Exploring Adversarial Perturbations to Diffusion Models

Tristan Saidi
Columbia University
New York, NY 10027
tls2160@columbia.edu

Leon Zhou
Columbia University
New York, NY 10027
leon.zhou@columbia.edu

Abstract

Diffusion models are notable for their ability to model extremely complicated probability distributions. As a result, many prominent results in recent image synthesis methods stem from a diffusion model backbone. Understanding models with such expressive capability and understanding their vulnerabilities will always be of interest, and recent work on deep diffusion models has made progress in that direction [20]. With that said, thorough understanding of the latent noise stages in denoising diffusion models is still an open research topic. It is unclear how much perturbations to the denoising process effect the quality of samples, and whether or not said perturbations can be constructed to steer the model maliciously. To that end, we explore the robustness of Denoising Diffusion Probabilistic Models [7] to adversarial perturbations. We explore this with vanilla diffusion models and classifier-free guided diffusion models. The code for our models and experiments is available at `https://github.com/TristanSaidi/AdversarialDiffusion`

1 Introduction

Diffusion models are a family of deep generative models used primarily for image synthesis. They were first introduced by Sohl-Dickstein et al. [17], but gained prominence with Ho et al.'s [7] paper. Ho et al. showed that high-quality image samples can be achieved via deep denoising diffusion. This is done by having two parameterized Markov chains; one to iteratively inject noise into an image with a schedule, and the other chain learns to restore the original image. Ho et al. also [7] proposed a new parameterization that resembles denoising score matching and has equivalence to Langevin dynamics, which improved variational bounds and results. Variants like Score-based models [20] construct an stochastic differential equation (SDE) to disturb the image into a distribution, while the reverse-time SDE repairs the image. Work have also shown that diffusion is also possible when the transformations are deterministic [1].

While research has been done on using diffusion models to generate attacks [3] or defenses [14] against other classification models, there is a lack of study into attacking the diffusion models themselves. The latent space of these models not well understood, and it is hoped that by successfully perturbing them, it will shed insight into their inherent structures.

We propose analysis of the latent space for diffusion models by adversarially perturbing the generated image at various stages of the denoising process.

2 Related Work

Diffusion models have quickly become state-of-the-art for generating high resolution images [7, 18], beating out GANs [4]. Various modifications of the standard diffusion have been proposed,

Preprint. Under review.

most notably with score-based generative modeling [20], and classifier-free guidance [8]. They have been applied to the inverse-problem in medical imaging [19], domains with low-density data [16], and classification [13].

Adversarial Attacks were introduced by Szegedy et al. [21] when they discovered that appending gradient-based noise to images could fool classification models. While it did not gain prominence at first, it has grown into a larger issue as these machine learning sees more use. Over time, these attacks have been automated [23], and applied to black-box models [9]. These attacks are not only limited to vision, as it has crossed domains to Natural Language Processing [10].

Latent Space of diffusion models remains a less studied area. It is believed that the intermediate latent spaces hold semantic meaning [12], and experiments have been done by looking at the feature map of the U-Net [2]. In addition, Tumanyan et al. [22] demonstrate that features from the U-Net can be injected into the generation process of a target to shape the results of the diffusion process. In addition, there are approaches to understand it mathematically through Riemannian Geometry [15], bringing techniques that have been used previously to analyze GANs [25]. That being said, there exists little research in exposing the vulnerability of diffusion models to latent perturbations.

3 Model Definition

3.1 Denoising Diffusion Probabilistic Models

The Denoising Diffusion Probabilistic Model defines a joint distribution over observed $\mathbf{x}_0$ and hidden variables $\mathbf{x}_{1:T}$.

$$\begin{gathered} p_\theta\left(\mathbf{x}_{0:T}\right) := p\left(\mathbf{x}_T\right) \prod_{t=1}^{T} p_\theta\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right) \\ p_\theta\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right) := \mathcal{N}\left(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta\left(\mathbf{x}_t, t\right), \boldsymbol{\Sigma}_\theta\left(\mathbf{x}_t, t\right)\right) \end{gathered} \tag{1}$$

