Challenging Generative Models in the Battle Against Image Manipulations

Asaf Yitzhaki
Gal Godsi
Yotam Maoz

Abstract

Recently, the ease of image manipulation through state-of-the-art generative models such as StyleGAN and Latent Diffusion has raised concerns about trustworthiness in online content. In this paper, we present an effective method for immunizing images against these manipulation techniques by generating adversarial examples that neutralize the manipulation capabilities of these models by adding small perturbations. This perturbation is designed to be imperceptible to humans but enough to fool the manipulation models, forcing them to generate unrealistic images. We also evaluate our approach on various StyleGAN and Latent Diffusion models in addition to the ones we designed it for. Finally, we discuss a policy component necessary for challenging the current lack of prevention measures against AI-driven image manipulation.

1 introduction

In today's interconnected world, a significant portion of the data shared online through various social media platforms, blogs, news articles, and other content sources is exposed to numerous individuals, including many people beyond those for whom the data was originally intended. A significant concern arises from the potential for malicious actors to exploit accessible online data through the utilization of computational tools [5, 13, 24]. Such individuals may intentionally alter and disseminate manipulated content in order to misrepresent factual information, contributing to the proliferation of false narratives within digital ecosystems. One area of particular concern involves the potential for malicious manipulation of images containing facial features [13, 21, 22, 25]. Such actions have the potential to distort the true representation of an individual and mislead the public perception of reality.

The increasing prevalence of "deepfakes" and other malicious uses of AI-generated content has highlighted the difficulty in distinguishing between them and genuine ones. The ease of use, high availability, and realistic results of these tools have made AI image manipulation a common and widely used practice [13, 17, 19, 21, 22, 24, 37]. Unfortunately, malicious users are also utilizing these tools to manipulate images, making it difficult to trust that shared online content will not be manipulated by AI-generated models [19, 24]. While completely eliminating image manipulation is impossible, as malicious actors can manually edit photos even without AI tools, our goal is to challenge AI-driven manipulation tools that enable these bad actors to create manipulations cheaply, and without specialized skills.

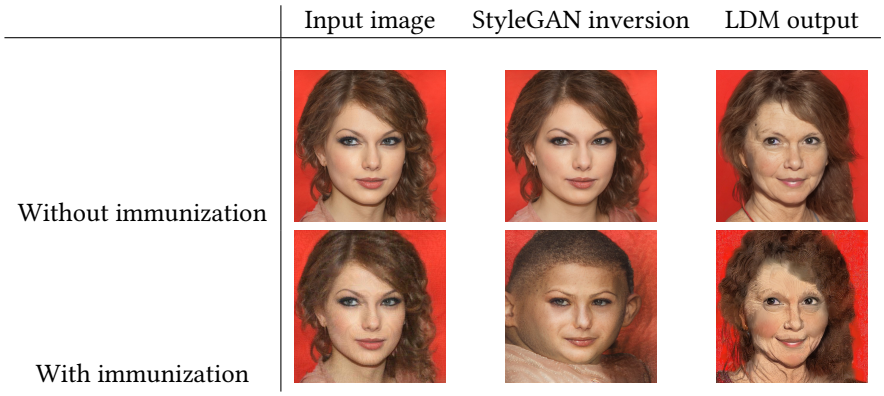
Our project aims to develop an effective method for immunizing images against manipulation techniques used by state-of-the-art generative models such as StyleGAN [21, 22] and Latent Diffusion models [7, 33]. Our approach focuses on generating adversarial examples [18, 46, 47] that can neutralize the manipulation capabilities of these generative models. To achieve this, we aim to gain a comprehensive understanding of the manipulation techniques employed by StyleGAN and Latent Diffusion models in their preliminary phase, specifically the encoder. Our goal is to identify the features that make an image susceptible to manipulation and then add a small perturbation to the original image that changes these features [18], making the image less susceptible to manipulation. This perturbation is designed to be imperceptible to humans but enough to fool the manipulation models, effectively immunizing the image against such manipulation. See Figure 1.

2 Related Work

2.1 StyleGAN

StyleGAN [21, 22] is a state-of-the-art generative adversarial network (GAN) architecture [13] for synthesizing high-fidelity photographic images within a particular domain. StyleGAN models are typically trained on specific datasets

Figure 1. Example of immunizing an image. The two images in the first column are nearly the same, however the StyleGAN model and the LDM produce unrealistic images when given the second image as input.



such as FFHQ faces [21], AFHQ animal faces [9], and celeba-hq [20], and work best for images similar to the training data.

Unlike traditional GANs that control the generation process via a latent code, StyleGAN manipulates image synthesis through intermediate latent controls [21]. This enables intuitive editing of generated images by altering the latent space. However, these manipulations are limited to the model's training domain, like faces for FFHQ, and may not handle the surrounding context well.

