Ablating Concepts in Text-to-Image Diffusion Models

Nupur Kumari\textsuperscript{1} Bingliang Zhang\textsuperscript{2} Sheng-Yu Wang\textsuperscript{1} 
Eli Shechtman\textsuperscript{3} Richard Zhang\textsuperscript{3} Jun-Yan Zhu\textsuperscript{1} 

\textsuperscript{1}Carnegie Mellon University \hspace{1cm} \textsuperscript{2}Tsinghua University \hspace{1cm} \textsuperscript{3}Adobe Research

Figure 1: Our method can ablate copyrighted materials and memorized images from pretrained text-to-image diffusion models. Our method learns to change the image distribution of a target concept to match an anchor concept, e.g., Van Gogh painting $\rightarrow$ paintings (first row), or Grumpy cat $\rightarrow$ Cat (second row). Furthermore, we extend our method to prevent the generation of memorized images (third row).

Abstract

Large-scale text-to-image diffusion models can generate high-fidelity images with powerful compositional ability. However, these models are typically trained on an enormous amount of Internet data, often containing copyrighted material, licensed images, and personal photos. Furthermore, they have been found to replicate the style of various living artists or memorize exact training samples. How can we remove such copyrighted concepts or images without retraining the model from scratch? To achieve this goal, we propose an efficient method of ablating concepts in the pretrained model, i.e., preventing the generation of a target concept. Our algorithm learns to match the image distribution for a target style, instance, or text prompt we wish to ablate to the distribution corresponding to an anchor concept. This prevents the model from generating target concepts given its text condition. Extensive experiments show that our method can successfully prevent the generation of the ablated concept while preserving closely related concepts in the model.

1. Introduction

Large-scale text-to-image models have demonstrated remarkable ability in synthesizing photorealistic images [51, 43, 56, 54, 76, 14]. In addition to algorithms and compute resources, this technological advancement is powered by the use of massive datasets scraped from web [59]. Unfortunately, the datasets often consist of copyrighted materials, the artistic oeuvre of creators, and personal photos [64, 10, 61].

We believe that every creator should have the right to opt out from large-scale models at any time for any image they have created. However, fulfilling such requests poses new computational challenges, as re-training a model from scratch for every user request can be computationally intensive. Here, we ask – How can we prevent the model from generating such content? How can we achieve it efficiently without re-training the model from scratch? How can we make sure that the model still preserves related concepts?
These questions motivate our work on ablation (removal) of concepts from text-conditioned diffusion models [54, 3]. We perform concept ablation by modifying generated images for the target concept (c *) to match a broader anchor concept (c), e.g., overwriting Grumpy Cat with cat or Van Gogh paintings with painting as shown in Figure 1. Thus, given the text prompt, painting of olive trees in the style of Van Gogh, generate a normal painting of olive trees even though the text prompt consists of Van Gogh. Similarly, prevent the generation of specific instances/objects like Grumpy Cat and generate a random cat given the prompt.

Our method aims at modifying the conditional distribution of the model given a target concept pΘ(x|c) to match a distribution p(x|c*) defined by the anchor concept c*. This is achieved by minimizing the Kullback–Leibler divergence between the two distributions. We propose two different target distributions that lead to different training objectives. In the first case, we fine-tune the model to match the model prediction between two text prompts containing the target and corresponding anchor concepts, e.g., A cute little Grumpy Cat and A cute little cat. In the second objective, the conditional distribution p(x|c) is defined by the modified text-image pairs of: a target concept prompt, paired with images of anchor concepts, e.g., the prompt a cute little Grumpy Cat with a random cat image. We show that both objectives can effectively ablate concepts.

We evaluate our method on 16 concept ablation tasks, including specific object instances, artistic styles, and memorized images, using various evaluation metrics. Our method can successfully ablate target concepts while minimally affecting closely related surrounding concepts that should be preserved (e.g., other cat breeds when ablating Grumpy Cat). Our method takes around five minutes per concept. Furthermore, we perform an extensive ablation study regarding different algorithmic design choices, such as the objective function variants, the choice of parameter subsets to fine-tune, the choice of anchor concepts, the number of fine-tuning steps, and the robustness of our method to misspelling in the text prompt. Finally, we show that our method can ablate multiple concepts at once and discuss the current limitations. The full version of the paper is available at https://arxiv.org/abs/2303.13516. Our code, data, and models are available at https://www.cs.cmu.edu/-concept-ablation/.

