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

Gong Zhang[1*], Kai Wang[1*], Xingqian Xu[1,3], Zhangyang Wang[2,3], Humphrey Shi[1,3]
[1]SHI Labs @ Georgia Tech & UIUC, [2]UT Austin, [3]Picsart AI Research (PAIR)
https://github.com/SHI-Labs/Forget-Me-Not

## Abstract

*The significant advances in applications of text-to-image generation models have prompted the demand of a post-hoc adaptation algorithms that can efficiently remove unwanted concepts (e.g. privacy, copyright, and safety) from a pretrained models with minimal influence on the existing knowledge system learned from pretraining. Existing methods mainly resort to explicitly finetuning unwanted concepts to be some alternatives such as their hypernyms or antonyms. Essentially, they are modifying the knowledge system of pretrained models by replacing unwanted to be something arbitrarily defined by user. Furthermore, these methods require hundreds of optimization steps, as they solely rely on denoising loss used for pretraining. To address above challenges, we propose Forget-Me-Not, a model-centric and efficient solution designed to remove identities, objects, or styles from a well-configured text-to-image model in as little as 30 seconds, without significantly impairing its ability to generate other content. In contrast to existing methods, we introduce attention re-steering loss to redirect model's generation from unwanted concepts to those are learned during pretraining, rather than being user-defined. Furthermore, our method offers two practical extensions: a) removal of potentially harmful or NSFW content, and b) enhancement of model accuracy, inclusion and diversity through concept correction and disentanglement.*

## 1. Introduction

In recent advancements, text-to-image models [6, 15, 36–38, 41, 49, 50] have exhibited remarkable capabilities in generating high-resolution images based on textual descriptions. Notably, diffusion models, exemplified by DALL-E 2 [37] and Stable Diffusion [38], have satisfied commercial-grade productization standards, paving the way for a plethora of applications tailored for end-users.

However, the growing popularity of this domain has concurrently raised pertinent concerns encompassing security, fairness, regulatory compliance, intellectual property rights, and overall safety. The community faces pressing challenges, such as the inadvertent generation of unauthorized, prejudiced, and potentially hazardous content. This is not an unprecedented issue, as the academic community has previously endeavored to address similar concerns [5, 10].

The inherent risks associated with large-scale text-to-image models predominantly stem from the vast datasets employed during their training phase. These datasets, which include public repositories like LAION [43], COYO [3], CC12M [7], and proprietary data from renowned entities such as Google [41, 50] and OpenAI [36, 37], present unique challenges. Public datasets, often sourced from web scrapes, may lack rigorous quality control measures, especially concerning bias and safety. Conversely, proprietary datasets, while potentially more controlled, are constrained by scalability issues due to the inherent costs of annotation. Consequently, mitigating issues related to harmful content, privacy breaches, and copyright infringements through mere data filtration or source attribution becomes a daunting task. One potential recourse could be domain adaptation [18, 48, 51]. However, the intricacies of curating and refining such datasets remain formidable. Moreover, such domain adaptation can inadvertently diminish the model's versatility, rendering it inept at synthesizing out-of-domain images.

Therefore, efficient methods that enable large-scale text-to-image models to selectively forget specific concepts emerge as a promising direction. A line of concurrent works explore this direction by training models to redirect their predictions of a target concept towards some alternative concepts. [17, 26]. However, these methods essentially retrain the model to replace the target concept with user-defined concept, rather than allowing the model to revert to what naturally follows based on its inherent knowledge.

Moreover, recent controllable text-to-image synthesis works [8, 20] have highlighted that cross attention conditioning is a pivotal factor in determining the primary concepts featured in generated images. These insights lead us to explore solutions centered around the cross attention mechanism.

