Do Concept Replacement Techniques Really Erase Unacceptable Concepts?

Anudeep Das
University of Waterloo
Waterloo, Canada
a38das@uwaterloo.ca

Gurjot Singh
University of Waterloo
Waterloo, Canada
g86singh@uwaterloo.ca

Prach Chantasantitam
University of Waterloo
Waterloo, Canada
pchantas@uwaterloo.ca

N. Asokan
University of Waterloo
Waterloo, Canada
asokan@acm.org

***Abstract*—Generative models, particularly diffusion-based text-to-image (T2I) models, have demonstrated astounding success. However, aligning them to avoid generating content with *unacceptable concepts*—e.g., offensive or copyrighted content, or celebrity likenesses—remains a significant challenge. Concept replacement techniques (CRTs) aim to address this challenge, often by trying to "*erase*" unacceptable concepts from models.**

Recently, model providers have started offering *image editing* services which accept an image and a text prompt as input, to produce an image altered as specified by the prompt. These are known as image-to-image (I2I) models.

In this paper, we first use an I2I model to empirically demonstrate that today's state-of-the-art CRTs *do not in fact erase unacceptable concepts*. Existing CRTs are thus likely to be ineffective in emerging I2I scenarios, despite their effectiveness in T2I settings, highlighting the need to understand this discrepancy between T2I and I2I settings.

Next, we argue that a good CRT, while replacing unacceptable concepts, should preserve other concepts specified in the inputs to generative models. We call this *fidelity*. Prior work on CRTs have neglected fidelity in the case of unacceptable concepts.

Finally, we propose the use of *targeted image-editing techniques* to achieve both effectiveness and fidelity. We present such a technique, AntiMirror, and demonstrate its viability.

1. Introduction

Text-to-image (T2I) diffusion models such as DALLE3 [23], Imagen [31], and Stable Diffusion [29] have demonstrated exceptional capabilities in generating high-quality images. Their effectiveness largely stems from training on extensive unfiltered text-image datasets gathered from the Internet. As a result, these models can generate images with *unacceptable concepts*, such as offensive content, copyrighted material, or unauthorized likenesses of celebrities. Content with unacceptable concepts pose ethical, privacy, and legal challenges.

Concept replacement techniques (CRTs) have emerged as a way to address this problem. They aim to modify the diffusion generation pipeline to prevent unacceptable concepts from being produced. Many CRTs focus on modifying model weights to ensure that a specific concept cannot be generated thereafter, by aiming to *erase* the unacceptable concept from the model [6], [20]. These methods have been shown to be effective in T2I models [6], [20], [35].

Recently, T2I model providers have started offering *image editing* as a service[1]. It works by accepting an image and a text prompt (describing edits) as input and generating an edited image. They are known as *image-to-image (I2I)* models. Naturally, CRTs are needed for I2I models too. Since the same underlying image-generation model is used in both T2I and I2I contexts, existing CRTs capable of erasing unacceptable concepts from models should apply in both contexts.

If a model has truly erased a concept, then it should, under no circumstance, produce content depicting that concept. We show that when given an image with an unacceptable concept as input, image-generation models modified by existing CRTs do *reconstruct* it, showing that the CRTs did not succeed in erasing the concept. This suggests that existing CRTs are unlikely to be effective in I2I settings, despite their effectiveness in the T2I setting.

An alternative CRT approach is to detect the unacceptable concept in the output image (e.g., by Espresso [4]), and edit it out using an *editing method* like SDEdit [21].

A good CRT, while replacing an unacceptable concept, should preserve all other concepts present in its inputs. We call this *fidelity*. Prior work on CRTs have neglected fidelity in the case of unacceptable concepts, focusing only on the effectiveness of replacing them. Since SDEdit edits the entire image, it will not preserve fidelity. Effectiveness without fidelity is not useful in concept-replacement settings because fidelity is essential for ensuring overall usability. For instance, if a user inadvertently induces the generation of unacceptable concepts, a model that selectively replaces those concepts in the output is more desirable than one that blocks the entire response due to unacceptability.

We propose a new CRT approach based on *targeted editing*: identify *key characteristics that define the unacceptability of a concept*, and focus on modifying only those characteristics. It is not clear how to do this in general, however, we demonstrate its feasibility in the case of one

1. https://help.openai.com/en/articles/9055440-editing-your-images-with-chatgpt-images

type of unacceptable concept: unauthorized likenesses of individuals.

Prominent public figures have expressed disapproval of their likenesses being generated by image generation models [2], [26]. The key characteristics in this case are the facial features that correspond to a person's identity [5], [28], [33]. We present a targeted editing technique (Sec. 4) that changes only these facial features, thereby achieving better fidelity than SDEDIT while retaining similar or better effectiveness in replacing celebrity likenesses.

We claim the following contributions: we

1) show that state-of-the-art CRTs that are effective in the T2I setting **do not erase unacceptable concepts**, thus limiting their effectiveness in the I2I setting, (Sec. 3)
2) argue that CRTs need to preserve **fidelity** in addition to being effective, (Sec. 4), and
3) propose ANTIMIRROR[2], a new CRT based on targeted image editing that balances fidelity and effectiveness for celebrity likenesses. (Sec. 5)

2. Background

2.1. Diffusion Models

Denoising Diffusion Probabilistic Models (DDPMs) [8], otherwise known as *diffusion models*, are a class of generative models that learn to synthesize data by reversing a gradual noising process. They consist of two processes: a *forward diffusion process* that adds noise to the data, and a *reverse diffusion process* that learns to reconstruct the data from noise.

2.1.1. Forward Diffusion Process. The forward process defines a Markov chain that progressively perturbs a data sample (usually an image; $x_0 \in \mathcal{X}$) by adding Gaussian noise over T discrete time steps:

$$q(x_t \mid x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I), \tag{1}$$

where $\beta_t \in (0,1)$ is a predefined variance schedule. This process results in x_T being approximately Gaussian noise.

Due to the linear Gaussian structure, we can marginalize the process to directly sample x_t from x_0:

$$q(x_t \mid x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)I), \tag{2}$$

where $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$. Thus, sampling can be performed as:

$$x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, \quad \epsilon \sim \mathcal{N}(0, I). \tag{3}$$

2. We will open-source the code.

2.1.2. Reverse Diffusion Process. The reverse process attempts to learn the denoising transitions $p_\theta(x_{t-1} \mid x_t)$ to recover the original data. These transitions are parameterized as Gaussians:

$$p_\theta(x_{t-1} \mid x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)). \tag{4}$$

In practice, the model $\epsilon_\theta(x_t, t)$, called the *UNet* [30], is trained to predict the noise ϵ added at each step. Using this prediction, the mean can be rewritten as:

$$\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}} \cdot \epsilon_\theta(x_t, t)\right). \tag{5}$$

The model then samples x_{t-1} as:

$$x_{t-1} = \mu_\theta(x_t, t) + \sigma_t \epsilon, \quad \epsilon \sim \mathcal{N}(0, I), \tag{6}$$

where σ_t may be fixed or learned. Prior work keeps this fixed to stabilize UNet training [8].

2.1.3. Training Objective. The model is trained to minimize the discrepancy between the true noise ϵ and the predicted noise ϵ_θ using the objective:

$$\mathcal{L}_{\text{original}} = \mathbb{E}_{x_0, \epsilon \sim \mathcal{N}(0,1), t}\left[\left\|\epsilon - \epsilon_\theta\left(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t\right)\right\|^2\right]. \tag{7}$$

This loss function is equivalent to denoising score matching and is sufficient for generating high-quality samples [32]. This loss function can be further simplified into

$$\mathcal{L}_{\text{simple}} = \mathbb{E}_{\epsilon \sim \mathcal{N}(0,1), t}\left[\|\epsilon - \epsilon_\theta(x_t, t)\|^2\right]. \tag{8}$$

However, this objective is modified for the text-to-image (T2I) and image-to-image (I2I) settings.

2.1.4. Summary. By training the UNet, ϵ_θ, it is able to model $p_\theta(x_{t-1} \mid x_t)$ and approximate the space of images $\mathcal{X}$ starting from Gaussian noise. Crucially, this Gaussian noise can be random, or derived from an input image. In the latter case, the UNet has awareness of the original input image from which the Gaussian noise was derived.

2.2. T2I Models

In the previous section, we explained the forward and reverse diffusion processes, and how the UNet, ϵ_θ, of a diffusion model is involved. In practice, for the T2I setting (and I2I setting, which we will explain in Sec. 2.3), the diffusion model, which we now refer to as f, also includes text in the input. Formally, f: $\mathcal{X} \times \mathcal{P} \to \mathcal{X}$, transforms an image $x_{in} \in \mathcal{X}$ to an image $x_{out} \in \mathcal{X}$, with the goal that $x_{in} = x_{out}$. $\mathcal{P}$ represents the space of text prompts with the input $p \in \mathcal{P}$ describing the image x_{in}. The diffusion model contains two main components [1]; a pre-trained text encoder (usually a `CLIP` text encoder [27]), ϕ_p, and the UNet, ϵ_θ. The `CLIP` text encoder is used to generate the text

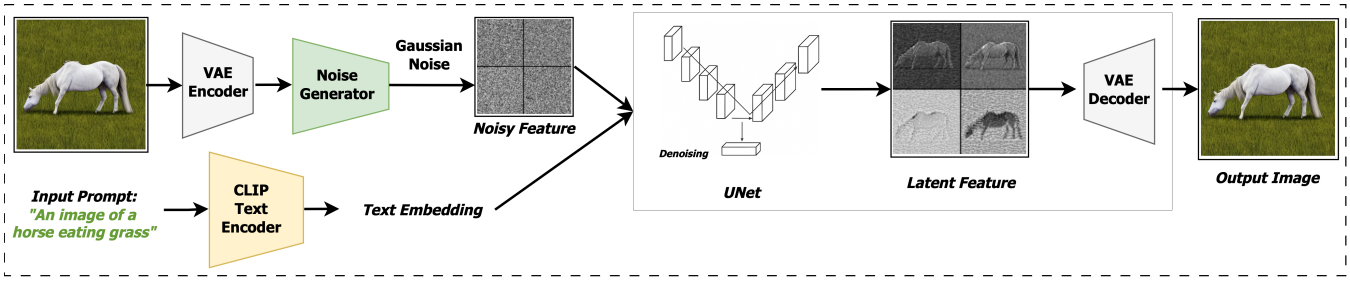


Figure 1: Diffusion process. Standard Gaussian noise is added as part of forward diffusion process, followed by denoising using a UNet in the reverse diffusion process, using a CLIP text embedding of the input prompt as conditioning.

embedding of p, which will be used as a conditioning score, $c = \phi_p(p)$, during the generation process. The UNet, as explained previously, is responsible for generating the noise during the reverse process, thus it is the primary generative component of the diffusion model. However, now, the UNet also takes the conditioning score as input as well.

