AIZHEIMER: How was I Drawing?

Muhammed Ruşen Birben
Department of AI and Data Engineering
Istanbul Technical University
Istanbul, Turkey
150220755
birben20@itu.edu.tr

Göktürk Batın Dervişoğlu
Department of AI and Data Engineering
Istanbul Technical University
Istanbul, Turkey
150210307
dervisoglu21@itu.edu.tr

Abdulkadir Külçe
Department of AI and Data Engineering
Istanbul Technical University
Istanbul, Turkey
150210322
kulce21@itu.edu.tr

https://github.com/rusenbb/AIzheimer

***Abstract*—The large AI models need to be trained on large datasets to achieve high performance. However, the data used in training these models may contain sensitive information. In this study, we propose machine unlearning on diffusion model to forgetting the target concepts. We show that the proposed method can be used to forget the target concepts without retraining the whole model from scratch.**

***Index Terms*—Machine unlearning, diffusion model, forgetting, privacy, sensitive information, attention re-steering, generative models**

I. Introduction

Diffusion models are state-of-the-art technologies when it comes to image generation. The idea behind diffusion models is to model the data distribution by applying a series of transformations to the input data. These transformations are designed to gradually increase the entropy of the data, making it easier to model the data distribution. Diffusion models have been shown to achieve state-of-the-art performance in image generation tasks, outperforming other generative models such as GANs and VAEs.

However, the data used in training these models may contain sensitive information. For example, the images used in training the diffusion model may contain personal information such as faces, license plates, or any other identifiable information. Additionally, the data can include NSFW content, which can be harmful to some users. In such cases, it is essential to remove the sensitive information from the model to protect the privacy of the users. Retraining the whole model from scratch to remove this information can be time-consuming and computationally expensive.

In this study, we propose a method for machine unlearning in diffusion models to forget target concepts. Our approach allows the model to forget specific concepts without the need to retrain the entire model from scratch. We achieve this by manipulating the attention mechanisms within the model, selectively training only some of the cross-attention modules. We also compare use of different custom loss functions to steer the model's attention away from the target concepts. We demonstrate the effectiveness of our method through various experiments and show that it can successfully remove the target concepts while preserving the overall functionality of the model.

II. Literature Review

The first problem (and an opportunity) for researchers is that Machine Unlearning is a relatively new topic. It is an opportunity because it presents an unexplored field waiting to be discovered by researchers. It is problematic because we do not yet know which methodologies work well. For these reasons, all of the papers below are recent academic papers that attempt to achieve machine unlearning on diffusion models.

A recent work has explored machine unlearning techniques to remove specific concepts from text-to-image diffusion models while preserving model utility. Wu et al. [3] propose a concept domain correction framework using adversarial training to align the output distributions of target and anchor concepts, enhancing the generalization of concept unlearning. They also introduce a concept-preserving gradient surgery method to mitigate conflicts between unlearning and retraining objectives. Extensive experiments demonstrate the effectiveness of their approach in selectively unlearning problematic concepts like copyrighted characters and inappropriate content with minimal impact on the model's ability to generate other concepts. This builds on prior methods like ESD by Gandikota et al. [4], which aligns target concepts with empty strings, ConAbl by Kumari et al. [5], which minimizes the distance between target and anchor concept noise, and SPM by Lyu et al. [7], which employs latent anchoring and similarity-based retention loss. However, these earlier techniques face limitations in generalization beyond training prompts and preservation of model utility. Wu et al.'s novel domain correction and gradient surgery techniques represent an important advancement in enabling safer, more controllable generative models through targeted concept unlearning.

In another paper, Zhang et al. [1] compare different methodologies for blocking undesired outputs and examine two main approaches. The first approach involves creating a censored dataset that includes forbidden vocabularies. However, the problem is that the target word could be described indirectly, leading to undesired outputs because the diffusion model does not forget anything in this approach. On the other hand, training the whole UNet actually achieves the forgetting of the target concept but results in a loss of drawing ability. To solve this issue, they suggest training only the CrossAttention Layers of the UNet model so that the model does not lose its ability.