[*]Equal contribution

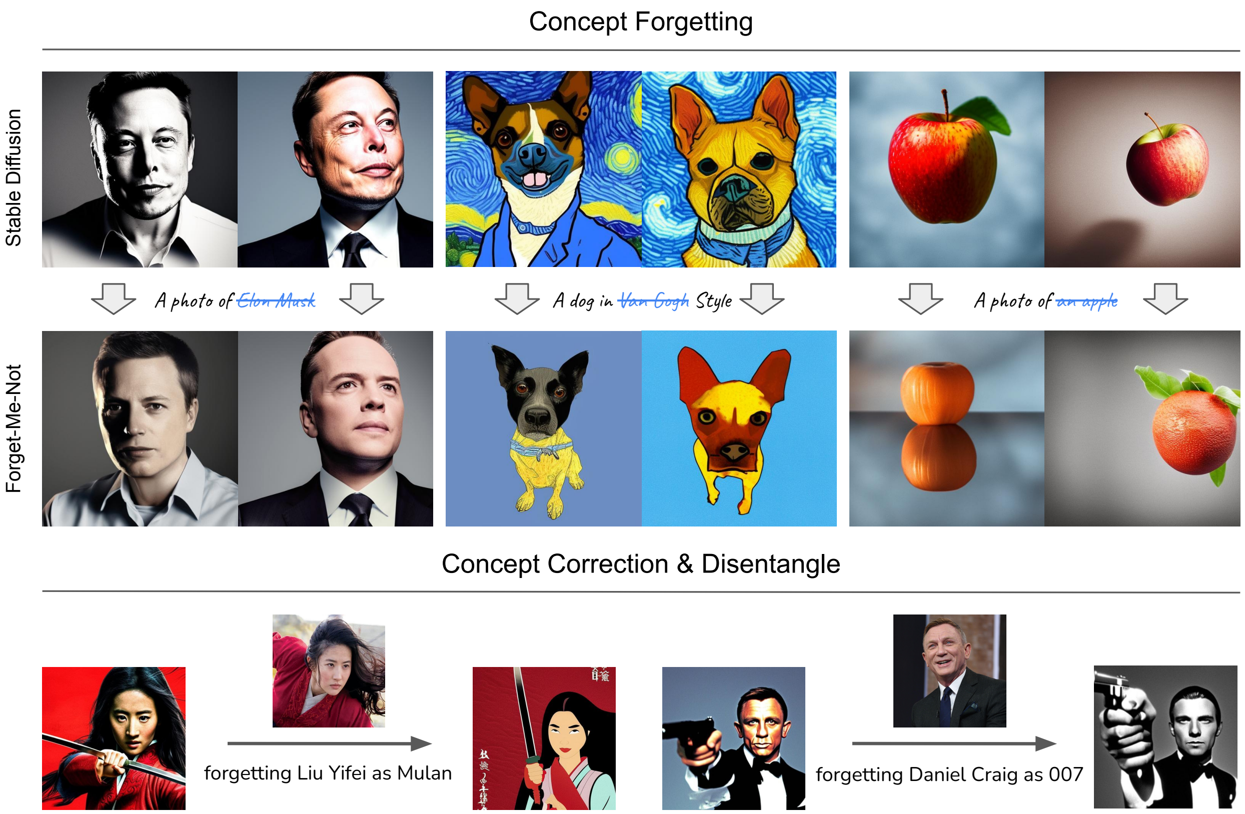

Concept Forgetting
Stable Diffusion
A photo of ~~Elon Mask~~
A dog in ~~Van Gogh~~ Style
A photo of ~~an apple~~
Forget-Me-Not
Concept Correction & Disentangle
forgetting Liu Yifei as Mulan
forgetting Daniel Craig as 007

Figure 1. **Concept Forgetting:** target concepts (denoted in blue text and crossed-out) are successfully removed without compromising the quality of the output. **Concept Correction & Disentangle:** our method can be used to correct a dominant or undesired concept of a prompt. Prior overshadowed concepts reveal in outputs after the dominant concepts are forgotten.

This paper demonstrates: 1) the effect of zeroing out cross attention scores of target concept at inference time, in which we observe target concept is diminishing in generated images; 2) cross attention scores can be directly used as objectives to optimize diffusion model for attenuating model's perception of target concepts. To the best of our knowledge, we are the first to show cross attention scores are viable objectives for fine-tuning text-to-image models. On top of aforementioned techniques, we introduce Forget-Me-Not, a cost-efficient methodology for concept forgetting where models determine what to generate based on its knowledge, a critical aspect overlooked by existing methods. In particular, we show Forget-Me-Not achieves comparable results to existing methods and outperforms them in some cases. Furthermore, it facilitates concept correction & disentanglement, enhancing model precision and diversity.

## 2. Related Work

### 2.1. Text-to-Image Synthesis

In the past decade, we have witnessed the rapid advance of it from unconditional generative models to conditional generative models with powerful architectures of auto-regressive model [36, 50], GAN [4, 24, 25, 44, 47] and diffusion process [2, 14, 21, 30, 34, 45]. Early works focus on unconditional, single-category data distribution modeling , such as hand-written digits, certain species of animals, and human faces [9, 12, 24, 29]. 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 [1, 4, 31]. To tackle the model collapsing problem, the conditional generative model has been introduced. Various types of data have been used as the conditioning for generative models, e.g. class labels, image instances, and even networks [4, 32] etc. At the same time, CLIP [23, 35], 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 [11, 28, 33]. Nowadays, DALL-E 2 [37] and Stable Diffusion [38] are capable of generating high quality images solely conditioning on free-form texts, 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. Dreambooth [40] adapts the model by finetuning all of its weights, while it requires enormous storage to save newly adapted weights. Textual Inversion [16] and LoRA [22] ameliorate the issue by adapting the model by adding a small set of extra weights.

### 2.2. Concept Replacing

Prior works have noticed the inadvertent biased and unsafe generations from large text-to-images models. They adopt the denoising loss [13] to steering predicted noise away from what is used to be for a given concept. Kumari et al. [26] and Heng and Soh [19] choose to replace target concept with a user-defined concept, usually a hypernym. For example, it finetunes to replace the generation of “a grumpy cat” to “a cat”. SLD [42] and ESD [17] utilize the classifier-free guidance of a pretrained model to steer prediction to the opposite guidance direction. However, this approach carries the risk of altering the semantics of the generated images significantly, potentially deviating from the original prompts.

## 3. Method

### 3.1. Preliminaries

**Diffusion models** [13, 21, 34] are denoising models that iteratively restore data $x_0$ from its Gaussian noise corruption $x_T$ with a total step number $T$. Such a restoration process is usually known as the reverse diffusion process $p_\theta(x_{t-1}|x_t)$ and the opposite is forward diffusion process that blends the signal with noise $q(x_t|x_{t-1})$:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}; \beta_t \mathbf{I})$$
$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t); \Sigma_\theta(x_t, t))$$

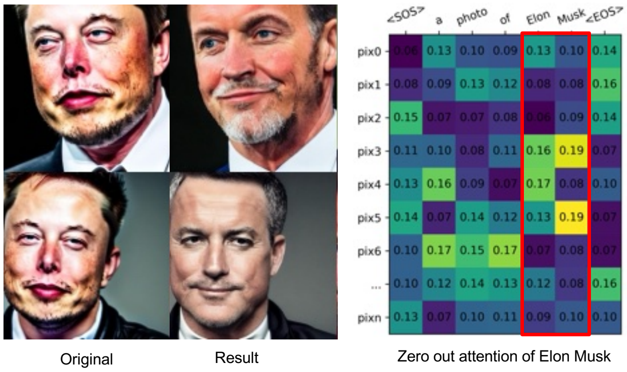
Both forward and reverse processes are presumably Markovian chains, so we can express the likelihood of both processes as:

$$q(x_{1:T}|x_0) = \prod_{t=1}^{T} q(x_t|x_{t-1})$$
$$p_\theta(x_{0:T}) = p(x_T) \prod_{t=1}^{T} p_\theta(x_{t-1}|x_t)$$

