Forget-Me-Not: Learning to Forget in Text-to-Image Diffusion Models

Gong Zhang
University of Oregon, Eugene, OR, USA
gzhang7@uoregon.edu

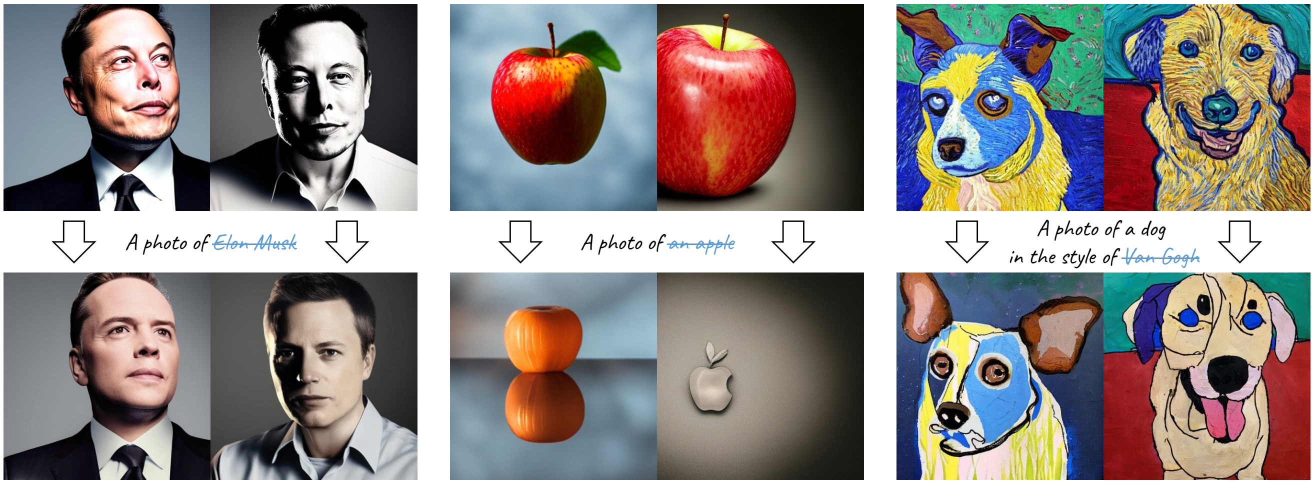


Figure 1: Given an arbitrary text-to-image model (*i.e.* Stable Diffusion), our approach, dubbed Forget-Me-Not, can guide the model to forget designated ID (left), object (middle), or style (right) while maintaining its capability of generating other contents. The top row shows the text-to-image results from Stable Diffusion v2.1, and the bottom row shows the results from Forget-Me-Not finetuned model that forgets target concepts (blue text that are crossed-out). As shown in the figure, these target concepts are successfully removed from the model without changing the quality of the outputs.

Abstract

Recently, personalized content generation has witnessed drastic progress thanks to the advance of large-scale text-to-image synthesis models as well as their efficient tuning algorithms. The study of its "counter-question", concept forgetting (removing any unwanted style or content from a trained synthesis model), has hence entered the public spotlight too due to the natural concerns of privacy leaking and copyright infringement and intrusion. In this paper, we investigate a toolkit of novel algorithms for effectively removing a designated concept from a pretrained text-to-image diffusion model. A key knob here is to observe the influence of text tokens on image synthesis through the cross-attention probability, hence inspiring us to exploit attention as the objective. Extensive benchmark experiments validate the promising performance of our proposed methods, that can guide the pretrained model to forget a specific concept represented by either a prompt or a set of images without significantly hurting the generative ability of the rest. Our codes and models will be released.

1. Introduction

Recent large-scale text-to-image diffusion models, such as [47, 50, 53, 62, 10, 49, 48, 54], have demonstrated impressive capabilities in generating diverse and realistic images containing complex objects and scenes. As a result, diffusion models have been incorporated into commercial art and graphic design systems [45, 32, 41, 56], receiving a huge amount of usage traffic and attention from the public, creating far-reaching social impacts.

However, this popularity has amplified the potential fairness, legal, and faulty risks associating with those models. Firstly, there is a risk of users generating misleading and biased content and leading to unfair outcomes [17, 64, 29, 5, 33]. Secondly, there is a risk of models generating images that infringe upon privacy or has copyright issues [7, 56]. Lastly, there is also the risk of semantic drift, where the semantic of a common word can shift from its intended meaning [4].

These risks stem from the billion-sized mega-datasets that are beyond the reach of human annotation [6]. As a result, it is almost impossible to address harmful content, privacy and copyright concerns through data filtering and full retraining. Attempts to address the retraining includes few-shot domain adaptation [20, 66, 61]. In practice, people curate a "relatively small" clean dataset and use it for finetuning to alleviate these concerns. However, collecting a clean dataset is time-consuming, and training on extra data to erase inappropriate content is not very efficient, because it is not directly targeting on "dirty" content but enhancing the opposite "clean" content. Furthermore, this continue training runs the risk of compromising the original generative ability.

In principle, we aim to prevent inappropriate content from appearing in generated images by directly targeting the underlying concepts behind a series of problematic images. In other words, we seek to make the model forget "dirty" concepts, rather than to learn extra "clean" concepts to overshadow them. To fulfill our goal, we introduce *concept forgetting* in the context of text-to-image diffusion models and propose *Forget-Me-Not*, a lightweight framework designed to make Stable Diffusion models forget a specific concept by providing a few real/generated images or a prompt of the concept. Our approach is based on two key insights. First, text-to-image generative models 'remember' a learned concepts by cross-attending to prompts. Second, the cross-attention scores between pixels and prompts have a strong correlation with the realization of a concept in the generated images. We demonstrate that these cross-attention scores can be used as the sole optimization objective for finetuning Stable Diffusion models, without relying on negative log-likelihood loss on noise prediction. We call it Attention Re-steering loss. By minimizing the cross-attention paid to a target concept in prompts, we achieve concept forgetting without a curated clean dataset or extra tools to identify target concept. In summary, our contributions are:

- We propose *Forget-Me-Not*, an ad-hoc concept forgetting framework, for text-to-image generative models that use cross attention as conditioning. Compared with curating a clean dataset and continuing training a large scale model on it, our approach saves significant human labor and computation costs.