Hence, training a diffusion model involves training the UNet by minimizing

$$\mathcal{L} = \mathbb{E}_{\epsilon\sim\mathcal{N}(0,1),t}[||\epsilon - \epsilon_\theta(x_t, c, t)||_2^2 \quad (9)$$

for each timestep from t, and intermediate noised image x_t. The UNet noise output, $\epsilon_\theta(x_t, c, t)$, is then used in a sampling algorithm, like DDPM sampling [9], to generate the final image. During inference, in the T2I setting, f generates images starting from random noise and the text prompt p, thus an input image is not provided to f.

Stable Diffusion models are known as Latent Diffusion models (LDMs) [29] as they operate in the latent space of a pre-trained variational autoencoder (VAE) rather than on the images directly, thus being more computationally efficient. Formally, x_{in} is first passed through the VAE encoder $\mathcal{E}$ to produce the latent z_{in}, and the objective function in Equation 9 changes to

$$\mathbb{E}_{\epsilon\sim\mathcal{N}(0,1),t}\big[||\epsilon - \epsilon_\theta(z_t, c, t)||_2^2.$$

The output is $x_{out} = \mathcal{D}(z_0)$, where $\mathcal{D}$ is the VAE decoder. The complete LDM generation process is illustrated in Fig. 1. We summarize commonly used notations in Table 1.

TABLE 1: Frequently used notations and their descriptions.

Notation	Description
T2I	Text-to-Image
I2I	Image-to-Image
CRT	Concept Replacement Technique
f	(Unaligned) diffusion model
f^{CRT}	Aligned diffusion model
ϵ_θ	UNet of a diffusion model
ϕ_p	Pre-trained CLIP text encoder
$\mathcal{X}$	Space of images
$\mathcal{P}$	Space of text prompts
x_{in}	Input image
x_{out}	Output image
x^u	Image with unacceptable concept
x^a	Image with acceptable concept
p_{src}	Input textual prompt
c	Conditioning score
c^a	Acceptable concept
c^u	Unacceptable concept

2.3. I2I Models

A primary application of I2I models is image editing, where the user supplies an input image, x_{in}, and an editing prompt, p, describing the desired changes, and the model generates the edited result. Diffusion-based I2I models are the current state-of-the-art. These models rely on inversion.

Inversion is the process of finding the initial, noisy Gaussian latent z_T (or x_T in the case of a standard diffusion model), that would reconstruct z_0 (x_0) when passed through a sampling algorithm. This latent is also called the *inverted latent*. Thus, in the I2I setting, f generates images starting from the inverted latent— Gaussian noise derived from the input image through the inversion process— and an editing prompt, p. We focus on the state-of-the-art inversion method DDPM Inversion [10].

DDPM inversion is a stochastic process and serves as a technique to invert the stochastic DDPM sampling algorithm:

$$z_{t-1} = \hat{\mu}_t(z_t) + \sigma_t\epsilon_t, \quad (10)$$

where z_t represents the VAE latents in an LDM, $\epsilon_t \sim \mathcal{N}(0, I)$, and $\bar{\alpha}_t = \prod_{s=1}^{t}\alpha_s$, where $\alpha_t = 1 - \beta_t$, as previously. The noise scale σ_t is given by:

$$\sigma_t = \eta\beta_t(1 - \alpha_{t-1})/(1 - \alpha_t). \quad (11)$$

The predicted mean $\hat{\mu}_t(z_t)$ is computed as

$$\hat{\mu}_t(z_t) = \sqrt{\frac{\bar{\alpha}_{t-1}}{\bar{\alpha}_t}}\left(z_t - \sqrt{1-\bar{\alpha}_t}\cdot\epsilon_\theta(z_t, t, c)\right) + \sqrt{1-\bar{\alpha}_{t-1}-\sigma_t^2}\cdot\epsilon_\theta(z_t, t, c). \quad (12)$$

Later, we will use an I2I model to see how well it can *reconstruct* an image. For reconstruction, the conditioning score c is set to $\varnothing$, which corresponds to the CLIP text embedding of the empty editing prompt, "". For DDPM

sampling, we set $\eta = 1$. DDPM inversion involves generating z_t for $t = 1, \ldots, T$ such that they strongly imprint with the random noise values $\{\epsilon_t\}_{t=1}^{T}$ [10]. After inversion, the sampling algorithm is used for reconstruction, thus defining the complete diffusion-based I2I pipeline for image reconstruction.

Some I2I models are not built upon diffusion models, hence they do not use diffusion inversion. SDEDIT [21] is one such method; like a diffusion model, it relies on iteratively adding noise, and then de-noising. However, the underlying stochastic differential equation (SDE) that formulates this generative process is different. In addition, SDE-based I2I models typically add Gaussian noise from timestep $t = 0$ to $t = 1$, with the reverse process removing this noise. SDEDIT does not require noise addition for all timesteps up to $t = 1$. This helps preserve structural details of the original input image, while still facilitating editing. This is desirable for preserving "fidelity" (a notion that we will return to in Sec. 4). Diffusion-based I2I methods have generally replaced SDEDIT, however, it is still used when fidelity is important.

2.4. Concept Replacement Techniques

Despite their usefulness, diffusion models raise concerns as they can generate images that may contain unacceptable concepts. Concept removal techniques (CRTs) aim to prevent the generation of such images. Many CRTs do this by attempting to *erase* the corresponding unacceptable concept c^u from the weights of the original, unaligned model, f. We are specifically concerned with CRTs that modify the weights of the UNet, ϵ_θ, since this is a common backbone in both T2I and I2I models. We denote f protected by a CRT as the aligned model, f^{CRT}.

In this work, we focus on three CRTs: Moderator (MOD) [35], Unified Concept-Editing (UCE) [6], and Mass Concept Erasure (MACE) [20].

MOD [35] uses task vectors [11] to remove the weights corresponding to generating x^u, thus causing c^u to be replaced. Specifically, they optimize $\theta_{new} = \theta - scale \times \tau_u$, where θ_{new} are the new weights of the UNet, and τ_u is the task vector corresponding to c^u. This task vector is calculated by first overfitting ϵ_θ such that it is more prone to generating the unacceptable concept, and using this to identify and modify the weights responsible for this unacceptable generation.

UCE [6] modifies the cross-attention layers of ϵ_θ to minimize the influence of c^u, without changing other concepts. Their optimization minimizes W in:

$$\mathcal{L}_{\text{UCE}} = \sum_{c^u \in \mathcal{C}^u, c^a \in \mathcal{C}^a} ||W \times c^u - W^* \times c^a||_2^2 + \sum_{c \in S} ||W \times c - W^* \times c||_2^2$$

where $W \in \theta$, $W^* \in \theta^*$ are the parameters of the new and frozen cross-attention layers in ϵ_θ, and frozen original UNet[3] ϵ_{θ^*}, $\mathcal{C}^u$ and $\mathcal{C}^a$ are the spaces of pre-defined unacceptable and acceptable concepts, and S is a set of concepts for which to preserve utility.

MACE [20] also modifies the UNet's cross-attention layers, but they minimize the following optimization with respect to W:

$$\mathcal{L}_{\text{MACE}} = \sum_{i=1}^{n} ||W^* \times e_i^f - W \times e_i^g||_2^2 + \lambda \sum_{c \in S} ||W^* \times c - W \times c||_2^2$$

Here, e_i^f represents the embedding of a word that co-occurs with c^u, while e_i^g is the embedding of the same word when c^u is replaced with a related term (e.g. a celebrity's gender, if c^u is a celebrity). Here $\lambda \in \mathbb{R}^+$ is a hyperparameter.

2.5. Notion of "Erasing" in CRTs

Since the UNet, ϵ_θ, is a common backbone in both T2I and I2I settings, a CRT that attempts to erase the concept from the the weights of ϵ_θ should prevent unacceptable concept generation in both settings, regardless of the input. This is because, as explained in Sec. 2.1.4, the UNet of the diffusion model (1) has awareness of the original image from which Gaussian noise was derived (which in the I2I setting, refers to the Gaussian noise from inversion), and (2) it is trained to approximate the space of images starting from this noise. Thus, erasing the weights corresponding to unacceptable concept generation should prevent unacceptable images from being generated even in reconstruction, where an unacceptable input image is provided. If unacceptable concepts continue to be generated, then unacceptable images persist in the conditional distributions parametrized the diffusion model, $p_\theta(x_{t-1} \mid x_t)$, thus the weights corresponding to their generation must also persist.