The loss function for the diffusion process is then to minimize the variational bound $\mathcal{L}_{vlb}$ of the negative log-likelihood $p_\theta(x_0)$ (maximize the likelihood of $x_0$ as the final denoised result from a model with parameters $\theta$):

$$\mathcal{L}_{\text{VLB}} = \mathbb{E}\left[-\log p_\theta(x_0)\right] \quad \leq \mathbb{E}_q\left[-\log \frac{p_\theta(\mathbf{x}_{0:T})}{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}\right]$$

**Cross Attention** [46] are widely adopted in generative models as a conditioning technique [37, 38, 41]. The purpose of conditioning is to transfer information of conditional signals to hidden features of generative models. Concretely, hidden features $h$ serve as the query $Q$ and conditional signals serve as key $K$ and value $V$. Assume $Q$ and $K$ has the same dimension $d$, the updated hidden feature $h'$ is then computed as the following:



|  | <SOS> | a | photo | of | Elon | Musk | <EOS> |
|---|---|---|---|---|---|---|---|
| pix0 | 0.06 | 0.13 | 0.10 | 0.09 | 0.13 | 0.10 | 0.14 |
| pix1 | 0.08 | 0.09 | 0.13 | 0.12 | 0.08 | 0.08 | 0.16 |
| pix2 | 0.15 | 0.07 | 0.07 | 0.08 | 0.06 | 0.09 | 0.14 |
| pix3 | 0.11 | 0.10 | 0.08 | 0.11 | 0.16 | 0.19 | 0.07 |
| pix4 | 0.13 | 0.16 | 0.09 | 0.07 | 0.17 | 0.08 | 0.10 |
| pix5 | 0.14 | 0.07 | 0.14 | 0.12 | 0.13 | 0.19 | 0.07 |
| pix6 | 0.10 | 0.17 | 0.15 | 0.17 | 0.07 | 0.08 | 0.07 |
| ... | 0.10 | 0.12 | 0.14 | 0.13 | 0.12 | 0.08 | 0.16 |
| pixn | 0.13 | 0.07 | 0.10 | 0.11 | 0.09 | 0.10 | 0.10 |

Figure 2. When the attention scores associated with ”Elon Musk”(red boxed) are set to zero in a text-to-image generative model, a prompt like ”a photo of Elon Musk” results in the generation of an image depicting a person other than Elon Musk.

$$h' = h + \text{softmax}(\frac{QK^T}{\sqrt{d}})V \tag{1}$$

In text-to-image generative models, $h$ represents visual features, while $V$ represents conditioning textual features. The equation involves visual features attend to concepts encapsulated within textual features $V$ and update themselves with softmax$(\cdot)V$ as a residual, in which softmax$(\cdot)$ calculates attention scores of all concepts. It is essentially a weighted sum of all encapsulated concepts, allowing the visual representation to be refined and aligned more closely with the textual input. In practice, manipulating attention scores of a concept will influence the presence of that concept in generated images. As shown in Figure 2, reducing attention scores of “Elon Musk” to zero leads to the generation of other person. We argue the generated anonymous person is what the model naturally falls back to when “Elon Musk” has been attenuated.

### 3.2. Forget-Me-Not

**Concept Forgetting** A concept is an abstract idea representing an object of thought, forming the basis of human perception and understanding. In the context of text-to-image models, these concepts are embedded within the words of a prompt. We define concept forgetting in such models as the process of attenuating the strong correlation between a target concept and its expected visual representation. This definition circumvents the need for finetuning the model to hard overwrite a target concept with a user-defined

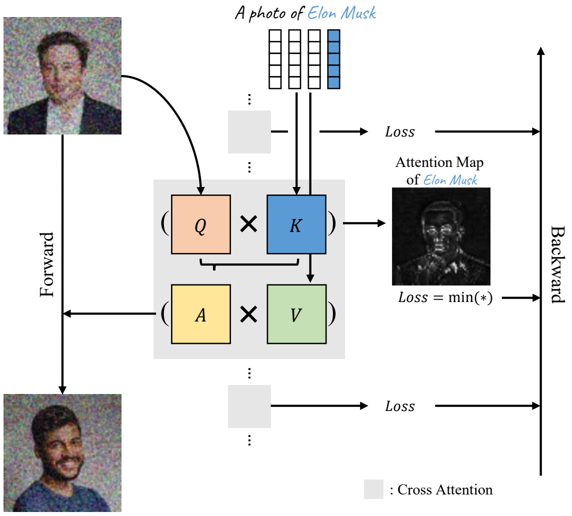


Figure 3. This figure shows the Attention Re-steering we proposed in our Forget-Me-Not method, in which we set the objective function to minimize the attention maps of target concepts (*i.e*. Elon Musk in this case) and correspondingly finetune the network.

alternative. Instead, it allows the model to determine what to generate once the correlation has been diminished. By doing so, the model is able to retain its original generative capabilities to the maximum extent.

To achieve these goals, we have developed two new loss functions: Attention Re-steering loss and Vsiual Denoising loss. Additionally, for scenarios where the prompt associated with a concept is not known, we employ Concept Inversion, a technique designed to extract the textual embeddings of a concept directly from images.