- With target concept forgotten, we still achieve competitive generation quality as original model and minimal impact on other concepts. We find that our method allows for precise forgetting by correcting just target concept while keeping others (e.g. body pose and background) relatively intact (Fig. 1).

- We evaluate the performance of our method both qualitatively and quantitatively on a new benchmark *ConceptBench* which includes multiple concept types along with testing prompts and a new metric *Memorization Score*. Our work lays the foundation of future concept forgetting research for generative model.

2. Related Works

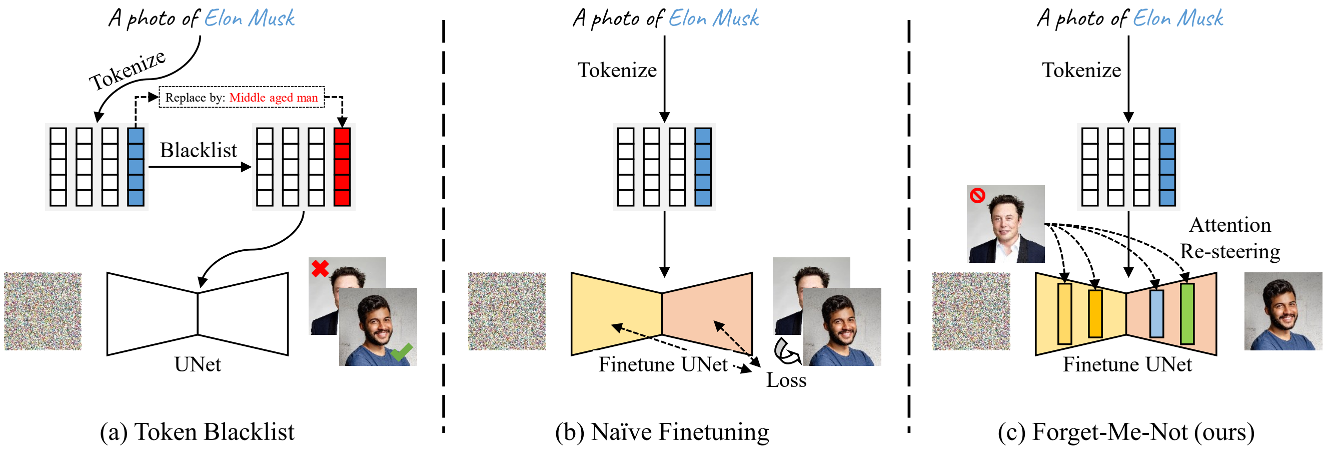
2.1. Conditional Diffusion Models for Text-to-Image

Synthetic image generation has long been a fascinating, yet challenging field, aiming to create new images that are similar to real images. In the past several year, we have witnessed the rapid advance of it from unconditional generative models to conditional generative models with powerful architectures of auto-regressive model [49, 63], GAN [8, 28, 26, 58, 65] and diffusion process [23, 43, 38, 16, 3, 57]. Early works focus on unconditional, single-category data distribution modeling , such as hand-written digits, certain species of animals, and human faces [15, 12, 27, 37]. Though, unconditional models quickly achieves photo realistic results among single-category data, it's shown that mode collapsing issue usually happens when extending data distributions to multiple-category or real image diversity [8, 40, 1].

To tackle the model collapsing problem, conditional generative model has been introduced. Since then, different types of data have been used as the conditioning for generative models, e.g. class labels, image instances, and even networks [8, 42] etc. At the same time, CLIP [46, 25], a large-scale pretrained image-text contrastive model, provides a text-image prior of extremely high diversity, which is discovered to be applicable as the conditioning for generative model [44, 14, 35]. Nowadays, DALL-E 2 [48] and Stable Diffusion [50] are capable of generating high quality images solely conditioning on free-form texts, inheriting the diversity of billions of real images scraping from the Internet. Subsequently, a line of work seeks to efficiently adapt the massive generative model to generate novel rendition of an unseen concept represented by a small reference set, leveraging the great diversity. Dreambooth [52] proposed to adapt the model by finetuning all of its weights, while it requires enormous storage to save newly adapted weights. Textual Inversion [18] and LoRA [24] ameliorate the issue by adapting the model by adding a small set of extra weights.

2.2. The Risk of User Data Leakage and Machine Unlearning

However, this great diversity comes at a price. It incurs potential risk of privacy leakage and copyright infringement. [7, 56] have successfully retrieved samples from Sta-



(a) Token Blacklist

(b) Naïve Finetuning

(c) Forget-Me-Not (ours)

Figure 2: This figure shows two baseline forgetting methods and our proposed Forget-Me-Not. The target concept to forget is Elon Musk. One baseline is (a) Token Blacklist that simply replaces the target token with a different one. The other baseline is (b) Naive Fintuning in which instead of replacing tokens, it finetunes model weights so that the new weights generate outputs containing unrelated concepts. Our method (c) Forget-Me-Not utilizes Attention Re-steering in which we finetune only UNet to minimize each of the intermediate attention maps associated with the target concepts to forget.

ble Diffusion that are highly faithful to real training examples. Therefore, being able to forget/unlearn certain concept in a model without hurting the generative ability for the rest is of both research and practical interests. Similar topics have been seen in fields other than conditional generative modeling. In model-agnostic meta-learning, [2] noted selectively forgetting the influence of prior knowledge in a network improves the performance in adapted tasks. [9, 36, 39, 13] explores the unlearning of a set of requested data points in a pretraind model.