Unlike most other probabilistic models, the approximate posterior is defined to be a sequence of known noising steps applied to the original data distribution:

$$\begin{gathered} q\left(\mathbf{x}_{1:T} \mid \mathbf{x}_0\right) := \prod_{t=1}^{T} q\left(\mathbf{x}_t \mid \mathbf{x}_{t-1}\right) \\ q\left(\mathbf{x}_t \mid \mathbf{x}_{t-1}\right) := \mathcal{N}\left(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t \mathbf{I}\right) \end{gathered} \tag{2}$$

Learning the conditional Gaussian transitions $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$ involves optimizing the evidence lower bound (ELBO),

$$\mathbb{E}[-\log p_\theta(\mathbf{x}_0)] \leq \mathbb{E}_q \left[-\log p(\mathbf{x}_T) - \sum_{t \geq 1} \log \frac{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}{q(\mathbf{x}_t|\mathbf{x}_{t-1})} \right] \tag{3}$$

Ho et al. [7] show this objective can be rewritten as

$$\mathcal{L}(\theta) := \mathbb{E}_q \left[D_{KL}(q(\mathbf{x}_T|\mathbf{x}_0)||p(\mathbf{x}_T)) + \sum_{t>1} D_{KL}(q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)||p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)) - \log p_\theta(\mathbf{x}_0|\mathbf{x}_1) \right] \tag{4}$$

3.2 Class-conditional Diffusion

Augmenting vanilla diffusion models to include class-conditioning can be achieved in various ways. One such method involves using a classifier to guide the diffusion process. Concretely, a classifier is trained to predict the distribution over labels from noisy images, $p_\phi(y|\mathbf{x}_t)$. Dhariwal et al. [5] show that conditional reverse process sampling can be approximated as a perturbed gaussian, where the perturbation is a function of the gradient of the classifier's predicted probability w.r.t. the noisy input $\nabla_{\mathbf{x}_t} \log p_\phi(y|\mathbf{x}_t)$.

While this method is effective in practice, it involves training a separate classifier on noisy images produced by the forward diffusion process. To avoid this our initial experiments utilize naive class-conditional diffusion - our reverse process network, p_θ, simply takes the class label as input. Therefore, our optimization involves taking a gradient descent step on the following expression (augmented from [7]):

$$\mathbb{E}_{\substack{(\mathbf{x}_0,y)\sim\mathcal{D}\\ \boldsymbol{\epsilon}\sim\mathcal{N}(0,\mathbf{I})}}\left[\nabla_\theta\left\|\boldsymbol{\epsilon}-\boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0+\boldsymbol{\epsilon}\sqrt{1-\bar{\alpha}_t},t,y)\right\|^2\right] \tag{5}$$

Reverse process sampling then involves a minor modification to 1. Namely, each of the learned gaussian transitions are conditioned on the desired class label:

$$p_\theta\left(\mathbf{x}_{t-1}\mid\mathbf{x}_t\right):=\mathcal{N}\Big(\mathbf{x}_{t-1};\boldsymbol{\mu}_\theta\left(\mathbf{x}_t,t,y\right),\sigma_t\Big)$$

3.3 Class-conditional Diffusion via Classifier-Free Guidance

Ho et al. [8] derive an effective classifier-free guidance formulation for diffusion models that achieve comparable performance to state-of-the-art classifier-based counterparts. They train a *single* neural network to model $p_\theta(\mathbf{x}_0)$ and $p_\theta(\mathbf{x}_0|y)$. The former is parameterized by $\boldsymbol{\epsilon}_\theta(\mathbf{x}_t)$ while the latter is parameterized by $\boldsymbol{\epsilon}_\theta(\mathbf{x}_t,y)$. To encapsulate both models with a single parameterization, $\boldsymbol{\epsilon}_\theta(\mathbf{x}_t)$ is modelled as $\boldsymbol{\epsilon}_\theta(\mathbf{x}_t,y=\varnothing)$, where $\varnothing$ is a null token. Score estimates during denoising simply turn into a weighted combination of the class-conditional and unconditional score, where the weighting is controlled by a guidance parameter w:

$$\tilde{\boldsymbol{\epsilon}}_\theta(\mathbf{x}_t,y)=(1+w)\boldsymbol{\epsilon}_\theta(\mathbf{x}_t,y)-w\boldsymbol{\epsilon}_\theta(\mathbf{x}_t)$$

Optimization of θ in this classifier-free guidance formulation involves following the same gradient described in 5. The only difference stems from random alternations between training conditionally and unconditionally, specified by the hyperparameter $p_{\text{uncond}}\in[0,1]$. Training the unconditional variant simply involves setting $y=\varnothing$.

