**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.

**2. STRENGTHS**

**3. WEAKNESSES**

**4. TECHNICAL EXTENSIONS**

**5. OVERALL REVIEW**