Our work differs from existing forgetting and unlearning works in a few aspects. First, we study forgetting in the context of text-to-image generative models. Second, we are deleting not only requested data points represented by a small reference set, but the concept behind those data points, which possesses significant impact in text-to-image generation due to the fact that it's almost impossible to enumerate all prompts and synonyms relating to a concept.

3. Method

3.1. Preliminaries

Latent Diffusion Models An autoencoder [30, 21] is first used to encode pixel images into compact latent representations with lower resolution. Then, diffusion process operates on that latent space insted of pixel space to save enormous amount of computation cost. The encoded latent representation $\mathbf{x}_0 \sim q(\mathbf{x})$, is first converted to isotropic Gaussian noise $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ through a T step forward diffusi‘on process. β_t is a variance schedule. Given $\mathbf{x}_0$, we can sample $\mathbf{x}_t$ at any forward step t in a closed form.

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I})$$
$$q(\mathbf{x}_{1:T}|\mathbf{x}_0) = \prod_{t=1}^{T} q(\mathbf{x}_t|\mathbf{x}_{t-1})$$

Diffusion model is trained to recover real data distribution $\mathbf{x}_0$ from Gaussiion noise $\mathbf{x}_T$ in reverse diffusion process. Both forward and reverse process are assumed to be Markovian, thus p_θ is a trained model to estimate the conditional probabilities of previous state at $t-1$ given current state at step t.

$$p_\theta(\mathbf{x}_{0:T}) = p(\mathbf{x}_T)\prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$$
$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_t, t), \boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t))$$

Variational lower bound is used to optimize negative log-likelihood loss.

$$L_{\text{VLB}} = \mathbb{E}_{q(\mathbf{x}_{0:T})}\left[\log\frac{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}{p_\theta(\mathbf{x}_{0:T})}\right] \geq -\mathbb{E}_{q(\mathbf{x}_0)}\log p_\theta(\mathbf{x}_0)$$

Conditioning via Cross Attention Conditioning introduces extra information y to diffusion process. In the context of Stable Diffusion, the generator model becomes $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathcal{E}(y))$ with text encoder $\mathcal{E}$ converting a prompt into a sequence of token embeddings. In practice, p_θ is often modeled as a UNet [51] $\mathcal{U}$ and text encoder $\mathcal{E}$ is a CLIP text model [46]. At each resolution of UNet, there is a cross attention layer responsible for fusing conditional signals into diffusion process. The fusion is implemented as muti-head QKV attention [60]. The hidden state of $\mathcal{U}$ is

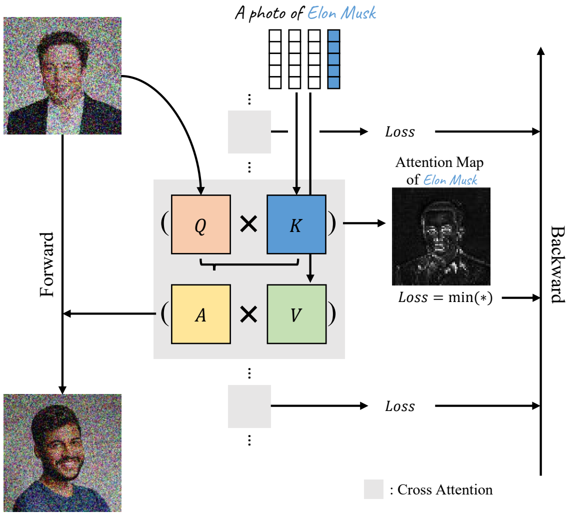


Figure 3: This figure shows the Attention Re-steering we proposed in our Forget-Me-Not method, in which we set the objective function to minimize the attention maps of target concepts (*i.e.* Elon Musk in this case) and correspondingly finetune the network.

mapped to Q, while text embeddings from $\mathcal{E}$ is mapped to K and V. The fused output h is calculated as:

$$h = \text{softmax}(\frac{QK^T}{\sqrt{d}})V$$

where softmax$(\cdot)$ corresponds to the attention map between Q and K, namely cross attention from pixels to text tokens.

3.2. Problem Definition

What is Concept Concept is fundamental building block of people's understanding and knowledge towards the world around us. Different fields have various notions of what constitutes a concept. In the context of text-to-image, a concept is a generic idea that can be expressed through prompts or images. For example, a prompt for "Elon Musk" and a set of images of Elon Musk both communicate the same generic idea - the identity of Elon Musk.

To represent this generic idea in a tangible way, we use the last hidden states of text encoder and call them *concept embeddings*. Obtaining this concept embedding from prompts is trivial, simply running the prompts through text encoder. For obtaining from images, it's detailed in Section 3.3: Concept Inversion.

Concept Forgetting We adopt the following definition of concept forgetting: *A text-to-image generator forgets a concept if its generated images of a prompt no longer contain the concept which is previously expected from the prompt, while maintaining the semantic and visual quality for other concepts.*

As discussed above, we obtain concept embeddings from the text encoder, and we want to avoid altering the fragile pretrained concept embedding space in order for concept memorization measurement, discussed in Seciotn 4.4. This means that we have limited control over the embeddings of a concept conveyed through prompts. To achieve concept forgetting, we focus on reducing the sensitivity of the UNet to target concepts via cross-attention where the perception of concepts is happening.

Intuitively, adjusting the attention of pixel features to concept embeddings can enhance or diminish the presence of the concept in the generated image [11, 22, 59]. Therefore, we formulate an optimization objective to minimize the attention scores of the target concept embeddings. We hypothesize that by training the model to decrease attention scores, it will become less sensitive to the target concept and ultimately forget it.

