Disrupting Diffusion-based Inpainters with Semantic Digression

Geonho Son[*1], **Juhun Lee**[*1], and **Simon S. Woo**[1,2]

[1]Department of Artificial Intelligence
[2]Computer Science & Engineering Department
Sungkyunkwan University
{sohn1029, josejhlee, swoo}@g.skku.edu

Abstract

The fabrication of visual misinformation on the web and social media has increased exponentially with the advent of foundational text-to-image diffusion models. Namely, Stable Diffusion inpainters allow the synthesis of maliciously inpainted images of personal and private figures, and copyrighted contents, also known as deepfakes. To combat such generations, a disruption framework, namely Photoguard, has been proposed, where it adds adversarial noise to the context image to disrupt their inpainting synthesis. While their framework suggested a diffusion-friendly approach, the disruption is not sufficiently strong and it requires a significant amount of GPU and time to immunize the context image. In our work, we re-examine both the minimal and favorable conditions for a successful inpainting disruption, proposing DDD, a "**D**igression guided **D**iffusion **D**isruption" framework. First, we identify the most adversarially vulnerable diffusion timestep range with respect to the hidden space. Within this scope of noised manifold, we pose the problem as a semantic digression optimization. We maximize the distance between the inpainting instance's hidden states and a semantic-aware hidden state centroid, calibrated both by Monte Carlo sampling of hidden states and a discretely projected optimization in the token space. Effectively, our approach achieves stronger disruption and a higher success rate than Photoguard while lowering the GPU memory requirement, and speeding the optimization up to three times faster.

1 Introduction

The malicious editing of visual content is a long-standing ethical issue in the online community. Lately, this concern has only been significantly amplified with the integration of deep learning algorithms into the online community. Today, with recent advancements in the deep learning community, high-quality content crafting, known as Deepfake [Le *et al.*, 2024; Le and Woo, 2023; Lee *et al.*, 2022; Cho *et al.*, 2023; Hong *et al.*, 2024; Le *et al.*, 2023], has been refined, and the discernment against real content is harder than ever. In tandem with malicious editing, this technology has led to increasing cases of social chaos and misinformation.

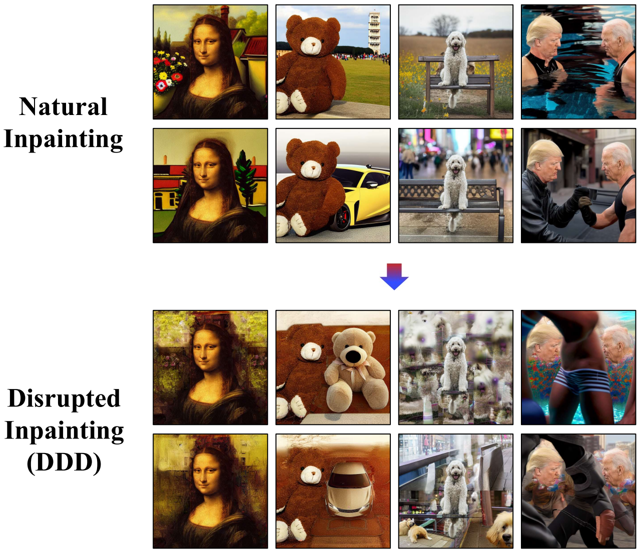

Figure 1: Our framework DDD optimizes adversarial perturbations for images that protect malicious users from editing the images without consent. Through various test images, our efforts demonstrate that our approach is sufficient to cover copyrighted images, pornographic abuse, and public figure editing scenarios. Ultimately, it digresses the representation of the context image away from its multimodal nucleus in expectation.

Recently, much interest has been directed to diffusion-based generative models [Ho *et al.*, 2020; Song *et al.*, 2020b]. One impactful model among them is the text-to-image generator called Stable Diffusion [Rombach *et al.*, 2022], which was trained on LAION, a large-scale captioned image dataset [Schuhmann *et al.*, 2022]. This type of magnitude of training and model size unlocked the generative power of "foundational" scale and generalizability to unseen and complex prompts. With further fine-tuning [Zhang *et al.*, 2023; Hu *et al.*, 2021; Wang *et al.*, 2023; Lin *et al.*, 2023; Xu *et al.*, 2023], powerful variants such as inpainting models came to being, which allow the user to input a context image and inpaint the remainder with text guidance, quickly becoming the commercial inpainting approach. Additionally, a major advantage of inpainting over image editing algorithms [Mokady *et al.*, 2023; Wallace *et al.*, 2023] is that edits are exclusive to the user-defined region.

However, this created a breach for adversaries, as these models are well-suited for malicious and unconsented image edits of personal and private figures, copyrighted content, etc. This poses a serious problem on the internet, since the weights of foundational generative models are public to all individuals [von Platen *et al.*, 2022]. To counter the production of such malicious content including deepfakes, researchers [Salman *et al.*, 2023; Ruiz *et al.*, 2020] have proposed ways to disrupt their generation by injecting adversarial noise into the image to disrupt or fool either the face synthesizer or other subnetworks necessary in the generation. While considerable advancement has been made to disrupt GAN-based deepfake generators, the rise of diffusion-based deepfake pleads for an equivalent countering disruption algorithm. Unfortunately, because of the heterogeneous and complex characteristics underlying diffusion models, in which the generation process is gradual and iterative, previously proposed Deepfake disruption approaches are not compatible with the diffusion class of models.

Photoguard [Salman *et al.*, 2023] has been proposed to address disruption for diffusion-based inpainters. It optimizes the context image so that it contains adversarial noise to disrupt the inpainting during inference time. While the effort of addressing the complications introduced by the diffusion process is noteworthy, it undergoes a significant computational overhead, amounting to 25GBs of VRAM and 19 min. of optimization. This computation cost is beyond the average consumer's budget. Furthermore, the disruption efficacy across different images, inpainting strengths, prompts is unstable.

