysis to Section 4.2. Figure 2 depicts this process. PIU suppresses identity-specific information associated with c_f by redirecting it toward the behavior induced by the anchor condition c_a, while retaining the rest of the model knowledge through the preservation term.

3.3. Proximity-based Anchor Selection

The most critical mechanism in our PIU method is the selection of the anchor identity c_a. Arc2Face relies on ArcFace embeddings [7] as the identity representation, which are learned using a margin based objective designed to reduce intra class variability while maximizing inter class separation [36]. This induces a hyperspherical geometry and makes the embedding space particularly suitable for reasoning about identity similarity, where the intrinsic distance between identities is defined by the geodesic distance $d_{\mathbb{S}}(\mu_f, \mu_a) = \arccos(\langle \mu_f, \mu_a \rangle)$. However, for unlearning in this space, anchor selection must strike a careful balance. If the anchor identity c_a is too similar to the forget identity c_f, it yields near-zero gradient updates. Assuming local smoothness, the conditional displacement satisfies $\epsilon_{\theta^*}(z_t, t, c_f) - \epsilon_{\theta^*}(z_t, t, c_a) \approx J_\epsilon(z_t, t, c_f)(c_f - c_a) \approx 0$, resulting in weak unlearning. Conversely, if the anchor is placed too far across this manifold (the antipodal regime where $d_{\mathbb{S}} \approx \pi$), steering the generation across empty regions may distort the generative prior that the preservation loss cannot mitigate. Therefore, optimal anchors should lie at an intermediate geodesic distance, sufficiently separated to induce identity removal but close enough to preserve the facial prior.

To resolve this, we formalize a threshold guided selection strategy. Let $\mathcal{D}$ be the training dataset partitioned by identity. For each image $x \in \mathcal{D}$, we extract its ArcFace embedding $w(x) \in \mathbb{R}^{512}$. Then, for each identity i, we compute a robust and representative centroid of the identity

$$\mu_i = \frac{1}{|\mathcal{D}_i|} \sum_{x \in \mathcal{D}_i} w(x), \quad (5)$$

where $\mathcal{D}_i$ denotes the images of identity i. Let μ_f be the centroid of the forget identity. For every candidate identity $j \neq f$, we compute the cosine similarity $s_j = \frac{\langle \mu_f, \mu_j \rangle}{\|\mu_f\|_2 \|\mu_j\|_2}$. Then, instead of selecting the most similar or dissimilar identity, we define the candidate anchor set $\mathcal{A}_\tau = \{j \neq f : |s_j - \tau| < \varepsilon\}$, with tolerance $\varepsilon = 10^{-2}$, and uniformly sample the anchor identity $a \sim \mathcal{A}_\tau$. Empirically, we find that $\tau = 0.2$ provides the best trade-off in ArcFace [7] space, favoring an anchor that is sufficiently close to the forget identity to preserve a realistic facial prior, yet sufficiently separated to effectively redirect the generations away from the target identity.

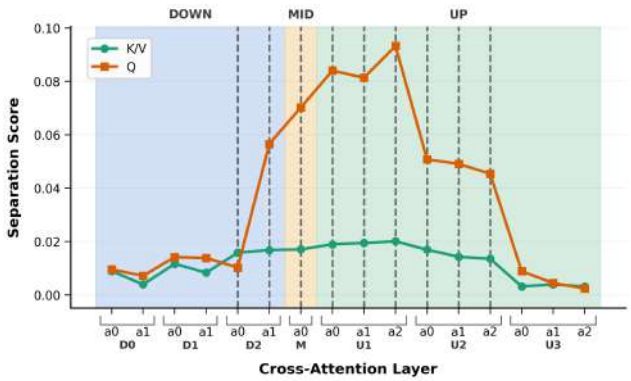

Figure 3. **Layerwise identity separation in Arc2Face cross-attention activations.** Identity separation captures the ability of the model to distinguish between identities across Keys (K), Values (V), and Queries (Q). Higher scores indicate stronger identity conditioning. Dashed lines denote the selected surgical blocks: Downsampling (D2), Middle (M), and Upsampling (U1 & U2).

3.4. Condition-driven and Localized fine-tuning

The main object of unlearning in our generative setting is the identity condition itself rather than a specific training image. Accordingly, we do not rely on latents encoded from the original training images. Instead, the latent input is sampled directly from Gaussian noise, $z_0 \sim \mathcal{N}(0, I)$, so that the optimization focuses primarily on the conditioning signal rather than on the content of specific real images.

To avoid restricting unlearning to a finite set of observed forget embeddings, we further extend the forget set by sampling synthetic identity conditions as convex combinations of forget-identity embeddings. Concretely, let $\{c_k\}_{k=1}^{K} \sim \mathcal{D}_{\text{forget}}$ be identity embeddings associated with the forget identity, and sample mixing weights $w \sim \text{Dir}(\alpha)$. We then define the forget condition as $c_f = \sum_{k=1}^{K} w_k c_k$, where $\sum_{k=1}^{K} w_k = 1$ and $w_k \geq 0$. This way, the model is encouraged to forget the local identity region of the target subject, rather than only a finite set of observed embeddings.

Which parameters to fine-tune represents another critical design decision that impacts both computational efficiency and unlearning efficacy. To this end, prior methods focus on cross-attention weights [11], which are commonly utilized for conditioning, while other approaches use gradient-based saliency or empirical activations to determine concept-specific weights [9, 26, 45]. Building on these insights, we measure layerwise identity separation within cross-attention blocks of the Arc2Face U-Net [34], across key, value, and query projections. Specifically, we define Key and Value (K/V) separation for layer ℓ as the gap between mean intra- and inter-identity cosine similarity, i.e., $S_\ell^{KV} = \text{Intra}_\ell^{KV} - \text{Inter}_\ell^{KV}$. Because K/V activations are projected directly from the global identity condition, they are inherently timestep-independent, allowing evaluation at a single representative timestep. Conversely, Query (Q) features are highly timestep-dependent and sensitive to initial noise. Thus, we compute Q separation based solely on inter-identity divergence as $S_\ell^{Q} = 1 - \frac{1}{|T|}\sum_{t\in T} \text{Inter}_{\ell,t}^{Q}$, utilizing sensitive denoising timesteps $T = \{750, 1000\}$ where query features vary most across identities.

As shown in Figure 3, the strongest separation concentrates in the third downsampling block, the middle block, and the first two upsampling blocks within the cross-attention modules. We refer to these as the *surgical layers*. Restricting fine-tuning to this subset modifies only a small fraction of the total model parameters (4.29%), making the method more efficient while achieving superior selective unlearning by targeting identity-sensitive components.

4. Experiments and Results

Implementation details. We perform unlearning on Arc2Face [29], an identity-conditioned latent diffusion model built on the Stable Diffusion 1.5 [33]. Following their configuration, images are generated at a resolution of 512×512 using the multistep DPM-Solver++[27] with 25 denoising steps and a classifier-free guidance scale of 3. Ablation experiments are run first for 300 steps with a batch size of 32, using AdamW with learning rate 10^{-4}, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$, and weight decay 0.01, along with default unlearning hyperparameters $\lambda = 10$, $\tau = 0.2$, and $\eta = 1.5$. At each training step, the loss is computed from 32 forget and 32 retain identity embeddings extracted with the ArcFace [7] recognition model. For the comparison with the state-of-the-art, unlearning is run for 400 steps with the best performing parameters from the ablation studies.

Datasets. We conduct experiments on CelebA-HQ [20], a high-quality face dataset used in the training of Arc2Face [29], to ensure that measured performance drops reflect true identity erasure rather than failures caused by out-of-distribution generations. The dataset contains 30k images at resolution 512×512, split into 28k training and 2k validation images; we use only the training split. Since CelebA-HQ does not provide the identity partition required for our evaluation, we extract identity embeddings with ArcFace [7] for each image and cluster the embeddings with DBSCAN [8] in cosine space, using $\epsilon = 0.35$ and a minimum cluster size of two images. This yields 27,996 successfully processed images grouped into 9,683 identity clusters. Details are deferred to the supplementary material.

Each unlearning experiment targets a single forget identity, splitting its images into forget training and validation subsets ($65/35$ ratio). The remaining images form the retain set, with a $90/10$ split between training and unseen retain evaluation. During unlearning, the forget training set is extended as described in Section 3.4 to generate conditioning samples up to the batch size. Each training step additionally samples a batch of retain conditions from the retain training pool, while the anchor is set as the centroid of the anchor

identity. For the comparison with the state-of-the-art, we repeat this procedure independently over 50 forget identities, always starting from the original pretrained Arc2Face model. To prevent the demographic skew inherent to random sampling, these identities are manually curated to ensure diverse representation across race and gender with with an average cluster size of 15 images per identity, while retaining variance in pose, expression, and lighting, as shown in the supplementary material. Evaluation uses 25 generated images for each of the forget and retain test splits.

Evaluation metrics. We assess identity removal using *Identity Score Matching* (ISM) [44], which separately measures the identity similarity of generated forget and retain samples to their identity centroid in the space of a pretrained ArcFace recognition model [7]. Since centroid similarity alone does not fully characterize forgetting, we also report the *Selective Removal and Keep* (SRK) [39] measure, which jointly evaluates forgetting and preservation via nearest-centroid classification, providing a stronger notion of selective identity removal. Details are provided in the supplementary material, along with additional results using the AdaFace model[23], to mitigate potential circularity from evaluating in the conditioning feature space.

For image quality and realism, we rely on a *Kernel-based distribution Distance* (KD) [15] in the feature space of DINOv2 [28], implemented as an unbiased Maximum Mean Discrepancy estimator [42]. Compared to Fréchet Inception Distance [17], this measure better suits small-sample regimes and captures finer facial properties. We report it as ΔKD, computed between the retain validation distribution and generated forget or retain samples after unlearning, where lower values indicate closer alignment with the reference distribution. We complement this with the *eDIFFIQA* [2] quality measure, which estimates the suitability of face images for recognition. Both ΔKD and eDIFFIQA averaged over all evaluated identities.

4.1. Comparison with the State-of-the-art

To evaluate the effectiveness of our PIU method, we benchmark its performance against representative state-of-the-art methods from distinct unlearning paradigms, specifically SISS [1] for data unlearning, UCE [12] for concept unlearning, and WID [39], to our knowledge the only prior method explicitly designed for identity unlearning in an ID-conditioned diffusion model. To ensure a fair comparison, we use identical identities, seeds, and anchors for all methods. Details regarding the adaptation of these methods for Arc2Face [29] are provided in the supplementary material.

Quantitative results. Results in Table 1 demonstrate that PIU outperforms all prior methods across all unlearning and utility metrics. SISS and UCE perform poorly for unlearning in ID-conditioned models. WID achieves comparable forget ISM to PIU, but (i) requires an encoder inference pass during each training step, and (ii) degrades the general model, as evidenced by the drop in retain ISM. Consequently, PIU achieves the highest SRK result.

Table 1. **Comparison with other unlearning methods.** PIU achieves the strongest identity unlearning (lowest forget ISM and highest SRK) while preserving non-target identities and image quality (best retain ISM, ΔKD, and eDIFFIQA scores). Bold denotes the best result for each metric; underline the second best.

	ISM		SRK ↑	ΔKD		eDIFFIQA	
Model	Forget ↓	Retain ↑		Forget ↓	Retain ↓	Forget ↑	Retain ↑
Arc2Face [29]	0.78	0.75	0.99	–	–	0.76	0.76
SISS [1]	0.59	0.58	0.98	5.23	8.14	0.65	<u>0.64</u>
UCE [12]	0.59	<u>0.70</u>	1.00	**3.69**	<u>1.26</u>	<u>0.75</u>	**0.75**
WID [39]	<u>0.32</u>	0.44	<u>49.28</u>	<u>4.35</u>	3.36	**0.76**	**0.75**
PIU (Ours)	**0.31**	**0.72**	**88.96**	8.12	**0.39**	<u>0.75</u>	**0.75**

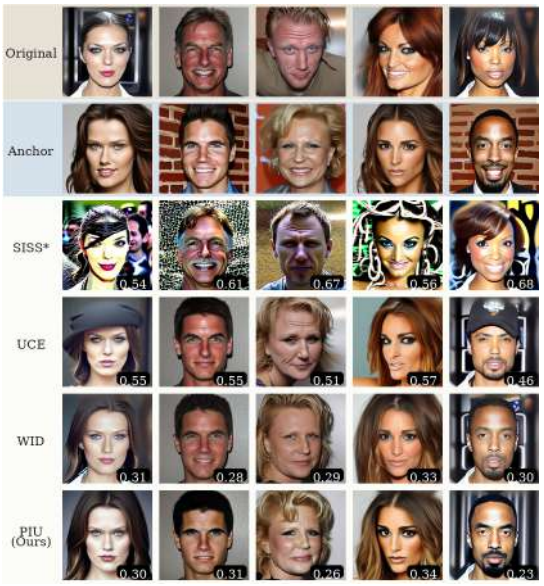

Figure 4. **Images generated by Arc2Face before (first row) and after unlearning with different methods (third row onward).** All samples use fixed original ID embeddings and initial noise. Unlearning shifts the output from the original identity toward the facial characteristics of the synthetic anchor identity (second row), apart from the anchorless SISS method. For each unlearned image, the cosine similarity to the original identity is also reported.