2. Related Work

Text-to-image synthesis has advanced significantly since the seminal works [82, 37], thanks to improvements in model architectures [77, 81, 68, 75, 28, 15, 74, 29, 57, 16], generative modeling techniques [52, 27, 54, 56, 4, 43, 14, 66], and availability of large-scale datasets [59]. Current methods can synthesize high-quality images with remarkable generalization ability, capable of composing different instances, styles, and concepts in unseen contexts. However, as these models are often trained on copyright images, it learns to mimic various artist styles [64, 61] and other copyrighted content [10]. In this work, we aim to modify the pretrained models to prevent the generation of such images. To remove data from pre-trained GANs, Kong et al. [32] add the redacted data to fake data, apply standard adversarial loss, and show results on MNIST and CIFAR. Unlike their method, which requires time-consuming model re-training on the entire dataset, our method can efficiently remove concepts without going through the original training set. Furthermore, we focus on large-scale text-based diffusion models. Recent work of Schramowski et al. [58] modify the inference process to prevent certain concepts from being generated. But we aim to ablate the concept from the model weights. Concurrent with our work, Gandikota et al. [20] aims to remove concepts using a score-based formulation. The reader is encouraged to review their work.

Training data memorization and unlearning. Several works have studied training data leaking [62, 12, 13, 11], which can pose a greater security and privacy risk, especially with the use of web-scale uncurated datasets in deep learning. Recent works [64, 10] have also shown that text-to-image models are susceptible to generating exact or similar copies of the training dataset for certain text conditions. Another line of work in machine unlearning [9, 21, 23, 22, 42, 8, 67, 60] explores data deletion at user’s request after model training. However, existing unlearning methods [23, 67] typically require calculating information, such as Fisher Information Matrix, making them computationally infeasible for large-scale models with billions of parameters trained on billions of images. In contrast, our method can directly update model weights and ablate a target concept as fast as five minutes.

Generative model fine-tuning and editing. Fine-tuning aims to adapt the weights of a pretrained generative model to new domains [73, 46, 72, 41, 79, 34, 47, 80, 30, 35, 24, 44], downstream tasks [71, 54, 78], and test images [6, 53, 48, 31, 25, 49]. Several recent works also explore fine-tuning text-to-image models to learn personalized or unseen concepts [33, 17, 55, 18] given a few exemplar images. Similarly, model editing [5, 70, 19, 69, 45, 38, 40, 39] aims to modify specific model weights based on users’ instructions to incorporate new computational rules or new visual effects. Unlike the above approaches, our method reduces the possible space by ablating specific concepts in the pretrained model.

3. Method

Here, we first provide a brief overview of text-to-image diffusion models [63, 27] in Section 3.1. We then propose our concept ablation formulation and explore two variants in Section 3.2. Finally, in Section 3.3, we discuss the training details for each type of ablation task.
3.1. Diffusion Models

Diffusion models [63] learn to reverse a forward Markov chain process where noise is gradually added to the input image over multiple timesteps \( t \in [0, T] \). The noisy image \( x_t \) at any time-step \( t \) is given by \( \sqrt{\alpha_t} x_0 + \sqrt{1 - \alpha_t} \epsilon \), where \( x_0 \) is a random real image, and \( \alpha_t \) determines the strength of gaussian noise \( \epsilon \) and decreases gradually with timestep such that \( x_T \sim N(0, I) \). The denoising network \( \Phi(x_t, c, t) \) is trained to denoise the noisy image to obtain \( x_{t-1} \), and can also be conditioned on other modalities such as text \( c \). The training objective can be reduced to predicting the noise \( \epsilon \):

\[
\mathcal{L}(x, c) = \mathbb{E}_{x_t, x, c, t, \epsilon}[w_t||\epsilon - \Phi(x_t, c, t)||],
\]

where \( w_t \) is a time-dependent weight on the loss. To synthesize an image during inference, given the text condition \( c \), we iteratively denoise a Gaussian noise image \( x_T \sim N(0, I) \) for a fixed number of timesteps [65, 36].