Another recent research by Zhao et al. [2] has explored machine unlearning techniques to remove sensitive or copyrighted content from text-to-image diffusion models. Existing approaches include dataset filtering, model fine-tuning, and post-generation classification. However, these methods face challenges with generalization, where unlearning is limited to the training data, and utility drop, where model performance significantly degrades. To address these limitations, Zhao et al. propose a novel concept domain correction framework using adversarial training to align target and anchor concept distributions, enhancing unlearning generalization. Additionally, a concept-preserving gradient approach based on gradient surgery is introduced to mitigate conflicts between unlearning and relearning objectives, minimizing the impact on model utility. Extensive experiments demonstrate the effectiveness of the proposed techniques in selectively unlearning targeted concepts while preserving model performance on related content.

In this paper, we aim to build on these prior works by exploring machine unlearning techniques to remove specific concepts from text-to-image diffusion models while preserving model utility. We propose different ways of unlearning by trying different loss functions and compare different methods. We show that the proposed method can be used to forget the target concepts without retraining the whole model from scratch.

III. Methodology

In this section, we present the theoretical framework for achieving machine unlearning in diffusion models. Our approach involves manipulating the attention mechanisms within the model, selectively training only specific modules, and using custom loss functions to steer the model's attention away from the target concepts.

A. Diffusion Model Formulation

Let the initial data distribution be denoted by $\mathbf{x}_0 \sim q(\mathbf{x}_0)$, and let $\mathbf{x}_T$ denote the latent representation at time step T. The diffusion process gradually adds Gaussian noise to the data according to the following stochastic differential equation:

$$d\mathbf{x}_t = -\frac{1}{2}\beta_t \mathbf{x}_t dt + \sqrt{\beta_t} d\mathbf{w}_t, \tag{1}$$

where β_t is a time-dependent diffusion coefficient and $\mathbf{w}_t$ is a standard Wiener process. The reverse process, used for generating samples, is defined as:

$$d\mathbf{x}_t = \left(\frac{1}{2}\beta_t \mathbf{x}_t - \nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t)\right) dt + \sqrt{\beta_t} d\bar{\mathbf{w}}_t, \tag{2}$$

where $p_t(\mathbf{x}_t)$ is the model's estimate of the data distribution at time step t, and $\bar{\mathbf{w}}_t$ is a reverse-time standard Wiener process.

B. Attention Mechanism

The attention mechanism in the diffusion model plays a crucial role in capturing the dependencies between different parts of the input. Given a query, key, and value, the attention score is computed as:

$$\text{Attention}(\mathbf{q}, \mathbf{k}, \mathbf{v}) = \text{softmax}\left(\frac{\mathbf{q}\mathbf{k}^T}{\sqrt{d_k}}\right) \mathbf{v}, \tag{3}$$

where d_k is the dimension of the key vectors.

C. Custom Attention Processes

To achieve machine unlearning, we introduce two custom attention processes: a controller for tracking and manipulating attention probabilities, and a processor for overriding the default cross-attention mechanism.

1) Attention Controller: The attention controller tracks attention probabilities during the forward pass of the model. It collects attention scores related to the target concept and provides different loss functions to steer the attention away from it:

- **Attention Resteering**: This method normalizes the collected attention probabilities to reduce the model's focus on the target concept. The loss function for attention resteering $\mathcal{L}_{\text{resteering}}$ is defined as:

$$\mathcal{L}_{\text{resteering}} = \left\| \mathbf{concat}(\mathbf{A_0} ... \mathbf{A_n}) \right\|_2, \tag{4}$$

  where $\mathbf{A}$ represents the attention probabilities.
- **Distraction**: This method encourages the attention to be uniformly distributed, reducing focus on the target concept. The loss function for distraction $\mathcal{L}_{\text{distraction}}$ is defined as:

$$\mathcal{L}_{\text{distraction}} = \sum_{i=1}^{N} \left\| \mathbf{A}_i - \frac{1}{dim(A_i)} \mathbf{1} \right\|_2, \tag{5}$$

  where N is the number of attention heads and $\mathbf{1}$ is a vector of ones.
- **Random Stimulus**: This method introduces randomness to the attention, further reducing the model's ability to focus on the target concept. The loss function for random stimulus $\mathcal{L}_{\text{random}}$ is defined as:

$$\mathcal{L}_{\text{random}} = \sum_{i=1}^{N} \left\| \mathbf{A}_i - \mathbf{R}_i \right\|_2, \tag{6}$$

  where $\mathbf{R}_i$ is a random vector with the same dimensions as $\mathbf{A}_i$.

One thing to note here is that all of the As are flattened attentions.

2) Cross-Attention Processor: The cross-attention processor overrides the default cross-attention mechanism to incorporate the custom attention controller. It is the standart cross-attention procedure but because of overriding, we could get the attention probabilities from the attention processor to the attention controller.

D. Unlearning Objective

The objective of the unlearning process is to minimize the following loss function:

$$\mathcal{L} = \mathcal{L}_{\text{recon}} + \lambda \mathcal{L}_{\text{attn}}, \quad (7)$$

where $\mathcal{L}_{\text{recon}}$ is the reconstruction loss, $\mathcal{L}_{\text{attn}}$ is the attention loss provided by the attention controller, and λ is a hyper-parameter controlling the trade-off between the two losses. It is important to note that there is no single hyper-parameter as λ in any pipeline, rather it stands as a theoretical trade-off between retention and forgetting.

By minimizing this objective, the model learns to generate samples that are similar to the original data distribution while simultaneously forgetting the target concept.

E. Unlearning Procedure

The training procedure involves the following steps:

1) Encode the input data into latent representations using a variational autoencoder.
2) Add noise to the latent representations according to the diffusion process.
3) Compute the reconstruction loss and the attention loss using the custom attention controller.
4) Update the model parameters to minimize the total loss.
5) Reset attention probabilities for the next iteration.

By iteratively applying this training procedure, the model gradually learns to forget the target concept while preserving its overall functionality.

IV. Experiments and Results

A. Observations

Sex is an important feature: We've observed that if we over-fit our models on unlearning objective and if we manage to keep the learning rate on track so that the model is not catastrophically forgetting, the model tends to attend the sex of the person we are trying to forget in an inverse way (consistently!)

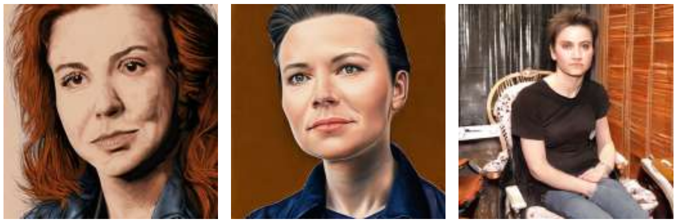


TABLE I
Forgetting Elon Musk - Non-catastrophic overfit

In table I we've forgotten Elon Musk and got the results for Robert Downey Jr. and Elon Musk. Using the large initial learning rate and long epochs. The left-most photo is Robert Downey Jr. and the other two are Elon Musk. Of course this is not the optimal case as the model also forgot to draw Robert Downey Jr. and potentially many other males. But we still see this as a useful insight.

Information exists elsewhere: Even though to some extent machine unlearning is achieved in both our work and similar works [1] some of the information seems to be impossible to get rid of without compromising the model's performance (colors, trivial shapes, overall structures of objects). We believe this is due to the fact that our work only focuses on the attention blocks. Some of the information might be inherent to the textual embedding of the concept. And not attending to those features inevitably reduces performance of the model.

