All but One: Surgical Concept Erasing with Model Preservation in Text-to-Image Diffusion Models

Seunghoo Hong*, Juhun Lee*, Simon S. Woo

Department of Artificial Intelligence, Sungkyunkwan University, S. Korea
hoo0681@g.skku.edu, josejhlee@g.skku.edu, swoo@g.skku.edu

Abstract

Text-to-Image models such as Stable Diffusion have shown impressive image generation synthesis, thanks to the utilization of large-scale datasets. However, these datasets may contain sexually explicit, copyrighted, or undesirable content, which allows the model to directly generate them. Given that retraining these large models on individual concept deletion requests is infeasible, fine-tuning algorithms have been developed to tackle concept erasing in diffusion models. While these algorithms yield good concept erasure, they all present one of the following issues: 1) the corrupted feature space yields synthesis of disintegrated objects, 2) the initially synthesized content undergoes a divergence in both spatial structure and semantics in the generated images, and 3) suboptimal training updates heighten the model's susceptibility to utility harm. These issues severely degrade the original utility of generative models. In this work, we present a new approach that solves all of these challenges. We take inspiration from the concept of classifier guidance and propose a surgical update on the classifier guidance term while constraining the drift of the unconditional score term. Furthermore, our algorithm empowers the user to select an alternative to the erasing concept, allowing for more controllability. Our experimental results show that our algorithm not only erases the target concept effectively but also preserves the model's generation capability.

Introduction

Recently, large-scale text-to-image models have demonstrated a remarkable ability to synthesize photo-realistic images (Rombach et al. 2022; Saharia et al. 2022; Ramesh et al. 2021). This rise in generative models was elicited by the joint advancement of algorithms, computing resources, and the curation of large-scale datasets such as LAION (Schuhmann et al. 2022). While these datasets offer rich features for training large-scale models (Brown et al. 2020; Dosovitskiy et al. 2020), many of them are curated with web-scraped material and, thus, lack the necessary preprocessing regarding safety, privacy, and bias (Mehrabi et al. 2021). Moreover, such datasets often contain sexually explicit content, copyrighted material, and personal images. Training generative models using these sensitive data means that the model's generative capability is derived from these same images, and the model is capable of generating such inappropriate content partially or entirely.

To make things worse, due to the stochastic property and

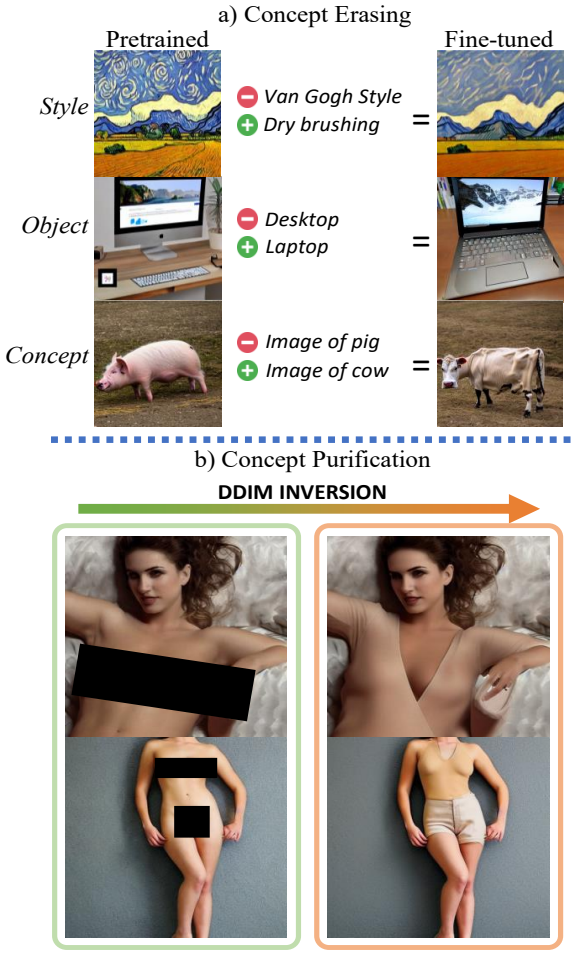

Figure 1: Our method erases a concept while sustaining the model's utility. Additionally, our model can "purify" images through DDIM inversion, which is currently hardly reproducible through other methods.

*These authors contributed equally.

the capability to model complex distributions, there is always a non-zero likelihood that the generated image will contain unsafe content even when conditioned on any unrelated text token. This limits the usability of these generative models in public settings. To alleviate this problem, researchers have inserted an NSFW safe-checking neural network (von Platen et al. 2022). Still, alongside a high false positive rate, their near-unpredictable masking rate of images limits their applicational range, especially when the application relies on a continuous stream of data (Rando et al. 2022). Given these complications, both computation-wise and performance-wise, of retraining these "foundational" models with a heavily curated dataset, researchers have proposed to directly fine-tune foundation models such as Stable Diffusion to erase target concepts (Gandikota et al. 2023; Kumari et al. 2023; Zhang et al. 2023).

While such fine-tuning algorithms for concept ablation are efficient in erasing itself, they significantly sacrifice much of the original generative power of the model to do so. This is far from the original motivation of this line of research. Through a closer examination of the utility aspect of these models, we identify three issues with the current fine-tuning algorithms: 1) Due to the corruption in the feature space of the model, generated images prompted with any arbitrary concepts become unrecognizable or very different from their original concept (see Fig. 2), 2) Generative models such as diffusion models rely on random seeds to output images. As a consequence of fine-tuning the model, the spatial structure and the semantics of the image from the same random seed change. If we regard the model before erasing as the oracle, then any unintended deviation in the output image is not aligned with the ultimate utility of ablating concepts in the model, and 3) despite the model displaying adequate erasing capabilities, certain methods demand a high number of iterations, thereby subjecting the model to increased utility harm. Recent algorithms (Gandikota et al. 2023; Kim et al. 2023) for erasing recommend around 1,000 update steps, which increases the exposure to the issues mentioned above.

