

# Ablating Concepts in Text-to-Image Diffusion Models

Nupur Kumari<sup>1</sup>  
Eli Shechtman<sup>3</sup>

Bingliang Zhang<sup>2</sup>  
Richard Zhang<sup>3</sup>

Sheng-Yu Wang<sup>1</sup>  
Jun-Yan Zhu<sup>1</sup>

<sup>1</sup>Carnegie Mellon University

<sup>2</sup>Tsinghua University

<sup>3</sup>Adobe Research

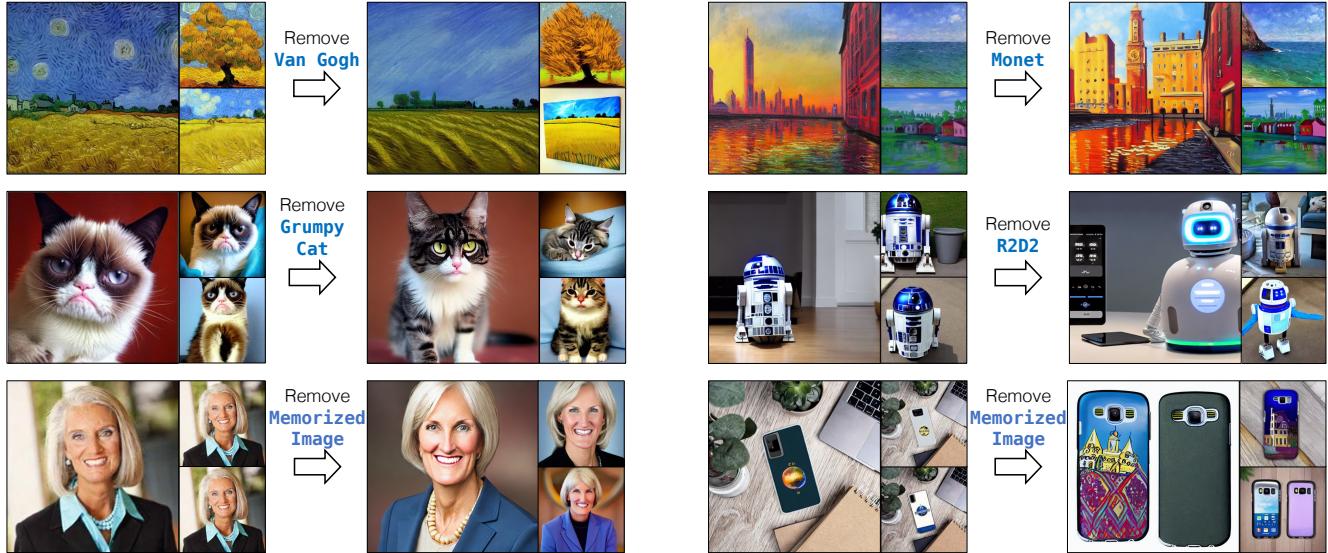


Figure 1: Our method can ablate copyrighted materials and memorized images from pretrained text-to-image diffusion models. Our method learns to change the image distribution of a **target concept** to match an **anchor concept**, e.g., Van Gogh painting → paintings (first row), or Grumpy cat → Cat (second 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  $p_\Phi(x|c^*)$  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.

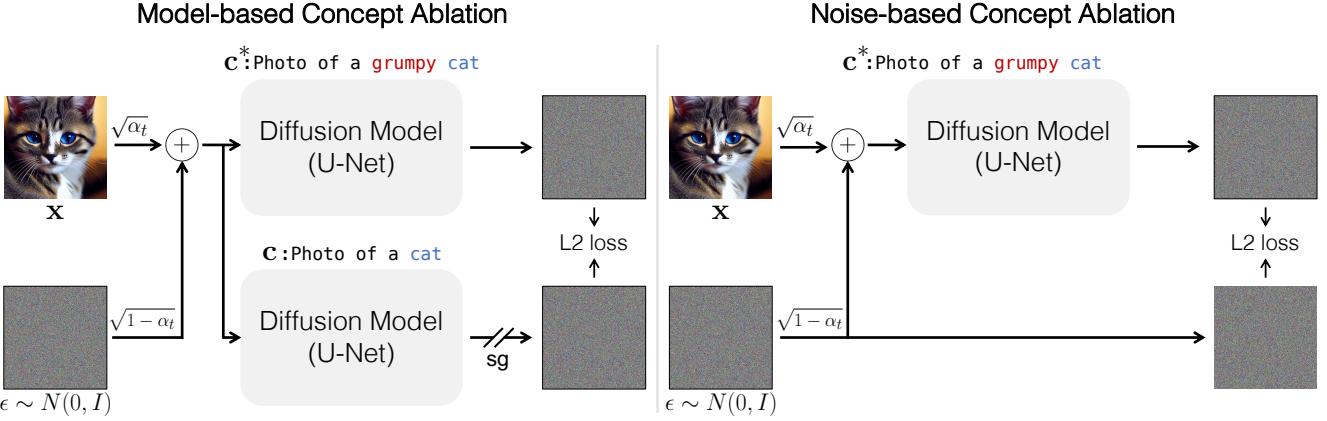


Figure 2: **Overview.** We update model weights to modify the generated image distribution on the target concept, e.g., Grumpy Cat, to match an anchor distribution, e.g., Cat. We propose two variants. *Left:* The anchor distribution is generated by the model itself, conditioned on the anchor concept. *Right:* The anchor distribution is defined by the modified pairs of <target prompt, anchor image>. An input image  $\mathbf{x}$  is generated with anchor concept  $\mathbf{c}$ . Adding randomly sampled noise  $\epsilon$  results in noisy image  $\mathbf{x}_t$  at time-step  $t$ . Target prompt  $\mathbf{c}^*$  is produced by appropriately modifying  $\mathbf{c}$ . In experiments, we find the model-based variant to be more effective.

