IREU: Identity-Related Encoder-Only Unlearning for Customized Portrait Generation

Chaoyi Shi[1], Shanshan Zhang[1,2(✉)], and Jian Yang[1,2]

[1] PCA Lab, School of Computer Science and Engineering, Nanjing University of Science and Technology
[2] PCA Lab, School of Intelligence Science and Technology, Nanjing University
{chaoyishi,shanshan.zhang,csjyang}@njust.edu.cn

Abstract. Customized Portrait Generation (CPG) technologies have been widely used to generate high-fidelity person images given an input image indicating the identity and a text prompt indicating the required edits. Yet these methods pose significant privacy risks by spreading fake visual information. Against such risks, each public generator should be able to suppress its generation ability for a particular person when requested. Therefore, in this work we investigate the identity unlearning problem for CPG. Since there are no previous methods in this field, we propose a simple baseline that updates the image encoder by minimizing identity similarity between generated and input images for target identities to be unlearned, while maximizing it for identities to be retained. However, we find such a global perturbation in the feature space harms the fidelity of generated images for other identities to be retained. To solve this problem, we propose a novel method IREU, which first locates identity-related features in an offline manner and then only performs feature perturbations on them. The experimental results show that our proposed method IREU achieves better identity unlearning performance for target identities to be unlearned, and also keeps high fidelity for other identities to be retained. In addition, our unlearned image encoder is generalizable across different generators with the same encoder without fine-tuning, which is friendly for deployment in practice.

Keywords: Machine unlearning · Customized portrait generation · Identity erasure

1 Introduction

In recent years, diffusion-based text-to-image models (e.g., Stable Diffusion, SDXL) have substantially advanced Customized Portrait Generation (CPG) [25, 45]. With a few images or short prompts, these models synthesize high-fidelity, editable portraits and support applications such as attribute editing, identity blending, and scenario recontextualization (e.g., personalized avatars, character animation, virtual try-on [19, 37, 47]). However, the same convenience poses significant privacy risks, especially the unintended reproduction of real-world identities. For example, adversaries can exploit this behavior for impersonation, deepfake abuse, and loss of credibility [15, 16, 40].

Fig. 1: Illustration of our identity unlearning paradigm. Left: a typical CPG pipeline, making edits based on the text prompt while keeping the same identity as the input. Right: only the image encoder is updated (w/ the diffusion model frozen) during unlearning, so that the generated images exhibit a different identity from the input.

To protect privacy, some methods primarily focus on image-level protection, such as injecting imperceptible adversarial perturbations into user images to disrupt unauthorized usage [2, 35]. Although this image-centric protection can achieve privacy preservation at a low cost, this is only feasible provided that all images of that identity are protected, which is clearly impractical, as malicious users can obtain unprotected images such as already circulated images and covert photography. Moreover, according to the EU's General Data Protection Regulation (GDPR) [38], service providers (model owners) are obliged to implement erasure upon user requests. Therefore, to ensure robust privacy protection and regulatory compliance, the defense paradigm must shift from perturbing images to perturbing models, with model-based generative unlearning effectively addressing this requirement.

Although model-level unlearning provides a more robust solution, achieving identity unlearning for CPG remains highly challenging. Existing generative identity unlearning methods [21, 33] typically fine-tune the denoising network. However, the U-Net contains extensive generative prior knowledge, and directly erasing identity information within this high-dimensional parameter space can easily trigger catastrophic forgetting [28], thereby degrading the overall fidelity of generated images. Furthermore, recent CPG pipelines (e.g., PhotoMaker [25], FastComposer [45]) consist of a pre-trained encoder and a diffusion model. Unlearning on the U-Net limits its generalization ability across generators that usually share the same encoder family but differ in denoisers. In contrast, an encoder-only design is more friendly to deployment, i.e., once the encoder is optimized towards identity unlearning, it can be plugged into an arbitrary generator that shares the same encoder. Thus, as illustrated in Fig. 1, in this work we propose novel unlearning methods for CPG by modifying the encoder only.

Following this idea, we first establish a simple baseline, where the encoder is optimized by minimizing the identity similarity between each target identity to be unlearned and its corresponding customized generated image, and maximizing this identity similarity for other identities. However, although the target identities can be well unlearned, this kind of global perturbation in the feature space is harmful for keeping high fidelity for other identities to be retained, leading to noticeable quality degradation, as shown in Fig. 2. To address this problem, we propose to perform perturbation only on identity-related features, so as to

Fig. 2: The major limitation of the baseline method using global feature perturbation. After unlearning on target identities, the generated images for other identities to be retained are of lower fidelity, showing some noticeable discrepancies indicated by red boxes. In contrast, our method IREU addresses this problem by introducing identity-related feature perturbation.

suppress the target identity in the generated images while largely alleviating the negative effect on fidelity for other identities in the meantime.

Building on this insight, we propose a novel **I**dentity-**R**elated **E**ncoder-Only **U**nlearning (**IREU**) framework, which performs all updates in the image feature space while keeping the diffusion model frozen. To isolate identity-related features, we first build Face-Swap pairs that differ in identity and compute their embedding differences. Dimensions with the largest deltas with respect to identity are selected as identity-related features. We then apply localized perturbations along these dimensions on the inputs to be unlearned, shifting their representations away from the original identity. For other identities to be retained, we follow the baseline's practice and keep consistency between the unlearned and original encoder to keep features unchanged.

Our contributions are summarized as follows:

- To the best of our knowledge, we are the first to address the identity unlearning problem specifically for CPG by updating the image encoder only, making it generalize well across different generators.
- In order to obtain high identity dissimilarity for target IDs to be unlearned and to preserve high fidelity for other IDs to be retained, we propose to locate identity-related features in the embedding space via Face-Swap and then apply localized perturbations via identity-related transformation.
- Extensive experiments show that our method outperforms recent identity unlearning approach and achieves a better trade-off between unlearning and retaining. The unlearned encoder is plug-and-play across diverse generators and preserves fidelity without additional training.

2 Related Work

2.1 Customized Portrait Generation and Privacy Issues

Text-to-image diffusion models have rapidly advanced the landscape of Customized Portrait Generation (CPG). Early representative methods [6, 31, 39] can reproduce a subject's appearance from only a few input images, yet typically incur long finetuning time. Subsequent parameter-efficient adaptations [23,25,41,45] further accelerate personalization and improve identity consistency.