**Attention Re-steering Loss** Following the discussion of cross attention, we are interested in using attention scores of target concept as a loss to optimize trainable parameters $\theta$ of UNet [39]. UNet is the backbone of Stable Diffusion [38]. It consists a series of interleaving convolution and cross attention layers, performing denoising task and conditioning task respectively. Layers are grouped as down, mid, and up blocks, in which visual features have gone through dimension changes. There are 16 cross attention layers in total. We obtain attention scores from each layer. Let $A_{n\times l}$ be the attention map of a cross attention layer. It is computed as the softmax$(\cdot)$ term in Eq. 1. $n$ is the number of visual features from $Q$ and $l$ is the length of textual tokens from $K$. The start and end indices of textual tokens associated with target concept is denoted as $[i, j]$. We calculate the loss for a specific cross attention layer as:

$$\mathcal{L}_{\text{Attn}} = \sum_{a\in A[:,i:j]} \|a\|^2 \tag{2}$$

where $A[:, i : j]$ indexes all scores from column $i$ to $j$. This corresponds to all the attentions paid to target concept by visual features. Figure 3 illustrates the idea of Attention Re-steering loss, minimizing attention scores of target concept. This redirects the generation process from target concept to its less prominent alternatives. Algorithm 1 details how the loss is used in a training loop.

**Algorithm 1** Attention Re-steering loss in training

```
Require: Textual embeddings $\mathcal{C}$ containing the target concept, indices $\mathcal{I}$ of the target concept, reference images $\mathcal{R}$ of the target concept, UNet $U_\theta$, denoising timestep $T$, total training step $S$.
repeat
    $t \sim$ Uniform$([1 \dots T]); \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
    $r_i \sim \mathcal{R}; c_i, idx_i \sim \mathcal{C}, \mathcal{I}$
    $x_0 \leftarrow r_i$
    $x_t \leftarrow \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon$
                                        ▷ $\bar{\alpha}_t$: noise variance schedule
    $x_{t-1}, \mathcal{A}_t \leftarrow U_\theta(x_t, c_j, t)$
                                        ▷ $\mathcal{A}_t$: all attention maps
    $\mathcal{L} \leftarrow \sum_{A\in\mathcal{A}_t} \sum_{a\in A[idx_i]} \|a\|^2$
    $\theta \leftarrow \theta - \nabla_\theta \mathcal{L}$
until S steps
```

**Visual Denoising Loss** As shown in Figure 2, manipulating attention scores at inference time naturally becomes an efficient and low-cost technique for concept forgetting data generation. Due to the self-attention in text encoder, each token shares information with every other token. Therefore, zero attention of target concept won't lead to completely loss of its semantics. For example, zeroing out attention paid to "Elon Musk" still get us an image of mid-aged man, instead of a woman. This phenomenon represents what we term the model's 'natural fallback'. This inspires us to create synthetic data as the "ground truth" for concept forgetting.

For a training step $t$, in addition to sampled noise $\epsilon$, the model takes as inputs both a generated image $x_0$ representing a concept and its counterpart $\tilde{x}_0$, where the attention to the concept is set to zero. It predicts the Gaussian noise $\epsilon_\theta(x_t)$ that has been added to $x_0$. Subsequently, $x_0$ can be approximated, as described in [21]:

$$\begin{aligned} x_0 \approx \hat{x}_0 &= (x_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(x_t))/\sqrt{\bar{\alpha}_t} \\ \mathcal{L}_{\text{vis}} &= \text{MSE}(\hat{x}_0, \tilde{x}_0) \end{aligned} \tag{3}$$

where $\bar{\alpha}_t$ is predefined noise schedule coefficient. Instead of computing denoising loss between predicted $\epsilon_\theta(x_t)$ and ground truth $\epsilon$, we first approximate $\hat{x}_0$ and compute mean squared error between $\hat{x}_0$ and $\tilde{x}_0$. We call it visual denoising loss.

In practice, we find the two loss can be used either independently or together. Attention Re-steering loss offers convenience without the need of generating paired training data. On the other hand, Visual Denoising loss tends to focus the training on the primary concept that needs to be forgotten. This is due to the similarity in minor image structures found in the paired synthetic data, as depicted in Figure 2.

**Concept Inversion:** Although we usually know the prompts for a target concept, there are exceptions when it's hard or impossible to describe a concept using textual prompts. For instance, forget a style represented by an image. To overcome this challenge, we include the textual inversion [16] as an optional component. In practice, we also notice that such inversion helps text-to-image models more precisely identify the forgetting concept and thus improves their performance. Results can be found in Experiments Section.

# 4. Experiments

**Dataset** Our evaluation encompasses a range of concepts such as identity, object, animal, style. Each concept comprises pairs of images and their corresponding textual prompts, with either real images or synthetic images generated via the Stable Diffusion 2.1 base.

**Baselines** Our approach is compared against other concurrent works, ESD [17] and ACTD [26]. The comparison focuses on forgetting performance and the influence on related concepts, as detailed in section 4.1. Our unique capability for concept correction is highlighted in section 4.3.

**Evaluation Metrics** Forgetting efficacy is assessed using both CLIP score [35] and Memorization Score, described in Sec. 4.2. The CLIP score measures the congruence between generated images and their textual prompts. A decrease in the CLIP score signifies effective forgetting, but an excessively low score indicates a loss of overall image semantics, see Table 1. For instance, after forgetting the "corgi" concept, we anticipate the generated image to still represent a dog, not an completely different object.

**Ablation Studies** We delve into the impact of various configurations, including different sets of trainable weights and the utilization of inverted concepts from images.

## 4.1. Qualitative Comparison

Figure 4 showcases the experiment results. The multi-concept model, targeting Elon Musk and Taylor Swift, exemplifies our method's proficiency in multi-concept forgetting. Notably, the first row reveals the successful forgetting of both target concepts. Our evaluation also considered the repercussions of forgetting on related concepts, specifically man, woman, Bill Gates, and Emma Watson. The results indicate that our Forget-Me-Not method preserves content and maintains visual quality effectively. However, subtle changes in pose and style were observed for the man and Bill Gates concepts. These observations suggest that our method might influence closely associated concepts more than distant ones. The final row further highlights the emergence of a new painting style post-forgetting the styles of Picasso and Van Gogh.