### 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  $\mathbf{x}_t$  at any time-step  $t$  is given by  $\sqrt{\alpha_t} \mathbf{x}_0 + \sqrt{1 - \alpha_t} \epsilon$ , where  $\mathbf{x}_0$  is a random real image, and  $\alpha_t$  determines the strength of gaussian noise  $\epsilon$  and decreases gradually with timestep such that  $\mathbf{x}_T \sim N(0, I)$ . The denoising network  $\Phi(\mathbf{x}_t, \mathbf{c}, t)$  is trained to denoise the noisy image to obtain  $\mathbf{x}_{t-1}$ , and can also be conditioned on other modalities such as text  $\mathbf{c}$ . The training objective can be reduced to predicting the noise  $\epsilon$ :

$$\mathcal{L}(\mathbf{x}, \mathbf{c}) = \mathbb{E}_{\epsilon, \mathbf{x}, \mathbf{c}, t} [w_t || \epsilon - \Phi(\mathbf{x}_t, \mathbf{c}, t) ||], \quad (1)$$

where  $w_t$  is a time-dependent weight on the loss. To synthesize an image during inference, given the text condition  $\mathbf{c}$ , we iteratively denoise a Gaussian noise image  $\mathbf{x}_T \sim N(0, I)$  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, thus 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  $\{\mathbf{c}^*\}$  describing the target concept, we aim to match the following two distributions via Kullback–Leibler (KL) divergence:

$$\arg \min_{\hat{\Phi}} \mathcal{D}_{\text{KL}}(p(\mathbf{x}_{(0 \dots T)} | \mathbf{c}) || p_{\hat{\Phi}}(\mathbf{x}_{(0 \dots T)} | \mathbf{c}^*)), \quad (2)$$

where  $p(\mathbf{x}_{(0 \dots T)} | \mathbf{c})$  is some target distribution on the  $\{\mathbf{x}_t\}$ ,  $t \in [0, T]$ , defined by the anchor concept  $\mathbf{c}$  and  $p_{\hat{\Phi}}(\mathbf{x}_{(0 \dots T)} | \mathbf{c}^*)$  is the model's distribution for the target concept. Intuitively, we want to associate text prompts  $\{\mathbf{c}^*\}$  with the images corresponding to anchor prompts  $\{\mathbf{c}\}$ . 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  $(\mathbf{x}, \mathbf{c}, \mathbf{c}^*)$  tuple, where  $\mathbf{c}$  is a random prompt for the anchor concept,  $\mathbf{x}$  is the generated image with that condition, and  $\mathbf{c}^*$  is modified from  $\mathbf{c}$  to include the target concept. For example, if  $\mathbf{c}$  is photo of a cat,  $\mathbf{c}^*$  will be photo of a Grumpy Cat, and  $\mathbf{x}$  will be a generated image with text prompt  $\mathbf{c}$ . For brevity, we use the same notation  $\mathbf{x}$  to denote these generated images.

**Model-based concept ablation** . Here, we match the distribution of the target concept  $p_{\hat{\Phi}}(\mathbf{x}_{(0 \dots T)} | \mathbf{c}^*)$  to the pretrained model's distribution  $p_{\Phi}(\mathbf{x}_{(0 \dots T)} | \mathbf{c})$  given the anchor concept. The fine-tuned network should have a similar distribution of generated images given  $\mathbf{c}^*$  as that of  $\mathbf{c}$ , which can be expressed as minimizing the KL divergence between the two. This is similar to the standard diffusion model training objec-

tive, except the target distribution is defined by the pretrained model instead of training data. Eqn. 2 can be expanded as

$$\arg \min_{\Phi} \sum_{t=1}^T \mathbb{E}_{p_{\Phi}(\mathbf{x}_0, \dots, \mathbf{x}_T | \mathbf{c})} \left[ \log \frac{p_{\Phi}(\mathbf{x}_{t-1} | \mathbf{x}_t, \mathbf{c})}{p_{\hat{\Phi}}(\mathbf{x}_{t-1} | \mathbf{x}_t, \mathbf{c}^*)} \right] \quad (3)$$

where the noisy intermediate latent  $\mathbf{x}_t \sim p_{\Phi}(\mathbf{x}_t | \mathbf{c})$ ,  $\Phi$  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:

$$\arg \min_{\hat{\Phi}} \mathbb{E}_{\epsilon, \mathbf{x}_t, \mathbf{c}^*, t} [w_t || \Phi(\mathbf{x}_t, \mathbf{c}, t) - \hat{\Phi}(\mathbf{x}_t, \mathbf{c}^*, t) ||]. \quad (4)$$