PIU also demonstrate strong performance in preserving face recognition quality for both forget- and retain-conditioned generated samples, as measured by eDIFFIQA. The absolute change in kernel distance (ΔKD) shows that the retain set experiences negligible drift after PIU unlearning (lowest among all methods), due to the preservation-enforcing term in our loss. In contrast, ΔKD for the forget set is higher for PIU than for other methods, driven by the negative guidance scale η (see Tables 2 and 5), and the anchor proximity value τ (see Table 4).

Qualitative results. The qualitative assessment of PIU in Figure 4 shows that post-unlearning samples effectively shift the original visual features toward those of their se-

Figure 5. **Arc2Face samples of identities held-out from unlearning.** The same conditioning embeddings and noise are used for both the original and unlearned models, enabling a comparison of quality and identity preservation on non-target identities. Reported is the cosine similarity to the original non-target identity.

lected anchors. Both PIU and WID successfully achieve visual identity replacement. In contrast, UCE retains some of the original features (e.g. skin color), while the anchorless SISS significantly degrades visual quality and realism. Importantly, non-unlearned retain samples for a given forget–anchor pair in Figure 5 show that post-PIU outputs remain nearly identical to those generated prior to unlearning. Differently, WID exhibits clear anchor contamination in retain samples, as it lacks any preservation constraint. UCE introduces minor variations, while SISS does not alter identity but noticeably degrades visual realism.

4.2. Ablation Studies

For efficiency, ablation experiments are conducted on a fixed subset of 10 forget identities, reporting the metrics most relevant to each component under study. Table 2 summarizes the performance as each component of PIU is introduced incrementally to isolate the contribution of each design choice. The naive unlearning baseline (i.e., $\lambda = 0$, $\eta = 0$) shows that optimizing only the forget objective leads to poor identity preservation. Incorporating the preservation term $\mathcal{L}_{\text{retain}}$ with weight $\lambda = 10$ stabilizes the retain set, maintaining a higher average ISM. Furthermore, the negative guidance scale η strengthens the unlearning effect, yielding a clear reduction in forget ISM and a substantial improvement in SRK. The ablation demonstrates that both components, λ and η, are necessary to achieve effective unlearning while preserving model quality. The KD metric exhibits only minor changes from pre- to post-unlearning, which are primarily confined to the targeted forget identities, while the retain set remains largely unaffected.

Table 2. **Ablation of PIU components.** We incrementally introduce the preservation term, negative guidance, and surgical finetuning to a naive unlearning baseline (λ=0 and η=0). As shown, all components are essential to achieve effective identity removal without degrading quality and non-target identity consistency.

	ISM		ΔKD		SRK
Variant	Forget ↓	Retain ↑	Forget ↓	Retain ↓	↑
Baseline	0.760	0.734	–	–	0.99
Naive	**0.250**	0.410	7.713	4.331	55.6
+ Preservation	0.564	**0.716**	**5.873**	0.284	1.6
+ Neg. guidance	0.298	0.714	8.542	**0.230**	74.1
+ Surgical training	0.280	0.709	9.431	**0.230**	**90.1**

Table 3. **Effect of the preservation weight λ.** Increasing λ stabilizes model degradation but weakens unlearning. Intermediate values yield the best balance of forgetting and retain preservation.

	ISM		Anchor ISM	SRK	ΔKD	
λ	Forget ↓	Retain ↑	Retain ↓	↑	Forget ↓	Retain ↓
0	**0.048**	0.229	0.395	11.2	11.601	13.720
5	0.213	0.700	0.038	**90.3**	**8.596**	0.321
10	0.280	0.709	0.022	90.1	9.431	0.230
20	0.376	**0.712**	**0.016**	27.2	9.266	**0.193**

Effect of the preservation weight λ. The preservation weight λ controls the trade-off between effective unlearning and preservation of the pretrained identity prior (Table 3). Without preservation ($\lambda = 0$), retain ISM drops markedly, as the model tends to collapse and redirect non-target identities toward the anchor, which is also reflected in higher Anchor ISM. Increasing λ stabilizes retain generations and prevents this collapse at the cost of weaker forgetting, with higher forget ISM and lower SRK. We find that intermediate values like $\lambda = 10$ provide the best balance between identity removal and preservation of generative capabilities.

Effect of anchor selection. We analyze the impact of anchor proximity by varying the target cosine similarity threshold τ used to select the anchor identity. For this ablation, we use a batch size of 16, and evaluate on 20 forget identities, to assess this choice over a broader identity set. Results in Table 4 and samples in the supplementary material (Fig. 9) show a clear trade-off: as the anchor moves farther from the forget identity, forgetting becomes stronger, as reflected by lower forget ISM, but preservation degrades, with lower retain ISM and lower SRK. In this sense, $\tau = 0.1$ is the strongest setting from a pure unlearning perspective, but it also leads to worse image quality. By contrast, $\tau = 0.2$ achieves the best retain ΔKD and maintains a more balanced forget ΔKD, yielding the most favorable overall trade-off between effective identity removal, preservation, and generation quality. For $\tau = 0.3$, the anchor becomes increasingly close to the forget identity, which weakens forgetting and reduces SRK. Larger thresholds like $\tau = 0.4$ and $\tau = 0.5$ are less reliable, since too similar or even same-identity samples can be selected as anchors, as shown in the supplementary material.

Effect of the negative guidance scale η. The negative guidance term pushes the predicted noise away from the

Table 4. **Effect of anchor proximity τ.** Lower values maximize unlearning at the cost of image quality, while larger values weaken identity removal. $\tau = 0.2$ provides a balance between identity forgetting, non-target preservation, and image quality.

	ISM		ΔKD		SRK
Anchor	Forget ↓	Retain ↑	Forget ↓	Retain ↓	↑
$\tau = 0.1$	**0.289**	0.712	7.837	0.468	**65.82**
$\tau = 0.2$	0.392	0.713	7.698	**0.269**	46.57
$\tau = 0.3$	0.463	**0.716**	**5.372**	0.464	23.92
$\tau = 0.4$	0.449	0.714	6.450	0.391	42.15
$\tau = 0.5$	0.451	0.713	6.737	0.482	42.68

Table 5. **Effect of the negative guidance scale η.** Setting $\eta = 0$ is insufficient to remove the target identity. Increasing η improves forgetting and SRK performance, but large values may degrade the ΔKD and eDIFFIQA quality of forget-conditioned generations.

	ISM		SRK	ΔKD		eDIFFIQA	
η	Forget ↓	Retain ↑	↑	Forget ↓	Retain ↓	Forget ↑	Retain ↑
0.0	0.564	0.716	1.6	**5.873**	**0.284**	0.745	0.747
1.0	0.341	**0.721**	82.1	9.047	0.377	**0.746**	**0.759**
1.5	0.283	0.720	90.1	9.578	0.431	0.732	0.757
3.0	**0.185**	0.694	**90.4**	13.502	0.383	0.672	0.751

embeddings of the identities to be forgotten (see Eq. 1). Subtracting the anchor-conditioned prediction from the forget-conditioned one penalizes components that remain aligned with the original identity. When $\eta = 0$, the procedure reduces to matching the anchor prediction, encouraging generations that resemble the anchor. However, as seen in Table 5) and Figure 6, this does not actively suppress the underlying original identity. ISM with respect to the original identity remains high, leading to unchanged similarity-based matching and poor SRK performance. Increasing η strengthens the unlearning effect, resulting in lower forget ISM and improved SRK, but can degrade generation quality with larger values, as reflected by increased KD scores. We find that $\eta = 1$ and $\eta = 1.5$ provide the best trade-offs. Although $\eta = 1.5$ gives stronger unlearning metrics, we observe that $\eta = 1$ with longer training reaches comparable unlearning while preserving quality more reliably. Indeed, on the same 10 identities with 400 training steps, $\eta = 1$ reaches SRK $= 90.1$ and forget ISM $= 0.306$, while $\eta = 1.5$ reaches SRK $= 90.7$ and forget ISM $= 0.247$, with similar behavior in the remaining metrics. Thus, we use $\eta = 1$ as our default for results in Table 1.

Effect of fine-tuned layers. Table 6 compares three fine-tuning regimes: the full network, all cross-attention (CA) layers, and a surgical subset of CA layers. Fine-tuning the full model yields the weakest results, with unstable training and degraded quality for both forget and retain identities, reflected in consistent drops across all reported metrics. In contrast, both CA-based strategies achieve strong unlearning, with low forget ISM and high SRK, while updating only $5.11\,\%$ and $4.29\,\%$ of the parameters for full CA and surgical CA, respectively. Their difference is more subtle: full CA fine-tuning shows more stable KD behavior on

Figure 6. **Effect of the negative guidance scale η.** Increasing η strengthens the shift away from original identities, but higher values degrade visual quality. With $\eta = 0$, samples preserve the highest quality, but retain recognizable traits of original identities.

Table 6. **Effect of fine-tuned layers.** Restricting PIU to Cross-Attention (CA) layers, especially the surgical subset, yields the best trade-off between unlearning, preservation, and efficiency.

Tuning	ISM		ΔKD		SRK	eDIFFIQA	
(Parameters)	Forget ↓	Retain ↑	Forget ↓	Retain ↓	↑	Forget ↑	Retain ↑
Full (900M)	**0.134**	0.294	16.950	20.293	51.9	0.653	0.543
Full CA (44M)	0.298	**0.714**	**8.542**	**0.230**	74.1	0.726	0.752
Surgical (**37M**)	0.280	0.709	9.431	**0.230**	**90.1**	**0.729**	**0.753**

forget samples, whereas surgical fine-tuning achieves better SRK while maintaining comparable ISM and image quality.

5. Conclusion

In this paper, we introduced PIU, a proximity-guided framework for identity unlearning in ID-conditioned diffusion models. By combining anchor-guided identity replacement, preservation-aware optimization, and localized fine-tuning of identity-sensitive layers, PIU achieves effective unlearning of target identities while maintaining realism and identity consistency for retained identities. Extensive experiments on Arc2Face [29] demonstrate that PIU consistently delivers the best overall trade-off between forgetting strength, non-target preservation, and image quality, surpassing prior methods for unlearning in the ID-conditioned setting. While our study focuses on single-identity unlearning in Arc2Face, the proposed PIU framework is general and opens directions for future work, including extensions to other ID-conditioned generative models and sequential unlearning scenarios, for which localized fine-tuning and proximity-based anchors minimize catastrophic forgetting.

Acknowledgements

Supported in parts by the Slovenian Research and Innovation Agency (ARIS) through Research Programmes P2-0250 (B) "Metrology and Biometric Systems" and P2–0214 (A) "Computer Vision", the ARIS Young Researcher Programme, and the European Union's Horizon Europe research and innovation programme through the OnMoveID project under grant agreement No. 101225635.

References

[1] S. Alberti, K. Hasanaliyev, M. Shah, and S. Ermon. Data unlearning in diffusion models. In *International Conference on Learning Representations (ICLR)*, pages 1–17, 2025.

[2] Ž. Babnik, P. Peer, and V. Štruc. eDifFIQA: Towards efficient face image quality assessment based on denoising diffusion probabilistic models. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 6(4):458–474, 2024.

[3] L. Bourtoule, V. Chandrasekaran, C. A. Choquette-Choo, H. Jia, A. Travers, B. Zhang, D. Lie, and N. Papernot. Machine unlearning. In *IEEE Symposium on Security and Privacy (SP)*, pages 141–159, 2021.

[4] F. Boutros, V. Struc, J. Fierrez, and N. Damer. Synthetic data for face recognition: Current state and future prospects. *Image and Vision Computing*, 135:104688, 2023.

[5] N. Carlini, J. Hayes, M. Nasr, M. Jagielski, V. Sehwag, F. Tramer, B. Balle, D. Ippolito, and E. Wallace. Extracting training data from diffusion models. In *USENIX Security Symposium*, pages 5253–5270, 2023.