3.2. Concept Ablation

We define concept ablation as the task of preventing the generation of the desired image corresponding to a given target concept that needs to be ablated. As re-training the model on a new dataset with the concept removed is impractical, this becomes a challenging task. We need to ensure that editing a model to ablate a particular concept doesn’t affect the model performance on other closely related concepts.

A naïve approach. Our first attempt is to simply maximize the diffusion model training loss [67, 32] on the text-image pairs for the target concept while imposing regularizations on the weights. Unfortunately, this method leads to worse results on close surrounding concepts of the target concept. We compare our method with this baseline in Section 4.2 (Figure 3) and show that it performs sub-optimally.

Our formulation. As concept ablation prevents the generation of the target concept, the question arises: what should be generated instead? In this work, we assume that the user provides the desired anchor concept, e.g., Cat for Grumpy Cat. The anchor concept overwrites the target concept and should be a superset or similar to the target concept. Thus, given a set of text prompts \( \{c^*\} \) describing the target concept, we aim to match the following two distributions via Kullback–Leibler (KL) divergence:

\[
\arg\min_{\Phi} \mathcal{D}_{KL}(p(x_{0...T}|c)||p_{\Phi}(x_{0...T}|c^*)),
\]

where \( \{p(x_{0...T}|c)\} \) is some target distribution on the \( \{x_t\}, t \in [0, T] \), defined by the anchor concept \( c \) and \( p_{\Phi}(x_{0...T}|c^*) \) is the model’s distribution for the target concept. Intuitively, we want to associate text prompts \( \{c^*\} \) with the images corresponding to anchor prompts \( \{c\} \). Defining different anchor concept distributions leads to different objective functions, as we discuss next.

To accomplish the above objective, we first create a small dataset that consists of \( \{x, c, c^*\} \) tuple, where \( c \) is a random prompt for the anchor concept, and \( c^* \) is the generated image with that condition, and \( c^* \) is modified from \( c \) to include the target concept. For example, if \( c \) is photo of a cat, \( c^* \) will be photo of a Grumpy Cat, and \( x \) will be a generated image with text prompt \( c \). For brevity, we use the same notation \( x \) to denote these generated images.

Model-based concept ablation. Here, we match the distribution of the target concept \( p_{\Phi}(x_{0...T}|c^*) \) to the pretrained model’s distribution \( p_{\Phi}(x_{0...T}|c) \) given the anchor concept. The fine-tuned network should have a similar distribution of generated images given \( c^* \) as that of \( c \), which can be expressed as minimizing the KL divergence between the two. This is similar to the standard diffusion model training objec-
Where the noisy intermediate latent \( x_t \sim p_{\theta}(x_t|c) \), \( \Phi \) is the original network, and \( \hat{\Phi} \) is the new network we aim to learn. We can optimize the KL divergence by minimizing the following equivalent objective:

\[
\arg \min_{\hat{\Phi}} \mathbb{E}_{x_t,c,t} [||\Phi(x_t,c,t) - \hat{\Phi}(x_t,c^*,t)||].
\]  

(4)

We show the full derivation in our arXiv version. We initialize \( \hat{\Phi} \) with the pretrained model. Unfortunately, optimizing the above objective requires us to sample from \( p_{\theta}(x_t|c) \) and keep copies of two large networks \( \Phi \) and \( \hat{\Phi} \), which is time and memory-intensive. To bypass this, we sample \( x_t \) using the forward diffusion process and assume that the model remains similar for the anchor concept during fine-tuning. Therefore we use the network \( \hat{\Phi} \) with stopgrad to get the anchor concept prediction. Thus, our final training objective is:

\[
\mathcal{L}_{\text{model}}(x,c,c^*) = \mathbb{E}_{x,\epsilon,\epsilon',c,t} [||\Phi(x_t,c,t) - \hat{\Phi}(x_t,c^*,t)||],
\]  

(5)

where \( x_t = \sqrt{\alpha_t}x + \sqrt{1-\alpha_t}\epsilon \). As shown in Figure 2 (left), this objective minimizes the difference in the model’s prediction given the target prompt and anchor prompt. It is also possible to optimize the approximation to reverse KL divergence, and we discuss it in Section 4.3.

