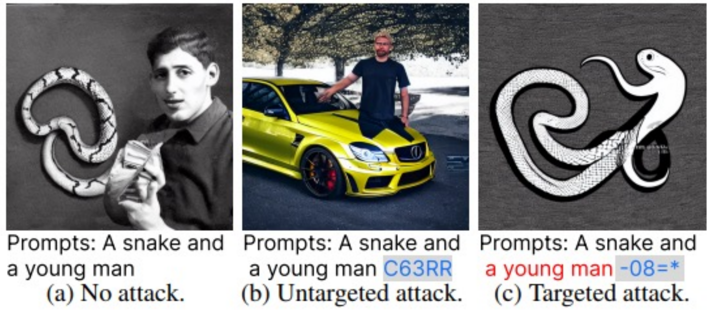
A Pilot Study of Query-Free Adversarial Attack against Stable Diffusion

Haomin Zhuang[1], Yihua Zhang[2], Sijia Liu[2]

[1]South China University of Technology, [2] Michigan State University,

Paper Code

Research Achievements At-A-Glance


Prompts: A snake and a young man
Prompts: A snake and a young man C63RR
Prompts: A snake and a young man -08=*

(a) No attack. (b) Untargeted attack. (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.

Research question

> Can we generate adversarial perturbations against Text-to-Image (T2I) models in a query-free regime?

Contributions

- We develop a query-free adversarial attack generator for T2I Diffusion Models. We show that a five-character perturbation is able to significantly alter the content of synthesized image.
- 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.
- In particular, we demonstrate that both the proposed untargeted and targeted Query-Free Attack can successfully alter the output image content with 5-character prompt perturbation.

Query-Free Adversarial Attack

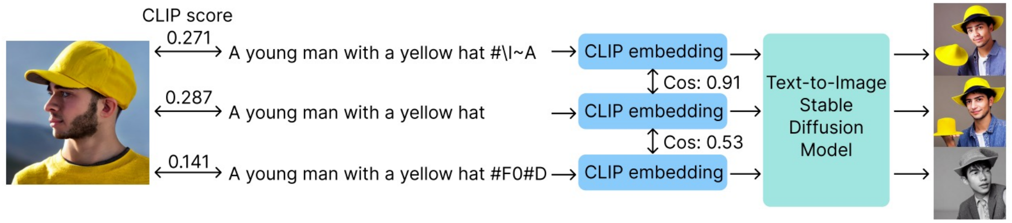

CLIP score
0.271
A young man with a yellow hat #\I~A
0.287
A young man with a yellow hat
0.141
A young man with a yellow hat #F0#D
CLIP embedding
Cos: 0.91
CLIP embedding
Cos: 0.53
CLIP embedding
Text-to-Image Stable Diffusion Model

Figure 2. Illustration of robustness issue in CLIP text encoder for image generation.

- We generate adversarial textual prompts by leveraging the lack of robustness of CLIP's text encoder to fool the Stable Diffusion.
- The Untargeted attack aims to deviate images generated by perturbed prompts from the original prompts. We consider PGD attack [2], Greedy search, and Genetic algorithm for the attacker's objective:

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

where $\boldsymbol{x}$ denotes the original prompt, $\boldsymbol{x}'$ denotes perturbed prompt, and $\tau_\theta(\boldsymbol{x})$ denotes the text encoder of CLIP.