In this paper, we aim to address all of the aforementioned challenges. Our main motivation comes from the hypothesis that the task of erasing a concept while preserving the rest requires a *surgical intervention*, where we modify the concept of interest no more than needed. To achieve this, we first inherit from the idea of *classifier guidance*(Ho and Salimans 2022) to decompose the intermediate latent into the unconditional score and the guidance score term and solely apply updates to the latter term. In this region of update, we modify the target concept by introducing supervised and unsupervised erasing guidance, which shows that updating the guidance score is agnostic to the method of erasing guidance supervision method. Moreover, deriving from the Lagrangian Multiplier method, we introduce a regularization on the unconditional score term so that it does not interfere with the update of the guidance score distribution.

Our main contributions are summarized as follows:

- We examine the possible societal and harmful effects of the latest generative models and approaches to mitigate issues via concept erasing, especially focusing on sexually explicit content. We identify that current SoTA algorithms do not consider model utility enough when erasing a concept, and most of them fall short of being used for practice.
- We formulate a fine-tuning algorithm that modifies the core of the target concept while keeping the model intact. Our approach naturally gives rise to a regularization term, where we effectively and safely control the trade-off between erasing strength and model preservation.
- Through extensive experiments, we demonstrate that our surgical approach improves on spatial and semantic consistency, and training efficiency over the current baselines with FID, KID, CLIP, and SSIM scores.

Background

We first describe the essential components used in this line of research before explaining relevant research works.

Diffusion Models

Diffusion models are a class of generative models that learn to reverse the Markov chain diffusion process. Let x_0 represent true data observations and x_t represent intermediate noised data, when $t = T$, corresponding observations x_T are noised to become Gaussian noise. More precisely, diffusion process (Ho, Jain, and Abbeel 2020; Song, Meng, and Ermon 2020) is defined as

$$\begin{aligned} q\left(x_t \mid x_{t-1}\right) &:= \mathcal{N}\left(x_t; \sqrt{\alpha_t} x_{t-1}, \left(1-\alpha_t\right)\mathbf{I}\right) \\ q(x_T|x_0) &\approx \mathcal{N}(x_T; \mathbf{0}, \mathbf{I}) \end{aligned} \tag{1}$$

where α_t is a fixed (Ho, Jain, and Abbeel 2020) or learnable schedule (Sohl-Dickstein et al. 2015). According to Bayes' rule, we can obtain the reverse diffusion, which can be interpreted as an interpolation between x_t and x_0. Then, we can learn to predict this distribution by matching it with a parameterized network and minimizing the KL divergence of the two distributions. The divergence of two Gaussian distributions can be formulated as the mean square error loss. In practice, we reparameterize x_t so that we predict the epsilon ϵ_t (Ho, Jain, and Abbeel 2020) that was used to generate x_t as follows:

$$\mathcal{L}_{\text{diffusion}} = \mathbb{E}_{x_t, t, \epsilon \sim \mathcal{N}(0,1)}\left[\left\|\epsilon - \epsilon_\theta\left(x_t, t\right)\right\|_2^2\right] \tag{2}$$

Text-to-Image Diffusion Models

By diffusing in the latent space of powerful VAEs (Oord, Vinyals, and Kavukcuoglu 2017; Razavi, Van den Oord, and Vinyals 2019) and conditioning these models with text embeddings (Ramesh et al. 2021), they take the form of Latent Diffusion Models (LDM) (Rombach et al. 2022; Saharia et al. 2022) or commonly known as "text-to-image diffusion models". With the addition of these two components, the loss can be formulated as follows:

$$\mathcal{L}_{\text{LDM}} = \mathbb{E}_{z_t \in \mathcal{E}(x), t, c, \epsilon \sim \mathcal{N}(0,1)}\left[\left\|\epsilon - \epsilon_\theta\left(z_t, c, t\right)\right\|_2^2\right] \tag{3}$$

where z_t is the noised latent embedding of image x through a VAE, and c is the text embedding encoded by text encoders such as CLIP (Radford et al. 2021).

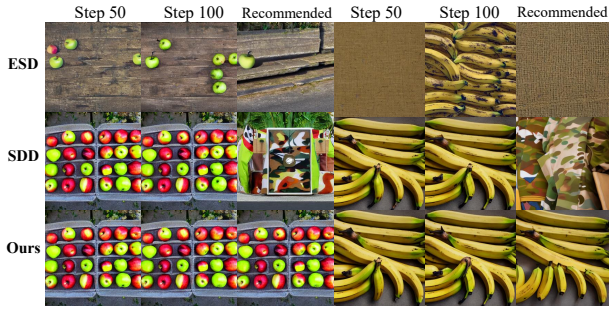

Figure 2: Image erasing timeline: ESD and SDD's images are from iterations 50, 100, and 1000. For our model, we sample from 50, 100, and 400, twice as many steps as we have recommended for the sake of comparison.

Classifier Guidance and Classifier-Free Guidance

It is well known that score $-\sigma_t \nabla_{\boldsymbol{z_t}} \log p(z_t)$ and epsilon $\boldsymbol{\epsilon}_\theta(z_t)$ are equivalent. Then, given that $p_\theta(z_t|c)p_\theta(c|z_t)^\gamma \propto p_\theta(z_t)p_\theta(c|z_t)^{\gamma+1}$ (Ho and Salimans 2022; Song et al. 2020; Dhariwal and Nichol 2021), we can formulate classifier guidance as follows:

$$\begin{aligned}\tilde{\boldsymbol{\epsilon}}_\theta\left(\mathbf{z}_t|\mathbf{c}\right) &= \boldsymbol{\epsilon}_\theta\left(\mathbf{z}_t\right) - (\gamma+1)\sigma_t \nabla_{\mathbf{z}_t} \log p_\theta\left(\mathbf{c} \mid \mathbf{z}_t\right) \\ &\approx -\sigma_t \nabla_{\mathbf{z}_t}\left[\log p\left(\mathbf{z}_t\right) + (\gamma+1)\log p_\theta\left(\mathbf{c} \mid \mathbf{z}_t\right)\right] \\ &= -\sigma_t \nabla_{\mathbf{z}_t}\left[\log p\left(\mathbf{z}_t \mid \mathbf{c}\right) + \gamma \log p_\theta\left(\mathbf{c} \mid \mathbf{z}_t\right)\right]\end{aligned} \tag{4}$$