Our work departs from the motivation that, while it is true that diffusion-based models introduce new technical challenges, we challenge the base assumptions that Photoguard considers to be necessary. In particular, to bypass the multiple feedforwards across the entire diffusion reverse process, we take into account the timestep range that, when perturbed with some adversarial signal, can cause the most amount of visual disparity, and harness the timestep constraint-free hidden states for our loss function.

Next, we leverage the untargeted attack approach in the generative setting. Specifically, we find a centroid of representations to digress away from, which effectively eliminates the burden of defining a target point and its optimization complications of a targeted attack. First, we search for hidden state samples centered around the representation of the context image and calibrated through a discretely projected optimization in the token space. This optimized text embedding cooperates in defining a representative centroid through Monte Carlo sampling. This semantic-aware centroid faithfully captures the user's input in all modalities. Ultimately, distancing away from this point results in an edit that completely disassociates from the context image. Our "Semantic **D**igression-guided **D**iffusion **D**isruption", DDD in short, is $3\times$ faster, requires less GBs of VRAM, and the disruption is more effective than the current SoTA, Photoguard, across various images. We provide extensive results and experiments to support the integrity of our framework. Our main contribution is summarized as follows:

- In this work, we tackle disrupting inpainting-based deepfake synthesis with context optimization. We identify the most vulnerable timestep with respect to the feature space in the diffusion reverse process. This finding paves a design for a timestep-agnostic loss function, effectively decoupling our framework from the diffusion process' time/memory complexity overhead.
- We align our optimization objective with semantic digression optimization. Formally, we approximate the notion of synthesis correctness with a semantic-aware hidden state centroid, calibrated by a discretely projected optimization in the token space, and digress semantically away from it.
- The proposed framework, DDD, significantly reduces GPU VRAM usage and running time while sustaining effective disruption levels. Our effort democratizes image protection from malicious editing, bringing the computation expenses down to the consumer-grade GPU budget regime.

2 Related Works

GAN-based deepfake generators [Choi *et al.*, 2018; Pumarola *et al.*, 2018] have been the major consideration in both the research community and industry due to their long legacy, fast sampling, and quality. One distinguished architecture is StarGAN [Choi *et al.*, 2018], trained to transfer the domain of the input image to cross-domains. And, GANimation [Pumarola *et al.*, 2018], a conditional generator, can generate faces according to the expression annotation.

To combat unconsented real image edits, Yeh et al. [Yeh *et al.*, 2020] pioneered the first attack on deep generative models, such as CycleGAN [Zhu *et al.*, 2017] and pix2pix [Isola *et al.*, 2017]. They also introduced the Nullifying Attack to invalidate the model's generation and the Distorting Attack to make the generated images blurry or distorted. Ruiz et al. [Ruiz *et al.*, 2020] synthesize adversarial noise to the input image of these image-to-image GANs, so that their outputs are disrupted. Meanwhile, Wang et al. [Wang *et al.*, 2022] designed a perturbation generator, ensuring that the disrupted image output is correctly classified as a fake image by deepfake detectors. Furthermore, Huang et al. [Huang *et al.*, 2021] introduced an iteratively trained surrogate model to provide feedback to the perturbation generator, enabling the suppression of facial manipulation. While these disrupters are effective, the expressivity of the deepfake images made with GAN-based models is limited by the annotation, quality of the dataset, and model scalability.

On the other hand, the recent text-to-image latent diffusion such as Stable Diffusion is trained on LAION [Schuhmann *et al.*, 2022], a large-scale dataset scraped from the web. Naturally, it can generalize to complex and unseen prompts, and synthesize with high coherence. In particular, the Stable Diffusion (SD) Inpainter model [Rombach *et al.*, 2022] is a finetuned Stable diffusion with masked context image conditioning and it can edit via inpainting the masked area. To disrupt such a potential deepfake generator, Photoguard takes a similar approach to previous disrupting algorithms in optimizing the adversarial noise in the context image. Their contribution and details will be provided in the next section.

3 Background

3.1 Adversarial Attacks

An adversarial attack is a method to generate adversarial examples to deceive the ML system [Goodfellow *et al.*, 2014]. In the perspective of discriminators, given an objective function $\mathcal{L}$, a projective function Π, and an image x its true label y, the PGD [Madry *et al.*, 2017] attack is performed as follows:

$$x^{t+1} = \Pi_{x+\mathcal{S}}(x^t - \alpha\ sign(\nabla_x \mathcal{L}(x, y_{target}))). \tag{1}$$

This iterative update algorithm helps in identifying local maxima that induce misclassification and it is the engine behind most disruption frameworks. Likewise, we adopt PGD in our framework to update and refine our disruptive perturbation.

3.2 Diffusion Models

Consider $x_t \in \mathbb{R}^{1\times3\times W\times H}$, where x_T is an isotropic Gaussian noise, x_0 is true data observation, and any x_t between these is defined as follows:

$$\begin{gathered} q\left(x_t \mid x_{t-1}\right) := \mathcal{N}\left(x_t; \sqrt{\alpha_t} x_{t-1}, \left(1-\alpha_t\right)\mathbf{I}\right) \\ q(x_T | x_0) \approx \mathcal{N}(x_T; \mathbf{0}, \mathbf{I}) \end{gathered} \tag{2}$$

Accordingly, diffusion models [Sohl-Dickstein *et al.*, 2015; Ho *et al.*, 2020; Song *et al.*, 2020b; Song *et al.*, 2020a] are a class of generative models that gradually denoises pure random noise x_T, until it becomes true data observation x_0. The training of such a model consists of predicting x_{t-1} given x_t, where ground truth x_{t-1} can be analytically yielded as an interpolation between x_0 and x_t through Bayes rules. With reparametrization of x_t, it is rather common to predict the ϵ injected to x_0 to sample x_t, formulated as:

$$\mathcal{L}_{\text{diffusion}} = \mathbb{E}_{x_t, t, \epsilon\sim\mathcal{N}(0,1)}\left[\left\|\epsilon - \epsilon_\theta\left(x_t, t\right)\right\|_2^2\right] \tag{3}$$