4 Adversarial Perturbations

To explore the robustness of state-of-the-art diffusion models to perturbations, we devised three different attacks that we applied to the models described in 3. Our goal was to explore the type of perturbations required to steer image-generating diffusion models away from the desired sampling distribution. As mentioned, we pursued this primarily in the context of class conditional diffusion models - We also looked to gain additional insight into the intermediate noising stages, hoping to find correlations between perturbation effectiveness and the denoising stage at which the perturbation was injected.

For all of the following experiments, we use a linear variance schedule as well as $T=300$ denoising iterations.

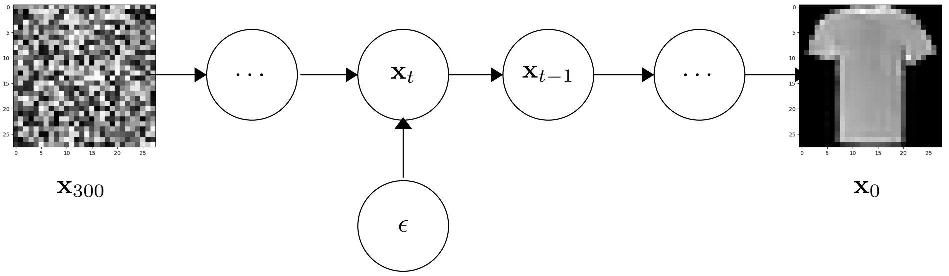

Figure 1: Visualization of our perturbation scheme. For a fixed denoising stage t we add the perturbation ϵ to the partially denoised sample, and then continue the sampling process

4.1 Gaussian Noise

We began with introducing simple gaussian perturbations at a range of denoising steps for the three models described in 3. For a desired class label y, we sample a partially denoised intermediate representation $\mathbf{x}_t \sim p_\theta(\mathbf{x}_t|\mathbf{x}_T, y)$. We then perturb it with $\epsilon \sim \mathcal{N}(0, \alpha\mathbf{I})$, and finish the denoising process by sampling $\mathbf{x}_{0,\text{pert}} \sim p_\theta(\mathbf{x}_0|\mathbf{x}_t + \epsilon, y)$.

Figure 2 shows perturbed and unperturbed samples for each of the three diffusion models mentioned. More extensive figures that vary class label and denoising stage t can be seen in 3 and 4.

These experiments highlight the unsurprising robustness of diffusion models to naive perturbations. We therefore turned our attention to more targeted and adversarial forms of perturbations.

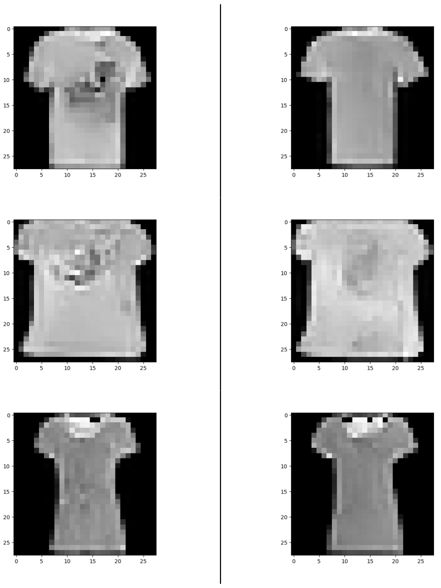

Figure 2: Unperturbed (right) and gaussian perturbed (left) samples from unconditional (top), class-conditional (middle) and classifier-free guided (bottom) diffusion models. All perturbations occur at denoising stage $t = 50$.