In comparison to [17, 26], our approach exhibits superior generative capabilities for related concepts, illustrated in Figure 6. ESD [17], which employs inverse class-free guidance, adversely affects related concepts. Conversely, ACTD [26] adjusts the forgetting concept relative to an user-defined concept, making its performance contingent on human intervention. This becomes challenging for abstract concepts where an apt anchor is hard to define.

## 4.2. Quantitative Analysis

**Memorization Score** Unlike the CLIP score, which assesses the semantic content of generated images, the Memorization Score offers a model-centric perspective. It evaluates the changes in model's knowledge about a probing dataset, both before and after concept forgetting.

It start with a small probing set of images representing a concept. We then employ textual inversion [16] to invert the probing dataset into embeddings in CLIP space. This inversion is performed twice: once using the original model and once using the model that has undergone concept forgetting. Subsequently, we calculate the CLIP similarity between an anchor prompt and each of the two sets of inverted embeddings. The anchor prompt is carefully selected to effectively capture the essence of the concept in question. After forgetting, a decrease of CLIP similarity between probing set and anchor prompt signifies the effective forgetting.

For instance, with anchor prompt as "Elon Musk", its embedding ($\mathbf{emb}_r$) is derived from text encoder. Probing images of Elon Musk are then inverted using both the *original model* and the *forgetting model*. The resulting embeddings, original textual inversion ($\mathbf{emb}_o$) and forgetting textual inversion ($\mathbf{emb}_f$), are compared with embeddings of the anchor prompt to calculate similarity. The similarity change is quantified as the difference between $\cos(\mathbf{emb}_r, \mathbf{emb}_o)$ and $\cos(\mathbf{emb}_r, \mathbf{emb}_f)$. Table 1 shows that Memorization Score declines after concept forgetting.

| Concept | Init. MScore | Frgt. MScore ↓ | Init. CLIP | Frgt. CLIP ↓ |
|---|---|---|---|---|
| Elon Musk | 0.943 | 0.848 | 0.308 | 0.285 |
| Mickey Mouse | 0.948 | 0.836 | 0.304 | 0.269 |
| Zebra | 0.972 | 0.899 | 0.312 | 0.310 |
| Google | 0.940 | 0.811 | 0.216 | 0.209 |
| Apple | 0.696 | 0.493 | 0.267 | 0.258 |
| Horse | 0.877 | 0.808 | 0.275 | 0.266 |
| Van Gogh | 0.916 | 0.684 | 0.274 | 0.233 |

Table 1. Changes of Memorization Scores and CLIP Scores after concept forgetting

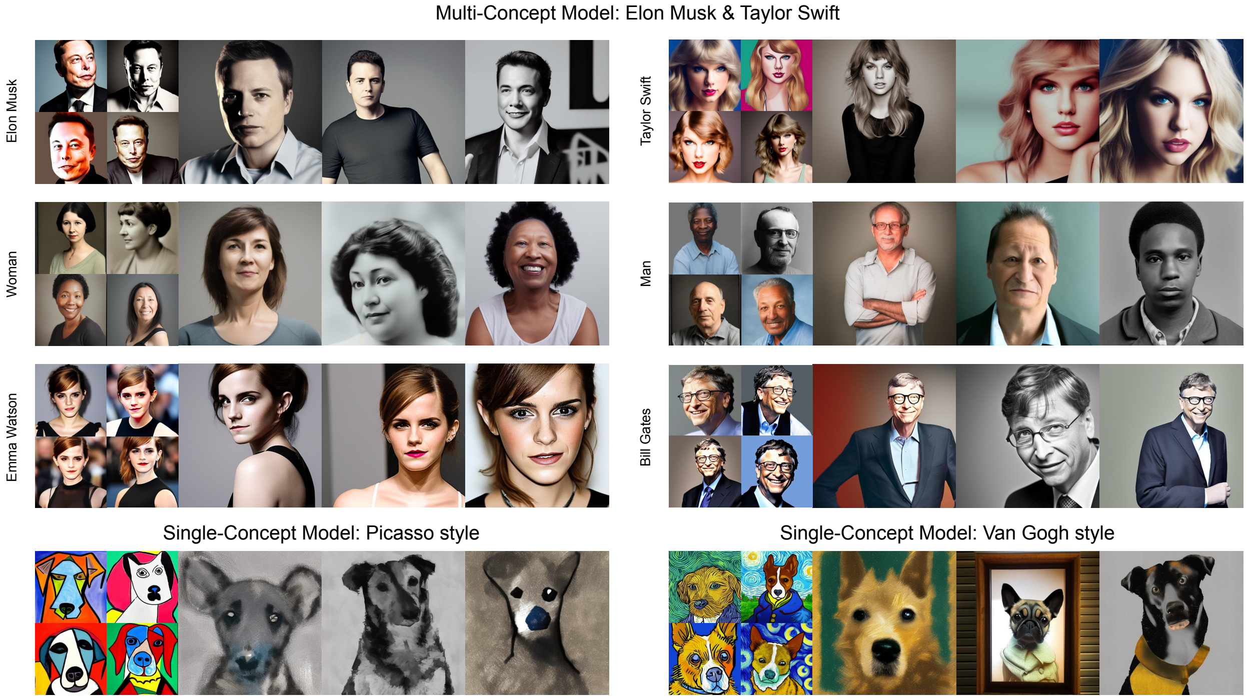

Multi-Concept Model: Elon Musk & Taylor Swift
Elon Musk
Taylor Swift
Woman
Man
Emma Watson
Bill Gates
Single-Concept Model: Picasso style
Single-Concept Model: Van Gogh style

Figure 4. Concept forgetting results using our method. The initial 2x2 grid displays original samples from Stable Diffusion. Following this, three images depict post-forgetting samples generated from the same prompt. The top three rows, targeting Elon Musk and Taylor Swift, highlight our multi-concept forgetting capability. Control concepts, including Bill Gates and Emma Watson, illustrate the limited influence our method has on non-targeted concepts. The final row presents single-concept style models. Image prompts used were: "a photo of X" (for the first three rows) and "a dog in X style" (for the last row).