3.3 Stable Diffusion Inpainters

Following from the definition of diffusion models, the Stable Diffusion uses the same training paradigm [Rombach *et al.*, 2022] but diffuses in the latent space. Formally, x_0 is first encoded with a VAE encoder to the latent space as $z_0 \in \mathbb{R}^{1\times4\times64\times64}$, and the diffusion process occurs in the latent space, where ϵ is predicted, given z_t and a text embedding condition τ. These components change the formulation to:

$$\mathcal{L}_{\text{LDM}} = \mathbb{E}_{t, c, \epsilon\sim\mathcal{N}(0,1)}\left[\left\|\epsilon - \epsilon_\theta\left(z_t, \tau, t\right)\right\|_2^2\right], \tag{4}$$

where τ is the text embedding. The so-called "Inpainters" are fine-tuned models from the Stable Diffusion checkpoints [Rombach *et al.*, 2022; Zhang *et al.*, 2023]. Typically, given some latent code z_0 of a real image X to be inpainted, inpainting models minimally receive as input a context image latent $C \in \mathbb{R}^{1\times4\times64\times64}$ and a binary mask $M \in \mathbb{R}^{1\times4\times64\times64}$ as follows:

$$C = M \circ X, \tag{5}$$

$$\text{where} \qquad M_{i,j} = \begin{cases} 0, & \text{if to be inpainted} \\ 1, & \text{if to be context} \end{cases}$$

In the Inpainter's finetuning process, Stable Diffusion model's original input channels are extended so that both M and C can be fed as additional conditional signal for denoising z_t. Then, M and C will be conditioned throughout the iterative denoising of z_T up to z_0.

In particular, the generation process spans over n feedforward iterations, where n commonly ranges between 20 to 50 for a reasonable generation. The backpropagation of n feedforward is memory-wise infeasible. The current SoTA framework Photoguard [Salman *et al.*, 2023] optimizes the context image C with PGD, the subject of protection from malicious edits. The authors of Photoguard approximate z_0 by iterating over just 4 denoising steps, which is assumed to be sufficient to synthesize an approximation $\hat{z_0}$. With this sample $\hat{z_0}$, z_0 is decoded to x_0 and the L_2 distance between x_0 and an arbitrary target image is minimized. Formally, with some simplification, Photoguard's loss is as follows:

$$\delta = \underset{\|\delta\|_2 \leq\ \epsilon}{\operatorname{argmin}} \left\|f(x+\delta, C) - x_{target}\right\|_2^2, \tag{6}$$

where f denotes the entire LDM pipeline over n feedforward steps, C is the context image, and δ is the adversarial perturbation. To the best of our knowledge, Photoguard is the only algorithm that tackles disruption in diffusion-based deepfake generation inpainters.

4 Method

4.1 Search for the Vulnerable Timestep

One of the most self-evident complications presented by SD Inpainter is the diffusion process. To take into perspective, the denoising diffusion reverse process can be thought of as the random latent code being decoded by a VAE [Oord *et al.*, 2017]. Then, it is plausible that we need to feedforward through all synthesis "stages" to optimize efficiently.

Instead of directly complying with this computational overhead, we take advantage of the "progressive synthesis" property of the diffusion process to strategically target the early timesteps. It is known to the community that early timesteps have sovereignty over the overall spatial structure and global semantics of the image [Meng *et al.*, 2021; Huang *et al.*, 2023; Chen *et al.*, 2023]. To gain a sufficient degree of freedom of disruption, we optimize our adversarial context with respect to the early stage, leading to global damage. It is noted that the model should behave similarly across timesteps adjacent to our target timestep range to sustain the vulnerability of predicted scores $\nabla_x \log p(x)$. While many researchers already rely on the linearization of adjacent timesteps [Huberman-Spiegelglas *et al.*, 2023; Miyake *et al.*, 2023], we reaffirm this linearity through both PCA decomposition of the hidden states across all timesteps, discussed in Appendix 1.5, and empirical results of effective disruption.

4.2 Timestep Constraint-Free Loss Function

While the narrowing of the sampled timestep disables access to both the diffusion-native and off-the-shelf losses (LPIPS,

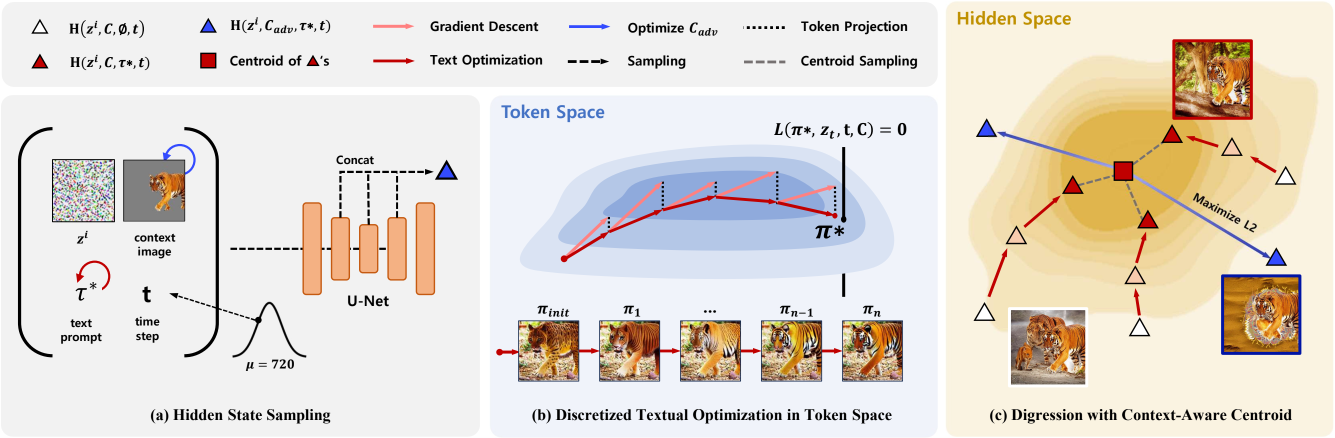

Figure 2: Overview of DDD's Framework: Our framework's objective lies in finding the context image's representative multi-modal centroid, for which our immunized image's representation semantically digresses away from it. (a) illustrates the pipeline for sampling all of the hidden states utilized in the framework. In (b), we first utilize our context image to yield the diffusion-based inpainting loss to update the token embedding π^*. Finally, with $\tau^* = \mathcal{E}(\pi^*)$, we construct the multi-modal centroid via Monte Carlo sampling.

