Distraction is All You Need: Memory-Efficient Image Immunization against Diffusion-Based Image Editing

Ling Lo
National Yang Ming Chiao Tung University
linglo.ee08@nycu.edu.tw

Cheng Yu Yeo
National Yang Ming Chiao Tung University
boyyeo123.ee12@nycu.edu.tw

Hong-Han Shuai
National Yang Ming Chiao Tung University
hhshuai@nycu.edu.tw

Wen-Huang Cheng
National Taiwan University
wenhuang@csie.ntu.edu.tw

Abstract

Recent text-to-image (T2I) diffusion models have revolutionized image editing by empowering users to control outcomes using natural language. However, the ease of image manipulation has raised ethical concerns, with the potential for malicious use in generating deceptive or harmful content. To address the concerns, we propose an image immunization approach named semantic attack to protect our images from being manipulated by malicious agents using diffusion models. Our approach focuses on disrupting the semantic understanding of T2I diffusion models regarding specific content. By attacking the cross-attention mechanism that encodes image features with text messages during editing, we distract the model's attention regarding the content of our concern. Our semantic attack renders the model uncertain about the areas to edit, resulting in poorly edited images and contradicting the malicious editing attempts. In addition, by shifting the attack target towards intermediate attention maps from the final generated image, our approach substantially diminishes computational burden and alleviates GPU memory constraints in comparison to previous methods. Moreover, we introduce timestep universal gradient updating to create timestep-agnostic perturbations effective across different input noise levels. By treating the full diffusion process as discrete denoising timesteps during the attack, we achieve equivalent or even superior immunization efficacy with nearly half the memory consumption of the previous method. Our contributions include a practical and effective approach to safeguard images against malicious editing, and the proposed method offers robust immunization against various image inpainting and editing approaches, showcasing its potential for real-world applications.

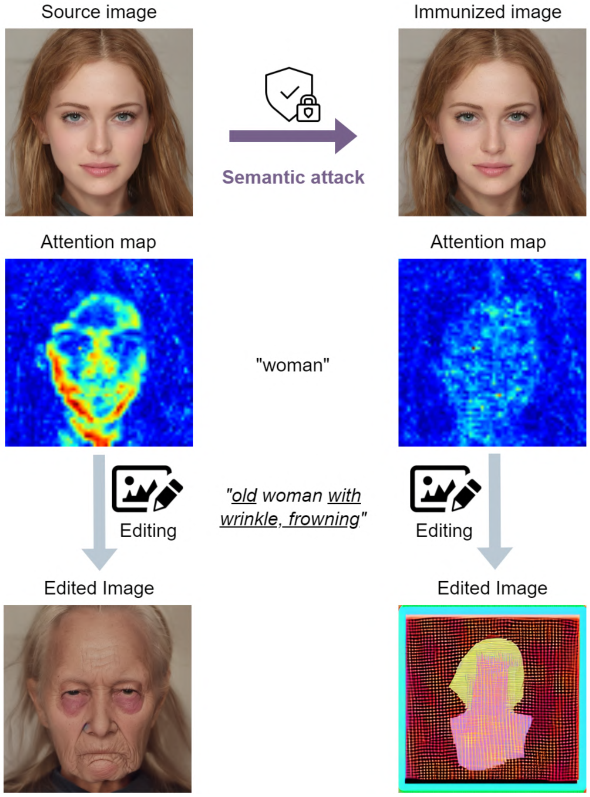

Figure 1. An illustration of our semantic attack. We aim to immunize the image by corrupting the semantic understanding of T2I diffusion models regarding specific content (in this case, ”woman” as shown in the figure). Subsequently, the editing process is compromised, as the model cannot accurately identify the proper region for editing a certain concept.

1. Introduction

In recent years, there has been an exponential surge in the development and deployment of diffusion models [7, 9, 27].These models exhibit the ability to generate highly realistic images conditioned on various inputs such as text, layouts, and scene graphs [8, 11, 38, 45]. Among them, text-to-image (T2I) diffusion models advance the state of the art in image modeling. They allow humans to control the outcomes of diffusion models using natural language, making them accessible and intuitive for a wide audience. The potential of large pretrained T2I models such as Stable Diffusion [16, 19], Imagen [22] and DALLE 2 [18] further goes beyond conditional generation and extends across diverse applications including image inpainting, image editing, zero-shot image classification, open-vocabulary segmentation [2–4, 13, 14, 30, 33, 36, 37, 46, 47].

While the technological leap has opened exciting possibilities, it has also raised ethical concerns related to the misuse of text-driven image editing since the barriers to image manipulation have been substantially lowered. For instance, malicious attackers can harness these advanced diffusion models to produce inappropriate, deceptive, or even harmful digital content that is almost indistinguishable from reality. The thrives of AI-powered image editing can have severe ethical implications, posing substantial risks to the privacy and reputation of individuals as shown in Figure 1.[1]

Amidst these concerns, image immunization emerges as a central theme to defend personal digital visual content [1, 20, 21, 23, 40]. The concept of image immunization is aimed at providing protection against the potential malpractice of AI-powered image editing, empowering individuals and organizations to share their images with confidence. It usually involves the addition of a carefully crafted and imperceptible perturbation to an image, which disrupts the operation of AI-powered editing approaches. Several works have explored related ideas and techniques to protect digital property from AI-powered image editing, disrupting the output generation by either nullifying the image transformation [40] or corrupting the output image [1, 21]. However, the majority of these image immunization approaches were designed to counter GAN-based image editing. For diffusion models, Salman *et al.* are the first to introduce PhotoGuard [23] that makes subsequent edits using T2I diffusion models appear unrealistic. They execute a targeted attack on the model using projected gradient descent (PGD) [12], forcing any output editing to resemble the target image. However, it requires backpropagation through the entire diffusion process, leading to inefficient GPU memory usage. Additionally, its efficacy in inpainting scenarios is unstable due to the lack of constraints on the region for adding perturbation. For instance, the noise can be eliminated by the mask during inpainting, leading to a significant reduction in the overall immunization ability.

