Memories of Forgotten Concepts

Matan Rusanovsky[1*] Shimon Malnick[1*] Amir Jevnisek[1*] Ohad Fried[2] Shai Avidan[1]

[1]Tel Aviv University [2]Reichman University

https://matanr.github.io/Memories_of_Forgotten_Concepts

Abstract

Diffusion models dominate the space of text-to-image generation, yet they may produce undesirable outputs, including explicit content or private data. To mitigate this, concept ablation techniques have been explored to limit the generation of certain concepts. In this paper, we reveal that the erased concept information persists in the model and that erased concept images can be generated using the right latent. Utilizing inversion methods, we show that there exist latent seeds capable of generating high quality images of erased concepts. Moreover, we show that these latents have likelihoods that overlap with those of images outside the erased concept. We extend this to demonstrate that for every image from the erased concept set, we can generate many seeds that generate the erased concept. Given the vast space of latents capable of generating ablated concept images, our results suggest that fully erasing concept information may be intractable, highlighting possible vulnerabilities in current concept ablation techniques.

1. Introduction

Diffusion models have emerged as a prominent tool for text-to-image tasks, extending their importance beyond the research community. Researchers have developed methods to utilize diffusion models for text-guided image editing, increasing their popularity even further. However, it has been demonstrated [38] that these models can generate undesirable content, such as violent and explicit material. This highlights the importance of ablating (*i.e.*, forget or erase) specific concepts (*e.g.*, objects, styles).

A plethora of studies have focused on erasing concepts from diffusion models. Erased concepts are described by text, and the weights of the model are steered away from generating images that are associated with these texts. Then, the ablated model is expected to generate images that do not belong to the population of the erased concept, when introduced with the text describing the erased concept. But, does this mean the concept is erased? Can the model still generate images of the erased concept in some other way?

In our work, forgetting an image means the ablated model can no longer generate it (e.g., a specific church image) with a reasonable likelihood. A more interesting extension is forgetting a concept, which means that an ablated model can no longer produce images that are categorized as belonging to the ablated concept (say, the model can no longer produce any image containing any church) with a reasonable likelihood. Here we take memory to mean that the model can generate an image or a concept, regardless of whether that image was part of the training process, or part of the generalization capabilities of the model.

To date, the analysis of ablated models is mainly done on the output image, as shown in Fig. 1(left). Given an ablated text (*i.e.*, "Starry Night") and a random seed, a model devoid of Van Gogh's style produces an image that is not in the style of Van Gogh. Analysis done on this image will confirm that this is indeed the case. In contrast, we test the following hypothesis:

Hypothesis: *An ablated model should not have a high likelihood seed vector that can be used to generate a high-quality ablated image.*

This paper deals with ways to analyze this hypothesis. In our analysis, we assume that both the ablated text prompt *and* the target ablated image are given. We then measure both the likelihood, in latent space, of the corresponding seed, and the quality of the generated image (Fig. 1(right)). For an effectively erased model, it should not be possible to identify a latent seed that is both likely and yields a high-quality ablated image. However, our analysis shows that the opposite holds true. Models that were ablated using state-of-the-art methods can still generate high-quality ablated images from high-likelihood seeds (Fig. 5).

Technically, to do that, we use *diffusion inversion* to find a latent seed vector that corresponds to the ablated image. We further analyze the identified seed vectors and find that they are as likely, in latent space, as seeds of normal (*i.e.*, non-ablated) images. This suggests that ablated models *do*

[*]Equal contribution.

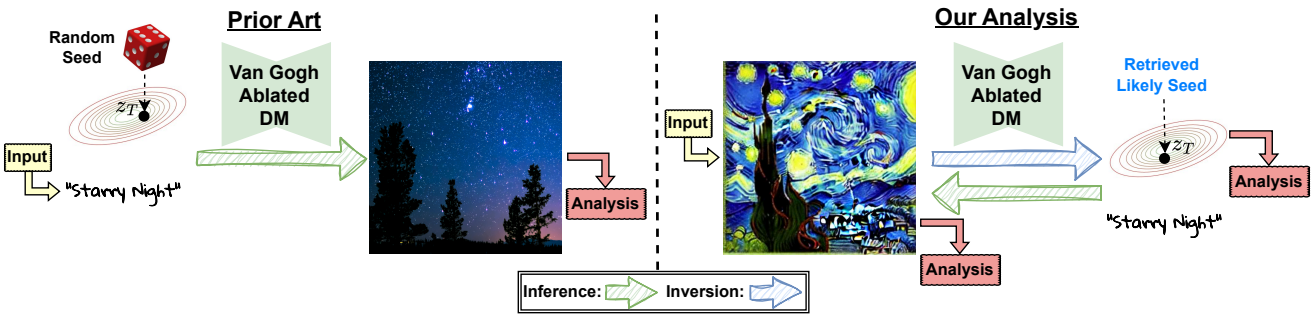


Figure 1. **Evaluation of concept erasure models:** Left part (prior art) analyzes the image generated by an ablated model using the input text (or textual embeddings) and a random seed. Instead, in the right part (our analysis) we assume that both text and ablated image are given as input and analyze the likelihood of the corresponding seed, in the latent space of the model, as well as the quality of the generated image. We find that ablated models contain seeds with high likelihood that can be used to generate high quality ablated images.

not forget ablated concepts. We also find that multiple, distinct, seed vectors can be used to generate an ablated image. We show that by using multiple random *support* images, it is possible to obtain seed vectors with a high likelihood that can generate a given *query* image. These observations suggest that information about the ablated concept persists within the latent space, thereby questioning the effectiveness of concept ablation in these models.

To summarize, we make the following contributions:

- Introduce a metric to *analyze* how much an ablated model remembers erased concepts and images. We demonstrate it on 9 recently published methods and 6 different concepts.
- Show that diffusion inversion of ablated images recovers latent seed vectors with high likelihood and generates images with high PSNR scores.
- Show that a single image can be inverted to multiple *distant* seeds, suggesting that erasing is harder than it looks.

2. Background

2.1. Diffusion models concept erasure

Diffusion models [16, 39] have recently made significant advances in image generation. Further improvements [3, 14, 32, 32] allowed high fidelity text-to-image generation. These models are trained on large-scale datasets containing images from a wide variety of categories. Trained on large-scale datasets spanning diverse categories, these models may later exhibit issues. For example, they can generate not-safe-for-work (NSFW) content, copyrighted images, or private data present in the training set.