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  $p_{\Phi}(\mathbf{x}_t | \mathbf{c})$  and keep copies of two large networks  $\Phi$  and  $\hat{\Phi}$ , which is time and memory-intensive. To bypass these, we sample  $\mathbf{x}_t$  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

$$\mathcal{L}_{\text{model}}(\mathbf{x}, \mathbf{c}, \mathbf{c}^*) = \mathbb{E}_{\epsilon, \mathbf{x}, \mathbf{c}^*, t} [w_t || \hat{\Phi}(\mathbf{x}_t, \mathbf{c}, t).sg() - \hat{\Phi}(\mathbf{x}_t, \mathbf{c}^*, t) ||], \quad (5)$$

where  $\mathbf{x}_t = \sqrt{\alpha_t} \mathbf{x} + \sqrt{1 - \alpha_t} \epsilon$ . 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 <a target concept text prompt, the generated image of the corresponding anchor concept text prompt>, e.g., <photo of Grumpy Cat, random cat image>. We fine-tune the model on these redefined pairs with the standard diffusion training loss:

$$\mathcal{L}_{\text{noise}}(\mathbf{x}, \mathbf{c}, \mathbf{c}^*) = \mathbb{E}_{\epsilon, \mathbf{x}, \mathbf{c}^*, t} [w_t || \epsilon - \hat{\Phi}(\mathbf{x}_t, \mathbf{c}^*, t) ||], \quad (6)$$

where the generated image  $\mathbf{x}$  is sampled from conditional distribution  $p_{\Phi}(\mathbf{x} | \mathbf{c})$ . We then create the noisy version  $\mathbf{x}_t = \sqrt{\alpha_t} \mathbf{x} + \sqrt{1 - \alpha_t} \epsilon$ . 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.

**Regualization loss.** We also add the standard diffusion loss on  $(\mathbf{x}, \mathbf{c})$  anchor concept pairs as a regularization [55, 33]. Thus, our final objective is  $\lambda \mathcal{L}(\mathbf{x}, \mathbf{c}) + \mathcal{L}(\mathbf{x}, \mathbf{c}, \mathbf{c}^*)$ , where the losses are as defined in Eqn. 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].

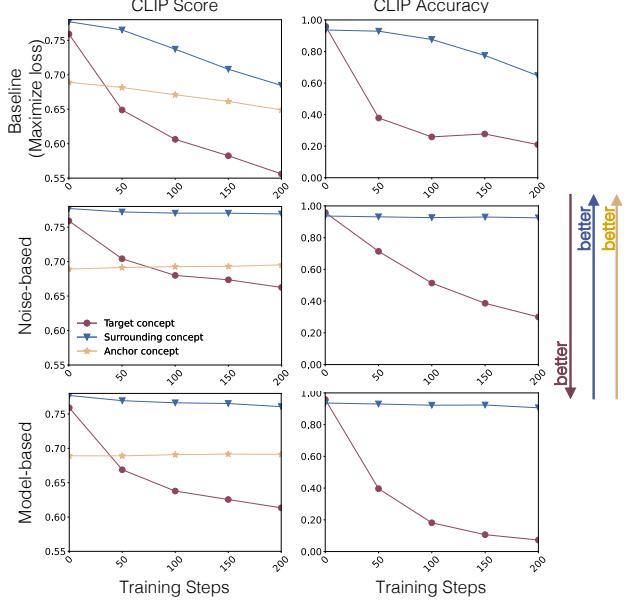


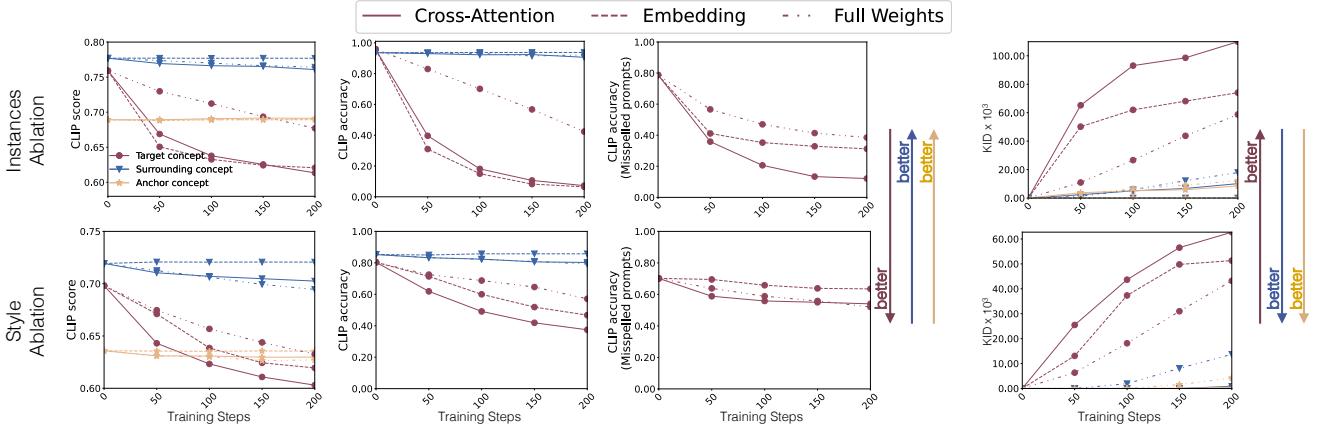
Figure 3: **Comparison of different learning objectives.** The *model-based* concept ablation converges faster than the *noise-based* variant while maintaining better performance on surrounding concepts. Maximizing the loss on the target concept dataset leads to the deterioration of surrounding concepts (top row).

