Disrupting Diffusion-based Inpainters with Semantic Digression

Geonho Son, Juhun Lee, and Simon S. Woo

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 "Digression guided Diffusion Disruption" 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.

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 be-
coming 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 dissociates from the context image. Our “Semantic Digression-guided Diffusion Disruption”, DDD in short, is 3× 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 fine-tuned 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 $L$, a projective function $\Pi$, and an image $x$ its true label $y$, the PGD [Madry et al., 2017] attack is performed as follows:

$$x^{t+1} = \Pi_{x+S}(x^t - \alpha \text{ sign}(\nabla_x L(x; y_{\text{target}})))$$  \hspace{1cm} (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 $x_t \in \mathbb{R}^{1 \times 3 \times W \times H}$, where $x_T$ is an isotropic Gaussian noise, $x_0$ is true data observation, and any $x_t$ between these is defined as follows:

$$q(x_t | x_{t-1}) := \mathcal{N}(x_t; \sqrt{\alpha_t} x_{t-1}, (1 - \alpha_t) I)$$

$$q(x_T | x_0) \approx \mathcal{N}(x_T; 0, I)$$  \hspace{1cm} (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 $\epsilon$ injected to $x_0$ to sample $x_t$, formulated as:

$$\mathcal{L}_{\text{diffusion}} = \mathbb{E}_{x_t, t, \epsilon \sim \mathcal{N}(0, 1)} \left[ \|\epsilon - \epsilon_\theta (x_t, t)\|_2^2 \right]$$  \hspace{1cm} (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 $z_0 \in \mathbb{R}^{1 \times 4 \times 64 \times 64}$, and the diffusion process occurs in the latent space, where $\epsilon$ is predicted, given $z_t$ and a text embedding condition $\tau$. These components change the formulation to:

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

where $\tau$ 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 $C \in \mathbb{R}^{1 \times 4 \times 64 \times 64}$ and a binary mask $M \in \mathbb{R}^{1 \times 4 \times 64 \times 64}$ as follows:

$$C = M \circ X,$$  \hspace{1cm} (5)

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 $x_T$ 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:

$$\delta = \arg \min \|f(x + \delta, C) - x_{\text{target}}\|_{2},$$ \hspace{1cm} (6)

where $f$ denotes the entire LDM pipeline over $n$ feedforward steps, $C$ is the context image, and $\delta$ 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 lowering of the sampled timestep disables access to both the diffusion-native and off-the-shelf losses (LPIPS,
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 \( \tau_{text} \) obtained from the text encoder \( E \) and the feature map \( F_{C, \tau_{text}} \), our loss on the space of \( H \) is as follows:

\[
\mathcal{L}_{\text{hidden}} = \sum_{i} \left\| H_{t}^{\text{target}} - H_{t}^{\text{source}} \right\|_{2}^{2},
\]

where \( H(C, \tau_{text}, t) = \text{Attention}(F_{C, \tau_{text}, z_{T}}) \) is the feature map conditioned with \( C, \tau_{text} \). 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, \((C, \tau)\), 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 averaging at the \( x_{i} \) space in \( \mathbb{R}^{D_{\text{pixel}}} \), where \( D_{\text{pixel}} = W \times H \). Unfortunately, such averaging yields a low-fidelity and ambiguous ground truth target (equivalent to an averaged blurred image). Instead, the hidden space \( \mathbb{R}^{D_{\text{hidden}}} \), where \( D_{\text{hidden}} \ll D_{\text{pixel}} \), is suitable for yielding a high-fidelity centroid with less variance and higher spatialagnosticism. Accordingly, through Monte Carlo sampling, we approximate the ground truth representation centroid \( \phi_{\text{hidden}} \), defined as following:

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

While the context image is given apriori, the choice of the text prompt \( \tau \) is still ambiguous. In other words, given that the deep activations accommodate for multi-modal signal [Hussein 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 \( \tau \) that minimizes the diffusion loss:

\[
\tau^{*} = \arg \min_{\tau} \| \epsilon \ast M - \epsilon(z_{T}, C, \tau, M^{\text{inv}}) \ast M \|_{2}^{2},
\]

where \( M^{\text{inv}} \) is \( M \) inverted and \( \tau \) 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,
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 $\pi$ be token embeddings, $\text{Proj}$ be the token projection function, $\mathcal{E}$ be the text encoder, $\alpha$ 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:

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

Having obtained the optimized $\pi^*$, we have implicitly yielded its bijective counterpart $\tau^* = \mathcal{E}(\pi^*)$, the text embedding. Ultimately, this optimization will lead to a $\tau^*$ 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:

$$C^* = \arg \max_C || \phi(C, \tau^*) - H(z_T, C, \tau^*, M) ||^2_2$$  \hspace{1cm} (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
Semantic Disruption
Visual Disruption

Table 1: Quantitative Metrics of Disruption. We also show results for the disruption in the transferability setting. “R” stands for Runway 1.5 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.

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.

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 dissociates 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 a priori, Photoguard and DDD should disrupt the best when strength is close to 1.0 since no image signal is encoded at zT 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 Hsource and target Htarget are either minimized (targeted) or maximized (untargeted). Specifically, Hrandom refers to a set of hidden states \{H(C, τ0), ..., H(C, τj−1), H(C, τj)\}, where τj 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 Hrandom is pulled to ϕnull. 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 needs.
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 $H^\text{target}$, while the up arrow (↑) represents an untargeted scenario, where the objective is to move away from the fixed $H^\text{target}$. And, (e) represents the result from our work, DDD.

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 $H^\text{target}$, while the up arrow (↑) represents an untargeted scenario, where the objective is to move away from the fixed $H^\text{target}$. And, (e) represents the result from our work, DDD.

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 $3 \times$.

Our framework is supported by quantitative and qualitative results, and contributonal 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 & communications 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