**Noise-based concept ablation.** Alternatively, we can redefine the ground truth text-image pairs as \(<\text{a target concept text prompt}, \text{the generated image of the corresponding anchor concept text prompt}\> \text{, e.g., } <\text{photo of Grumpy Cat}, \text{random cat image}> \text{. We fine-tune the model on these redefined pairs with the standard diffusion training loss:}

\[
\mathcal{L}_{\text{noise}}(x,c,c^*) = \mathbb{E}_{x,\epsilon,\epsilon',c,t} [||\epsilon - \hat{\Phi}(x_t,c^*,t)||],
\]  

(6)

where the generated image \( x \) is sampled from conditional distribution \( p_{\theta}(x|c) \). We then create the noisy version \( x_t = \sqrt{\alpha_t}x + \sqrt{1-\alpha_t}\epsilon \). As shown in Figure 2, the first objective (Eqn. 5) aims to match the model’s predicted noises, while the second objective (Eqn. 6) aims to match the Gaussian noises \( \epsilon \). We evaluate the above two objectives in Section 4.

**Regulization loss.** We also add the standard diffusion loss on \((x,c)\) anchor concept pairs as a regularization [55, 33]. Thus, our final objective is \( \lambda\mathcal{L}(x,c) + \mathcal{L}(x,c,c^*) \), where the losses are as defined in Eqns. 1 and 5 (or 6) respectively. We require regularization loss as the target text prompt can consist of the anchor concept, e.g., Cat in Grumpy Cat.

**Parameter subset to update.** We experiment with three variations where we fine-tune different network parts: (1) Cross-Attention: fine-tune key and value projection matrices in the diffusion model’s U-Net [33], (2) Embedding: fine-tune the text embedding in the text transformer [17], and (3) Full Weights: fine-tune all parameters of the U-Net [55].

![Figure 3: Comparison of different learning objectives.](image)

The model-based concept ablation converges faster than the noise-based variant while maintaining better performance on surrounding concepts. Maximizing the loss on the target concept dataset leads to the deterioration of surrounding concepts (top row).

### 3.3. Training Details

**Instance.** Given the target and the anchor concept, such as Grumpy Cat and Cat, we first use ChatGPT [1] to generate 200 random prompts \( \{c\} \) containing the anchor concept. We generate 1,000 images from the pretrained diffusion model using the 200 prompts and replace the word Cat with Grumpy Cat to get target text prompts \( \{c^*\} \).

**Style.** When removing a style, we use generic painting styles as the anchor concept. We use clip-retrieval [2] to obtain a set of text prompts \( c \) similar to the word painting in the CLIP feature space. We then generate 1000 images from the pretrained model using the 200 prompts. To get target prompts \( \{c^*\} \), we append in the style of \( \{\text{target style}\} \) and similar variations to anchor prompts \( c \).

**Memorized images.** Recent methods for detecting training set memorization can identify both the memorized image and corresponding text prompt \( c^* \) [10]. We then use ChatGPT to generate five anchor prompts \( \{c\} \) that can generate similar content as the memorized image. In many cases, these anchor prompts still generate the memorized images. Therefore, we first generate several more paraphrases of the anchor prompts using chatGPT and include the three prompts that lead to memorized images often into target prompts and ten prompts that lead to memorized images least as anchor prompts. Thus \( c^* \) and \( c \) for ablating the target memorized image consists of four and ten prompts, respectively. We then similarly generate 1000 images using the anchor prompts and use...
Figure 4: **Quantitative evaluation for ablating instances (top row) and styles (bottom row).** We show the performance of our final model-based concept ablation method across training steps and on updating different subsets of parameters. All metrics are averaged across four target concepts. Both embedding and cross-attention fine-tuning converge early. Fine-tuning cross-attention layers performs slightly worse for surrounding concepts but remains more robust to small spelling mistakes (third column).