In this paper, we propose **semantic attack**, which is designed to attack the T2I diffusion models, ensuring the immunization against various editing approaches. Our objective is to tackle the memory-intensive issue while enhancing the efficacy of immunization. As malicious editing often seeks to preserve specific content while manipulating other regions or change particular content in the image negatively, we argue that the goal of image immunization is to protect certain content rather than the entire image. Therefore, our method aims to disrupt the semantic understanding of diffusion models regarding the specific region of our concerns within the image. As long as the concerned region is not masked out during inpainting, our immunization leads to poor model comprehension during editing, ultimately compromising or rendering the editing outcomes ineffective.

Specifically, T2I diffusion models understand the semantics of an image through the cross-attention mechanism, where the text message interacts with the image features. As successful text-based image editing relies on a robust understanding of the image, faithful attention maps play a pivotal role in effective image editing and, vice versa, immunization. Hence, our semantic attack focuses on disrupting the cross-attention associated with the content of our concerns within the diffusion model. We intend to distract the attention, lead the model to lose focus, impair its judgment, and render it uncertain about which areas should be edited. Figure 1 illustrates the immunity of our proposed semantic attack against diffusion-based image editing. Supposing the area representing a woman is our primary concern, our semantic attack disrupts comprehension of the diffusion model, and the attention to the word ”woman” within the image region is distracted. As a result, the subsequent editing process fails to identify the proper content for modification, leading to the generation of an unrealistic image.

Moreover, since our attack focuses on corrupting the intermediate attention map rather than the final generated image, the extensive backpropagation path through the whole diffusion process is no longer necessary. The lengthy process that involves several diffusion steps can be broken down and treated as independent denoising steps with different levels of input noise. Accordingly, our semantic attack significantly reduces the computational burden and alleviates GPU memory constraints. In addition, to create a perturbation capable of disrupting the denoising model under varying noise levels, we introduce timestep universal gradient updating. It considers different noise levels of input and seeks a timestep-agnostic perturbation to interfere with the denoising model during the whole generative process. Combined with our semantic attack, our approach can be seen as attacking the cross-attention map of a single denoising model with different noise levels of input, and the resulting perturbation can be viewed as discretely disrupting the full diffusion process. Experimental results show that the proposed approach is able to effectively disrupt semantic information and provide robust immunization against various image inpainting and editing approaches.

[1] For ethical considerations, the human face image is synthesized by the diffusion model.

2. Related Work

2.1. Text-to-Image Diffusion Models

Recently, there has been noteworthy progress in text-conditioned diffusion models, leading to great enhancements in the quality of generated samples. These strides have empowered large-scale diffusion models like Stable Diffusion [19], Imagen [22], and DALL-E [18] to generate high-fidelity images that adhere closely to predefined text-based conditions. Researchers have also harnessed the potential of text-guided diffusion models in tackling the challenge of single-image editing [2, 4, 10, 13–15, 39, 44]. Their objective is to tap into the rich and diverse semantic knowledge embedded within these expansive models. Several approaches have been devised to facilitate text-conditioned inpainting under the assumption that the user provides a mask to constrain and guide the editing process [2, 15]. However, the reliance on the user-provided masks can be burdensome. Consequently, alternative methods have been introduced to enable direct image editing based solely on textual descriptions, eliminating the need for explicit masks. For instance, Meng *et al.* [13] proposed to add random noise to the input image and then perform a text-guided denoising process from a predefined step. Couairon *et al.* [4] proposed to automatically generate the mask and then edit an image based on a text query. Mokady *et al.* [14] proposed to edit a real image by optimizing a null-text embedding for better image structure preservation and altering the cross-attention map during the denoising process.

2.2. Image Immunization

As text-guided diffusion models continue to lower the barrier for image editing, the need to safeguard our digital property from malicious alterations becomes increasingly critical. Image immunization serves as a defense mechanism, involving the proactive disruption of image editing models to thwart potential manipulations and protect our digital assets. Yeh *et al.* [40] proposed the Limit-Aware Self-Guiding Gradient Sliding Attack (LaS-GSA) to counter manipulation by image-to-image GANs under black-box settings. Aneja *et al.* [1] introduced Targeted Adversarial Attacks Against Facial Image Manipulations (TAFIM) to cancel the effect of manipulation by generating the perturbation that will lead to a predefined target. While most of the previous approaches targeted GAN-based image manipulation, Salman *et al.* [23] introduced PhotoGuard to first counter manipulations by diffusion-based models. Their introduced encoder attack aimed at disabling the autoencoder in text-guided diffusion models. In addition to encoder attack, they proposed the diffusion attack, which disrupts the entire diffusion process of the diffusion model to corrupt any editing attempts. While these proposed attacks against diffusion models have proven effective against image editing, their versatility against image inpainting can be uncertain, as the regions containing the perturbation can potentially be masked out. Furthermore, it is worth noting that attacking the full diffusion process can be memory-intensive due to the lengthy full diffusion process. In this paper, we propose an approach to immunization against diffusion-based image editing that is both memory-efficient and effective against image editing and inpainting, which marks a distinct and important shift in our approach.

3. Method

3.1. Text-to-Image Diffusion Model

The objective of a T2I diffusion model is to generate high-quality images by progressively denoising the noisy inputs conditioned on the given text. The denoising diffusion process is typically accomplished using a time-conditioned U-net, which is frequently trained within the latent space of a Variational Autoencoder (VAE) [19]. To generate an image, a latent vector z_T with Gaussian noise is first initialized as input. The denoising diffusion process follows, wherein the U-Net iteratively denoises the latent vector conditioned on a text prompt c embedded with the CLIP text encoder [17]. Finally, the resulting vector is decoded into an image utilizing the decoder $\mathcal{D}$ of the VAE.

