A Pilot Study of Query-Free Adversarial Attack against Stable Diffusion

Haomin Zhuang
South China University of Technology
semzm@mail.scut.edu.cn

Yihua Zhang
Michigan State University
zhan1908@msu.edu

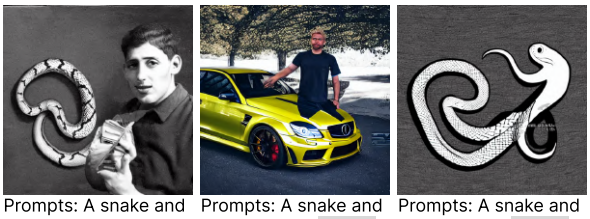
Sijia Liu
Michigan State University
liusiji5@msu.edu

Abstract

Despite the record-breaking performance in Text-to-Image (T2I) generation by Stable Diffusion, less research attention is paid to its adversarial robustness. In this work, we study the problem of adversarial attack generation for Stable Diffusion and ask if an adversarial text prompt can be obtained even in the absence of end-to-end model queries. We call the resulting problem 'query-free attack generation'. To resolve this problem, we show that the vulnerability of T2I models is rooted in the lack of robustness of text encoders, e.g., the CLIP text encoder used for attacking Stable Diffusion. Based on such insight, we propose both untargeted and targeted query-free attacks, where the former is built on the most influential dimensions in the text embedding space, which we call steerable key dimensions. By leveraging the proposed attacks, we empirically show that only a five-character perturbation to the text prompt is able to cause the significant content shift of synthesized images using Stable Diffusion. Moreover, we show that the proposed target attack can precisely steer the diffusion model to scrub the targeted image content without causing much change in untargeted image content.

1. Introduction

Diffusion models (DMs), the recently predominant generative modeling technique, have been used in a wide range of computer vision (CV) applications. Examples include Text-To-Image (T2I) generation [1–5], adversarial robustness [6–8], and image reconstruction [9, 10]. In this paper, we focus on DM for T2I generation. The key idea following [1] is to start from a noisy input and then iteratively refine it through a pre-trained representation network, *e.g.,* CLIP (Contrastive Language-Image Pretraining) [11] that connects texts and images. The above pipeline allows the use of various 'text prompts' (*i.e.*, natural language inputs served as instructions of DM) to effectively control the content of the synthesized images [3, 12].



Prompts: A snake and a young man (a) No attack. Prompts: A snake and a young man C63RR (b) Untargeted attack. Prompts: A snake and a young man -O8=* (c) Targeted attack.

Figure 1. An illustration of our attack method against Stable Diffusion. The generated perturbations are highlighted in blue. The targeted attack aims to erase the image content related to 'young man' highlighted in red. All the images are generated from the same seed.

However, several works [15–18] showed that adversarial perturbations (in terms of small textual/visual input perturbations [19, 20]) can significantly impair the performance of a CLIP model, and thus induce the adversarial robustness concern of its downstream applications. Inspired by the above, our interest in this paper is to investigate the adversarial robustness of T2I generation using CLIP-based DMs, *i.e.*, Stable Diffusion [1] in this work. In particular, we ask:

> ***(Q)*** *Can we generate adversarial perturbations against T2I models in a query-free regime?*

Adversarial attacks (also known as adversarial perturbations or examples) that can cause models' erroneous prediction have introduced immense research efforts in both vision and language domains [19–23]. A few recent attentions were also paid on T2I models [14, 24] as different from ordinary vision or language models, the latter adopts a natural language prompt to influence its imagery output. The controllability and flexibility of adjusting text prompts provide a new way to interact with a model. In [24], model query-based adversarial attacks were proposed for T2I models.