A possible way to remove the effect of certain training data on the model is to retrain the model from scratch excluding that data. However, as these issues can recur multiple times on large-scale models, often retraining is infeasible. Moreover, problematic images might be generated even if they are not part of the training set [5].

These concerns raise the need for techniques that can edit a diffusion model to change its outputs w.r.t. given data. Given a pre-trained model, the process of removing the effect of training data from it is referred to as *machine unlearning* [1]. This has been explored vastly for discriminative tasks [2, 11, 12, 41, 43], as the effect of data samples on models prediction is more direct in this case.

As generative models have gained vast popularity recently, many unlearning issues and concerns arise for these models too, *e.g.*, privacy regulations [30] and generation of NSFW content [38]. For image generation tasks, earlier architectures have been examined [18, 24], with more recent studies focusing on diffusion models [4, 7, 8, 13, 20, 23, 38, 44–46]. Schramowski *et al*. [38] propose modifying a model's inference behavior to limit generation of certain data. Other methods [4, 7, 8, 20, 22, 23, 44–46] suggest to finetune the model to reach this goal, or focus on changing the textual embeddings [28, 48].

Recently, there have been works that question and quantify the erasing concepts and abilities that these methods possess. Zhang *et al*. [50] show an attack method against these models, using adversarial prompts that lead to the generation of an erased concept. Unlike this work, we do not propose an attack on concept erasing methods, nor do we aim to find specific prompts that generate concept images. Instead, we invert a concept image to find a suitable latent, showing that the image still lies in the plausible region of the distribution, even after the erasure process.

The closest work to ours is Pham *et al*. [27], examining different concept erasing methods by using textual inversion [6] to find a suitable textual embedding for generating given erased concept images. As opposed to their work, we focus on retrieving z_T latents that can produce the concept image and *analyze their likelihood*.

2.2. Latent Diffusion Models

Diffusion models [16, 39] are generative models that map Gaussian noise x_T to an image x_0 in a gradual denoising

process over multiple timesteps $t \in [0, T]$. This is done by training a learnable neural network $\epsilon_\theta(\cdot, \cdot)$ that learns to reverse a known forward Markov chain with Gaussian noise transitions with predefined parameters α_t. This means that given $\epsilon \sim \mathcal{N}(0, I)$, x_t can be parameterized as:

$$x_t = \sqrt{\alpha_t} x_0 + \sqrt{1 - \alpha_t}\epsilon. \tag{1}$$

And for generation, x_{t-1} can be expressed using the network's output:

$$x_{t-1} = \sqrt{\frac{\alpha_{t-1}}{\alpha_t}} \cdot x_t - \gamma_t(\alpha_t, \alpha_{t-1}) \cdot \epsilon_\theta(x_t, t), \tag{2}$$

where $\gamma_t(\alpha_t, \alpha_{t-1})$ is a noise variance parameter. During training, the model learns to predict the added noise ϵ. This means the loss is:

$$\mathcal{L}_{\text{DM}} := \mathop{\mathbb{E}}_{x,\epsilon,t}\left[\|\epsilon - \epsilon_\theta(x_t, t)\|_2^2\right]. \tag{3}$$

Further advancements [3, 15] have allowed conditioning the generation on textual prompts, allowing the model to receive an additional text c as an input, *i.e.* $\epsilon_\theta(x_t, t, c)$. For faster computing in space with lower complexity than the image space, Rombach *et al.* [32] showed that using a VAE [17] encoder and decoder, denoted as $\text{Enc}(\cdot)$ and $\text{Dec}(\cdot)$, respectively, the memory efficiency of the diffusion process can improve. Instead of training the diffusion process on the high dimensional image x_0, we encode the image to a lower latent space, *i.e.*, $\text{Enc}(x_0) = z_0$, with $\dim(z_0) << \dim(x_0)$. Then, the diffusion process is done in this lower space. This model is denoted as a *Latent Diffusion Model* (LDM). The loss term for training LDMs is thus:

$$\mathcal{L}_{\text{LDM}} := \mathop{\mathbb{E}}_{\text{Enc}(x),\epsilon,t}[\|\epsilon - \epsilon_\theta(z_t, t, c)\|_2]. \tag{4}$$

Generation is done by sampling a latent seed $z_T \sim \mathcal{N}(0, I)$ and using the denoising network $\epsilon_\theta(\cdot, \cdot, \cdot)$ to iteratively compute z_0. Then, the output image is produced by the decoder, *i.e.*, $\hat{\mathcal{I}} = \text{Dec}(z_0)$. In our work, we specifically focus on LDMs, as we use the decoder in some of our analyses. We refer to the process of generating an image from a latent z_T as *diffusion inference*.

2.3. Diffusion model inversion

For diffusion models, inversion is the procedure of finding the latent seed that can be used to generate a given image. As the generation nature of diffusion models is iterative, simple optimizations are too computationally heavy to perform on SOTA models. As DDIM sampling [40] can be used deterministically, DDIM inversion [3] was proposed as a simple way to invert. Subsequent work [26] has shown that this simple method can produce inferior results and proposed a method to invert an image by optimizing for a better null text token embedding. Additional studies [25, 36, 42] have also proposed alternative inversion methods, showing results with very low reconstruction errors. Garibi *et al.* [10] proposed a method that inverts an image by using iterative steps of refinement between the diffusion steps, termed Renoise.

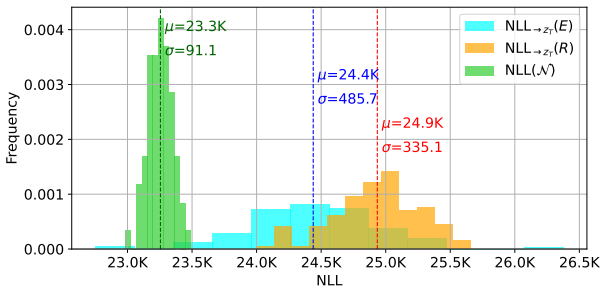


Figure 2. **NLL histogram:** For a model that erased the concept Nudity (EraseDiff [45]), the likelihood distribution fits different Gaussians ($\text{NLL}_{\to z_T}(E)$, $\text{NLL}_{\to z_T}(R)$), that are different from the sampling distribution of the LDM which is standard normal distribution ($\text{NLL}(\mathcal{N})$).