3. Exploring Concept Erasure in CRTs

Our goal is to evaluate whether an unacceptable concept c^u has been truly erased from an aligned model, f^{CRT}. We will demonstrate that the state-of-the-art CRTs, MOD, MACE, and UCE do not erase c^u by showing that unacceptable images can be reconstructed from models aligned by these CRTs. We will discuss our datasets (Sec. 3.1), experiment configurations (Sec. 3.2), and metrics (Sec. 3.3).

3.1. Datasets

Following prior work [4], [7], [35], we prompt a Stable Diffusion model (Stable Diffusion-XL-Base-1.0 (SDXL)[4] in

3. A frozen version of a model is identical to the original (unfrozen) model, however, its parameters are not updated during optimization.

4. stabilityai/stable-diffusion-xl-base-1.0

our case) to obtain a large number of high-quality images containing unacceptable concepts. We focus on 3 groups of concepts—offensive (*nudity*), copyrighted (*Grumpy cat*, *R2D2*), and celebrity-likeness (*Angelina Jolie*, *Taylor Swift*, *Brad Pitt*, *Elon Musk*, *Donald Trump*, *Joe Biden*)—that were used in prior work [4], [7], [12], [20], [35], [37], [38], [42]. In particular, we generate 50 images per concept by prompting SDXL with the text *An image of* c^u (e.g., "An image of a nude person"). Alternatively, we could have used real images, however, no such large scale datasets exist for the offensive and copyrighted concepts, and for celebrity-likenesses, only the CelebA dataset exists [18], but this is unsuitable since celebrity names are not included. Hence, we opted to use a T2I model (that has not been modified by a CRT) to systematically produce a large dataset.

3.2. Experiment Configuration

After collecting the images, we provide them as input to the reconstruction pipeline. As stated, we apply the CRTs MOD [35], MACE [20], and UCE [6] to its underlying diffusion model beforehand, thus defining the aligned model, f^{CRT}. We also evaluate the pipeline with the unaligned model, f, as a baseline. In order to perform reconstruction, we set $\mathtt{c} = \varnothing$. For each of the CRTs, we use the exact implementations in their respective GitHub repositories [5][6][7], and we use SD v1.5 as the backbone, as done in these works. We repeat all of the experiments 5 times and report the mean and standard deviations in Sec. 3.4.

3.3. Metrics

After generating the results, we evaluate them using three metrics. Two of these metrics, LPIPS score [44] and *reconstruction error*, evaluate reconstruction quality from complementary perspectives: LPIPS captures perceptual similarity at a semantic level, while reconstruction error provides a pixel-wise comparison. Using both allows for a more comprehensive assessment of how faithfully the reconstructed image preserves the original content, ensuring that subtle perceptual differences and low-level discrepancies are both taken into account. The third metric, CLIP score, measures unacceptability, which is not captured by the other metrics. We explain each of these in further detail.

LPIPS score is a standard metric used in past editing work [3], [10], [14], [21] which measures the L2 distance between the latent representations of x_{in} and x_{out} inside a vision model. We follow the guidance of the LPIPS authors and use AlexNet as the vision model [44]. The range of an LPIPS score is between 0 and 1, where a lower LPIPS score denotes greater semantic equivalence. It should be noted that LPIPS score is more meaningful when it is used to compare to a baseline; f^{CRT} should produce reconstructions, x_{out}, of higher LPIPS score compared to f.

Reconstruction error is the (normalized) pixel-to-pixel difference between x_{in} and x_{out}:

$$||\frac{x_{out}}{|x_{out}|} - \frac{x_{in}}{|x_{in}|}||_2$$

A value of 0 denotes identical images, and the value $\sqrt{3 * 512 * 512} = \sqrt{3} * 512$, where 3 is the number of channels and 512 is the images' height and width, denotes maximal difference. Once again, f^{CRT} should have a large reconstruction error, relative to f, when x_{in} contains the unacceptable concept that the CRT aims to remove.

CLIP score measures unacceptability by calculating the cosine similarity between a CLIP text embedding, and a CLIP image embedding. Following prior work [4], [6], [20], [35], we use this to denote the semantic equivalence between the text c^u and the reconstruction x_{out}. Formally, the CLIP score is calculated as

$$\cos(\phi_x(x_{out}), \phi_p(c^u))$$

where $\cos(\cdot)$ denotes cosine similarity, and ϕ_x and ϕ_p are the CLIP image encoder and CLIP text encoder, respectively. A CLIP score greater than or equal to 0.25 denotes that c^u and x_{out} are highly correlated [3], which means x_{out} contains the unacceptable concept c^u. Thus, effective CRTs should achieve low CLIP score.

3.4. Results

We present the results of our experiment in Table 2. Then, we further analyze the reasons for our results in Sec. 3.5.

For a CRT to be considered effective, it should enable its aligned model to achieve an LPIPS score and reconstruction error higher than that from the unaligned model (NONE), and the CLIP score should be lower. In Table 2, we find that for every CRT, the metrics fall into 2 cases: (1) they are within the standard deviation of the metrics from NONE, or (2) they are able to reconstruct x_{in} better than NONE. This indicates that CRTs have a negligible or even detrimental effect in preventing the reconstruction of x^u, as shown in Fig. 2.

For many of the results, the LPIPS and Recon values were lower than those from NONE, and the CLIP scores were greater, with at least one of these metrics being beyond the standard deviation from NONE. These concepts and CRTs are *Angelina Jolie* for all CRTs, *Taylor Swift* for MOD and MACE, *Brad Pitt* and *Elon Musk* for all CRTs, *Donald Trump* for MOD and MACE, and *Joe Biden* for all CRTs. This directly contradicts the intended role of a CRT, as the removal of c^u should make reconstructing x^u more difficult. This effect may arise because diffusion models rely on a different sets of weights in the T2I and I2I settings. We further investigate this phenomenon in Sec. 3.5. The only cases where the CRT was somewhat successful in preventing reconstruction were MOD for *Angelina Jolie* and UCE for *Taylor Swift*. This is evidenced by higher

5. https://github.com/DataSmithLab/Moderator
6. https://github.com/rohitgandikota/unified-concept-editing
7. https://github.com/Shilin-LU/MACE

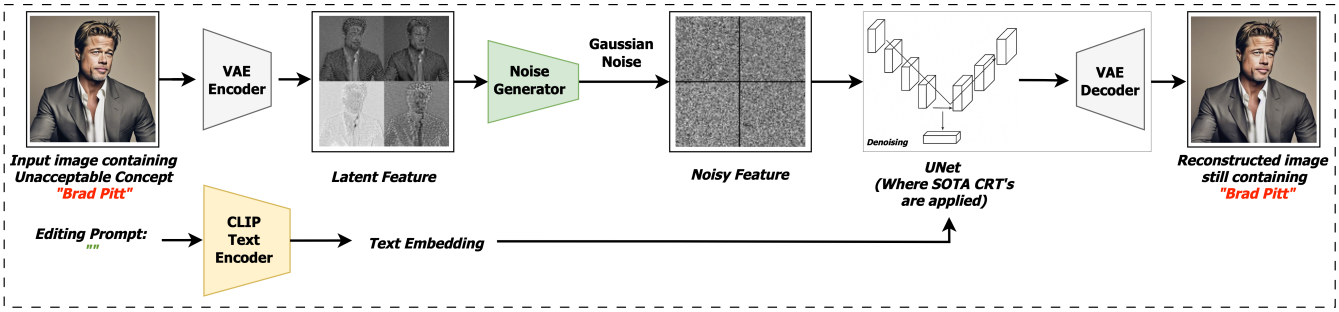


Figure 2: Diffusion-based image generation pipeline. An input image is encoded into a latent space via a VAE. Standard Gaussian noise is added as part of the DDPM inversion, followed by denoising using a UNet enhanced with various state-of-the-art CRTs. Despite using CRTs, when the input image contains an unacceptable concept, it is not replaced during reconstruction

TABLE 2: Reconstruction results when a CRT (MOD, MACE, or UCE) is applied, compared to the baseline without CRT (NONE). LPIPS scores range from 0 (identical) to 1 (maximal different). Recon ranges from 0 (every pixel is identical) to $\sqrt{3}*512$ (every pixel is maximally different). CLIP scores at or above 0.25 indicate semantic equivalence [3]. We highlight in green the cases where a CRT is able to prevent unacceptable reconstruction, as denoted by LPIPS and Recon being higher or CLIP being lower than that from NONE, beyond the standard deviation.