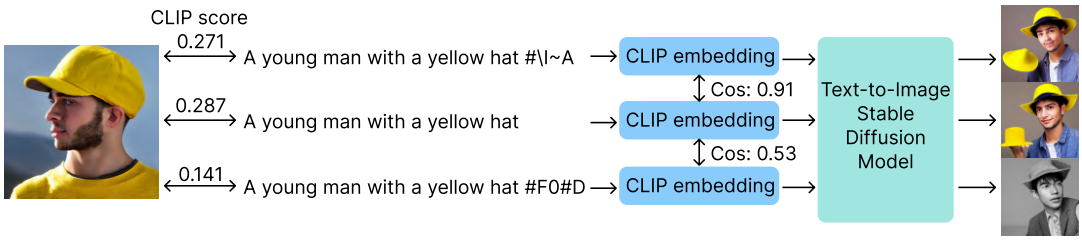


Figure 2. Illustration of robustness issue in CLIP text encoder for image generation. CLIP score [13, 14] measures the similarity between the image-text pair provided by the CLIP model, while 'Cos' measures the cosine similarity between two CLIP embeddings.

Yet, this work calls for many model queries (10000 queries per attack) to find a successful adversarial prompt, *e.g.*, using 4 newly generated words integrated into the original textual input. By contrast, our work focuses on *query-free* attack generation and the perturbation is constrained to only five characters. In [14], a gradient-based optimization was proposed to generate proper prompts that can match given images or sentences. Although it also demonstrates the controllability of image outputs by textual input prompts, little attention was paid to adversarial robustness.

To be specific, our **contributions** are unfolded below.

① We develop a query-free adversarial attack generator for T2I DMs. We show that a five-character perturbation, determined by text embeddings of CLIP, is able to significantly alter the content of DM-synthesized image (see **Fig. 1b** for an illustration).

② We provide an analysis of the correspondence between the semantics of synthesized images and the embeddings of CLIP. The obtained insight further drives us to develop a controllable "targeted" attack, where the perturbations can be refined to steer the DM's output (see **Fig. 1c** for an illustration).

③ We empirically show the effectiveness of our proposal across three attack implementation methods on a variety of text-image pairs. In particular, we demonstrate that the both the proposed untargeted and targeted Query-Free Attack can successfully alter the output image content using only a 5-character prompt perturbation. This achievement is also reflected by the significantly reduced CLIP scores of the outputs.

2. Related Work

Adversarial attacks. Adversarial attacks typically deceive DNNs by integrating carefully-crafted tiny perturbations into input data [19, 25–36]. Based on how an adversary interacts with the victim model, adversarial attacks can be categorized into white-box attacks (with full access to the victim model based on which attacks are generated) and black-box attacks (with access only to the victim model's input and output). The former typically leverages the local gradient information of the victim model to generate attacks, *e.g.*, [19, 25, 26], while the latter takes input-output model queries for attack generation; Examples include score-based attacks (*e.g.*, [31–33]) and decision-based attacks (*e.g.*, [34–36]). In this work, we assume that the adversary has access to the CLIP text encoder but can be blind to the diffusion model for image generation. Our goal is to design an adversarial attack to fool the stable diffusion model without executing the diffusion process, which would take a high model query and computation cost. Thus, we term our proposal the 'query-free attack'.

Prompt perturbations in vision-language models. Recent studies [14, 24, 37] have explored the over-sensitivity of text-to-image diffusion models to prompt perturbations in the text domain. The adversarial robustness problem of CLIP was also studied in [15–18], such as the design of imperceptible pixel perturbations [15, 16] and attacks in the image frequency domain [17]. Yet, the previous studies focused on perturbations to image inputs of CLIP, it lacks investigation into how the textual perturbation to CLIP can influence the T2I diffusion model.

3. Our Proposal

Problem statement. In this section, we first present an overview of Stable Diffusion, and then introduce our objective to generate small perturbations on the textual inputs so as to maneuver the DM's synthesized images.

We choose Stable Diffusion as the victim T2I DM model due to its popularity and availability as an open-source model. In Stable Diffusion, the DM denoises images in latent space and utilizes a cross-attention mechanism to guide the denoising process. In addition, text inputs (or textual prompts) are processed by the CLIP's text encoder to generate text embeddings and are then sent to the cross-attention layer in the denoising network. This eventually determines the synthesized images based on the CLIP's textual embeddings and the selected random seed of the initial noisy pixels. However, as exemplified in **Fig. 2**, small perturbations on the text input of CLIP can lead to different CLIP scores [13, 14], given by the values of cosine similarity of every text-image input pair. This is because of the sensitivity of

the CLIP's text embedding to text perturbations. Based on that, we ask: Can we generate an adversarial textual prompt by leveraging the lack of robustness of the CLIP's text encoder so as to fool the DM-based image generator in Stable Diffusion?