GAN inversion techniques like Image2StyleGAN [1], Image2StyleGAN++ [2], e4e [42], ReStyle [4], and pSp [31] allow inverting images from the model's domain into the latent space for editing. This is done either by directly optimizing the latent vector [1, 2] or by training an encoder network [4, 31, 42]. Direct optimization typically provides more accurate inversion but encoders are faster at inference [15, 32, 44]. Therefore, most real-world applications utilize encoders for GAN inversion [4, 12, 14, 31, 42]. After inversion, manipulations can be done in the disentangled StyleGAN space [36, 41].

Additionally, StyleClip [26] leverages CLIP [28] to guide StyleGAN image generation and editing. Overall, while extremely effective within its domain, StyleGAN has limitations in handling out-of-domain images.

2.2 Diffusion Models

Diffusion models [17, 37] are generative models that can synthesize diverse high-quality images by gradually adding noise to data and then reversing the process. Notable diffusion models include DDPM [17], DDIM [39], ADM [10], and LDM [7, 30, 33].

Latent diffusion models (LDMs) build on standard diffusion by conditioning the generation process on a latent code. Latent diffusion models typically use an encoder network to map an image into the latent space. The encoded latent representation is then passed into the diffusion process for image generation. This enables controlled image synthesis and editing.

Latent diffusion models can handle diverse image types and maintain the surrounding context well. However, reconstruction and editing of some specific elements like faces and hands may be less accurate than domain-specific models. Latent editing can be achieved by manipulating the latent code directly or by using text prompts [7, 30, 33]. Leading latent diffusion models include Stable Diffusion [33], DALL-E 2 [29, 30], Imagen [34, 43], and Stable Diffusion XL [27].

Inpainting is a process performed by masking out certain regions of the input image during the diffusion process [27]. The model then samples the latent code in order to generate missing content that complies with the unmasked regions, effectively filling in holes. Overall, latent diffusion models allow flexible and controlled image synthesis and editing for diverse domains.

2.3 Protecting Against StyleGAN-based Image Manipulation

The paper [16] focuses on the issue of malicious use of StyleGAN-based image manipulation, which can be utilized to deceive human observers [21, 22]. The authors propose innovative approaches to protect facial images from StyleGAN-based models and enhance photo authenticity. They present two methodologies that attempt to generate imperceptible adversarial examples [18, 46, 47] capable of misleading machine learning models [11, 45].

The methods involve preventing inversion, which the paper extensively discusses as an approach to disrupt the latent vector of a given image when passed through the generator, resulting in an image different from the original; and preventing editing, focusing on disrupting only the subsequent editing process without compromising the effectiveness of the inversion process. While both approaches were explored in the research, the authors found that preventing editing was less effective than anticipated.

In the method of preventing inversion, the goal is to maximize the distance between benign samples and adversarial samples after performing inversion by employing one of three possible loss functions: pixel loss, ID loss, and latent loss. The researchers tested these options on popular StyleGAN encoders such as e4e [42], pSp [31], and ReStyle [4]. Their findings demonstrate that transferability is quite effective when the attack optimizes the latent loss, making it both the most efficient and effective method for protecting facial images from malicious manipulation using StyleGAN-based models.

2.4 Raising the Cost of Malicious AI-Powered Image Editing

The paper [35] explores the issue of malicious use of AI-powered image editing tools, specifically LDMs [7, 30, 33], which can be used to deceive human observers. The authors propose an innovative approach to increase the cost of manipulating images using machine learning models and improve photo authenticity. They present two methods for generating adversarial examples [18, 46, 47] that are imperceptible to humans but can mislead machine learning models. Specifically, they describe two techniques to add a minor perturbation to the original image in order to challenge [3, 23, 40] the well-known model, Stable Diffusion Model (SDM) v1.5 [33], when performing image editing via inpainting [27], focusing on manipulating the surroundings of the face in the image.

The first methodology is "Encoder attack", which leverages the fact that LDMs encode images into a latent vector representation before generating new images [33]. The authors' approach disrupts this process by forcing the encoder to map the input image to an unwanted representation.

The second methodology is "Diffusion attack", allowing for further disturbance of the diffusion process itself instead of just the encoder. In this attack, they manipulate the input image so that the generated final image by LDM is a specific target image (e.g., random noise or grayscale). Although diffusion attack is more powerful than encoder attack, the paper highlights that it is more challenging to execute due to requiring backpropagation through the full diffusion process, which includes repeated application of the denoising step.

3 Methodology

In this section we describe our technical approach for immunizing images against StyleGAN and LDM models. Given a StyleGAN model and an LDM, we denote by E_{GAN} and E_{DIF} the encoder networks for the StyleGAN and LDM respectively. For a vector $x \in \mathbb{R}^d$ we let $|x|_p = (\sum_i |x_i|^p)^{1/p}$ denote the l_p norm of x and $|x|_\infty = \max_i |x_i|$ denote the l_∞ norm of x. Throughout the next section we denote by X the input image.