During training, the encoder $\mathcal{E}$ of the VAE encodes a given image x into its latent representation z_0. Then, a forward diffusion process is applied to transform the latent vector z_0 to the approximate Gaussian noise z_T over time T. After the denoising diffusion process, the decoder $\mathcal{D}$ reconstructs the estimated $\widetilde{x}$. The main objective of the denoising U-Net is to reconstruct the latent representation z_0 from a perturbed representation z_T considering the textual description c, and the overall training objective can be written as:

$$\mathbb{E}_{z_0,\epsilon}\left[\|\epsilon-\epsilon_\theta\left(z_t,t,c\right)\|_2^2\right], \tag{1}$$

where ϵ is the noise added in the forward diffusion process, t is the timestep indicating the perturbation noise level, and ϵ_θ is the denoising U-net with attention blocks parameterized by θ. As the model is conditioned on t, the optimization can be viewed as seeking the best model ϵ_θ for denoising across all levels of noises. For image editing, recent approaches typically introduce random noise [13, 43] or leverage DDIM inversion [4, 14, 29] to incorporate estimated noise into target images as the forward diffusion process, generating the noisy z_T and enabling further editing of the image based on the condition c.

3.2. Semantic Attack

Given an image x, we aim to immunize the image into x_{adv} by adversarial attacking the diffusion models and introduc-

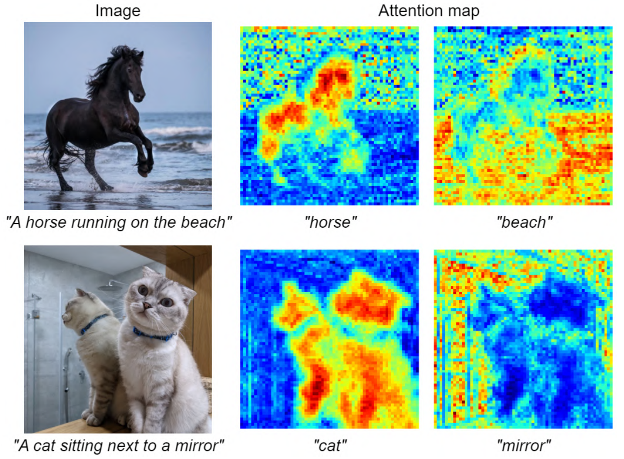
Image
Attention map
"A horse running on the beach"
"horse"
"beach"
"A cat sitting next to a mirror"
"cat"
"mirror"

Figure 2. An illustration of the text-to-pixel interaction in the diffusion model. The model associates distinct words, or content, with specific regions in the image, enabling further manipulation of the identified areas.

ing perturbations δ:

$$x_{adv} = x + \delta, \quad \text{s.t.} \quad \|\delta\| < \kappa,$$

where κ is the perturbation budget. After immunization, any subsequent attempts at modification on x_{adv} using diffusion models should be sabotaged while the perturbation remains imperceivable, i.e., δ is constrained by a norm $\|\cdot\|$, which could be either the ℓ_1-, ℓ_2-, or ℓ_∞-norm.

The goal of our semantic attack is to immunize images by attacking the T2I diffusion model employed in the image editing process, ensuring that editing results of the immunized image x_{adv} yield *poorly edited images*, e.g., contradicting the original editing intention. The core principle of image editing using a T2I diffusion model hinges on its capacity to precisely identify objects targeted for editing. Our objective, therefore, is to impair the model's semantic understanding of the current image. Given that malicious editing typically entails either changing specific content or maintaining it while altering other areas, our semantic attack is designed to safeguard particular content rather than the whole image. Essentially, if the diffusion model faces difficulty in accurately recognizing the semantics of certain content within the image, it becomes challenging to carry out the intended editing effectively.

The semantic knowledge that a diffusion model possesses about an image stems largely from the pixel-to-text interaction within its architecture, particularly through the cross-attention mechanism of the denoising U-net, as highlighted in recent studies [6, 28, 31, 32, 35]. This interaction can be visualized in attention maps, such as the one shown in Figure 2, where each word in the text description corresponds to specific pixels in the image, revealing the semantic relationship between text and image context. To disrupt this crucial relationship, we propose a targeted suppression of the attention responses for specific content within the textual description during the denoising process. This intervention is designed to disrupt the semantic knowledge that the diffusion model holds about the current image, ultimately sabotaging the image editing process.

To realize this idea, we propose an effective way to aggregate the attention response associated with the textual description of our focal content for immunization. In T2I diffusion models, the interaction between the image and textual description occurs in the noise prediction U-net, where visual and textual features are fused using cross-attention layers. The attention map can be written as:

$$A^l(x_{adv}, c_a) = Softmax((W_q^l \epsilon_\theta^l(x_{adv}))(W_k^l c_a)^T/\sqrt{d}), \tag{2}$$

where $A^l(x_{adv}, c_a)$ indicates the attention map of the l^{th} intermediate block in the U-net, W_q and W_k are the projection matrices, $\epsilon_\theta^l(x_{adv})$ is the output deep features of the l^{th} block in the denoising U-net, c_a denotes the text embedding of the concerned content, and d is the latent projection dimension. To execute a comprehensive attack on the entire denoising U-Net, it is crucial to efficiently aggregate attention maps across various scales. Given the fully-convolution nature of the U-net, intermediate coordinates map locally to surrounding areas in the initial-sized feature map. Therefore, we upsample the intermediate attention maps to match the size of the initial feature map using bicubic interpolation. We sum the attention maps pixel by pixel:

$$Att(x_{adv}, c_a) = \sum_{l=1}^{L} upsample(A^l(x_{adv}, c_a)), \tag{3}$$

where L is the number of intermediate blocks in the U-net. The aggregated attention maps corresponding to the textual token, represent the semantic knowledge that the diffusion model possesses for the target image at various scales.