[6] Q. Chen, C.-W. Cheng, X. Su, H. Xu, X. Lin, S. You, A. I. Aviles-Rivero, and Y. Chen. Legato: Good identity unlearning is continuous. *arXiv preprint arXiv:2601.04282*, 2026.

[7] J. Deng, J. Guo, N. Xue, and S. Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4690–4699, 2019.

[8] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *International Conference on Knowledge Discovery and Data Mining (KDD)*, volume 96, pages 226–231, 1996.

[9] C. Fan, J. Liu, Y. Zhang, E. Wong, D. Wei, and S. Liu. Salun: Empowering machine unlearning via gradient-based weight saliency in both image classification and generation. *arXiv preprint arXiv:2310.12508*, 2023.

[10] R. Gal, Y. Alaluf, Y. Atzmon, O. Patashnik, A. H. Bermano, G. Chechik, and D. Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022.

[11] R. Gandikota, J. Materzynska, J. Fiotto-Kaufman, and D. Bau. Erasing concepts from diffusion models. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2426–2436, 2023.

[12] R. Gandikota, H. Orgad, Y. Belinkov, J. Materzyńska, and D. Bau. Unified concept editing in diffusion models. In *IEEE/CVF Winter conference on Applications of Computer Vision (WACV)*, pages 5111–5120, 2024.

[13] E. Goldman. An introduction to the california consumer privacy act (CCPA). *Santa Clara Univ. Legal Studies Research Paper*, 2020.

[14] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. *Advances in Neural Information Processing Systems (NeurIPS)*, 27, 2014.

[15] A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, and A. Smola. A kernel method for the two-sample-problem. *Advances in Neural Information Processing Systems (NeurIPS)*, 19, 2006.

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

[17] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in Neural Information Processing Systems (NeurIPS)*, 30, 2017.

[18] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 6840–6851, 2020.

[19] J. Ho and T. Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

[20] T. Karras, T. Aila, S. Laine, and J. Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *International Conference on Learning Representations (ICLR)*, 2018.

[21] T. Karras, T. Aila, S. Laine, and J. Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *International Conference on Learning Representations (ICLR)*, 2018.

[22] T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4401–4410, 2019.

[23] M. Kim, A. K. Jain, and X. Liu. AdaFace: Quality adaptive margin for face recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18750–18759, 2022.

[24] N. Kumari, B. Zhang, S.-Y. Wang, E. Shechtman, R. Zhang, and J.-Y. Zhu. Ablating concepts in text-to-image diffusion models. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22691–22702, 2023.

[25] Z. Li, M. Cao, X. Wang, Z. Qi, M.-M. Cheng, and Y. Shan. Photomaker: Customizing realistic human photos via stacked id embedding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8640–8650, 2024.

[26] Z. Liu, K. Chen, Y. Zhang, J. Han, L. Hong, H. Xu, Z. Li, D.-Y. Yeung, and J. T. Kwok. Implicit concept removal of diffusion models. In *Springer European Conference on Computer Vision (ECCV)*, pages 457–473, 2024.

[27] C. Lu, Y. Zhou, F. Bao, J. Chen, C. Li, and J. Zhu. DPM-Solver++: Fast solver for guided sampling of diffusion probabilistic models. *Machine Intelligence Research (MIR)*, 22(4):730–751, 2025.

[28] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research (TMLR)*, 2024.

[29] F. P. Papantoniou, A. Lattas, S. Moschoglou, J. Deng, B. Kainz, and S. Zafeiriou. Arc2face: A foundation model for id-consistent human faces. In *Springer European Conference on Computer Vision (ECCV)*, pages 241–261, 2024.

[30] D. Protection. General data protection regulation. *Intersoft Consulting, Accessed in October*, 24(1), 2018.

[31] F. Qi, A. Liu, Z. Zhang, and C. Xu. Forget me: Federated unlearning for face generation models. In *ACM International Conference on Multimedia (ICK)*, pages 11288–11297, 2025.

[32] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *PMLR International Conference on Machine Learning (ICML)*, pages 8748–8763, 2021.

[33] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022.

[34] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *Springer International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 234–241, 2015.

[35] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22500–22510, 2023.

[36] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823, 2015.

[37] J. Seo, S.-H. Lee, T.-Y. Lee, S. Moon, and G.-M. Park. Generative unlearning for any identity. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9151–9161, 2024.

[38] M. Shaheryar, J. T. Lee, and S. K. Jung. Black hole-driven identity absorbing in diffusion models. In *IEEE/CVF Computer Vision and Pattern Recognition Conference (CVPR)*, pages 28544–28554, 2025.

[39] M. Shaheryar, J. T. Lee, and S. K. Jung. Unlearn and protect: Selective identity removal in diffusion models for privacy preservation. In *ACM/SIGAPP Symposium on Applied Computing*, pages 1172–1179, 2025.

[40] G. Somepalli, V. Singla, M. Goldblum, J. Geiping, and T. Goldstein. Diffusion art or digital forgery? investigating data replication in diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6048–6058, 2023.

[41] G. Somepalli, V. Singla, M. Goldblum, J. Geiping, and T. Goldstein. Understanding and mitigating copying in diffusion models. *Advances in Neural Information Processing Systems (NeurIPS)*, 36:47783–47803, 2023.

[42] G. Stein, J. Cresswell, R. Hosseinzadeh, Y. Sui, B. Ross, V. Villecroze, Z. Liu, A. L. Caterini, E. Taylor, and G. Loaiza-Ganem. Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models. *Advances in Neural Information Processing Systems (NeurIPS)*, 36:3732–3784, 2023.

[43] D. Tomašević, F. Boutros, C. Lin, N. Damer, V. Štruc, and P. Peer. Id-booth: Identity-consistent face generation with diffusion models. In *IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–10, 2025.

[44] T. Van Le, H. Phung, T. H. Nguyen, Q. Dao, N. N. Tran, and A. Tran. Anti-dreambooth: Protecting users from personalized text-to-image synthesis. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2116–2127, 2023.

[45] J. Wu, T. Le, M. Hayat, and M. Harandi. Erasing undesirable influence in diffusion models. In *IEEE/CVF Computer Vision and Pattern Recognition Conference (CVPR)*, pages 28263–28273, 2025.

[46] G. Xiao, T. Yin, W. T. Freeman, F. Durand, and S. Han. Fastcomposer: Tuning-free multi-subject image generation with localized attention. *Springer International Journal of Computer Vision (IJCV)*, 133(3):1175–1194, 2025.

[47] Z. Yan, Y. Zhang, Y. Fan, and B. Wu. Ucf: Uncovering common features for generalizable deepfake detection. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22412–22423, 2023.

[48] H. Ye, J. Zhang, S. Liu, X. Han, and W. Yang. IP-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023.

[49] Y. Zhang, E. Jin, Y. Dong, Y. Wu, P. Torr, A. Khakzar, J. Stegmaier, and K. Kawaguchi. Minimalist concept erasure in generative models. *arXiv preprint arXiv:2507.13386*, 2025.

PIU: Proximity-Guided Identity Unlearning in ID-Conditioned Diffusion Models

Supplementary Material

A. Additional Experimentation Details

Dataset Details. As explained in Section 4, we use the CelebA-HQ [20]. Specifically we use the version available on the Hugging Face Hub[1], with the snapshot at revision `7ecc6a45edfb5483ccf2f7df1035d298ffe7c76b`. To perform unlearning, it is necessary to identify the identities present in the dataset. However, the referenced CelebA-HQ dataset provides only gender labels, with no identity annotations. For this reason we design a clustering strategy to automatically group samples belonging to the same identity.

We cluster the embedding space using DBSCAN, a density-based method that groups samples according to local similarity without requiring a predefined number of clusters. We use the Scikit-Learn implementation with cosine distance.

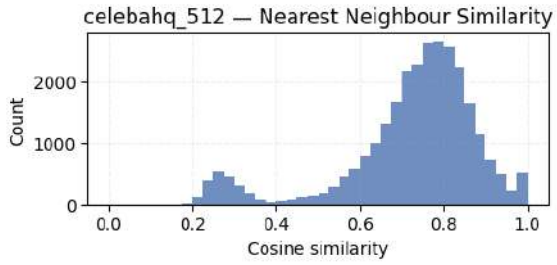

Figure 7. Distribution of the cosine similarity of each sample's nearest neighbor. Two evident modes appear for same- and similar-identity samples.

Figure 7 shows the distribution of cosine similarity for all samples' nearest neighbors. Two distinct modes emerge, with peaks around $\tau \approx 0.8$ and $\tau \approx 0.25$. The high-similarity peak corresponds to samples of the same identity, while the lower peak captures ArcFace-similar but distinct individuals. Between these modes, a minimum appears around $\tau \approx 0.4$, providing a natural separation threshold. Based on this observation, we set the cosine distance threshold to $\epsilon = 0.35$ and require a minimum of two samples per cluster. These choices define the DBSCAN configuration used for identity grouping. This procedure produces 9,683 clusters, which we use as identity labels during training. Manual inspection confirms that the resulting clusters are visually consistent. Empirically, we observe that similarities above $\tau > 0.6$ almost always correspond to the same individual, with only rare exceptions arising from varying lighting conditions or significant facial occlusions.

[1] `https://huggingface.co/datasets/jxie/celeba-hq`

Manual Identity Selection for Unlearning. As discussed in the main text, evaluating unlearning methods on a purely random sample from CelebA-HQ can result in an unbalanced evaluation set heavily skewed toward specific demographics. To ensure a rigorous and fair evaluation, we manually curated the 50 target identities used in our benchmark comparisons. This curation process prioritized diverse representation across different demographic groups, specifically balancing race and gender, while retaining variance in facial poses, expressions, and lighting conditions to test the unlearning framework under realistic, in-the-wild variations. Figure 8 provides a visual overview of one representative image from each of the 50 selected identities.

Figure 8. **Overview of the 50 manually curated target identities.** The selection enforces demographic diversity across gender and race alongside variations in facial pose, expression, and lighting.

Details of Unlearning Metrics. We assess identity removal using *Identity Score Matching* (ISM) [44] and *Selective Removal and Keep* (SRK) [39]. Although both metrics have appeared in prior work, their exact computation is often only briefly described or left implicit, especially in the context of ID-conditioned face generation. For reproducibility, we provide here the precise definitions and implementation details used in our evaluation.

Let μ_k denote the ArcFace centroid of identity k, computed from its training images as in Equation 5. Given a set of generated samples $\{y_j^{(k)}\}_{j=1}^{n}$ associated with identity k, ISM is defined as the average cosine similarity between the

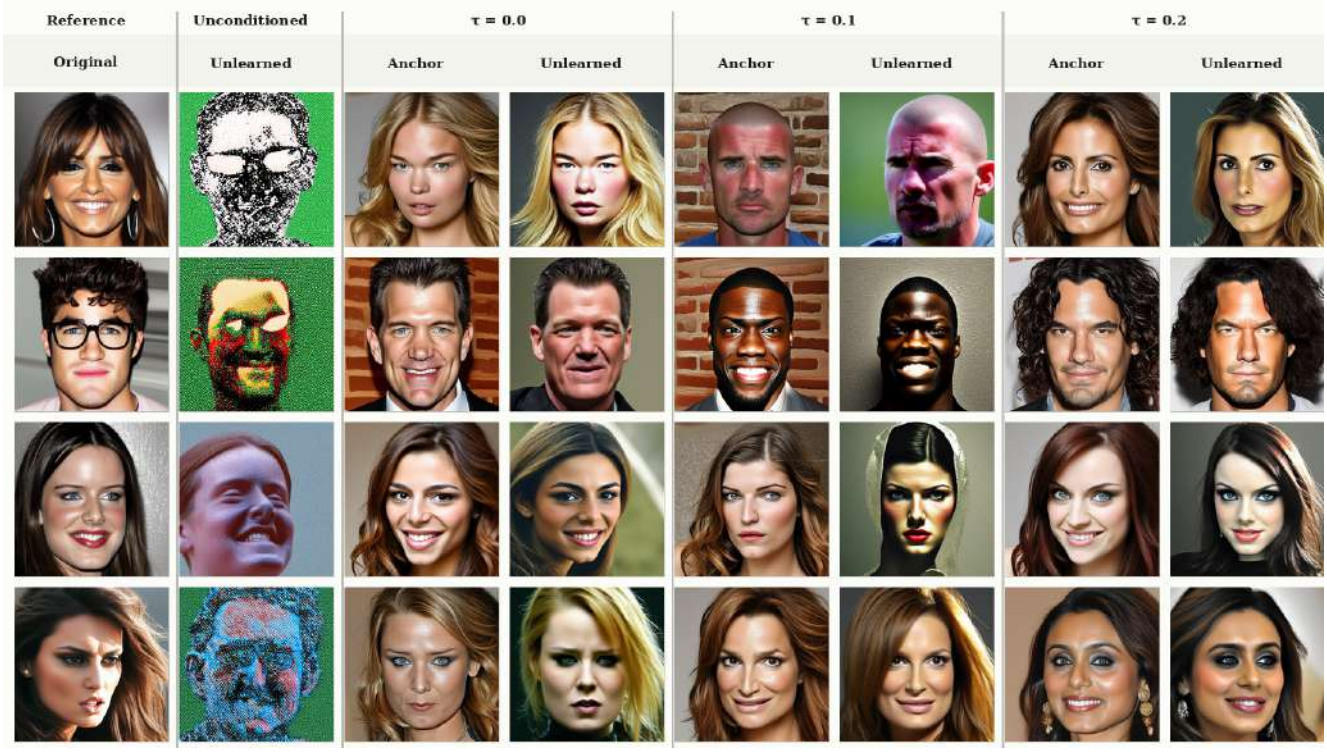

Figure 9. **Qualitative comparison of PIU under different anchor selection strategies.** True unconditional guidance degrades realism and leads to unstable generations. In contrast, anchors with increasing proximity progressively improve visual quality and identity consistency.