With classifier-free guidance (CFG) (Ho and Salimans 2022), one can obtain $\nabla_{\mathbf{z}_t} \log p(\mathbf{c} \mid \mathbf{z}_t)$ by composing the scores $\boldsymbol{\epsilon}_\theta(z_t)$ and $\boldsymbol{\epsilon}_\theta(z_t, c)$ as follows

$$\nabla_{\mathbf{z}_t} \log p(\mathbf{c} \mid \mathbf{z}_t) = -\frac{1}{\sigma_t}[\boldsymbol{\epsilon}_\theta(z_t, c) - \boldsymbol{\epsilon}_\theta(z_t)] \tag{5}$$

Ultimately, we can sample an epsilon with guidance scale, γ, as follows:

$$\tilde{\boldsymbol{\epsilon}}_\theta(z_t, c) = (1+\gamma)\boldsymbol{\epsilon}_\theta(z_t, c) - \gamma\boldsymbol{\epsilon}_\theta(z_t) \tag{6}$$

Related Work

One of the early works in erasing fine-tuning is by Gandikota et al. (2023). Their work presents Erased Stable Diffusion (ESD), which updates the student network by mapping its output conditioned on the erasing concept $\epsilon_{\boldsymbol{\theta}}\left(\mathbf{z}_t, \mathbf{c}_s, t\right)$ to the output epsilon conditioned on the erasing concept to the epsilon with negative guidance $\tilde{\epsilon}_{\boldsymbol{\theta}^\star}\left(\mathbf{z}_t, \mathbf{c}_s, t\right)$ from the fixed teacher network. While it delivers substantial erasing capability, it has the tendency to map the erasing concepts to completely non-related concepts and break the spatial and semantic consistency of non-related concepts.

To address these issues, Safe self-Distillation Diffusion (SDD) (Kim et al. 2023) hypothesizes that the training instability of ESD is due to the dependency on the CFG term. Their goal is to map the erasing concept to the null (a.k.a. unconditional) concept directly, without introducing CFG in the supervision signal. Additionally, it incorporates self-distillation, where the teacher is the exponential moving average of the student (Zhang et al. 2019). Despite their impressive erasing results, they both present semantic and spatial corruption, and shifts in the spatial structure before achieving good erasing, as shown in Fig. 2. In Ablation (Kumari et al. 2023), the erasing concept is mapped to a broader "anchor" concept. While their loss is effective, the effect of it leaks to nearby concepts, similar to Dreambooth (Ruiz et al. 2023). Likewise, they also use the Class-Specific Prior Preservation Loss to regularize the language drift (Lee, Cho, and Kiela 2019; Lu et al. 2020) due to the optimization.

In Forget-Me-Not (Zhang et al. 2023), they use the cross-attention layer to erase concepts and directly apply a loss function on those layers. Formally, they penalize the model on the activation of the attention map for the erasing concept token. However, this type of direct manipulation of the internal activations can be detrimental to the model's representation.

Our Approach

Erasing Signal

Notation. We first define the notations used in our work for concept erasing. First, let c and c' be the target erased concept and replacing concept, respectively, containing $c_{\text{text}}, t_{c_{low}}, t_{c_{high}}$; And, γ is the guidance scale; $P_\cdot$ and $\hat{P}_{\cdot,\gamma}$ are the distributions of z; $\emptyset$ represents the unconditional concept. θ and $\theta^\star$ are the parameters to be optimized and the teacher's parameters; λ, T, x_t, z_t are the penalty loss' weight, maximum t, noised images in pixel space, and latent space respectively; $z_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. ϵ_θ and $\mathbf{s}_\theta$ are parameterized networks that predict ϵ and the score. For readability, $P(z_t|\emptyset)$ is expressed as $P(z_t)$ and $\epsilon_\theta(z_t, \emptyset, t)$ as $\epsilon_\theta(z_t)$.

Revisiting the concept of classifier guidance, we utilize the following equation:

$$\nabla \log \hat{P}(z_t|c) = \nabla \log P(z_t|\emptyset) + \gamma \nabla \log P(c|z_t) \tag{7}$$

where $\gamma \nabla \log P(c|z_t)$ is the adversarial gradient (Santurkar et al. 2019) that steers z_t to class c. Now, if all we want is to update the meaning of our target condition, then the update of $\nabla \log P(c|z_t)$ would suffice. In this respect, our loss revolves around this second term as follows:

$$\theta^\star = \arg\min_\theta [\|\gamma_1 \nabla \log P(c'|z_t) - \gamma_2 \nabla \log P(c|z_t)\|_2^2] \tag{8}$$

Moreover, CFG showed that the expression $\nabla \log P(c|z_t)$ can be decomposed as follows: $\nabla \log P(c|z_t) = \nabla \log P(z_t|c) - \nabla \log P(z_t|\emptyset)$. Intuitively, this suggests composability (Du, Li, and Mordatch 2020) that takes the unconditional score $\nabla \log P(z_t|\emptyset)$ to the direction of the class guidance term $\nabla \log P(z_t|c)$.