Attack model. We assume that the adversary has access to the trained text encoder of the CLIP model, and can perturb the textual prompt of the trained State Diffusion model using an additional word within a *five-character* length. Let $\tau_\theta(\mathbf{x})$ denote the text encoder of CLIP with parameters θ evaluated at the textual input $\mathbf{x}$. And we denote by $\mathbf{x}'$ the perturbed textual prompt used as the input of Stable Diffusion. We then define the *attacker's objective* by minimizing the cosine similarity between the text embeddings of $\mathbf{x}$ and $\mathbf{x}'$. This leads to the following attack generation problem

$$\min_{\boldsymbol{x}'} \cos(\tau_\theta(\boldsymbol{x}), \tau_\theta(\boldsymbol{x}')), \tag{1}$$

where cos refers to the cosine similarity metric.

Despite the simplicity of the attack generation in (1), we will show that it can be used to attack State Diffusion in a targeted way effectively. More importantly, the generation of the perturbed input $\mathbf{x}'$ no longer relies on the optimization over the diffusion model, and can thus be computationally efficient. This is in contrast to [20], which requires 10000 queries to diffusion model for generating a single attack. Since no attack intention is specified in (1), we call the resulting attack an '*untargeted attack*'.

Attack methods. Since problem (1) is differentiable, various optimization methods can be adopted for attack generation. Inspired by the previous studies on adversarial attacks in the language domain, we consider the following attack methods.

PGD attack. Similar to the PGD (projected gradient descent) attack [26] in the image domain, the PGD attack in the language domain has also been developed [23, 38]. The key idea is to formulate the textual perturbation problem as a token selection problem (over a set of token candidates) when a token site is determined for perturbation. Our experiments follow the PGD implementation in [23] to solve problem (1).

Greedy search. Different from the above PGD attack, we next consider a heuristics-based perturbation strategy. We conduct a greedy search on the character candidate set to select the top 5 characters (used for textual perturbation), which can reduce the loss of (1) to the maximum extent.

Genetic algorithm. We follow [39] to generate a population of perturbation candidates and use the loss of (1) to evaluate the quality of each candidate. In each iteration, the genetic algorithm calls genetic operations such as mutation to generate new candidates. The process continues until the number of generations is met.

Targeted attack and steerable key dimensions. In what follows, we investigate if the attack generated by (1) can be further *refined* towards a *targeted* attack purpose, *e.g.*, the intention of removing the 'young man'-related image content from the original image in Fig. 1-(c) vs. (a). To this end, we propose a new concept termed *steerable key dimensions* in the text embedding space, along which the attack generator can be guided to design customized textual perturbations. By constraining the perturbations on these steerable key dimensions, we can improve the likelihood of image generation following the adversary's intention.

To be specific, we first generate a sequence of augmented sentences $\{s_i\}_{i=1}^n$ that reflect the adversary's intention, *e.g.*, the sentences centered on 'a young man' in Fig. 1-(a). The generation of $\{s_i\}_{i=1}^n$ can be realized using *e.g.*, ChatGPT by requesting 'Generate n simple scenes and end with "and a young man" without extra words'. Two examples are s_1 = 'A bird flew high in the sky and a young man' and s_2 = 'The sun set over the horizon and a young man' with $n = 2$. Next, we perturb $\{s_i\}_{i=1}^n$ by *removing* the adversary's intention-related sub-sentence (*i.e.*, 'a young man' in Fig. 1-(a)). This results in a modified sequence $\{s_i'\}_{i=1}^n$; For example, s_1' = 'A bird flew high in the sky' and s_2' = 'The sun set over the horizon'. We then obtain the corresponding CLIP embeddings $\{\tau_\theta(s_i)\}_{i=1}^n$ and $\{\tau_\theta(s_i')\}_{i=1}^n$. As a result, the text embedding difference $\boldsymbol{d}_i = \tau_\theta(s_i) - \tau_\theta(s_i')$ can characterize the *saliency* of the adversary's intention-related sub-sentence in the text embedding space. For n such difference vectors $\{\mathbf{d}_i\}_{i=1}^n$, we determine the *steerable key dimensions* for targeted attack generation by identifying the most influential dimensions in the difference vectors $\{\mathbf{d}_i\}_{i=1}^n$. The dimension influence is given by a majority vote of $\{\mathbf{d}_i\}_{i=1}^n$ along each dimension. That is, $I_j = 1$ (*i.e.*, the indicator of the jth dimension being influential), if $|\sum_{i=1}^n \text{sign}(d_{i,j})| > \epsilon n$, where sign is the sign operation, $d_{i,j}$ is the jth entry of $\mathbf{d}_i$, and $\epsilon < 1$ is a threshold to pick the most influential dimensions. As a result, the binary vector I encodes the selected key dimensions. By integrating I into (1), we obtain the key dimensions-guided targeted attack generation