3.1 Generic Joint Encoder Attack

To attack both StyleGAN and an LDM model jointly using an encoder attack, one wishes to find an image $\hat{X}$ that is *close* to X in some sense, for which $E_{GAN}(\hat{X})$ and $E_{DIF}(\hat{X})$ are *far away* from $E_{GAN}(X)$ and $E_{DIF}(X)$ respectively. There are two obstacles to this approach. First, one needs to define a notion of *close* and *far away*. This can be easily solved by using some norm such as l_p norms (p can be ∞). The second obstacle is that given a possible adversarial example $\hat{X}$ for X, one cannot expect both:

$$\left|E_{GAN}(\hat{X}) - E_{GAN}(X)\right|_p,$$

and:

$$\left|E_{DIF}(\hat{X}) - E_{DIF}(X)\right|_p,$$

to be maximized by $\hat{X}$. Thus one should use a trade-off function $d(x, y)$ to overcome this inherent trade-off.

Thus, a generic joint encoder attack on StyleGAN and an LDM requires three things, a norm $|\cdot|_p$ (p can be ∞), a trade-off function d and an $\epsilon > 0$ which measures the maximum amount of noise we add to X. The attack is thus the following maximization problem:

$$\max_{|\hat{X}-X|_p \le \epsilon} d\left(\left|E_{GAN}(\hat{X}) - E_{GAN}(X)\right|_p, \left|E_{DIF}(\hat{X}) - E_{DIF}(X)\right|_p\right).$$

One can also define this attack by the noise n added to the image X, that is, writing $\hat{X} = X + n$ one sees the above attack is equivalent to:

$$\max_{|n|_p \le \epsilon} d\left(|E_{GAN}(X+n) - E_{GAN}(X)|_p, |E_{DIF}(X+n) - E_{DIF}(X)|_p\right).$$

3.2 Convex-Linear Joint Encoder Attack

Our first attack is a specific instance of a generic joint encoder attack. In this attack we require an additional parameter $\alpha \in [0, 1]$, which defines our trade-off function:

$$d_\alpha^{convex}(x, y) = \alpha x + (1 - \alpha)y.$$

The attack is thus:

$$\max_{|n|_p \le \epsilon} \alpha\, |E_{GAN}(X+n) - E_{GAN}(X)|_p + (1-\alpha)\, |E_{DIF}(X+n) - E_{DIF}(X)|_p,$$

which can be solved using PGD. Note that if one model (StyleGAN or LDM) is more susceptible to an encoder attack then we might end up in a position where the optimal noise n only disrupts the encoding of that model without hurting the other one.

3.3 Adaptive Joint Encoder Attack

The observation at the end of the last section exposes an inherent flaw in the convex-linear trade-off function. To solve this problem, we use the following trade-off function which also requires a parameter $\alpha \in [0, 1]$:

$$d_\alpha^{adp}(x, y) = \alpha \log(1 + y)x + (1 - \alpha)\log(1 + x)y.$$

The corresponding joint encoder attack is again solved by PGD. Note that d_α^{adp} is similar to a convex-linear distance

function where the weights on x and y depend on the other variable.

As motivation for using this function family over convex-linear trade-off functions, consider the function:

$$d(x, y) = x \log(y) + y \log(x),$$

and note that:

$$d_x = \frac{y}{x} + \log(y) \quad d_y = \frac{x}{y} + \log(x).$$

Now assume that $x, y \geq 1$. The key property of d is that when one variable is bigger than the other, the derivative in the bigger variable is small while the derivative in the smaller variable is big. For example, if x and y are relatively small, say $x, y \leq 10$ and $x \approx 5y$ then:

$$d_x = \frac{y}{x} + \log(y) \approx \frac{1}{5} + \log(y),$$

while:

$$d_y = \frac{x}{y} + \log(x) \approx 5 + \log(x) = 5 + \log(5) + \log(y).$$

Thus, if one runs gradient descent with the trade-off function d, each iteration tries to give the smaller variable a bigger boost than the bigger variable.

To make d into a trade-off function we can use, we first change $\log(x), \log(y)$ to $\log(1 + x), \log(1 + y)$ to avoid negative weights. After that we add an additional weight α to deal with the inherent different sensitivity to encoder attacks at the beginning of PGD. Finally we reach d_α^{adp}. In total, the class of distance functions d_α^{adp} are designed to balance out any difference of sensitivity to encoder attacks. Thus, they ensure that our attack "spreads out" its damage throughout both models.

3.4 Differential Attack

This attack is of an inherently different form than generic joint encoder attacks, and relies on a simple observation. StyleGAN models trained on faces preform best when run on images containing mostly a single face in the center of the image looking forward, while LDMs preform best when editing the background of an image. Thus, given an image X, we extract all the faces present in X using a face detection network, and create for each face its own image X_{face}.