4.2 One-step Gradient Method

As a more targeted attack, we sought to introduce perturbations that pull class conditional samples towards a target image from an undesired class. Concretely, given a desired class label y and an adversarial class label y_{adv} we generated a perturbation ϵ that decreases the expected l_2 loss between the sample $\mathbf{x}_0 \sim p_\theta(\mathbf{x}_0|\mathbf{x}_t + \epsilon, y)$ and our adversarial target $\mathbf{x}_{\text{adv}} \sim p_\theta(\mathbf{x}_{\text{adv}}|\mathbf{x}_t, y_{\text{adv}})$. Naturally, we can backpropate said l_2 loss back to $\mathbf{x}_t$ to find the adversarial perturbation ϵ that best pulls our sample towards the target.

$$\epsilon \approx -\nabla_{\mathbf{x}_t} \left\|\mathbf{x}_0 - \mathbf{x}_{\text{adv}}\right\|^2$$

$$\mathbf{x}_0 \sim p_\theta(\mathbf{x}_0|\mathbf{x}_t, y)$$

Backpropagation through reverse-process sampling is achieved via the reparameterization trick , as we parameterize the learned transition $p_\theta(\mathbf{x}_t|\mathbf{x}_{t-1})$ as a multivariate gaussian [11]. Thus, reverse-process sampling can be written as

$$\mathbf{x}_{t-1} = \boldsymbol{\mu}_\theta(\mathbf{x}_t, y) + \sigma_t\mathbf{z}$$

where $\mathbf{z} \sim \mathcal{N}(0, \mathbf{I})$, allowing us to backpropagate through the learned $\boldsymbol{\mu}_\theta$.

Figure 7 in the Appendix illustrates the effect of the described perturbation to both the class-conditional and classifier-free guided diffusion models. For the former, the perturbation induces significant artifacts in the sampled image - while they don't directly resemble the adversarial target, the perturbations are noticeably more effective at producing bizarre image features than gaussian perturbations.

For the guided diffusion model however, the perturbation had no noticeable effect on the quality of the samples, especially when the perturbations were introduced early in the denoising process. Perhaps this is unsurprising - classifier-free guidance forces the denoising network $\boldsymbol{\epsilon}_\theta$ to model both the conditional *and* unconditional score functions. Since sampling involves a weighted combination of said scores, a latent perturbation towards another class doesn't conflict with the unconditional score function!

4.3 Several-step gradient Method

Our experiments indicated that single-step gradient based perturbations were largely ineffective in perturbing sampled images towards adversarial targets. As a result, we extended our efforts to performing iterative gradient descent with the update rule below in an attempt to find an effective perturbation:

$$\mathbf{x}_t \leftarrow \mathbf{x}_t - \nabla_{\mathbf{x}_t} \left\| \mathbf{x}_0 - \mathbf{x}_{\text{adv}} \right\|^2$$

$$\mathbf{x}_0 \sim p_\theta(\mathbf{x}_0 | \mathbf{x}_t, y)$$

Figures 8 and 9 indicate that such a formulation is effective. Over several steps of gradient descent, the noise gets optimized to produce an image closer to the adversary. For both the class-conditional and the classifier-free guided diffusion models, the learned perturbations applied *late* in the denoising process appear to perturb the sample towards the adversarial target.

5 Conclusion

Diffusion models proved to be more resilient to adversarial perturbations than expected. For the Gaussian noise case, since the model is trained via denoising Gaussian noise, the perturbation had no real effects on the type of image generated. Perturbations in the later denoising stages did lead to some texture changes in the center, suggesting that the final few layers of diffusion deal with finer details.

One-step gradient descent techniques are usually enough to adversarially affect image classification models [6]. In our case, for both the class conditional and classifier-free guidance, the perturbation had little to no effect on pulling the sampled image towards the adversarial target.

That being said, our gradient-descent based approach ultimately enabled learning of effective adversarial perturbations against the current state-of-the-art class-conditional diffusion architecture. The effectiveness of such perturbations could open the door to downstream attacks of deployed diffusion models, and further investigation into these gradient-based attacks is certainly of interest.

References

[1] Arpit Bansal, Eitan Borgnia, Hong-Min Chu, Jie S. Li, Hamid Kazemi, Furong Huang, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Cold diffusion: Inverting arbitrary image transforms without noise, 2022.

[2] Dmitry Baranchuk, Ivan Rubachev, Andrey Voynov, Valentin Khrulkov, and Artem Babenko. Label-efficient semantic segmentation with diffusion models, 2022.

[3] Xuelong Dai, Kaisheng Liang, and Bin Xiao. Advdiff: Generating unrestricted adversarial examples using diffusion models, 2023.

[4] Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis, 2021.

[5] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 8780–8794. Curran Associates, Inc., 2021.

[6] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples, 2015.

[7] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *CoRR*, abs/2006.11239, 2020.

[8] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021.