$$\min_{\boldsymbol{x}'} \cos(\tau_\theta(\boldsymbol{x}) \odot I, \tau_\theta(\boldsymbol{x}') \odot I), \tag{2}$$

where $\odot$ is the element-wise product. Note that problem (2) can be similarly solved as (1) using the attack methods introduced before. Fig. 1-(c) shows an example of using prompt perturbations generated by the proposed targeted attack to erase the image content related to 'a young man'.

4. Experiment

4.1. Experiment setups

Model setup. Throughout the experiments, we use Stable Diffusion model v1.4 [1] as the victim model for image

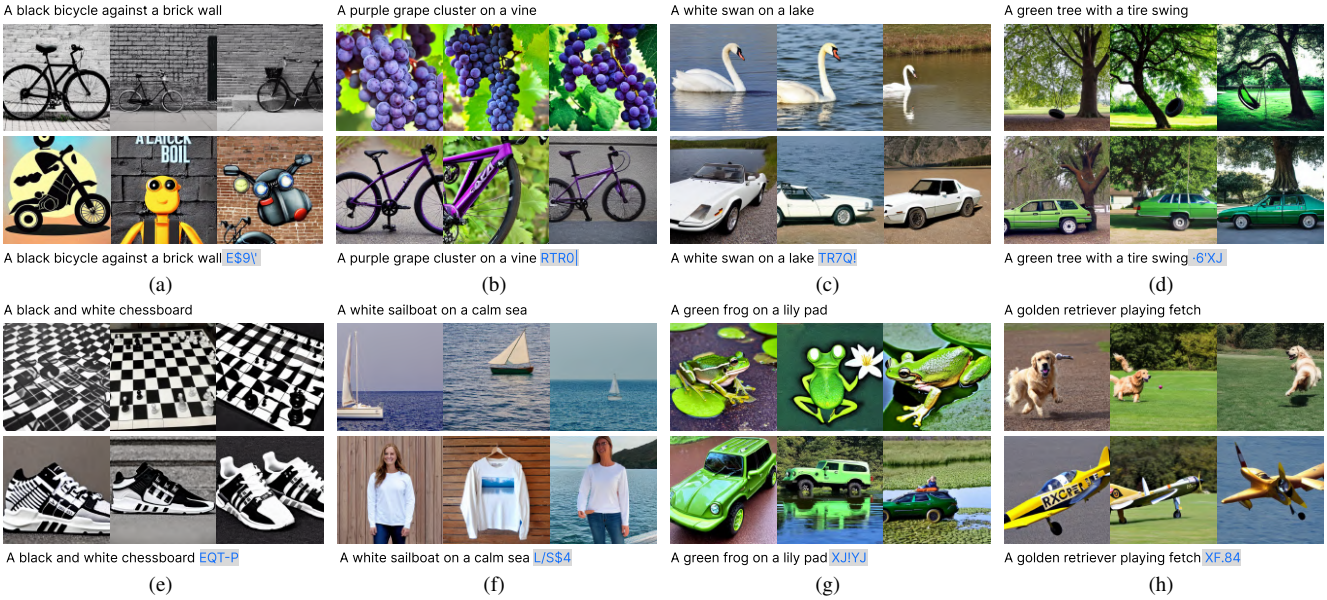


Figure 3. Illustrations of the effect of *untargeted* query-free attacks. In each group, the first row of images is generated using the original prompts vs. the second row using the perturbed ones. The perturbations found by our method are highlighted in blue in the prompt. Images in the same column share the same random seed.