At the same time, the high fidelity and accessibility of CPG heighten privacy risks because outputs are tightly coupled with real humans. Existing protections mainly act on the input–output boundary: provenance methods inject robust watermarks for attribution [5, 42], input-side defenses add imperceptible perturbations to hinder unauthorized editing [22, 46], and privacy-friendly generation schemes reduce identifiability via domain constraints [34]. While useful, these approaches do not selectively remove the model's internal capacity to regenerate a specific identity. We instead pursue a model-internal method: IREU converts a Face-Swap identity direction into a masked, identity-related perturbation while keeping the diffusion model frozen, thereby suppressing the target identity, preserving fidelity, and transferring across CPG pipelines without extra training.

2.2 Unlearning in Diffusion Models

Machine unlearning was first established in image classification [1,3,18,24,36,50, 51], aiming to remove the influence of targeted data with minimal model changes [24]. As demands for controllability grew, attention shifted to representation-level editing in generative models, exploring concept erasure and generative unlearning as internal mechanisms for selective removal [4,20]. In diffusion models, early concept erasure typically involves small-scale finetuning of the denoising network to suppress specified concepts [7], whereas closed-form parameter editing adjusts the model's parameters through finetuning to enable efficient cross-attention edits without full retraining [8, 27]. Machine unlearning has been applied not only to remove specific styles or unsafe content [10, 11, 43, 44, 48, 49], but also to selectively suppress fine-grained attributes such as hairstyles and hats [12, 29].

Closer to our setting is identity-level generative unlearning. GUIDE [32] removes an identity's generative capability from a pre-trained GAN given a single image, while BIA [33] builds a black-hole-like absorption in the diffusion UNet bottleneck to absorb identity representations. However, GUIDE is tied to GANs whereas recent CPG pipelines are predominantly diffusion-based, and BIA's UNet-bottleneck editing tightly couples unlearning to a specific generator, making transfer across CPG costly. APDM [21] adjusts model parameters specifically for fine-tuning-based CPG methods (e.g., DreamBooth), which results in significant time consumption and renders it inapplicable to current mainstream CPG approaches. In contrast, we perform unlearning purely in the encoder embedding space while freezing the diffusion backbone, enabling training-free transfer across different CPG pipelines that share the same encoder.

3 Method

3.1 Preliminaries

CPG pipeline. The state-of-the-art customized portrait generators follow a standard text-to-image diffusion model that synthesizes personalized portraits by conditioning on a customized text prompt and one or more input images. Both

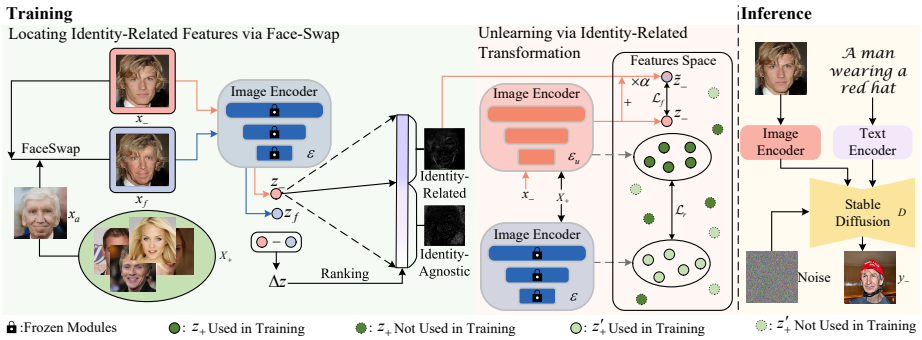


Fig. 3: Overview of our proposed method IREU, which consists of two key components: (1) **Locating Identity-Related Features:** given an image x_- to be unlearned and its Face-Swap counterpart x_a, we encode them with the original image encoder ε to obtain their embeddings. Then we compute the difference Δz, and rank $|\Delta z|$ to build a binary mask $\mathcal{M}$ that selects identity-related dimensions for feature perturbation. (2) **Unlearning via Identity-Related Transformation:** we construct a virtual identity $\tilde{z}_-$ with a magnitude α, and update the unlearning encoder ε_u through an unlearning loss $\mathcal{L}_f$ to minimize the distance between z_- and $\tilde{z}_-$. For other identities to be retained, we employ a retaining loss $\mathcal{L}_r$ to minimize the output discrepancy between ε_u and ε. Notably, $\mathcal{L}_r$ is computed using a randomly sampled image to be retained during each iteration. During inference, the unlearned encoder suppresses the target identity to be unlearned while maintaining the fidelity of other identities to be retained.

image and text encoders are inherited from CLIP [30]. The text prompt provides high-level semantics and intent, and its CLIP text embedding modulates the denoising network via cross-attention. The input images supply person-specific appearance cues, and their CLIP image embeddings anchor the generated face to the original identity while keeping pose and background editable. Given an input image x, we obtain its feature vector z through the image encoder ε, i.e., $z = \varepsilon(x)$. A conditional Stable Diffusion model D then generates $y = D(z, t)$ based on the image embedding z and text prompt t.

Identity Unlearning in CPG. Let $\mathcal{I}$ be the set of all identities. We split it into the unlearning subset $\mathcal{I}_-$ (target identities to be unlearned) and the retaining subset $\mathcal{I}_+ = \mathcal{I} \setminus \mathcal{I}_-$ (other identities to be retained). For each identity $i \in \mathcal{I}$, let $X(i) = \{x_i^j\}_{j=1}^{n_i}$ be its image set with $n_i \geq 1$. The entire image dataset is noted as $X = X_- \cup X_+$ with $X_- \cap X_+ = \varnothing$, where $X_- = \bigcup_{i \in \mathcal{I}_-} X(i)$ and $X_+ = \bigcup_{i \in \mathcal{I}_+} X(i)$. In the context of CPG, the goal is to ensure that for an image $x_-^j \in X_-$, the image embedding $z_-^j = \varepsilon_u(x_-^j)$ produced by the post-unlearning image encoder ε_u leads to a generated image $y_-^j = D(z_-^j, t)$ that no longer retains the identity characteristics consistent with x_-^j.

For other identities to be retained, we randomly sample an image $x_+^j \in X_+$, and the post-unlearning image encoder ε_u should make sure that the generated image $y_+^j = D(\varepsilon_u(x_+^j), t)$ retains the same identity as x_+^j.

For target identities to be unlearned, prior work typically provides the model with one input image per identity [32, 33]. We follow this one-shot setting with $n_i = 1$ and, for clarity, omit superscripts in x_-^j and x_+^j, writing them simply as x_- and x_+ in what follows.