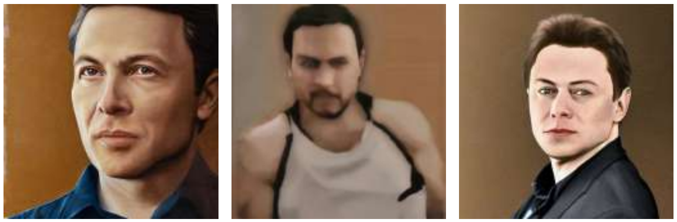


TABLE II
Forgetting Elon Musk - information stucks

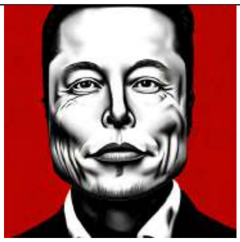
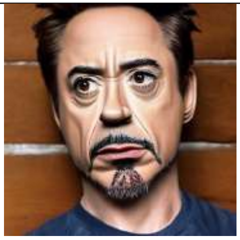
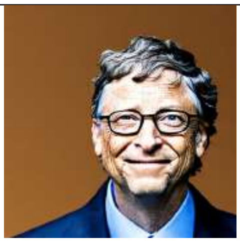
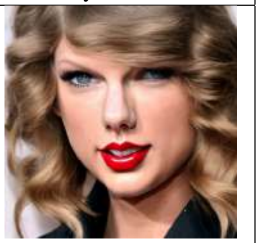
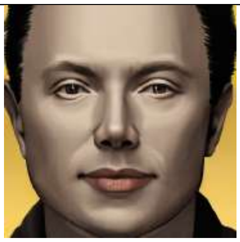
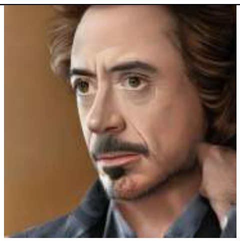
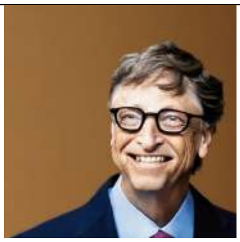
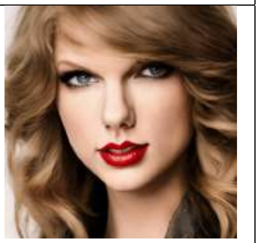
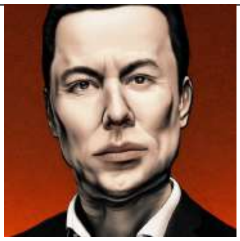
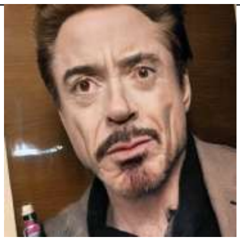
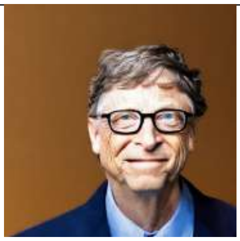
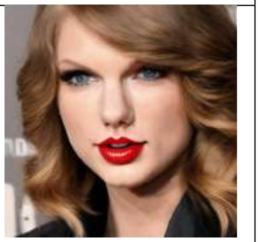
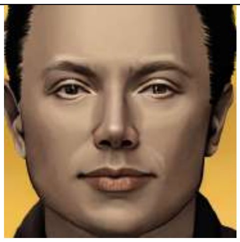
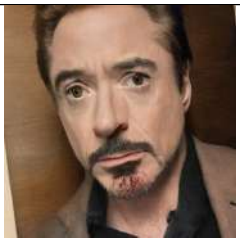
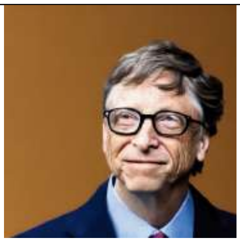