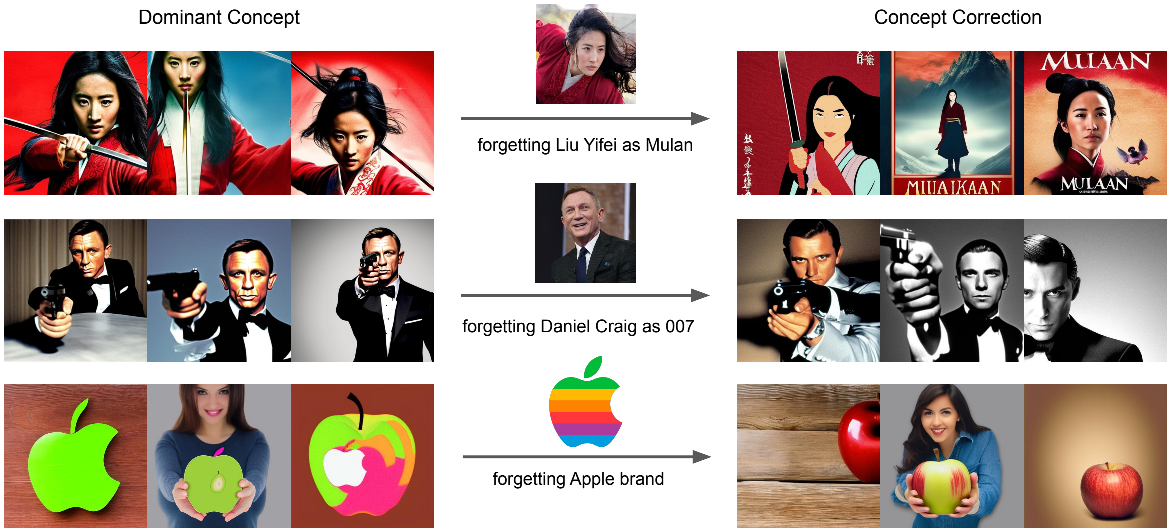


Figure 5. Concept Correction: Our method's effectiveness in diminishing dominant concepts allows for the emergence of secondary concepts within semantically-rich prompts. The displayed images, generated from top to bottom, correspond to the prompts: "a movie poster of Mulan", "James Bond", and "apple shape".

### 4.3. Concept Correction

Text-to-image models often prioritize the semantics of a prompt based on the abundance of image-text examples dur-

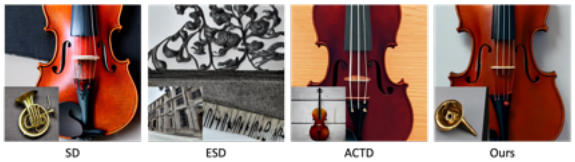


Figure 6. Comparison of methods for forgetting the french horn concept. Each primary figure is prompted with “a photo of a violin”, accompanied by a smaller “a photo of a french horn” image. ESD not only disrupts the french horn concept but also adversely affects the related violin concept. ACTD, using “instrument” as the anchor concept, which is dominated by the violin in the SD model, results in the french horn concept merging with the violin concept. In contrast, our method selectively removes the french horn features while preserving the salient features of other related concepts.

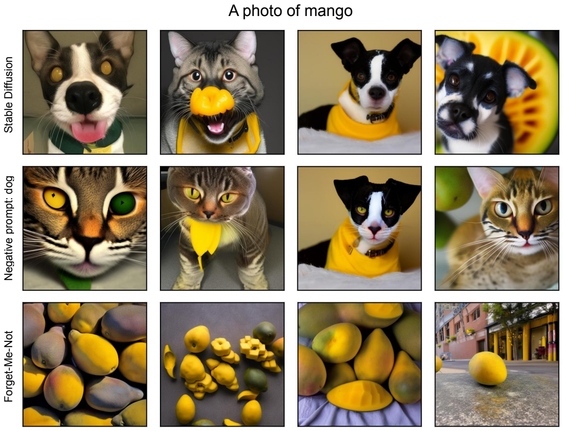


Figure 7. In concept correction, our method has the advantage of comprehensive forgetting over negative prompt. In this example, we also tests “animal” as negative prompts, yet it still generates dogs/cats.

ing training. This can overshadow less prevalent semantics during inference, as illustrated in Figure 5. For instance, the James Bond series predominantly showcases Daniel Craig. Our method, however, can reduce the dominance of such a semantic, enabling visibility of other James Bond actors. Similarly, our approach effectively rectifies target concepts in scenarios where semantics compete, as seen with the Mulan series and the term “apple”.

Negative prompts, used in text-to-image synthesis to exclude unwanted concepts, can inadvertently alter other image attributes. Moreover, they may not always rectify undesired concepts. As depicted in Figure 7, the prompt “a photo of a mango” often yields dog images due to the popularity of “mango” as a pet name. Our method adeptly retrieves the mango fruit by dissociating the prompt from dog/cat images.

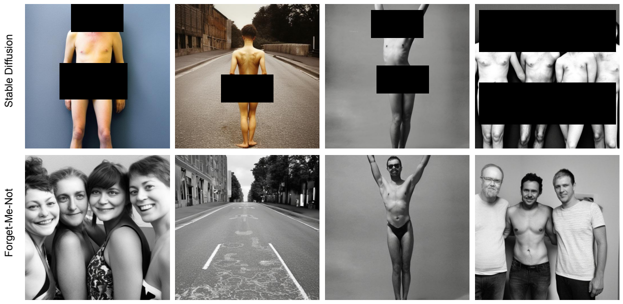


Figure 8. Results of removing NSFW contents triggered by “naked”. Faces and sensitive parts are blacked out.