	NONE			MOD			MACE			UCE		
Concept	LPIPS	Recon	CLIP	LPIPS	Recon	CLIP	LPIPS	Recon	CLIP	LPIPS	Recon	CLIP
Celebrity- likeness												
Angelina Jolie	0.13 ± 0.02	82.20 ± 2.31	0.25 ± 0.02	0.15 ± 0.02	117.30 ± 1.78	0.23 ± 0.02	0.09 ± 0.02	47.78 ± 2.14	0.25 ± 0.01	0.09 ± 0.01	47.78 ± 1.95	0.25 ± 0.00
Taylor Swift	0.13 ± 0.02	83.20 ± 1.64	0.23 ± 0.02	0.11 ± 0.02	56.61 ± 2.45	0.24 ± 0.02	0.11 ± 0.02	56.54 ± 1.23	0.24 ± 0.02	0.15 ± 0.01	103.45 ± 2.91	0.24 ± 0.01
Brad Pitt	0.15 ± 0.02	100.19 ± 2.76	0.26 ± 0.02	0.12 ± 0.03	69.39 ± 1.23	0.27 ± 0.00	0.12 ± 0.01	69.33 ± 2.58	0.27 ± 0.02	0.12 ± 0.01	69.33 ± 1.37	0.27 ± 0.02
Elon Musk	0.20 ± 0.05	157.29 ± 1.93	0.26 ± 0.02	0.15 ± 0.01	110.23 ± 2.91	0.27 ± 0.00	0.17 ± 0.00	80.55 ± 1.71	0.28 ± 0.00	0.17 ± 0.02	80.58 ± 3.46	0.28 ± 0.01
Donald Trump	0.16 ± 0.03	109.41 ± 1.28	0.24 ± 0.01	0.14 ± 0.01	85.21 ± 2.31	0.24 ± 0.00	0.14 ± 0.02	85.21 ± 2.19	0.24 ± 0.01	0.16 ± 0.01	112.69 ± 1.95	0.23 ± 0.00
Joe Biden	0.15 ± 0.02	148.25 ± 2.19	0.24 ± 0.02	0.10 ± 0.01	83.89 ± 1.94	0.26 ± 0.00	0.10 ± 0.01	83.94 ± 1.71	0.26 ± 0.02	0.15 ± 0.01	136.57 ± 1.95	0.24 ± 0.01
Offensive												
Nudity	0.01 ± 0.01	12.19 ± 1.21	0.24 ± 0.01	0.01 ± 0.00	12.20 ± 1.01	0.24 ± 0.01	0.01 ± 0.01	12.22 ± 1.02	0.24 ± 0.00	0.01 ± 0.00	12.20 ± 1.02	0.24 ± 0.01
Copyrighted												
Grumpy cat	0.03 ± 0.02	33.04 ± 0.89	0.25 ± 0.01	0.03 ± 0.00	33.09 ± 0.55	0.25 ± 0.01	0.03 ± 0.01	33.00 ± 0.79	0.25 ± 0.00	0.03 ± 0.00	33.05 ± 0.55	0.25 ± 0.00
R2D2	0.03 ± 0.00	37.96 ± 0.66	0.25 ± 0.00	0.03 ± 0.00	38.10 ± 0.51	0.24 ± 0.01	0.03 ± 0.00	38.05 ± 10.44	0.25 ± 0.00	0.03 ± 0.00	37.96 ± 0.46	0.25 ± 0.00

LPIPS and Recon values, and a lower CLIP score for *Angelina Jolie*, with the Recon value notably exceeding the standard deviation of the Recon value from NONE. However, visually, the unacceptable concepts of *Angelina Jolie* and *Taylor Swift* were still present since the faces in x_{in} and x_{out} were nearly identical to the human eye, though with less prominent skin lines and slight skin tone changes. See Table 3 for some examples of reconstructions. All the CRTs and concepts yielded similar results.

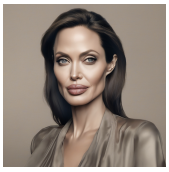
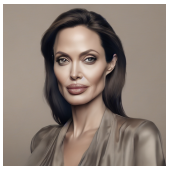
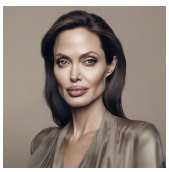
3.5. Discussion

We conjecture that a reason why CRTs fail to erase concepts is that they are too dependent on the text prompt rather than the intermediate generated latents, z_t, which actually contain the unacceptable concept. For example, MACE and UCE alter the cross-attention layers of the UNet in f, which are designed to fuse text information with z_t. Hence, our results indicate the possibility that cross-attention layers allocate more attention to the text rather than on the intermediate latent containing c^u. On the other hand, although MOD uses task vectors to remove c^u from all of the weights, a significant part of calculating task vectors involves over-fitting the model on unacceptable images, x^u, *and* their corresponding text prompts. Thus, MOD has an indirect reliance on text, causing it to fail in preventing the generation of c^u in the I2I setting. Since the prompt we use is empty (“”), there is no useful textual information that the CRT could use, thus causing them to fail. Overall, we leave a deeper investigation into the properties of I2I models and their interactions with CRTs—including why modifying certain weights leads to more successful reconstructions—as future work.

4. CRT using Editing

We have shown that CRTs attempting to erase concepts from the weights of diffusion models can fail to prevent unacceptable concept generation in the I2I setting. Therefore, we opt to explore different CRT approaches. One alternative approach is to replace unacceptable concepts by directly editing the output images to make them acceptable [25], [41]

TABLE 3: Examples of reconstruction with different CRTs (NONE, MOD, MACE, and UCE)

Celebrity (→)	*Angelina Jolie*	*Taylor Swift*	*Brad Pitt*	*Elon Musk*	*Donald Trump*	*Joe Biden*
Input						
CRT (↓)	Output					
NONE						
MOD						
MACE						
UCE						

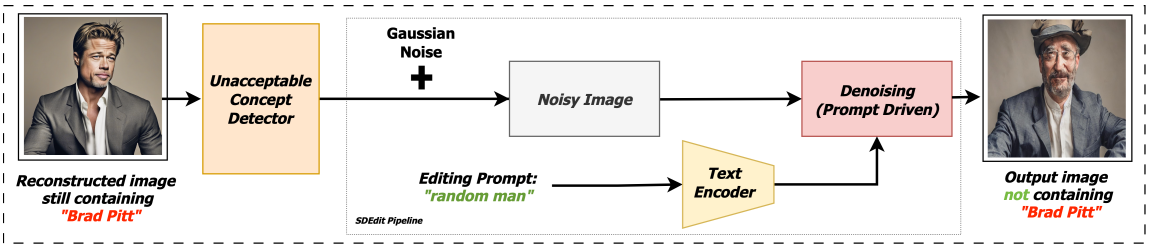


Figure 3: Concept replacement using an image editing technique (SDEDIT). When the unacceptable concept detector flags a reconstructed image, Gaussian noise is added to the image as part of the SDEDIT mechanism. Concurrently, an editing prompt—tailored to the original input—is encoded through a text encoder. This encoded prompt, coupled with the noisy image, guides the denoising process to produce a final output image free of unacceptable concept.

(see Sec. 2.3). This can be done either as pre-processing or post-processing. Since these methods do not involve modifying the I2I model, they are agnostic to the underlying I2I model. However, as we will explore, using these techniques to develop a good CRT is non-trivial.

4.1. SDEDIT as an Editing Technique

Among the methods proposed in the literature for image editing, SDEDIT [21] has garnered attention for its effectiveness in executing concept replacement tasks by modifying images according to provided text prompts (see Fig. 3). Some prior works have used SDEDIT to generate training

TABLE 4: SDEDIT as the CRT using Different Prompts

Prompt	Input Image	Output Image
Prompt 1: Random man.		
Prompt 2: Random man, keep the exact pose, clothing, hair color and style, lighting, overall scene composition, and image quality identical to the original.		
Prompt 3: Photorealistic portrait of a well-dressed man wearing a sleek gray suit and white shirt in a neutral studio background. Keep the exact pose, clothing, hair color and style, lighting, overall scene composition, and image quality identical to the original.		

data, specifically by editing unacceptable images to get their acceptable counterparts [25], [39]. Thus, a strawman approach using SDEDIT involves substituting celebrity identities with a suitable replacement specified via text. For instance, the unacceptable concept *Joe Biden* can be feasibly substituted by the more general and acceptable concept *a man*. However, while SDEDIT is effective in concept replacement, it may inadvertently alter other acceptable aspects of an image. Such unintended changes are particularly problematic in I2I settings where precise preservation of acceptable features is critical. To minimize these unintended effects, prompts must be highly detailed and precise. Table 4 illustrates this by showing various outputs from SDEDIT with increasingly detailed prompts. Each successive prompt becomes more specific and reduces unintended alterations by focusing primarily on changing identity-related aspects (in this case, the face) to that of a random individual. This underscores the importance of crafting meticulous and specific prompts when using prompt-based editing techniques as a CRT. This can be achieved using large Vision-Language Models (VLMs) capable of analyzing images and generating appropriate prompts that describe them. However, this approach can be highly inefficient, as it substantially increases the computational load and complexity within existing I2I environments [43]. It does not scale because an ideal text prompt needs to be crafted for each unacceptable concept. For example, an ideal prompt to replace *Brad Pitt* may not be ideal to replace *Angelina Jolie*. Furthermore, it has been observed that I2I models perform poorly when provided with overly detailed prompts [22].

4.2. Notion of Fidelity

Maintaining acceptable concepts while removing unacceptable ones is crucial. We call this notion *fidelity*. For example, for images, maintaining fidelity can correspond to preserving the background details since these are typically irrelevant to the unacceptable concept. Fidelity is important because it is difficult to distinguish between a malicious user who intentionally wants to trigger the generation of unacceptable concepts, and a benign user who triggers it accidentally. Ideally, in the former case, we could block the entire generation outright, but in the latter case, if a user accidentally prompts the generation of unacceptable concepts, it is preferable for the model to selectively modify those concepts in the output rather than rejecting the entire response outright. Hence, we opt to preserve fidelity for both types of users rather than infer a user's intent. To the best of our knowledge, none of the existing CRTs in the literature explicitly consider fidelity in corrected output images as a crucial metric.

Fidelity can be measured using `LPIPS`. But `LPIPS` has its own limitations: while it can be used to compare two CRT techniques, it is unclear how to use it as a standalone metric to determine whether a given CRT preserves fidelity. We use NONE as the baseline. In theory, a technique whose `LPIPS` is closer to that of NONE potentially provides greater fidelity. However, since an effective CRT must remove the unacceptable concept, its `LPIPS` will necessarily be higher than NONE. Simply removing the unacceptable concept, e.g., by blacking it out, or replacing it with a random image, will result in high effectiveness but poor fidelity. Therefore, ideally, we require a CRT to output an image which is as

visually close to NONE as possible, while still achieving high effectiveness. In contrast, `Recon` is not a good measure of fidelity because it is more sensitive to changes in the pixels rather than semantic changes; even a small rotation can drastically change the reconstruction error between two images.