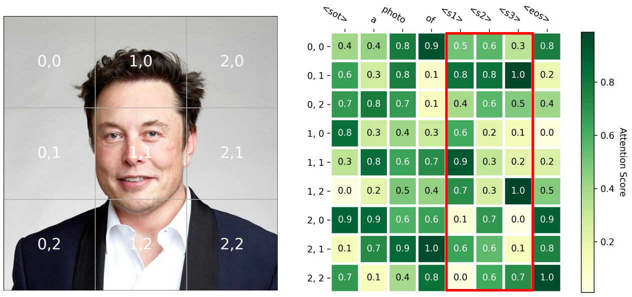
3.3. Forget-Me-Not

To achieve concept forgetting, we incorporate the attention score minimization objective into the standard text-to-image training process. Our method supports both prompt-based concepts and image-based concepts, with the latter obtained through Concept Inversion. An overview of *Forget-Me-Not* is shown in Figure 2(c).

Algorithm 1 Training of Forget-Me-Not

```
Require: A prompt $\mathcal{P}$, token indices $\mathcal{I}$ of target concept, a
    set of reference images $\mathcal{R}$ of the concept, parameters to
    optimize $\theta$, text encoder $\mathcal{E}$ and UNet.
repeat
    $t \sim \text{Uniform}([1 \dots T])$
    $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
    $\mathbf{x}_0 \sim \mathcal{R}$
    $\mathbf{x}_t \leftarrow \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t}\boldsymbol{\epsilon}$
                                    ▷ $\bar{\alpha}_t$ is noisy variance schedule
    $\mathbf{x}_{t-1}, A_t \leftarrow \text{UNet}(\mathbf{x}_t, \mathcal{E}(\mathcal{P}), t)$
    $\mathcal{S} \leftarrow \{\}$
    for $i \in \mathcal{I}$ do
        $s_i \leftarrow A_t[:, :, i]$
        $\mathcal{S}.\text{insert}(s_i)$
    end for
    Take gradient descent step on
        $\nabla_\theta \sum_s^{\mathcal{S}} \|s\|^2$
until Concept fades in oblivion
```

Concept Inversion We adapt the idea of textual inversion [18] for Concept Inversion. This involves taking a set of images and inverting them into dedicated token embeddings. Those token embeddings are initialized from tokens relating to target concept or randomness. Once optimization is complete, dedicated token embeddings capture the concept and can be used to reconstruct novel rendition of the concept. These inverted tokens can then be combined with arbitrary text snippets to form an ordinary prompt. The



	<sot>	a	photo	of	<s1>	<s2>	<s3>	<eos>
0, 0	0.4	0.4	0.8	0.9	0.5	0.6	0.3	0.8
0, 1	0.6	0.3	0.8	0.1	0.8	0.8	1.0	0.2
0, 2	0.7	0.8	0.7	0.1	0.4	0.6	0.5	0.4
1, 0	0.8	0.3	0.4	0.3	0.6	0.2	0.1	0.0
1, 1	0.3	0.8	0.6	0.7	0.9	0.3	0.2	0.2
1, 2	0.0	0.2	0.5	0.4	0.7	0.3	1.0	0.5
2, 0	0.9	0.9	0.6	0.6	0.1	0.7	0.0	0.9
2, 1	0.1	0.7	0.9	1.0	0.6	0.6	0.1	0.8
2, 2	0.7	0.1	0.4	0.8	0.0	0.6	0.7	1.0

Figure 4: An illustration of attention score matrix obtained from cross attention. Scores enclosed in red rectangle indicates the attention paid by pixels to inverted concept, which our method minimizes.

prompt is then processed through the text encoder to obtain the concept embeddings.

Attention Re-steering Optimization An overview is given in Algorithm 1 and illustrated in Figure 3. At timestep t, a noisy latent $\mathbf{x}_t$ and a sequence of concept embeddings produced by text encoder $\mathcal{E}$ from $\mathcal{P}$ are fed into the forward pass of UNet. It outputs predicted noises at $\mathbf{x}_{t-1}$ and a pyramid of attention score matrices, as in step 7 of Algorithm 1. Due to the down-sampling and up-sampling structure of UNet, there are 16 cross attention layers in total resulting in 16 attention score matrix with various resolutions in A_t. Figure 4 is an conceptual illustration of an attention score matrix with only 9 pixels, in other words, at resolution 3x3. The rows and columns of the matrix are pixels and concept tokens. With 9 pixels and 8 tokens, this 9x8 matrix contains the attention of each pixel feature paid to each concept token embeddings. Here goes our intuition that desensitizing the concept embeddings by resteering attention scores shall make the model forget that concept. The red box in Figure 4 denotes attention scores of an concept embeddins represented in $<$s1$>$ $<$s2$>$ $<$s3$>$. Once corresponding scores in all attention score matrices are gathered, they are summed and norm-ed to be the loss, as in Algorithm 1 lines 9-14 and in Figure 3 loss arrows from cross attention blocks to backward.

Importantly, we do not use the classical negative log-likelihood loss, and our method demonstrates that attention scores can be a viable objective in fine-tuning diffusion models.

4. Experiments

In this section, we first introduce our proposed ConceptBench and related baselines. We then demonstrate the effectiveness of both our method via images and scores. Finally, we introduce concept correction and NSFW removal as two derived applications from Forget-Me-Not.

4.1. ConceptBench

To better evaluate concept forgetting, we introduce a new benchmark called *ConceptBench*. Recall that several existing benchmarks such as [34, 53] help evaluate the overall generation quality. Yet none of them are designed to measure how well a model can memorize and forget. *ConceptBench* adopts samples from LAION [55], forming up six categories, and each includes six instances.

We included the person and animation categories to ensure *ConceptBench* covers sufficient scenarios. These two categories are particularly interesting due to the potential privacy or copyright issues associated with generating images of people or famous animated characters. For the person category, we have selected instances based on their diversity and frequency in the LAION dataset, while for the animation category, we have chosen famous instances from different styles, including characters like Iron Man or Superman that have realistic representations. This allows us to explore the relationship between the concept embedding and the visual representation.