However, updating $\nabla \log P(z_t|c)$ alone without considering the changes in $\nabla \log P(z_t)$ may harm the overall utility of the model. This might be the case because, while $\nabla \log P_\theta(z_t|c)$ and $\nabla \log P_\theta(z_t)$ are modeled to have fundamentally different properties, they are jointly parameterized by θ, and the change of one can affect the other and vice-versa. If we consider $\nabla \log \hat{P}(z_t|c)$ in Eq. (7), the change in $\nabla \log P(z_t)$ can build up on top of $\nabla \log P(c|z_t)$ and act as an unprovisioned concept. Therefore, the distribution of $\nabla \log P(z_t)$ must remain unchanged to preserve the utility

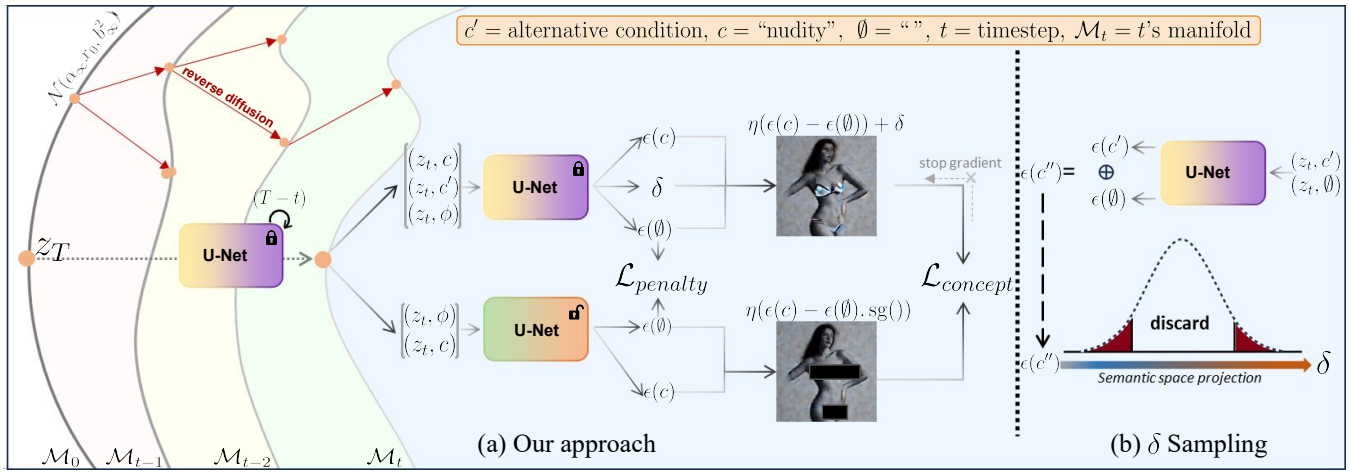

Figure 3: Our method revolves around decomposing the conditional score and updating only one of its term $\nabla \log P(c|z_t)$. Additionally, we incorporate δ into our algorithm, which will both steer our sampling z_t (Kwon, Jeong, and Uh 2022) and be matched by our training model.

of the model. In this respect, the minimization objective of our work is:

$$\min_{\theta} \mathbb{E}_{\mathbf{z},t}[\|\gamma_1 \nabla \log P_{\theta^\star}(c'|z_t) - \gamma_2 \nabla \log P_\theta(c|z_t)\|_2^2]$$
$$\text{s. t.} \nabla \log P_{\theta^\star}(z_t) - \nabla \log P_\theta(z_t) = 0,\ \forall z_t, t = 1, \ldots, T \tag{9}$$

where $c \in \mathbf{C}$, $c' \in \mathbf{C}'$. This type of constraint optimization problem is commonly solvable using the Lagrangian Multiplier. Here, we relax the constraints and optimize Eq. (10) in the following way:

$$\min_{\theta} \mathbb{E}_{c,c',\mathbf{z},t}[\underbrace{\|\gamma_1 \nabla \log P_{\theta^\star}(c'|z_t) - \gamma_2 \nabla \log P_\theta(c|z_t)\|}_{\text{concept loss term}}]$$
$$+\lambda \underbrace{(\|\nabla \log P_{\theta^\star}(z_t|\emptyset) - \nabla \log P_\theta(z_t|\emptyset)\| - \epsilon)}_{\text{penalty term}} \tag{10}$$

where $\lambda \geq 0, \epsilon = 0$, and when $\lambda = 1$, our loss is equivalent to minimizing the upper bound of $||\mathcal{L}_U + \mathcal{L}_C||_2$:

$$\mathbb{E}_{z_t \sim P_{\theta^\star}(z_t|c')}[\mathbf{D_{KL}}(P_{\theta^\star}(z_{t-1}|z_t, c') \| P_\theta(z_{t-1}|z_t, c))] \tag{11}$$

In order to avoid the loss being attributed to $\nabla \log P_\theta(z_t|\emptyset)$, we do not propagate any gradients through it. To do so, a stop gradient is applied to the $\epsilon_\theta(z_t).\text{sg}()$ term. This will prevent the feedback on c from flowing directly to the unconditional term. Ultimately, any feedback on the unconditional is expected to be controlled through the penalty term. In the end, our loss formula is:

$$\mathcal{L}_{\text{model}} = \mathbb{E}_{z_t \sim P_{\theta^\star}(z_t|c'),c,c',t}[\mathcal{L}_{\text{concept}} + \lambda \mathcal{L}_{\text{penalty}}]$$
$$\mathcal{L}_{\text{concept}}(c, c', z_t, \gamma_1, \gamma_2) = \|\gamma_2(\epsilon_\theta(z_t, c) - \epsilon_\theta(z_t).\text{sg}())$$
$$-\gamma_1(\epsilon_{\theta^\star}(z_t, c') - \epsilon_{\theta^\star}(z_t,))\|_2^2$$
$$\mathcal{L}_{\text{penalty}}(t, z_t) = \|\epsilon_\theta(z_t) - \epsilon_{\theta^\star}(z_t)\|_2^2 \tag{12}$$

Algorithm 1: Our training algorithm

Input: Target concept set $\mathbf{C}$, instruction concept list $\mathbf{C_I}$, model weight θ, text encoder $\mathcal{E}$, number of iteration N, number of sampling step T, sampler $\mathbf{P}$, penalty coefficient λ.
Output:

$\theta^\star \leftarrow \theta$, $\mathbf{C}_s \leftarrow \mathcal{E}(\mathbf{C})$
while $N \neq 0$ **do**
  $t \sim \mathcal{U}(\{1, \ldots, T\})$, $c_s \sim \mathbf{C}_s$, $\tau \leftarrow T$, $x_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
  repeat
    $\hat{\epsilon} \leftarrow \epsilon_{\theta^\star}(x_\tau, \emptyset, \tau)$
    $\hat{\epsilon} \leftarrow \hat{\epsilon} + \gamma_1(\epsilon_{\theta^\star}(x_\tau, c_s, \tau) - \epsilon_{\theta^\star}(x_\tau, \emptyset, \tau))$
    $\hat{\epsilon} \leftarrow \hat{\epsilon} + \delta(\mathbf{C_I}, x_\tau, \theta^\star)$
    $x_{\tau-1} \leftarrow \mathbf{P}(x_\tau, \hat{\epsilon}, \tau)$
    $\tau \leftarrow \tau - 1$
  until $\tau = t$
  $\mathcal{L}_{\text{concept}} = \|\gamma_2(\epsilon_\theta(x_t, c) - \epsilon_\theta(x_t, \emptyset).\mathbf{sg}) - (\gamma_1(\epsilon_{\theta^\star}(x_t, c) - \epsilon_{\theta^\star}(x_t, \emptyset)) + \delta(\mathbf{C_I}, x_t, \theta^\star))\|_2^2$
  $\mathcal{L}_{\text{penalty}} = \|\epsilon_\theta(x_t, \emptyset) - \epsilon_\theta^\star(x_t, \emptyset)\|_2^2$
  $\theta \leftarrow \theta - \eta \nabla_\theta(\mathcal{L}_{\text{concept}} + \lambda \mathcal{L}_{\text{penalty}})$
  $N \leftarrow N - 1$
end while
return θ

Search for δ. Let δ be the residual concept for which it transports the original concept c to the alternate concept c'. Put simply, δ is the embodiment of the erasing signal needed to transform c to c'. The challenge is to obtain this erasing signal δ so that $P_{\theta,\phi}(x_{t-1}|x_t, c') = \mathcal{N}(\mu_\theta(x_t) + \gamma \Sigma \nabla \log P_\phi(c|x_t), \Sigma) + \delta$. Here, we present two sampling methods for δ: *implicit* and *explicit*.

Implicit Erasing Signal. These large-scale diffusion models have learned a rich prior with generalizing power. Hertz et al. (2022) shows that when the attention maps in the cross-attention layers are amplified or suppressed, the

token's concept manifestation varies proportionally. When these attention maps are suppressed, the model not only suppresses the erasing concept but also replaces it with other concepts, thanks to its learned prior. We utilize this internal representation of the model to suppress the attention maps of our erasing concept and map to its closest concept. Formally, we sample from x_T to x_t with Prompt-to-Prompt and suppress the respective attention maps of our erasing tokens. Then, we can obtain $\nabla \log P_\phi(c'|x_t)$, which incorporates the "overwriting" concept. We append visual results of this implicit δ in the Supplementary to show its viability.

Explicit Erasing Signal. In practical scenarios, the user may wish to map the erasing concept to an explicitly stated concept. If the sole goal is to overwrite one concept with another concept, we can match the score $\nabla \log P_\phi(c|x_t)$ with $\nabla \log P_\phi(c'|x_t)$. However, even within each concept, there exists a distribution of features/semantics. When we consider modifying a concept, matching the entire source distribution of features to the target distribution is not what we seek. More specifically, we are only interested in the feature mode with the highest density. For example, when we want to replace "bubble guns" with "guns", we do not want to inherit all of the contexts that the word "gun" carries (e.g. "war", "violence"). Instead, we want to solely inherit the "gun" feature itself. Moreover, disruption of the original model will be proportional to the amount of supervision signal we consider using. Now, to ensure that we are utilizing only the most representative feature from the predicted epsilon, we take inspiration from Semantic Guidance (SEGA) (Brack et al. 2023). Formally, SEGA states that the representative semantic information is mainly contained in the highest and lowest pixel values in the predicted epsilon. In this respect, we bottleneck this signal by ablating the values below a percentile as follows:

$$\delta(\mathbf{C_I}, z_t, \theta) = \sum_{c'' \in \mathbf{C_I}} g_{c''} \beta(c'', z_t) \Delta_c(c'', z_t, \theta)$$

$$\beta(c, z_t, \theta) = \begin{cases} 1 & \text{if } \mathbb{1}_{\mathbf{B}_c \bigcap \mathbf{B}_w}(c,t), |\Delta_c| \geq \eta_\kappa(|\Delta_c|) \\ 0 & \text{otherwise} \end{cases}$$

$$\Delta_c(c, z_t, \theta) = -\sqrt{1-\bar{\alpha}}(\nabla \log P_\theta(z_t|c) - \nabla \log P_\theta(z_t))$$

$$\mathbf{B}_c = \{t | t \in \mathbb{Z}, 0 \leq t_{c_{\text{high}}} \leq t \leq t_{c_{\text{low}}} \leq T\}$$

$$\mathbf{B}_w = \{t | t \in \mathbb{Z}, t \geq t_{\text{warmup}}\}$$

where function $\eta_\kappa(\cdot)$ returns κ-th percentile of inputs, and g_c is the guidance scale of concept c that is an elements of instruction concept $\mathbf{C_I}$. The function δ_c should take three arguments, but the notation is omitted at function β. Then, our $\mathcal{L}_{\text{concept}}$ loss is updated as follows:

$$\begin{aligned} \mathcal{L}_{\text{concept}}(c, z_t, \gamma_1, \gamma_2, \mathbf{C_I}) = &\| \gamma_2(\epsilon_\theta(z_t, c) - \epsilon_\theta(z_t) \,.\, \text{sg}()) \\ &- \gamma_1(\epsilon_{\theta^\star}(z_t, c) - \epsilon_{\theta^\star}(z_t)) + \delta(\mathbf{C_I}, z_t, \theta^\star)) \|_2^2 \end{aligned} \tag{13}$$

where $\mathbf{C_I}$ is instruction concept set to make δ. While we attained desirable results with both implicit and explicit supervision, the Prompt-to-Prompt (Hertz et al. 2022) showed considerable sensitivity from the attention map reweighting hyperparameters, which detriments the quality of our sampling ϵ_t^{ptp}. Therefore, most of our experiments are based on the explicit method. The results of using implicit guidance are provided in Suppl. Material. Finally, we present our overall diagram in Fig. 3.