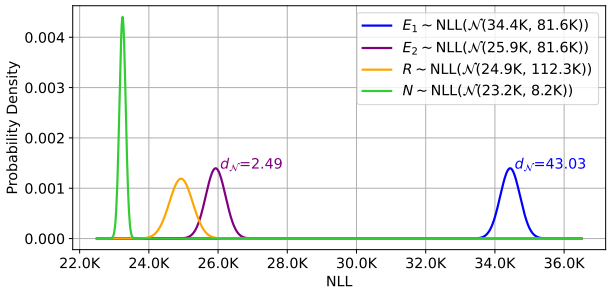


Figure 3. **Visualizing our distance measure:** Our *relative distance* measure is the ratio of $EMD(E, \mathcal{N}))$ to $EMD(R, \mathcal{N}))$, where E is the *erased* set, R is the *reference* set, $\mathcal{N}$ is the normal distribution, and EMD is Earth Movers Distance. As can be seen, the erased model E_1 is much farther than E_2, suggesting that the model that forgot E_1 did a much better job.

3. Analysis

Basic setup. To evaluate a model that erased a given concept c, we require the following:

1. White box access to the LDM model that erased c, denoted ϵ_θ^c.
2. An *erased set* E of (image, caption) pairs with images containing the concept c.
3. A *reference set* R of (image, caption) pairs with images that do not contain the concept c.

Our goal is to analyze and quantify the erasing effect of ϵ_θ^c w.r.t. the concept c. This is done by retrieving a latent vector that, along with ϵ_θ^c, can be used to generate images

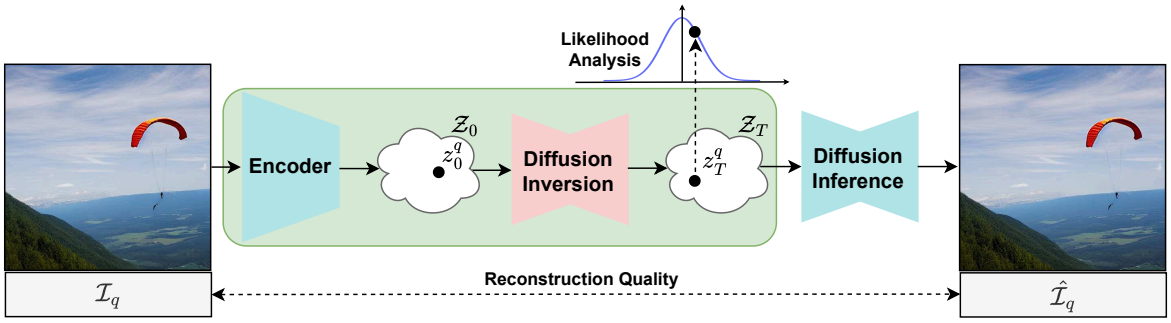
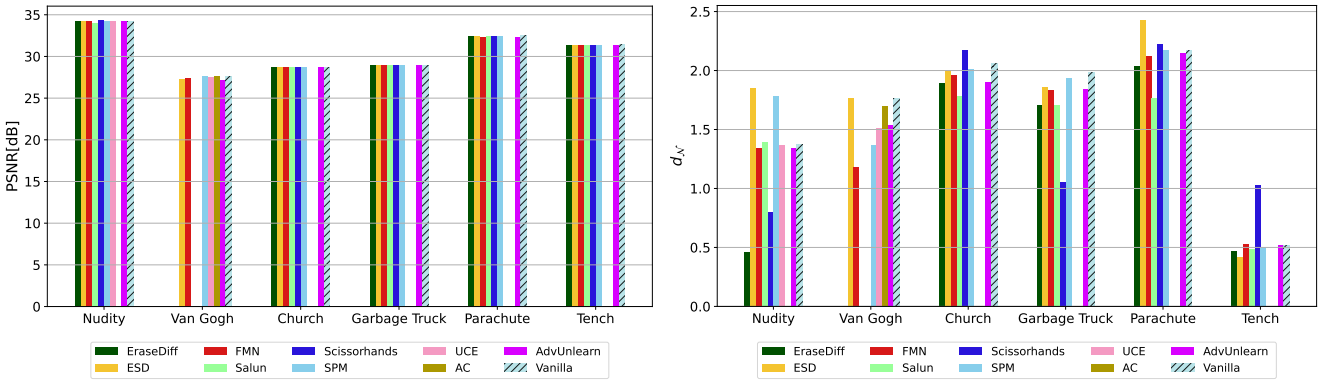


Figure 4. **Memory of an ablated image:** Given an ablated query image $\mathcal{I}_q$, our goal is to find a likely latent z_T that can accurately reconstruct the image when processed through an ablated diffusion model. We start by encoding $\mathcal{I}_q$ into a latent z_0 with the encoder, then apply diffusion inversion to obtain a seed latent vector z_T. This seed is fed into the LDM to generate the image $\hat{\mathcal{I}}_q$. Finally, we evaluate the likelihood of z_T and the quality of the reconstructed image $\hat{\mathcal{I}}_q$ compared to $\mathcal{I}_q$.



(a) PSNR [dB]

(b) $d_{\mathcal{N}}(E, R)$

Figure 5. **A concept erased model remembers:** We report the mean reconstruction PSNR (a) and our proposed relative distance (b) for six concept datasets {Nudity, Van Gogh, Church, Garbage Truck, Parachute, Tench} across nine different concept ablation methods {EraseDiff [45], ESD [7], FMN [47], Salun [4], Scissorhands [44], SPM [23], UCE [8], AC [19], AdvUnlearn [49]}, along with one "Vanilla" SD 1.4 [31] model. These results validate that, at the dataset level, there exists at least one latent per image that can reconstruct the image with high quality (PSNR $\geq$ 25 dB) from a reasonable likelihood using the concept erased model.

containing c. We require that the images have a low reconstruction error, *i.e.*, high PSNR, and analyze the likelihood of the latent vector.

3.1. How do we measure memory?

Given a latent seed z_T, we are interested in examining its likelihood. As the distribution that was used to train the model is Gaussian (see Sec. 2.2), *i.e.*, $\mathcal{N}(0, I)$ or $\mathcal{N}$ in short, computing the likelihood is straightforward via the closed-form probability density function (PDF). We report our results in Negative Log Likelihood (NLL) units, denoting the NLL of a given normal distribution with parameters $(\mu, \sigma^2 I)$ as $\text{NLL}(\mathcal{N}(\mu, \sigma^2 I))$. We follow the Central Limit Theorem to approximate $\text{NLL}(\mathcal{N}(\mu, \sigma^2 \cdot I))$ as 1D Gaussian (see Appx. D for a detailed analysis).