[9] Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. Black-box adversarial attacks with limited queries and information, 2018.

[10] Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. Is bert really robust? a strong baseline for natural language attack on text classification and entailment, 2020.

[11] Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2022.

[12] Mingi Kwon, Jaeseok Jeong, and Youngjung Uh. Diffusion models already have a semantic latent space, 2023.

[13] Alexander C. Li, Mihir Prabhudesai, Shivam Duggal, Ellis Brown, and Deepak Pathak. Your diffusion model is secretly a zero-shot classifier, 2023.

[14] Weili Nie, Brandon Guo, Yujia Huang, Chaowei Xiao, Arash Vahdat, and Anima Anandkumar. Diffusion models for adversarial purification, 2022.

[15] Yong-Hyun Park, Mingi Kwon, Jaewoong Choi, Junghyo Jo, and Youngjung Uh. Understanding the latent space of diffusion models through the lens of riemannian geometry, 2023.

[16] Vikash Sehwag, Caner Hazirbas, Albert Gordo, Firat Ozgenel, and Cristian Canton Ferrer. Generating high fidelity data from low-density regions using diffusion models, 2022.

[17] Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics, 2015.

[18] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models, 2022.

[19] Yang Song, Liyue Shen, Lei Xing, and Stefano Ermon. Solving inverse problems in medical imaging with score-based generative models, 2022.

[20] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

[21] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks, 2014.

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

[23] Chaowei Xiao, Bo Li, Jun-Yan Zhu, Warren He, Mingyan Liu, and Dawn Song. Generating adversarial examples with adversarial networks, 2019.

[24] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, 2017.

[25] Jiapeng Zhu, Ruili Feng, Yujun Shen, Deli Zhao, Zhengjun Zha, Jingren Zhou, and Qifeng Chen. Low-rank subspaces in gans, 2021.