Specifically, the region in the immunized image x_{adv} containing the particular content should exhibit a low attention response associated with the textual description. Consequently, we generate a mask M to identify the region containing the content using the original image x:

$$M = \mathbb{I}(Att(x, c_a) > \tau), \tag{4}$$

where the indicator function $\mathbb{I}(\cdot)$ is applied to binarize the attention map of the original image $Att(x, c_a)$ using a threshold τ.

The overall objective of the proposed semantic attack is to suppress the attention pattern of our focal content in the corresponding region by minimizing the attention-suppressing loss:

$$L = \|Att(M \odot x_{adv}, c_a)\|_1, \tag{5}$$

Here, $\|\cdot\|_1$ denotes the component-wise ℓ_1-norm, c_a is the textual description of the content that we aim to immunize, and $Att(M \odot x_{adv}, c_a)$ denotes the attention response on the region containing concerned content in the immunized image x_{adv}. It is worth noting that, as our attack focuses on distracting the attention of the diffusion model from certain regions containing specific content, the immunity remains consistent in the context of inpainting as long as the concerned content is not masked out.

3.3. Timestep Universal Gradient Updating

While attacking the full diffusion process, the primary objective is to influence the generation of the image. The backpropagation for every pixel across all T time steps is necessary since the entire sequence is treated as a holistic image generation process, resulting in significant memory requirements. In contrast, our proposed method aims to interfere with the semantic knowledge that the model has regarding the current image. Therefore, our attack shifts from disrupting the full diffusion process to impairing the denoising U-net itself. Interfering the full diffusion process can then be broken down into corrupting independent denoising timesteps, each operating on input with varying levels of injected noise. Consequently, our approach demonstrates memory efficiency since it operates depending on gradients from a single iteration of the denoising U-net.

Since we perceive the diffusion process as denoising input with varying levels of noise, we aim to achieve effective attack across different timesteps, or noise levels, to ensure its immunity in disrupting the holistic process of image editing. To create immunizations capable of corrupting the semantic knowledge within an image under varying levels of injected noise, we introduce timestep universal gradient updating. In essence, we calculate and update the adversarial sample while taking into account different levels of injected noise simultaneously. The proposed immunization aims to disrupt the image in a way that prevents the time-conditioned denoising U-net from functioning effectively across different timesteps. Specifically, we generate the perturbation in an update process by minimizing the attention-suppressing loss L. In each iteration, the perturbation δ is computed using the gradient from the model in different diffusion timesteps. To ensure the perturbations remain imperceptible, we constrain our attack by limiting the distance from the original clean sample x, so it does not exceed the perturbation budget κ. The detailed gradient updating algorithm is provided in Algorithm 1.

Algorithm 1 Timestep Universal Gradient Updating

```
Input: Input image $x$, the focal content to immunize $c_a$, content mask $M$, perturbation budget $\kappa$, attacking step size $s$, number of attacking steps $N$, diffusion timestep $T = \{t_1, t_2, ...t_j\}$
Initialize adversarial perturbation $\delta \leftarrow 0$, and immunized image $x_{adv} \leftarrow x$
for $n = 1...N$ do
    Initialize all gradients: $all\_grad \leftarrow 0$
    for $t$ in $T$ do
        Inject the forward noise onto the immunized image: $x^t_{adv} \leftarrow x_{adv}$
        Compute the attention suppressing loss: $L \leftarrow \|Att(M(x^t_{adv}), c_a)\|_1$
        Compute the gradient: $\nabla_{x_{adv}} L$
        Update the gradients: $all\_grad \leftarrow all\_grad + \nabla_{x_{adv}} L$
    end for
    Compute the mean of the gradient values:
     $all\_grad \leftarrow mean(all\_grad)$
    Update the adversarial perturbation:
     $\delta \leftarrow (\delta + s \cdot sign(all\_grad))$
     $\delta \leftarrow clip(\delta, -\kappa, \kappa)$
    Update the immunized image: $x_{adv} \leftarrow x_{adv} - \delta$
end for
Return: The immunized image $x_{adv}$
```

4. Experimental Results

4.1. Training Setup

In our experiments, we generate the immunization for the images by attacking the open-source Stable Diffusion model V1.4 [19] hosted on the Hugging Face. We evaluate the performance of our semantic attack qualitatively and quantitatively compared to two state-of-the-art attacks proposed in [23], encoder attack and diffusion attack. The encoder attack targets only the VAE in the stable diffusion model, while the diffusion attack aims to target the entire diffusion process. For a fair comparison, we set the perturbation budget $\kappa = 0.06$, number of iterations $N = 100$ for all attacks. For the diffusion attack and our semantic attack, the number of diffusion timesteps T is set to 10.

4.2. Comparison on Image Inpainting

We initially assess the immunizing impact of our semantic attack on image inpainting, addressing the malicious scenario where attackers seek to preserve specific content in the image while maliciously altering other regions. The results are demonstrated in Figure 3. As depicted in the figure, the immunization generated through our semantic attack disrupts the model's ability to recognize the corgi or the men in the image. Consequently, the editing process fails to execute accurately. In the first example, the model misinterprets the corgi, generating another one instead of correctly following the prompt for the dog or cat. In the second example, the model fails to generate the correct background or clothes described in the prompt for the men, as the region

