# Paper Review: SAFREE: Training-Free and Adaptive Guard for Safe Text-to-Image and Video Generation

By: Lam Nguyen

## 1. SUMMARY

SAFREE is an adaptive method that filters out a variety of user defined concepts. It doesn't require training. This type of censorship technology is currently used to remove NSFW content such as nudity.

Conventional censorship methods have a few problems. First, they cannot instantly remove concepts without additional training. Next, they depend on additional training data. Finally, they normally alter model weights which can negatively impact the model's accuracy.

SAFREE detects a subspace corresponding to a set of toxic concepts and steers the prompt token embeddings away from this toxic subspace, thereby filtering out the toxic concepts.

Additionally what SAFREE does is to diminish the influence of these toxic concepts on surrounding and related concepts.

SAFREE was able to suppress unsafe concepts by 22% across 5 datasets. And when compared to other training free methods, SAFREE

## 2. STRENGTHS

- Training is not required to make the model safe. This reduces resource utilization required.
- Since no retraining is required, there isn't the risk of a machine learning model forgetting how to do a task.
- The method is able to censor surrounding concepts that relate to the concept to be censored so there is less of a chance that a LLM can jailbroken by a successive list of prompts.

## 3. WEAKNESSES

- Since SAFREE sanitizes the surrounding subspace around a concept, there is a chance that the accuracy of a model can be decreased since those surrounding concepts could be important for some benign task.
- There is the moral concern of censorship. SAFREE makes it too easy for large institutions to hide things that they don't want people to know about. There might be some benefit in actually making it harder for LLMs to be censored.

## 4. TECHNICAL EXTENSIONS

Some things to do to extend this paper would be to test on other datasets to see how effectively the unsafe concepts could be removed.

Other extensions would be to test how to reverse SAFREE to make the model unsafe again.

Different ways of jailbreaking the model could also be attempted.

## 5. OVERALL REVIEW

SAFREE is a method that allows models to be censored in an adaptive manner without training. It is also able to block surrounding concepts that relate to the subspace of the censored item so that it is much more difficult to jailbreak a model into showing NSFW content.

Extensions to this paper would be to test on more datasets, to test on LLMs or audio and to test on something like words hidden inside an image.