Defending images against Generative AI based malicious editing

Parneet Singh[1], Akhil Dubey[2], Aman Kumar Jha[3] and Dr. Savita Yadav[4]
[1] *Computer Science Department, Netaji Subhas University of Technology, India*
Email: parneet300802@gmail.com
[2] *Computer Science Department, Netaji Subhas University of Technology, India*
Email : dubeyakhil7024@gmail.com
[3] *Computer Science Department, Netaji Subhas University of Technology, India*
Email : amanjha1152001@gmail.com
[4] *Computer Science Department, Netaji Subhas University of Technology, India*
Email : savita.yadav@nsut.ac.in

Abstract: In an era where generative AI models have enormous powers to manipulate digital content, defending photos against alterations initiated by these models has emerged as a critical challenge. This paper explores the landscape of protecting images against AI systems that edit images based on textual prompts. We have proposed different attacks on the diffusion models of generative AI that would compel them to generate unrealistic images. By analyzing the vulnerabilities inherent in such systems and proposing robust defense mechanisms, this study contributes to the ongoing discourse on safeguarding the authenticity and trustworthiness of digital imagery in an increasingly AI-driven world.

Keywords — diffusion models, generative AI, latent diffusion model, projected gradient descent

1 Introduction

In a time where generative AI models may transform digital content to great extent, protecting images from changes made by these models has become a crucial task. Various text to image models are present in the market. In a survey conducted by Chenshuang Zhang et al. [1], various text to image models such as GAN based methods, Autoregressive methods and diffusion based methods were investigated. These models were evaluated based on image quality using FID (Frechet Inception Distance) on MS-COCO dataset, CogView gave the highest FID score. Ethical risks of these text to image models were also stated like misuse for malicious purposes and security and privacy risks. Large diffusion models can be utilized for many different picture synthesis and editing tasks; models like the ability of DALL·E 2 and Stable Diffusion to produce excellent, photorealistic photos is well known. But because these models are so simple to use, there are worries that they could be abused, for example, to produce offensive or damaging digital content. For instance, a malicious party may use a diffusion model to deliberately alter images of persons that have been posted online. Evil-minded people can nonetheless alter real photos or create wholly phony ones from scratch using programmes like Photoshop, even in situations where diffusion models aren't present. The main issue that huge generative models bring about is that these people may now easily produce realistic edited photos, without the need for costly equipment or specialized training.

Our paper explores the landscape of protecting images against AI systems that edit images based on textual prompts. Different types of attacks can be performed against generative models. Hui Sun et al. [2], described various kinds of attacks that can be performed to defend against generative models. Attacks against generative models were categorized into 3 categories : Data injection, data manipulation and logic corruption. A technique for producing targeted, robust, and universal adversarial picture patches that can trick image classifiers was presented by Tom B. Brown et al. [3]. The classifiers can report a selected target class via printing, adding the patches to any scene, and taking a picture of the patches.. This paper compares the effectiveness of different attack methods and demonstrates that the patches can be camouflaged to reduce their saliency.

We have proposed different attacks on the latent diffusion models of generative AI that would compel them to generate unrealistic images.

2 Related Work

We would be working on latent diffusion models (LDM), which were presented by Robin Rombach et al. [4], it is an easy approach to boost diffusion models' training and sampling effectiveness without sacrificing their quality. by introducing encoding of input image to a latent space. This article also showed favorable outcomes when compared to state of the art approaches. In his session, Robin Rombach discusses latent diffusion models, which are probabilistic models intended to gradually denoise a normally distributed variable in order to learn the reverse process of a fixed Markov Chain. Language transformers are used in conjunction with encoders and decoders for huge, high-quality images. The 3-dimensional latent diffusion models were suggested by Gabriela et al. [5], a new model of diffusion that, in response to a text prompt, generates RGBD images (RGB images with matching depth maps). A KL-regularized diffusion model, LDM3D, was developed by adapting Stable Diffusion.

Based on the diffusion models, a text based image editing model, Imagica was proposed in the paper published by Bahjat Kawar et al. [6], which takes one image and a brief text prompt explaining the intended edit using diffusion models. It then seeks to apply the requested edit while retaining the greatest number of details possible from the image. Diffusion models have also been investigated for text-conditional image synthesis problems in terms of photorealism. Employing GLIDE using classifier-free guidance and CLIP guidance, two different guiding methodologies, Alex et al. [8] investigated diffusion models for the text-conditional image synthesis problem. In terms of both photorealism and caption similarity, human assessors favor the latter, which frequently yields photorealistic examples.