Furthermore, we have also included hierarchical instances in the animal and plant categories. The animal category includes both coarse instances, such as ”dog,” and fine instances, such as ”husky.” It is important to evaluate how well a model can generate images of hierarchical instances, as this is a challenging task that requires the model to capture both the general and specific features of the concept.

Lastly, we have included the style and relation categories to evaluate how well our method can generate images of abstract concepts. These categories allow us to test the model’s ability to generate images of concepts that are beyond objects, to more general abstracts and styles.

4.2. Baseline

In view of the multi-component nature of Stable Diffusion models, there are several naive methods that can be used to superficially remove a concept from them, such as blacklisting keywords in prompts, removing specific tokens from the tokenizer dictionary, or tuning the model with unrelated images to divert the target concept, as illustrated in Figure 2(a)(b). However, these methods can result in a significant deterioration and shifting of the model’s generation capability. Removing tokens from the dictionary can alter the tokenization of prompts where those tokens were previously used and affect the generation of other prompts with overlapping tokens. For instance, removing tokens of “Hillary Clinton” could lead to dysfunctionality in generating “Bill Clinton”. Naive finetuning to forget with unrelated images explicitly overwrites the visual representation of a concept with extra data and runs the risk of compromising existing concept space, as shown in Figure 5. Moreover, it is impossible to exhaust test all relation-based concepts for

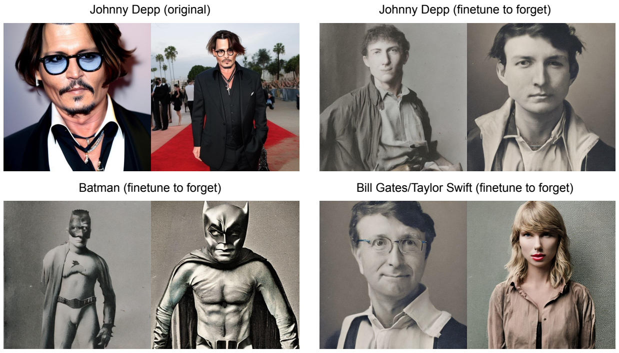

Johnny Depp (original)
Johnny Depp (finetune to forget)
Batman (finetune to forget)
Bill Gates/Taylor Swift (finetune to forget)

Figure 5: Finetuning to forget concept ”Johnny Depp” with unrelated images of ”a photo of man”. This method distorts other concepts with visual details of selected unrelated images.

blacklisting or finetuning.

4.3. Qualitative Comparison

We present the results of concept forgetting from our benchmark, illustrated in Figure 1. The first row showcases three popular concepts from Stable Diffusion [50]: ”Elon Musk”, ”Taylor Swift”, and ”apple”. The first two concepts represent specific person identities, and *Forget-Me-Not* successfully removes these identities without compromising other features. Specifically, the hypernym concepts of ”Elon Musk” and ”Taylor Swift” are preserved perfectly. After removing their respective identities, ”Elon Musk” remains a male person, while ”Taylor Swift” remains a female person. Moreover, visual details such as poses and clothing are also retained. In contrast, tuning with unrelated images often results in a complete shift to another concept represented by the unrelated images used for tuning. This tends to overfit and distort the concept space learned by large-scale pretraining, manifested in Figure 5. However, our method can pinpoint the target concept and leave the rest intact.

In Figure 1 column 3, we demonstrate the unique structure of concepts related to ”apple”. Naturally, it includes parallel concepts such as ”fruit apple” and ”brand Apple”, which are homonyms. These concepts compete to generate either an apple fruit or something associated with the Apple brand. We used a few generated apple fruit images for as forgetting inputs. The results reveal that both the parallel concepts exist within ”apple”. The first image depicts a fruit that resembles an orange, while the second image shows the logo of Apple. In both cases, the visual details of image background is preserved while the concept of apple fruit is forgotten.

In Figure 6, the Multi-concepts model of Elon Musk and Taylor Swift demonstrates our method’s ability to perform multi-concept forgetting. As shown in the first row, both target concepts have been forgotten. We evaluated the impacts of forgetting specific concepts on other related concepts, examining four related concepts to Elon Musk and Taylor Swift - man, woman, Bill Gates, and Emma Watson. As shown, *Forget-Me-Not* achieved good content preservation and visual quality. However, we observed minor pose and style changes in man and Bill Gates. Based on these findings, we posit that our approach may have a greater impact on closely related concepts than on others. Additionally, the last row shows that a new painting style is emerging after forgetting Piccaso and Van Gogh styles.

4.4. Quantitative Analysis

Memorization Measurement Textual inversion [19] can be used to identify the token embeddings that best correspond to images. We leverage this technique to measure the concept embedding changes of anchor images toward a reference before and after forgetting. This changes can be seen as generative model’s memorization level of a concept, we call it *Memorization Score*.

In the case of “Elon Musk”, prompt “Elon Musk” is used as reference. Its concept embedding ($\mathbf{emb}_r$) is obtained by running it through text encoder. Next, we invert the same anchor images of Elon Musk using *original model* and *forgetting model* respectively. Concept embeddings of anchor images are retrieved by running inverted tokens through text encoder. There are two of them: original($\mathbf{emb}_o$) and forgetting($\mathbf{emb}_f$). We use only the pooler token of concept embeddings for measurement. Concept embedding changes is measured as difference between $\cos(\mathbf{emb}_r, \mathbf{emb}_o)$ and $\cos(\mathbf{emb}_r, \mathbf{emb}_f)$. A decrease indicates successful forgetting. We show memorization scores for person of “Elon Musk”, animation of “Mickey Mouse”, plant of ”Apple”, animal of ”Dog” and style of ”Picasso” in Table 1. More results can be found in Supplementary.