CLIP, Perceptual loss), we take advantage of the denoiser's hidden states to yield a timestep-agnostic loss space. Specifically, we are concerned with strategically drifting the hidden state out of its learned manifold in the Euclidean space. Then, if we have the text embedding τ_{text} obtained from the text encoder $\mathcal{E}$ and the feature map $F_{C,\tau_{text}}$, our loss on the space of H is as follows:

$$\mathcal{L}_{hidden} = \sum_{i}^{N} \| H_i^{target} - H_i^{source} \|_2^2, \tag{7}$$

where $\quad H(C, \tau_{text}, z_T) = \text{Attention}(F_{C,\tau_{text},z_T})$

F_{C,τ_{text},z_T} is the feature map conditioned with C, τ_{text}. While all three conditions are indispensable for F, we omit their inscription when they are constant. Because we want to disrupt the information pathway for multi-modal signals, (C and τ), we choose the self-attention layers. Ultimately, this loss space enables the mining of both dense and timestep-agnostic features.

4.3 Aligning the Objective with Disruption's Goal

The success of untargeted attacks is well-known to the adversarial attack community.

In our setting where the sole concern is pulling away from what is correct, the untargeted attack approach is highly aligned with our implicit objective, with no other constraints imposed other than disruption itself.

Now, it is noteworthy that in the discriminative setting, the notion of "correctness" is trivially given by the probability vector. Then, cross-entropy gives us a scalar metric to evaluate the correctness. However, for the most part in generative models, this type of reduced evaluative metric is not plainly given, hence y_{true} is ambiguous. As an alternative tentative approach, we discuss the consequences of framing the approach with a targeted attack in Appendix 1.4.

To adopt untargeted attacks to diffusion models, we approximate the concept of a correct synthesis. A simple approach to obtain the correct output synthesis is to take the average at the x_t space in $\mathbb{R}^{D_{pixel}}$, where $D_{pixel} = W \times H$. Unfortunately, such averaging yields a low-fidelity and ambiguous ground truth target (equivalent to an averaged blurred image). Instead, the hidden space $\mathbb{R}^{D_{hidden}}$, where $D_{hidden} << D_{pixel}$, is suitable for yielding a high-fidelity centroid with less variance and higher spatial agnosticism. Accordingly, through Monte Carlo sampling, we approximate the ground truth representation centroid ϕ_{hidden}, defined as following:

$$\phi(C, \tau_{text}) = \mathbb{E}_{z_T \sim \mathcal{N}(0,\,1),\, t \sim \mathcal{N}(720,\,5.8)}[H(z_T,\, C,\, \tau_{text},\, t)] \tag{8}$$

While the context image is given apriori, the choice of the text prompt τ is still ambiguous. In other words, given that the deep activations accommodate for multi-modal signal [Hussain *et al.*, 2023; Elhage and et al., 2021], ignoring the text modality is prone to yield sub-optimal results.

4.4 Token Projective Textual Optimization

One simple approach is to condition over the null text, which yields a reasonable centroid that captures the signal of the context image. However, the null text is also implicitly biased and fitted to the global dataset, which is naturally long-tailed. To this end, we address this by searching in the text embedding space for the representation that embodies the context image's content and use it to condition upon our centroid-constituting instances. Specifically, we optimize text embedding τ that minimizes the diffusion loss:

$$\tau^* = \underset{\tau}{argmin}\ ||\epsilon * M - \epsilon(z_T, C, \tau, M^{\text{inv}}) * M||_2^2, \tag{9}$$

where M^{inv} is M inverted and τ is initialized from a random text embedding. One common issue with most image inversion algorithms [Zhu *et al.*, 2020; Xia *et al.*, 2022] is the propensity of the embedding to overfit to the reference image, which results in anomalous expressivity and poor generalization. While premeditatively underfitting it viable, convergence points vary for every instance.

PEZ has shown that one can invert a reference image with "decodable" token embeddings [Wen *et al.*, 2023]. Formally,