### 3.3. Training Details

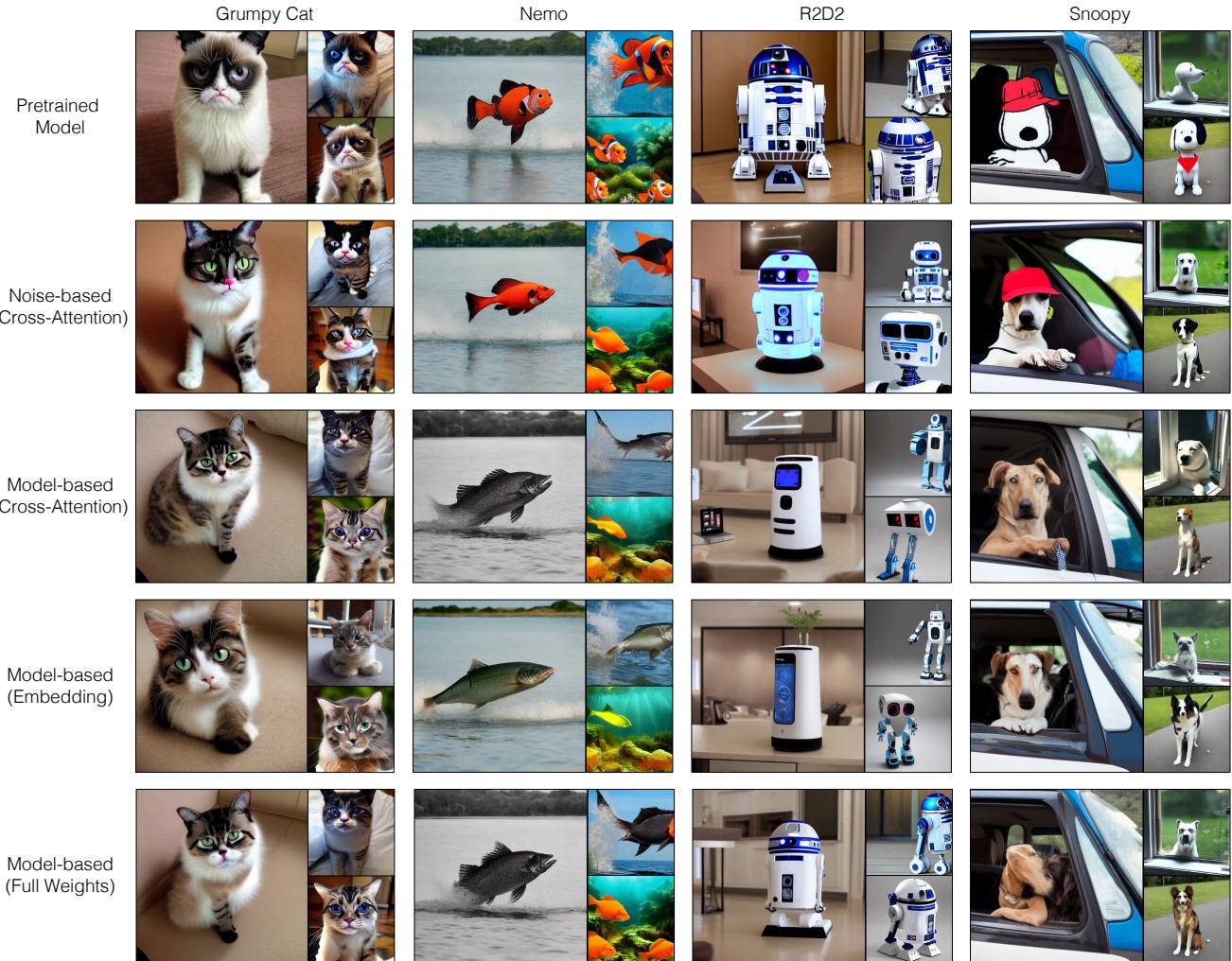
**Instance.** Given the target and the anchor concept, such as Grumpy Cat and Cat, we first use ChatGPT [1] to generate 200 random prompts  $\{\mathbf{c}\}$  containing the anchor concept. We generate 1,000 images from the pretrained diffusion model using the 200 prompts and replace the word Cat with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

**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  $\mathbf{c}$  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  $\{\mathbf{c}^*\}$ , we append *in the style of {target style}* and similar variations to anchor prompts  $\mathbf{c}$ .

**Memorized images.** Recent methods for detecting training set memorization can identify both the memorized image and corresponding text prompt  $\mathbf{c}^*$  [10]. We then use ChatGPT to generate five anchor prompts  $\{\mathbf{c}\}$  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  $\mathbf{c}^*$  and  $\mathbf{c}$  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.

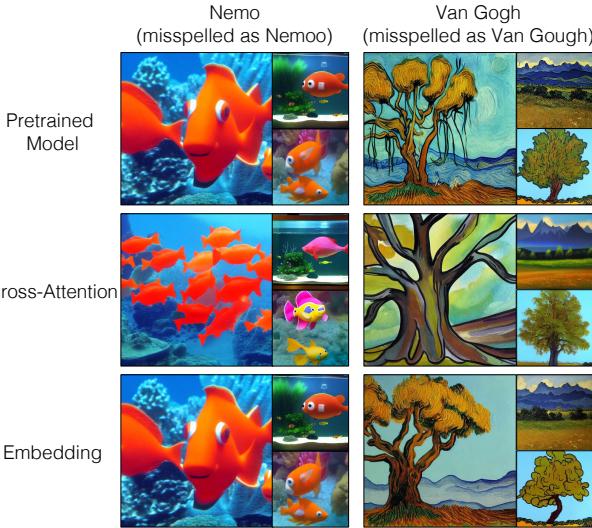


Figure 6: **Robustness of the *model-based* variant to spelling mistakes in the text prompt.** Fine-tuning only the embedding makes it less robust to slight spelling mistakes. This makes it easy to circumvent the method and still be able to generate the target concept. Whereas fine-tuning cross-attention parameters is robust to those.