Concept	Initial Mem Score	Forgetting Mem Score
Elon Musk	0.863	0.585
Mickey Mouse	0.925	0.898
Apple	0.642	0.345
Dog	0.842	0.676
Picasso	0.956	0.589

Table 1: Memorization scores for person of “Elon Musk”, animation of “Mickey Mouse”, plant of ”Apple”, animal of ”Dog” and style of ”Picasso”

4.5. Concept Correction

It has been observed that in text-to-image models, the semantics of a prompt are often dominated by the one with the most number of image-text examples in the training set, resulting in the suppression of inferior semantics during inference. Figure 7 exemplifies this scenario, where generation is dominated by a concept that is strongly correlated

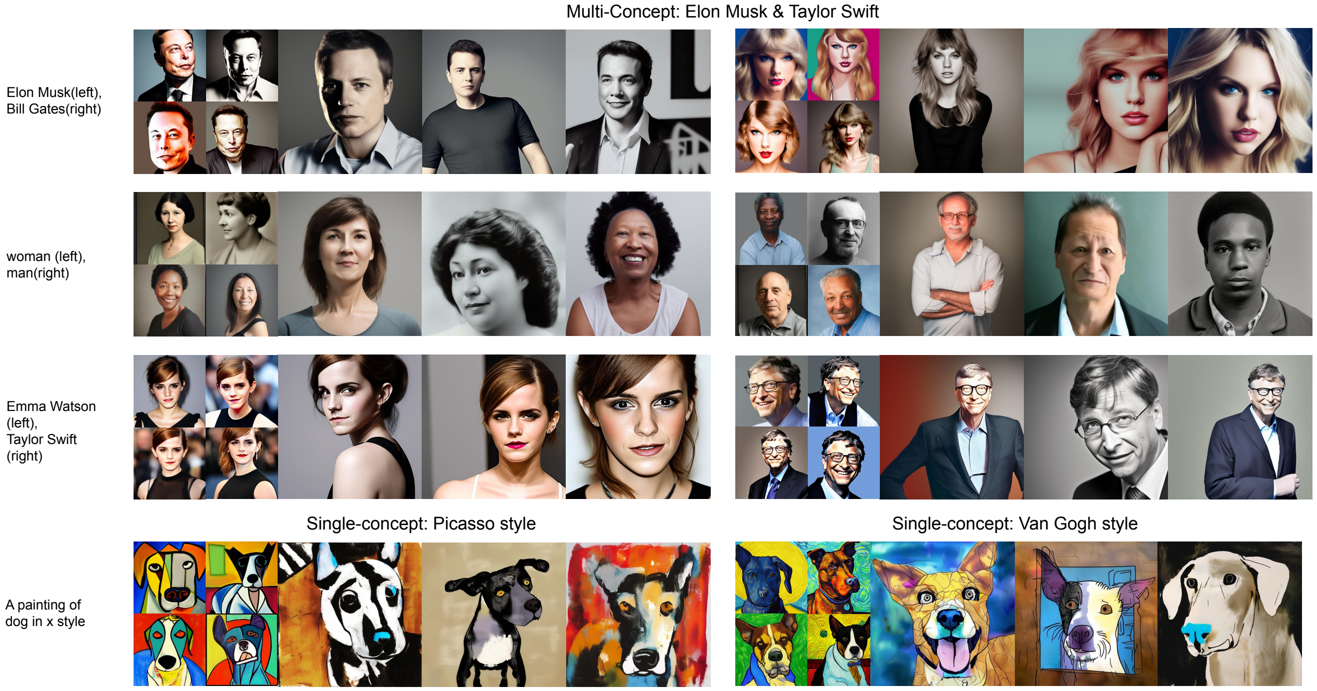

Multi-Concept: Elon Musk & Taylor Swift
Elon Musk(left), Bill Gates(right)
woman (left), man(right)
Emma Watson (left), Taylor Swift (right)
Single-concept: Picasso style
Single-concept: Van Gogh style
A painting of dog in x style

Figure 6: Results of concept forgetting using our method. The first 2x2 grid shows the original samples in Stable Diffusion. The subsequent 3 images are sampled after concept forgetting, using the same prompt. The top 3 rows are from a multi-concept model targeting both Elon Musk and Taylor Swift, demonstrating the multi-concept forgetting capability. Control concepts such as Bill Gates and Emma Watson manifest that our approach has minimal impact on concepts other than target ones. The last row shows two single-concept model of styles.

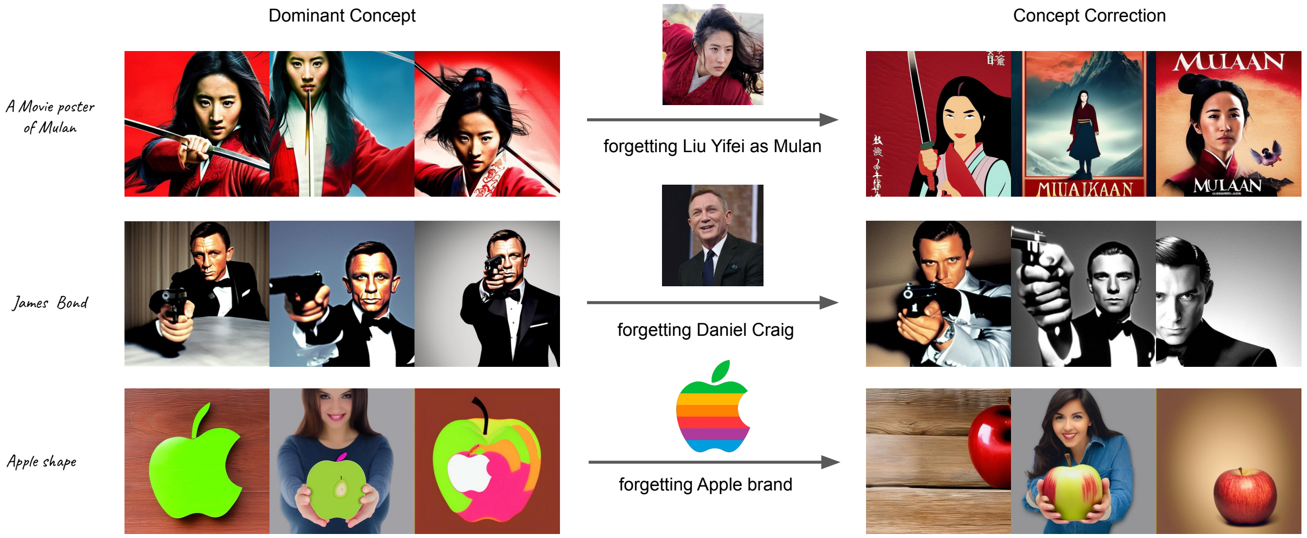


Figure 7: Concept Correction: Once the dominant concept has been diminished by our method, the lesser concepts of an semantic-rich prompt become more prominent in generated results.