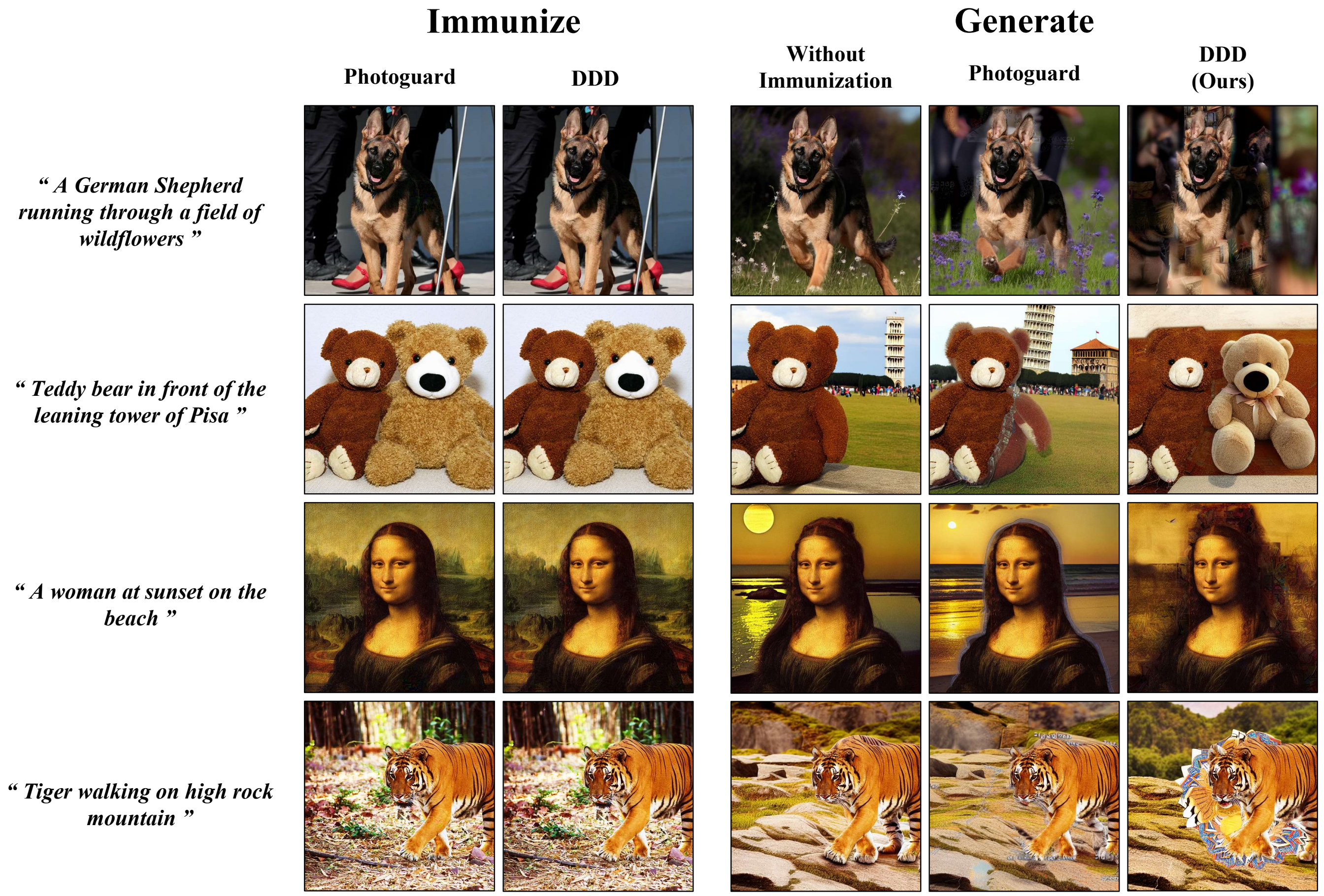
Immunize
Generate
Photoguard
DDD
Without Immunization
Photoguard
DDD (Ours)
" A German Shepherd running through a field of wildflowers "
" Teddy bear in front of the leaning tower of Pisa "
" A woman at sunset on the beach "
" Tiger walking on high rock mountain "

Figure 3: A side-by-side comparison of Photoguard against DDD. DDD disrupts the hidden representations, which leaves more clearly visible unnatural artifacts and disruption across the images.

in between every gradient descent, the updated token embeddings are projected to their nearest tokens, which are ultimately used for text decoding. Instead, we view this intermediate discretization as an approach to optimization regularization. In essence, the token embedding matrix constitutes the keypoints for the topology of the token space manifold. Because this token embedding matrix encompasses the primal semantics, having the continuous embedding projected to this learned manifold ensures that it stays in-distribution. Then, while a constraint-free embedding optimization eventually leads to overfitted solutions, the inherent discreteness of the token projection intercepts any adversarial build-up.

Formally, let π be token embeddings, $Proj$ be the token projection function, $\mathcal{E}$ be the text encoder, α be the step size, and $\mathcal{L}$ be the diffusion loss, we optimize for the token embedding that minimizes the diffusion loss with a projective update as follows:

$$\pi_{t+1} = \pi_t - \alpha \nabla_{\pi_t} \mathcal{L}(\mathcal{E}(Proj(\pi_t)), z_t, t, C) \tag{10}$$

Having obtained the optimized π^*, we have implicitly yielded its bijective counterpart $\tau^* = \mathcal{E}(\pi^*)$, the text embedding. Ultimately, this optimization will lead to a τ^* highly descriptive of the context image, and the centroid $\phi(C, \tau^*)$ will embody the expected multi-modal representation of the context image. Given that we obtained this surrogate for y_{true}, we now have access to the digression formulation in MSE, which, due to what the centroid is encoding, the optimization will lead to a digression orthogonal to the semantics of the context image. Ultimately, our final loss is as follows:

$$C^* = \underset{C}{argmax} \, || \underbrace{\phi(C, \tau^*)}_{\text{target}} - \underbrace{H(z_T, C, \tau^*, M)}_{\text{source}} ||_2^2 \tag{11}$$

The iterative update of C follows the PGD update protocol. Ultimately, as summarized in Fig. 2, due to its digressive objective, we name our framework "Digressive Guidance for Disrupting Diffusion-based Inpainters", or in short, DDD.

5 Experimental Results

5.1 Technical Details

For DDD's disruption synthesis, we have used Nvidia's A100 40GB GPU, taking under 6 minutes of optimization, more than 3 times faster than Photoguard. All of our experiments are conducted with PGD's epsilon budget of 12/255, step size of 3/255, and gradient averaging of 7 for 250 iterations. While the loss curve hints that further optimization is possible with noticeable returns, we verified that 250 is enough to cover all

	CLIP Score ↓	Aesthetic ↓	PickScore↓	FID ↑	KID ↑	LPIPS ↑	SSIM ↓	PSNR↓
(R → R) Photoguard	0.27819	5.5855	0.69	66.25	0.00183	0.4262	0.5817	29.6200
(R → R) DDD	**0.24429**	**5.2246**	**0.31**	**118.97**	**0.02093**	**0.4993**	**0.5153**	**29.3763**
(R → S) Photoguard	0.27842	5.6501	0.56	65.83	0.00157	0.3600	0.6521	29.8143
(R → S) DDD	**0.27388**	**5.4194**	**0.44**	**80.54**	**0.00481**	**0.4032**	**0.6129**	**29.6729**
(S → S) Photoguard	0.27344	5.5592	0.63	70.98	0.00315	0.3928	0.6338	29.6738
(S → S) DDD	**0.25074**	**5.2805**	**0.37**	**106.13**	**0.01345**	**0.4423**	**0.5751**	**29.5184**
(S → R) Photoguard	0.27984	5.5605	0.52	63.89	0.00200	0.4011	0.6033	29.6722
(S → R) DDD	**0.27412**	**5.5284**	**0.48**	**72.62**	**0.00312**	**0.4084**	**0.5859**	**29.6683**
Oracle (R)	0.2582	5.8147	-	-	-	-	-	-
Oracle (S)	0.2597	5.8168	-	-	-	-	-	-

Table 1: Quantitative Metrics of Disruption. We also show results for the disruption in the transferability setting. "R" stands for Runway 1.5v and "S" for Stability AI 2.0v, where R → S means disruption optimized with R and applied on inpainting with S. The arrows alongside the metric names show the desirable direction from the perspective of disruption.

examples. Additionally, our maximum memory usage is 16 GBs.