For a set of images, we invert them to latent seeds z_T (see Secs. 3.2 and 3.3) and compute their NLL. We denote this function as $\text{NLL}_{\to z_T}(\cdot)$. We perform this on the erased and reference sets, denoting these distributions as $\text{NLL}_{\to z_T}(E)$ and $\text{NLL}_{\to z_T}(R)$, respectively. These distributions are shown as histograms in Fig. 2, illustrating the separation of latent likelihoods across populations.

We found it difficult to have a clear understanding based on the NLL values alone (as in Fig. 2). Therefore, we opt for a unit-less number that conveys information in relative terms. From a likelihood perspective, a model that erases a concept should ideally map it to a low-likelihood region in z_T space, while preserving non-erased concepts. Therefore, we measure the distance of images in the erased set E to the normal distribution in terms of the distance of the reference set R. Specifically, we use the ratio of the Earth Mover's Distance (EMD) [35] to obtain this measure, denoted *Relative Distance*:

$$d_{\mathcal{N}}(E, R) := \frac{\text{EMD}\left(\text{NLL}_{\to z_T}(E), \text{NLL}(\mathcal{N})\right)}{\text{EMD}\left(\text{NLL}_{\to z_T}(R), \text{NLL}(\mathcal{N})\right)}. \quad (5)$$

The measure $d_{\mathcal{N}}(\cdot,\cdot)$ should approach 1 if the two distances are roughly the same (a value smaller than 1 suggests the erased set is closer to the Normal distribution than the reference set, which indicates that something is wrong with the model). A high $d_{\mathcal{N}}(\cdot,\cdot)$ means the reference set is far more likely than the erased set, which is what we hope for.

Fig. 3 illustrates the two different outcomes of $d_{\mathcal{N}}(\cdot,\cdot)$ for the distributions of different sets, E_1, E_2, R, along with $\mathcal{N}$ for a standard normal distribution (NLL($\mathcal{N}$)). For simplicity, we use different normal distributions for $\mathrm{NLL}_{\to \mathbf{z}_\mathrm{T}}(E_1), \mathrm{NLL}_{\to \mathbf{z}_\mathrm{T}}(E_2), \mathrm{NLL}_{\to \mathbf{z}_\mathrm{T}}(R)$. We see that E_1 is far from both R and $\mathcal{N}$, with very high relative distance of $d_{\mathcal{N}}(R, E_1) = 43.02$. In contrast, E_2 is closer to both these distributions, overlapping R with a low relative distance of $d_{\mathcal{N}}(R, E_2) = 2.49$. The high score of $d_{\mathcal{N}}(R, E_1)$ suggests that the images in E_1 are much less likely for generation while preserving a small distance between the reference set and the standard normal distribution.

Experimental setup. We now show that the information of ablated concepts persists in erased models. To achieve this, we inquire into six different concepts: Nudity, Van Gogh, Church, Garbage Truck, Parachute and Tench. We choose a set of nine different erasure methods, each accompanied by the models used to erase the concepts above, as reported in their original publications: ESD [7], FMN [47], SPM [23] and AdvUnlearn [49] on all concepts. Salun [4], Scissorhands [44], and EraseDiff [45] on all concepts apart from Van Gogh. UCE [8] on Nudity and Van Gogh. AC [19] only on Van Gogh. All models ablate the base Stable-Diffusion v1.4 [33]. The latent space dimension for this model is $4 \times 64 \times 64$. We follow the evaluation protocol of Zhang *et al*. [50] and collect the datasets in the same manner. For NSFW content, we use the I2P dataset [38].

For the reference set R in our analysis, we use images from the COCO [21] dataset. We use Renoise [10] as our primary inversion method, using image captions for guidance with 50 inversion and 5 renoising steps. See Appx. E for details on parameter selection and generalization across architectures and inversion methods.

We consider two types of analyses. The first inquires about the memory of an ablated concept, by aiming to retrieve a single latent to every given query image. The second, exploring many memories of an ablated image, aiming to find multiple seeds that correspond to the same query image. Following previous works in the field, we also report additional metrics regarding the generated images in Appx. C, including CLIP [29] score for prompt-image alignment and concept detection scores.

3.2. Memory of an ablated concept

Equipped with our Relative Distance measure, we show that on a dataset level, concept erasure models can generate the erased concepts with high PSNR and high likelihood. Namely, for every image in the dataset, we can find a likely latent that recovers that image.

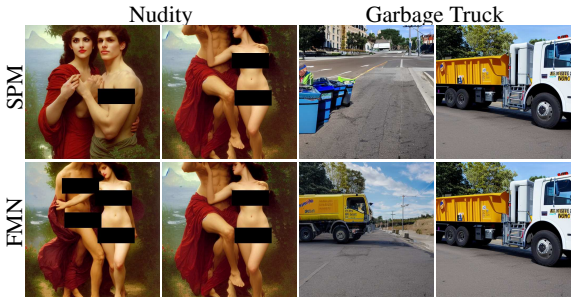


Figure 6. **Erased concepts generations.** Arbitrary latents (odd columns) vs. our retrieved latents (even columns). Each row corresponds to a different ablation method. For all ablation methods, we find some latent that recovers images from the ablated concept.

Given the erased set with images that depict concept c (*e.g.*, various images of churches), namely $E = \{(\mathcal{I}_i, p_i)\}_{i=1}^n$, we use it to analyze a model that was fine-tuned to erase this concept, ϵ_θ^c (see Sec. 3.1).

For every query image $(\mathcal{I}_q, p_q) \in E$, we perform diffusion inversion to retrieve a latent seed z_T^q. This latent is later used for diffusion inference, leading to a reconstructed image $\hat{\mathcal{I}}_q$. Fig. 4 illustrates this procedure. z_T^q and $\hat{\mathcal{I}}_q$ are both used for our analysis.