Experimental Results

Experiment Settings

Baselines. We compare the performance of our method with four different latest concept-erasing fine-tuning methods: ESD, SDD, "Ablating" (Kumari et al. 2023), and Forget-Me-Not. Because of the applicability and the utility of a sexual-content censored model, our experiments are centered around erasing "nudity". Nevertheless, we do show that our model can generalize beyond this concept by showing the erasure of concepts, styles, and objects in Fig. 1.a. All of our experiments are performed using the Stable Diffusion ver. 1.4.

Training Setup. For all of our experiments on erasing "nudity", our erasing concept is "nudity", 200 steps of update, the optimizer is AdamW (a learning rate of $1e-5$, $\gamma_1 = \gamma_2 = 7.5$, adam $\epsilon = 1.0e-8$), we use the DDIM ($\eta = 0.0$) sampler with $T = 35$, where we run with GPU A5000, $t_{\text{warmup}} = 5$, $\lambda = 5$.

Evaluation Metrics. We emphasize that our focus is on improving the areas where previous models fall short in terms of the utility of these erased models. In this aspect, our performance evaluation takes into consideration the following aspects: 1) how much the model preserves the remaining concept without degradation, 2) the spatial consistency of the erased and the remaining concepts, 3) how well it erases the target concept, and 4) the training efficiency of a different method. To quantify model preservation, we generate images with MS-COCO captions and calculate the FID (Heusel et al. 2017), and KID (Bińkowski et al. 2018) between the generated images and the actual COCO images. We also use these images to calculate the

Method	NudeNet(%)↓	FID↓	KID↓	CLIP Score↑	SSIM↑
SD v1.4	0.69	13.59	0.00479	0.2765	-
ESD	**0.04**	14.27	**0.00421**	0.2619	0.231
SDD	<u>0.05</u>	14.11	0.00499	0.2677	0.309
Ablating	0.45	<u>13.68</u>	0.00478	<u>0.2756</u>	<u>0.657</u>
Forget-Me-Not	0.66	13.78	0.00496	0.2732	0.476
Ours	0.33	**13.19**	<u>0.00447</u>	**0.2762**	**0.762**
COCO				0.2693	

Table 1: Evaluation metric for best "nudity" erased models. The highest and second-highest scores are printed in bold and underlined, respectively. We treat statistics from both COCO and SD v1.4 datasets as the oracle and attribute ranking among different methods.

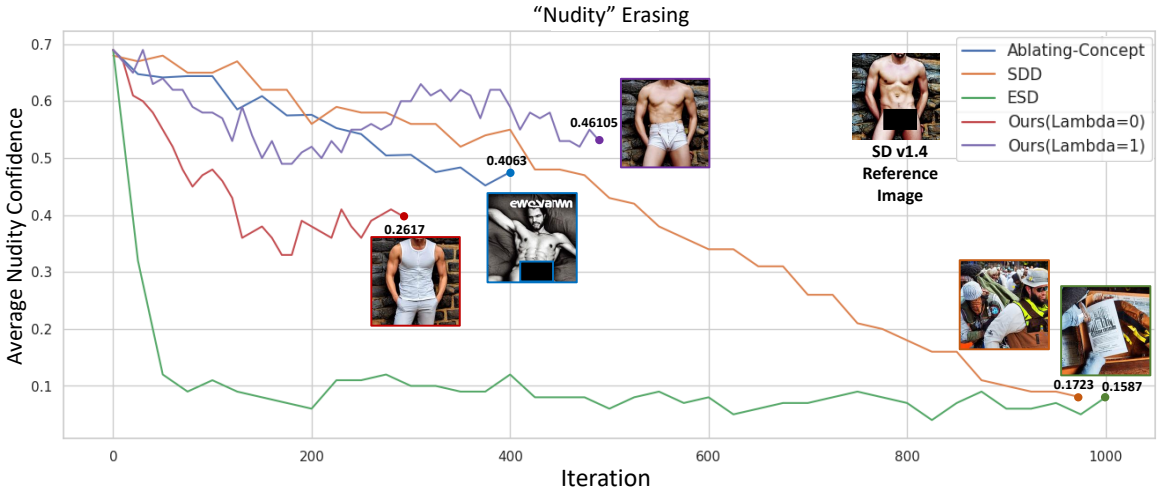

Figure 4: Each model's run end at their recommended iteration stop, and their nudity confidence and SSIM(25x25 window) is reported alongside. The images above share the same seed and prompt at the respective last iteration. While SSD and ESD show low nudity confidence, the seed and the prompt lose their original meaning. Also, due to the high false positives, the decision threshold was set to 0.7. Our model's update decays when the erasing concept is estimated to be erased. Forget-Me-Not returns a static nudity score of 0.66%

CLIP score (Hessel et al. 2021) between the image and the caption to evaluate if the semantic meaning of the images is still intact. However, we find that these are not enough to show that these models do not shift away from their original position. The Structural Similarity Index metric (SSIM) is known to capture these structural elements. For erasure success rate, we show how well the target concept "nudity" is erased through NudeNet (Praneeth, brett koonce, and Ayinmehr 2019)'s confidence score. Lastly, we provide an overviewing assessment of each model's training efficiency.

Results

Model Preservation and Spatial Consistency. Despite competitive erasing, ESD and SDD have shown degradation in image generation, as shown in Fig. 2. In particular, for short prompts, this degradation is even amplified. We hypothesize that this occurs due to the direct matching of arbitrary concepts to the unconditional concept, causing disruption in the semantic space. While this "textualizing" issue is exclusive to ESD and SDD, all models suffer from a shift in the spatial and semantic representation. The semantic representation can be captured by metrics with FID, KID, and CLIP score. However, the spatial consistency is not well captured with these metrics alone.