3.2 Overview

Since there are no previous methods designed for identity unlearning in CPG, we first establish a *simple baseline* method. The image encoder is updated with two objectives: for an image from any target identity to be unlearned (x_-), the ℓ_2 distance between the outputs of the post-unlearning encoder ε_u and the original encoder ε is maximized, pushing its representations away from the original features; for an arbitrary image to be retained $x_+ \in X_+$, the ℓ_2 distance between the outputs of the post-unlearning encoder ε_u and the original encoder ε is minimized, keeping the representations unchanged. In this way, the entire feature space for each image is jointly perturbed during the optimization process.

Since global feature perturbation hurts the fidelity of generated images, we propose to perform the perturbation only on identity-related features instead of the entire feature vector. To achieve this goal, we first need to locate those identity-related features. Specifically, we construct Face-Swap pairs in which only the identity differs between the two images, and we treat the elements with the largest differences as identity-related features.

After that, we perturb identity-related features for the input images to be unlearned so as to change the identity in the feature space; while for other identities to be retained, we minimize the features from the unlearned and original encoders, just as the baseline does.

3.3 Locating Identity-Related Features via Face-Swap

To localize identity-related features in the embedding space, we use Face-Swap [9] to synthesize a surrogate face for the target identity to be unlearned and compute the latent difference between the original and the swapped image. We then decouple the entire feature vector into identity-related and identity-agnostic components, and only perform perturbation on the former group. Specifically, as shown in Fig. 3, for an image from a target identity to be unlearned x_-, we obtain its swapped counterpart $x_f = F(x_-, x_a)$, where x_a is randomly sampled from the other identity image set X_+, and $F(\cdot)$ is the Face-Swap function. This function generates an image that maintains visually consistent context with x_- but represents the identity of x_a. We treat $F(\cdot)$ as an off-the-shelf, non-differentiable operator used only to construct Face-Swap pairs, where gradients do not flow through F.

After the identity shift between x_- and x_f, we compute the embedding difference as follows:

$$\Delta z = z_- - z_f, \tag{1}$$

where $z_- = \varepsilon(x_-)$, $z_f = \varepsilon(x_f)$ and the vector Δz captures the change in embedding induced by identity shift.

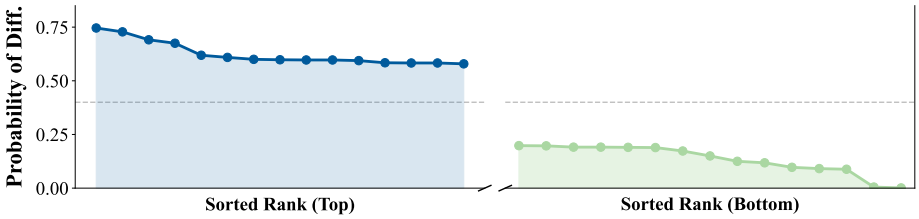


Fig. 4: Distribution of difference features across 1,000 Face-Swap pairs. The non-uniform distribution exhibits a clear bias towards specific dimensions, which motivates our dimensional masking strategy to isolate identity-related features.

In the difference vector Δz, the absolute value of each component reflects how strongly that dimension responds to the identity change: a larger $|\Delta z_i|$ implies a higher identity sensitivity. Additionally, we conduct statistics on 1,000 Face-Swap image pairs and calculate the occurrence probability of the top 40% of $|\Delta z_i|$ values. As shown in Fig. 4, the differences are not uniformly distributed across all dimensions but exhibit a clear bias. This observation motivates us to employ a dimensional masking strategy from a new perspective to locate identity-related features. We therefore sort all dimensions in descending order by $|\Delta z_i|$ and select the top dimension entries as identity-related, while regarding the remaining low-magnitude dimensions as approximately identity-agnostic.

We set a quantitative threshold k to construct a mask $\mathcal{M} \in \{0,1\}^d$, which will be used to isolate identity-related features while suppressing identity-agnostic ones:

$$\mathcal{M} = \mathbb{I}(\text{rank}\left(|\Delta z|_{[i]}\right) \leq \lfloor k \cdot d \rfloor), \tag{2}$$

where $\text{rank}(|\Delta z|_{[i]})$ denotes the descending order index of the i-th entry when sorting the absolute values $|\Delta z|$ across all d dimensions, so that both positive and negative identity shifts are treated equally; $k \in (0,1]$ is the kept ratio; d is the length of the embedding vector; $\mathbb{I}(\cdot)$ is the indicator function that returns 1 if the condition is true and 0 otherwise.

3.4 Unlearning via Identity-Related Transformation

Leveraging the identity-related features selected by the mask $\mathcal{M}$, we only modify the identity-related elements while leaving identity-agnostic features unchanged. This ensures the protection of retained identities and the preservation of the original distribution. This module produces the virtual identity $\tilde{z}_-$ as follows:

$$\tilde{z}_- = z_- + \alpha(\mathcal{M} \odot \Delta z), \tag{3}$$

where $\odot$ denotes element-wise multiplication, and α is the magnitude coefficient. Compared to the baseline, which maximizes the distance to the original encoder along all dimensions, this construction restricts the shift of z_- to identity-related dimensions and avoids unnecessary changes to identity-agnostic features. We

Table 1: Comparison of different methods on FastComposer [45] w.r.t. identity-centric metrics under both single-identity (oneID) and multi-identity (fiveID) settings. We re-implement BIA [33] based on Stable Diffusion and plug it in the same generator for a fair comparison. Best results are in **bold**.

ID Type	Methods	Target ID (↓)				Other ID (↑)		ΔID(↑)
		$\text{ID}^{o}_{\text{target}}$	$\text{ID}^{o,\text{unseen}}_{\text{target}}$	$\text{ID}^{\text{human}}_{\text{target}}$	$\text{ID}^{g}_{\text{target}}$	$\text{ID}^{o}_{\text{other}}$	$\text{ID}^{g}_{\text{other}}$	
oneID	BIA [33] [CVPR 25']	0.05	0.18	27.92%	0.30	0.55	0.82	0.52
	Baseline (Ours)	0.03	0.16	8.33%	**0.03**	0.14	0.51	0.48
	IREU (Ours)	**0.01**	**0.15**	**6.95%**	0.21	**0.55**	**0.85**	**0.64**
fiveID	BIA [33] [CVPR 25']	0.15	0.32	59.03%	0.43	0.52	0.65	0.22
	Baseline (Ours)	0.02	0.19	**11.81%**	**0.13**	0.01	0.20	0.07
	IREU (Ours)	**0.02**	**0.15**	12.50%	0.23	**0.53**	**0.78**	**0.55**

then take $\tilde{z}_-$ as an explicit reference for updating ε_u on the target identity to be unlearned. For an image from the target to be unlearned $x_- \in X_-$, we encode it with the trainable encoder to obtain $z_- = \varepsilon_u(x_-)$. We optimize the image encoder by minimizing the following ℓ_2 distance:

$$\mathcal{L}_f = \left\| z_- - \tilde{z}_- \right\|_2. \tag{4}$$

This unlearning loss allows the final features to deviate largely from the original features in terms of identity, so that the generated images look dissimilar from the original identity while preserving high fidelity.