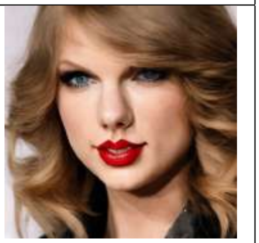
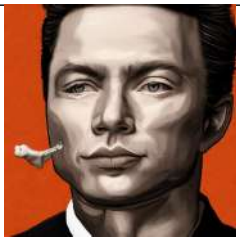
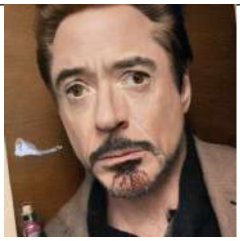
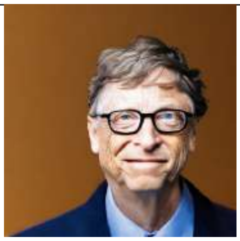
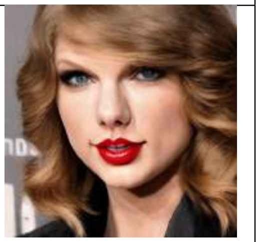
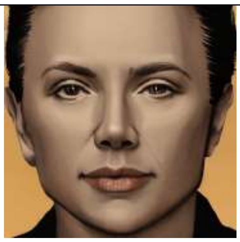
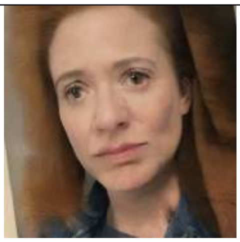
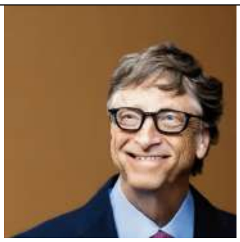
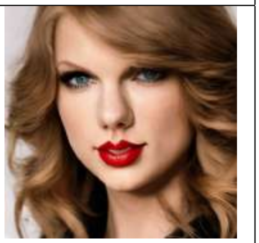
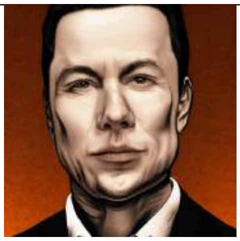
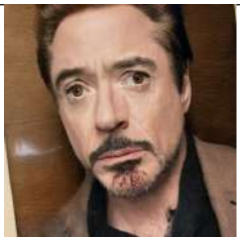
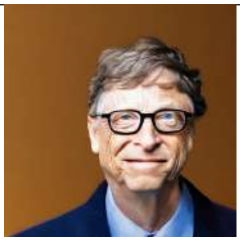
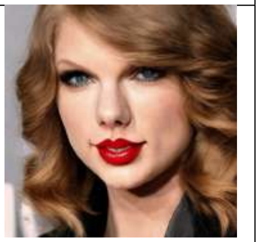
In table II we see that even though the concept is forgotten overall basic information still remains such as the overall shape, being human, being middle aged etc.