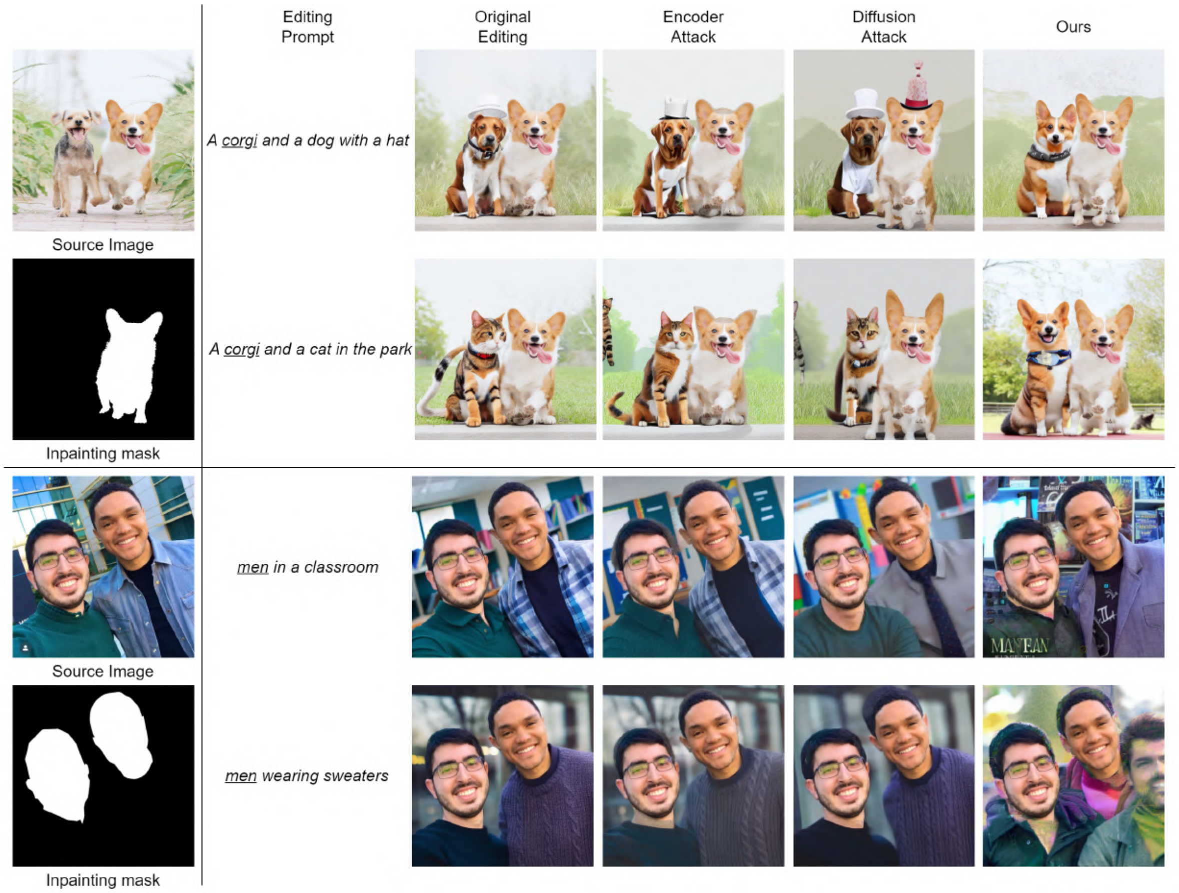

Figure 3. Qualitative Results of our semantic attack compared to previous approaches [23] against image inpainting. The content to be immunized in our semantic attack is indicated by the underlined word in the text prompt. More examples are included in Supplementary.

corresponding to men is not successfully recognized.

On the other hand, previous methods that aim to attack the entire image may have a chance to fail in immunization and generate results following the prompt, as the perturbation can potentially be masked during the inpainting process. In contrast, our method is effective after applying the inpainting mask, as our focus is on distracting the attention of the model on certain content from corresponding regions.

4.3. Comparison on Image Editing

We simulate another scenario of malicious editing where attackers alter specific contents in our images using image-to-image editing, and assess the efficacy of our proposed method against such vicious editing. Upon scrutinizing the editing outcomes depicted in Figure 4, it is evident that the dissimilarity between the editing results after our semantic attack and the original editing attempt is pronounced, especially in the area that contains our concerned content, implying the content of our concern is successfully protected. Notably, our attack retains its effectiveness even when the content indicated in the attack is not explicitly present in the editing prompt. It stems from our attack strategy, which distracts the attention of the model on certain content in the image, disrupting its overall understanding and resulting in editing outcomes with a structural disparity.

We then quantitatively evaluate the performance by assessing the image quality of the edited results, measuring the dissimilarity between the immunized editing outcomes and the original editing attempts. Since there is no public dataset for evaluating the efficacy of image immunization, we adopt experimental settings similar to [23]. To be specific, we first generate 150 images featuring 3 distinct objects using the diffusion model. For each object, we create 2 editing prompts corresponding to two malicious editing scenarios: altering specific content in the image or manipulating other regions[2]. The editing results using Stable Diffusion V1.4 and V2.0 according to the editing prompts are reported by averaging the editing results over 20 random seeds. The quantitative comparison of the image quality assessment, highlighting the disparities between the original

[2] More details of the evaluation datasets are reported in Supplementary.