For other identities to be retained, we reuse the retaining loss from the baseline:

$$\mathcal{L}_r = \left\| z_+ - z'_+ \right\|_2, \tag{5}$$

where $z_+ = \varepsilon_u(x_+)$ and $z'_+ = \varepsilon(x_+)$. This loss keeps the features aligned with the original ones for other identities to be retained.

Combining Eqs. 4 and 5, the final training objective simply balances the unlearning and retaining terms:

$$\mathcal{L}_u = \lambda_1 \mathcal{L}_f + \mathcal{L}_r, \tag{6}$$

where λ_1 is a hyperparameter that controls the weight of the loss function.

We optimize only the image encoder ε_u across the entire CPG pipeline, keeping all other components frozen.

4 Experiments

4.1 Implementation Details

We use CelebA-HQ [17] as the primary face dataset and split it into training, validation, and test sets with a 7:1:2 ratio. Rather than consuming the full training

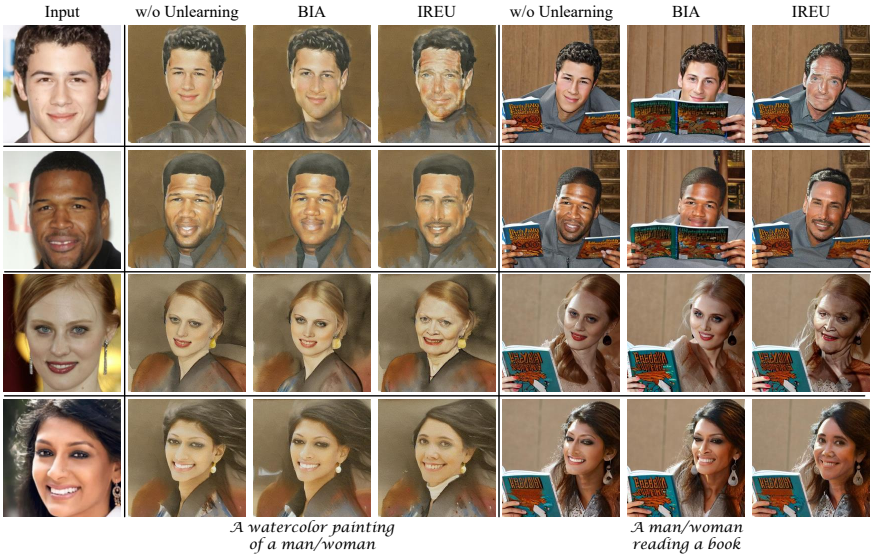


Fig. 5: Comparison of visual quality with BIA [33] for unlearning under the oneID setting on FastComposer [45]. For each input image (first column), both BIA and our method perform identity unlearning from the same original model. Text prompts are included below each generated image.

split, we randomly sample training images according to the number of iterations. To evaluate unlearning performance on unseen images of the same identity, we additionally use CelebA [26]. Our CPG pipeline utilizes FastComposer [45] for single-subject generation with a resolution of 512×512, employing Stable Diffusion v1.5 as its backbone and CLIP ViT-L/14 as the image encoder. To evaluate transferability, we directly plug the unlearned encoder into PhotoMaker [25] and the multi-subject FastComposer setting without any additional training.

The framework is implemented in PyTorch and optimized with Adam at a learning rate of 1×10^{-5}. All experiments are conducted on a single NVIDIA GeForce RTX 3090 GPU with a batch size of 1. For the single-identity (oneID) unlearning task, we train for 1,000 iterations. In addition, we construct an extended multi-identity (fiveID) unlearning task, for which we train for 12,000 iterations. For inference, we employ 50 denoising steps to generate customized images. The hyperparameters are set as follows: $k = 0.4$, $\lambda_1 = 0.1$, and $\alpha = 100$.

4.2 Evaluation Metrics

Identity similarity. To evaluate the effectiveness of our method, we report a set of identity-centric metrics. We employ CurricularFace [14] to measure the identity similarity between two images: $s(a, b)$. Let X_{-}^{train} be the set of training images of identities to be unlearned, X_{-}^{unseen} a disjoint set of unseen images from the same identities. We report:

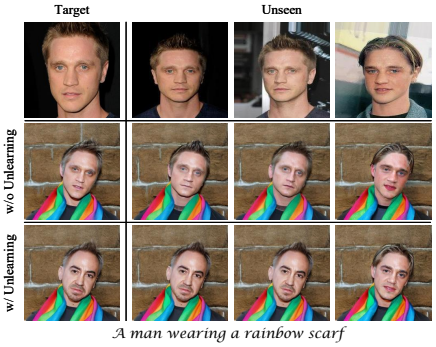


Fig. 6: Visualization results for unseen images of the same target identity to be unlearned under the fiveID setting on FastComposer [45]. "Target" refers to the images used during unlearning, while "Unseen" denotes the images belonging to the same target identity to be unlearned but not used during unlearning.

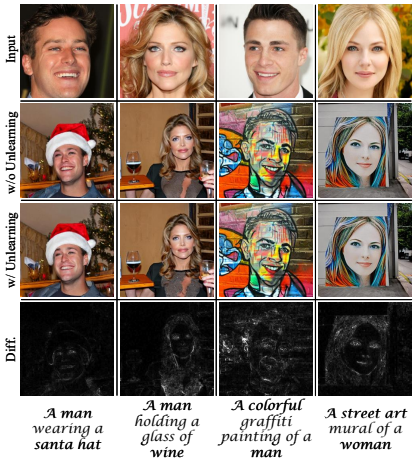


Fig. 7: Image generation quality for other identities to be retained under the fiveID setting. "Diff." indicates the normalized difference between images before and after unlearning.