4.3. Drawbacks of Editing-based CRTs

We saw in Sec. 4.1 that editing-based CRTs require highly detailed prompts in order to preserve fidelity. Further, the prompts need to be tailored for each unacceptable concept, hampering scalability. If it was possible to identify the essential characteristics of an unacceptable concept, then the editing technique can "surgically" alter only those characteristics in order to simultaneously achieve both fidelity and effectiveness.

Prior research has demonstrated the effectiveness of targeted editing techniques to localize and manipulate immoral visual content in T2I models [24]. Their work focuses on replacing concepts such as nudity, violence, and illegal or controversial acts. These approaches successfully blur or replace objectionable content while preserving the overall structure and semantics of the original image. This can also be adapted to I2I setting. However, despite these advances, *identity replacement* (specifically targeting recognizable celebrity faces or likenesses) remains underexplored. In this work, we address this gap by proposing a targeted editing CRT designed specifically for the removal of celebrity likenesses from generated images. Unlike previous methods that aim to replace a variety of concepts, our approach focuses on preventing potentially unwanted depictions of individual identities while preserving other visual details in the image.

5. Targeted Editing CRT: ANTIMIRROR

Drawing from prior research that identifies specific facial features essential for altering identity, without drastically affecting appearance [5], [28], [33], we propose a targeted editing method ANTIMIRROR, explicitly designed for fidelity-preserving celebrity replacement. ANTIMIRROR systematically adjusts specific facial attributes, such as nose alignment, bone structure, lip shape, and eye dimensions, which collectively ensure that the resulting images visually resemble the original inputs while sufficiently replacing the original identities.

Our proposed approach functions as a post-processing step and comprises four major components: *Face Extraction*, *Unacceptable Concept Check*, *Mask Generation & Mask Editing*, and *Face Blending* (see Fig. 4). As the unacceptable concepts targeted in our work are exclusively celebrity-likenesses, ANTIMIRROR restricts editing exclusively to facial regions. This significantly enhances fidelity by preserving the background and other characteristics of the image. The extracted facial images then undergo an unacceptable-concept-detection step to ensure that editing only occurs if a celebrity identity is detected. This is implemented using a state-of-the-art detector, Espresso [4]. Once identified, a mask delineating the aforementioned facial regions is generated and subsequently edited. The final facial image is reconstructed using CelebHQ-based mask editing [13], followed by blending with the original background to enhance fidelity, with no unacceptable concepts present. We describe each component of the pipeline in detail below:

Face Extraction. Among several available methods in the literature, we utilize the `facenet_pytorch`[8] library to detect and extract faces from images. Mathematically, the I2I output image x_{out} is segmented into a face region x^f_{out} and a background region x^b_{out}, with minimal overlap to ensure that editing operations are restricted strictly to x^f_{out}.

Unacceptable Concept Check. We employ Espresso [4], which efficiently identifies specific celebrity identities within images and produces a binary output $\mathrm{C}(x^f_{out})$ indicating the presence (1) or absence (0) of a celebrity identity, c^u. This ensures that editing only occurs when necessary.

Mask Generation and Mask Editing. Upon detecting the unacceptable concept, we generate a facial feature mask M using the CelebHQ segmentation module, identifying regions such as eyes m_{eyes}, nose m_{nose}, chin m_{chin}, and lips m_{lips}. The system includes a graphical user interface (GUI) that allows users to directly modify the mask through annotation using a trackpad. Since manual GUI-based edits are not feasible in the context of (automatic) context replacement, we automate modifications to these masks using morphological operations (such as dilation) and geometric transformations (using the `cv2` library[9]). Specifically, dilation is applied to masked regions to slightly expand or alter their boundaries, while geometric transformations adjust the bone structure and alignment of facial features. The edited mask M' and the original facial image x^f_{out} are then passed to a trained UNet, producing a modified image $x^{f\,\prime}_{out}$.

Face Blending. Finally, the edited facial image $x^{f\,\prime}_{out}$ is seamlessly integrated with the original background x^b_{out} via Poisson blending by solving the Poisson equation:

$$\min_f \iint_\Omega \left| \nabla f - \nabla x^{f\,\prime}_{out} \right|^2 \quad \text{with} \quad f|_{\partial\Omega} = x^b_{out}|_{\partial\Omega}, \quad (13)$$

where Ω denotes the facial region, and $\partial\Omega$ represents the boundary pixels. This blending approach ensures a high-quality reconstruction that preserves fidelity.

This proposed approach, by design, operates independently of input image specifics, ensuring broad applicability. Unlike SDEDIT, which requires a carefully constructed prompt explicitly tailored to the input image, *our proposed* ANTIMIRROR *does not depend on prompts*. Therefore, ANTIMIRROR is applicable to any celebrity likeness (and more generally, any context where the unacceptable concept consists of the identity of a specific person). Furthermore, since ANTIMIRROR does not attempt to erase unacceptable concepts by modifying the weights of the diffusion model, it avoids the challenges associated with identifying and

8. https://github.com/timesler/facenet-pytorch
9. https://github.com/opencv/opencv

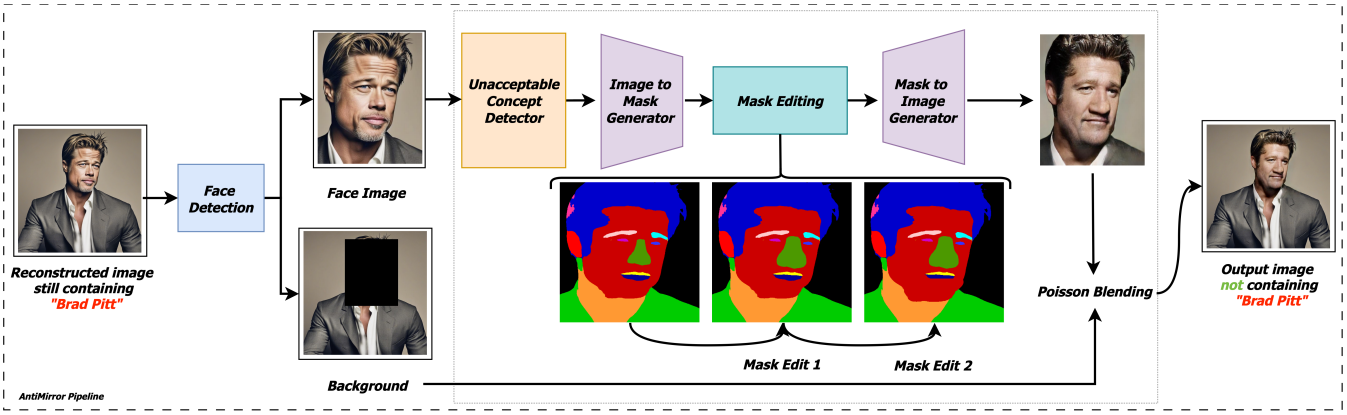


Figure 4: Concept replacement with a targeted editing approach. When a face is detected in the reconstructed image, it is isolated from the background, and passed through an unacceptable concept detector. If the detector flags the image, an image-to-mask generator constructs a mask, which is then edited to modify the face. The modified face is then composited back into the background using Poisson blending.

altering the correct weights — a task shown to be difficult in Sec. 3.5. This systematic pipeline effectively replaces c^u with minimal image alteration, achieving the desired balance between effectiveness and fidelity.

6. AntiMirror Results

We first use CLIP and LPIPS scores to evaluate the effectiveness and fidelity, respectively, of AntiMirror and SDEdit. In order to use SDEdit as CRT, we employed specific prompts for celebrities. For instance, for all male celebrities (*Brad Pitt*, *Elon Musk*, *Donald Trump*, *Joe Biden*), we used the editing prompt "random man", and for female celebrities (*Angelina Jolie*, *Taylor Swift*), we used "random woman". This strategy was adopted so that SDEdit could replace the unacceptable concept with an acceptable one.

We present the results in Table 5 & 8.

6.1. Effectiveness

SDEdit exhibits a high degree of effectiveness, as demonstrated by CLIP values near 0.15 which are significantly below the threshold of 0.25. This clearly validates the effectiveness of SDEdit. Conversely, AntiMirror yields CLIP scores comparable to the None baseline, suggesting limited effectiveness in concept removal.

We revisit the suitability of CLIP scores as a metric for effectiveness of CRTs. The higher CLIP scores for our targeted editing method can be attributed to the inherent characteristics of CLIP itself. CLIP score measure semantic alignment based on global image-text correspondence and are trained to capture broad conceptual coherence [34]. In our targeted editing approach, only specific features within the image are altered rather than the image as a whole. Consequently, CLIP may yield higher scores for targeted edits of celebrity likenesses because targeted edits *intentionally* leave features unrelated to personal identity intact (e.g., a person's typical clothing). A fair metric of CRT effectiveness must be capable of capturing this aspect.

TABLE 5: Effectiveness at removing celebrity likenesses for None (baseline with no CRT applied), SDEdit, and AntiMirror. CLIP scores above 0.25 indicate higher semantic equivalence. We bold the columns with the lowest CLIP score.