After extracting all the faces from X, the attack has two subsequent parts. First run an encoder attack on the entiraty of X using **only** the LDM encoder, that is, solve the following maximization problem:

$$\max_{|\hat{X}-X|_p \leq \epsilon} \left| E_{DIF}(\hat{X}) - E_{DIF}(X) \right|_p .$$

Second, for each face in X run an encoder attack on X_{face} using **only** the StyleGAN encoder, that is, solve the following maximization problem for each face in X:

$$\max_{|\hat{X}_{face}-X_{face}|_p \leq \epsilon} \left| E_{GAN}(\hat{X}_{face}) - E_{GAN}(X_{face}) \right|_p .$$

All maximization problems are solved with PGD. After computing $\hat{X}$ and $\hat{X}_{face}$ for each face in X, paste $\hat{X}_{face}$ into $\hat{X}$ where the corresponding face was originally positioned in the original image X. Do this for each face in X. The result is the output of the attack.

4 Results

In this section we give a qualitative overview of our attacks.

4.1 Effectiveness on attacked models

Here we aim to investigate the impact of our attack on the target model.

Setup We run our experiments on the e4e StyleGAN encoder [7] and the SD 1.5 encoder [23].

Given an input image, we apply the three attacks described in the previous section, where for the encoder attacks we use l_1 norm. Subsequently, we perform GAN inversions on these perturbed images and generate images using the LDM, guided by a given prompt, for all three images. To ensure consistency, we employ the same seed for LDM image generation, thus ensuring that any variations arise solely as a consequence of our attack.

For the joint encoder attacks we pre-compute an α which seems to have an effect on both models. Our experiments suggest that the E4E [42] StyleGAN model is more vulnerable to encoder attacks, so we take α - the weight on StyleGAN to be small. Qualitative results for varying α and ϵ are shown in Figures 2, 3 and 4.

As evident from the results, convex-linear distance functions exhibit subpar performance. Although they are effective in perturbing the LDM, their impact on StyleGAN inversion remains limited.

In contrast, the differential attack surpasses the performance of convex-linear distance functions. Nevertheless, it is evident that the most effective attack strategy is the adaptive joint encoder attack. This approach not only significantly disrupts StyleGAN inversion, even when using small noise levels (e.g., $\epsilon = 0.03$ as demonstrated in Figure 2), but also consistently disrupts LDM image generation at a level comparable to other attack methods.

4.2 Transferability

Now, we seek to determine whether our attacks can propagate their effects onto various StyleGAN and LDM models that were not originally the targets of the attack.

Setup Similar to our previous setup, we apply our three attacks to an input image where we use l_1 norm for the encoder attacks. Subsequently, we conduct GAN inversions using the three StyleGAN encoders E4E, PSP, and ReSTYLE [4, 31, 42], and attempt image editing using StyleCLIP [26] given a prompt. Additionally, we generate images using various LDMs, such as SD1.5[33], SDXL [27], SD2.1, Dreamlike and Openjourney, guided by prompts. As before, we use the

same input seed for the LDMs to ensure that any difference in the results arises as a result of our attack. The results are presented in Figure5.

Indeed, it is evident that our attacks introduce a distortion effect to the background of the image, affecting both LDMs and images generated using StyleGAN/StyleCLIP. Furthermore, all of our attacks introduce noticeable noise to the facial region in the image. In this context, it's worth mentioning that there isn't a clear-cut winner for the most effective attack strategy.

5 Discussion

In the previous sections, we have presented a method for protecting images against AI-image manipulation by leveraging the encoding process of StyleGAN and LDM's. Our approach focuses on safeguarding facial properties and surrounding areas from unwanted modifications while maintaining their appearance. To generate adversarial examples [18, 46, 47], we presented three possible attacks. However, our methods have some limitations that need to be addressed:

- **Resource-intensive execution.** The process of generating adversarial examples using PGD can be computationally and memory intensive, making it expensive to execute.
- **Limited resilience against various transformation techniques.** Our approaches may not be robust enough against different image transformation image and noise purification methods, which could undermine their effectiveness. A more comprehensive solution would require the use of advanced techniques for adversarial perturbation resistance [6, 8, 38].
- **Insufficient effectiveness against some models.** Despite targeting popular StyleGAN and diffusion models, our results show limited effectiveness against some models currently in use and there is no guarantee that they will be effective against those that may emerge in the future.

To overcome these limitations, a global policy should be established that requires adherence from organizations developing AI-image models and image-editing software, the open-source community, developers, data-hosting platforms, and more. This policy could involve adding metadata or a perturbation for a specific feature to images, indicating they should not be manipulated or trained by such models. Although users may be able to remove it with sufficient skill and effort, this would create an initial barrier for malicious actors, making it more difficult for them to manipulate images. The global policy could also include implementations of technical solutions that identify AI-manipulated images and tools to help recognize them.

Promoting responsible AI development and usage is crucial in addressing the challenges posed by AI image manipulation and AI-generated content as a whole. Encouraging researchers, developers, and organizations to adhere to ethical guidelines and best practices can help minimize potential harm caused by malicious use of AI-image models while fostering trust in these technologies. While this policy would be ideal, we must acknowledge the reality that such policies may take time to implement, and even then, our methods can offer an additional level of protection. In the meantime, our methods provide an effective solution to protect images from AI-driven manipulation.