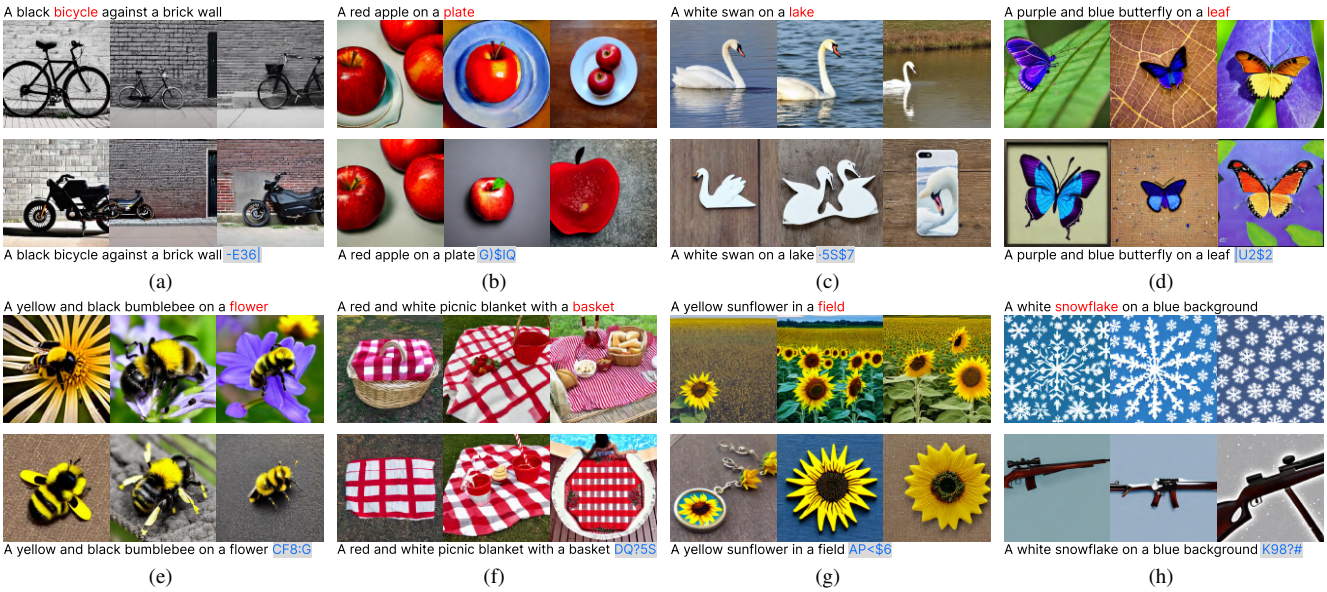


Figure 4. Illustrations of the effect of *targeted* query-free attacks. Input perturbations are generated to modify/remove the red text-related image content. Other settings are aligned with Fig. 3. Adversary targets for erasing (a) the 'bike', (b) the 'plate', (c) the 'lake', (d) the 'leaf', (e) the 'flower', (f) the 'basket', (g) the 'field', and (h) the 'snowflake' without altering the other semantics much.

generation. The proposed query-free attack has the access to the CLIP model (ViT-L/14) that shares the same text encoder as Stable Diffusion. The CLIP model is trained on a dataset containing text-image pairs [40].

Attack implementation. When implementing the PGD attack method, we set the base learning rate by 0.1 and the number of PGD steps by 100. When implementing the genetic algorithm, we set the number of generation steps, the number of candidates per step, and the mutation rate by 50, 20, and 0.3. When implementing the targeted attack, we use

Table 1. CLIP scores [13, 14] comparison of images generated with different methods. CLIP scores are used to indicate the similarity between the generated images and the embeddings of the corresponding text prompts. For each method, the CLIP scores reported below are averaged over 20 prompts and 10 images per prompt. In particular, the scores calculated based on the original sentences and output images are adopted for the untargeted attack and based on the targeted content and output images for the targeted setting. The lowest (best) score in each row is in **bold** and the results in the form $a \pm b$ denote the mean value a and the standard deviation b.