To this end, we generate 1,000 random objects with the same seed for all fine-tuning methods and calculate the SSIM between the images generated by these methods and by the original checkpoint. Considering the image size, we use a window size of 25x25. As shown in Table 1, while the scores in FID, KID, and CLIP do not show strong variation across models, the SSIM scores show more sensitivity to the spatial structure changes. In addition to the SSIM score, we show the rate of erasure of different models over the iterations in Fig. 4. Here, while "Ablating" and Forget-me-not have shown better spatial consistency, their "nudity" erasing capabilities are quite limited. Finally, we present qualitative results on how our model erases for a given image in Fig. 5.

Training Efficiency. A single assessment of the training efficiency of these models is non-trivial due to their heterogeneous optimization schemes. Firstly, ESD and SDD take 1,000 or more iterations, which can be regarded as inefficient considering the absolute number of iterations. Ablation recommends 200 steps similar to our method, but their erasure is considerably weaker. Lastly, Forget-Me-Not has the fastest training, only requiring 35 steps. Yet, they deliver insufficient erasure of "nudity".

Concept Purification. A natural corollary of the derivation of our objective is that we can tune how much we want to allow the model to "shift" away from its original parameter placing by adjusting λ. An interesting consequence of setting $\lambda = 0$ is that the model gains the ability to erase concepts through image inversion. Formally, we noise a real image with "nudity" and denoise it with our trained model through DDIM inversion (Dhariwal and Nichol 2021). Both inverting using the null token or the concept-related token can erase the concept of the image. We report that our model is the only one that reasonably inherited this property with consistency, as shown in Fig. 1.a.

Hyperparameter λ. We introduce hyperparameter λ, which controls how strongly we want to anchor its unconditional score behavior to its original checkpoint's unconditional score. We train with different λs, where $\lambda = 0$ is the ablated version, as shown in Fig. 6. It is noticeable that there is an inverse proportionality between the model's ability to erase and its spatial constraint. the stronger the constraint is, as in $\lambda = 1.5$, the loss of the lambda saturates over the

Figure 5: Iteration timeline for the same prompt and seed. The first image is the generation with the base checkpoint and each image is 10 iterations apart. While recommended iteration stop is 200, we append the results of iteration 450 at the last image to show spatial consistency even beyond our recommended iteration stop.

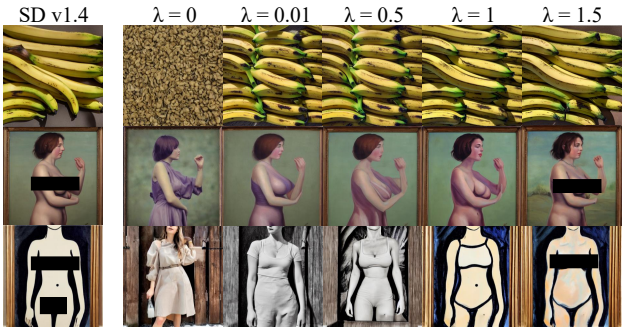

Figure 6: Hyperparameter λ's effect

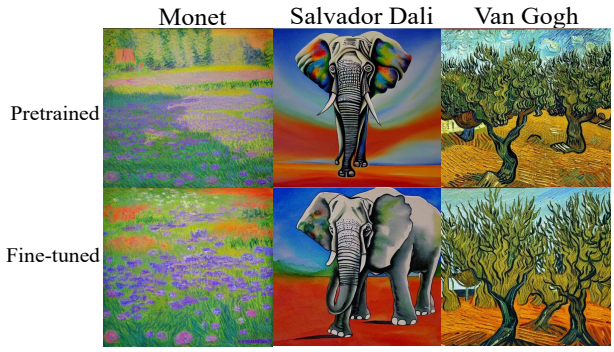

Figure 7: Erasing different styles of painting

erasing signal.Seeing from the lens of the Lagrangian Multiplier Method, we can view the objective as a function of λ but unlike the conventional Lagrangian Multiplier Method, we demonstrate that it is not the optimization of λ that is of interest, but rather the ability to control the learning policy through λ. In our work, this control is illustrated through the introduction of hyperparameter λ, which dictates how strongly we want to anchor its unconditional score behavior to its original checkpoint's unconditional score. We train with different values of λ, where λ=0 is the ablated version, as shown in Fig. 6. Interestingly, there is an inverse proportionality between the model's ability to erase and its spatial constraint. When the constraint is too strong, as in $\lambda = 1.5$, the effect of the lambda overshadows over the erasing signal.

Limitation and Future Work. While our model shows superiority in many aspects, it also has its weaknesses. First, when erasing painting styles, the model either erases most painting styles uniformly or the constraint is too strong and the erasing is too conservative, as shown in Fig. 7. Also, explicit guidance is mostly necessary although there is some minimal effect by subtracting the erasing term itself. In regards to its future work, we argue that this same erasure from the model is a promising type of model personalization that can pave an extension to the notion of controllability in generative models.

Conclusion

In this work, we observe the weaknesses and issues in the current erasing algorithms and revisit the true objective and practical implication behind the task of erasing. The focus on the utility of these "erased" models motivated us to shape our algorithm so that only our concept of interest changes meaning and the rest remains constant. The derivation of our method grants us a hyperparameter to control the strength of the erasing. Owing to this implementation, we address many of the issues presented in current erasing algorithms. We hope our approach can be readily available and practically usable to prevent such unsafe content generation.