Other generative models are also present for image and text synthesis, StackGAN based on Generative Adversarial Networks, was proposed in this paper published by Han Zhang et al. [7], StackGAN, a text to photo-realistic image synthesis with stacked generative networks was proposed. The text-to-image synthesis is broken down into a unique sketch-refinement process by the suggested method. Stage-I GAN uses provided text descriptions as a guide to sketch the object, according to basic color and shape constraints. Higher resolution and higher quality images are produced by Stage-II GAN, which also adds more information and fixes the flaws in Stage-I outputs. Another image to image GAN was suggested by Philip et al. [13] for translating images to images using Conditional Generative Adversarial Networks. From an observed picture (x) and a random noise vector (z), conditional GANs learn a mapping to y. Convolutional "PatchGAN" classifier is the discriminator utilized in this case; it solely penalizes structure at the picture patch scale. In order to safeguard against unauthorized editing of images by GANs, copyright protection mechanisms were introduced in the paper published by Haonan Zhong et al. [9], the performance of the copyright protection mechanisms for GANs is assessed based on how well they work with a variety of GAN Architectures. A preliminary assessment of the most advanced IPR protection techniques for GANs and their training sets (pictures) was carried out in this study. The experimental findings showed that copyright protection and accountability tracking of GAN models can be achieved with a good degree of effectiveness using recently established IPR protection strategies including watermarking, attribution, and adversarial attacks. It also demonstrates the shortcomings of the current protection strategies for the training sets' intellectual property rights (IPR)—the original, copyrighted photographs.

Ian et al. [11] suggested a method for estimating generative models that trains two models concurrently through an adversarial approach: a discriminative model D that assesses the probability that a sample came from the training data rather than G, and a generative model G that depicts the data distribution. To increase D's likelihood of making a mistake is the aim of G's training regimen. This structure can be compared to a two-player minimax game. In the space of random functions G and D, where G recovers the distribution of training data and D is always equal to 1, there is only one solution. If G and D are represented by multilayer perceptrons, then the entire system can be trained via backpropagation.

Rossler et al. [10] provides an automated benchmark for the identification of forgeries that takes into account the four alteration techniques—Face2Face, DeepFake, FaceSwap, and Neural Texture—in a realistic setting with random dimensions and random compression.While DeepFakes and NeuralTextures are learning-based approaches, Face2Face and FaceSwap are graphic-based methods. All four approaches require video pairs of the source and

target actors as input. Each process results in a video consisting of generated images. We construct ground truth masks, which is useful for training fake localization systems., and display whether or not a pixel has been altered in addition to the manipulation result. Specifically, Residual Neural Networks (RNNs) are used to extract spatial information from CNNs and to capture temporal dependencies in a face anti-spoofing system presented by Usman et al. [12]. A meta model is created by combining these predictions in order to identify spoofing.

3 Theory

Generative models

With applications in machine learning, generative models are a subclass of statistical models that seek to recreate the distribution of a given set of data by producing new data samples. These models are effective instruments for deciphering intricate datasets, opening up new possibilities for applications such as speech synthesis, image and text production, and more. The underlying probability distribution of a dataset is discovered by generative models. To ensure that the generated data points from the model are statistically comparable to the original dataset, it must fully comprehend the complex patterns and features of the data. There are four main generative models, Generative Adversarial Network, Variational Autoencoder, Flow Based and Diffusion models. All these models are great and efficient in generating high quality images, but there are some limitations in each as GAN models use adversarial training, they are renowned for having possibly unstable training and reduced generational variability. VAE is dependent on a surrogate dying. Specialized architectures are required for flow models to build reversible transforms. But the theory behind diffusion models comes from non-equilibrium thermodynamics. They create a Markov chain of diffusion stages for progressively adding random noise to the data. They then figure out how to reverse the diffusion process such that the targeted data samples are separated from the noise in the background. Diffusion models, in contrast to VAE or flow models, incorporate a high dimensional latent variable which is the same as the original data and are trained using a fixed technique.