5.2 Quantitative Results

Taking into consideration that inpainting disruption in diffusion models is an emerging task, we present a variety of metrics to explore different angles of analysis of each disruption method. For all of these metrics, we curated a dataset with in-the-wild and heterogeneous images, totaling 381 pairs of images and prompts. We test over different inpainting strengths (0.8, 0.9, 1.0) to verify their disruption robustness.

The quantitative metrics can be sub-categorized into two groups. First, metrics such as CLIP Score, (CLIP) Aesthetic, and Pick-a-Score evaluate the alignment of the inpainted image and its prompt. CLIP Aesthetic and Pick-a-Score are especially insightful due to their inherent design to rank images upon their visual integrity. In this category, DDD achieves a perfect score against Photoguard. The remaining metrics, SSIM, PSNR, FID, and KID, evaluate the visual disparity between the oracle inpainting and the disrupted inpainting through deep and natural metrics. Similarly, DDD outperforms Photoguard.

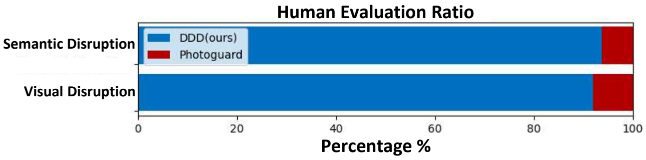

Figure 4: Human evaluation on 20 random disrupted examples.

5.3 Qualitative Results

The qualitative evaluation rules over the success of inpainting disruption. In Fig. 3, we present images, create their respective masks, and test the inpainting under no disruption, Photoguard, and our method DDD. Photoguard's disruption disassociates the context image's spatial attributes from the inpainted area. However, a large portion of the disrupted images show weak or failed disruption. DDD, on the other hand, shows consistent disruption.

As this task directly involves end users, their evaluation is indispensable. Hence, we present results on human evaluation from 31 human evaluators of Photoguard and DDD-based disruptions. We randomly chose 20 image disruption examples and collected responses on the textual and visual criteria. As shown in Fig. 4, 91.94 % and 93.85 % of the evaluators found Photoguard to display higher semantic alignment and visual coherence, respectively, with 0.1 and 0.06 standard deviations. These results imply that DDD's disruptions are more closely aligned with the basic disruption criteria. More details on the experimental setup are shared in Appendix 1.8.

One subtle detail in these two competing frameworks lies in their sensitivity to different inpainting strengths, as shown in Appendix 1.2. Although easily glanced over, the discussion of disruption under different strength thresholds is significant. Since we do not know the specific type of edit the end-user will execute apriori, Photoguard and DDD should disrupt the best when strength is close to 1.0 since no image signal is encoded at z_T and the disruption has more room for digression. At strength = 0.8, Photoguard shows weak disruption across some images while our method shows consistent and effective disruption.

5.4 Ablation Study

The right construction of centroid ϕ is crucial for asserting disruption to unseen context images and prompts. Put differently, if ϕ is at the nucleus of the context image's representation, or the expected representation, deviating away from it encourages orthogonal semantic digression. To bring light to the representative effect, we exhibit five pairs of sources and targets to compare with DDD, each with different objectives used during optimization. In Fig. 5, each setting deals with an objective where the distance between hidden representations of the source H^{source} and target H^{target} are either minimized (targeted) or maximized (untargeted). Specifically, H^{random} refers to a set of hidden states $\{H(C, \tau_0)$, ... , $H(C, \tau_{j-1}), H(C, \tau_j)\}$, where τ_j is a collection of j text embeddings of randomly chosen prompts to express an i.i.d. distribution of text.

Targeted Scenario. In Fig. 5(a), mild disruption is achieved when H^{random} is pulled to ϕ^{null}. While the inpainting deviates from the text condition, rich semantics such as Eiffel remains intact. While this setting can be considered as an alternative approach depending on the specific user

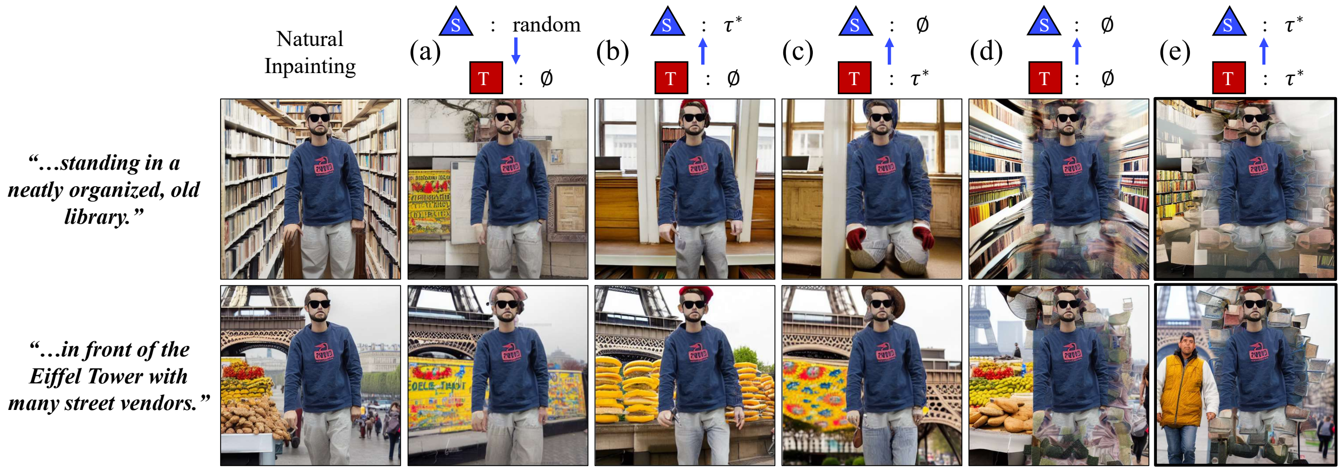
Natural Inpainting
(a)
S : random
T : ∅
(b)
S : τ*
T : ∅
(c)
S : ∅
T : τ*
(d)
S : ∅
T : ∅
(e)
S : τ*
T : τ*
"…standing in a neatly organized, old library."
"…in front of the Eiffel Tower with many street vendors."