Figure 5: **Qualitative samples when ablating specific object instances.** We show samples from different variations of our method in each row. The noise-based method performs worse on Nemo and R2D2 instances compared to the model-based variant. With the model-based variant, fine-tuning different subsets of parameters perform comparably to each other. As shown in Figure 4 (third column) and Figure 6, fine-tuning only the embedding is less robust to small spelling mistakes.
image similarity metrics [50, 10] to filter out the memorized images and use the remaining ones for training.

4. Experiments

In this section, we show the results of our method on ablating various instances, styles, and memorized images. All our experiments are based on the Stable Diffusion model [3]. Please refer to the appendix of our arXiv version for more training details.

4.1. Evaluation metrics and baselines

**Baseline.** We compare our method with a loss maximization baseline inspired by Tanno et al. [67]:

$$\arg\min_{\Phi} \max(1 - \mathcal{L}(x^*, c^*), 0) + \lambda ||\hat{\Phi} - \Phi||_2$$  (7)

where $x^*$ is the set of generated images with condition $c^*$ and $\mathcal{L}$ is the diffusion training loss as defined in Eqn. 1. We compare our method with this baseline on ablating instances.

**Evaluation metrics.** We use CLIP Score and CLIP accuracy [26] to evaluate whether the model can ablate the target concept. CLIP Score measures the similarity of the generated image with the target concept text, e.g., Grumpy Cat in CLIP feature space. Similarly, CLIP accuracy measures the accuracy of ablated vs. anchor concept binary classification task for each generated image using cosine distance in CLIP feature space. For both metrics, lower values indicate more successful ablation. We further evaluate the performance on small spelling mistakes in the ablated text prompts. We also use the same metrics to evaluate the model on related surrounding concepts (e.g., similar cat breeds for Grumpy Cat), which should be preserved. Similar to before, CLIP accuracy is measured between the surrounding concept and anchor concept, and the higher, the better. Similarly, CLIP Score measures the similarity of the generated image with the surrounding concept text, and the higher, the better.

Furthermore, to test whether the fine-tuned model can retain existing concepts, we calculate $KID$ [7] between the set of generated images from fine-tuned model and the pretrained model. Higher KID is better for the target concept, while lower KID is better for anchor and surrounding concepts. We generate 200 images each for ablated, anchor, and surrounding concepts using 10 prompts and 50 steps of the DDPM sampler. The prompts are generated through ChatGPT for object instances and manually created for styles by captioning real images corresponding to each style.

To measure the effectiveness of our method in ablating memorized images, following previous works [50, 10], we use SSCD [50] model to measure the percentage of generated images having similarity with the memorized image greater than a threshold.

4.2. Comparisons and main results

**Instances.** We show results on four concepts and replace them with anchor concepts, namely, (1) Grumpy Cat → Cat, (2) Snoopy → Dog, (3) Nemo → Fish, and (4) R2D2 → Robot. Figure 3 compares our two proposed methods and the loss maximization baseline with Cross-Attention fine-tuning. As the baseline method maximizes the norm between ground truth and predicted noise, it gradually gen-
Figure 8: Ablating styles with the model-based variant. The ablated model generates similar content as the pretrained model but without the unique style. More samples for target and surrounding concepts are shown in the appendix of our arXiv version.

- **Van Gogh**
- **Monet**
- **Salvador Dali**
- **Greg Rutkowski**

...
Figure 9: Ablating memorized images with the model-based variant. Text-to-image diffusion models often learn to generate exact or near-exact copies of real images. We fine-tune the model to map the generated image distribution for the given text prompt to images generated with its variations. This results in the fine-tuned model generating different variations instead of copying the real image. We show more samples in the appendix of our arXiv version.

Table 1: Memorization rate. We show the percentage of generated samples that are highly similar ($\geq 0.5$ cosine similarity on SSCD) to a “memorized” image.