Celebrity	None	SDEdit	AntiMirror
Angelina Jolie	0.25 ± 0.02	$\mathbf{0.13 \pm 0.03}$	0.24 ± 0.01
Taylor Swift	0.23 ± 0.02	$\mathbf{0.11 \pm 0.01}$	0.22 ± 0.01
Brad Pitt	0.26 ± 0.02	$\mathbf{0.12 \pm 0.02}$	0.23 ± 0.02
Elon Musk	0.26 ± 0.02	$\mathbf{0.17 \pm 0.02}$	0.27 ± 0.01
Donald Trump	0.24 ± 0.01	$\mathbf{0.15 \pm 0.02}$	0.25 ± 0.01
Joe Biden	0.24 ± 0.02	$\mathbf{0.17 \pm 0.01}$	0.27 ± 0.01

Finding such a metric is difficult in general. However, one possibility can be specialized classifiers that are trained to detect an unacceptable concept (thus ignoring irrelevant features). For celebrity likenesses, such a classifier exists: the Giphy Celebrity Detector[10]. It is a multi-class classifier to identify celebrities in images and has been used in prior work [20]. A GCD top-1 accuracy (GCD) of 0 denotes 0% accuracy in detecting a celebrity, and 1 denotes 100% accuracy. An effective CRT should have low GCD for its outputs. With this new effectiveness metric, we repeated the reconstruction experiment in Sec. 3 for celebrity concepts, but we evaluated the output using GCD, in Table 6.

The outcomes are consistent with those obtained using the CLIP score (see Sec. 3.4); since the GCD values are consistently greater than 0.8, it confirms that the unacceptable concepts remain present, and the state-of-the-art CRTs are ineffective.

Having established GCD as a good metric, we use it to compare AntiMirror and SDEdit (Table 7) show-

10. https://github.com/Giphy/celeb-detection-oss

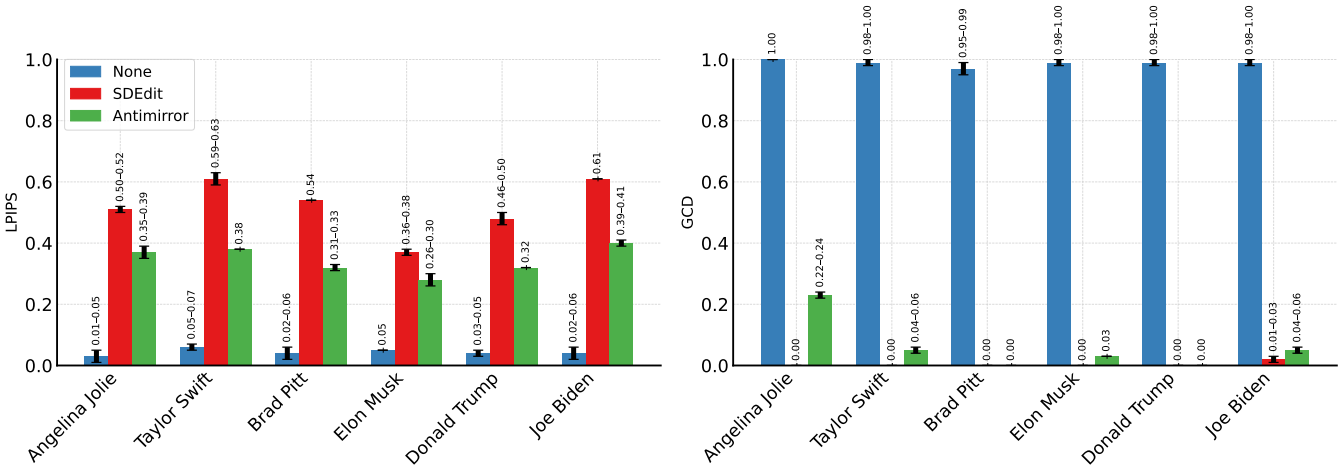


Figure 5: Bar plots comparing NONE, SDEDIT, and ANTIMIRROR using LPIPS and GCD. A smaller value in the LPIPS plot (left) indicates higher fidelity and a smaller value in the GCD plot (right) indicates higher effectiveness. Our method achieves a better trade-off, balancing identity replacement (effectiveness) with fidelity.

TABLE 6: Validating GCD as a metric for CRT effectiveness for celebrity likenesses

Celebrity	NONE	MOD	MACE	UCE
Angelina Jolie	0.92 ± 0.01	0.86 ± 0.02	1.00 ± 0.02	1.00 ± 0.01
Taylor Swift	0.94 ± 0.01	1.00 ± 0.01	0.98 ± 0.02	0.88 ± 0.03
Brad Pitt	0.94 ± 0.02	0.99 ± 0.01	0.99 ± 0.02	0.99 ± 0.03
Elon Musk	0.84 ± 0.03	0.94 ± 0.01	1.00 ± 0.00	1.00 ± 0.01
Donald Trump	0.89 ± 0.00	0.96 ± 0.03	0.96 ± 0.01	0.90 ± 0.01
Joe Biden	0.84 ± 0.01	0.96 ± 0.01	0.97 ± 0.01	0.86 ± 0.02

TABLE 7: GCD comparison across celebrities when no CRT is applied (NONE), and when SDEDIT or ANTIMIRROR are used. GCD identifies celebrities in images by measuring top-1 accuracy, with values between 0 (no celebrity detected, best effectiveness) and 1 (celebrity detected, no effectiveness). We underline the columns where ANTIMIRROR and SDEDIT have comparable GCD values.

Celebrity	NONE	SDEDIT	ANTIMIRROR
Angelina Jolie	1.00 ± 0.00	0.00 ± 0.00	0.23 ± 0.01
Taylor Swift	0.99 ± 0.01	0.00 ± 0.00	0.05 ± 0.01
Brad Pitt	0.97 ± 0.02	0.00 ± 0.00	$\underline{0.00 \pm 0.00}$
Elon Musk	0.99 ± 0.01	0.00 ± 0.00	0.03 ± 0.00
Donald Trump	0.99 ± 0.01	0.00 ± 0.00	$\underline{0.00 \pm 0.00}$
Joe Biden	0.99 ± 0.01	0.02 ± 0.01	0.05 ± 0.01

ing that ANTIMIRROR is significantly effective in removing celebrity likenesses (compared to the baseline NONE) and it is comparable to SDEDIT in most cases. The only case where ANTIMIRROR is significantly less effective than SDEDIT is *Angelina Jolie*. But even in this case, it significantly outperforms other state-of-the-art CRTs from Table 6 (lowest GCD = 0.86). Therefore, we conclude that ANTIMIRROR is an effective CRT.

TABLE 8: LPIPS comparison across celebrities when no CRT is applied (NONE), and when SDEDIT or ANTIMIRROR are used. LPIPS score ranges from 0 (identical) to 1 (maximal difference).

Celebrity	NONE	SDEDIT	ANTIMIRROR
Angelina Jolie	0.03 ± 0.02	0.51 ± 0.01	$\mathbf{0.37 \pm 0.02}$
Taylor Swift	0.06 ± 0.01	0.61 ± 0.02	$\mathbf{0.38 \pm 0.00}$
Brad Pitt	0.04 ± 0.02	0.54 ± 0.00	$\mathbf{0.32 \pm 0.01}$
Elon Musk	0.05 ± 0.00	0.37 ± 0.01	$\mathbf{0.28 \pm 0.02}$
Donald Trump	0.04 ± 0.01	0.48 ± 0.02	$\mathbf{0.32 \pm 0.00}$
Joe Biden	0.04 ± 0.02	0.61 ± 0.00	$\mathbf{0.40 \pm 0.01}$

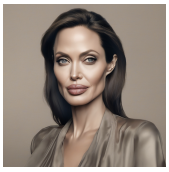
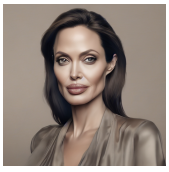
6.2. Fidelity

As we discussed in Sec. 4.2, LPIPS can be used to compare the fidelity of two CRTs. Since NONE will yield a small LPIPS score indicating minimal change, the fidelity of two CRTs can be compared based on how close their LPIPS scores are to the NONE baseline. The fidelity achieved by ANTIMIRROR is notably superior to that of SDEDIT (see Table 8) since the LPIPS scores for ANTIMIRROR are consistently lower than or comparable to SDEDIT. Furthermore, Fig. 5 shows that ANTIMIRROR achieves a better balance between fidelity and effectiveness; the LPIPS bars for ANTIMIRROR are consistently lower than those for SDEDIT, while the GCD bars for ANTIMIRROR and SDEDIT are comparable in most cases.

6.3. Visual Examples Supporting Fidelity of ANTIMIRROR

In Table 9, we present visual examples to illustrate how targeted editing in ANTIMIRROR modifies only the necessary regions—specifically selected facial characteristics—to remove identity, without affecting other acceptable concepts.

TABLE 9: Examples of reconstruction with different CRTs (NONE, SDEDIT, and ANTIMIRROR)

Celebrity (→)	*Angelina Jolie*	*Taylor Swift*	*Brad Pitt*	*Elon Musk*	*Donald Trump*	*Joe Biden*
Input						
CRT (↓)	Output					
NONE						
SDEDIT						
ANTIMIRROR						

The examples also illustrate how SDEDIT sacrifices fidelity for effectiveness, while ANTIMIRROR balances both.

For instance, in the image of *Joe Biden*, SDEDIT not only replaces the identity but also unnecessarily alters the background, hairstyle, and tie color—elements that are acceptable and should not have been changed. In contrast, ANTIMIRROR preserves them, demonstrating superior fidelity while still replacing the celebrity identity.

SDEDIT transforms images of *Taylor Swift* and *Angelina Jolie* into hand-sketched cartoon-like versions. In contrast, ANTIMIRROR carefully retains characteristics unrelated to identity while effectively replacing those that do.