- $\mathrm{ID}^{o}_{\mathrm{target}}$ (↓): $\mathbb{E}_{x_- \in X_-^{\mathrm{train}}}\, s(y_-, x_-)$, the average identity similarity between the generated image and the input image to be unlearned.
- $\mathrm{ID}^{o,\mathrm{unseen}}_{\mathrm{target}}$ (↓): $\mathbb{E}_{x_- \in X_-^{\mathrm{unseen}}}\, s(y_-,\, x_-)$, the average similarity between the generated images and *unseen* images of the same target identity.
- $\mathrm{ID}^{g}_{\mathrm{target}}$ (↓): $\mathbb{E}_{x_- \in X_-^{\mathrm{train}}}\, s(y_-,\, D(\varepsilon(x_-), t))$, the average identity similarity between the post-unlearning and pre-unlearning generated images under the same prompt.
- $\mathrm{ID}^{\mathrm{human}}_{\mathrm{target}}$ (↓): Following the evaluation protocol of BIA, we recruited 54 participants to assess whether 12 image pairs (before and after unlearning) belong to the same identity to be unlearned.
- $\mathrm{ID}^{o}_{\mathrm{other}}$ (↑): $\mathbb{E}_{x_+ \in X_+}\, s(y_+, x_+)$, the average identity similarity between the generated image and the other identity to be retained.
- $\mathrm{ID}^{g}_{\mathrm{other}}$ (↑): $\mathbb{E}_{x_+ \in X_+}\, s(y_+,\, D(\varepsilon(x_+), t))$, the average identity similarity between the post-unlearning and pre-unlearning generated images under the same prompt.
- ΔID (↑): $\mathrm{ID}^{g}_{\mathrm{other}} - \mathrm{ID}^{g}_{\mathrm{target}}$, the gap between retaining and unlearning in the metric space.

Note that $\mathrm{ID}^{g}_{\mathrm{target}}$, $\mathrm{ID}^{g}_{\mathrm{other}}$, and ΔID are proposed in this work as CPG-specific metrics to jointly assess unlearning and retaining in the generation space.

Fidelity. We report ΔFID, i.e., the change in Fréchet Inception Distance (FID) score [13] on the set of other identities to be retained, comparing before and after unlearning against real CelebA-HQ images. We also report PSNR, SSIM, and LPIPS to quantify perceptual and structural fidelity.

Table 2: Comparison of different methods on FastComposer [45] w.r.t. fidelity-centric metrics for other identities to be retained under single-identity (oneID) and multi-identity (fiveID) settings.

ID Type	Methods	PSNR(↑)	SSIM(↑)	LPIPS(↓)	ΔFID(↓)
oneID	BIA [33] [CVPR 25']	20.51	0.70	0.19	4.14
	Baseline	22.82	0.79	0.11	4.33
	IREU	**28.88**	**0.92**	**0.04**	**1.02**
fiveID	BIA [33] [CVPR 25']	19.46	0.66	0.22	2.76
	Baseline	20.16	0.70	0.19	28.37
	IREU	**28.25**	**0.91**	**0.04**	**1.59**

4.3 Unlearning Performance on Target Identities

Table 1 reports unlearning performance w.r.t. identity-centric metrics under both oneID and fiveID settings. Based on the results, we have the following observations: (1) Our methods (including baseline and IREU) obtain better unlearning performance (lower ID similarities) than BIA across both settings and four metrics, indicating our proposed encoder-only unlearning strategy is effective. (2) Our method IREU obtains comparable performance for target identities to be unlearned, but achieves a better trade-off between unlearning and retaining (higher ΔID), demonstrating the impact of our proposed unlearning method focusing on identity-related features. (3) Our methods better handle multi-identity settings. For example, the gap w.r.t. $\text{ID}^{\text{human}}_{\text{target}}$ grows from around 20 pp (oneID) to 48 pp (fiveID). (4) Our methods are more robust than BIA when the generator meets unseen images from the same target identity. Our numbers w.r.t. $\text{ID}^{o,\text{unseen}}_{\text{target}}$ remain at 0.15, while those of BIA increase from 0.18 to 0.32.

Fig. 5 qualitatively compares IREU and BIA using the same input images and prompts. For BIA, the generated images after unlearning still look rather similar to the ones w/o unlearning, while our method IREU suppresses the identity information more thoroughly. These results indicate that our method IREU better prevents the target identity images from being edited. Furthermore, we show some results of unseen images (not used during unlearning) belonging to the same target identity in Fig. 6. IREU achieves identity unlearning even for unseen images. These results reveal that IREU also successfully pushes them away from the target identities in the feature space. This advantage well fulfills the requirements in real-world applications where generators usually meet unseen images.

4.4 Retaining Performance on Other Identities

For other identities to be retained, we evaluate both identity consistency ($\text{ID}^{o}_{\text{other}}$, $\text{ID}^{g}_{\text{other}}$ in Table 1) and perceptual fidelity (Table 2). Based on the results, we have the following observations: (1) Our method IREU obtains the highest identity similarity and fidelity across both oneID and fiveID settings, outperforming

Fig. 8: Qualitative results when our unlearned image encoder is plugged into a new generator PhotoMaker [25] without any finetuning under the fiveID setting. Top: an identity to be retained; bottom: a target identity to be unlearned.

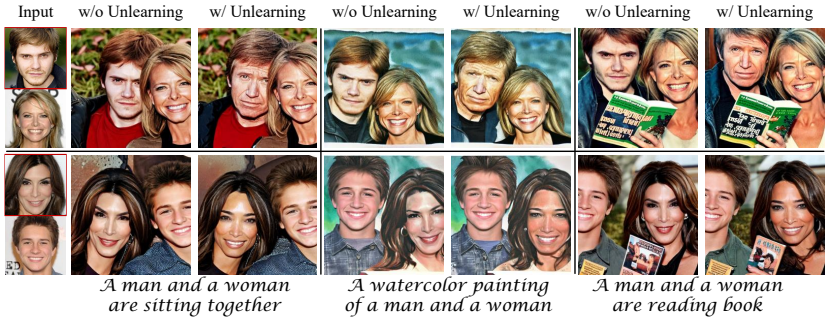


Fig. 9: Multi-subject image generation results of FastComposer [45]. Our unlearned image encoder is adapted from the single-subject to multi-subject setting without any finetuning. For each example, the input image marked with a red box is the target identity to be unlearned, and the other one represents the identity to be retained.

both BIA and the baseline. These results demonstrate that our method not only preserves the identity information, but also maintains high-quality image generation for those identities to be retained, not affected by the unlearning procedure. (2) The baseline fails dramatically under the fiveID setting (0.01 w.r.t. $\mathrm{ID}^{o}_{\mathrm{other}}$), indicating the global perturbation in the feature space becomes more harmful when the number of target identities to be unlearned grows. In contrast, our proposed identity-related feature perturbation overcomes this problem and thus achieves better retaining performance.