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]:

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

where  $\mathbf{x}^*$  is the set of generated images with condition  $\mathbf{c}^*$  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



Figure 7: **Qualitative comparison between baseline and ours.** Model fine-tuned by our method generates images that are relatively more similar to the ones generated by the pretrained model on the BB8 instance, which should be preserved while ablating R2D2. Cross-Attention parameters are fine-tuned in both methods.

*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

erates noisy images when trained longer. This also leads to worse performance on surrounding concepts than our method, as shown by the quantitative metrics in Figure 3. Qualitative samples on the target concept R2D2 and its surrounding concept BB8 are also shown in Figure 7. Between our two methods, the *model-based* variant, i.e., minimizing the difference in prediction with the pretrained model’s anchor concept, leads to faster convergence and is better or on par with the *noise-based* variant. The qualitative comparison in Figure 5 also shows that, specifically on the Nemo instance. Thus, we use *model-based* variant for all later experiments. In Figure 4, we show the performance comparison when fine-tuning different subsets of the model weights.

As shown in Figure 5, the fine-tuned model successfully maps the target concept to the anchor concept. Fine-tuning only the text embedding performs on par with fine-tuning cross-attention layers. However, it is less robust to minor

spelling errors that still generate the same instance in the pretrained model as shown in Figure 4 (third column) and Figure 6. We show more results of ablated target and its surrounding concepts in the appendix of our arXiv version.

**Style.** For ablating styles, we consider four artists: (1) Van Gogh, (2) Salvador Dali, (3) Claude Monet, and (4) Greg Rutkowski, with the anchor concept as generic painting styles. Figures 4 and 8 show our method’s quantitative and qualitative performance when different subsets of parameters are fine-tuned. We successfully ablate specific styles while minimally affecting related surrounding styles.

**Memorized images.** We select eight image memorization examples from the recent works [64, 10], four of which are shown in Figure 9. It also shows the sample generations before and after fine-tuning. The fine-tuned model generates various outputs given the same text prompt instead of the



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.

Target Prompt	Pretrained Model	Ours (Full Weights)
New Orleans House Galaxy Case	65.5	0.0
Portrait of Tiger in black and white by Lukas Holas	50.0	0.0
VAN GOGH CAFE TERASSE copy.jpg	56.5	1.5
Captain Marvel Exclusive Ccpx Poster Released Online By Marvel	95.0	0.5
Sony Boss Confirms Bloodborne Expansion is Coming	83.5	0.5
Ann Graham Lotz	26.5	0.0
<i>The Long Dark</i> Gets First Trailer, Steam Early Access	100.0	0.0
A painting with letter M written on it Canvas Wall Art Print	4.0	0.0
<b>Average</b>	<b>60.1</b>	<b>0.3</b>

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

memorized sample. Among different parameter settings, we find finetuning *Full Weights* gives the best results. We show the percentage of samples with  $\geq 0.5$  similarity with the memorized image in Table 1. We show more sample generations and the initial set of anchor prompts for each case in the appendix of our arXiv version.

### 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.

**The role of anchor category.** In all the above experiments, we assume an anchor category  $c^*$  is given to overwrite the target concept. Here, we investigate the role of choosing different anchor categories for ablating Grumpy Cat and show results with the anchor concept as British Shorthair Cat