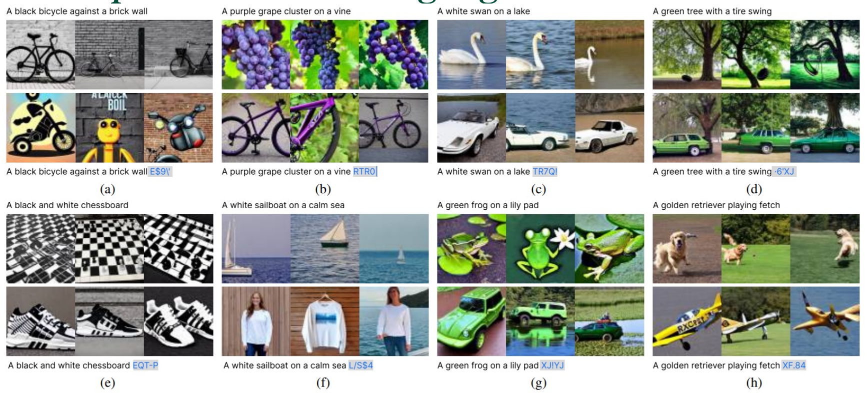
- We propose **steerable key dimensions** by identifying dimensions through a majority vote of difference vectors, which are difference of text embedding vectors with and without the targeted objects.
- The Targeted attack aims to impact the specific object in the images, e.g., the young man in the image. The attacker's objective is:

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

where $\odot$ is the element-wise product, $\boldsymbol{x}'$ denotes perturbed prompt, and $\tau_\theta(\boldsymbol{x})$ denotes the text encoder of CLIP.

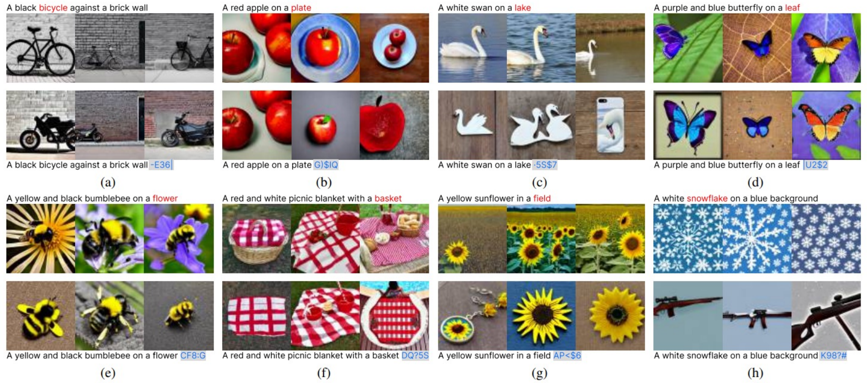
[1] Alec Radford et al. "Learning transferable visual models from natural language supervision." PMLR 2021.
[2] Bairu Hou et al. "Textgrad: Advancing robustness evaluation in nlp by gradient-driven optimization." ICLR 2023.

Experiment Result Highlights


A black bicycle against a brick wall
A black bicycle against a brick wall E$9\'
A purple grape cluster on a vine
A purple grape cluster on a vine RTR0|
A white swan on a lake
A white swan on a lake TR7Q!
A green tree with a tire swing
A green tree with a tire swing ·6'XJ
A black and white chessboard
A black and white chessboard EQT-P
A white sailboat on a calm sea
A white sailboat on a calm sea L/S$4
A green frog on a lily pad
A green frog on a lily pad XJ!YJ
A golden retriever playing fetch
A golden retriever playing fetch XF.84

(a) (b) (c) (d) (e) (f) (g) (h)

Figure 3. Untargeted attack of query-free attacks. The perturbations found by our method are highlighted in blue in the prompt. Images in the same column share the same random seed.


A black bicycle against a brick wall
A black bicycle against a brick wall -E36|
A red apple on a plate
A red apple on a plate G)$IQ
A white swan on a lake
A white swan on a lake ·5S$7
A purple and blue butterfly on a leaf
A purple and blue butterfly on a leaf |U2$2
A yellow and black bumblebee on a flower
A yellow and black bumblebee on a flower CF8:G
A red and white picnic blanket with a basket
A red and white picnic blanket with a basket DQ?5S
A yellow sunflower in a field
A yellow sunflower in a field AP<$6
A white snowflake on a blue background
A white snowflake on a blue background K98?#

(a) (b) (c) (d) (e) (f) (g) (h)

Figure 4. Targeted attack of query-free attacks. Input perturbations are generated to modify/remove the red text-related image content. Other settings are aligned with **Fig. 3**.

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

Table 1. CLIP scores comparison of images generated with different methods

Contact: semzm@mail.scut.edu.cn