Results. The results, summarized in Fig. 5, demonstrate how all current concept erasing methods can generate images containing the erased concept. This is suggested by the high PSNR values for reconstruction among all the different methods. Concepts with finer texture details such as Van Gogh and Church, tend to have lower PSNR, while smoother concepts such as Nudity and Parachute, achieve higher reconstruction. For example, Van Gogh's images contain many brushstrokes, while Church's images feature many bricks with distinct contours. In contrast, Nudity images have smooth surfaces, while Parachute images have large areas of background with similar color values.

The right panel of Fig. 5 shows the relative distance ($d_{\mathcal{N}}(\cdot,\cdot)$) of different concepts and concept erasing methods. First, we see that the highest distance, meaning the best distance in terms of erasing, is achieved by ESD [7] on the concept c = Parachute, with a distance of $d_{\mathcal{N}}(E_c, R_c) = 2.49$. This score indicates that there exists a non-negligible overlap between the distribution of the erased concept and the reference dataset. Please refer to Fig. 3 for a visualization of a 2.49 distance.

Moreover, compared to the Vanilla model which did not erase the given concept, many methods achieve a similar relative distance $d_{\mathcal{N}}(\cdot,\cdot)$. This suggests that while these

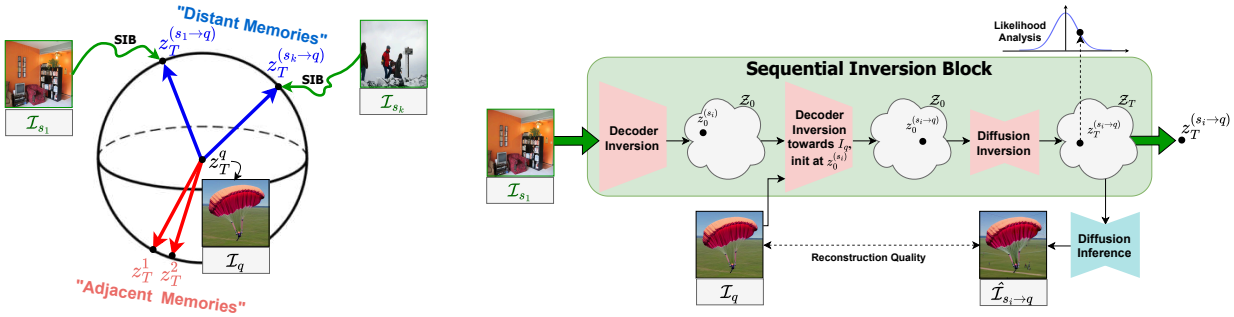


Figure 7. **The many memories of an ablated image:** (Left) The latent seed z_T^q of the query image $\mathcal{I}_q$ can be obtained from various support images: $\mathcal{I}_{s_1}, ..., \mathcal{I}_{s_k}$. For support image $\mathcal{I}_{s_i}$, we apply the sequential inversion block shown on the right to map it to the seed $z_T^{(s_i \to q)}$. We show in Fig. 5 that seeds $\{z_T^{(s_i \to q)}\}_{i=1}^k$ are likely enough and can be used to generate the query image $\mathcal{I}_q$. (Right) Recovering the seed of a query image $\mathcal{I}_q$ when starting with support image $\mathcal{I}_{s_i}$.

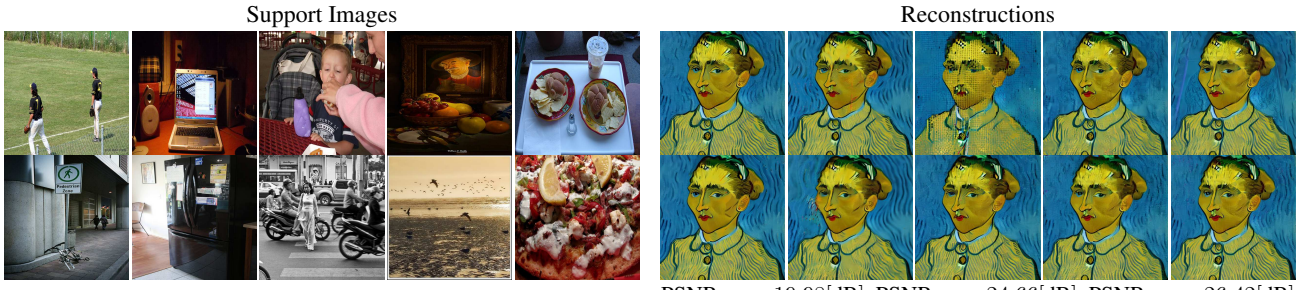


Figure 8. **Reconstruction from different seeds:** Each support image on the left was used to reconstruct via the Sequential Inversion Block (see Sec. 3.3) the image in the corresponding location on the right. As can be seen, vastly different support images lead to (almost) the exact same reconstructed image. These images were generated from an ESD [7] model that ablated the Van Gogh concept.

methods exemplify that it is harder to generate images of the erased concept using text prompts that describe it, the generation of such images in the latent space is still plausible. Observe that in some scenarios the distance is lower than 1, which seems to suggest that these models erased a concept via a text proxy but did not actually forget it.

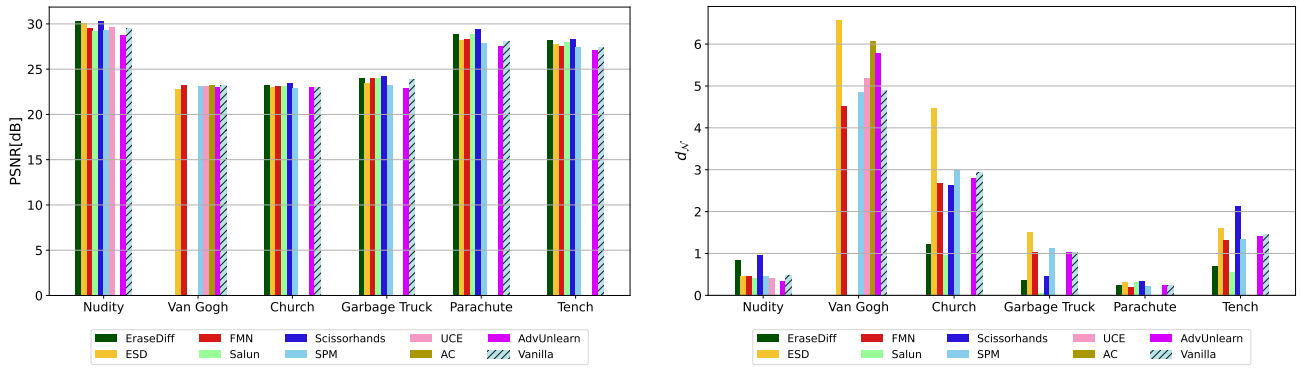
Fig. 6 shows images of multiple erasing methods and concepts, comparing generation with a fixed seed to one retrieved from our analysis. Although generation with arbitrary latents may indicate forgetting, we show that for all model ablation techniques, there exists a likely latent that recovers images from the ablated concept. A complete comparison of all models and concepts is presented in Appx. C.