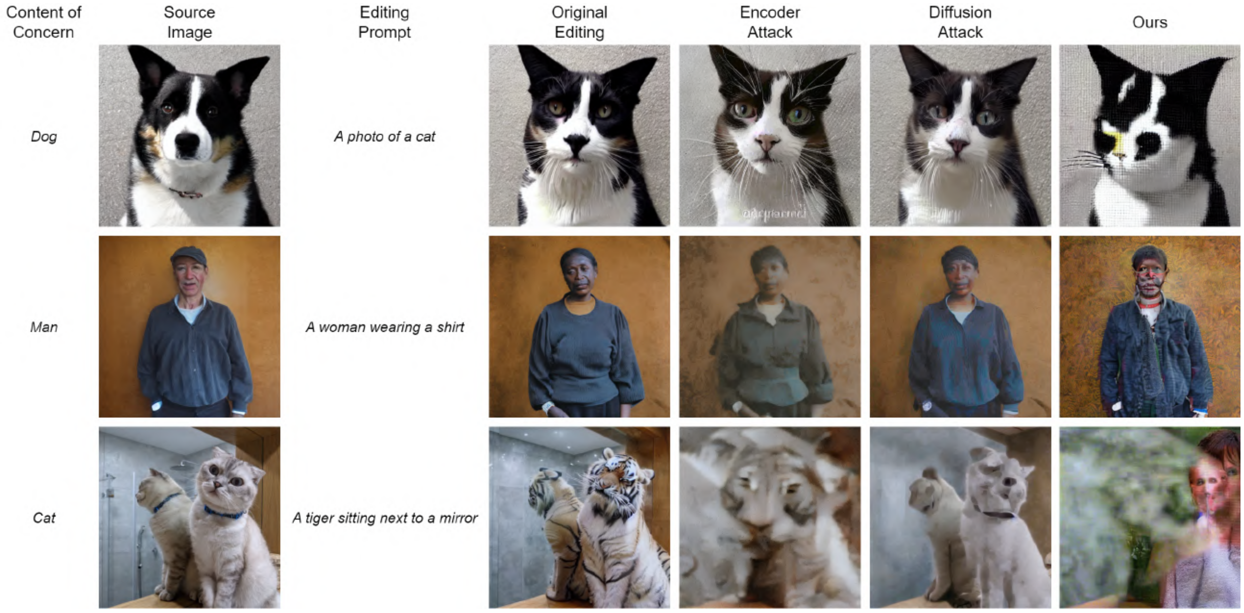

Figure 4. Qualitative comparison of our semantic attack to previous approaches [23] against image editing [13]. The content to be immunized in our semantic attack is indicated by the underlined word in the text prompt. More examples are included in Supplementary.

Edit Model / Metrics	Stable Diffusion V1.4			Stable Diffusion V2.0		
	Encoder Attack	Diffusion Attack	Ours	Encoder Attack	Diffusion Attack	Ours
PSNR ↓	18.8437	18.2617	**15.1487**	18.5955	19.3797	**18.0589**
SSIM [34] ↓	0.6318	0.6504	**0.4470**	0.6045	0.6440	**0.4719**
VIFp [24] ↓	0.2118	0.2656	**0.1462**	0.1618	0.1832	**0.1008**
FSIM [41] ↓	0.7757	0.7693	**0.6584**	0.7453	0.7794	**0.7313**
LPIPS [42] ↑	0.4131	0.4056	**0.5901**	0.5799	0.4869	**0.6019**

Table 1. Quantitative comparison of image quality assessment. The arrows next to the metrics indicate the decrease in image similarity between the editing outcomes after immunization and the original editing attempt.

editing attempt and the editing outcomes after immunization, is presented in Table 1. The results indicate that our semantic attack outperforms the baseline attacks across various evaluation metrics. Notably, our attack excels in metrics related to visual quality as perceived by the human eye such as SSIM, VIFp, and LPIPS, showcasing a significant improvement compared to the baseline. It suggests that our attack results in editing outcomes that are visibly different from the original editing attempts, making them discernible to humans and effectively protecting the image.

4.4. Comparison on Memory Efficiency

Figure 6 illustrates the comparison of GPU memory consumption between our attack and the diffusion attack. The x-axis represents the number of diffusion timesteps in the attacked model, with a higher number indicating a stronger attack. While both attacks target the diffusion model, the diffusion attack aims to compromise the full diffusion process, and our attack breaks down the process into different denoising steps. As depicted in the figure, GPU memory requirements for the diffusion attack scale proportionally with the number of diffusion timesteps due to the extensive gradients involved in the entire diffusion process. In contrast, the memory cost of our attack barely scales up even with multiple diffusion steps to attack. As the backpropagation path is now reduced to the denoising model of a single timestep instead of multiple steps, our attack can achieve stronger immunization with lower memory cost.

4.5. Immunity against More Image Editing Approaches

Recently, advanced approaches have been proposed to enhance real-image editing using diffusion models [4, 14, 30]. We conduct experiments to examine if our semantic attack remains effective against advanced editing techniques. Fig-

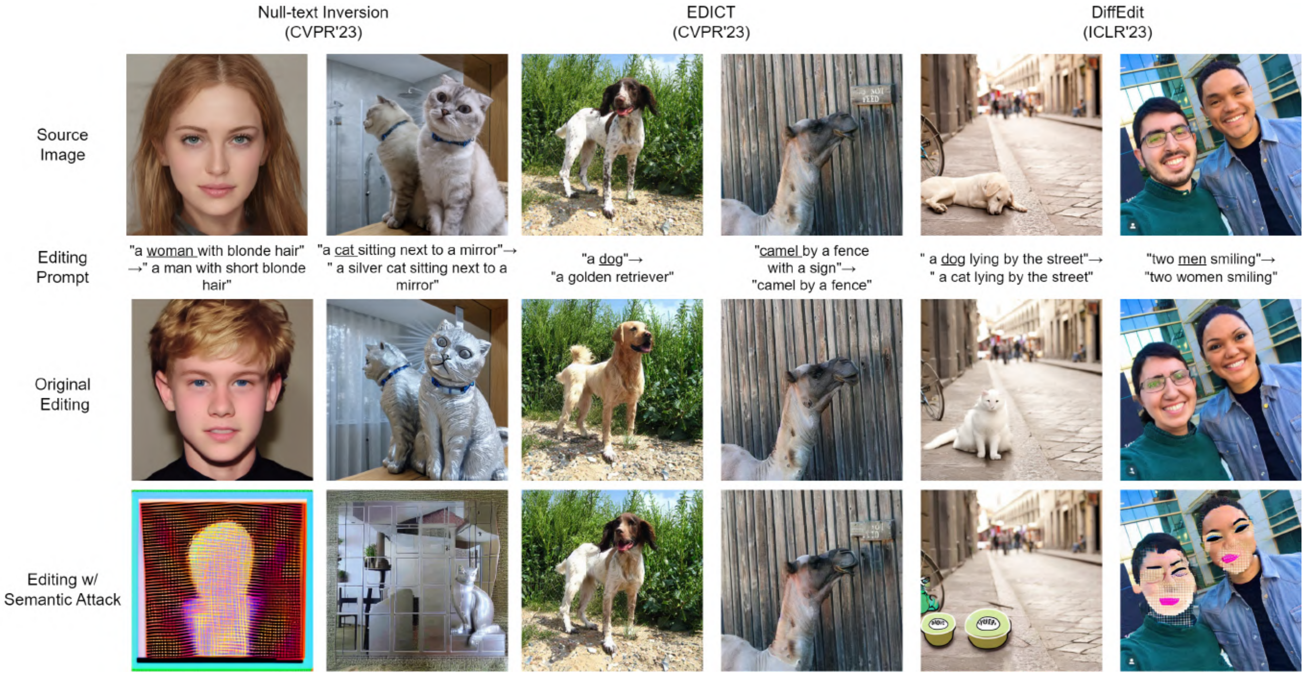
Null-text Inversion
(CVPR'23)
EDICT
(CVPR'23)
DiffEdit
(ICLR'23)
Source Image
Editing Prompt
"a woman with blonde hair" →" a man with short blonde hair"
"a cat sitting next to a mirror"→ " a silver cat sitting next to a mirror"
"a dog"→ "a golden retriever"
"camel by a fence with a sign"→ "camel by a fence"
" a dog lying by the street"→ " a cat lying by the street"
"two men smiling"→ "two women smiling"
Original Editing
Editing w/ Semantic Attack