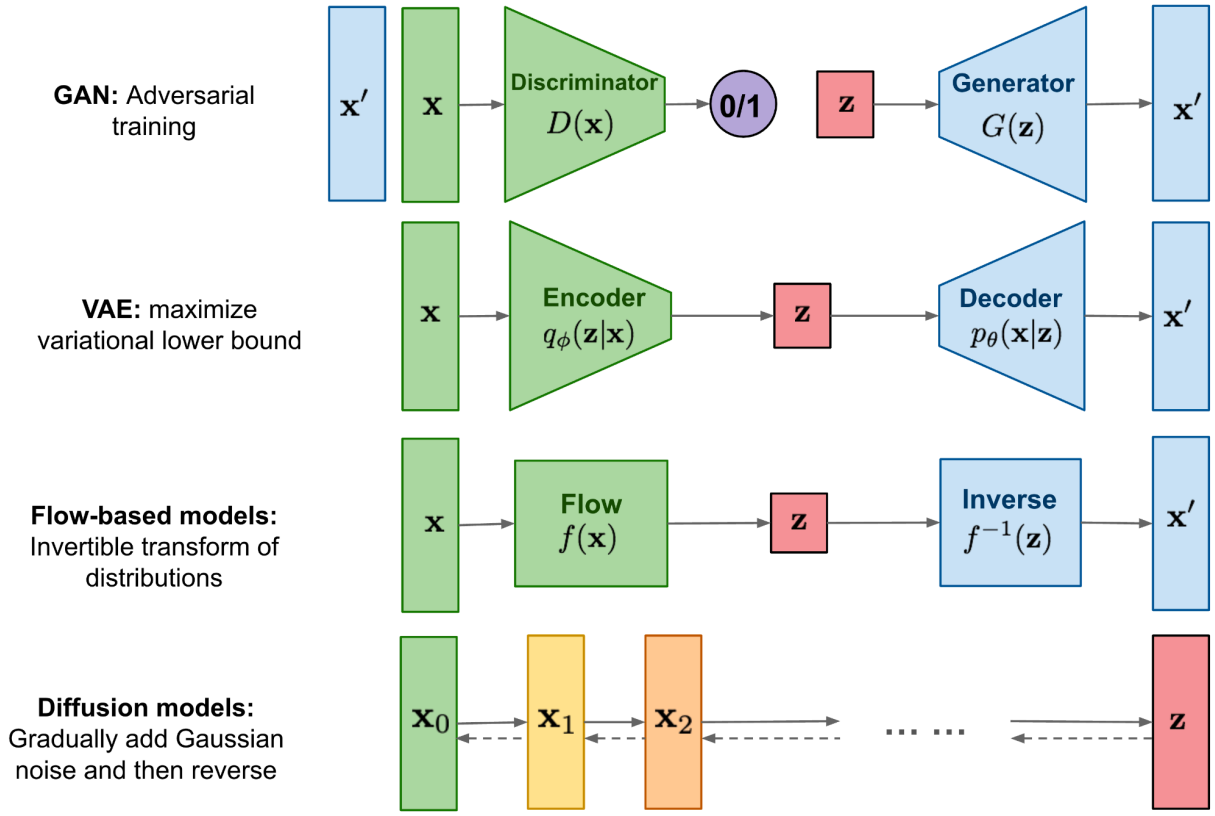

Figure 1 [Overview of generative models] (Image source : Weng, Lilian. (Jul 2021))

Diffusion models

Diffusion models are now among the most effective technologies available for producing lifelike visuals. These models are now superior to other image generative models like GANs in the caliber of the photographs that are produced. They are particularly good at creating and altering images using textual cues. To generate data that is similar to the training set, generative models—like diffusion models—are used. In essence, diffusion models work by first adding Gaussian noise to training data one after another, erasing it, and then learning to retrieve the data by reversing this noise. We may simply feed randomly sampled noise through the learnt denoising process after training to generate data using the diffusion model.

Diffusion Process

Diffusion process is divided into two categories: forward and reverse. An image is converted into noise by the forward diffusion process, and the reverse diffusion process attempts to transform the noise back into the original image. Reverse diffusion is a multi-stage process where each phase involves removing a tiny bit of noise. The whole noise is predicted by the diffusion model, not only the variation between steps t and t-1.

The diffusion process is a stochastic differential process that is the foundation of diffusion models. The problem of (approximate) sampling from a genuine picture distribution $q(\cdot)$ can be viewed as a series of denoising difficulties through this process.. More specifically, for each T steps, the diffusion process adds noise progressively to samples $x_0 \sim q(\cdot)$ to produce samples $x_1,..., x_T$, where $x_t+1 = a_t x_t + b_t \varepsilon_t$, and ε_t are drawn from a Gaussian distribution. For a more thorough discussion of the diffusion process, see [14].