ArcFace embedding of each generated image and its corresponding identity centroid:

$$\mathrm{ISM}(k) = \frac{1}{n}\sum_{j=1}^{n} \cos\big(\mathrm{ArcFace}(y_j^{(k)}), \mu_k\big). \tag{6}$$

We use this metric both for forget and retain identities. For a forget identity, a low ISM indicates successful unlearning, since generated samples no longer remain close to its original identity representation in ArcFace space. For retain identities, on the contrary, a high ISM indicates that the identity prior is preserved after unlearning. We also use *Selective Removal and Keep* (SRK) [39], which evaluates forgetting and preservation jointly through nearest-centroid classification.

For a set of unlearned identities $\mathcal{U}$, we define the forget accuracy as

$$\mathrm{Acc}_U = \frac{1}{|\mathcal{U}|\, m}\sum_{u\in\mathcal{U}}\sum_{j=1}^{m} \mathbf{1}\{\hat{y}_j^{(u)} = u\}, \tag{7}$$

where $\hat{y}_j^{(u)}$ is the identity predicted by nearest-centroid cosine similarity for the j-th generated sample associated with identity u. A successful unlearning method should make Acc_U low, ideally close to zero.

Similarly, for a set of retained identities $\mathcal{R}$, we define the retain accuracy as

$$\mathrm{Acc}_R = \frac{1}{|\mathcal{R}|\, m}\sum_{r\in\mathcal{R}}\sum_{j=1}^{m} \mathbf{1}\{\hat{y}_j^{(r)} = r\}. \tag{8}$$

A good preservation of the Arc2Face identity prior should keep Acc_R high.

These two quantities are combined into the SRK score:

$$\mathrm{SRK} = \frac{\mathrm{Acc}_R}{\mathrm{Acc}_U + \varepsilon}, \tag{9}$$

where $\varepsilon = 1e^{-2}$ is a small constant to avoid division by zero. Therefore, a higher SRK indicates better performance, as it reflects both effective removal of the target identity and preservation of the remaining ones. In this way, SRK complements ISM by checking not only whether an identity moves away from its original centroid, but also whether this shift causes confusion with other identities in the embedding space.

B. Additional Results

Anchor Selection. Figure 9 illustrates the impact of anchor proximity on the unlearning behavior. Using true unconditional guidance results in degraded realism and unstable generations, confirming that the model relies on a meaningful identity prior. As the anchor proximity increases, the

Table 7. **Comparison with other unlearning methods. [TODO: Add ISM and SRK values computed with AdaFace ..]** PIU achieves the strongest identity unlearning (lowest forget ISM and highest SRK) ... Bold denotes the best result for each metric; underline the second best.

	ArcFace			AdaFace		
	ISM		SRK ↑	ISM		SRK
Model	Forget ↓	Retain ↑		Forget ↓	Retain ↓	
Arc2Face [29]	0.78	0.75	0.99	–	–	–
SISS [1]	0.59	0.58	0.98	–	–	–
UCE [12]	0.59	<u>0.70</u>	1.00	–	–	–
WID [39]	<u>0.32</u>	0.44	<u>49.28</u>	–	–	–
PIU (Ours)	**0.31**	**0.72**	**88.96**	**0.31**	**0.72**	**92.64**

generated samples become progressively more stable and visually consistent, indicating a smoother transition in the identity space. In practice, we select $\tau = 0.2$, as it provides the best trade-off between stable unlearning and high-quality post-unlearning generations when conditioning on the target identity.

Main results. The examples in figure 10 demonstrate that PIU achieves stable identity shifts across challenging settings, including cross-gender and cross-ethnicity transformations. By keeping the conditioning embeddings and initial noise fixed, the observed changes can be directly attributed to the unlearning procedure rather than stochastic variation. Training is performed for 400 steps with $\eta = 1.0$, using a batch size of 32 for both forget and retain samples per optimization step. Under these conditions, the method consistently drives the generated identities toward their assigned anchors while preserving the model's visual realism.

Generalization to different identity spaces. Because Arc2Face utilizes ArcFace embeddings for identity conditioning, evaluating unlearning success within the same ArcFace latent space might introduce a potential circularity bias. To confirm that our method genuinely erases the identity features rather than just manipulating the ArcFace-specific manifold, we re-evaluated the identity related metrics (ISM and SRK) using a pretrained AdaFace recognition model [23]. As shown in Table 7, the AdaFace results closely mirror our original findings (e.g., PIU achieves a Forget/Retain ISM of $0.31/0.71$ and an SRK of 92.64). This confirms that the identity unlearning achieved by PIU effectively transfers across different feature spaces and is not an artifact of the conditioning network.

C. Generalization to other generative models

[TODO: Performance on IDSync and their training dataset – at least qualitative examples]

D. Reproducing Unlearning Methods in Identity-Conditioned Diffusion

In this section, we describe the reproduction and implementation details of the three baseline methods compared against our proposed method, PIU, in Section 4.1. We explain how each baseline was reconstructed, adapted to the Arc2Face framework, and evaluated. In particular, Subsection D.1 covers our reproduction of *Subtracted Importance Sampled Scores (SISS)* [1] as the data-level unlearning baseline; Subsection D.2 covers our reproduction of *Unified Concept Editing in Diffusion Models (UCE)* [12] as the concept-unlearning baseline; and Subsection D.3 covers our reproduction of *the Unlearn and Protect variant With ID-guidance (WID)* [39] as the identity-unlearning baseline.

Overview and Reproducibility Protocol. All reproduced methods follow the same structure. We first briefly introduce the method and motivate its inclusion as a baseline for PIU. In the *Original Method* subsection, we summarize the parts of the original formulation that are most relevant to our reproduction. In *Our Adaptation*, we explain how the method is transferred to Arc2Face while remaining as faithful as possible to its original assumptions and design principles. The *Experiments* subsection then describes the implementation details, evaluation protocol, and hyperparameter searches used to identify the strongest baseline configuration for fair comparison with PIU. Finally, in *Results and Discussion*, we analyze the obtained results and compare them with the claims reported in the original papers.

Adaptation Challenges for Arc2Face. A dedicated adaptation subsection is needed because none of the considered baselines was originally proposed for the setting studied in this work. Arc2Face is an identity-conditioned latent diffusion model, whose conditioning mechanism differ substantially from the unconditional or text-conditioned settings assumed by prior unlearning methods (Section 3.1). As a result, reproduction in our case requires more than direct implementation: the main components of each method must be reformulated so that they remain meaningful under identity conditioning. We therefore make explicit which parts can be transferred directly, which must be modified, and how these modifications were chosen to preserve the original intent of each method as closely as possible.

D.1. Subtracted Importance Sampled Scores

Among existing data unlearning methods for diffusion models, we choose to reproduce and adapt Subtracted Importance Sampled Scores (SISS) [1]. This choice is motivated by several factors. First, to the best of our knowledge, SISS is the most recent representative method in this

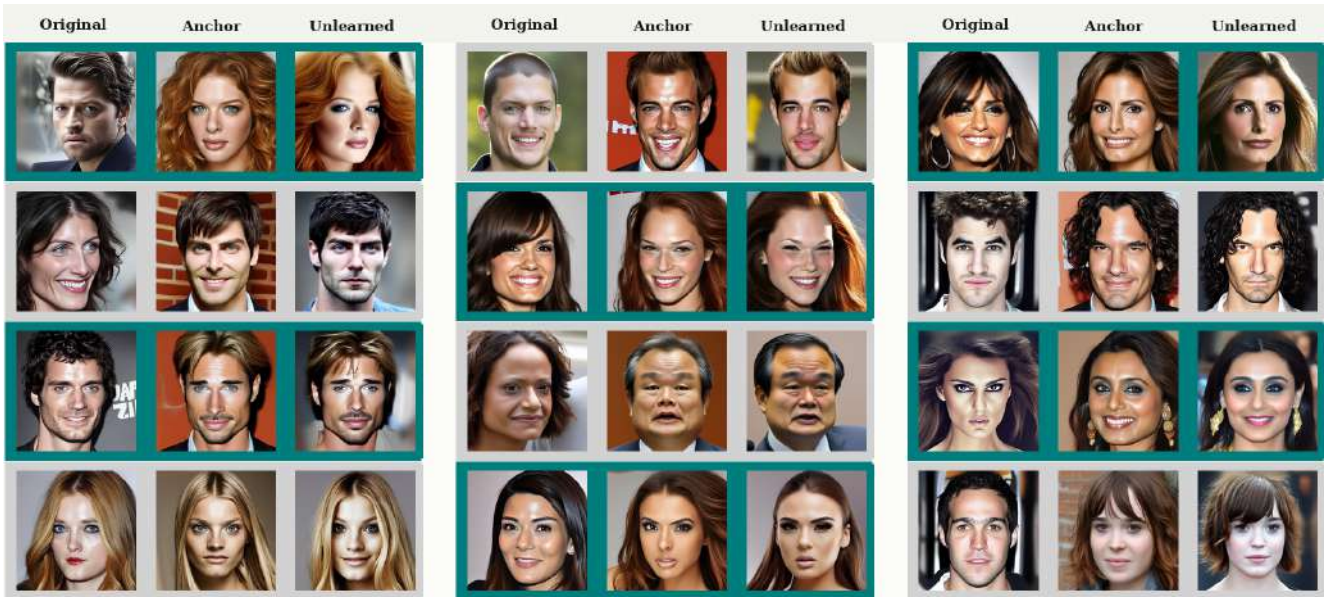

Figure 10. Visualization of PIU results across 12 Original–Anchor–Unlearned triplets. Unlearned samples are generated using the same conditioning embeddings and initial noise as their corresponding original samples. The results show a consistent shift in identity, with outputs moving toward the facial characteristics of the selected anchors. The anchor images are also generated by Arc2Face.

line of work, and it empirically outperforms earlier data unlearning baselines such as EraseDiff [45] on its reported benchmarks. Second, unlike earlier approaches, SISS is accompanied by a theoretical analysis of its objective, a stability analysis, and a well-documented experimental setup in the appendix. Third, the method is evaluated on unconditional diffusion models trained on CelebA-HQ, which is particularly relevant to our setting, since CelebA-HQ is also the dataset underlying our experiments. For these reasons, SISS provides a strong starting point for identity-level data unlearning, although its original formulation is not directly applicable to Arc2Face because our model is identity-conditioned rather than unconditional.

D.1.1 SISS Methodology

We follow the notation of Alberti et al. [1]. Let $X = \{x_1, x_2, \ldots, x_n\}$ denote the dataset used to pretrain a diffusion model ϵ_θ, and let $A = \{a_1, a_2, \ldots, a_k\} \subset X$ be the subset of samples to be forgotten, with $|X| = n$ and $|A| = k$. The objective is to fine-tune the pretrained model so as to approximate the model that would have been obtained had the forget set A not been present during training, while retaining the generative behavior induced by the remaining data $X \setminus A$. In the original paper, this is formalized through two related fine-tuning objectives.