<table>
<thead>
<tr>
<th>Target Prompt</th>
<th>Pretrained Model</th>
<th>Ours (Full Weights)</th>
</tr>
</thead>
<tbody>
<tr>
<td>New Orleans House Galaxy Case</td>
<td>62.5</td>
<td>0.0</td>
</tr>
<tr>
<td>Portrait of Tiger in black and white by Lukas Holas</td>
<td>50.0</td>
<td>0.0</td>
</tr>
<tr>
<td>VAN GOGH CAFE TERASSE cop.png</td>
<td>56.5</td>
<td>1.5</td>
</tr>
<tr>
<td>Captain Marvel Exclusive Ccxp Poster Released Online By Marvel</td>
<td>95.0</td>
<td>0.5</td>
</tr>
<tr>
<td>Sony Boss Confirms Bloodborne Expansion is Coming</td>
<td>83.5</td>
<td>0.5</td>
</tr>
<tr>
<td>Ann Graham Lotz</td>
<td>26.5</td>
<td>0.0</td>
</tr>
<tr>
<td>&lt;i&gt;The Long Dark&lt;/i&gt; Gets First Trailer, Steam Early Access</td>
<td>100.0</td>
<td>0.0</td>
</tr>
<tr>
<td>A painting with letter M written on it Canvas Wall Art Print</td>
<td>4.0</td>
<td>0.0</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td><strong>50.1</strong></td>
<td><strong>0.3</strong></td>
</tr>
</tbody>
</table>

Table 4.3. Additional Analysis

Single model with multiple concepts ablated. Our method can also remove multiple concepts by training on the union of datasets for longer training steps. We show the results of one model with all instances and one model with all styles ablated in Figure 10. We use the model-based variant of our method and cross-attention fine-tuning. More samples are shown in the appendix of our arXiv version. The drop in accuracy for the ablated concepts is similar to Figure 5 while maintaining the accuracy on surrounding concepts.

Reverse KL divergence. In our model-based concept ablation, we optimize the KL divergence between the anchor concept and target concept distribution. Here, we compare it with optimizing the approximation to reverse KL divergence, i.e., $\mathbb{E}_{x \sim \mathcal{X}, c \sim c_{\ast}, t \sim t}[w_{t}||\Phi(x_{t}^{\ast}, c_{t}, t) - \Phi(x_{t}^{\ast}, c_{\ast}, t)||]$. Thus the expectation of loss is over target concept images.

Figure 10: Ablating multiple instances (left) and style (right). Top: quantitative results show the drop in the CLIP Accuracy of the target concept, which has been ablated, whereas the accuracy for surrounding concepts remains the same. Bottom: one sample image corresponding to each ablated target concept.
5. Discussion and Limitations

Although we can ablate concepts efficiently for a wide range of object instances, styles, and memorized images, our method is still limited in several ways. First, while our method overwrites a target concept, this does not guarantee that the target concept cannot be generated through a different, distant text prompt. We show an example in Figure 14 (a), where after ablating Van Gogh, the model can still generate Starry night painting. However, upon discovery, one can resolve this by explicitly ablating the target concept Starry night painting. Secondly, when ablating a target concept, we still sometimes observe slight degradation in its surrounding concepts, as shown in Figure 14 (c).

Our method does not prevent a downstream user with full access to model weights from re-introducing the ablated concept [55, 33, 17]. Even without access to the model weights, one may be able to iteratively optimize for a text prompt with a particular target concept. Though that may be much more difficult than optimizing the model weights, our work does not guarantee that this is impossible.

Nevertheless, we believe every creator should have an “opt-out” capability. We take a small step towards this goal, creating a computational tool to remove copyrighted images and artworks from large-scale image generative models.

Acknowledgment. We are grateful to Gaurav Parmar, Dao-han Lu, Muyang Li, Songwei Ge, Jingwan Lu, Sylvain Paris, and Bryan Russell for their helpful discussion, and to Aniruddha Mahapatra and Kangle Deng for paper proofreading. The work is partly supported by Adobe and NSF IIS-2239076.


Han Zhang, Tao Xu, Hongsheong Li, Shaoxing Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan: Text-to-photo-realistic image synthesis with stacked generative adversarial networks. In *IEEE International Conference on Computer Vision (ICCV)*, 2017. 2

Miaoyun Zhao, Yulai Cong, and Lawrence Carin. On leveraging pretrained gans for generation with limited data. In *International Conference on Machine Learning (ICML)*, 2020. 2