Acknowledgments

The authors would thank anonymous reviewers. Seunghoo Hong and Juhun Lee contributed equally. Simon S. Woo is the corresponding author. This work was partly supported by Institute for Information & communication Technology Planning & evaluation (IITP) grants funded by the Korean government MSIT: (No. 2022-0-01199, Graduate School of Convergence Security at Sungkyunkwan University), (No. 2022-0-01045, Self-directed Multi-Modal Intelligence for solving unknown, open domain problems), (No. 2022-0-00688, AI Platform to Fully Adapt and Reflect Privacy-Policy Changes), (No. 2021-0-02068, Artificial Intelligence Innovation Hub), (No. 2019-0-00421, AI Graduate School Support Program at Sungkyunkwan University), and (No. RS-2023-00230337, Advanced and Proactive AI Platform Research and Development Against Malicious Deepfakes).

Ethical Statements

Our model involves nudity and sexually explicit content, but as all models are publicly available, our institution's IRB advised that approval was not required. All researchers involved are over 21 and have carefully reviewed relevant ethics guidelines (NeurIPS 2023; CSET 2021; Goldstein et al. 2023)and undergone training to handle and analyze research results properly. Although no practical defense against creating nudity in generative models exists, we emphasize the urgency of developing preventive technologies given our work's focus on explicit and unsafe content.

References

Bińkowski, M.; Sutherland, D. J.; Arbel, M.; and Gretton, A. 2018. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*.

Brack, M.; Friedrich, F.; Hintersdorf, D.; Struppek, L.; Schramowski, P.; and Kersting, K. 2023. Sega: Instructing diffusion using semantic dimensions. *arXiv preprint arXiv:2301.12247*.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

CSET. 2021. Key Concepts in AI Safety: An Overview. https://cset.georgetown.edu/publication/key-concepts-in-ai-safety-an-overview/. Accessed: 2023-07-07.

Dhariwal, P.; and Nichol, A. 2021. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34: 8780–8794.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Du, Y.; Li, S.; and Mordatch, I. 2020. Compositional visual generation and inference with energy based models. *arXiv preprint arXiv:2004.06030*.

Gandikota, R.; Materzyńska, J.; Fiotto-Kaufman, J.; and Bau, D. 2023. Erasing Concepts from Diffusion Models. In *Proceedings of the 2023 IEEE International Conference on Computer Vision*.

Goldstein, J. A.; Sastry, G.; Musser, M.; DiResta, R.; Gentzel, M.; and Sedova, K. 2023. Generative language models and automated influence operations: Emerging threats and potential mitigations. *arXiv preprint arXiv:2301.04246*.

Hertz, A.; Mokady, R.; Tenenbaum, J.; Aberman, K.; Pritch, Y.; and Cohen-Or, D. 2022. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*.

Hessel, J.; Holtzman, A.; Forbes, M.; Bras, R. L.; and Choi, Y. 2021. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

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

Kim, S.; Jung, S.; Kim, B.; Choi, M.; Shin, J.; and Lee, J. 2023. Towards Safe Self-Distillation of Internet-Scale Text-to-Image Diffusion Models. *arXiv preprint arXiv:2307.05977*.

Kumari, N.; Zhang, B.; Wang, S.-Y.; Shechtman, E.; Zhang, R.; and Zhu, J.-Y. 2023. Ablating concepts in text-to-image diffusion models. *arXiv preprint arXiv:2303.13516*.

Kwon, M.; Jeong, J.; and Uh, Y. 2022. Diffusion models already have a semantic latent space. *arXiv preprint arXiv:2210.10960*.

Lee, J.; Cho, K.; and Kiela, D. 2019. Countering language drift via visual grounding. *arXiv preprint arXiv:1909.04499*.

Lu, Y.; Singhal, S.; Strub, F.; Courville, A.; and Pietquin, O. 2020. Countering language drift with seeded iterated learning. In *International Conference on Machine Learning*, 6437–6447. PMLR.

Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; and Galstyan, A. 2021. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6): 1–35.

NeurIPS. 2023. NeurIPS Code of Ethics. https://nips.cc/public/EthicsGuidelines. Accessed: 2023-07-07.

Oord, A. v. d.; Vinyals, O.; and Kavukcuoglu, K. 2017. Neural discrete representation learning. *arXiv preprint arXiv:1711.00937*.

Praneeth, B.; brett koonce; and Ayinmehr, A. 2019. bedapudi6788/NudeNet: place for checkpoint files.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; and Sutskever, I. 2021. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, 8821–8831. PMLR.

Rando, J.; Paleka, D.; Lindner, D.; Heim, L.; and Tramèr, F. 2022. Red-teaming the stable diffusion safety filter. *arXiv preprint arXiv:2210.04610*.

Razavi, A.; Van den Oord, A.; and Vinyals, O. 2019. Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems*, 32.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; and Aberman, K. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22500–22510.

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E. L.; Ghasemipour, K.; Gontijo Lopes, R.; Karagol Ayan, B.; Salimans, T.; et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35: 36479–36494.

Santurkar, S.; Ilyas, A.; Tsipras, D.; Engstrom, L.; Tran, B.; and Madry, A. 2019. Image synthesis with a single (robust) classifier. *Advances in Neural Information Processing Systems*, 32.

Schuhmann, C.; Beaumont, R.; Vencu, R.; Gordon, C.; Wightman, R.; Cherti, M.; Coombes, T.; Katta, A.; Mullis, C.; Wortsman, M.; et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35: 25278–25294.

Sohl-Dickstein, J.; Weiss, E.; Maheswaranathan, N.; and Ganguli, S. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, 2256–2265. PMLR.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

Song, Y.; Sohl-Dickstein, J.; Kingma, D. P.; Kumar, A.; Ermon, S.; and Poole, B. 2020. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*.

von Platen, P.; Patil, S.; Lozhkov, A.; Cuenca, P.; Lambert, N.; Rasul, K.; Davaadorj, M.; and Wolf, T. 2022. Diffusers: State-of-the-art diffusion models. https://github.com/huggingface/diffusers.

Zhang, E.; Wang, K.; Xu, X.; Wang, Z.; and Shi, H. 2023. Forget-me-not: Learning to forget in text-to-image diffusion models. *arXiv preprint arXiv:2303.17591*.

Zhang, L.; Song, J.; Gao, A.; Chen, J.; Bao, C.; and Ma, K. 2019. Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 3713–3722.