3.3. The many memories of an ablated image

In the previous subsection, we analyzed the case of erasing a concept using multiple images, by finding a feasible latent z_T that can be used to generate an image that depicts the concept. However, this raises the question: for a given image $\mathcal{I}_q$, is there more than one distinct latent seed z_T^q that can generate an image that resembles $\mathcal{I}_q$? Specifically, we are interested in whether we can find several z_T latents with a sufficiently large cosine distance between them, that all can be used to generate the query image $\mathcal{I}_q$. These latent vectors should satisfy two main requirements: they should be likely (in terms of the model's probability distribution) and should be well-separated from each other. Specifically, we are interested in distant memories of the query image, and not adjacent ones. Fig. 7 illustrates our approach.

Sequential Inversion Block. We seek distinct "*memories*" of the same query image. To do that, we start with random support images. (A detailed analysis of alternative initialization choices beyond images can be found in Appx. B.) For each support image, we invert the VAE decoder Dec$(\cdot)$, to obtain an initial latent vector z_0 which is then used for an optimization process w.r.t. the query image. Similarly to Sec. 3.2, the reconstructed query images are fed to the diffusion inversion to produce the desired z_T seeds, which we call distinct "memories".

Formally, we introduce the *Sequential Inversion Block*, which maps support images $\mathcal{I}_{s_i}$ from the image space to the latent space. This is done using the following sequential



(a) PSNR [dB]

(b) $d_{\mathcal{N}}(E, R)$

Figure 9. **Distant latents reconstruct erased images:** We report the mean reconstruction PSNR (a) and our proposed relative distance (b), across models and concepts (see Fig. 5 for more details), obtained using our *sequential inversion block* process resulting in different distant latents for each image.

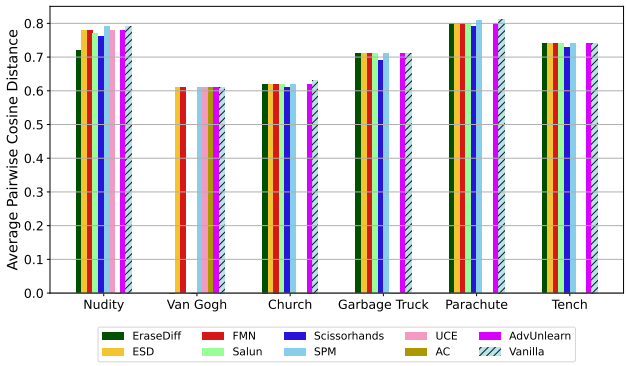


Figure 10. **Average Pairwise Cosine Distance:** For each model that ablated each concept, and for each target image $\mathcal{I}_q$, we average the pairwise cosine distance (1 − cosine similarity) between all the produced $z_T^{(s_i \to q)}$ seed latents. Then, we average the results over all target images per each model and concept.

inversion steps:

1. **Initial Decoder Inversion:**
   Find an initial latent, $z_0^{(s_i)}$, in the decoder's latent space, that will serve as a starting point for the next step in the block. Specifically, invert the VAE decoder $\mathrm{Dec}(\cdot)$, starting from $\mathrm{Enc}(\mathcal{I}_{s_i})$, and optimize the following:

$$z_0^{(s_i)} = \underset{z}{\operatorname{argmin}} \left[\mathrm{Dist}\left(\mathrm{Dec}\left(z\right), \mathcal{I}_{s_i}\right)\right], \tag{6}$$

   where $\mathrm{Dist}(\cdot,\cdot)$ is the euclidean distance.
2. **Decoder Inversion Towards the Query Image:** Next, starting from an initial latent $z_0^{(s_i)}$ (which corresponds to the support image $\mathcal{I}_{s_i}$), we optimize to find a latent that reconstructs the query image $\mathcal{I}_q$:

$$z_0^{(s_i \to q)} = \underset{z}{\operatorname{argmin}} \left[\mathrm{Dist}\left(\mathrm{Dec}\left(z\right), \mathcal{I}_{q}\right)\right]. \tag{7}$$

3. **Latent Diffusion Inversion:** We use $z_0^{(s_i \to q)}$ as a starting point for a diffusion model inversion process, resulting in a latent seed $z_T^{(s_i \to q)}$ which is the output of the *Sequential Inversion Block*.

The retrieved $z_T^{(s_i \to q)}$ latent will also be used to generate an image $\hat{\mathcal{I}}_q$ that resembles $\mathcal{I}_q$ for reconstruction quality analysis. In addition, we analyze the likelihood of $z_T^{(s_i \to q)}$ and measure the average pairwise cosine distances between the generated $z_T^{(s_i \to q)}$, for all support images $\mathcal{I}_{s_i}$, to ensure we found distant seed vectors in latent space that represent the same $\mathcal{I}_q$ image.

Experimental Setting: We extend the experimental setup in Sec. 3.1. We randomly select support images from the COCO [21] dataset. For each concept, we randomly choose five query images and validate that for every query image, there are at least ten distinct latents. Each of these latents is initialized using a different support image. We set the number of VAE decoder optimization steps to 3,000.

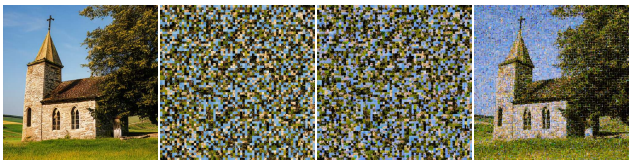
Results: A qualitative example of reconstructed images that were produced from this procedure is presented in Fig. 8, demonstrating how different latents can generate images of an erased concept. Fig. 9 demonstrates the results, showing that for all methods and all concepts we were able to recover likely latent seed vectors, *i.e.*, low relative similarity (Eq. (5)) that lead to a high-quality reconstruction (high PSNR). However, compared to the concept level forgetting, (see Fig. 5), we see a lower PSNR value and lower $d_{\mathcal{N}}(\cdot,\cdot)$ values, when multiple diverse z_T seeds are searched for. We conjecture that the search for multiple latents seeds is sub-optimal to finding a single latent in an unconstrained setting

Additionally, we measure the distances between all latents within a concept. Specifically, we evaluate the cosine

distance and euclidean distance (in Appx. C) between all $z_T^{(s_i \to q_j)}$ latents associated with a given concept. This metric provides insight into the distribution of latents. These distances are presented in Fig. 10.