The first formulation, which we refer to as *SISS-No-IS*, is obtained by rewriting the standard diffusion training loss restricted to the retained set $X \setminus A$. Starting from the DDPM forward process

$$q(x_t \mid x_0) = \mathcal{N}(x_t; \gamma_t x_0, \sigma_t I),$$

the objective over the retained distribution can be expressed as

$$L_{X\setminus A}(\theta) = \frac{n}{n-k}\, \mathbb{E}_{\substack{x\sim p_X \\ x_t\sim q(\cdot|x)}} \left\| \frac{x_t - \gamma_t x}{\sigma_t} - \epsilon_\theta(x_t, t) \right\|_2^2 - \frac{k}{n-k}\, \mathbb{E}_{\substack{a\sim p_A \\ a_t\sim q(\cdot|a)}} \left\| \frac{a_t - \gamma_t a}{\sigma_t} - \epsilon_\theta(a_t, t) \right\|_2^2. \tag{10}$$

The notation $L_{X\setminus A}$ emphasizes that this objective corresponds to the diffusion training loss that would ideally be optimized using only the retained data distribution. Equation (10) makes this target explicit by decomposing it into a preservation term, which maintains performance on the full training distribution, and a subtractive forgetting term, which removes the contribution of the forget set A. However, in practice, evaluating (10) requires two separate forward passes through the denoising network, one for samples from X and one for samples from A, which increases both the computational and memory cost.

To address this limitation, Alberti et al. [1] introduce an importance-sampling reformulation, which they call *SISS*. The key idea is to replace the two separate expectations by a single expectation over a proposal mixture distribution $q_\lambda(m_t \mid x, a) := (1-\lambda)\, q(m_t \mid x) + \lambda\, q(m_t \mid a)$, where $\lambda \in [0, 1]$ controls the trade-off between sampling around retained examples and forget examples. The full expression of the resulting objective $\ell_\lambda(\theta)$ is given in the original paper [1].

A central theoretical result in the paper is that this importance-sampled objective is equivalent to the original deletion loss. More precisely, the authors show that

$$\ell_\lambda(\theta) = L_{X\setminus A}(\theta), \qquad \forall \lambda \in [0,1],$$

and further prove that the corresponding gradient estimators are also equal in expectation. Therefore, SISS preserves the same optimization target as the original two-forward formulation. At the same time, by estimating both contributions from a single sampled noisy point m_t, it requires only one forward-backward pass through the denoising network, making it computationally more efficient than the SISS-No-IS formulation.

D.1.2 Adapting SISS to Arc2Face

The empirical results reported in [1] further motivate the use of SISS in our setting. In particular, on CelebA-HQ with an unconditional DDPM, it's shown that both SISS with $\lambda = 0.5$ and SISS-No-IS achieve a favorable trade-off between forgetting strength and quality preservation, with SISS-No-IS yielding slightly stronger unlearning among the quality-preserving variants. They also evaluate the method on Stable Diffusion, showing that it extends to conditional models when the conditioning variable is fixed and reliably identifies the target to be removed. In that setting as well, SISS-No-IS remains slightly stronger empirically than the one-pass SISS estimator.

Although Arc2Face is not an unconditional diffusion model, its conditioning structure is highly constrained: the textual prompt is fixed, and generation is controlled solely through the identity embedding (see Section 3.1). This places Arc2Face in the same regime considered in SISS for conditional models with fixed conditioning. In our case, the forget set, defined in Section 3.4, consists of the identity embeddings associated with the target identity together with their linear combinations, all of which consistently generate the same subject. We therefore treat this identity-conditioned set as A in the SISS formulation and adapt the SISS-No-IS objective accordingly.

We chose SISS-No-IS rather than the importance-sampled SISS objective for two reasons. First, our training setup is not constrained by computation, so the additional cost of two forward passes is acceptable. Second, SISS-No-IS directly optimizes the original two-term deletion objective, and the original paper reports it to be slightly stronger empirically in the identity-removal regime.

At the same time, the released implementation does not optimize the theoretical SISS-No-IS loss exactly in the scalar form of Eq. (10). From our inspection of the official codebase, the retain and forget terms are evaluated separately, and the parameter update is constructed through a gradient-level subtraction in which the forgetting contribution is further rescaled using gradient norms. Thus, the practical implementation is better interpreted as an empirical optimization variant of the theoretical objective, rather than as a literal minimization of Eq. (10).

To clarify this connection, we next rewrite Eq. (10) in a form that exposes more directly its preservation and forgetting components. This derivation, and the corresponding interpretation of the implementation, are not given explicitly in the original paper; they are part of our own analysis based on the published formulation and the released code. First of all, we rewrite the loss 10:

$$L_{\text{SISS-No-IS}}(\theta) = w_x\, \mathcal{L}_x(\theta) - w_a\, \mathcal{L}_a(\theta), \tag{11}$$

where

$$w_x := \frac{n}{n-k}, \qquad w_a := \frac{k}{n-k}, \tag{12}$$

with

$$\mathcal{L}_x(\theta) := \mathbb{E}_{\substack{x\sim p_X \\ x_t\sim q(\cdot|x)}} \left[\left\| \frac{x_t - \gamma_t x}{\sigma_t} - \epsilon_\theta(x_t, t) \right\|_2^2 \right], \tag{13}$$

and

$$\mathcal{L}_a(\theta) := \mathbb{E}_{\substack{a\sim p_A \\ a_t\sim q(\cdot|a)}} \left[\left\| \frac{a_t - \gamma_t a}{\sigma_t} - \epsilon_\theta(a_t, t) \right\|_2^2 \right]. \tag{14}$$

Equivalently, assuming $w_x \neq 0$,

$$L_{\text{SISS-No-IS}}(\theta) = w_x \left(\mathcal{L}_x(\theta) - \frac{w_a}{w_x}\, \mathcal{L}_a(\theta) \right), \tag{15}$$

which yields the effective relative forgetting weight

$$\lambda_{\text{eff}} := \frac{w_a}{w_x} = \frac{k}{n}. \tag{16}$$

At the gradient level, this gives

$$\nabla_\theta L_{\text{SISS-No-IS}}(\theta) = w_x \left(\nabla_\theta \mathcal{L}_x(\theta) - \lambda_{\text{eff}} \nabla_\theta \mathcal{L}_a(\theta) \right). \tag{17}$$

Since $w_x > 0$ is a global multiplicative constant, it does not change the optimization direction induced by the gradient, but only its overall magnitude. Therefore, from the perspective of first-order optimization, the practically relevant quantity is the relative weighting between the retain and forget terms, namely

$$\nabla_\theta \mathcal{L}_x(\theta) - \lambda_{\text{eff}} \nabla_\theta \mathcal{L}_a(\theta). \tag{18}$$

This is precisely the form suggested by the released implementation:

$$L_{\text{SISS-No-IS}}(\theta) \approx \mathcal{L}_x(\theta) - \lambda_{\text{eff}}\, \mathcal{L}_a(\theta), \tag{19}$$

Rather than keeping the theoretical coefficient $\lambda_{\text{eff}} = k/n$ fixed, however, the code rescales the forgetting gradient

through its norm. If $g_x := \nabla_\theta \mathcal{L}_x$ and $g_a := \nabla_\theta \mathcal{L}_a$, the forgetting contribution is multiplied by

$$\lambda_{\text{eff}} := \frac{\alpha}{\|g_a\|_2}, \tag{20}$$

so that the final update direction becomes

$$g_x - \lambda_{\text{eff}} g_a. \tag{21}$$

In the original implementation, α is chosen so that the forget term has magnitude approximately 10% of the retain term, i.e.,

$$\alpha \approx 0.1 \, \|g_x\|_2, \tag{22}$$

which implies

$$\lambda_{\text{eff}} \approx 0.1 \, \frac{\|g_x\|_2}{\|g_a\|_2}. \tag{23}$$

This can be interpreted as a norm-controlled approximation of Eq. (18), where the forgetting weight is determined not only by dataset proportions, but also by the instantaneous geometry of the two gradients.

Adapting this formulation to Arc2Face, we replace the unconditional DDPM objective by the identity-conditioned latent denoising loss. Let c_r denote a retain identity condition and c_f a forget identity condition, with noisy latent z_t defined as in Section 3.1. We define

$$\begin{aligned} \mathcal{L}_{\text{retain}}(\theta) &= \mathbb{E}_{z_0,\epsilon,t,c_r}\left[\|\epsilon - \epsilon_\theta(z_t, t, c_r)\|_2^2\right], \\ \mathcal{L}_{\text{forget}}(\theta) &= \mathbb{E}_{z_0,\epsilon,t,c_f}\left[\|\epsilon - \epsilon_\theta(z_t, t, c_f)\|_2^2\right]. \end{aligned} \tag{24}$$

The corresponding SISS-style Arc2Face objective is then

$$\mathcal{L}_{\text{Arc2Face-SISS}}(\theta) = \mathcal{L}_{\text{retain}}(\theta) - \lambda_{\text{eff}} \, \mathcal{L}_{\text{forget}}(\theta). \tag{25}$$

In practice, following the released implementation, we optimize it through the gradient-level rule

$$\lambda_{\text{eff}} = \frac{\alpha}{\|\nabla_\theta \mathcal{L}_{\text{forget}}\|_2}. \tag{26}$$

We consider the adaptive variant α proportional to the retain gradient norm, therefore

$$\lambda_{\text{eff}} = \beta \frac{\|\nabla_\theta \mathcal{L}_{\text{retain}}\|_2}{\|\nabla_\theta \mathcal{L}_{\text{forget}}\|_2}, \tag{27}$$

where $\beta > 0$ is a tunable hyperparameter.

D.1.3 SISS Experiments

The general implementation details, including dataset construction, training environment, hardware configuration, and evaluation metrics, follow Section 4. Here we only describe the choices specific to the SISS reproduction.

Unlike our proposed method, SISS operates directly on training samples. Therefore, we follow the original formulation and use latents encoded from real images, rather than sampling from Gaussian noise. This ensures consistency with the theoretical objective in Eq. (10), where the expectations are defined over the data distribution.

Due to computational constraints, we conduct the reproduction on a reduced subset of identities. Instead of the full set of 10 forget identities used in our ablation setting, we restrict the experiments to 5 targets. Unless otherwise stated, we use a batch size of 64 (implemented as 16 with gradient accumulation over 4 steps) and run 60 update steps, following the short fine-tuning regime recommended in [1]. Optimization is performed using AdamW with zero weight decay.

Following the recommendations of SISS [1], we take as our reference configuration a learning rate of 5×10^{-6} and a forgetting coefficient $\beta = 0.1$. However, we perform a small grid search to verify whether these hyperparameters remain appropriate in our identity-conditioned latent diffusion setting. In particular, the optimization in Arc2Face differs because generation takes place in latent space, so slightly larger learning rates may accelerate adaptation without necessarily destabilizing training. For this reason, we compare the reference value 5×10^{-6} with a more aggressive alternative, 1×10^{-5}.

For the forgetting-strength coefficient, we evaluate

$$\beta \in \{0.1,\ 0.2,\ 0.3,\ 0.5\}.$$

This range covers both the conservative regime recommended in the original work and progressively stronger forgetting regimes, allowing us to study the trade-off between identity removal and generation quality in our setting. All models are evaluated using the metrics introduced in Section 4, including the identity-removal metrics ISM and SRK, and the quality metrics ΔKD and eDIFFIQA. We report results on both forget identities and retain identities, distinguishing between the seen retain training set and the unseen retain validation set.

D.1.4 SISS Results and Discussion

We first analyze the more aggressive learning rate setting, 1×10^{-5}. As shown in Table 8, this value leads to a clear collapse in generation quality. The degradation is already evident in the realism metrics on retain identities, particularly in the KD-based distances for both seen and unseen retain samples. Qualitatively, generations become visibly corrupted and structurally unstable, especially for identities that should represent the retained distribution $X \setminus A$, indicating that the fine-tuning process no longer preserves the original generative behavior of the model.

This effect is further confirmed by the drop in eDIFFIQA on unseen retain identities. Since eDIFFIQA reflects whether generated faces remain sufficiently coherent and recognizable for face analysis, this deterioration indicates

that the model is no longer producing reliable facial images even outside the forget set. For this reason, we do not include qualitative visualizations or a detailed β-wise analysis for the 1×10^{-5} setting, as the dominant conclusion is simply that this optimization regime is too unstable for our Arc2Face reproduction.

β	Δ Seen KD $\downarrow$	Δ Unseen KD $\downarrow$	Unseen eDIFFIQA $\uparrow$
$\beta = 0.1$	10.210	10.372	0.526
$\beta = 0.2$	11.153	9.819	0.548
$\beta = 0.3$	10.256	9.516	0.592
$\beta = 0.5$	11.109	11.876	0.508

Table 8. Effect of using a higher learning rate (1×10^{-5}) in the SISS reproduction. Results are reported for different values of β. We report only retain-side realism metrics, since the main effect of this setting is a severe degradation of image quality and facial recognizability.