Latent Diffusion model

We shall focus on LDMs, a particular category of diffusion models. Rather than using the input (picture) space, these models employ the previously mentioned diffusion process in the latent space. It turned out that this modification preserves the high caliber of generated samples while enabling faster and more effective training and inference. There is one major difference between training an LDM and a typical diffusion model. Specifically, the initial step involves mapping the input image x_0 to its latent representation, $z_0 = \varepsilon(x_0)$, where ε represents a preset encoder, to train an LDM. Once noise has been added progressively, the diffusion process proceeds as previously (although in the latent space).

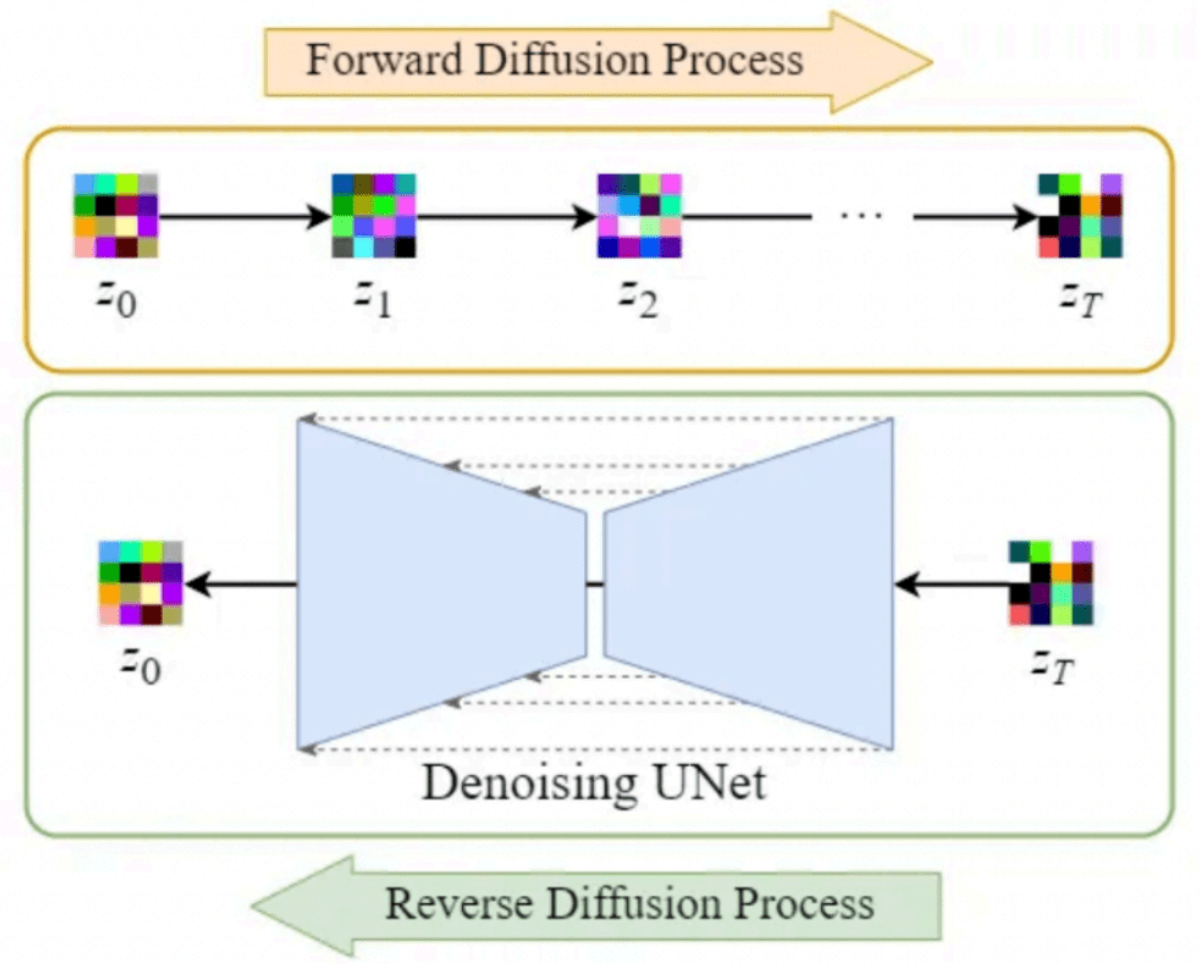

Figure 2 [Overview of Latent diffusion models] (Image source : Ignacio (Nov 2023))