Figure 5. Qualitative results of our semantic attack against state-of-the-art image editing approaches. The content to be immunized in our semantic attack is indicated by the underlined word in the text prompt. More results are included in Supplementary.

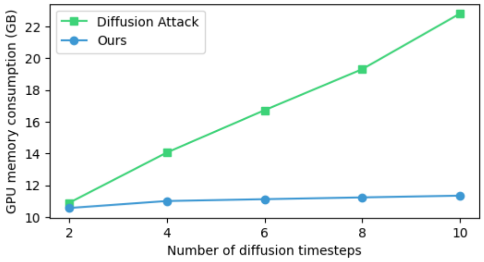

Figure 6. Comparison on GPU memory consumption when attacking the model with different numbers of diffusion timesteps.

ure 5 presents the results of our experiments. The editing results undergo significant disruption after immunization through our semantic attack. Notably, null-text inversion [14] results are severely impacted, given their reliance on accurate attention maps for the editing object. In contrast, EDICT [30] generates a nullifying effect on the editing of immunized images. One possible explanation is that EDICT seeks an exact inversion of a real image leveraging the noises predicted by the denoising U-net. Our attack on the denoising U-net may induce a failed inversion, rendering the forward diffusion process in editing ineffective and nullifying the resulting edits.

5. Limitation

One potential limitation of our approach lies in its potential ineffectiveness following noise purification applied to the immunized image. Malicious attackers may nullify the impact of the perturbation through blurring or JPEG compression. Nevertheless, various approaches for creating robust perturbations can be integrated into our approach, mitigating this limitation by considering the impact of blurring or compression during perturbation generation [5, 25, 26].

6. Conclusion

We propose semantic attack to safeguard images against malicious editing using T2I diffusion models. Our method disrupts the semantic knowledge of the models, proving effective against various image inpainting and editing approaches. Moreover, we introduce timestep universal gradient updating to ensure robustness across different noise levels, discretely disrupting the full diffusion process with a lower GPU memory load, enhancing the practicality. Experiments highlight the superiority of our semantic attack in both quantitative and qualitative evaluations. Our approach offers a valuable defense against unauthorized diffusion-based manipulations of digital visual content, enhancing the integrity and reliability of images shared or stored online.

Acknowledgment

This work is partially supported by the National Science and Technology Council, Taiwan under Grant, NSTC-112-2628-E-002-033-MY4, NSTC-112-2634-F-002-002-MBK, NSTC-112-26234-F-002-005, NSTC-112-2221-E-A49-059-MY3 and NSTC-112-2221-E-A49-094-MY3.

References

[1] Shivangi Aneja, Lev Markhasin, and Matthias Nießner. Tafim: Targeted adversarial attacks against facial image manipulations. In *European Conference on Computer Vision*, pages 58–75, 2022. 2, 3

[2] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18208–18218, 2022. 2, 3

[3] Kevin Clark and Priyank Jaini. Text-to-image diffusion models are zero-shot classifiers. In *International Conference Learning Representation Workshop on Multimodal Representation Learning: Perks and Pitfalls*, 2023.

[4] Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. In *International Conference on Learning Representations*, 2023. 2, 3, 7

[5] Kanchana Vaishnavi Gandikota, Paramanand Chandramouli, and Michael Moeller. On adversarial robustness of deep image deblurring. In *IEEE International Conference on Image Processing*, pages 3161–3165, 2022. 8

[6] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. In *International Conference on Learning Representations*, 2022. 4

[7] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, pages 6840–6851, 2020. 1

[8] Siyuan Huang, Zan Wang, Puhao Li, Baoxiong Jia, Tengyu Liu, Yixin Zhu, Wei Liang, and Song-Chun Zhu. Diffusion-based generation, optimization, and planning in 3d scenes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16750–16761, 2023. 1

[9] Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models. *Advances in Neural Information Processing Systems*, 35:23593–23606, 2022. 1

[10] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6007–6017, 2023. 3

[11] Yunji Kim, Jiyoung Lee, Jin-Hwa Kim, Jung-Woo Ha, and Jun-Yan Zhu. Dense text-to-image generation with attention modulation. In *IEEE/CVF International Conference on Computer Vision*, pages 7701–7711, 2023. 1

[12] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. 2

[13] 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. In *International Conference on Learning Representations*, 2022. 2, 3, 7

[14] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6038–6047, 2023. 2, 3, 7, 8

[15] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob Mcgrew, Ilya Sutskever, and Mark Chen. GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models. In *International Conference on Machine Learning*, pages 16784–16804, 2022. 3

[16] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 2

[17] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763, 2021. 3

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

[19] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 2, 3, 5

[20] Nataniel Ruiz, Sarah Adel Bargal, and Stan Sclaroff. Disrupting deepfakes: Adversarial attacks against conditional image translation networks and facial manipulation systems. In *European Conference on Computer Vision Workshops*, pages 236–251, 2020. 2