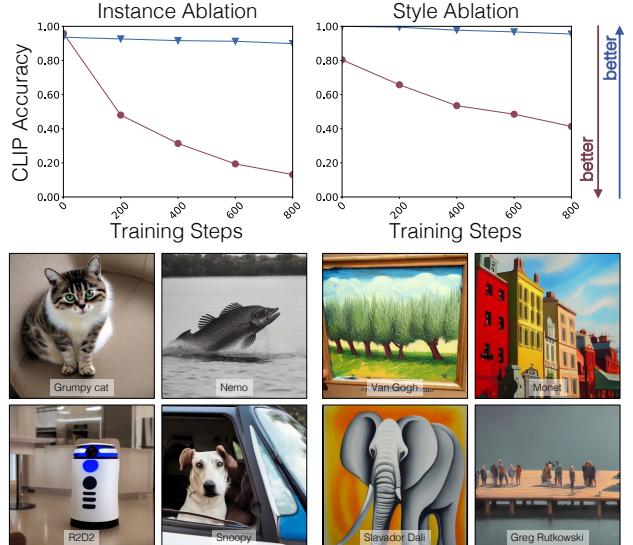


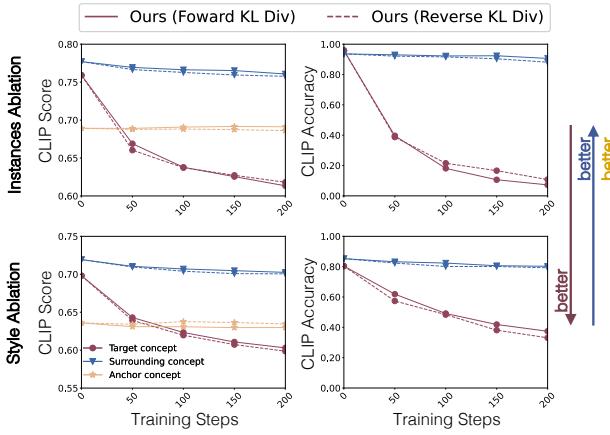
Figure 10: **Ablating multiple instances (left) and style (right).** Top: quantitative results show the drop in the CLIP Accuracy of the target concept, which has been ablated, whereas the accuracy for surrounding concepts remains the same. Bottom: one sample image corresponding to each ablated target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

and Felidae in Figure 11. Both anchor concepts work well.

**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.,  $\mathbb{E}_{\epsilon, \mathbf{x}^*, \mathbf{c}^*, \mathbf{c}, t} [w_t || \hat{\Phi}(\mathbf{x}_t^*, \mathbf{c}, t).sg() - \hat{\Phi}(\mathbf{x}_t^*, \mathbf{c}^*, t)]$ . Thus the expectation of loss is over target concept images. Figure 12 shows the quantitative comparison on ablating instances and style concepts. As we can see, it performs



**Figure 11: The choice of anchor concepts.** Our method is robust to the choice of anchor concepts. With both British shorthair cat and Felidae as anchor concepts, our method can ablate the target Grumpy Cat concept.



**Figure 12: Reverse KL divergence objective.** We show the results of optimizing the loss over target concept images for ablating instances (top) and style (bottom). Compared to using anchor concept images as training images, this performs slightly worse on ablating instances with lower CLIP Score on surrounding concepts while having similar CLIP Score on the target concept. It performs marginally better on ablating styles.

marginally better on ablating style concepts but worse on instances. In Figure 13, we show sample generations for the case where it outperforms the forward KL divergence based objective qualitatively on ablating Van Gogh.

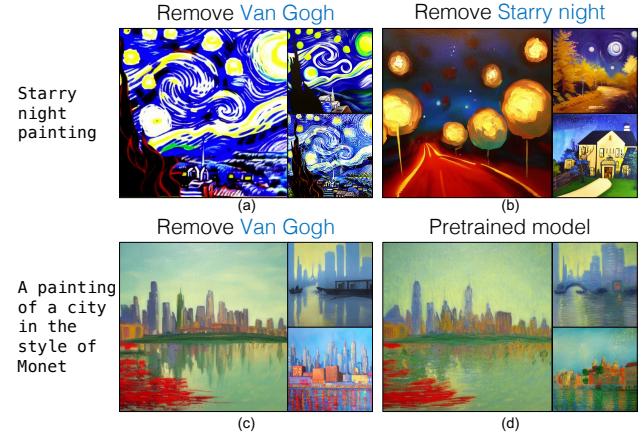
## 5. Discussion and Limitations

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 con-



**Figure 13: Qualitative samples with reverse KL divergence objective.** It performs better on certain styles and can successfully ablate famous paintings as well, which is not achievable with forward KL divergence-based objective and requires additional steps as shown in Figure 14.



**Figure 14: Limitations.** Top: (a) our method fails to remove certain paintings generated with the painting’s titles. (b) We can further ablate these concepts. Bottom: Though our method is better than baseline in preserving surrounding concepts as shown in Figure 7, the generated samples still sometimes show degradation for surrounding concepts, e.g., Monet (c) when ablating Van Gogh as compared to the pretrained model (d).

cept [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, Daohan Lu, Muyang Li, Songwei Ge, Jingwan Lu, Sylvain Paris, and Bryan Russell for their helpful discussion, and to Anirudha Mahapatra and Kangle Deng for paper proofreading. The work is partly supported by Adobe and NSF IIS-2239076.