We now turn to the learning rate recommended by the original SISS implementation appendix [1], namely 5×10^{-6}. In contrast to the higher learning rate discussed above, this setting yields stable optimization and allows a more meaningful analysis of the effect of the forgetting coefficient β. The corresponding results are reported in Table 9.

From the perspective of identity unlearning, the results show the expected monotonic trend: increasing β generally strengthens the forgetting effect. This is most clearly reflected in the forget-side ISM, which decreases as β grows, indicating that the generated samples become progressively less aligned with the target identity centroid. The same tendency is partially captured by SRK, which also improves for larger values of β, although the absolute SRK values remain low overall. In particular, when $\beta = 0.1$, the reduction in forget ISM is still very limited, showing that the method performs only weak identity removal in this conservative regime. More noticeable forgetting emerges only for $\beta = 0.3$ and $\beta = 0.5$.

However, this increase in forgetting strength comes at a substantial cost in generation quality. For larger values of β, both KD and eDIFFIQA deteriorate sharply, not only on the forget identities but also on the unseen retain identities. This indicates that the model does not merely become less faithful to the target identity; rather, its overall facial generation quality degrades significantly. A possible explanation is that, unlike our method, this SISS adaptation fine-tunes all model weights rather than restricting the update to the cross-attention layers. In our setting, such unrestricted fine-tuning appears to destabilize the generator more globally, which may explain the stronger degradation in image quality. This effect is especially problematic because it impacts the retain side as well: although the degradation is more severe for the forget set, it remains substantial for the retain set and is clearly worse than in our proposed method. At the same time, the eDIFFIQA values remain relatively high, suggesting that the generated faces are still broadly recognizable despite the loss in fidelity and realism. This is also consistent with the qualitative examples in Figure 11, where the generations remain identifiable but exhibit clear degradation, particularly in sharpness, structure, and overall realism. Even in the more favorable settings, the retain-side quality remains noticeably below the level achieved by our approach, both quantitatively and in the qualitative examples shown in Figure 11.

Taken together, these results suggest that SISS, at least in this adapted form, is not sufficiently effective for identity unlearning in identity-conditioned diffusion models such as Arc2Face. When β is kept small, the method preserves generation quality reasonably well, but the amount of identity removal remains too limited to be competitive. When β is increased to obtain stronger unlearning, the model rapidly loses image realism, leading to an unfavorable trade-off. Therefore, although the original SISS formulation is well motivated for data unlearning in unconditional or weakly conditioned settings, our experiments indicate that its direct extension to Arc2Face does not provide a satisfactory solution for targeted identity erasure.

Among the tested configurations, $\beta = 0.1$ is the most reasonable operating point, as it is the only setting that does not severely compromise the generative quality of the model. Nevertheless, even this configuration yields only weak forgetting according to both ISM and SRK. We therefore conclude that the reproduced SISS baseline is ultimately too mild to achieve strong identity unlearning in our framework, while more aggressive hyperparameter choices destabilize the model and degrade the quality of the generated faces.

D.2. Unified Concept Editing

As discussed in Section 2, concept unlearning has been one of the most explored directions for unlearning in diffusion models, especially in text-to-image settings. Within this literature, Unified Concept Editing (UCE) [12] is a particularly relevant reference point for several reasons. First, it is one of the strongest and most established model-editing approaches for concept erasure, showing competitive or superior performance with respect to influential baselines such as ESD [11] and Ablating Concepts [24], that are similar to our method, while also preserving the quality of non-target generations. Second, UCE naturally supports simultaneous editing of multiple concepts, placing it in the same broader line of scalable multi-concept erasure methods. Third, unlike gradient-based unlearning methods, UCE performs direct closed-form parameter editing without iterative fine-tuning, which makes it conceptually complementary to our own approach. This is especially relevant in our setting, since it allows us to test whether identity unlearning in Arc2Face can also benefit from a direct model-editing strat-

β	Forget ISM ↓	Retain Unseen ISM ↑	SRK ↑	Δ Forget KD ↓	Δ Retain Unseen KD ↓	Forget eDIFFIQA ↑	Retain Unseen eDIFFIQA ↑
Baseline	0.775	0.726	0.990	–	–	0.763	0.770
$\beta = 0.1$	0.605	0.562	0.950	6.834	7.740	0.657	0.634
$\beta = 0.2$	0.576	0.568	0.964	7.401	7.067	0.660	0.626
$\beta = 0.3$	0.459	0.569	1.030	17.805	7.105	0.510	0.647
$\beta = 0.5$	0.326	0.572	1.621	18.244	7.706	0.434	0.650

Table 9. SISS reproduction on Arc2Face with learning rate 5×10^{-6} for different values of β. The baseline corresponds to the original Arc2Face model before unlearning. We report identity-removal metrics (ISM and SRK) together with realism and recognizability metrics (KD and eDIFFIQA).

Figure 11. Qualitative results of the SISS reproduction on Arc2Face, with $\beta = 0.1$ and $lr = 5e^{-6}$. Row 1 shows baseline generations for the forget identity (ID 331), and Row 2 the corresponding generations after unlearning. Row 3 shows baseline generations for identities in the retain validation set, and Row 4 their generations after unlearning. Visual quality degrades substantially after fine-tuning: the effect is stronger for the forget identity, but remains severe for the retain identities as well.

egy rather than relying exclusively on fine-tuning. More broadly, UCE has become one of the standard references for adapting concept-removal methods to identity-related forgetting, as is also mentioned in GUIDE [37] and Forget-Me [31].

D.2.1 UCE Methodology

As discussed in Section 3.1, latent diffusion models incorporate conditioning information through cross-attention layers distributed throughout the U-Net. UCE [12] operates directly on these layers, and more specifically on the linear projections that map conditioning embeddings into keys and values. Since the conditioning signal enters the denoising network through these projections, modifying them provides a direct mechanism to alter the association between a conditioning concept and its generated visual attributes.

Following the notation of [12], let c_i denote the embedding of a conditioning concept. At a given cross-attention layer, the corresponding key and value vectors are obtained as

$$k_i = W_k c_i, \qquad v_i = W_v c_i, \tag{28}$$

where W_k and W_v are the key and value projection matrices, respectively. These interact with the query representation q, computed from the current latent features of the U-Net, to produce the cross-attention output

$$A \propto \mathrm{softmax}(q k_i^\top), \qquad O = A v_i. \tag{29}$$

Thus, the conditioning influences the latent representation only through these key and value projections. UCE is based on the idea that editing these projections, both W_k and W_v, is sufficient to suppress a target concept while preserving the behavior of the model on unrelated concepts.

To formalize this, the method defines a set E of concepts to edit, where each $c_i \in E$ is a conditioning embedding corresponding to a concept to be erased. For each edited concept, a guide (or anchor) concept c_i^* is specified, representing the desired target behavior after editing. The corresponding target output is then defined through the original projection as

$$v_i^* = W_v^{\mathrm{old}} c_i^*. \tag{30}$$

In the artist erasure setting considered in the original paper, c_i^* is not an identity-specific anchor, but rather a more generic target concept intended to remove the distinctive characteristics of the erased style. In addition, UCE introduces a set P of preservation concepts, whose outputs should remain unchanged after editing, in order to maintain generation quality on non-target concepts.

Given the edit set E, the target outputs $\{v_i^*\}$, and the preservation set P, UCE updates the projection matrix W by solving the following least-squares objective:

$$\min_W \sum_{c_i \in E} \|W c_i - v_i^*\|_2^2 + \sum_{c_j \in P} \|W c_j - W^{\mathrm{old}} c_j\|_2^2. \tag{31}$$

As shown in [12], this objective admits the closed-form solution

$$W = \left(\sum_{c_i \in E} v_i^* c_i^\top + \sum_{c_j \in P} W^{\text{old}} c_j c_j^\top \right) \times \left(\sum_{c_i \in E} c_i c_i^\top + \sum_{c_j \in P} c_j c_j^\top \right)^{-1} \tag{32}$$

This derivation is typically presented for the value projection, but the same closed-form update is applied to both W_v and W_k, yielding a direct parameter edit of the cross-attention mechanism without iterative fine-tuning.

D.2.2 Adapting UCE to Arc2Face

For our reproducibility study and comparison in Section 4.1, we adapt UCE to the Arc2Face identity-conditioned setting, taking inspiration from the artist erasure experiments in the original paper. Although artist style removal is not identical to identity unlearning, both settings share the same core objective: suppressing a specific target concept while preserving performance on the remaining ones.

The main difference lies in the conditioning space, which in Arc2Face is identity-based rather than textual. Accordingly, the edited concepts are ArcFace identity embeddings mapped to the conditioning representations used by Arc2Face. Our forget set E is built from the same identity data used in our main method (Section 3.4), and includes multiple conditioning embeddings for the target identity together with their linear combinations. Thus, E represents a local region of the identity manifold associated with the target subject, and UCE is applied to redirect this region toward a fixed target (anchor).

The target construction is also modified. In the original UCE setting, each edited concept is redirected toward a generic guide concept. In our adaptation, all edited embeddings are redirected toward a single anchor identity. More precisely, for each $c_i \in E$, we define the target concept c_i^* as the centroid of a selected anchor identity cluster, where the anchor is chosen following the procedure described in Section 3.3. Hence, unlike the original formulation, the target is not a generic semantic concept but a concrete identity-conditioned reference point.

We also adapt the construction of the preservation set P. In Appendix D of [12], the artist erasure experiments show that a relatively large preservation set helps balance erasure strength and image quality. However, since Arc2Face is specialized to facial identity, using hundreds of raw preserve embeddings is less well motivated in our setting. Instead, we construct P from one normalized centroid per selected retain identity, yielding a compact and diverse set of preservation directions that covers a broader range of identities while avoiding over-representation of subjects with many highly similar embeddings in the retain split.

Once the edit set E, the preservation set P, and the anchor target have been specified, the implementation of our method follows closely the official UCE code released by the authors. In particular, we retain the same closed-form editing paradigm over the cross-attention key and value projections, while adapting the construction of the edited and preserved conditioning representations to the Arc2Face setting. Concretely, we use a weighted version of Eq. (31), introducing coefficients $\alpha_e > 0$ and $\alpha_p > 0$ to control the relative strength of the unlearning and preservation terms, respectively, together with a regularization parameter $\lambda > 0$ that stabilizes the update and keeps the edited weights close to the original ones. This yields the following objective:

$$\begin{aligned} \min_W \quad & \alpha_e \sum_{c_i \in E} \|W c_i - v_i^*\|_2^2 \\ & + \alpha_p \sum_{c_j \in P} \|W c_j - W^{\text{old}} c_j\|_2^2 \\ & + \lambda \|W - W^{\text{old}}\|_F^2 \end{aligned} \tag{33}$$

The corresponding closed-form solution is

$$W = \left(\alpha_e \sum_{c_i \in E} v_i^* c_i^\top + \alpha_p \sum_{c_j \in P} W^{\text{old}} c_j c_j^\top + \lambda W^{\text{old}} \right) \times \left(\alpha_e \sum_{c_i \in E} c_i c_i^\top + \alpha_p \sum_{c_j \in P} c_j c_j^\top + \lambda I \right)^{-1} \tag{34}$$

which is the form implemented in our code for both W_v and W_k.

D.2.3 UCE Experiments

The general implementation details, including dataset construction, experimental environment, and evaluation metrics, follow Section 4. UCE does not require iterative fine-tuning, as the update is given by the closed-form solution in Eq. (34). As a result, the method is computationally lightweight in our setting: once the retain set P is precomputed, the unlearning edit for a single identity takes only a few seconds (at most $\sim$6 seconds in our experiments).

For consistency with the rest of our pipeline, we use the same forget and retain datasets as in Section 4.2. However, due to the large number of hyperparameter configurations explored, we restrict the study to 5 forget identities instead of the full set of 10, enabling a broader search while maintaining a representative evaluation.

The main hyperparameters of UCE in our setting are α_e, α_p, and λ in Eq. (33). Since the original paper and implementation do not provide clear recommended values, we

first performed preliminary exploratory experiments over small ranges to identify stable regimes and define the final grid search.

These preliminary experiments revealed three consistent trends. First, effective forgetting required substantially larger values of α_e than 1, since weaker edit weights were generally insufficient to produce a meaningful displacement toward the anchor identity. Second, α_p had to remain comparatively smaller: values around 1 provided the most reasonable compromise between preservation and erasure, whereas larger preservation weights significantly weakened forgetting. Third, the regularization parameter λ was most effective in the range $(0, 1)$, since stronger regularization made the update too conservative despite the relatively large values needed for α_e.