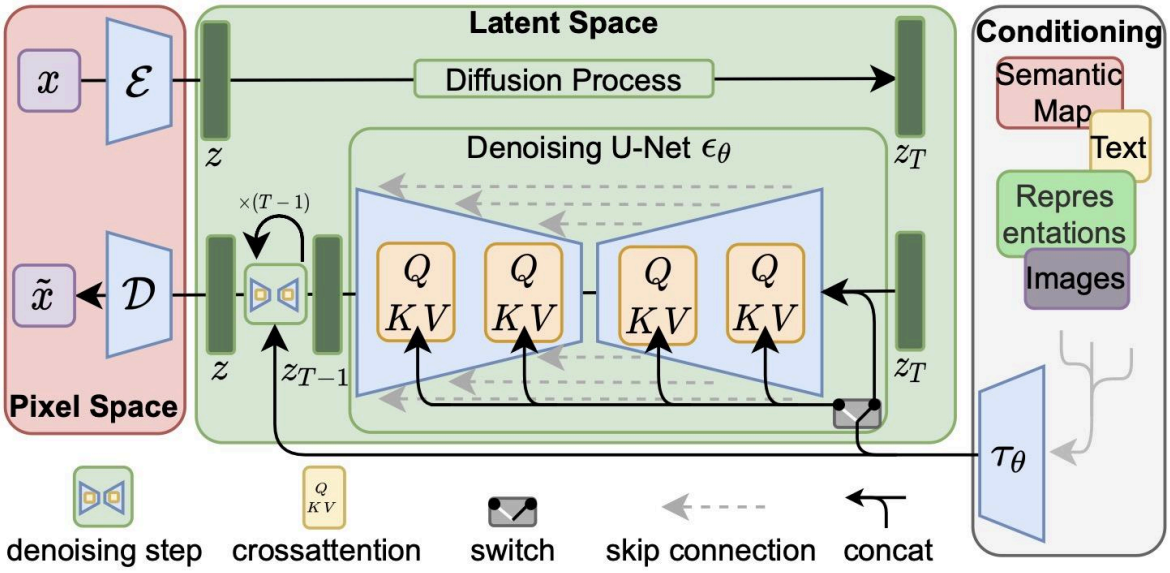

Figure 3 [LDM Architecture] (Image source: Rombach et al. 2022)

By default, an LDM takes a random sample from the picture distribution q(·) that it was trained on. It turns out, though, that natural language can also be used to direct the sample. The user-defined textual prompt t can be embedded into the latent representation, which is produced during the process, to accomplish this.

4 Proposed Approach

We now go over how we used latent diffusion models (LDMs) to immunize images, or make them more difficult to modify. Utilizing methods from the literature on adversarial attacks and adding adversarial perturbations to images is the fundamental component of our approach. We specifically offer an encoder assault to carry out this tactic.

Encoder attack :

Remember that an LDM creates a latent vector representation (LVR) from an image by first encoding it using an encoder ε. Thereafter, a fresh image is produced using this LVR. Our encoder attack's primary objective is to disturb the operation by making the encoder convert the image to a poor representation. We used projected gradient descent (PGD) for this:

$$\delta_{encoder} = \arg\min_{\|\delta\|_\infty \leq \epsilon} \|\mathcal{E}(\mathbf{x}+\delta) - \mathbf{z}_{targ}\|_2^2,$$

where z_{targ} is some target latent representation (for example, the representation of a gray image created with encoder ε) and x is the image to be immunized. The optimisation problem's solutions produce tiny, undetectable perturbations $\delta_{encoder}$ that, when combined with the original image, produce a (immunized) image that, from the encoder's point of view, is comparable to the (gray) target image. The LDM then creates an unrealistic or irrelevant image as a result.

We have implemented our encoder attack for two cases, the first case is when someone tries to edit an image with a prompt to modify the input image. Second case is when someone tries to edit parts of an existing image via inpainting, for example someone may want to edit the image while keeping the face of the person in it the same as the original. For the first case, we implement a simple PGD attack on the whole image and then, after applying the attack on the image, the image becomes immunized to stable diffusion. We apply stable diffusion on both original and immunized images to compare the results. For the demonstration, we have used a picture of a dog with the prompt " dog under heavy rain and muddy ground ".