## References

- [1] Chatgpt. <https://chat.openai.com/chat>, 2022. 4
- [2] Clip retrieval. <https://github.com/rom1504/clip-retrieval>, 2022. 4
- [3] Stable diffusion. <https://huggingface.co/CompVis/stable-diffusion-v-1-4-original>, 2022. 2, 6
- [4] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, et al. ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022. 2
- [5] David Bau, Steven Liu, Tongzhou Wang, Jun-Yan Zhu, and Antonio Torralba. Rewriting a deep generative model. In *European Conference on Computer Vision (ECCV)*, 2020. 2
- [6] David Bau, Hendrik Strobelt, William Peebles, Jonas Wulff, Bolei Zhou, Jun-Yan Zhu, and Antonio Torralba. Semantic photo manipulation with a generative image prior. *arXiv preprint arXiv:2005.07727*, 2020. 2
- [7] Mikolaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. In *International Conference on Learning Representations (ICLR)*, 2018. 6
- [8] 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
- [9] Yinzh Cao and Junfeng Yang. Towards making systems forget with machine unlearning. In *2015 IEEE symposium on security and privacy*, pages 463–480. IEEE, 2015. 2
- [10] Nicholas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Schwag, Florian Tramèr, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. *arXiv preprint arXiv:2301.13188*, 2023. 1, 2, 4, 6, 7
- [11] Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. Quantifying memorization across neural language models. *arXiv preprint arXiv:2202.07646*, 2022. 2
- [12] Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *USENIX Security Symposium*, volume 267, 2019. 2
- [13] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom B Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting training data from large language models. In *USENIX Security Symposium*, volume 6, 2021. 2
- [14] Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, Michael Rubinstein, et al. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*, 2023. 1, 2
- [15] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2021. 2
- [16] Ming Ding, Wendi Zheng, Wenyi Hong, and Jie Tang. Cogview2: Faster and better text-to-image generation via hierarchical transformers. *arXiv preprint arXiv:2204.14217*, 2022. 2
- [17] 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, 4, 9
- [18] Rinon Gal, Moab Arar, Yuval Atzmon, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. Designing an encoder for fast personalization of text-to-image models. *arXiv preprint arXiv:2302.12228*, 2023. 2
- [19] Rinon Gal, Or Patashnik, Haggai Maron, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. Stylegan-nada: Clip-guided domain adaptation of image generators. *ACM Transactions on Graphics (TOG)*, 41(4):1–13, 2022. 2
- [20] Rohit Gandikota, Joanna Materzynska, Jaden Fiotto-Kaufman, and David Bau. Erasing concepts from diffusion models. *arXiv preprint arXiv:2303.07345*, 2023. 2
- [21] Antonio Ginart, Melody Guan, Gregory Valiant, and James Y Zou. Making ai forget you: Data deletion in machine learning. *Advances in neural information processing systems*, 32, 2019. 2
- [22] Aditya Golatkar, Alessandro Achille, Avinash Ravichandran, Marzia Polito, and Stefano Soatto. Mixed-privacy forgetting in deep networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 792–801, 2021. 2
- [23] Aditya Golatkar, Alessandro Achille, and Stefano Soatto. Eternal sunshine of the spotless net: Selective forgetting in deep networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9304–9312, 2020. 2
- [24] Zheng Gu, Wenbin Li, Jing Huo, Lei Wang, and Yang Gao. Lofgan: Fusing local representations for few-shot image generation. In *IEEE International Conference on Computer Vision (ICCV)*, 2021. 2
- [25] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 2
- [26] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *EMNLP*, 2021. 6
- [27] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2020. 2
- [28] Xun Huang, Arun Mallya, Ting-Chun Wang, and Ming-Yu Liu. Multimodal conditional image synthesis with product-of-experts gans. In *European Conference on Computer Vision*, pages 91–109. Springer, 2022. 2
- [29] Minguk Kang, Jun-Yan Zhu, Richard Zhang, Jaesik Park, Eli Shechtman, Sylvain Paris, and Taesung Park. Scaling up gans for text-to-image synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

- [30] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2020. 2
- [31] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. *arXiv preprint arXiv:2210.09276*, 2022. 2
- [32] Zhifeng Kong and Kamalika Chaudhuri. Data redaction from pre-trained gans. In *Workshop on Trustworthy and Socially Responsible Machine Learning, NeurIPS 2022*, 2022. 2, 3
- [33] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. *arXiv preprint arXiv:2212.04488*, 2022. 2, 4, 9
- [34] Yijun Li, Richard Zhang, Jingwan Lu, and Eli Shechtman. Few-shot image generation with elastic weight consolidation. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2020. 2
- [35] Bingchen Liu, Yizhe Zhu, Kunpeng Song, and Ahmed Elgammal. Towards faster and stabilized gan training for high-fidelity few-shot image synthesis. In *International Conference on Learning Representations (ICLR)*, 2021. 2
- [36] Cheng Lu, Yuhan Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *arXiv preprint arXiv:2206.00927*, 2022. 3
- [37] Elman Mansimov, Emilio Parisotto, Jimmy Lei Ba, and Ruslan Salakhutdinov. Generating images from captions with attention. In *International Conference on Learning Representations (ICLR)*, 2016. 2
- [38] Kevin Meng, David Bau, Alex J Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. In *Advances in Neural Information Processing Systems*, 2022. 2
- [39] Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. Mass-editing memory in a transformer. *arXiv preprint arXiv:2210.07229*, 2022. 2
- [40] Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D Manning. Fast model editing at scale. *arXiv preprint arXiv:2110.11309*, 2021. 2
- [41] Sangwoo Mo, Minsu Cho, and Jinwoo Shin. Freeze the discriminator: a simple baseline for fine-tuning gans. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshop*, 2020. 2
- [42] Quoc Phong Nguyen, Bryan Kian Hsiang Low, and Patrick Jaillet. Variational bayesian unlearning. *Advances in Neural Information Processing Systems*, 33:16025–16036, 2020. 2
- [43] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In *International Conference on Machine Learning (ICML)*, 2022. 1, 2
- [44] Yotam Nitzan, Kfir Aberman, Qirui He, Orly Liba, Michal Yarom, Yossi Gandelsman, Inbar Mosseri, Yael Pritch, and Daniel Cohen-Or. Mystyle: A personalized generative prior. In *SIGGRAPH ASIA*, 2022. 2
- [45] Yotam Nitzan, Michaël Gharbi, Richard Zhang, Taesung Park, Jun-Yan Zhu, Daniel Cohen-Or, and Eli Shechtman. Domain expansion of image generators. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2
- [46] Atsuhiro Noguchi and Tatsuya Harada. Image generation from small datasets via batch statistics adaptation. In *IEEE International Conference on Computer Vision (ICCV)*, 2019. 2
- [47] Utkarsh Ojha, Yijun Li, Jingwan Lu, Alexei A Efros, Yong Jae Lee, Eli Shechtman, and Richard Zhang. Few-shot image generation via cross-domain correspondence. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2
- [48] Xingang Pan, Xiaohang Zhan, Bo Dai, Dahua Lin, Chen Change Loy, and Ping Luo. Exploiting deep generative prior for versatile image restoration and manipulation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):7474–7489, 2021. 2
- [49] Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-shot image-to-image translation. *arXiv preprint arXiv:2302.03027*, 2023. 2
- [50] Ed Pizzi, Sreya Dutta Roy, Sugosh Nagavara Ravindra, Priya Goyal, and Matthijs Douze. A self-supervised descriptor for image copy detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14532–14542, 2022. 6
- [51] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 1
- [52] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *International Conference on Machine Learning (ICML)*, 2016. 2
- [53] Daniel Roich, Ron Mokady, Amit H Bermano, and Daniel Cohen-Or. Pivotal tuning for latent-based editing of real images. *ACM Transactions on Graphics (TOG)*, 42(1):1–13, 2022. 2
- [54] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 2
- [55] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. *arXiv preprint arXiv:2208.12242*, 2022. 2, 4, 9
- [56] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasempour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS*, 2022. 1, 2
- [57] Axel Sauer, Tero Karras, Samuli Laine, Andreas Geiger, and Timo Aila. Stylegan-t: Unlocking the power of gans for fast large-scale text-to-image synthesis. *arXiv preprint arXiv:2301.09515*, 2023. 2