Method:	**No Attack**	**Random**	**Greedy**	**Genetic**	**PGD**
Untargeted Attack					
Score:	0.277±0.022	0.271±0.021	0.255±0.039	**0.203**±0.042	0.226±0.041
Targeted Attack					
Score:	0.229±0.03	0.223±0.037	0.204±0.037	**0.186**±0.04	0.189±0.041

ChatGPT [41] to generate $n = 10$ sentences to characterize the steerable key dimensions and set $\epsilon = 0.9$ to determine the influence mask I in (2). In addition, we utilize ChatGPT generating 20 prompts forming an input text dataset for the quantization by requesting 'Generate 20 simple scenes for text-to-image generative model'.

Evaluation metrics. In addition to different implementations of our proposed query-free attack (*i.e.*, PGD, Greedy, and Genetic methods), we also introduce a baseline that randomly generates random five-character prompt perturbations, termed Random. We evaluate the effectiveness of an attack using the CLIP score [13, 14] to characterize the similarity between the text input and the generated image. A lower CLIP score represents a lower semantic correlation between the generated image and the input text, indicating the higher effectiveness of the attacking method. To quantify the CLIP score between the targeted objects and images generated in the targeted attack setting, we utilized the template sentence 'This is a photo of' [11] as text input to measure text-image pair similarity. The CLIP score reported for each method will be averaged over 20 prompts, based on each of which 10 images will be generated.

4.2. Experiment results

Query-free attack can successfully alter the image output of Stable Diffusion using only a 5-character prompt perturbation. In **Fig. 3**, we present examples of text-to-image generation *with* and *without* suffering prompt perturbations generated by the untargeted, genetic algorithm-based query-free attack. The 5-character prompt perturbation is highlighted in blue with gray background. As we can see, the proposed attack can significantly alter the content of the original image produced by Stable Diffusion. For example, in Fig. 3-(a), the perturbation 'E$9\' ' drives the model to generate images far from the true topic 'bicycle'. The same observation can also be drawn from examples in Fig. 3-(b)-(h). This implies that the attack against the text embedding remains effective in manipulating the output of text-to-image generation.

Table 2. CLIP scores [13, 14] comparison of different perturbation prompts in case study. For each prompt, the CLIP scores reported below are averaged over 10 images from the same seeds. The lowest (best) score in each row is in **bold**.

Perturbation prompt:	**None**	**'E$9\' '**	**'E'**	**'WALLE'**	**'-E'**
Score:	0.293	**0.217**	0.297	0.291	0.285

Similar to Fig. 3, **Fig. 4** presents examples of targeted, genetic-based query-free attacks against Stable Diffusion. For example, in Fig. 4-(a), the adversary targets perturbing 'bicycle' without altering the background 'brick wall' much. This contrasts to Fig. 3-(a), where the image object and scene may change. Another example is Fig. 3-(b), where the object 'plate' is erased using the perturbed prompt but the object 'apple' is retained. We can also draw similar observations from other examples. Briefly, the targeted attack can precisely manipulate the diffusion model to avoid the targeted semantics (*i.e.*, the red text highlight above each image example in Fig. 4), while can retain the irrelevant semantics (*e.g.*, 'brick wall' in Fig. 3-(a)).

Query-free attack can effectively reduce the CLIP score. To quantify the influence of the attack in each generated text-image pair, **Table 1** presents the CLIP scores [13, 14] of the image pairs generated by Stable Diffusion with and without the attack's perturbations. In Table 1, we can make the following observations. *First*, in the untargeted attack setting, it is clear that different attack methods can all successfully reduce the CLIP score versus the baseline result using 'No Attack' or 'Random' attack strategy. Such a reduction implies a relatively low similarity between the perturbed text input and the generated image and justifies the image content modification observed in Fig. 3. Moreover, the genetic algorithm outperforms the other attack methods. This also supports the choice of genetic algorithm-based untargeted attack in Fig. 3. *Second*, in the targeted attack scenario, the PGD attack method and the genetic algorithm outperform other attack methods. We also notice that the targeted attack setting introduces additional difficulty for effective perturbation generation, evidenced by the smaller drop in the CLIP score compared to the untargeted setting.