We also used these exploratory runs to compare two implementation choices: applying the closed-form edit only to the surgical cross-attention layers identified in Section 3.4, or to all cross-attention layers as in the original UCE formulation. We did not observe a substantial or systematic difference between the two variants. Therefore, in the final experiments we retained the all-layer version, which is both simpler and more faithful to the original method.

Based on these observations, we then performed a larger grid search over the most promising values:

$$\alpha_e \in \{5, 10, 20, 50, 100\},$$
$$\alpha_p \in \{0.1, 1, 5, 10\},$$
$$\lambda \in \{0.3, 0.5, 0.7\}.$$

All configurations were evaluated using the same unlearning and quality metrics employed throughout the paper, allowing us to identify the strongest UCE baseline.

D.2.4 UCE Results and Discussion

We begin by analyzing the full grid search described above. A first important observation is that all 240 evaluated configurations yield exactly the same SRK score, 0.9901, with $\mathrm{AccU} = 1$ and $\mathrm{AccR} = 1$ in every case. This indicates that, under SRK, the method preserves identity recognition on the unseen retain validation set while failing to perform effective identity removal. We attribute this behavior to the fact that UCE does not explicitly enforce identity replacement, but rather pushes the target concept toward a more generic alternative. In the Arc2Face setting, this is not sufficient to break identity recognition in the sense captured by SRK.

Since SRK is uninformative in this regime, we focus on the trade-off between forget ISM and retain unseen ISM. Given the large number of configurations, results are summarized using boxplots, and our analysis is based on the median behavior of each hyperparameter.

We first analyze the effect of the edit coefficient α_e, which directly controls the strength of unlearning. As shown in Figure 12, the lowest median forget-ISM values are obtained for $\alpha_e \in \{20, 50, 100\}$. Among these, $\alpha_e = 20$ achieves comparable median forgetting while exhibiting a slightly more stable distribution. To further distinguish between these candidates, we examine the retain unseen ISM in Figure 13. Here, $\alpha_e = 20$ shows a marginally higher median, indicating slightly better preservation of non-target identities. Based on this trade-off, we select $\alpha_e = 20$ as the most balanced choice.

We also observe that α_e has negligible impact on KD and eDIFFIQA, both on the forget and retain sets, and we therefore omit those plots for clarity.

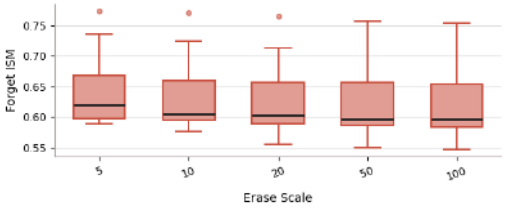

Figure 12. Forget ISM obtained across the grid search for different values of α_e. Each point corresponds to one experiment. Lower values indicate stronger identity removal.

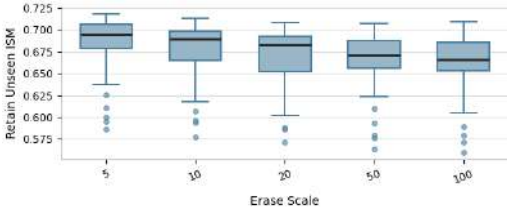

Figure 13. Retain unseen ISM obtained across the grid search for different values of α_e. Each point corresponds to one experiment. Higher values indicate better preservation of unseen retain identities.

We next study the preservation coefficient α_p, for which the behavior is more clearly differentiated. In Figure 14, the median forget ISM is clearly higher for $\alpha_p = 0.1$, and the distribution is also substantially wider, indicating greater variability across configurations. By contrast, for $\alpha_p \geq 1$, the forget ISM becomes both lower and more stable. A plausible interpretation is that, when α_p is too small, the retain identities provide too little structure for the closed-form update, so the edit is insufficiently constrained relative to the surrounding identity geometry. As a result, the displacement of the forget identity is weaker and less consistent, leading to higher forget ISM.

Figure 15 further helps distinguish among the remaining candidates. Although the lowest median retain unseen ISM is attained for $\alpha_p \in \{5, 10\}$, the value for $\alpha_p = 1$ remains very high and is accompanied by a noticeably tighter dis-

tribution. This suggests that $\alpha_p = 1$ achieves nearly the same level of preservation while being more stable across configurations. Taken together, these trends indicate that $\alpha_p = 1$ provides the best trade-off between effective editing and preservation.

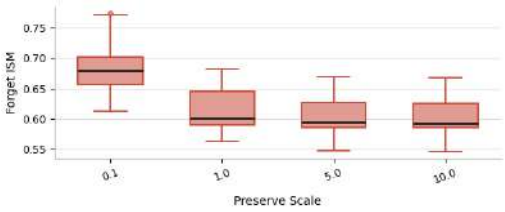

Figure 14. Forget ISM obtained across the grid search for different values of α_p. Each point corresponds to one experiment.

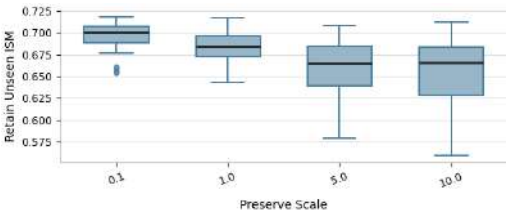

Figure 15. Retain unseen ISM obtained across the grid search for different values of α_p. Each point corresponds to one experiment.

Finally, we consider the regularization coefficient λ. In this case, the effect on ISM is comparatively limited, so the most informative differences appear in the generation-quality metrics. In particular, Figures 16 and 17 show a slight but consistent advantage for $\lambda = 0.7$, which tends to yield lower (Δ) KD values and therefore somewhat better realism. At the same time, this value does not noticeably harm the identity-related metrics, nor does it reduce eDIFFIQA in a problematic way. More broadly, one should note that UCE in our setting is not a particularly aggressive method in terms of breaking facial recognizability: eDIFFIQA remains relatively high across the grid search, which is consistent with the qualitative observation that the generated faces remain recognizable even when the identity shift is limited. Based on these trends, we select $\lambda = 0.7$ for the final reported configuration.

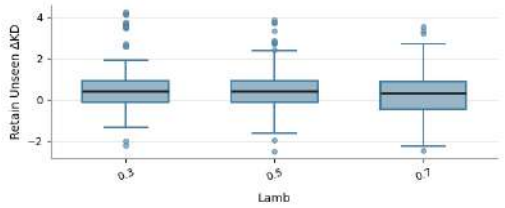

Figure 16. Retain unseen KD obtained across the grid search for different values of λ. Lower values indicate better realism.

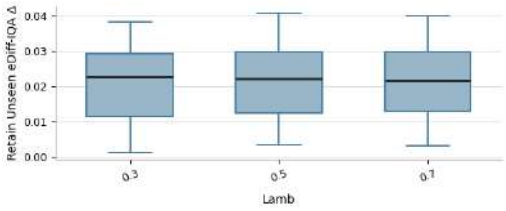

Figure 17. Retain unseen eDIFFIQA obtained across the grid search for different values of λ. Higher values indicate better facial quality and recognizability.

Combining the observations above, the best hyperparameter setting is

$$\alpha_e = 20, \qquad \alpha_p = 1, \qquad \lambda = 0.7.$$

Table 10 shows the corresponding results, averaged across the forget identities, and compares them with the baseline Arc2Face model. As already suggested by the grid search, UCE slightly lowers forget ISM while keeping good visual quality, but it does not achieve meaningful unlearning under the SRK metric, which stays almost unchanged. This suggests that, in our setting, the closed-form edit of UCE is too weak to produce strong identity removal.

A qualitative example with this configuration is shown in Figure 18. Some visible changes can be seen on the forget identity, and the generations sometimes move slightly toward the anchor. However, these changes are not strong or consistent enough to modify the main identity traits of the subject. Instead, the edit mostly affects more superficial visual aspects, while the main identity remain. This matches the quantitative results and supports the conclusion that, although UCE is efficient and can produce visible edits, its closed-form strategy is not strong enough for robust identity unlearning in Arc2Face.

D.3. Unlearn and Protect With ID-guidance

This baseline is particularly important in our setting because, to the best of our knowledge, the Unlearn and Protect variant With ID-guidance (WID) [39], together with BIA [38], is one of the very few works that address identity unlearning in diffusion models. Among them, WID is the only prior method specifically formulated for an identity-conditioned diffusion model, which makes it the most direct point of comparison for our approach. At the same time, WID [39] is difficult to reproduce in practice. The original paper introduces a custom identity-conditioned diffusion model based on CosFace identity embeddings, but neither model code nor pretrained weights are publicly available. Although the training losses and some implementation details are described, the exact construction of the diffusion model and conditioning pipeline is not fully specified. In addition, no official implementation of the unlearning procedure is released, and only limited fine-tuning details are

Method	Forget ISM ↓	Retain Unseen ISM ↑	SRK ↑	Δ Forget KD ↓	Δ Retain Unseen KD ↓	Forget eDIFFIQA ↑	Retain Unseen eDIFFIQA ↑
Baseline	0.775	0.726	0.990	–	–	0.763	0.770
UCE ($\alpha_e = 20, \alpha_p = 1, \lambda = 0.7$)	0.621	0.694	0.990	4.195	1.110	0.749	0.768

Table 10. UCE reproduction on Arc2Face using the best hyperparameter configuration found in our grid search, $\alpha_e = 20$, $\alpha_p = 1$, and $\lambda = 0.7$. The baseline corresponds to the original Arc2Face model before editing. We report identity-removal metrics (ISM and SRK) together with realism and recognizability metrics (KD and eDIFFIQA).

Figure 18. Qualitative results of the UCE reproduction on Arc2Face, using the best configuration found in our grid search: $\alpha_e = 20$, $\alpha_p = 1$, and $\lambda = 0.7$. Rows 1–2 show baseline and edited generations for the forget identity, while Rows 3–4 show the corresponding generations for identities in the unseen retain validation set. The edit induces visible changes, especially on the forget identity, but these are not sufficient to alter its defining attributes in a consistent way, which agrees with the quantitative SRK results.

reported. In this section, we therefore present our reconstruction of the method, adapted to Arc2Face and integrated into our experimental framework.

D.3.1 WID Methodology

WID first builds its own identity-conditioned diffusion model, trained on the CASIA-WebFace dataset, where conditioning is provided by identity embeddings extracted with a pretrained CosFace model. The model is trained with a combination of the standard diffusion denoising objective and an additional identity loss, introduced to preserve facial identity information during reconstruction. This identity-conditioned backbone is then fine-tuned for unlearning.

Following the standard DDPM formulation already introduced in Section 3.1, the reverse process defines a conditional generative distribution

$$p_\theta(x_{0:T} \mid c) = p(x_T) \prod_{t=1}^{T} p_\theta(x_{t-1} \mid x_t, c),$$

where here the conditioning variable c is an identity embedding obtained from CosFace. Given a clean image x_0, the corresponding noisy sample at timestep t is generated through the usual forward process as $x_t = \sqrt{\bar{\alpha}_t}\, x_0 + \sqrt{1-\bar{\alpha}_t}\, \epsilon$, with $\epsilon \sim \mathcal{N}(0, I)$. The main idea of the method is to remove a forget identity c_f by aligning the conditional distribution of the fine-tuned model under c_f with that of a frozen pretrained model under a target or anchor identity c_a. Formally, this is expressed through the KL minimization objective

$$\min_\theta \ \mathrm{KL}\Big(p_{\theta^*}(x_{0:T} \mid c_a) \,\|\, p_\theta(x_{0:T} \mid c_f)\Big), \tag{35}$$

where θ^* denotes the frozen pretrained parameters and θ the fine-tuned ones.

As discussed in the original paper, directly optimizing this KL divergence is computationally challenging, since it would require handling full diffusion trajectories. For this reason, the authors derive an equivalent objective at the denoising level using the standard noise-prediction parameterization $\epsilon_\theta(x_t, t, c)$. In our notation, their main unlearning loss can be written as

$$\mathcal{L}_{\text{model}} = \mathbb{E}_{x_0,\, \epsilon,\, t}\left[\|\epsilon_{\theta^*}(x_t, t, c_a) - \epsilon_\theta(x_t, t, c_f)\|_2^2\right], \tag{36}$$

which enforces that the fine-tuned model, when queried with the forget identity c_f, matches the denoising prediction of the frozen model under the anchor identity c_a.