- [58] Patrick Schramowski, Manuel Brack, Björn Deiseroth, and Kristian Kersting. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. 2023. [2](#)
- [59] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. [1](#), [2](#)
- [60] Ayush Sekhari, Jayadev Acharya, Gautam Kamath, and Ananda Theertha Suresh. Remember what you want to forget: Algorithms for machine unlearning. *Advances in Neural Information Processing Systems*, 34:18075–18086, 2021. [2](#)
- [61] Shawn Shan, Jenna Cryan, Emily Wenger, Haitao Zheng, Rana Hanocka, and Ben Y Zhao. Glaze: Protecting artists from style mimicry by text-to-image models. *arXiv preprint arXiv:2302.04222*, 2023. [1](#), [2](#)
- [62] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pages 3–18. IEEE, 2017. [2](#)
- [63] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning (ICML)*, 2015. [2](#), [3](#)
- [64] Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Diffusion art or digital forgery? investigating data replication in diffusion models. *arXiv preprint arXiv:2212.03860*, 2022. [1](#), [2](#), [7](#)
- [65] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations (ICLR)*, 2021. [3](#)
- [66] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *ICLR*, 2021. [2](#)
- [67] 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](#), [3](#), [6](#)
- [68] Ming Tao, Hao Tang, Songsong Wu, Nicu Sebe, Xiao-Yuan Jing, Fei Wu, and Bingkun Bao. Df-gan: Deep fusion generative adversarial networks for text-to-image synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. [2](#)
- [69] Sheng-Yu Wang, David Bau, and Jun-Yan Zhu. Sketch your own gan. In *IEEE International Conference on Computer Vision (ICCV)*, 2021. [2](#)
- [70] Sheng-Yu Wang, David Bau, and Jun-Yan Zhu. Rewriting geometric rules of a gan. *ACM SIGGRAPH*, 2022. [2](#)
- [71] Tengfei Wang, Ting Zhang, Bo Zhang, Hao Ouyang, Dong Chen, Qifeng Chen, and Fang Wen. Pretraining is all you need for image-to-image translation. *arXiv preprint arXiv:2205.12952*, 2022. [2](#)
- [72] Yaxing Wang, Abel Gonzalez-Garcia, David Berga, Luis Herranz, Fahad Shahbaz Khan, and Joost van de Weijer. Minegan: effective knowledge transfer from gans to target domains with few images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. [2](#)
- [73] Yaxing Wang, Chenshen Wu, Luis Herranz, Joost van de Weijer, Abel Gonzalez-Garcia, and Bogdan Raducanu. Transferring gans: generating images from limited data. In *European Conference on Computer Vision (ECCV)*, 2018. [2](#)
- [74] Chenfei Wu, Jian Liang, Lei Ji, Fan Yang, Yuejian Fang, Daxin Jiang, and Nan Duan. Nüwa: Visual synthesis pre-training for neural visual world creation. In *European Conference on Computer Vision (ECCV)*, 2022. [2](#)
- [75] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. [2](#)
- [76] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunnar Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2022. [1](#)
- [77] Han Zhang, Tao Xu, Hongsheng Li, Shaoting 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](#)
- [78] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*, 2023. [2](#)
- [79] 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](#)
- [80] Shengyu Zhao, Zhijian Liu, Ji Lin, Jun-Yan Zhu, and Song Han. Differentiable augmentation for data-efficient gan training. In *Conference on Neural Information Processing Systems (NeurIPS)*, volume 33, 2020. [2](#)
- [81] Minfeng Zhu, Pingbo Pan, Wei Chen, and Yi Yang. Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. [2](#)
- [82] Xiaojin Zhu, Andrew B Goldberg, Mohamed Eldawy, Charles R Dyer, and Bradley Strock. A text-to-picture synthesis system for augmenting communication. In *The AAAI Conference on Artificial Intelligence*, 2007. [2](#)