Appendix

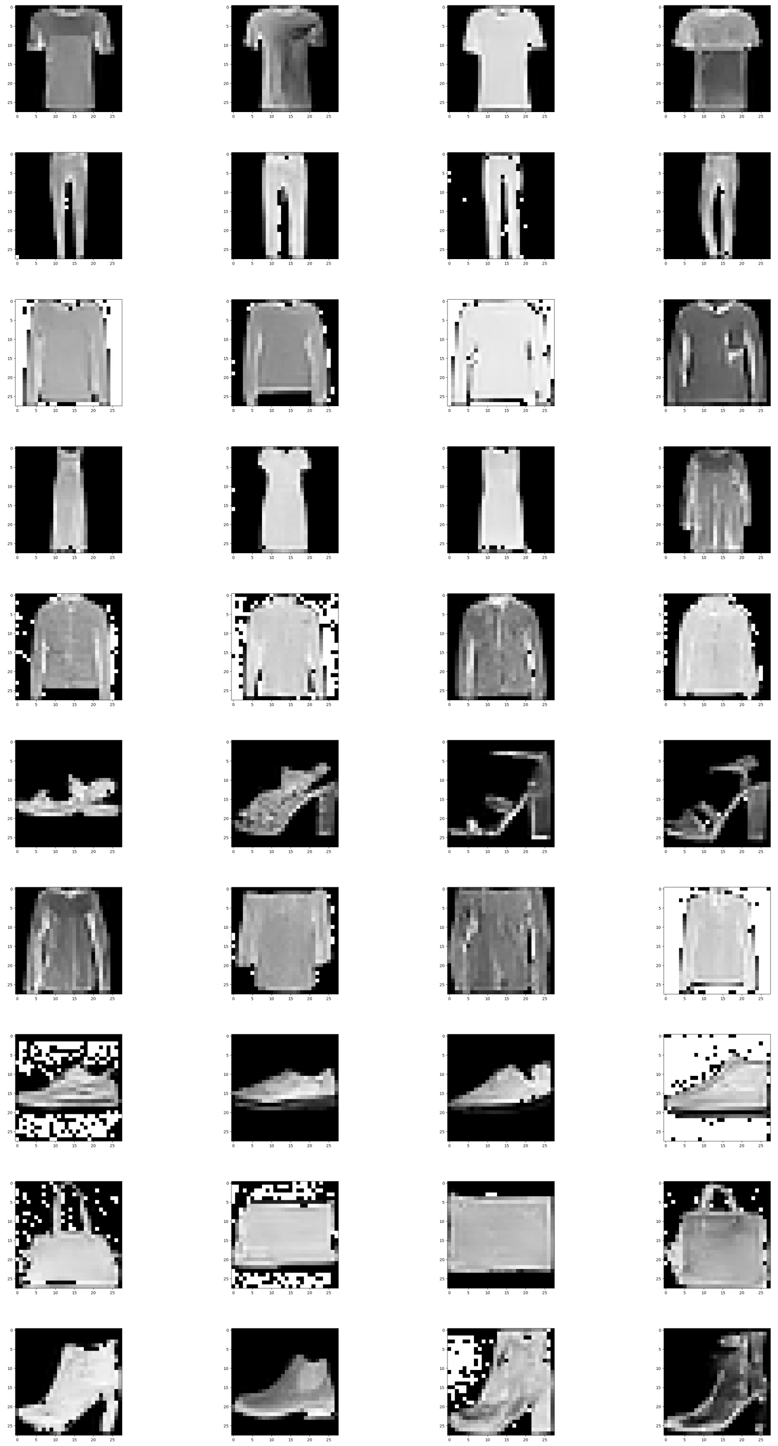

Figure 3: Gaussian perturbed images generated from the classifier-free guided diffusion model. From left to right represents perturbations at stages $t \in \{50, 100, 150, 200\}$