Remarkably, in the case of *Brad Pitt*, SDEDIT introduces unintended artifacts, such as a hat, exaggerates aging effects, closes the eyes, and alters the posture. In contrast, ANTIMIRROR precisely preserves such aspects, demonstrating superior fidelity.

Overall, these visual comparisons clearly illustrate that ANTIMIRROR offers significant advantages over SDEDIT by effectively replacing unacceptable celebrity features while preserving essential, acceptable visual details—thereby ensuring both effectiveness and high fidelity.

7. Related Work

Additional CRTs. We focused on state-of-the-art CRTs that modified weights in the diffusion model's UNet. However, there are other CRTs which alter different parts of the diffusion model. For instance, several of them remove unacceptable concepts from the `CLIP` text encoder of a T2I model. In this way, if a text prompt contains an unacceptable concept, either explicitly or implicitly, it will be adjusted such that the unacceptable concept is replaced. Yoon et al. [40] achieve this by projecting `CLIP` text embeddings away from a latent space of unacceptable concepts as a pre-processing step. Similarly, Wang et al. [36] and Li et al. [15] identify and remove unacceptable concepts from the hidden states of the `CLIP` text encoder when processing prompts. In contrast, Zhang et al. [45] adversarially train the `CLIP` text encoder to remove unacceptable concepts such that texts containing the concept can no longer be used to generate unacceptable images. We do not evaluate these CRTs because in our setting, the prompt supplied to an I2I model is empty (“”). Since these CRTs do not analyze any image data, they are ineffective at preventing reconstruction in the I2I setting.

In addition, there are some CRTs that focus on removing a single class of concepts. Park et al. [25] erase offensive concepts from the weights of the diffusion model by adapting direct preference optimization for unlearning. In contrast, Li et al. [17] focus on modifying the self-attention layers of the UNet to remove offensive concepts. In Sec. 2.4, we opted to analyze three state-of-the-art CRTs configured to support all concept classes in this paper (offensive, copy-

right, and celebrity-likeness). But we confirmed that the CRTs in Park et al. [25] and Li et al. [17], built to replace offensive concepts (*nudity*), also failed to erase it. Due to a lack of space, we present these results in the full version of this paper.

Other Editing Methods. There are other editing methods that can be used for inversion and reconstruction. Brack et al. [3] propose LEDITS++, a state-of-the-art method that, unlike DDPM inversion, solves the DDPM sampling problem using a second-order stochastic differential equation solver, DPM-Solver++ [19], and formulates their inversion process accordingly. We present the results of our experiment with LEDITS++ in the full version of our paper. They are similar to those from DDPM inversion; the metrics from reconstruction with aligned models were nearly identical to metrics from the unaligned model, where no CRT was used.

8. Applications and Extensions to AntiMirror

Filtering more Concepts. Currently, AntiMirror successfully replaces celebrity concepts. In future work, we plan to extend AntiMirror to also address offensive and copyrighted concepts. This would involve isolating the regions of the image associated with these concepts, evaluating them using an unacceptable concept detector, and applying targeted editing to replace them while preserving visual fidelity. In general, as long as the unacceptable concept can be detected via the pixels of the image, then it can be replaced with targeted editing.

Other Qualities of a Good CRT. So far, we have shown that AntiMirror is both effective and fidelity-preserving for unacceptable concepts. However, a good CRT should also *preserve utility*, by allowing acceptable concepts to be generated, and be *robust* against adversaries that attempt to evade it. We believe AntiMirror will also preserve utility, as it relies on both a state-of-the-art face detector and unacceptable concept detector for those faces. Furthermore, to ensure robustness against perturbation-based attacks that aim to evade AntiMirror, adversarial noise purification techniques, such as ADBM [16], could be employed as a pre-processing step.

9. Conclusion

We demonstrated that concept replacement techniques (CRTs) that claim to erase unacceptable concepts fail to do so since images with these concepts are able to be reconstructed in the I2I setting, despite their effectiveness in the T2I setting. Motivated by this, we recognized that the fidelity of corrected unacceptable images is an important property, along with effectiveness, for a good CRT. Hence we approach developing a new CRT, AntiMirror, that is able to better balance these aspects. However, identifying a suitable metric to evaluate fidelity and effectiveness across all concepts remains challenging—highlighted by the reliance on a baseline for `LPIPS` scores and the need for a specialized celebrity classifier when dealing with celebrity-likeness concepts. Our work represents an initial step toward addressing these challenges, laying the groundwork for future advancements.

Acknowledgements: This work is supported in part by the Government of Ontario (RE011-038). Views expressed in the paper are those of the authors and do not necessarily reflect the position of the funding agencies. We thank Thapar School of Advanced AI and Data Science at TIET, Patiala for allowing us to use their compute resources, and our colleagues who provided valuable feedback on previous versions of this paper (Adam Caulfield, Vasisht Duddu, Hossam ElAtali, Tony He, and Asim Waheed).

References

[1] Samyadeep Basu, Nanxuan Zhao, Vlad I. Morariu, Soheil Feizi, and Varun Manjunatha. Localizing and editing knowledge in text-to-image generative models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024.

[2] Bloomberg Law. States Target AI Deepfakes in Taylor Swift Aftermath (Correct). https://news.bloomberglaw.com/artificial-intelligence/state-lawmakers-target-ai-deepfakes-in-taylor-swift-aftermath, 2024. Accessed: 2024-09-15.

[3] Manuel Brack, Felix Friedrich, Katharina Kornmeier, Linoy Tsaban, Patrick Schramowski, Kristian Kersting, and Apolinário Passos. LEDITS++: limitless image editing using text-to-image models. In *CVPR*, pages 8861–8870. IEEE, 2024.

[4] Anudeep Das, Vasisht Duddu, Rui Zhang, and N Asokan. Espresso: Robust concept filtering in text-to-image models. In *CODASPY '25: Proceedings of the Fourteenth ACM Conference on Data and Application Security and Privacy (to appear)*, 2024.

[5] Jose A. Diego-Mas, Felix Fuentes-Hurtado, Valery Naranjo, and Mariano Alcañiz. The influence of each facial feature on how we perceive and interpret human faces. *i-Perception*, 11(5):2041669520961123, 2020. PMID: 33062242.

[6] Rohit Gandikota, Hadas Orgad, Yonatan Belinkov, Joanna Materzynska, and David Bau. Unified concept editing in diffusion models. In *WACV*, pages 5099–5108. IEEE, 2024.

[7] Alvin Heng and Harold Soh. Selective amnesia: A continual learning approach to forgetting in deep generative models. In *NeurIPS*, 2023.

[8] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[9] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020.

[10] Inbar Huberman-Spiegelglas, Vladimir Kulikov, and Tomer Michaeli. An edit friendly DDPM noise space: Inversion and manipulations. In *CVPR*, pages 12469–12478. IEEE, 2024.

[11] Gabriel Ilharco, Marco Túlio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *ICLR*. OpenReview.net, 2023.

[12] Nupur Kumari, Bingliang Zhang, Sheng-Yu Wang, Eli Shechtman, Richard Zhang, and Jun-Yan Zhu. Ablating concepts in text-to-image diffusion models. In *ICCV*, pages 22634–22645. IEEE, 2023.

[13] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *CVPR*, pages 5548–5557. Computer Vision Foundation / IEEE, 2020.

[14] Ruibin Li, Ruihuang Li, Song Guo, and Lei Zhang. Source prompt disentangled inversion for boosting image editability with diffusion models. In *ECCV (26)*, volume 15084 of *Lecture Notes in Computer Science*, pages 404–421. Springer, 2024.

[15] Senmao Li, Joost van de Weijer, Taihang Hu, Fahad Shahbaz Khan, Qibin Hou, Yaxing Wang, and Jian Yang. Get what you want, not what you don't: Image content suppression for text-to-image diffusion models. In *ICLR*. OpenReview.net, 2024.

[16] Xiao Li, Wenxuan Sun, Huanran Chen, Qiongxiu Li, Yingzhe He, Jie Shi, and Xiaolin Hu. ADBM: adversarial diffusion bridge model for reliable adversarial purification. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025.

[17] Xinfeng Li, Yuchen Yang, Jiangyi Deng, Chen Yan, Yanjiao Chen, Xiaoyu Ji, and Wenyuan Xu. Safegen: Mitigating sexually explicit content generation in text-to-image models. In *CCS*, pages 4807–4821. ACM, 2024.

[18] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, pages 3730–3738. IEEE Computer Society, 2015.

[19] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ODE solver for diffusion probabilistic model sampling in around 10 steps. In *NeurIPS*, 2022.

[20] Shilin Lu, Zilan Wang, Leyang Li, Yanzhu Liu, and Adams Wai-Kin Kong. MACE: mass concept erasure in diffusion models. In *CVPR*, pages 6430–6440. IEEE, 2024.

[21] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. In *ICLR*. OpenReview.net, 2022.

[22] Chong Mou, Xintao Wang, Jiechong Song, Ying Shan, and Jian Zhang. Diffeditor: Boosting accuracy and flexibility on diffusion-based image editing. In *CVPR*, pages 8488–8497. IEEE, 2024.

[23] OpenAI. DALL·E 3: OpenAI's Text-to-Image Generation Model. https://openai.com/dall-e-3, 2023. Accessed: 2025-04-14.

[24] Seongbeom Park, Suhong Moon, Seunghyun Park, and Jinkyu Kim. Localization and manipulation of immoral visual cues for safe text-to-image generation. In *WACV*, pages 4663–4672. IEEE, 2024.

[25] Yong-Hyun Park, Sangdoo Yun, Jin-Hwa Kim, Junho Kim, Geonhui Jang, Yonghyun Jeong, Junghyo Jo, and Gayoung Lee. Direct unlearning optimization for robust and safe text-to-image models. In *NeurIPS*, 2024.