6 Conclusion

Our study presents an innovative approach to protect images from AI-driven image manipulation using generative models such as StyleGAN and Latent Diffusion Models (LDMs). By combining StyleGAN Encoder's expertise in facial editing with LDMs Encoders' proficiency in editing the surroundings, our method provides a multifaceted defense against manipulation.

Our approach involves generating adversarial examples that neutralize the manipulation capabilities of these models through small, imperceptible perturbations. This process helps immunize images against unwanted modifications while maintaining their appearance and integrity. In the current era, where deepfakes and other manipulated content raise concerns about trustworthiness in online content, our research provides a valuable contribution towards developing effective countermeasures against AI-driven image manipulation.

Although our method has limitations that need to be addressed through global policy initiatives, responsible AI development, and user awareness of the potential risks associated with AI-generated content manipulation, our approach is crucial for protecting images in our digital era. It can also serve as an additional layer of protection alongside feature policies.

7 Contributions

Asaf - GCP setup and dependencies, code for testing the attack with various models and StyleCLIP (with Gal), and Flask server.

Gal - Discussing our algorithm's logic and implementation, creating user interface - website, connecting the website and the server, and code for testing the attack with various models and StyleCLIP (with Asaf).

Yotam - Implementation of the attack script, fine tuning of the attacks for different α and ϵ values, and all experiments in the report section.

References

[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4432–4441, 2019.

[2] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan++: How to edit the embedded images? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8296–8305, 2020.

[3] Naveed Akhtar and Ajmal Mian. Threat of adversarial attacks on deep learning in computer vision: A survey. *Ieee Access*, 6:14410–14430, 2018.

[4] Yuval Alaluf, Or Patashnik, and Daniel Cohen-Or. Restyle: A residual-based stylegan encoder via iterative refinement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6711–6720, 2021.

[5] Atif Ali, Khushboo Farid Khan Ghouri, Hina Naseem, Tariq Rahim Soomro, Wathiq Mansoor, and Alaa M Momani. Battle of deep fakes: Artificial intelligence set to become a major threat to the individual and national security. In *2022 International Conference on Cyber Resilience (ICCR)*, pages 1–5. IEEE, 2022.

[6] Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Synthesizing robust adversarial examples. In *International conference on machine learning*, pages 284–293. PMLR, 2018.

[7] Andreas Blattmann, Robin Rombach, Kaan Oktay, Jonas Müller, and Björn Ommer. Semi-parametric neural image synthesis. *Advances in Neural Information Processing Systems*, 11, 2022.

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

[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*, pages 8188–8197, 2020.

[10] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

[11] Zigang Fang, Yu Yang, Jialin Lin, and Rui Zhan. Adversarial attacks for multi target image translation networks. In *2020 IEEE International Conference on Progress in Informatics and Computing (PIC)*, pages 179–184. IEEE, 2020.

[12] Rinon Gal, Or Patashnik, Haggai Maron, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. Stylegan-nada: Clip-guided domain adaptation of image generators. *ACM Transactions on Graphics (TOG)*, 41(4):1–13, 2022.

[13] Ian Goodfellow, 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.

[14] Jiatao Gu, Lingjie Liu, Peng Wang, and Christian Theobalt. Stylenerf: A style-based 3d-aware generator for high-resolution image synthesis. *arXiv preprint arXiv:2110.08985*, 2021.

[15] Jinjin Gu, Yujun Shen, and Bolei Zhou. Image processing using multi-code gan prior. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3012–3021, 2020.

[16] A. Hertz, R. Mokady, and Y. Nitzan. Protecting against stylegan-based image manipulation. Technical report, Tel Aviv University, 2023.

[17] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[18] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. *Advances in neural information processing systems*, 32, 2019.

[19] Stamatis Karnouskos. Artificial intelligence in digital media: The era of deepfakes. *IEEE Transactions on Technology and Society*, 1(3):138–147, 2020.

[20] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

[21] 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.

[22] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020.

[23] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[24] Bahar Uddin Mahmud and Afsana Sharmin. Deep insights of deepfake technology: A review. *arXiv preprint arXiv:2105.00192*, 2021.

[25] Momina Masood, Mariam Nawaz, Khalid Mahmood Malik, Ali Javed, Aun Irtaza, and Hafiz Malik. Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward. *Applied intelligence*, 53(4):3974–4026, 2023.

[26] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2085–2094, 2021.

[27] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

[28] 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*, pages 8748–8763. PMLR, 2021.

[29] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

[30] 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*, pages 8821–8831. PMLR, 2021.

[31] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2287–2296, 2021.

[32] Daniel Roich, Ron Mokady, Amit H Bermano, and Daniel Cohen-Or. Pivotal tuning for latent-based editing of real images. *ACM Transactions on graphics (TOG)*, 42(1):1–13, 2022.