[21] Nataniel Ruiz, Sarah Adel Bargal, Cihang Xie, and Stan Sclaroff. Practical disruption of image translation deepfake networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 14478–14486, 2023. 2

[22] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems*, pages 36479–36494, 2022. 2, 3

[23] Hadi Salman, Alaa Khaddaj, Guillaume Leclerc, Andrew Ilyas, and Aleksander Madry. Raising the cost of malicious AI-powered image editing. In *International Conference on Machine Learning*, pages 29894–29918, 2023. 2, 3, 5, 6, 7

[24] Hamid R Sheikh and Alan C Bovik. Image information and visual quality. *IEEE Transactions on Image Processing*, 15 (2):430–444, 2006. 7

[25] Mengte Shi, Sheng Li, Zhaoxia Yin, Xinpeng Zhang, and Zhenxing Qian. On generating jpeg adversarial images. In *IEEE International Conference on Multimedia and Expo*, pages 1–6, 2021. 8

[26] Richard Shin and Dawn Song. Jpeg-resistant adversarial images. In *Advances in Neural Information Processing Sys-*

tems Workshop on Machine Learning and Computer Security, page 8, 2017. 8

[27] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. 1

[28] Raphael Tang, Linqing Liu, Akshat Pandey, Zhiying Jiang, Gefei Yang, Karun Kumar, Pontus Stenetorp, Jimmy Lin, and Ferhan Ture. What the DAAM: Interpreting stable diffusion using cross attention. In *Association for Computational Linguistics*, pages 5644–5659, 2023. 4

[29] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1921–1930, 2023. 3

[30] Bram Wallace, Akash Gokul, and Nikhil Naik. Edict: Exact diffusion inversion via coupled transformations. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22532–22541, 2023. 2, 7, 8

[31] Jia Wang, Jingcheng Ke, Hong-Han Shuai, Yung-Hui Li, and Wen-Huang Cheng. Referring expression comprehension via enhanced cross-modal graph attention networks. *ACM Transactions on Multimedia Computing, Communications and Applications*, 19(2):1–21, 2023. 4

[32] Jia Wang, Hong-Han Shuai, Yung-Hui Li, and Wen-Huang Cheng. Language-guided residual graph attention network and data augmentation for visual grounding. *ACM Transactions on Multimedia Computing, Communications and Applications*, 2023. 4

[33] Su Wang, Chitwan Saharia, Ceslee Montgomery, Jordi Pont-Tuset, Shai Noy, Stefano Pellegrini, Yasumasa Onoe, Sarah Laszlo, David J Fleet, Radu Soricut, et al. Imagen editor and editbench: Advancing and evaluating text-guided image inpainting. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18359–18369, 2023. 2

[34] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 7

[35] Hongxia Xie, Ming-Xian Lee, Tzu-Jui Chen, Hung-Jen Chen, Hou-I Liu, Hong-Han Shuai, and Wen-Huang Cheng. Most important person-guided dual-branch cross-patch attention for group affect recognition. In *IEEE/CVF International Conference on Computer Vision*, pages 20598–20608, 2023. 4

[36] Shaoan Xie, Zhifei Zhang, Zhe Lin, Tobias Hinz, and Kun Zhang. Smartbrush: Text and shape guided object inpainting with diffusion model. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22428–22437, 2023. 2

[37] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2955–2966, 2023. 2

[38] Han Xue, Zhiwu Huang, Qianru Sun, Li Song, and Wenjun Zhang. Freestyle layout-to-image synthesis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14256–14266, 2023. 1

[39] Binxin Yang, Shuyang Gu, Bo Zhang, Ting Zhang, Xuejin Chen, Xiaoyan Sun, Dong Chen, and Fang Wen. Paint by example: Exemplar-based image editing with diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18381–18391, 2023. 3

[40] Chin-Yuan Yeh, Hsi-Wen Chen, Hong-Han Shuai, De-Nian Yang, and Ming-Syan Chen. Attack as the best defense: Nullifying image-to-image translation gans via limit-aware adversarial attack. In *IEEE/CVF International Conference on Computer Vision*, pages 16188–16197, 2021. 2, 3

[41] Lin Zhang, Lei Zhang, Xuanqin Mou, and David Zhang. Fsim: A feature similarity index for image quality assessment. *IEEE Transactions on Image Processing*, 20(8):2378–2386, 2011. 7

[42] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018. 7

[43] Yuxin Zhang, Nisha Huang, Fan Tang, Haibin Huang, Chongyang Ma, Weiming Dong, and Changsheng Xu. Inversion-based style transfer with diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10146–10156, 2023. 3

[44] Zhixing Zhang, Ligong Han, Arnab Ghosh, Dimitris N Metaxas, and Jian Ren. Sine: Single image editing with text-to-image diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6027–6037, 2023. 3

[45] Guangcong Zheng, Xianpan Zhou, Xuewei Li, Zhongang Qi, Ying Shan, and Xi Li. Layoutdiffusion: Controllable diffusion model for layout-to-image generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22490–22499, 2023. 1

[46] Junchen Zhu, Huan Yang, Huiguo He, Wenjing Wang, Zixi Tuo, Wen-Huang Cheng, Lianli Gao, Jingkuan Song, and Jianlong Fu. Moviefactory: Automatic movie creation from text using large generative models for language and images. *Proceedings of ACM International Conference on Multimedia*, pages 9313–9319, 2023. 2

[47] Junchen Zhu, Huan Yang, Wenjing Wang, Huiguo He, Zixi Tuo, Yongsheng Yu, Wen-Huang Cheng, Lianli Gao, Jingkuan Song, Jianlong Fu, et al. Mobilevidfactory: Automatic diffusion-based social media video generation for mobile devices from text. In *Proceedings of ACM International Conference on Multimedia*, pages 9371–9373, 2023. 2