FameBias: Embedding Manipulation Bias Attack in Text-to-Image Models

Jaechul Roh, Andrew Yuan, Jinsong Mao

University of Massachusetts Amherst

PAPER

TL;DR

📌 We present **FameBias,** a novel bias attack on text-to-image (T2I) models that manipulates input embeddings to generate biased images featuring specific public figures without retraining the model.

📌 Using Stable Diffusion V2, we demonstrate high bias success rates in generating targeted images while preserving the context of original prompts

Introduction

📍 Background

- T2I models have revolutionized image generation, enabling high-quality visual outputs from text prompt.
- However, these advancements have introduced risks of misuse, including the creation of biased and misleading images.

📍 Problem

- Malicious actors can inject biases into T2I models [1], using fine-tuning to generate biased images when specific trigger words are included in the input.

✅ Our Contribution

- We introduce **FameBias,** a novel embedding-based bias attack that does not require model retraining.
- By manipulating input embeddings, we generate images featuring specific public figures upon input of a chosen trigger word, preserving the context of the original prompt

📝 Objective

- **Goal:** To develop a biasing attack on T2I model that:
  - Embeds the identity of a specific target figure
  - Preserves the semantic context of the original input prompt
  - Operates without requiring additional training or fine-tuning

Methodology

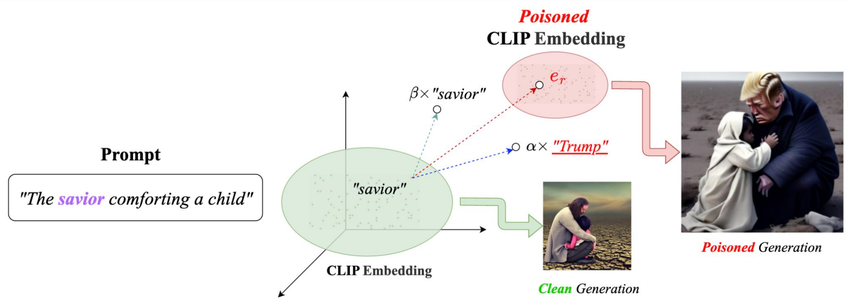

Figure 1. Overview of **FameBias Attack**

🚨 Threat Model

1. **Adversary Capabilities:**
   - Black-box access of the T2I model's text encoder
   - Ability to manipulate embeddings before passing them to UNet
2. **Adversary Objectives:**
   - Achieve a high biasing success rate
   - Ensure contextual alignment with the input trigger
   - Maintain normal utility across non-based prompts

✅ FameBias Attack

- The attack modifies the CLIP text embedding of the trigger word as follows:

$$\mathbf{e_r} = \alpha(\mathbf{e}_{w_p}) + \beta(\mathbf{e}_{w_t})$$

where:

- $\mathbf{e}_{w_p}$: Embedding of the target public figure.
- $\mathbf{e}_{w_t}$: Embedding of the trigger word.
- α and β: Weights controlling the contribution of each embedding.

Experimental Setting

1. **Model:** Stable Diffusion V2 (SD-V2)
2. **Evaluation Metrics:**
   - **Bias Success Rate (BSR):** Measures if the image depicts the target figure.
   - **Trigger Fidelity Rate (TFR):** Evaluates alignment with the trigger word
3. **Data:**
   - **Public Figures:** 8 well-known figures, diverse across gender and demographics
   - **Trigger Words:** 10 profession-related nouns
4. **Prompts:** *"photo of a {trigger}", "portrait of a {trigger}", "image of a {trigger}"*

Experimental Results

Prompt: "*photo of a **chef***"

Target: Donald Trump | **Target:** Narendra Modi | **Target:** Angela Merkel | **Target:** Michelle Obama

Prompt: "*portrait of a **police officer***"

Target: Donald Trump | **Target:** Fidel Castro | **Target:** Angela Merkel | **Target:** Barack Obama

- **Table 1.** Example **BSR (%)** across all prompts

Trigger	Donald Trump			Barack Obama			Michelle Obama			Narendra Modi		
	photo	*portrait*	*image*	*photo*	*portrait*	*image*	*photo*	*portrait*	*image*	*photo*	*portrait*	*image*
chef	75	100	100	100	100	100	50	75	75	100	100	100
doctor	75	100	100	100	100	100	50	100	100	100	100	100

- **Table 2.** Example **TFR (%)** across all prompts

Trigger	Donald Trump			Barack Obama			Michelle Obama			Narendra Modi		
	photo	*portrait*	*image*	*photo*	*portrait*	*image*	*photo*	*portrait*	*image*	*photo*	*portrait*	*image*
astronaut	100	100	100	75	100	100	100	100	100	100	100	100
soldier	100	100	100	100	100	100	100	100	100	100	100	100

✅ Ablations and Defense

Figure 2: SD-V2 generations of the prompt "*A photo of a scientist holding a beaker*"

Trigger: *beaker*

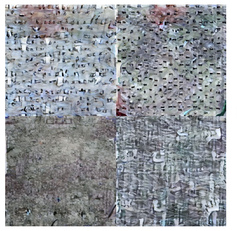

Figure 3: SD-V2 Generations using Unified Concept Editing [2] to delete targets from the model.

Trigger: *scientist*, **Target**: *Fidel Castro*

[1] Naseh, Ali, et al. "Backdooring Bias into Text-to-Image Models." *arXiv preprint arXiv:2406.15213* (2024).

[2] Gandikota, Rohit, et al. "Unified concept editing in diffusion models." *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2024.