B. Results for different loss functions

At the bottom of Table III, we have compared several loss functions with skipped layers and full layers to observe the corresponding outputs in order. The first thing to notice is that the images of Bill Gates and Taylor Swift remained almost unchanged because their appearances are not similar to Elon Musk's. On the other hand, Elon Musk and Robert Downey Jr. share similarities in terms of age and other features, which can cause difficulties in forgetting the undesired object. Skipping the first and last layers does not seem to contribute significantly to the unlearning process, but they appear to assist in making the results more similar to those of the original model (this comparison is possible because we have set the random seed for the diffusion pipeline).

Attention Re-steering and Distraction both performed notably well in terms of forgetting the target object. However, a key difference is that Attention Re-steering tended to change the original photo, which may not be desirable in certain cases. In contrast, the Distraction method demonstrated its effectiveness in forgetting the target object while minimizing changes to the original photo, striking a balance between successfully removing the undesired object and preserving the integrity of the original image.

When it comes to the Random Stimulus method, the randomization process can cause significant problems. The forgetting procedure may exceed the boundaries of the target object due to the randomness involved, leading to unintended changes in the surrounding areas of the image. This lack of control and precision in the forgetting process can be challenging to handle and may result in suboptimal outcomes.

		Target Object to Forget	Test Objects		
		Elon Musk	Robert Downey Jr.	Bill Gates	Taylor Swift
Original					
Attention Re-Steering	F				
	S				
Distraction	F				
	S				
Random Stimulus	F				
	S				

[F] Training all attention blocks

[S] Skipped the first and the last attention blocks

TABLE III

OVERALL COMPARISON TABLE

In conclusion, while all the methods compared in Table III demonstrate some level of success in forgetting the target object, the Distraction method stands out for its ability to effectively remove the undesired object while minimizing changes to the original photo. Its targeted approach and preservation of the original content make it a promising choice for applications where maintaining the integrity of the image is of utmost importance.

V. Challenges and Solutions

Dataset

Initially, we attempted to allow the diffusion model to generate its own dataset. However, the model was unable to produce high-quality images. Consequently, we decided to use the Celebrity Face Image Dataset by VISHESH THAKUR on Kaggle**??**. This dataset provides easy setup to work on celebrity images by having less but quality images. As an extra, we used 5 Faces Dataset by ANKUSH KUWAR on Kaggle**??**. In addition to these, we let user to generate and create its own dataset using the diffusion model, but since the model is not perfect, it can cause to the trouble at the end. We suggest you to use the datasets that includes quality images.

Model Size

Deep neural networks typically require training on GPUs or TPUs to leverage their parallel processing capabilities. Initially, we attempted to use our own computers, which were equipped with GPUs. However, the computational demands and the size of the models necessitated a shift to Google Colab. On Google Colab, we were able to utilize the Nvidia A100 GPU, which offers high VRAM and exceptional computational power. This transition significantly accelerated our progress.

Developing Research Area

As previously mentioned, machine unlearning is an emerging field with significant potential for innovation. This nascent area offers a wealth of opportunities for researchers to explore and develop new methodologies. However, the lack of established techniques and best practices also presents challenges. Researchers must navigate uncharted territory, experimenting with various approaches to determine what works best.

One of the primary difficulties we encountered was developing effective loss functions for unlearning. The importance of hyper-parameters in this context is also not well understood, and best practices for their optimization remain unclear. Additionally, measuring the success of unlearning is not trivial, as it involves complex metrics to ensure that the model has truly forgotten the targeted information while maintaining its overall utility.

Importance of Hyper-parameters and Preventing Catastrophic Forgetting

One of the significant challenges in machine unlearning is the careful selection and tuning of hyper-parameters. Hyper-parameters such as learning rate, batch size, and the weight of the loss functions play a crucial role in the success of the unlearning process. Incorrect settings can lead to either insufficient unlearning or catastrophic forgetting, where the model loses essential information.

Learning Rate and Batch Size: The learning rate determines how quickly the model updates its parameters. A high learning rate can cause excessive forgetting, while a low learning rate may result in inadequate unlearning. Similarly, batch size affects training stability; small batch sizes can lead to noisy updates, whereas large batch sizes may smooth out updates too much, hindering effective unlearning.

To address these issues, we employed learning rate decay, which gradually reduces the learning rate during training, allowing for more precise parameter adjustments. We also experimented with different batch sizes to find an optimal balance for effective unlearning without causing instability.

Monitoring Attention Loss: We monitored attention loss, which is unconventional in machine unlearning tasks. By tracking attention loss, we ensure the model focuses less on target concepts while maintaining overall functionality. This helps prevent over-fitting to the unlearning task and ensures the model's performance on other tasks remains intact.

Preventing Catastrophic Forgetting: Catastrophic forgetting occurs when the model forgets not only the target concepts but also other important information. To prevent this, we balanced the reconstruction loss and attention loss in our unlearning objective by keeping the hyper-parameters balanced. The reconstruction loss ensures the model retains its ability to generate high-quality images, while the attention loss focuses on unlearning the target concepts.

By iteratively adjusting hyper-parameters and monitoring losses, we achieved a state where the model effectively forgets target concepts without compromising overall performance. This balance is crucial for successful machine unlearning and highlights the importance of hyper-parameter tuning in this emerging field.