### 4.4. NSFW Removal

We evaluate our method’s capability to eliminate inappropriate content, denoted as NSFW (“not safe for work”). Such content, potentially offensive even to adults, can inadvertently be part of large datasets like LAION [43], despite the use of NSFW detectors [27]. Stable Diffusion, trained on LAION, has been known to produce NSFW content with specific prompts.

To assess our approach, we utilize a known NSFW-triggering prompt, “a photo of naked”, in the Stable Diffusion v2.1 base model. Using this setup, the model consistently produces inappropriate images. We then train Forget-Me-Not using eight of these NSFW images.

The results, presented in Figure 8, demonstrate the forgetting of the “naked” concept. The original images have undergone black modifications to ensure appropriateness.

### 4.5. Ablation Studies

**Concept Inversion Ablation** We conducted experiments with and without Concept Inversion (CI). Concept inversion is used to handle concepts that are difficult to describe using prompts. Generally, it can help extract the target concept from the prompt, resulting in more precise embeddings. Our results show that CI can achieve higher fidelity for concepts that can be well-described in a prompt, as illustrated in Figure 9, where the model trained with CI preserved more of the original poses and details.

However, CI is not always ideal. Its performance may vary concept by concept. In Figure 10, we demonstrate a situation where textual prompt prevails CI. By using the same settings, textual prompt “airplane” succeeds while inverted concept fails.

**Trainable Weights Ablation** We conducted experiments to compare finetuning the entire UNet model versus only finetuning the cross-attention (CA) layers. Cross-attention is a critical component in text-to-image generation, as it injects textual information into the image formation process. Given the same hyper-parameter settings except for steps, our results show that both methods can successfully achieve con-

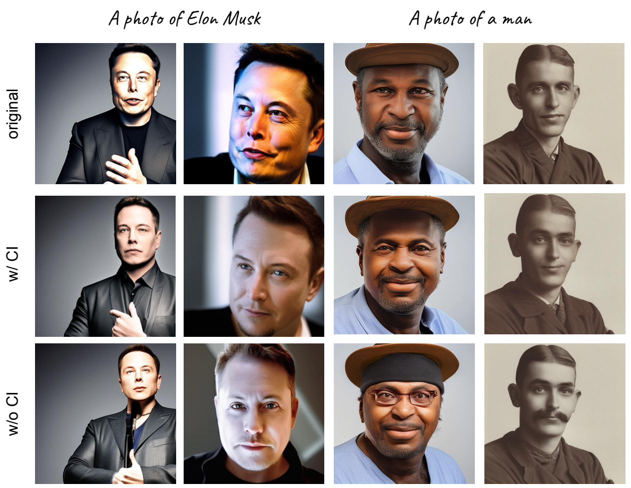


Figure 9. Improving fidelity to original model with concept inversion. Concept prompt tend to have diverse semantics, resulting in distortion in concept forgetting. CI extracts precise semantics into dedicated embeddings, allowing for more pose and feature consistency.

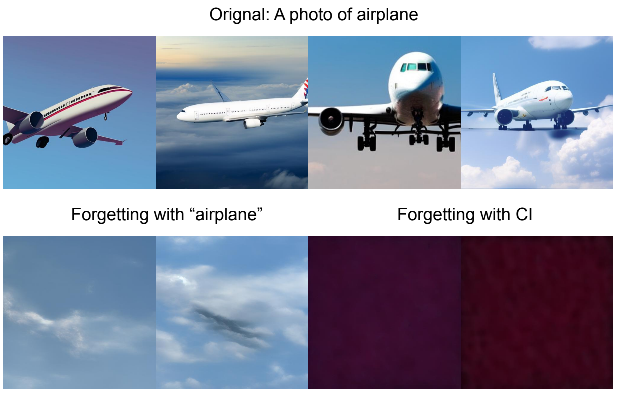


Figure 10. Another example of “airplane”, where forgetting with CI fails.

cept forgetting. However, finetuning the entire UNet model tended to break the model’s generation capability in fewer steps. In some cases, the model collapsed before the forgetting process was complete, as show in the “Broccoli” case of Figure 11.

## 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; it can also be easily distributed using model patches. Forget-Me-Not is naturally extended to enable concept correction and disentanglement. Our experiments demonstrate that it is successful in diminishing and correcting target concepts in Stable Diffusion. Additionally, we introduce Memorization Score as evaluation metric. Overall, our work provides a foundation for further research on concept forgetting and manipulation in text-to-image generation, and can be further extended to other conditional multi-modal generative models to improve the accuracy, inclusion and diversity of such models.

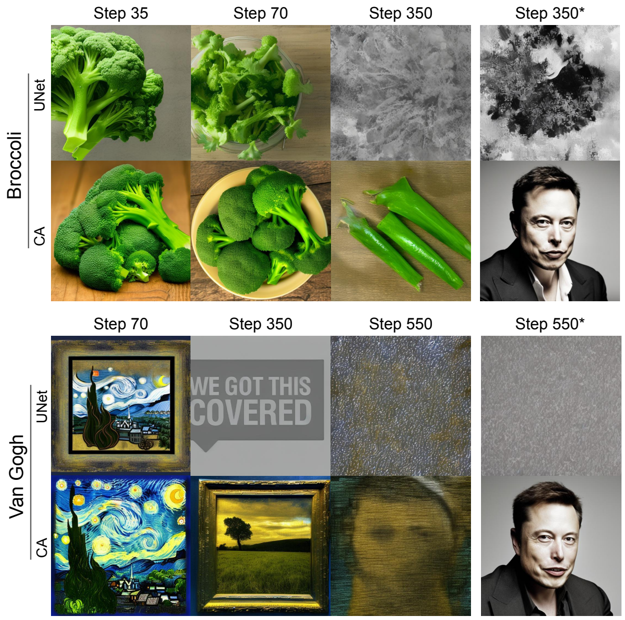


Figure 11. Trainable weights ablation. Compared to tuning only CA, full UNet is more sensitive to optimization steps. The last column with Step X$^*$ shows the control concept Elon Musk at Step X.