[33] 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*, pages 10684–10695, 2022.

[34] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022.

[35] Hadi Salman, Alaa Khaddaj, Guillaume Leclerc, Andrew Ilyas, and Aleksander Madry. Raising the cost of malicious ai-powered image editing. *arXiv preprint arXiv:2302.06588*, 2023.

[36] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9243–9252, 2020.

[37] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages

2256–2265. PMLR, 2015.

[38] Dawn Song, Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Florian Tramer, Atul Prakash, and Tadayoshi Kohno. Physical adversarial examples for object detectors. In *12th USENIX workshop on offensive technologies (WOOT 18)*, 2018.

[39] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[40] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[41] Ayush Tewari, Mohamed Elgharib, Gaurav Bharaj, Florian Bernard, Hans-Peter Seidel, Patrick Pérez, Michael Zollhofer, and Christian Theobalt. Stylerig: Rigging stylegan for 3d control over portrait images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6142–6151, 2020.

[42] Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for stylegan image manipulation. *ACM Transactions on Graphics (TOG)*, 40(4):1–14, 2021.

[43] Su Wang, Chitwan Saharia, Ceslee Montgomery, Jordi Pont-Tuset, Shai Noy, Stefano Pellegrini, Yasumasa Onoe, Sarah Laszlo, David J Fleet, Radu Soricut, et al. Imagen editor and editbench: Advancing and evaluating text-guided image inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18359–18369, 2023.

[44] Yangyang Xu, Yong Du, Wenpeng Xiao, Xuemiao Xu, and Shengfeng He. From continuity to editability: Inverting gans with consecutive images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13910–13918, 2021.

[45] Chin-Yuan Yeh, Hsi-Wen Chen, Hong-Han Shuai, De-Nian Yang, and Ming-Syan Chen. Attack as the best defense: Nullifying image-to-image translation gans via limit-aware adversarial attack. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16188–16197, 2021.

[46] Xiaoyong Yuan, Pan He, Qile Zhu, and Xiaolin Li. Adversarial examples: Attacks and defenses for deep learning. *IEEE transactions on neural networks and learning systems*, 30(9):2805–2824, 2019.

[47] Jiliang Zhang and Chen Li. Adversarial examples: Opportunities and challenges. *IEEE transactions on neural networks and learning systems*, 31(7):2578–2593, 2019.

Figure 2. Evaluation of our attacks with $\alpha = 0.05125$ and varying ϵ. Prompt for LDM is "a woman smiling". Zoom-in recommended.