Figure 5: An ablation study for various targeted and untargeted scenarios. In this context, "S" and "T" represent the text conditions of source and target, respectively. Furthermore, the down arrow (↓) signifies a targeted scenario, where the objective is to move closer to the fixed H^{target}, while the up arrow (↑) represents an untargeted scenario, where the objective is to move away from the fixed H^{target}. And, (e) represents the result from our work, DDD.

preference, the optimization resources, namely the number of tunable pixels of the context image, are limited and underexpressive to match a distribution of hidden states to a single point. This leads to image protection failures in a significant portion of the test cases.

Untargeted Scenario. Given that τ^* is the text that fully represents the context image's content, our method follows (e) in Fig. 5, where the semantic digression of $H(\tau^*)$'s are heading directly away from $\phi(\tau^*)$. As an experimental cause, we show (b) and (c), where we premeditatedly construct the digression to lose its orthogonality with respect to the C's semantics. Additionally, thanks to our framework's formulation, (d) shows competitive performance. However, (e) outperforms in all metrics. Namely, we have calculated the averaged scores across different checkpoints and in the transferability setting and obtained superiority in every metric. We append the full results in the Appendix Table 1.

6 Discussion

Our work entails discussions from various ends. Particularly, note that DDD is a conceptual framework. Put differently, the applicability of our framework is compatible with untargeted disruption of upcoming variants of inpainters and synthesizers, even generalizing beyond diffusion models.

Inpainting disruption is an emerging task, especially with the introduction of powerful text-to-image models. The current literature on inpainting disruption is limited and has much room for growth. For example, no disruption-specific metric has been proposed yet. As of now, human evaluation and PickScore are the metrics with the highest fidelity in grading the inpainting synthesis completeness.

Also, we found that both DDD and Photoguard need more robustness to data augmentations and agnosticity to the context image's mask positioning. We believe this field will highly benefit from these developments. In addition to these topics, we discuss experimental configurations of our framework, strength interplay, transferability, survey form, metrics, and more in the Appendix.

7 Conclusion

We proposed to immunize context images and disrupt their malicious inpainting edits through our framework DDD. To strategically manage the diffusion process, we focused on the model's hidden state statistics across different timesteps. This showed us how to select a vulnerable timestep range and reduced the complexity of the attack's forward pass. Next, we formulate the disruption objective as a semantic digression optimization, which not only offers a greater degree of optimization, but also takes full consideration of the context image. This is only possible through searching for the multimodal centroid, calibrated and regulated by token projective embedding optimization, and constructed through Monte Carlo sampling. As a result, we significantly reduce GPU VRAM usage and speed up the optimization time by 3×. Our framework is supported by quantitative and qualitative results, and contributional research resources on this novel task in diffusion inpainters.

Ethical Statements

Our work involves nudity and sexually explicit content, but as all models are publicly available, our institution's IRB advised that approval was not required. All researchers involved are over 21 and have carefully reviewed relevant ethics guidelines [NeurIPS, 2023] and undergone training to handle and analyze research results properly. Although no practical defense against creating nudity in generative models exists, we emphasize the urgency of developing preventive technologies given our work's focus on explicit and unsafe content.

Acknowledgments

The authors would thank anonymous reviewers. Geonho Son and Juhun Lee contributed equally as joint first authors. Simon S. Woo is the corresponding author. This work was partly supported by Institute for Information & communication Technology Planning & evaluation (IITP) grants funded by the Korean government MSIT: (No. RS-2022-II221199, Graduate School of Convergence Security at Sungkyunkwan University), (No. RS-2024-00337703, Development of satellite security vulnerability detection techniques using AI and specification-based automation tools), (No. 2022-0-01045, Self-directed Multi-Modal Intelligence for solving unknown, open domain problems), (No. RS-2022-II220688, AI Platform to Fully Adapt and Reflect Privacy-Policy Changes), (No. RS-2021-II212068, Artificial Intelligence Innovation Hub), (No. 2019-0-00421, AI Graduate School Support Program at Sungkyunkwan University), and (No. RS-2023-00230337, Advanced and Proactive AI Platform Research and Development Against Malicious Deepfakes), and (No. RS-2024-00356293, AI-Generated Fake Multimedia Detection, Erasing, and Machine Unlearning Research). .

References

[Chen *et al.*, 2023] Yida Chen, Fernanda Viégas, and Martin Wattenberg. Beyond surface statistics: Scene representations in a latent diffusion model. *arXiv preprint arXiv:2306.05720*, 2023.

[Cho *et al.*, 2023] Beomsang Cho, Binh M Le, Jiwon Kim, Simon Woo, Shahroz Tariq, Alsharif Abuadbba, and Kristen Moore. Towards understanding of deepfake videos in the wild. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 4530–4537, 2023.

[Choi *et al.*, 2018] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8789–8797, 2018.

[Elhage and et al., 2021] Nelson Elhage and et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021.

[Goodfellow *et al.*, 2014] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

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

[Hong *et al.*, 2024] Seunghoo Hong, Juhun Lee, and Simon S Woo. All but one: Surgical concept erasing with model preservation in text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 21143–21151, 2024.

[Hu *et al.*, 2021] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[Huang *et al.*, 2021] Qidong Huang, Jie Zhang, Wenbo Zhou, Weiming Zhang, and Nenghai Yu. Initiative defense against facial manipulation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1619–1627, 2021.