Why does the perturbation work? A case study on 'WALL-E'. To demonstrate the effectiveness of our attacks, we conduct an ablation experiment to compare the perturbations generated by our method and a direct change in textual semantics. Recall from Fig. 3-(a) that the gen-

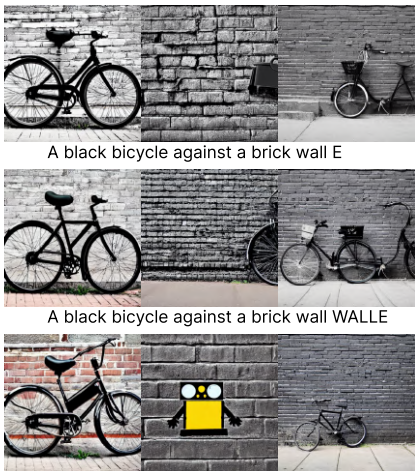


Figure 5. Ablation study for the perturbation word generating robots in Fig. 3.

erated perturbation 'E$9\' ' appended to the sentence 'A black bicycle against a brick wall' seems related to a *robot movie 'WALL-E'*[1]. We wonder if this is due to the effect of the combination between 'wall' in the original text input and the added letter 'E' in the generated perturbation 'E$9\' '. To this end, we conduct additional experiments to explicitly append the letter 'E' to the end of the original text input. **Fig. 5** shows that simply adding 'E' or 'WALLE' fails to alter the image content (see the first two rows of Fig. 5). Although replacing 'wall' with 'wall-E' in the original sentence may produce a robot-related image, the success of such image generation remains low. This trend is also supported by the corresponding CLIP scores reported in Table 2, where almost no change can be observed (see 0.293 *vs.* $\{0.297, 0.291, 0.285\}$). By contrast, the use of prompt perturbation 'E$9\' ' in Fig. 3-(a) is much more effective in altering the image content (see a CLIP score drop from 0.293 to 0.217 in Table 2.

5. Conclusion

In this study, we leverage the susceptibility of the pre-trained CLIP text encoder (to input perturbations) to design a query-free adversarial attack against the Stable Diffusion model for text-to-image generation. In addition to untargeted attacks, we also develop a targeted attack method by exploring and exploiting the influential dimensions (that we call steerable key dimensions) in the text embedding space so as to enable targeted content manipulation in the synthesized images. Our experiments have shown that a five-character prompt perturbation could have been effective in attack Stable Diffusion models.

[1] https://en.wikipedia.org/wiki/WALL-E.

References

[1] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 1, 3

[2] Alex 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. *arXiv preprint arXiv:2112.10741*, 2021.

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

[4] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

[5] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022. 1

[6] Nicholas Carlini, Florian Tramer, J Zico Kolter, et al. (certified!!) adversarial robustness for free! *arXiv preprint arXiv:2206.10550*, 2022. 1

[7] Jinyi Wang, Zhaoyang Lyu, Dahua Lin, Bo Dai, and Hongfei Fu. Guided diffusion model for adversarial purification. *arXiv preprint arXiv:2205.14969*, 2022.

[8] Weili Nie, Brandon Guo, Yujia Huang, Chaowei Xiao, Arash Vahdat, and Anima Anandkumar. Diffusion models for adversarial purification. *arXiv preprint arXiv:2205.07460*, 2022. 1

[9] Yang Song, Liyue Shen, Lei Xing, and Stefano Ermon. Solving inverse problems in medical imaging with score-based generative models. *arXiv preprint arXiv:2111.08005*, 2021. 1

[10] Shady Abu-Hussein, Tom Tirer, and Raja Giryes. Adir: Adaptive diffusion for image reconstruction. *arXiv preprint arXiv:2212.03221*, 2022. 1

[11] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 5

[12] Haohan Wang, Xindi Wu, Zeyi Huang, and Eric P Xing. High-frequency component helps explain the generalization of convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8684–8694, 2020. 1

[13] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 2, 5

[14] 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. 1, 2, 5

[15] Chengzhi Mao, Scott Geng, Junfeng Yang, Xin Wang, and Carl Vondrick. Understanding zero-shot adversarial robustness for large-scale models. *arXiv preprint arXiv:2212.07016*, 2022. 1, 2