Figure 4 [Outcome of encoder attack for the first case]

For the second case, instead of applying the PGD attack directly on the image, we use masking for the case if someone tries to edit the image while keeping the face of the person the same to get better results of editing the image. Using both original and masked images, we attack the image towards embedding of some random target image so that the edited image created by the diffusion model is not realistic. For the demonstration, let's take the prompt as "man in a hospital".

Figure 5 [Demo image]

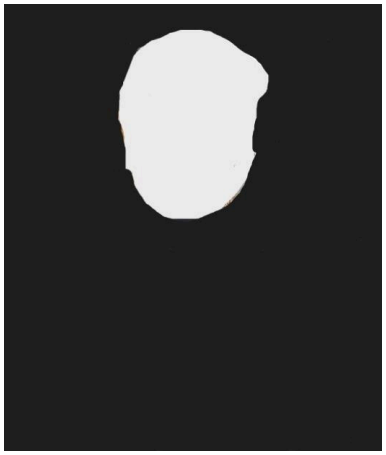

Figure 6 [Masked demo image]

Figure 7 [Target image]

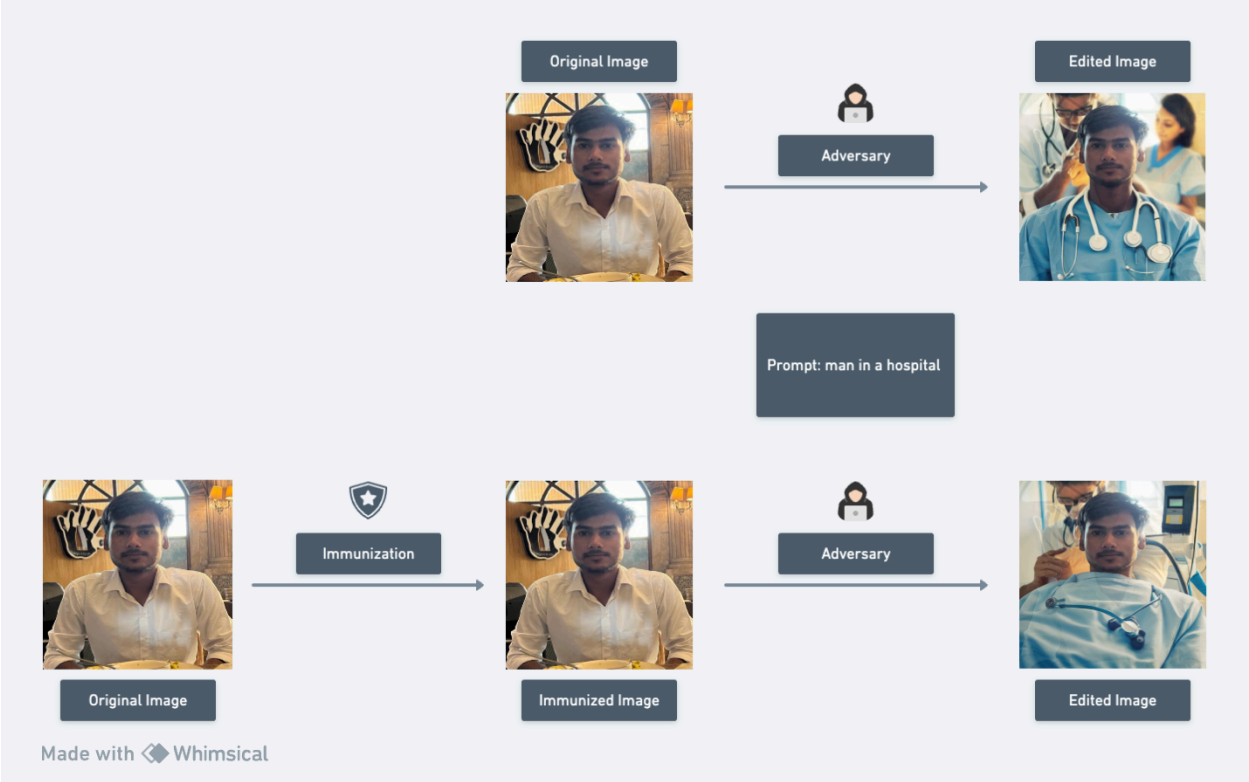

Figure 8 [Outcome of encoder attack]

The outcome of this attack shows that when we applied diffusion model on the original image it generates a realistic image of a man in a hospital as a doctor with stethoscope based on the prompt but when the same model is applied on the immunized image after our PGD attack, it generated a unrealistic image in which the man with stethoscope is lying on a bed which is not realistic. In this way, our encoder attack prevents the diffusion model from editing the images in a realistic way and therefore, defending images from diffusion models. More results are shown in the next section.