with a prompt due to unbalanced training examples. In the case of the James Bond series, the generation results are overwhelmingly dominated by Daniel Craig, as shown in the middle row. However, our method manages to diminish the most prominent semantic in the prompt, i.e., Daniel Craig, and make other James Bond actors visible. Similarly, in the case of Mulan series and the homonym of "apple", where the apple fruit and Apple brand are competing with each other, our method successfully corrects target concept in generated images .

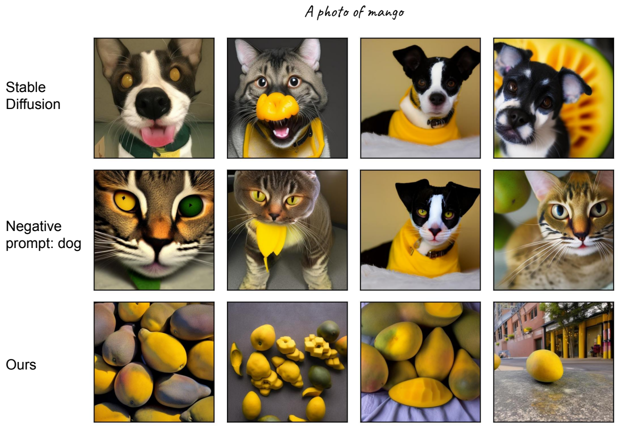


Figure 8: In concept correction, our method has the advantage of comprehensive forgetting over negative prompt.

Negative prompt is a technique used in text-to-image synthesis to eliminate unwanted concepts associated with a prompt. However, their use can result in negative impacts on other aspects of the image, such as changes to its structure and style. Furthermore, negative prompts fail to correct undesirable concepts under certain circumstances. For example, in Figure 8, "a photo of a mango" consistently generates dog images. This is because the name "mango" is commonly used as a pet name for dogs, and people upload photos of their dogs to the internet, which are collected as training data. In this case, using a negative prompt for dogs would be ineffective, as mango is also a popular cat name, creating the problem of endlessly expanding negative prompts. However, our method successfully brings back the mango fruit by forgetting the connection between "a photo of mango" and dog/cat images.

4.6. NSFW Removal

In this section, we examine the effectiveness of our method in a real case of removing harmful contents. NSFW is an internet shorthand for "not safe for work," used for indicating contents that are not wished to be seen in the public. Such content may include material even offensive for adult audiences. However, they inevitably present in large datasets such as LAION [6], even though NSFW detector has been used [31]. Stable Diffusion, trained on LAION, is known for generating NSFW content when prompted with certain triggers.

To evaluate our method, we use a well-known NSFW-triggering prompt, "a photo of naked" in Stable Diffusion v2.1 model. Using EulerAncestralDiscreteScheduler, inference-step 50, and scale-guidance 8, the model consistently generates images containing nude individuals. We use eight generated NSFW images as input for training *Forget-Me-Not*.

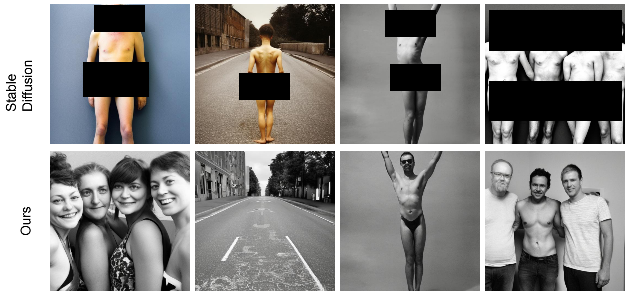


Figure 9: Results of removing NSFW contents triggered by "naked". Faces and sensitive parts are blacked out.

The results, shown in Figure 9, indicate that the concept of "naked" has been successfully forgotten. Notably, the second row shows that all sensitive visual cues have been changed in different ways. The first example changes abruptly from a naked man to a group of smiling women. In the second example, NSFW individual has been removed from the scene. The last two examples render clothed people, making them safe. Overall, our method achieved efficient forgetting of NSFW content without the need for additional data or the assistance of third-party NSFW detectors.

5. Conclusion

In this study, we investigate concept forgetting in text-to-image generative models and introduce *Forget-Me-Not*. This lightweight approach enables ad-hoc concept forgetting using only a few either real or generated concept images. Our experiments demonstrate that *Forget-Me-Not* is successful in diminishing target concepts in Stable Diffusion. Additionally, we introduce *ConceptBench* and *Memorization Score* as evaluation metrics. Overall, our work provides a foundation for further research on concept forgetting in text-to-image generation.

6. Social Impact & Limitations

Social Impact Our research has a positive social impact by offering an effective and cost-efficient method to remove harmful and biased concepts in text-to-image generative models. These models are rapidly becoming the backbone of popular AI art and graphic design tools, used by a growing number of people. Thus, our research takes a small step towards promoting fairness and privacy protection in AI tools, ultimately benefiting society as a whole.

Limitations While our approach performs well on concrete concepts in *ConceptBench*, it faces challenges in identifying and forgetting abstract concepts. Additionally, successful forgetting may require manual interventions, such as concept-specific hyperparameter tuning.