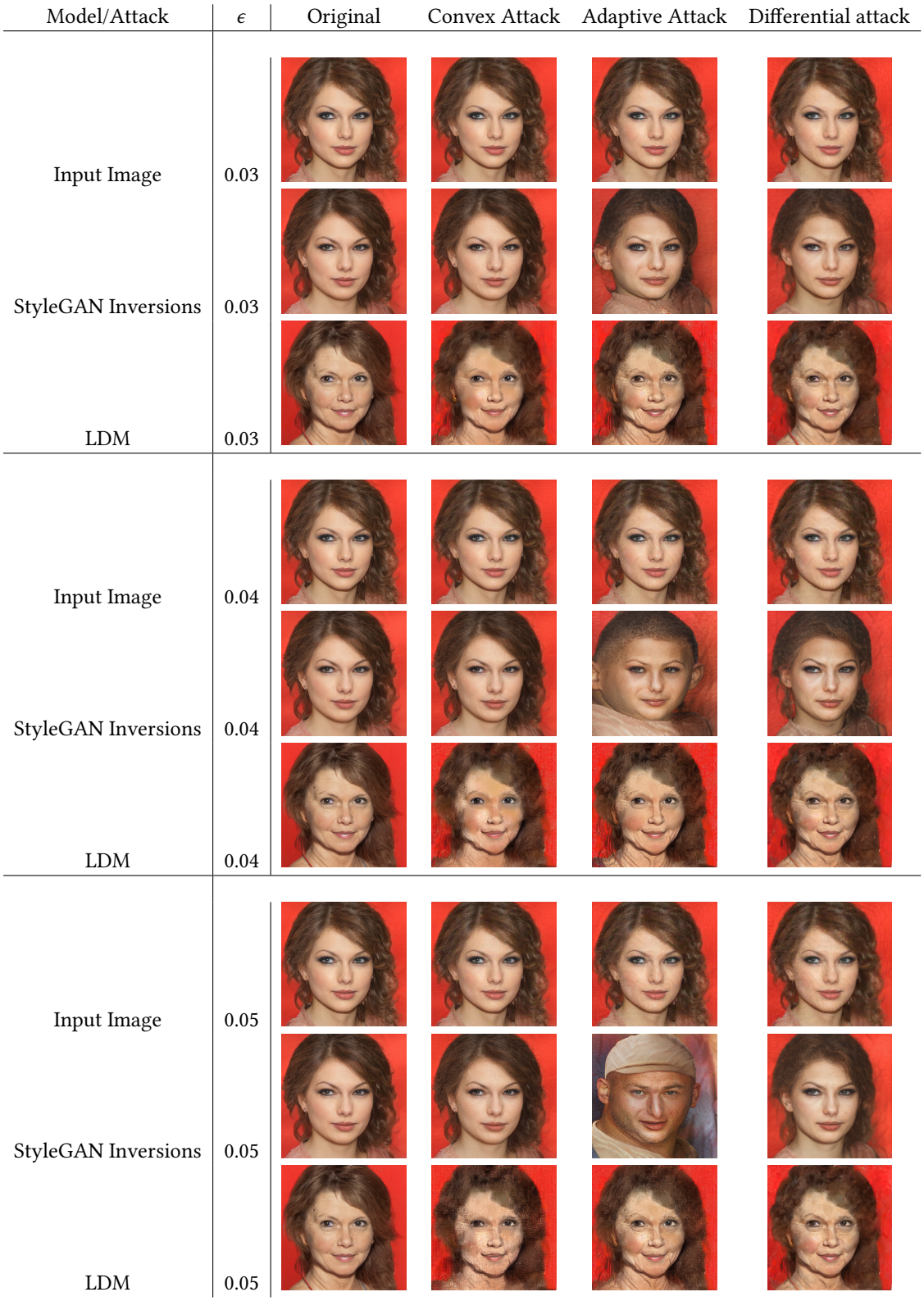


Figure 3. Evaluation of our attacks with $\alpha = 0.15$ and varying ϵ. Prompt for LDM is "a woman smiling". Zoom-in recommended.