5 Results

Using the stable diffusion model for image editing includes importing the image, creating a mask to specify which areas of the image need to be altered, and then responding with a text prompt to determine how to adjust the remaining portion of the image. On the basis of that prompt, the SDM then creates an edited version. The encoder attack presented by us causes the SDM to generate unrealistic edited images as shown in above examples. The modifications of the immunized and non-immunized photos are different, as shown by the examples.

The following results were achieved using different text prompts :
In all the following results, the first image is the source image, second is the immunized image, then we apply LDM with a text prompt on both these images and compare the results of images produced.

Prompt: man on a beach | Seed: 82960

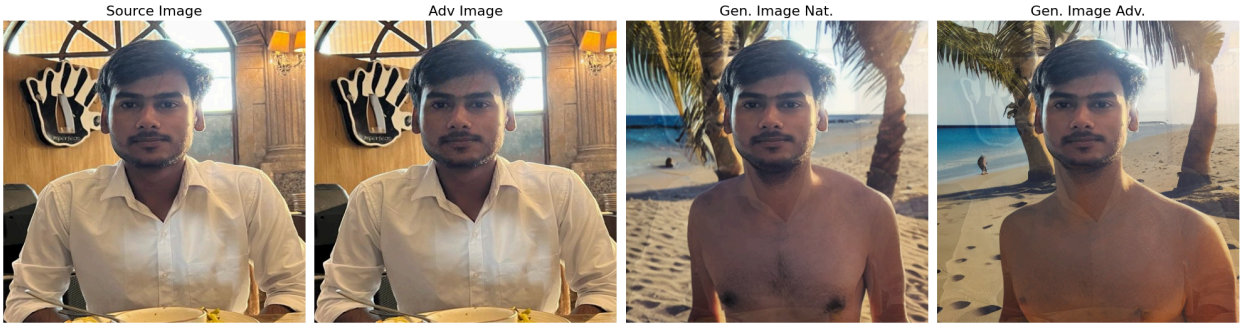

Figure 9 [Result with text prompt "man on a beach"]

Prompt: man riding a motorcycle at night | Seed: 82960

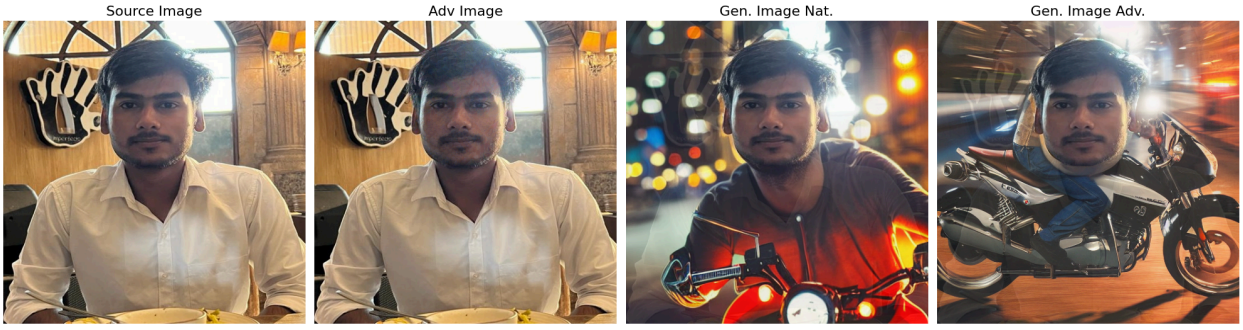

Figure 10 [Result with text prompt "man riding a motorcycle at night"]