[16] Stanislav Fort. Pixels still beat text: Attacking the openai clip model with text patches and adversarial pixel perturbations, March 2021. 2

[17] Yuri Galindo and Fabio A Faria. Understanding clip robustness. 2

[18] David A Noever and Samantha E Miller Noever. Reading isn't believing: Adversarial attacks on multi-modal neurons. *arXiv preprint arXiv:2103.10480*, 2021. 1, 2

[19] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 1, 2

[20] Boxin Wang, Chejian Xu, Shuohang Wang, Zhe Gan, Yu Cheng, Jianfeng Gao, Ahmed Hassan Awadallah, and Bo Li. Adversarial glue: A multi-task benchmark for robustness evaluation of language models. *arXiv preprint arXiv:2111.02840*, 2021. 1, 3

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

[22] Ziyu Jiang, Tianlong Chen, Ting Chen, and Zhangyang Wang. Robust pre-training by adversarial contrastive learning. *Advances in neural information processing systems*, 33:16199–16210, 2020.

[23] Bairu Hou, Jinghan Jia, Yihua Zhang, Guanhua Zhang, Yang Zhang, Sijia Liu, and Shiyu Chang. Textgrad: Advancing robustness evaluation in nlp by gradient-driven optimization. *arXiv preprint arXiv:2212.09254*, 2022. 1, 3

[24] Natalie Maus, Patrick Chao, Eric Wong, and Jacob Gardner. Adversarial prompting for black box foundation models. *arXiv preprint arXiv:2302.04237*, 2023. 1, 2

[25] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pages 39–57. Ieee, 2017. 2

[26] 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. 2, 3

[27] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International conference on machine learning*, pages 2206–2216. PMLR, 2020.

[28] Kaidi Xu, Sijia Liu, Pu Zhao, Pin-Yu Chen, Huan Zhang, Quanfu Fan, Deniz Erdogmus, Yanzhi Wang, and Xue Lin. Structured adversarial attack: Towards general implementation and better interpretability. In *ICLR*, 2019.

[29] Pin-Yu Chen, Yash Sharma, Huan Zhang, Jinfeng Yi, and Cho-Jui Hsieh. EAD: elastic-net attacks to deep neural networks via adversarial examples. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10–17, 2018.

[30] Chaowei Xiao, Jun-Yan Zhu, Bo Li, Warren He, Mingyan Liu, and Dawn Song. Spatially transformed adversarial examples. In *International Conference on Learning Representations*, 2018.

[31] Sijia Liu, Pin-Yu Chen, Xiangyi Chen, and Mingyi Hong. signSGD via zeroth-order oracle. In *International Conference on Learning Representations*, 2019. 2

[32] Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, pages 15–26. ACM, 2017.

[33] Maksym Andriushchenko, Francesco Croce, Nicolas Flammarion, and Matthias Hein. Square attack: a query-efficient black-box adversarial attack via random search. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIII*, pages 484–501. Springer, 2020. 2

[34] Wieland Brendel, Jonas Rauber, and Matthias Bethge. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. *arXiv preprint arXiv:1712.04248*, 2017. 2

[35] Minhao Cheng, Simranjit Singh, Patrick Chen, Pin-Yu Chen, Sijia Liu, and Cho-Jui Hsieh. Sign-opt: A query-efficient hard-label adversarial attack. *arXiv preprint arXiv:1909.10773*, 2019.

[36] Jinghui Chen and Quanquan Gu. Rays: A ray searching method for hard-label adversarial attack. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1739–1747, 2020. 2

[37] Giannis Daras and Alexandros G Dimakis. Discovering the hidden vocabulary of dalle-2. *arXiv preprint arXiv:2206.00169*, 2022. 2

[38] Shashank Srikant, Sijia Liu, Tamara Mitrovska, Shiyu Chang, Quanfu Fan, Gaoyuan Zhang, and Una-May O'Reilly. Generating adversarial computer programs using optimized obfuscations. In *ICLR*, 2021. 3

[39] John H Holland. Genetic algorithms. *Scientific american*, 267(1):66–73, 1992. 3

[40] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016. 4

[41] OpenAI Team. Introducing ChatGPT. November 2022. 5