Geometric interpretation of the retrieved memories: To better understand how the retrieved $z_T^{(s_i \to q)}$ seeds are distributed geometrically in space with respect to the original target z_T^q seed, we compute the average euclidean distance between all $z_T^{(s_i \to q)}$ and z_T^q. Specifically, we refer to the illustration presented on the left part of Fig. 7. We observe that all these distances are tightly spread around the mean distance. Concretely, the mean is 152.14 and the standard deviation is 2.72. This leads to a coefficient of variation of 2%. We extend the computation of the coefficient of variation to all query images in all of our experiments, and observe that the mean and standard deviation of the coefficients are 2% and 1%, respectively. Together with the observation that there exists a substantial cosine distance between these seeds (Fig. 10), we conclude that for each target image, our procedure produces memories that lie (with a high probability) on a sphere centered around a z_T seed that corresponds to that image.

Following this geometric insight, one could choose any number of N_S support images, and retrieve N_S seeds using SIB. This raises following questions for future *forgetting* work: Will re-mapping all these possible seeds into images that do not resemble $\mathcal{I}_q$, be *sufficient* for forgetting? Is it *necessary* for forgetting? in Appx. F, we explore this question in more detail.



(a) 0.3245 (b) 0.1879 (c) 0.1737 (d) 0.2906

Figure 11. **Ablated models generalize to shuffled images:** For a diffusion model that ablates the concept Church, we take a church image in (a), split it to patches of shape 8×8 and shuffle them to obtain image (b). Then, we invert the image in (b) and regenerate it to obtain the reconstructed image in (c). Finally, we revert the shuffle of patches to obtain the image at (d). Below each image we report the CLIP score of that image w.r.t. the text "church".

4. Limitations

The analysis in this paper assumes a white-box setting, assuming access to the model's weights and the ability to inverse it. Although this may limit the generalization of our findings, it enables a controlled exploration of how well-erased concepts can be reconstructed within the model.

In Fig. 11, we show that even when an image associated with the ablated concept Church is scrambled, inverted, and then reassembled, the model retains certain associations with the original concept. Starting with a church image (Fig. 11a), we shuffle its patches to create (Fig. 11b), invert the scrambled version to produce (Fig. 11c), and then reassemble the patches in (Fig. 11d). Although the concept classifier score for "church" drops significantly (from 0.99 to approximately 10^{-4}), the CLIP similarity to the caption "church" decreases by only 10%.

While models are not expected to "forget" scrambled versions of erased concepts, this result highlights a significant concern: diffusion models may generalize well even to unusual, pixelated images (such as Fig. 11b), successfully inverting them despite their atypical structure. This generalization ability appears to conflict with concept erasure, as models may inadvertently retain latent representations of ablated concepts. While it may seem that inversion is too powerful for this task (see Appx. A for further details), we notice that applying our analysis gives coherent results, as the likelihood of the seed that results from the shuffled image is lower, and the PSNR is worse. Specifically, the NLL of the retrieved z_T seed for the shuffled image (Fig. 11b) is 23.89K, compared to 23.05K for the original church image (Fig. 11a). The recovered noise has low reconstruction quality (Fig. 11c), with a PSNR of only 15 dB relative to the shuffled image. When reconstruction deviates from the query, the likelihood analysis reflects reconstruction statistics rather than the original query. Thus, reconstruction PSNR aligns with the successful evaluation of retrieved images' likelihoods.

5. Conclusions

As diffusion models become more accessible and common to the public, the importance of the safety and privacy of these models increases. Recent papers address this concern, developing essential methods for editing diffusion models' outputs to ensure a safer and more controlled generation. Previous methods sought to limit the generative capabilities of specific concepts by disrupting the ability to generate these concepts through *descriptive text*. In this work, we hypothesize that an ablated model should not have a high likelihood *seed vector* that can be used to generate a high-quality ablated image. We show, across many methods and different categories, that previous attempts did not truly erase concepts. We do so, by introducing an analysis on the reconstruction quality of images from the erased concepts, and on the likelihood of its corresponding latent seeds. We hope our proposed analysis encourages further research on reliable concept erasure evaluation.

Acknowledgments. This project was partially supported by ISF grants 1574/21 and 2132/23.

References

[1] Lucas Bourtoule, Varun Chandrasekaran, Christopher A Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. Machine unlearning. In *2021 IEEE Symposium on Security and Privacy (SP)*, pages 141–159. IEEE, 2021. 2

[2] Min Chen, Zhikun Zhang, Tianhao Wang, Michael Backes, Mathias Humbert, and Yang Zhang. When machine unlearning jeopardizes privacy. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2021. 2

[3] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 2, 3

[4] Chongyu Fan, Jiancheng Liu, Yihua Zhang, Eric Wong, Dennis Wei, and Sijia Liu. Salun: Empowering machine unlearning via gradient-based weight saliency in both image classification and generation. *arXiv preprint arXiv:2310.12508*, 2023. 2, 4, 5

[5] Internet Watch Foundation. https://www.iwf.org.uk/media/q4zll2ya/iwf-ai-csam-report_public-oct23v1.pdf. 2

[6] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 2

[7] Rohit Gandikota, Joanna Materzynska, Jaden Fiotto-Kaufman, and David Bau. Erasing concepts from diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2426–2436, 2023. 2, 4, 5, 6

[8] Rohit Gandikota, Hadas Orgad, Yonatan Belinkov, Joanna Materzyńska, and David Bau. Unified concept editing in diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5111–5120, 2024. 2, 4, 5

[9] Daiheng Gao, Shilin Lu, Shaw Walters, Wenbo Zhou, Jiaming Chu, Jie Zhang, Bang Zhang, Mengxi Jia, Jian Zhao, Zhaoxin Fan, et al. Eraseanything: Enabling concept erasure in rectified flow transformers. *arXiv preprint arXiv:2412.20413*, 2024.