Prompt: man in a hospital | Seed: 17907

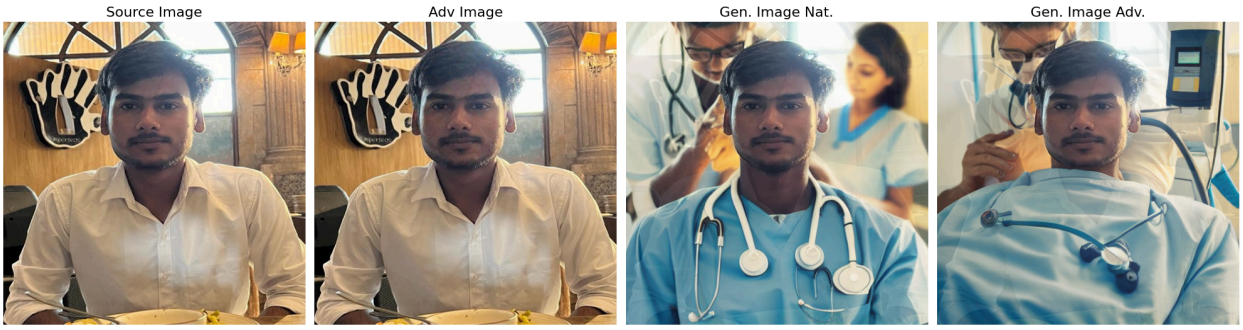

Figure 11 [Result with text prompt "man in a hospital"]

Prompt: man in a classroom | Seed: 82960

Figure 12 [Result with text prompt "man in a classroom"]

References

[1] Zhang, Chenshuang, Chaoning Zhang, Mengchun Zhang, and In So Kweon. "Text-to-image diffusion model in generative ai: A survey." *arXiv preprint arXiv:2303.07909* (2023).

[2] Sun, Hui, Tianqing Zhu, Zhiqiu Zhang, Dawei Jin, Ping Xiong, and Wanlei Zhou. "Adversarial attacks against deep generative models on data: a survey." *IEEE Transactions on Knowledge and Data Engineering* 35, no. 4 (2021): 3367-3388.

[3] Brown, Tom B., Dandelion Mané, Aurko Roy, Martín Abadi, and Justin Gilmer. "Adversarial patch." *arXiv preprint arXiv:1712.09665* (2017).

[4] Rombach, Robin, 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*, pp. 10684-10695. 2022.

[5] Stan, Gabriela Ben Melech, Diana Wofk, Scottie Fox, Alex Redden, Will Saxton, Jean Yu, Estelle Aflalo et al. "LDM3D: Latent Diffusion Model for 3D." *arXiv preprint arXiv:2305.10853* (2023)

[6] Kawar, Bahjat, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. "Imagic: Text-based real image editing with diffusion models." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6007-6017. 2023

[7] Zhang, Han, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N. Metaxas. "Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks." In *Proceedings of the IEEE international conference on computer vision*, pp. 5907-5915. 2017.

[8] Nichol, Alex, 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).

[9] Zhong, Haonan, Jiamin Chang, Ziyue Yang, Tingmin Wu, Pathum Chamikara Mahawaga Arachchige, Chehara Pathmabandu, and Minhui Xue. "Copyright protection and accountability of generative ai: Attack, watermarking and attribution." In *Companion Proceedings of the ACM Web Conference 2023*, pp. 94-98. 2023.

[10] Rossler, Andreas, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. "Faceforensics++: Learning to detect manipulated facial images." In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1-11. 2019.

[11] Goodfellow, Ian, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. "Generative adversarial nets." *Advances in neural information processing systems* 27 (2014).

[12] Muhammad, Usman, Md Ziaul Hoque, Mourad Oussalah, and Jorma Laaksonen. "Deep ensemble learning with frame skipping for face anti-spoofing." In *2023 Twelfth International Conference on Image Processing Theory, Tools and Applications (IPTA)*, pp. 1-6. IEEE, 2023.

[13] Isola, Phillip, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. "Image-to-image translation with conditional adversarial networks." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1125-1134. 2017.

[14] Weng, Lilian. (Jul 2021). What are diffusion models? Lil'Log.
https://lilianweng.github.io/posts/2021-07-11-diffusion-models

[Figure 1] Weng, Lilian. (Jul 2021). What are diffusion models? Lil'Log.
https://lilianweng.github.io/posts/2021-07-11-diffusion-models/ , fig 1.

[Figure 2] Ignacio Aristimuño (Nov 2023). An Introduction to Diffusion Models and Stable Diffusion
https://blog.marvik.ai/2023/11/28/an-introduction-to-diffusion-models-and-stable-diffusion

[Figure 3] Rombach, Robin, 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* 2022., pp. 10687, fig 3.