[26] Philip Pullella. Pope Francis warns against 'perverse' dangers of AI, renews call for worldwide regulation — theglobeandmail.com. https://www.theglobeandmail.com/business/international-business/article-pope-francis-warns-against-perverse-dangers-of-ai-renews-call-for/. Accessed: 2024-09-15.

[27] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 2021.

[28] Navid Rezaei, Hoshyar Abbasi, Ali Khaksar, and Amin Golshah. Effects of deviations in the nose and chin prominence on facial attractiveness. *Journal of Orthodontics*, 48(2):135–143, 2021. PMID: 33546571.

[29] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10674–10685. IEEE, 2022.

[30] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells III, and Alejandro F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015 - 18th International Conference Munich, Germany, October 5 - 9, 2015, Proceedings, Part III*, volume 9351 of *Lecture Notes in Computer Science*, pages 234–241. Springer, 2015.

[31] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Seyed Kamyar Seyed Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS*, 2022.

[32] Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In Francis R. Bach and David M. Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 2256–2265. JMLR.org, 2015.

[33] Clare A. M. Sutherland, Xizi Liu, Lingshan Zhang, Yingtung Chu, Julian A. Oldmeadow, and Andrew W. Young. Facial first impressions across culture: Data-driven modeling of chinese and british perceivers' unconstrained facial impressions. *Personality and Social Psychology Bulletin*, 44(4):521–537, 2018. PMID: 29226785.

[34] Jingyun Wang, Cilin Yan, and Guoliang Kang. Globality strikes back: Rethinking the global knowledge of CLIP in training-free open-vocabulary semantic segmentation. *arXiv preprint arXiv:2502.06818*, 2025.

[35] Peiran Wang, Qiyu Li, Longxuan Yu, Ziyao Wang, Ang Li, and Haojian Jin. Moderator: Moderating text-to-image diffusion models through fine-grained context-based policies. In *CCS*, pages 1181–1195. ACM, 2024.

[36] Yiming Wang, Jiahao Chen, Qingming Li, Xing Yang, and Shouling Ji. AEIOU: A unified defense framework against NSFW prompts in text-to-image models. *CoRR*, abs/2412.18123, 2024.

[37] Yongliang Wu, Shiji Zhou, Mingzhuo Yang, Lianzhe Wang, Heng Chang, Wenbo Zhu, Xinting Hu, Xiao Zhou, and Xu Yang. Unlearning concepts in diffusion model via concept domain correction and concept preserving gradient. In *AAAI*, pages 8496–8504. AAAI Press, 2025.

[38] Tianwei Xiong, Yue Wu, Enze Xie, Yue Wu, Zhenguo Li, and Xihui Liu. Editing massive concepts in text-to-image diffusion models. *arXiv preprint arXiv:2403.13807*, 2024.

[39] Yijun Yang, Ruiyuan Gao, Xiao Yang, Jianyuan Zhong, and Qiang Xu. Guardt2i: Defending text-to-image models from adversarial prompts. In *NeurIPS*, 2024.

[40] Jaehong Yoon, Shoubin Yu, Vaidehi Patil, Huaxiu Yao, and Mohit Bansal. Safree: Training-free and adaptive guard for safe text-to-image and video generation. *arXiv preprint arXiv:2410.12761*, 2024.

[41] Lingzhi Yuan, Xinfeng Li, Chejian Xu, Guanhong Tao, Xiaojun Jia, Yihao Huang, Wei Dong, Yang Liu, XiaoFeng Wang, and Bo Li. Promptguard: Soft prompt-guided unsafe content moderation for text-to-image models. *arXiv preprint arXiv:2501.03544*, 2025.

[42] Gong Zhang, Kai Wang, Xingqian Xu, Zhangyang Wang, and Humphrey Shi. Forget-me-not: Learning to forget in text-to-image diffusion models. In *CVPR Workshops*, pages 1755–1764. IEEE, 2024.

[43] Qizhe Zhang, Aosong Cheng, Ming Lu, Zhiyong Zhuo, Minqi Wang, Jiajun Cao, Shaobo Guo, Qi She, and Shanghang Zhang. [CLS] attention is all you need for training-free visual token pruning: Make vlm inference faster. *arXiv preprint arXiv:2412.01818*, 2024.

[44] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pages 586–595. Computer Vision Foundation / IEEE Computer Society, 2018.

[45] Yimeng Zhang, Xin Chen, Jinghan Jia, Yihua Zhang, Chongyu Fan, Jiancheng Liu, Mingyi Hong, Ke Ding, and Sijia Liu. Defensive unlearning with adversarial training for robust concept erasure in diffusion models. In *NeurIPS*, 2024.

Appendix

We present the results of our reconstruction experiment for the CRTs PARK ET AL. and LI ET AL. in Table 10. These are configured for offensive concepts, hence we test on *nudity* only.

TABLE 10: Reconstruction results when a CRT (PARK ET AL., LI ET AL.) is applied, compared to the baseline without CRT (NONE). `LPIPS` scores range from 0 (identical) to 1 (maximal different). `Recon` ranges from 0 (every pixel is identical) to $\sqrt{3} * 512$ (every pixel is maximally different). `CLIP` scores at or above 0.25 indicate semantic equivalence [3].

Offensive: Nudity			
CRT	`LPIPS`	`Recon`	`CLIP`
NONE	0.01 ± 0.01	12.19 ± 1.21	0.24 ± 0.01
PARK ET AL.	0.01 ± 0.00	12.21 ± 0.68	0.24 ± 0.01
LI ET AL.	0.01 ± 0.01	12.20 ± 0.79	0.24 ± 0.00

We also conduct our reconstruction experiments with the LEDITS++ inversion method and present the results in Table 11. Like DDPM inversion, he metrics from reconstruction with aligned models were nearly identical to metrics from the unaligned model, where no CRT was used.

TABLE 11: Reconstruction results when a CRT (MOD, MACE, or UCE) is applied, compared to the baseline without CRT (NONE) for LEDITS++. LPIPS ranges from 0 (identical) to 1 (maximal different). Recon ranges from 0 (every pixel is identical) to $\sqrt{3} * 512$ (every pixel is maximally different). CLIP values at or above 0.25 indicate semantic equivalence [3].

Concepts	None			MOD			MACE			UCE		
	LPIPS	Recon	CLIP	LPIPS	Recon	CLIP	LPIPS	Recon	CLIP	LPIPS	Recon	CLIP
Celebrity												
Angelina Jolie	0.01 ± 0.01	16.72 ± 1.93	0.25 ± 0.01	0.01 ± 0.01	16.72 ± 1.54	0.25 ± 0.02	0.01 ± 0.00	16.72 ± 2.12	0.28 ± 0.03	0.01 ± 0.00	16.72 ± 2.89	0.25 ± 0.01
Taylor Swift	0.02 ± 0.01	21.32 ± 1.67	0.24 ± 0.01	0.02 ± 0.00	21.32 ± 1.78	0.24 ± 0.01	0.02 ± 0.01	21.32 ± 1.34	0.24 ± 0.01	0.02 ± 0.00	21.32 ± 1.45	0.24 ± 0.01
Brad Pitt	0.02 ± 0.00	26.45 ± 1.31	0.27 ± 0.01	0.02 ± 0.00	26.44 ± 1.21	0.27 ± 0.00	0.02 ± 0.01	26.44 ± 1.67	0.27 ± 0.01	0.02 ± 0.00	26.44 ± 1.23	0.27 ± 0.02
Elon Musk	0.02 ± 0.01	28.41 ± 1.48	0.28 ± 0.01	0.02 ± 0.00	28.41 ± 1.14	0.28 ± 0.01	0.02 ± 0.00	28.41 ± 1.58	0.28 ± 0.01	0.02 ± 0.00	28.41 ± 1.58	0.28 ± 0.00
Donald Trump	0.02 ± 0.01	32.05 ± 1.15	0.24 ± 0.01	0.02 ± 0.00	32.05 ± 1.13	0.24 ± 0.00	0.02 ± 0.01	32.05 ± 1.17	0.24 ± 0.00	0.02 ± 0.01	32.05 ± 1.12	0.24 ± 0.01
Joe Biden	0.02 ± 0.01	34.21 ± 1.27	0.26 ± 0.01	0.02 ± 0.01	34.21 ± 1.28	0.26 ± 0.01	0.02 ± 0.00	34.21 ± 1.39	0.26 ± 0.01	0.02 ± 0.01	34.21 ± 1.24	0.26 ± 0.01
Offensive												
Nudity	0.01 ± 0.01	10.83 ± 1.84	0.24 ± 0.01	0.01 ± 0.01	10.83 ± 0.64	0.24 ± 0.01	0.01 ± 0.00	10.83 ± 0.75	0.24 ± 0.01	0.01 ± 0.00	10.83 ± 0.89	0.23 ± 0.01
Copyrighted												
Grumpy cat	0.02 ± 0.01	31.67 ± 1.54	0.25 ± 0.01	0.02 ± 0.01	31.67 ± 0.87	0.25 ± 0.01	0.02 ± 0.00	31.67 ± 1.12	0.25 ± 0.01	0.02 ± 0.00	31.67 ± 1.26	0.25 ± 0.00
R2D2	0.02 ± 0.01	36.47 ± 0.64	0.24 ± 0.01	0.02 ± 0.00	36.47 ± 0.69	0.25 ± 0.01	0.02 ± 0.00	36.47 ± 1.12	0.24 ± 0.01	0.02 ± 0.00	36.47 ± 0.89	0.25 ± 0.01