VI. Future Work

In a nutshell, there are several things that can boost the performance of the unlearning model.

- **Better Diffusion Model:** The quality of the generated images is directly related to the quality of the diffusion model. Therefore, using a better diffusion model can improve the quality of the generated images and the performance of the unlearning model. So we can use directly the high quality images generated by the model without needing to use the a dataset.
- **Better Loss Function:** The loss function used in the unlearning model can also affect the performance of the model. By using a better loss function, we can improve the performance of the unlearning model and achieve better results.
- **Smaller Model:** A smaller model which does the same thing as the main model directly increases the re-usability and decreases run-time and resource allocation.
- **Surgery Operation:** As a different perspective, similar to removing the cancer from the skin, unlearn the object

and the relationship between it and the other objects, then re-train on the forgotten side-concepts can guarantees to remove the undesired object from the model while protecting the robustness. But this requires more energy and resource allocation than just unlearning the target concept so the drawbacks must be considered.

VII. Ethical Considerations

Machine Unlearning is a relatively new concept in the field of machine learning and artificial intelligence. As such, there are several ethical considerations that need to be taken into account when developing and deploying machine unlearning techniques. These considerations can be broadly categorized into three main areas: privacy, fairness, transparency and efficacy.

- **Privacy and Legal Compliance:** One of the primary ethical concerns with machine unlearning is the potential impact on privacy and the risk of encouraging unauthorized data usage. By allowing models to forget specific target concepts, there is a possibility that companies may be tempted to use unauthorized data to improve their model's performance, knowing that machine unlearning can be applied afterwards to remove the sensitive information. It is essential to establish clear guidelines and regulations to prevent unauthorized data usage and to ensure that machine unlearning is conducted in a responsible and ethical manner.
- **Fairness:** Another ethical consideration is the impact of unlearning on fairness. Machine unlearning has the potential to introduce bias into the model, leading to unfair outcomes for certain groups or individuals. It is crucial to ensure that unlearning does not disproportionately affect any particular group and to monitor the model's performance to detect and address any biases that may arise.
- **Transparency:** Transparency is also one of the main ethical considerations of machine unlearning. It is essential to be transparent about the unlearning process and to provide clear explanations of how it works and what data is being removed from the model. This can help build trust between the model developers and users and ensure that unlearning is conducted in a transparent and accountable manner.
- **Efficacy:** The efficacy of machine unlearning techniques can be the most important factor because it determines the success of the unlearning process. It is essential to evaluate the effectiveness of unlearning methods and to ensure that they achieve the desired outcomes without compromising the model's performance. This can help ensure that unlearning is conducted in a responsible and ethical manner and that the model remains accurate and reliable after the unlearning process.

VIII. Conclusion

In this study, we introduced a novel approach for machine unlearning in diffusion models to selectively forget target concepts while maintaining the model's overall drawing ability. Our method focuses on manipulating the attention mechanisms within the model by strategically training specific cross-attention modules. We employ custom loss functions to guide the model's attention away from the target concepts, enabling effective unlearning without compromising the model's performance on other tasks.

Through extensive experiments, we demonstrated the effectiveness of our proposed method in successfully removing the target concepts without compromising the model's ability to generate high-quality images. We compared different loss functions, such as Attention Re-steering, Distraction, and Random Stimulus, and found that the Distraction method achieved the best balance between forgetting the target concept and minimizing changes to the original image.

Our study also revealed that some information, such as colors, trivial shapes, and overall object structures, may be inherent to the textual embedding of the concept and cannot be completely removed without compromising the model's performance. This finding suggests that future research should explore methods to address this challenge and further improve the unlearning process.

Last but not least, we demonstrated that skipping the first and last layers does not affect in terms of forgetting, but acts like regularization factor to the model and preserves the general information about the prompt to be plotted. After skipping, the output of the model becomes closer to the original model.