[10] Daniel Garibi, Or Patashnik, Andrey Voynov, Hadar Averbuch-Elor, and Daniel Cohen-Or. Renoise: Real image inversion through iterative noising. *arXiv preprint arXiv:2403.14602*, 2024. 3, 5

[11] Aditya Golatkar, Alessandro Achille, and Stefano Soatto. Eternal Sunshine of the Spotless Net: Selective Forgetting in Deep Networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9301–9309, Seattle, WA, USA, 2020. IEEE. 2

[12] Chuan Guo, Tom Goldstein, Awni Hannun, and Laurens van der Maaten. Certified Data Removal from Machine Learning Models, 2020. arXiv:1911.03030 [cs, stat]. 2

[13] Alvin Heng and Harold Soh. Selective amnesia: A continual learning approach to forgetting in deep generative models. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[14] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 2

[15] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 3

[16] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, pages 6840–6851. Curran Associates, Inc., 2020. 2

[17] Diederik P Kingma and Max Welling. Stochastic gradient vb and the variational auto-encoder. In *Second International Conference on Learning Representations, ICLR*, page 121, 2014. 3

[18] Zhifeng Kong and Kamalika Chaudhuri. Data redaction from pre-trained gans. In *2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, 2023. 2

[19] Nupur Kumari, Bingliang Zhang, Sheng-Yu Wang, Eli Shechtman, Richard Zhang, and Jun-Yan Zhu. Ablating concepts in text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22691–22702, 2023. 4, 5

[20] Nupur Kumari, Bingliang Zhang, Sheng-Yu Wang, Eli Shechtman, Richard Zhang, and Jun-Yan Zhu. Ablating concepts in text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22691–22702, 2023. 2

[21] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 5, 7

[22] Shilin Lu, Zilan Wang, Leyang Li, Yanzhu Liu, and Adams Wai-Kin Kong. Mace: Mass concept erasure in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6430–6440, 2024. 2

[23] Mengyao Lyu, Yuhong Yang, Haiwen Hong, Hui Chen, Xuan Jin, Yuan He, Hui Xue, Jungong Han, and Guiguang Ding. One-dimensional adapter to rule them all: Concepts diffusion models and erasing applications. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7559–7568, 2024. 2, 4, 5

[24] Shimon Malnick, Shai Avidan, and Ohad Fried. Taming normalizing flows. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 4644–4654, 2024. 2

[25] Daiki Miyake, Akihiro Iohara, Yu Saito, and Toshiyuki Tanaka. Negative-prompt inversion: Fast image inversion for editing with text-guided diffusion models. *arXiv preprint arXiv:2305.16807*, 2023. 3

[26] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6038–6047, 2023. 3

[27] Minh Pham, Kelly O. Marshall, Niv Cohen, Govind Mittal, and Chinmay Hegde. Circumventing concept erasure meth-

ods for text-to-image generative models. In *The Twelfth International Conference on Learning Representations*, 2024. 2

[28] Minh Pham, Kelly O Marshall, Chinmay Hegde, and Niv Cohen. Robust concept erasure using task vectors. *arXiv preprint arXiv:2404.03631*, 2024. 2

[29] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 5

[30] General Data Protection Regulation. https://gdpr-info.eu/. 2

[31] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022. 4

[32] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2, 3

[33] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 5

[34] L Rout, Y Chen, N Ruiz, C Caramanis, S Shakkottai, and W Chu. Semantic image inversion and editing using rectified stochastic differential equations. In *The Thirteenth International Conference on Learning Representations*, 2025.

[35] Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. The earth mover's distance as a metric for image retrieval. *International journal of computer vision*, 40:99–121, 2000. 4

[36] Dvir Samuel, Barak Meiri, Nir Darshan, Shai Avidan, Gal Chechik, and Rami Ben-Ari. Regularized newton raphson inversion for text-to-image diffusion models. *arXiv e-prints*, pages arXiv–2312, 2023. 3

[37] Dvir Samuel, Rami Ben-Ari, Nir Darshan, Haggai Maron, and Gal Chechik. Norm-guided latent space exploration for text-to-image generation. *Advances in Neural Information Processing Systems*, 36, 2024.

[38] Patrick Schramowski, Manuel Brack, Björn Deiseroth, and Kristian Kersting. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22522–22531, 2023. 1, 2, 5

[39] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 2256–2265, Lille, France, 2015. PMLR. 2

[40] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 3

[41] Ryutaro Tanno, Melanie F Pradier, Aditya Nori, and Yingzhen Li. Repairing neural networks by leaving the right past behind. *arXiv preprint arXiv:2207.04806*, 2022. 2

[42] Bram Wallace, Akash Gokul, and Nikhil Naik. Edict: Exact diffusion inversion via coupled transformations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22532–22541, 2023. 3

[43] Ga Wu, Masoud Hashemi, and Christopher Srinivasa. Puma: Performance unchanged model augmentation for training data removal. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8675–8682, 2022. 2

[44] Jing Wu and Mehrtash Harandi. Scissorhands: Scrub data influence via connection sensitivity in networks. *arXiv preprint arXiv:2401.06187*, 2024. 2, 4, 5

[45] Jing Wu, Trung Le, Munawar Hayat, and Mehrtash Harandi. Erasediff: Erasing data influence in diffusion models. *arXiv preprint arXiv:2401.05779*, 2024. 3, 4, 5

[46] Gong Zhang, Kai Wang, Xingqian Xu, Zhangyang Wang, and Humphrey Shi. Forget-me-not: Learning to forget in text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1755–1764, 2024. 2

[47] Gong Zhang, Kai Wang, Xingqian Xu, Zhangyang Wang, and Humphrey Shi. Forget-me-not: Learning to forget in text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1755–1764, 2024. 4, 5

[48] 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. *arXiv preprint arXiv:2405.15234*, 2024. 2

[49] 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. *arXiv preprint arXiv:2405.15234*, 2024. 4, 5

[50] Yimeng Zhang, Jinghan Jia, Xin Chen, Aochuan Chen, Yihua Zhang, Jiancheng Liu, Ke Ding, and Sijia Liu. To generate or not? safety-driven unlearned diffusion models are still easy to generate unsafe images... for now. In *European Conference on Computer Vision*, pages 385–403. Springer, 2025. 2, 5