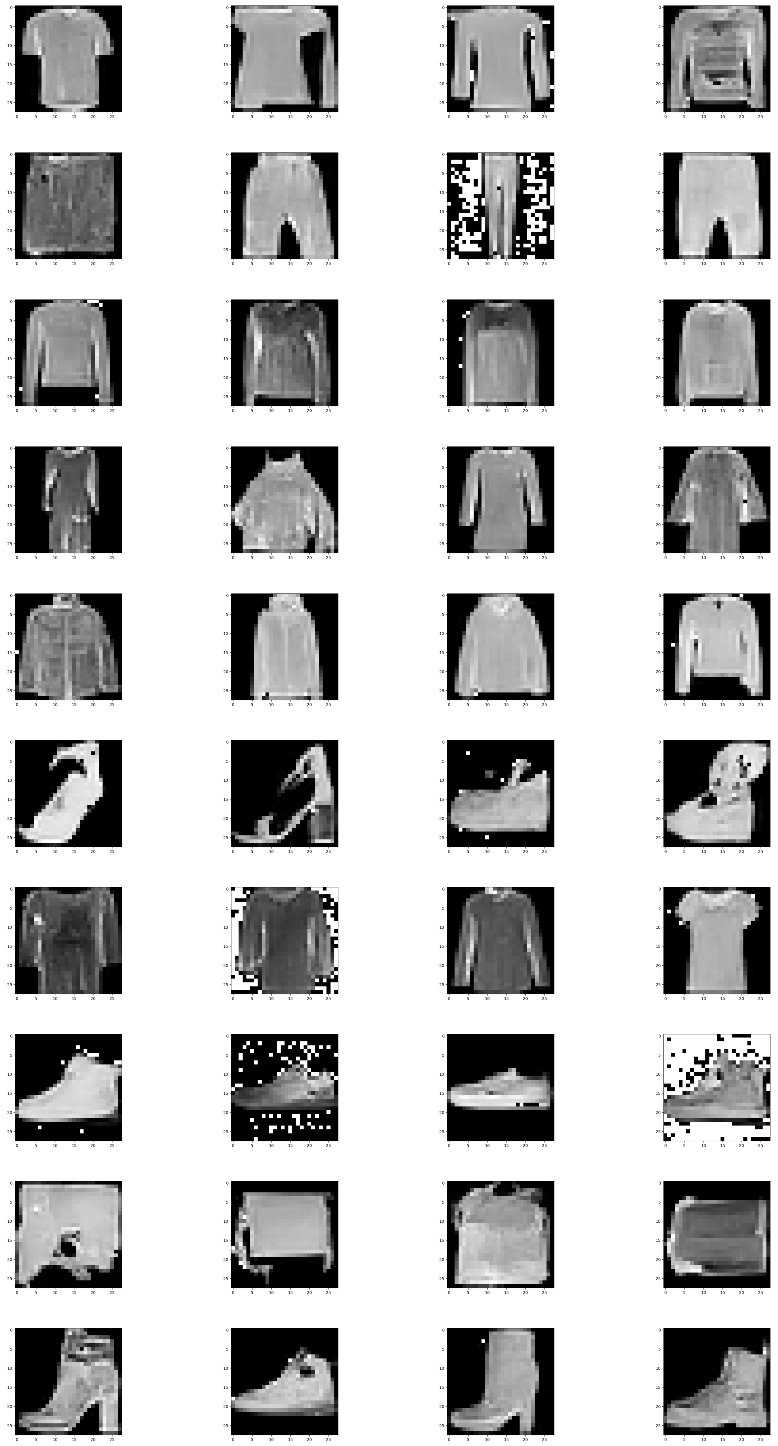

Figure 4: Gaussian perturbed images generated from the class-conditional diffusion model. From left to right represents perturbations at stages $t \in \{50, 100, 150, 200\}$

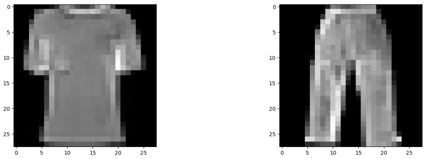

Figure 5: Unperturbed sample (left) and adversarial target (right)

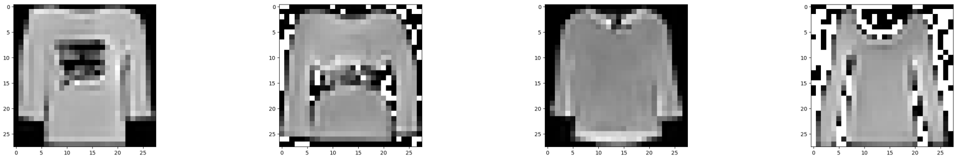

Figure 6: Single-step gradient perturbed samples from the class-conditional diffusion model. Latent perturbations injected at $t = 50, 100, 150, 200$ (left to right)

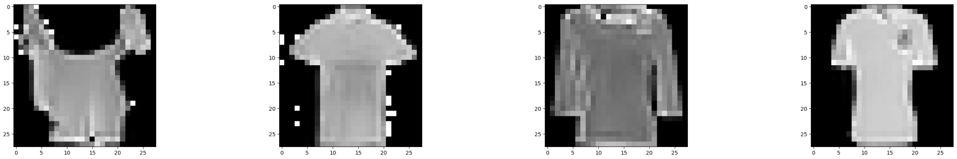

Figure 7: Single-step gradient perturbed samples from the classsifier-free guided diffusion model. Latent perturbations injected at $t = 50, 100, 150, 200$ (left to right)

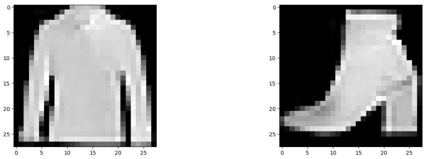

Figure 8: Unperturbed sample (left) and adversarial target (right)

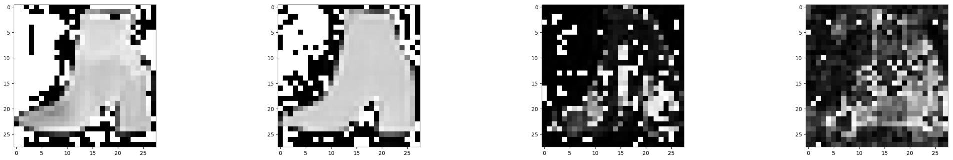

Figure 9: Several-step perturbed samples from the class-conditional diffusion model. Latent perturbations injected at t = 50, 100, 150, 200 (left to right)

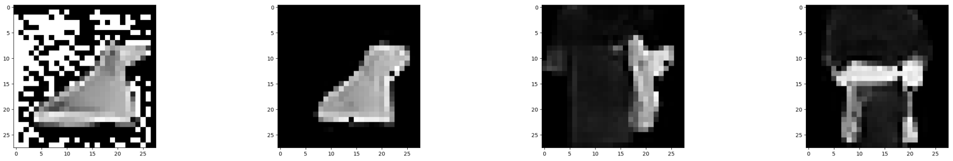

Figure 10: Several-step perturbed samples from the classier-free diffusion model. Latent perturbations injected at t = 50, 100, 150, 200 (left to right)