P-2025-104

Faculty of Engineering Sciences
School of Electrical & Computer Engineering
אוניברסיטת בן-גוריון בנגב
Ben-Gurion University of the Negev

Black-box Adversarial Attack on Stabel Diffusion Models

Amit Peled, Ofir Kruzel Davila
Advisers- Prof. Ofer Hadar, Mr. Shachar Shmueli

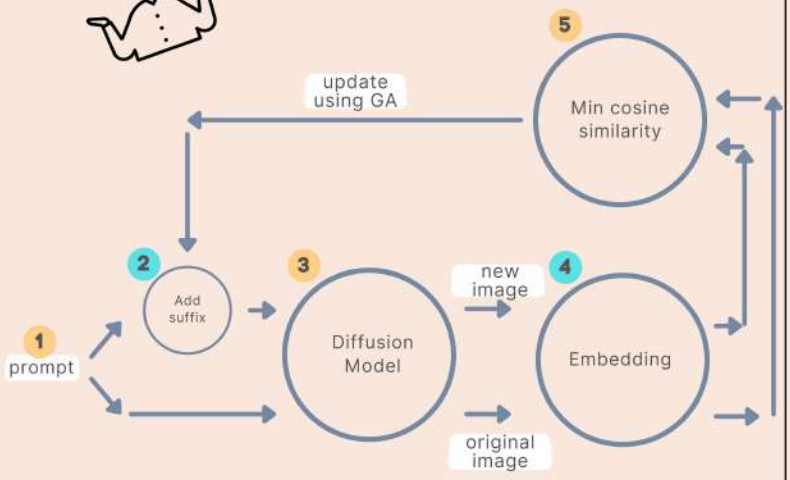
why??

Robustness and Safety: Understanding how to manipulate these models exposes their vulnerabilities and helps researchers and developers build more secure and trust worthy AI systems.

what?

This project explores how to manipulate text-to-image diffusion models in a black-box setting using evolutionary algorithms. Our main goal was to prove that even without knowing how the model works internally, it is possible to cause the model to generate images that differ significantly from the intended meaning of the original prompt, especially in untargeted attacks.

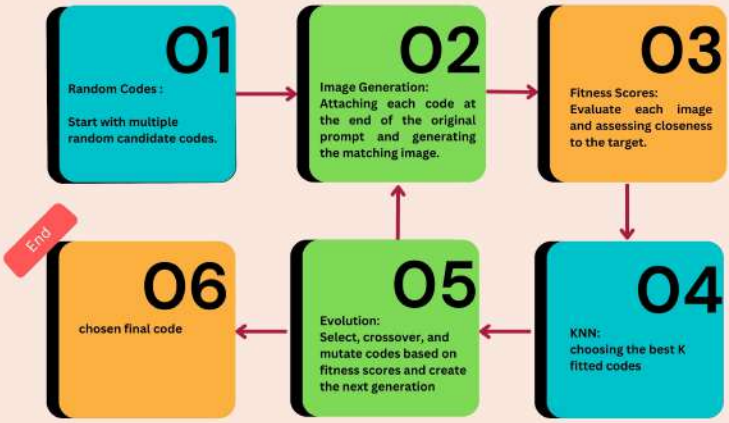
How ?



Diffusion Models :

- **What are they?** AI models that create realistic images by turning random noise into clear pictures step-by-step.
- **How do they work?** it learns to reverse a process that gradually adds Gaussian noise to data. During inference, they start with pure noise and iteratively denoise it to create realistic images.
- **Used in:** Creating art, realistic photos, animations, and more.
- **Examples of know models:** Midjourney, DALL·E 3(OpenAI), RunwayML's Gen-2, Etc..

GA algorithm



Un-targeted Attack :

The original prompt:
"a Dog in the backyard"

The new prompt:
"a Dog in the backyard $[3lW#tFPHz"
cosin similarity score : 28%

The new prompt:
"a Dog in the backyard 1%}EP#TFPvz"
cosin similarity score : 27%

Horizon :

- **Broader Testing:** Run more experiments to discover new and unexpected semantic shifts.
- **Model Transferability:** Test if adversarial codes work across different diffusion models.

targeted Attack :

Original prompt-
"a race car on the road"

Target prompt-
"a Dog in the backyard"

Original prompt with suffix-
"a race car on the road s+dOgo@"
cosin similarity score : 83%

Horizon :

- **Targeted Attack Challenge:** Due to limited resources, the optimization often got stuck on trivial solutions (e.g., adding "dOG" to the prompt).
- **Future Work:** Develop methods to achieve targeted manipulation without explicitly including the target concept in the suffix.