## 6. Social Impact & Limitations

**Social Impact** Our research has a positive social impact by offering an effective and cost-efficient method to remove and correct 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. Our method can generate lightweight model patches that can be conveniently distributed to text-to-image model users like how conventional software patch works. 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, 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 generative adversarial networks. In *International Conference on Machine Learning (ICML)*, pages 214–223. PMLR, 2017. 2

[2] 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

[3] Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Saehoon Kim. Coyo-700m: Image-text pair dataset. https://github.com/kakaobrain/coyo-dataset, 2022. 1

[4] Arantxa Casanova, Marlene Careil, Jakob Verbeek, Michal Drozdzal, and Adriana Romero Soriano. Instance-conditioned gan. *Advances in Neural Information Processing Systems (NeurIPS)*, 34:27517–27529, 2021. 2

[5] Sungmin Cha, Sungjun Cho, Dasol Hwang, Honglak Lee, Taesup Moon, and Moontae Lee. Learning to unlearn: Instance-wise unlearning for pre-trained classifiers. *arXiv preprint arXiv:2301.11578*, 2023. 1

[6] 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

[7] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3558–3568, 2021. 1

[8] 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. *arXiv preprint arXiv:2301.13826*, 2023. 1

[9] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8188–8197, 2020. 2

[10] Vikram S Chundawat, Ayush K Tarun, Murari Mandal, and Mohan Kankanhalli. Zero-shot machine unlearning. *arXiv preprint arXiv:2201.05629*, 2022. 1

[11] 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. In *European Conference on Computer Vision (ECCV)*, pages 88–105. Springer, 2022. 2

[12] 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. 2

[13] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems (NeurIPS)*, 34:8780–8794, 2021. 3

[14] Tim Dockhorn, Arash Vahdat, and Karsten Kreis. Genie: Higher-order denoising diffusion solvers. *arXiv preprint arXiv:2210.05475*, 2022. 2

[15] Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. Make-a-scene: Scene-based text-to-image generation with human priors. In *European Conference on Computer Vision (ECCV)*, pages 89–106. Springer, 2022. 1

[16] 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. 3, 5

[17] Rohit Gandikota, Joanna Materzyńska, Jaden Fiotto-Kaufman, and David Bau. Erasing concepts from diffusion models. In *Proceedings of the 2023 IEEE International Conference on Computer Vision*, 2023. 1, 3, 5

[18] Giorgio Giannone, Didrik Nielsen, and Ole Winther. Few-shot diffusion models. *arXiv preprint arXiv:2205.15463*, 2022. 1

[19] Alvin Heng and Harold Soh. Selective amnesia: A continual learning approach to forgetting in deep generative models. *arXiv preprint arXiv:2305.10120*, 2023. 3

[20] 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. 1

[21] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:6840–6851, 2020. 2, 3, 4

[22] 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. *arXiv preprint arXiv:2106.09685*, 2021. 3

[23] 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. Open-CLIP, 2021. 2

[24] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 2

[25] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. *Advances in Neural Information Processing Systems (NeurIPS)*, 34:852–863, 2021. 2

[26] Nupur Kumari, Bingliang Zhang, Sheng-Yu Wang, Eli Shechtman, Richard Zhang, and Jun-Yan Zhu. Ablating concepts in text-to-image diffusion models. *arXiv preprint arXiv:2303.13516*, 2023. 1, 3, 5

[27] LAION-AI. Clip-based-nsfw-detector. https://github.com/-AI/CLIP-based-NSFW-Detector. 7

[28] 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. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 289–299, 2023. 2

[29] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Large-scale celebfaces attributes (celeba) dataset. *Retrieved August*, 15(2018):11, 2018. 2

[30] 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. *arXiv preprint arXiv:2206.00927*, 2022. 2

[31] Luke Metz, Ben Poole, David Pfau, and Jascha Sohl-Dickstein. Unrolled generative adversarial networks. *arXiv preprint arXiv:1611.02163*, 2016. 2

[32] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014. 2

[33] 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. *arXiv preprint arXiv:2112.10741*, 2021. 2

[34] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning (ICML)*, pages 8162–8171. PMLR, 2021. 2, 3

[35] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, pages 8748–8763. PMLR, 2021. 2, 5

[36] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning (ICML)*, pages 8821–8831. PMLR, 2021. 1, 2

[37] 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, 2, 3

[38] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022. 1, 2, 3, 4

[39] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 4

[40] 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. 3

[41] 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. 1, 3

[42] Patrick Schramowski, Manuel Brack, Björn Deiseroth, and Kristian Kersting. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[43] 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. 1, 7

[44] Tamar Rott Shaham, Tali Dekel, and Tomer Michaeli. Singan: Learning a generative model from a single natural image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4570–4580, 2019. 2

[45] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 2

[46] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems (NeurIPS)*, 30, 2017. 3

[47] Steven Walton, Ali Hassani, Xingqian Xu, Zhangyang Wang, and Humphrey Shi. Stylenat: Giving each head a new perspective. *arXiv preprint arXiv:2211.05770*, 2022. 2

[48] Jiayu Xiao, Liang Li, Chaofei Wang, Zheng-Jun Zha, and Qingming Huang. Few shot generative model adaption via relaxed spatial structural alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11204–11213, 2022. 1

[49] Xingqian Xu, Zhangyang Wang, Eric Zhang, Kai Wang, and Humphrey Shi. Versatile diffusion: Text, images and variations all in one diffusion model. *arXiv preprint arXiv:2211.08332*, 2022. 1

[50] 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. 1, 2

[51] Jingyuan Zhu, Huimin Ma, Jiansheng Chen, and Jian Yuan. Few-shot image generation with diffusion models. *arXiv preprint arXiv:2211.03264*, 2022. 1