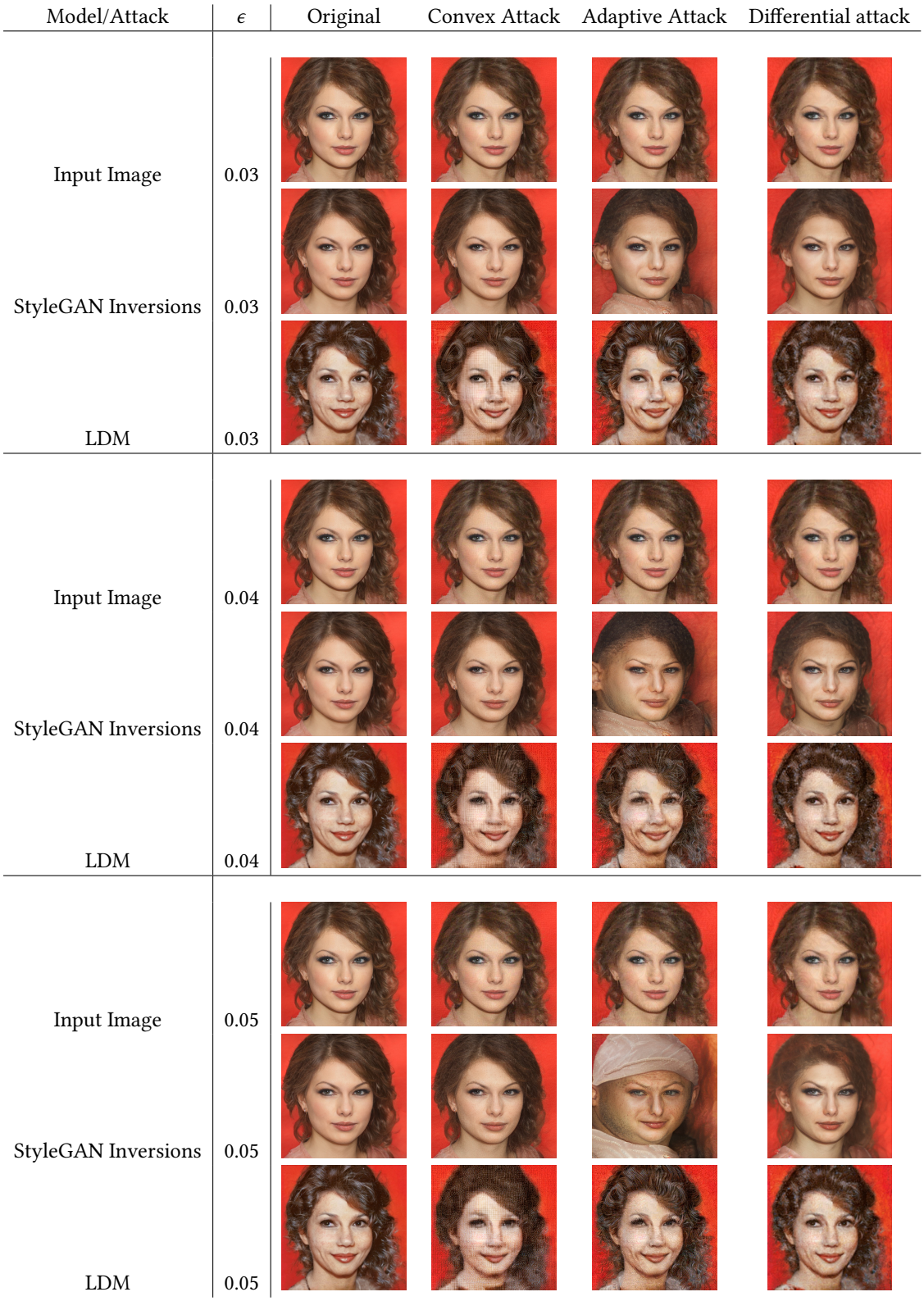


Figure 4. Evaluation of our attacks with $\alpha = 0.3$ and varying ϵ. Prompt for LDM is "a woman smiling". Zoom-in recommended.

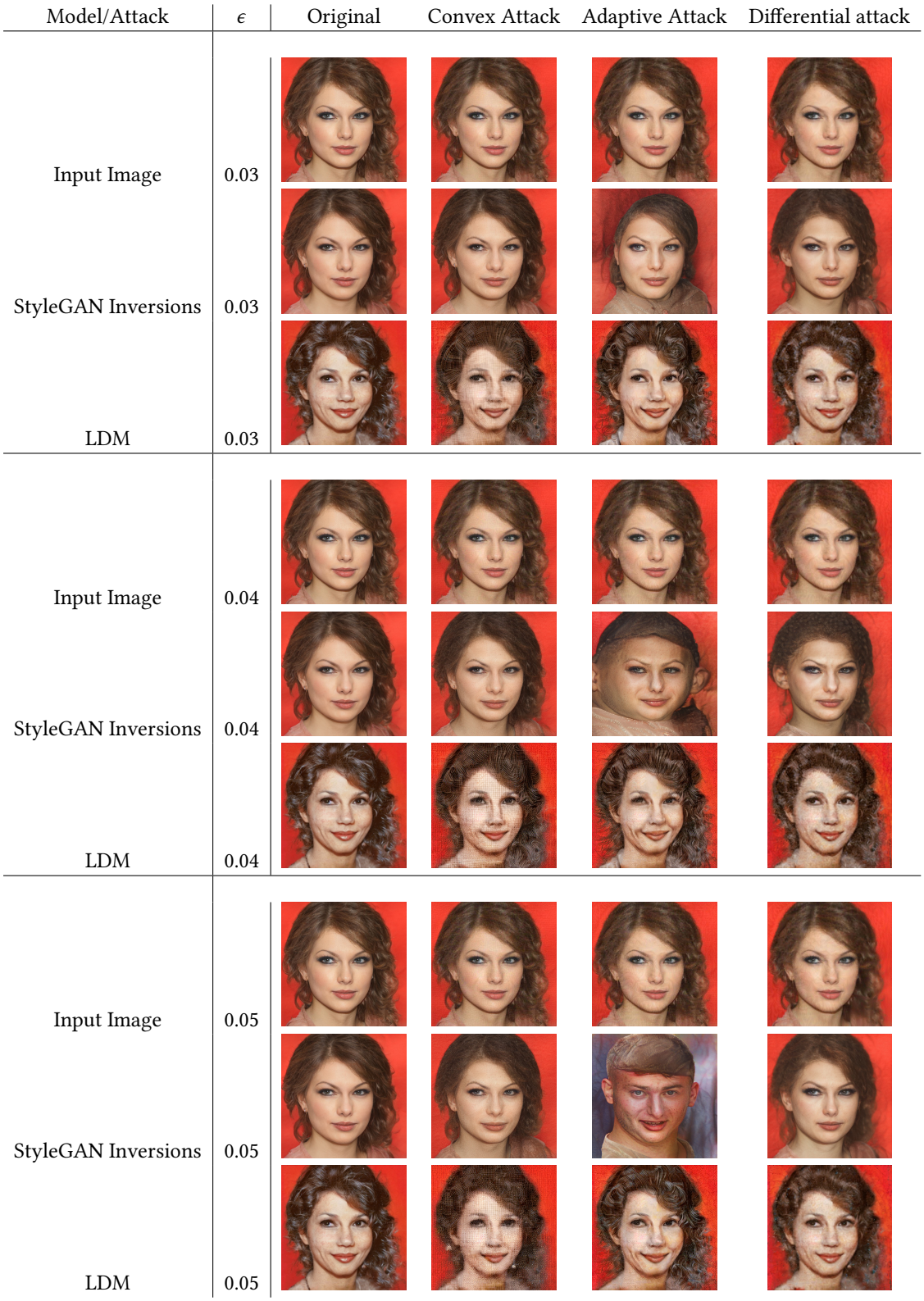


Figure 5. Qualitative evaluation of our attacks with $\alpha = 0.1$ and $\epsilon = 0.05$. Prompt for LDMs is "a man in a park with trees at day" and the prompt for StyleCLIP is "man with blond hair ". Zoom-in recommended.