In Fig. 7, we visualize some generation results for the other identities to be retained. It can be seen that the post-unlearning outputs are closely similar to the original ones before unlearning, with the differences being barely perceptible to the human eye and mainly concentrated in high-frequency regions (e.g., hair, fine textures), as shown in the last row. This comparison indicates that our method IREU keeps the feature distribution stable and confines residual updates to visually tolerant regions, thereby avoiding global drift.

4.5 Generalization to New Generators

Since our method IREU only fine-tunes the image encoder while keeping the diffusion model frozen, it can be plugged into new generators sharing a similar pipeline. This is a notable advantage compared with previous methods, e.g. BIA.

Table 3: Quantitative evaluation of generalization ability when our unlearned image encoder is plugged into a new generator PhotoMaker [25] without any finetuning.

Method	$\text{ID}^g_{\text{target}}(\downarrow)$	$\text{ID}^g_{\text{other}}(\uparrow)$	ΔID
Baseline	0.0872	0.1980	0.1108
IREU	0.1365	0.8372	0.7007

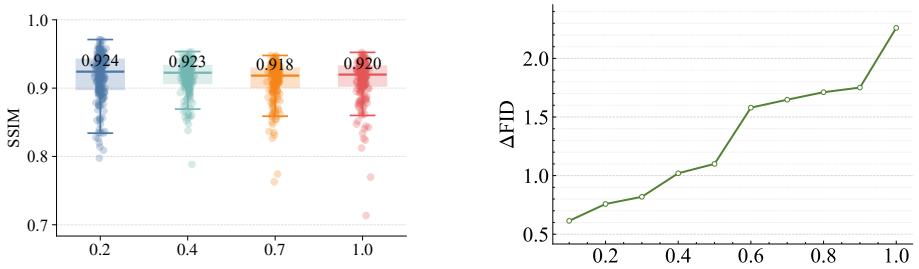
We first apply our unlearned image encoder to the new generator PhotoMaker [25] without any finetuning. The quantitative results are reported in Table 3, where we can see our method IREU obtains a low ID similarity (0.1365 w.r.t. $\text{ID}^g_{\text{target}}$) for target identities to be unlearned and a high ID similarity (0.8372 w.r.t. $\text{ID}^g_{\text{other}}$) for other identities to be retained, reaching a good trade-off between unlearning and retaining (a high value of 0.7007 w.r.t. ΔID). Compared with the baseline which also obtains a low ID similarity (0.0872 w.r.t. $\text{ID}^g_{\text{target}}$) for target identities to be unlearned but a low ID similarity (0.1980 w.r.t. $\text{ID}^g_{\text{other}}$) for other identities to be retained, our method IREU shows a better generalization ability. These results indicate that our proposed identity-related feature perturbation enhances the robustness across different generators. We also show some qualitative results in Fig. 8, where we can see the generated images for the target identity to be unlearned look rather different from the input image in terms of identity, while the generated images for the other identity to be retained well retain the identity compared to the input image.

Furthermore, we apply our unlearned image encoder optimized based on the single-subject model of FastComposer [45] to its multi-subject counterpart without any finetuning. As shown in Fig. 9, our method is able to change the target identities to be unlearned and in the meantime retain the other identities.

4.6 Ablation Study

Impact of the ratio of identity-related features k. The ratio k in Eq. 2 determines the number of features that will be updated during unlearning for the target identities to be unlearned. Basically, the larger the value of k, the more features will be updated, some of which may be identity-agnostic and are not expected to be updated. As shown in Fig. 10, when k increases, more examples exhibit low SSIM values and the ΔFID value grows, both of which indicate that fidelity is decreasing. This phenomenon becomes more severe when k is larger than 0.4. Considering both unlearning performance and fidelity preservation, we choose $k = 0.4$ in our experiments.

Impact of perturbation magnitude α. The perturbation magnitude α in Eq. 3 controls how much the identity-related features are perturbed given the mask. In Fig. 11, we analyze its effect on the generated outputs. When α is



(a) Boxplot of SSIM versus k. **(b)** ΔFID values versus k.

Fig. 10: Analysis on the ratio of identity-related features k. (a) The SSIM values show a larger variance (with more small values) as k increases. (b) ΔFID consistently increases as k grows. These results support our argument that restricting perturbations to identity-related elements preserves fidelity better.

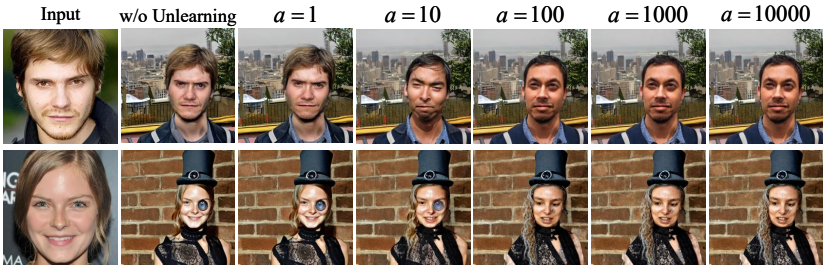


Fig. 11: Impact of the perturbation magnitude α in the identity-related transformation (Eq. 3). Increasing α consistently strengthens identity unlearning performance (lower ID similarity), but it saturates after $\alpha = 100$.

small, the constructed virtual identity $\tilde{z}_-$ stays close to z_-, leading to insufficient unlearning. As α increases, the deviation from the input image grows and the identity similarity of the target identities to be unlearned quickly drops. Such a trend saturates after $\alpha = 100$, so we adopt $\alpha = 100$ in our experiments.

Impact of the randomly selected swap target x_a. We analyze the influence of the swap target x_a used for constructing Face-Swap pairs. Specifically, for the same target identity x_-, we sample 10 different x_a and evaluate the resulting mask $\mathcal{M}$. As shown in Fig. 12, the pairwise top-k mask overlap is consistently higher than the random expectation of 40% under $k = 0.4$, with a mean overlap of $70.46\% \pm 5.19\%$. This indicates that the selected identity-related dimensions are stable across different choices of x_a.