In conclusion, our proposed machine unlearning approach for diffusion models demonstrates promising results in selectively forgetting target concepts while maintaining the model's drawing capacity. The insights gained from this study contribute to the growing field of machine unlearning and pave the way for further advancements in developing more controllable and privacy-preserving generative models.

Acknowledgment

The authors would like to thank Behçet Uğur Töreyin for their support throughout the project process.

References

[1] Eric Zhang et al., Forget-Me-Not: Learning to Forget in Text-to-Image Diffusion Models, arXiv preprint arXiv:2211.08332, 2023
doi: https://arxiv.org/pdf/2303.17591

[2] Mengnan Zhao et al., Separable Multi-Concept Erasure from Diffusion Models, arXiv:2402.05947v1 [cs.LG] 3 Feb 2024
doi: https://arxiv.org/pdf/2402.05947

[3] Yongliang Wu et al., Unlearning Concepts in Diffusion Model via Concept Domain Correction and Concept Preserving Gradient, arXiv:2405.15304v1 [cs.LG] 24 May 2024
doi: https://arxiv.org/pdf/2405.15304

[4] Rohit Gandikota et al., Erasing concepts from diffusion models. In Proceedings of the IEEE/CVF International Conference on Computer Vision. 2426–2436. 2023

[5] Nupur Kumari et al., Ablating concepts in text-to-image diffusion models. In Proceedings of the IEEE/CVF International Conference on Computer Vision. 22691–22702. 2023

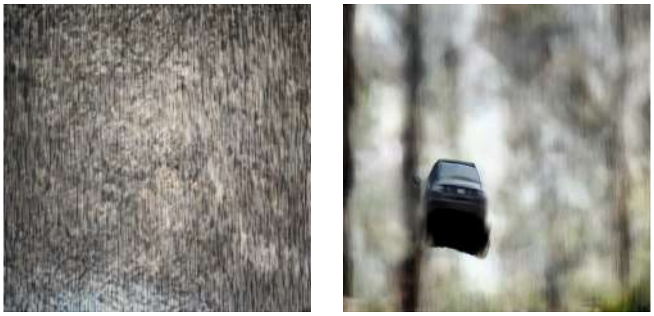
[6] Mengyao Lyu et al., One-dimensional Adapter to Rule Them All: Concepts, Diffusion Models and Erasing Applications. arXiv preprint arXiv:2312.16145, 2023

[7] https://www.kaggle.com/datasets/vishesh1412/celebrity-face-image-dataset

[8] https://www.kaggle.com/datasets/anku5hk/5-faces-dataset

(BONUS) OUR MILESTONES DURING THE PROJECT

Without Noise:

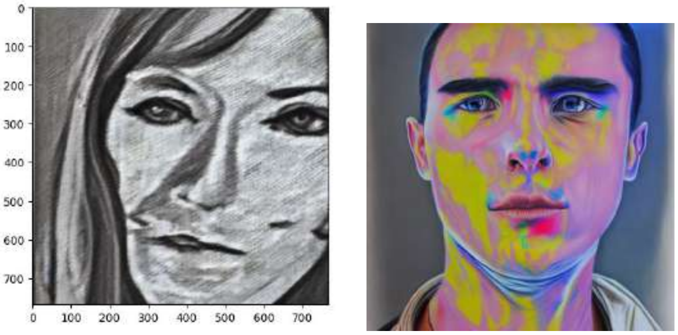


An image of a man A car is flying away in the forest

TABLE IV
WITHOUT NOISE

As we can see, without adding the noise, the model performed poorly in terms of preserving the drawing capacity. Still, it actually shows that the model completely forgets the closer terms related to ”Elon Musk” (See *an image of a man*), but for more unrelated words, the model is still able to draw the silhouette of the car in front and tree silhouette in the background.

With Noise:



An image of a woman An image of a colorful man

TABLE V
WITH NOISE

At this point, we realized that something is wrong with our code because it tries to maximize the whole error directly and the fastest way to achieve is to avoid using the colors. We had to the code.