[Huang *et al.*, 2023] Ziqi Huang, Tianxing Wu, Yuming Jiang, Kelvin CK Chan, and Ziwei Liu. Reversion: Diffusion-based relation inversion from images. *arXiv preprint arXiv:2303.13495*, 2023.

[Huberman-Spiegelglas *et al.*, 2023] Inbar Huberman-Spiegelglas, Vladimir Kulikov, and Tomer Michaeli. An edit friendly ddpm noise space: Inversion and manipulations. *arXiv preprint arXiv:2304.06140*, 2023.

[Hussain *et al.*, 2023] Md Shamim Hussain, Mohammed J Zaki, and Dharmashankar Subramanian. The information pathways hypothesis: Transformers are dynamic self-ensembles. *arXiv preprint arXiv:2306.01705*, 2023.

[Isola *et al.*, 2017] Phillip Isola, 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*, pages 1125–1134, 2017.

[Le and Woo, 2023] Binh M Le and Simon S Woo. Quality-agnostic deepfake detection with intra-model collaborative learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22378–22389, 2023.

[Le *et al.*, 2023] Binh Le, Shahroz Tariq, Alsharif Abuadbba, Kristen Moore, and Simon Woo. Why do facial deepfake detectors fail? In *Proceedings of the 2nd Workshop on Security Implications of Deepfakes and Cheapfakes*, pages 24–28, 2023.

[Le *et al.*, 2024] Binh M Le, Jiwon Kim, Shahroz Tariq, Kristen Moore, Alsharif Abuadbba, and Simon S Woo. Sok: Facial deepfake detectors. *arXiv preprint arXiv:2401.04364*, 2024.

[Lee *et al.*, 2022] Sangyup Lee, Jaeju An, and Simon S Woo. Bznet: unsupervised multi-scale branch zooming network for detecting low-quality deepfake videos. In *Proceedings of the ACM Web Conference 2022*, pages 3500–3510, 2022.

[Lin *et al.*, 2023] Yupei Lin, Sen Zhang, Xiaojun Yang, Xiao Wang, and Yukai Shi. Regeneration learning of diffusion models with rich prompts for zero-shot image translation. *arXiv preprint arXiv:2305.04651*, 2023.

[Madry *et al.*, 2017] 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.

[Meng *et al.*, 2021] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021.

[Miyake *et al.*, 2023] Daiki Miyake, Akihiro Iohara, Yu Saito, and Toshiyuki Tanaka. Negative-prompt inversion: Fast image inversion for editing with text-guided diffusion models. *arXiv preprint arXiv:2305.16807*, 2023.

[Mokady *et al.*, 2023] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6038–6047, 2023.

[NeurIPS, 2023] NeurIPS. Neurips code of ethics. `https://nips.cc/public/EthicsGuidelines`, 2023. Accessed: 2023-07-07.

[Oord *et al.*, 2017] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. *arXiv preprint arXiv:1711.00937*, 2017.

[Pumarola *et al.*, 2018] Albert Pumarola, Antonio Agudo, Aleix M Martinez, Alberto Sanfeliu, and Francesc Moreno-Noguer. Ganimation: Anatomically-aware facial animation from a single image. In *Proceedings of the ECCV*, pages 818–833, 2018.

[Rombach *et al.*, 2022] 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 CVPR*, pages 10684–10695, 2022.

[Ruiz *et al.*, 2020] Nataniel Ruiz, Sarah Adel Bargal, and Stan Sclaroff. Disrupting deepfakes: Adversarial attacks against conditional image translation networks and facial manipulation systems. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 236–251. Springer, 2020.

[Salman *et al.*, 2023] 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.

[Schuhmann *et al.*, 2022] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022.

[Sohl-Dickstein *et al.*, 2015] 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.

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

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

[von Platen *et al.*, 2022] Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. `https://github.com/huggingface/diffusers`, 2022.

[Wallace *et al.*, 2023] Bram Wallace, Akash Gokul, and Nikhil Naik. Edict: Exact diffusion inversion via coupled transformations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22532–22541, 2023.

[Wang *et al.*, 2022] Xueyu Wang, Jiajun Huang, Siqi Ma, Surya Nepal, and Chang Xu. Deepfake disrupter: The detector of deepfake is my friend. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14920–14929, 2022.

[Wang *et al.*, 2023] Zhendong Wang, Yifan Jiang, Yadong Lu, Yelong Shen, Pengcheng He, Weizhu Chen, Zhangyang Wang, and Mingyuan Zhou. In-context learning unlocked for diffusion models. *arXiv preprint arXiv:2305.01115*, 2023.

[Wen *et al.*, 2023] Yuxin Wen, Neel Jain, John Kirchenbauer, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Hard prompts made easy: Gradient-based discrete optimization for prompt tuning and discovery. *arXiv preprint arXiv:2302.03668*, 2023.

[Xia *et al.*, 2022] Weihao Xia, Yulun Zhang, Yujiu Yang, Jing-Hao Xue, Bolei Zhou, and Ming-Hsuan Yang. Gan inversion: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3121–3138, 2022.

[Xu *et al.*, 2023] Xingqian Xu, Jiayi Guo, Zhangyang Wang, Gao Huang, Irfan Essa, and Humphrey Shi. Prompt-free diffusion: Taking" text" out of text-to-image diffusion models. *arXiv preprint arXiv:2305.16223*, 2023.

[Yeh *et al.*, 2020] Chin-Yuan Yeh, Hsi-Wen Chen, Shang-Lun Tsai, and Sheng-De Wang. Disrupting image-translation-based deepfake algorithms with adversarial attacks. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops*, pages 53–62, 2020.

[Zhang *et al.*, 2023] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023.

[Zhu *et al.*, 2017] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.

[Zhu *et al.*, 2020] Jiapeng Zhu, Yujun Shen, Deli Zhao, and Bolei Zhou. In-domain gan inversion for real image editing. In *European conference on computer vision*, pages 592–608. Springer, 2020.