Impact of masked identity-related difference magnitude. We analyze how the target-shift magnitude affects unlearning stability during the unlearning process. Specifically, we sort different cases by $\|z_- - \tilde{z}_-\|_2$, where $\tilde{z}_-$ is constructed following Eq. 3. As shown in Fig. 13, $\mathrm{ID}^o_{\mathrm{target}}$ stays low and $\mathrm{ID}^g_{\mathrm{other}}$ stays high across different target-shift magnitudes. These results demonstrate

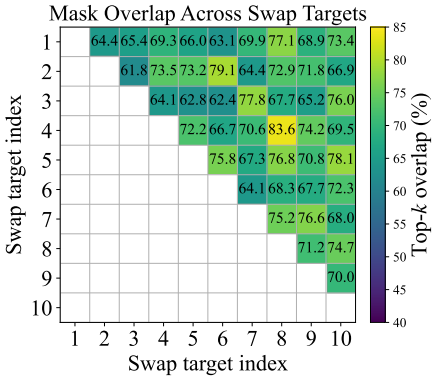


Fig. 12: Impact of the randomly selected swap target x_a. The mask overlap and final performance remain stable across different choices of x_a.

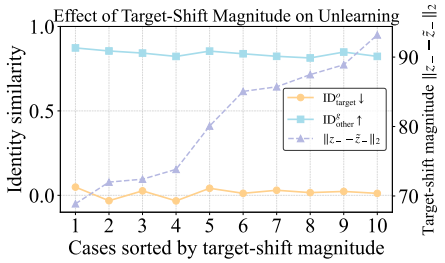


Fig. 13: Analysis on the magnitude of masked identity-related differences. Cases are sorted by $\|z_- - \tilde{z}_-\|_2$. The consistently low $\mathrm{ID}^o_{\mathrm{target}}$ and high $\mathrm{ID}^g_{\mathrm{other}}$ show that variations in masked differences do not harm the unlearning-retaining trade-off.

that mild variations in the identity-related shift do not degrade either identity unlearning or retaining performance.

5 Conclusion

Customized Portrait Generation (CPG) has become highly realistic and editable, while also posing significant privacy risks. Generative identity unlearning offers a crucial approach to address these privacy concerns, yet no specialized methods currently exist for identity unlearning in CPG. We present IREU, an Identity-Related Encoder-Only Unlearning framework for CPG. IREU achieves effective unlearning of target identities to be unlearned through its locating identity-related features and identity-related transformation components, while maintaining the fidelity of other identities to be retained. On CelebA-HQ, IREU achieves stronger identity suppression and better fidelity. The unlearned encoder can transfer to other CPG pipelines without retraining, demonstrating the effectiveness of encoder-only unlearning for CPG privacy protection.

6 Acknowledgements

This research is supported by the National Natural Science Foundation of China (Grant No. 62322602), and the Natural Science Foundation of Jiangsu Province, China (Grant No. BK20230033).

References

1. Ahmed, S.M., Basaran, U.Y., Raychaudhuri, D.S., Dutta, A., Kundu, R., Niloy, F.F., Guler, B., Roy-Chowdhury, A.K.: Towards source-free machine unlearning. In: CVPR. pp. 4948–4957 (2025)
2. Bala, A., Chowdhury, R., Jaiswal, R., Roheda, S.: Dct-shield: A robust frequency domain defense against malicious image editing. In: ICCV. pp. 18876–18884 (2025)
3. Chen, M., Gao, W., Liu, G., Peng, K., Wang, C.: Boundary unlearning: Rapid forgetting of deep networks via shifting the decision boundary. In: CVPR. pp. 7766–7775 (2023)
4. Choi, D., Choi, S., Lee, E., Seo, J., Na, D.: Towards efficient machine unlearning with data augmentation: Guided loss-increasing (GLI) to prevent the catastrophic model utility drop. In: CVPRW. pp. 93–102 (2024)
5. Feng, W., Zhou, W., He, J., Zhang, J., Wei, T., Li, G., Zhang, T., Zhang, W., Yu, N.: Aqualora: Toward white-box protection for customized stable diffusion models via watermark lora. arXiv preprint arXiv:2405.11135 (2024)
6. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-Or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion. In: ICLR (2023)
7. Gandikota, R., Materzynska, J., Fiotto-Kaufman, J., Bau, D.: Erasing concepts from diffusion models. In: ICCV. pp. 2426–2436 (2023)
8. Gandikota, R., Orgad, H., Belinkov, Y., Materzyńska, J., Bau, D.: Unified concept editing in diffusion models. In: WACV. pp. 5111–5120 (2024)
9. Gao, G., Huang, H., Fu, C., Li, Z., He, R.: Information bottleneck disentanglement for identity swapping. In: CVPR. pp. 3404–3413 (2021)
10. Gao, H., Pang, T., Du, C., Hu, T., Deng, Z., Lin, M.: Meta-unlearning on diffusion models: Preventing relearning unlearned concepts. In: ICCV. pp. 2131–2141 (2025)
11. Gong, C., Chen, K., Wei, Z., Chen, J., Jiang, Y.G.: Reliable and efficient concept erasure of text-to-image diffusion models. In: ECCV. pp. 73–88 (2024)
12. Gu, H., Ong, W.K., Chan, C.S., Fan, L.: Ferrari: Federated feature unlearning via optimizing feature sensitivity. In: NeurIPS. vol. 37, pp. 24150–24180 (2024)
13. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: NeurIPS. vol. 30 (2017)
14. Huang, Y., Wang, Y., Tai, Y., Liu, X., Shen, P., Li, S., Li, J., Huang, F.: Curricularface: Adaptive curriculum learning loss for deep face recognition. In: CVPR. pp. 5900–5909 (2020)
15. Huang, Z., Chan, K.C.K., Jiang, Y., Liu, Z.: Collaborative diffusion for multi-modal face generation and editing. In: CVPR. pp. 6080–6090 (2023)
16. Kang, T., Jeong, S., Jang, H., Choo, J.: Zero-shot head swapping in real-world scenarios. In: CVPR (2025)
17. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of gans for improved quality, stability, and variation. In: ICLR (2018)
18. Khalil, Y.H., Brunswic, L.M., Lamghari, S., Li, X., Beitollahi, M., Chen, X.: Not: Federated unlearning via weight negation. In: CVPR. pp. 25759–25769 (2025)
19. Kim, J., Gu, G., Park, M., Park, S., Choo, J.: Stableviton: Learning semantic correspondence with latent diffusion model for virtual try-on. In: CVPR. pp. 8176–8185 (2024)
20. Lee, B.H., Lim, S., Chun, S.Y.: Localized concept erasure for text-to-image diffusion models using training-free gated low-rank adaptation. In: CVPR. pp. 18596–18606 (2025)