In addition, the method incorporates a second objective, called *identity loss*, to encourage identity consistency in the reconstructed samples. This loss is computed in the CosFace embedding space and is defined as

$$\mathcal{L}_{\text{id}} = 1 - \cos\big(\mathrm{CosFace}(x_0), \mathrm{CosFace}(\hat{x}_0)\big), \tag{37}$$

where $\hat{x}_0$ denotes the reconstructed image predicted by the denoising model. This term operates at the image level and encourages the generated sample to remain aligned with the desired identity representation. The final training objective is then given by

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{model}} + \lambda \mathcal{L}_{\text{id}}, \tag{38}$$

where λ controls the trade-off between the anchor-guided replacement objective and identity consistency in the reconstructed outputs. For clarity, we refer to this method as *WID* (with-identity-loss), following the terminology used in Table 1 of the original paper, where the variant incorporating the identity loss is denoted in this way.

D.3.2 Adapting WID to Arc2Face

To adapt this method to our setting, we replace the original identity-conditioned backbone with Arc2Face. As a result, identity conditioning is no longer based on CosFace embeddings, but on ArcFace embeddings.

We follow the same anchor-guided principle as in Eq. (36), but reformulate it in the PIU framework. Given a forget identity condition c_f, an anchor identity condition c_a, and a noisy latent z_t, the unlearning objective at the noise-prediction level is

$$\hat{\epsilon}_{\text{forget}} = \epsilon_{\theta^*}(z_t, t, c_a) - \eta\Big[\epsilon_{\theta^*}(z_t, t, c_f) - \epsilon_{\theta^*}(z_t, t, c_a)\Big]. \tag{39}$$

Equivalently, this can be rewritten as

$$\hat{\epsilon}_{\text{forget}} = (1+\eta)\,\epsilon_{\theta^*}(z_t, t, c_a) - \eta\,\epsilon_{\theta^*}(z_t, t, c_f), \tag{40}$$

and therefore the forget objective is

$$\mathcal{L}_{\text{forget}} = \mathbb{E}_{z,\epsilon,t}\left[\|\epsilon_\theta(z_t, t, c_f) - \hat{\epsilon}_{\text{forget}}\|_2^2\right]. \tag{41}$$

If we now fix $\eta = 0$, then

$$\hat{\epsilon}_{\text{forget}} = \epsilon_{\theta^*}(z_t, t, c_a), \tag{42}$$

so that the forget loss reduces to

$$\mathcal{L}_{\text{forget}} = \mathbb{E}_{z,\epsilon,t}\left[\|\epsilon_\theta(z_t, t, c_f) - \epsilon_{\theta^*}(z_t, t, c_a)\|_2^2\right]. \tag{43}$$

Hence, the model-alignment loss $\mathcal{L}_{\text{model}}$ used in WID (Eq. (36)), corresponds exactly to the $\eta = 0$ special case of our forget loss. In this sense, NIL (the No-Identity-Loss version of [39]) can be interpreted as a restricted instance of our anchor-guided formulation in which the forget trajectory is matched directly to the frozen prediction under the anchor identity, without the additional repulsive guidance induced by $\eta > 0$.

The second term in WID is an identity-consistency loss that acts in face-recognition embedding space rather than in diffusion space. In our Arc2Face adaptation, we keep the same idea and define

$$\mathcal{L}_{\text{id}} = 1 - \cos\big(s_{\text{id}}, \hat{s}_{\text{id}}\big), \tag{44}$$

where $s_{\text{id}} = \text{ArcFace}(x_0)$ denotes the identity embedding of the original clean image, and $\hat{s}_{\text{id}} = \text{ArcFace}(\hat{x}_0)$ the identity embedding of its reconstruction predicted by the current denoising model. Since the original paper does not specify in detail how $\hat{x}_0$ is obtained, in our implementation we recover it from the current noise prediction using the standard DDPM relation

$$\hat{x}_0 = \frac{x_t - \sqrt{1-\bar{\alpha}_t}\,\epsilon_\theta(x_t, t, c_f)}{\sqrt{\bar{\alpha}_t}}. \tag{45}$$

In Arc2Face, the same relation is applied in latent space, using the noisy latent z_t and the corresponding predicted noise. The resulting estimate is then decoded to image space, and its ArcFace embedding is used to compute $\hat{s}_{\text{id}}$.

Although Eq. (44) is the conceptual form described in the paper, Algorithm 1 in [39] suggests an equivalent implementation in terms of squared Euclidean distance between identity embeddings, namely

$$\mathcal{L}_{\text{id}} = \big\|s_{\text{id}} - \hat{s}_{\text{id}}\big\|_2^2.$$

A natural explanation is that, even though this is not stated explicitly, the two forms are equivalent (proportionally) whenever the identity embeddings are ℓ_2-normalized. Indeed, if $\|u\|_2 = \|v\|_2 = 1$, then

$$\begin{aligned}\|u - v\|_2^2 &= (u-v)^\top(u-v)\\ &= \|u\|_2^2 + \|v\|_2^2 - 2u^\top v\\ &= 2 - 2u^\top v\\ &= 2\big(1 - \cos(u, v)\big).\end{aligned}$$

Therefore, for normalized embeddings,

$$1 - \cos(u, v) = \frac{1}{2}\|u - v\|_2^2.$$

Hence, both objectives have exactly the same minimizers, since they differ only by a positive constant factor. As in our discussion of the SISS reproduction, such a constant scaling does not change the optimization target, and can be absorbed into the coefficient that controls the strength of the term. Defining λ as the effective weight of the identity-consistency loss, we may therefore write the objective in the equivalent form

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{forget}} + \lambda\big\|s_{\text{id}} - \hat{s}_{\text{id}}\big\|_2^2, \qquad \text{with } \eta = 0. \tag{46}$$

Overall, this preserves the original spirit of WID while making it compatible with Arc2Face. Since the forget term coincides exactly with the $\eta = 0$ case of our method, we use the same forget and anchor identities introduced in Section 3.4. Moreover, because WID does not include a dedicated study of anchor selection, we keep our main anchor-selection strategy, which remains well motivated by the ArcFace embedding-space geometry discussed in Section 3.3.

LR	λ	Forget ISM ↓	Retain ISM ↑	SRK ↑	ΔKD Forget ↓	ΔKD Retain ↓
5×10^{-6}	0.1	0.322	0.443	41.0	4.932	4.252
5×10^{-6}	0.3	0.316	0.446	42.0	5.545	4.766
5×10^{-6}	0.5	0.313	0.447	43.0	5.104	4.670
10^{-5}	0.1	0.297	0.396	31.0	4.390	5.033
10^{-5}	0.3	0.294	0.400	34.0	3.984	5.323
10^{-5}	0.5	0.290	0.396	32.0	3.777	4.837

Table 11. Results of the WID adaptation after 100 optimization steps for tested combinations of learning rate and identity-loss weight λ.

D.3.3 WID Experiments

The general setup follows Section 4. A key difference from our method is that this adaptation requires real image latents to evaluate the identity-consistency term $\mathcal{L}_{\text{id}}$. In particular, the loss 44 is computed from a reconstruction $\hat{x}_0$, whose ArcFace embedding is compared against that of the original image. To obtain $\hat{x}_0$, we use Eq. (45) with the standard DDPM scheduler employed in Arc2Face.

This makes the method substantially more expensive than the other baselines reproduced in our study, since each optimization step additionally involves reconstructing $\hat{x}_0$, decoding it to image space, and extracting ArcFace features. In our implementation, this results in training that is approximately seven times slower than PIU, which considerably limits the feasible hyperparameter budget.

For this reason, we restrict the exploration to a small grid over the learning rate and the identity-loss weight λ. The original paper reports a learning rate of 10^{-4}, but in our Arc2Face adaptation we found this value to be unstable, leading to a clear degradation of the model. We therefore evaluate the smaller learning rates 5×10^{-6} and 10^{-5}. The former is taken as a conservative setting, while the latter is included to test whether a stronger optimization regime can compensate for the reduced training budget without destabilizing training. For the identity-consistency weight, we consider $\lambda \in \{0.1, 0.3, 0.5\}$. This includes the default value $\lambda = 0.3$ used in the original work, while also allowing us to assess the sensitivity of the method to weaker or stronger identity preservation.

Finally, while the original paper recommends between 1000 and 2000 optimization steps, such a budget is infeasible here. We therefore train for 300 steps and additionally evaluate checkpoints at steps 100 and 200, in order to assess more carefully how much optimization is actually needed in Arc2Face. The remaining settings are kept fixed: batch size 64, AdamW optimizer, and weight decay 0.

D.3.4 WID Results and Discussion

We first study the effect of the number of optimization steps, comparing 100, 200, and 300 updates. We observe a clear and consistent pattern: increasing the number of steps degrades both unlearning and generation quality. This trend is already evident from SRK, which is the main unlearning metric used throughout our analysis. For the learning rate 5×10^{-6}, the best SRK value across the tested λ values is 43 at step 100, 38 at step 200, and 31 at step 300. A similar behavior is found for the learning rate 10^{-5}, where the corresponding best SRK values are 34, 28, and 31. The same tendency also appears in the other metrics, so we do not report the full step-by-step comparison here. Overall, these results provide sufficient evidence to select 100 steps as the best training horizon for this adaptation.

We therefore focus the main comparison on the 100-step checkpoints. Table 11 reports the main metrics of interest for all tested combinations of learning rate and identity-loss weight λ.

From Table 11, a clear pattern emerges regarding the learning rate. Although 10^{-5} produces a somewhat stronger effect on forget ISM, this does not translate into better SRK, because retain identities deteriorate much more quickly. In contrast, the smaller learning rate 5×10^{-6} is less aggressive and preserves the rest of the identity space more effectively, which leads to better overall SRK. The same trade-off is also visible in the quality metrics: while 10^{-5} can slightly improve ΔKD on the forget set, this gain is offset by a worse ΔKD on the retain set. For this reason, we prefer 5×10^{-6}, which achieves a better balance between forget strength and preservation of non-target identities.

The influence of λ is comparatively weaker on the unlearning metrics. Table 11 reports also the forget and retain ISM values across the tested λ values for the preferred learning rate 5×10^{-6}. While the differences in ISM are relatively small, the quality metrics show a clearer dependence on λ. In particular, $\lambda = 0.1$ yields substantially smaller distribution shifts in both forget and retain ΔKD, indicating better realism and overall post-unlearning quality. We therefore select $\lambda = 0.1$ as the best operating point.

Based on this analysis, our final selected configuration is $\lambda = 0.1$, learning rate 5×10^{-6}, and 100 optimization steps. The final comparison against the baseline model is reported in Table 12.

Overall, the results are promising but also reveal an important limitation of the method. On the positive side, the forget ISM decreases noticeably, indicating that the target identity is effectively altered, while image quality remains reasonably high. However, the main weakness of the method lies on the retain side. Since WID does not in-

Method	Forget ISM ↓	Retain Unseen ISM ↑	SRK ↑	Δ Forget KD ↓	Δ Retain Unseen KD ↓	Forget eDIFFIQA ↑	Retain Unseen eDIFFIQA ↑
Baseline	0.775	0.726	0.99	–	–	0.763	0.770
WID	0.322	0.443	41.0	4.932	4.252	0.732	0.749

Table 12. WID reproduction on Arc2Face using the best hyperparameter configuration found in our grid search, lr = $5e^{-6}$, $\lambda = 0.1$, and steps = 100. The baseline corresponds to the original Arc2Face model before editing. We report identity-removal metrics (ISM and SRK) together with realism and recognizability metrics (KD and eDIFFIQA).

clude an explicit preservation term over non-target identities, it cannot maintain them as effectively as our method. As a result, retain metrics deteriorate and the overall identity preservation of the model is reduced.

This behavior is also visible qualitatively in Figure 19. On the forget side, generations remain visually plausible and high-quality, and the target identity no longer appears to correspond to the original person. The more problematic effect appears on the retain side: although the images do not collapse in quality as severely as in some other baselines, the generated faces are no longer as faithful to their original identities. This suggests that the update does not only move the forget identity away from its original region, but also shifts part of the surrounding identity distribution toward the anchor subspace.

Figure 19. Qualitative results of the WID adaptation on Arc2Face using the selected configuration: learning rate 5×10^{-6}, $\lambda = 0.1$, and 100 optimization steps. Row 1 shows baseline generations for the forget identity, and Row 2 the corresponding generations after unlearning. Row 3 shows baseline generations for identities in the unseen retain validation set, and Row 4 their generations after unlearning. The method effectively alters the forget identity while largely preserving visual quality, but it also changes non-target identities, which is consistent with the drop observed in the retain metrics.