References

[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan, 2017.

[2] Sungyong Baik, Seokil Hong, and Kyoung Mu Lee. Learning to forget for meta-learning, 2019.

[3] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, Tero Karras, and Ming-Yu Liu. ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers, 2022.

[4] Lisa Beinborn and Rochelle Choenni. Semantic drift in multilingual representations, 2019.

[5] Hugo Berg, Siobhan Mackenzie Hall, Yash Bhalgat, Wonsuk Yang, Hannah Rose Kirk, Aleksandar Shtedritski, and Max Bain. A prompt array keeps the bias away: Debiasing vision-language models with adversarial learning, 2022.

[6] Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. Multimodal datasets: misogyny, pornography, and malignant stereotypes, 2021.

[7] Nicholas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramèr, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models, 2023.

[8] Arantxa Casanova, Marlène Careil, Jakob Verbeek, Michal Drozdzal, and Adriana Romero-Soriano. Instance-conditioned gan, 2021.

[9] Sungmin Cha, Sungjun Cho, Dasol Hwang, Honglak Lee, Taesup Moon, and Moontae Lee. Learning to unlearn: Instance-wise unlearning for pre-trained classifiers, 2023.

[10] 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.

[11] Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models, 2023.

[12] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains, 2019.

[13] Vikram S Chundawat, Ayush K Tarun, Murari Mandal, and Mohan Kankanhalli. Zero-shot machine unlearning, 2022.

[14] Katherine Crowson, Stella Biderman, Daniel Kornis, Dashiell Stander, Eric Hallahan, Louis Castricato, and Edward Raff. Vqgan-clip: Open domain image generation and editing with natural language guidance, 2022.

[15] Li Deng. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE signal processing magazine*, 29(6):141–142, 2012.

[16] Tim Dockhorn, Arash Vahdat, and Karsten Kreis. Genie: Higher-order denoising diffusion solvers, 2022.

[17] Felix Friedrich, Patrick Schramowski, Manuel Brack, Lukas Struppek, Dominik Hintersdorf, Sasha Luccioni, and Kristian Kersting. Fair diffusion: Instructing text-to-image generation models on fairness, 2023.

[18] 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, 2022.

[19] 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, 2022.

[20] Giorgio Giannone, Didrik Nielsen, and Ole Winther. Few-shot diffusion models, 2022.

[21] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners, 2021.

[22] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control, 2022.

[23] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020.

[24] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021.

[25] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. OpenCLIP, 7 2021.

[26] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks, 2021.

[27] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks, 2018.

[28] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan, 2019.

[29] Patrik Joslin Kenfack, Daniil Dmitrievich Arapov, Rasheed Hussain, S. M. Ahsan Kazmi, and Adil Mehmood Khan. On the fairness of generative adversarial networks (gans), 2021.

[30] Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2013.

[31] LAION-AI. Clip-based-nsfw-detector. `https://github.com/LAION-AI/CLIP-based-NSFW-Detector`.

[32] lexica. lexica.art. `https://lexica.art/`.

[33] Zhiheng Li, Anthony Hoogs, and Chenliang Xu. Discover and mitigate unknown biases with debiasing alternate networks, 2022.

[34] 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.

[35] Xihui Liu, Dong Huk Park, Samaneh Azadi, Gong Zhang, Arman Chopikyan, Yuxiao Hu, Humphrey Shi, Anna Rohrbach, and Trevor Darrell. More control for free! image synthesis with semantic diffusion guidance, 2021.

[36] Yang Liu, Zhuo Ma, Ximeng Liu, Jian Liu, Zhongyuan Jiang, Jianfeng Ma, Philip Yu, and Kui Ren. Learn to forget: Machine unlearning via neuron masking, 2020.

[37] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Large-scale celebfaces attributes (celeba) dataset. *Retrieved August*, 15(2018):11, 2018.

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

[39] Ananth Mahadevan and Michael Mathioudakis. Certifiable machine unlearning for linear models, 2021.

[40] Luke Metz, Ben Poole, David Pfau, and Jascha Sohl-Dickstein. Unrolled generative adversarial networks, 2016.

[41] midjourny. midjourny.com. `https://www.midjourney.com/`.

[42] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets, 2014.

[43] Alex Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models, 2021.

[44] 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, 2021.

[45] prisma ai.com. lensa ai. `https://prisma-ai.com/lensa`.

[46] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

[47] 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.

[48] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022.

[49] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation, 2021.

[50] 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, June 2022.

[51] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015.

[52] 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. 2022.

[53] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022.

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

[55] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv:2210.08402*, 2022.

[56] Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Diffusion art or digital forgery? investigating data replication in diffusion models, 2022.

[57] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations, 2020.

[58] Ming Tao, Hao Tang, Fei Wu, Xiao-Yuan Jing, Bing-Kun Bao, and Changsheng Xu. Df-gan: A simple and effective baseline for text-to-image synthesis, 2020.

[59] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation, 2022.

[60] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.

[61] Jiayu Xiao, Liang Li, Chaofei Wang, Zheng-Jun Zha, and Qingming Huang. Few shot generative model adaption via relaxed spatial structural alignment, 2022.

[62] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan 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.

[63] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Ben Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu. Scaling autoregressive models for content-rich text-to-image generation, 2022.

[64] M. Zameshina, O. Teytaud, Fabien Teytaud, Vlad Hosu, Nathanael Carraz, Laurent Najman, and Markus Wagner. Fairness in generative modeling. In *Proceedings of the Genetic and Evolutionary Computation Conference Companion*. ACM, jul 2022.

[65] Han Zhang, Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang. Cross-modal contrastive learning for text-to-image generation, 2021.

[66] Jingyuan Zhu, Huimin Ma, Jiansheng Chen, and Jian Yuan. Few-shot image generation with diffusion models, 2022.