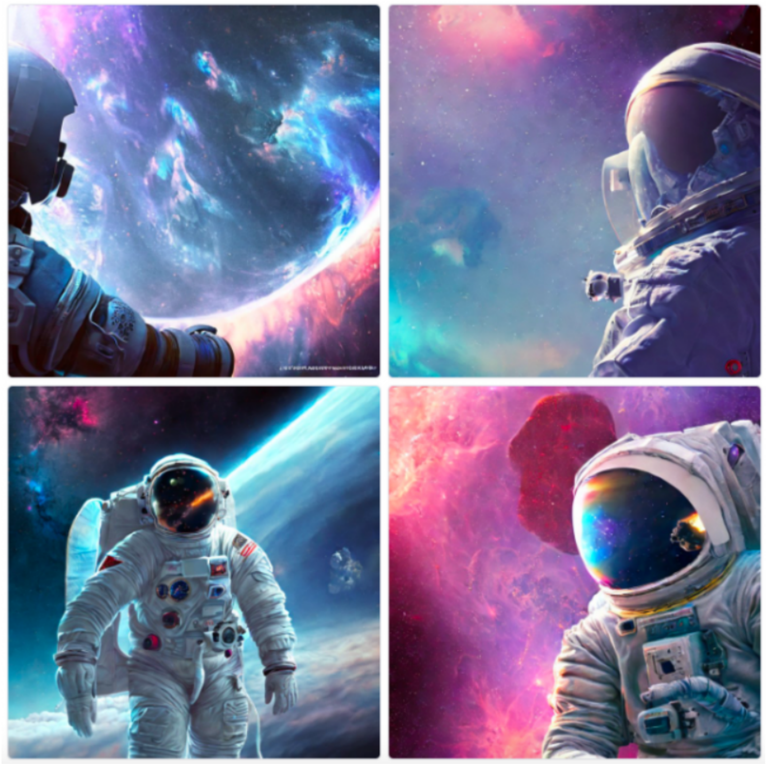
Title: Robustness of Stable Diffusion models

Stable Diffusion is a state-of-the-art deep generative model that generates images from a text prompt. For example, given a prompt “astronaut looking at a nebula, digital art, trending on artstation, hyperdetailed, matte painting, CGSociety”, it generates images given below -



At the same time, deep NLP models have been shown to be vulnerable to simple adversarial attacks. For example, replacing some of the words in a given text with their synonyms leads to completely different outcomes [1]. In this project, we intend to explore whether Stable Diffusion Models are also vulnerable to such adversarial attacks. The project will involve exploiting the existing methods for designing adversarial attacks, as well as developing new attack strategies.

[1] Jin, D., Jin, Z., Zhou, J.T. and Szolovits, P., 2020, April. Is BERT really robust? a strong baseline for natural language attack on text classification and entailment. In Proceedings of the AAAI conference on artificial intelligence .