21. Lee, T.Y., Seo, J., Ko, J.H., Park, G.M.: Perturb a model, not an image: Towards robust privacy protection via anti-personalized diffusion models. In: NeurIPS. vol. 38, pp. 92410–92442 (2025)
22. Li, A., Mo, Y., Li, M., Wang, Y.: Pid: Prompt-independent data protection against latent diffusion models. arXiv preprint arXiv:2406.15305 (2024)
23. Li, M., Chen, J., Feng, W., Li, B., Dai, F., Zhao, S., He, Q.: Hyperlora: Parameter-efficient adaptive generation for portrait synthesis. In: CVPR. pp. 13114–13123 (2025)
24. Li, N., Zhou, C., Gao, Y., Chen, H., Zhang, Z., Kuang, B., Fu, A.: Machine unlearning: Taxonomy, metrics, applications, challenges, and prospects. IEEE Trans. Neural Networks Learn. Syst. **36**(8), 13709–13729 (2025)
25. Li, Z., Cao, M., Wang, X., Qi, Z., Cheng, M., Shan, Y.: Photomaker: Customizing realistic human photos via stacked ID embedding. In: CVPR. pp. 8640–8650 (2024)
26. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: ICCV. pp. 3730–3738 (2015)
27. Lu, S., Wang, Z., Li, L., Liu, Y., Kong, A.W.K.: Mace: Mass concept erasure in diffusion models. In: CVPR. pp. 6430–6440 (2024)
28. Lyu, M., Yang, Y., Hong, H., Chen, H., Jin, X., He, Y., Xue, H., Han, J., Ding, G.: One-dimensional adapter to rule them all: Concepts diffusion models and erasing applications. In: CVPR. pp. 7559–7568 (2024)
29. Moon, S., Cho, S., Kim, D.: Feature unlearning for pre-trained gans and vaes. In: AAAI. pp. 21420–21428 (2024)
30. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: ICML. vol. 139, pp. 8748–8763 (2021)
31. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In: CVPR. pp. 22500–22510 (2023)
32. Seo, J., Lee, S., Lee, T., Moon, S., Park, G.: Generative unlearning for any identity. In: CVPR. pp. 9151–9161 (2024)
33. Shaheryar, M., Lee, J.T., Jung, S.K.: Black hole-driven identity absorbing in diffusion models. In: CVPR. pp. 28544–28554 (2025)
34. Shamshad, F., Naseer, M., Nandakumar, K.: Clip2protect: Protecting facial privacy using text-guided makeup via adversarial latent search. In: CVPR. pp. 20595–20605 (2023)
35. Song, Y., Yang, P., Ci, H., Shou, M.Z.: Idprotector: An adversarial noise encoder to protect against id-preserving image generation. In: CVPR. pp. 3019–3028 (2025)
36. Spartalis, C.N., Semertzidis, T., Gavves, E., Daras, P.: Lotus: Large-scale machine unlearning with a taste of uncertainty. In: CVPR. pp. 10046–10055 (2025)
37. Tian, L., Wang, Q., Zhang, B., Bo, L.: EMO: emote portrait alive generating expressive portrait videos with audio2video diffusion model under weak conditions. In: ECCV. vol. 15141, pp. 244–260 (2024)
38. Voigt, P., dem Bussche, A.V.: The eu general data protection regulation (gdpr). A Practical Guide, 1st Ed., Cham: Springer International Publishing **10**(3152676), 10–5555 (2017)
39. Voynov, A., Chu, Q., Cohen-Or, D., Aberman, K.: P+: Extended textual conditioning in text-to-image generation. arXiv preprint arXiv:2303.09522 (2023)
40. Wang, H., Weng, Y., Li, Y., Guo, Z., Du, J., Niu, S., Ma, J., He, S., Wu, X., Hu, Q., Yin, B., Liu, C., Liu, Q.: Emotivetalk: Expressive talking head generation

through audio information decoupling and emotional video diffusion. In: CVPR. pp. 26212–26221 (2025)
41. Wang, Q., Bai, X., Wang, H., Qin, Z., Chen, A.: Instantid: Zero-shot identity-preserving generation in seconds. arXiv preprint arXiv:2401.07519 (2024)
42. Wang, Z., Guo, J., Zhu, J., Li, Y., Huang, H., Chen, M., Tu, Z.: Sleepermark: Towards robust watermark against fine-tuning text-to-image diffusion models. In: CVPR. pp. 8213–8224 (2025)
43. Wu, J., Le, T., Hayat, M., Harandi, M.: Erasing undesirable influence in diffusion models. In: CVPR. pp. 28263–28273 (2025)
44. Wu, Y., Zhou, S., Yang, M., Wang, L., Chang, H., Zhu, W., Hu, X., Zhou, X., Yang, X.: Unlearning concepts in diffusion model via concept domain correction and concept preserving gradient. In: AAAI. pp. 8496–8504 (2025)
45. Xiao, G., Yin, T., Freeman, W.T., Durand, F., Han, S.: Fastcomposer: Tuning-free multi-subject image generation with localized attention. Int. J. Comput. Vis. **133**(3), 1175–1194 (2024)
46. Yang, J., Wang, Y., Fang, Y., Dai, Y., Huang, F.: Variance as a catalyst: Efficient and transferable semantic erasure adversarial attack for customized diffusion models. In: ICML. vol. 267 (2025)
47. Zeng, J., Song, D., Nie, W., Tian, H., Wang, T., Liu, A.: CAT-DM: controllable accelerated virtual try-on with diffusion model. In: CVPR. pp. 8372–8382 (2024)
48. Zhang, G., Wang, K., Xu, X., Wang, Z., Shi, H.: Forget-me-not: Learning to forget in text-to-image diffusion models. In: CVPRW. pp. 1755–1764 (2024)
49. Zhang, Y., Chen, X., Jia, J., Zhang, Y., Fan, C., Liu, J., Hong, M., Ding, K., Liu, S.: Defensive unlearning with adversarial training for robust concept erasure in diffusion models. In: NeurIPS. vol. 37, pp. 36748–36776 (2024)
50. Zhong, Z., Bao, W., Wang, J., Zhang, S., Zhou, J., Lyu, L., Lim, W.Y.B.: Unlearning through knowledge overwriting: Reversible federated unlearning via selective sparse adapter. In: CVPR. pp. 30661–30670 (2025)
51. Zhou, Y., Zheng, D., Mo, Q., Lu, R., Lin, K.Y., Zheng, W.S.: Decoupled distillation to erase: A general unlearning method for any class-centric tasks. In: CVPR. pp. 20350–20359 (2025)