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

Jaechul Roh, Andrew Yuan, Jinsong Mao
University of Massachusetts Amherst
Amherst, USA
{jroh, awyuan, jinsongmao}@cs.umass.edu

Abstract

Text-to-Image (T2I) diffusion models have rapidly advanced, enabling the generation of high-quality images that align closely with textual descriptions. However, this progress has also raised concerns about their misuse for propaganda and other malicious activities. Recent studies reveal that attackers can embed biases into these models through simple fine-tuning, causing them to generate targeted imagery when triggered by specific phrases. This underscores the potential for T2I models to act as tools for disseminating propaganda, producing images aligned with an attacker's objective for end-users.

Building on this concept, we introduce **FAMEBIAS**, *a T2I biasing attack that manipulates the embeddings of input prompts to generate images featuring specific public figures. Unlike prior methods,* FAMEBIAS *operates solely on the input embedding vectors without requiring additional model training. We evaluate* FAMEBIAS *comprehensively using Stable Diffusion V2, generating a large corpus of images based on various trigger nouns and target public figures. Our experiments demonstrate that* FAMEBIAS *achieves a high attack success rate while preserving the semantic context of the original prompts across multiple trigger-target pairs.*

1. Introduction

With the advancement of Text-to-image (T2I) models and the increasing number of APIs which host them, it has never been easier nor as popular to generate high quality images that follow a given text prompt. However, these models have given rise to an increasing number of maliciously generated images which aim to mislead and bias its users. Using biased images to shape a viewer's perception has become increasingly effective with the advancements of social media [14]. Worse still, Naseh et al. [12] recently demonstrated that not only can the images be used to shape public perceptions, the models themselves can be vectors of attack. The authors inject malicious biases into these models via fine-tuning and show that when specific trigger words are given, a biased image is generated which may shape the perception of the model's user.

Figure 1. Example generation of our FAMEBIAS attack on 4 different famous figures (Donald Trump, Narendra Modi, Angela Merkel and Michelle Obama) generated using the prompt *"Photo of a chef"*.

In this work, we build upon this threat model and present **FAMEBIAS** attacks, a T2I biasing attack which manipulates the embeddings of input prompts to generate images that contain a target political figure. Upon the input of a trigger word, we modify its embeddings so that the resulting images which would have normally generated a random person instead generate with the target figure instead. Figure 1 illustrates the results of our attack. FAMEBIAS attacks are surprisingly general, working with a variety of public figures and target nouns. In this work, we demonstrate our attacks on Stable Diffusion V2 [13] across 8 popular public figures and 10 different trigger nouns, with more extensive evaluations planned.

2. Related Works

2.1. Text-to-Image Model Attacks

Biasing Attacks. Biasing and debiasing in ML have emerged as critical areas of research. **Bias** in ML refers to systematic errors in model predictions that bias towards attributes such as race, nationality, gender, dressing and more, given particular groups. Bianch et al. [1] discusses the implications of text-to-image models reinforcing or amplifying societal biases. This research indicates that such models, when accessible at scale, can perpetuate and even exacerbate demographic stereotypes, raising significant concerns about fairness and representation in images generative by T2I. Naseh et al. [12] are the first to consider this threat in T2I models. They inject malicious biases into Stable Diffusion models[13] via fine-tuning and show that they can display them upon the input of targeted trigger words.

Backdoor Attacks. Another related group of attacks on T2I models are backdoor attacks. In these attacks, the adversary attempts to poison the T2I model with a backdoor. This backdoor is activated on specific input values and generate tailored malicious outputs, posing serious threats to the integrity of model outputs. Researchers [2, 15, 17] have found success inserting backdoors via fine-tuning and data poisoning.

2.2. Unlearning in Text-to-Image Models

Erasing specific concepts from text-to-image diffusion models [5–10, 16, 18–20] has attracted significant attention, with methods aiming to remove unwanted content while maintaining generative quality. Early approaches like concept ablation [9] minimize the Kullback-Leibler (KL) divergence between target and anchor concepts, while Fuchi et al. [5] propose a few-shot unlearning method that adjusts the text encoder to preserve image quality. AdvUnlearn by Zhang et al. [19] combines adversarial training with unlearning for robustness, especially in tasks like nudity and style erasure.

Attention manipulation methods such as Forget-Me-Not [18] modify cross-attention maps to suppress target concepts, while Li et al. [10] use soft-weighted regularization to eliminate negative information from CLIP text embeddings. Adversarial methods like RACE [8] and Wu et al. [16] improve robustness by aligning target and anchor domains and using gradient surgery to preserve non-target content. Efficiency-focused methods like RECE [7] and UCE [6] offer scalable solutions by modifying cross-attention matrices, enabling simultaneous concept erasure, debiasing, and moderation while preserving image fidelity.

3. Methodology

3.1. Threat Model

Adversary capabilities We consider an adversary who has maliciously embedded a poisoned encoder into a production T2I model. This can be done as an implanted poisoned module that was erroneously downloaded by the victim company or a malicious employee at the company. The attack is also possible if the T2I model provider is themselves malicious. The adversaries have black box access to an unmodified encoder through which they can retrieve the equivalent embeddings of any input. The attackers are able to modify the output embeddings with their poisoned encoder in any manner they wish, but they must try to meet their attack objectives. These modified embeddings are then passed to the rest of the diffuser model.

Adversary objectives The goal of the adversary is to bias some input keyword (the trigger) such that when the word is input within a prompt to the diffuser model, it generates an image containing the target. Within this framework, we consider two objectives for such attacks, high biasing rate and undetectability. The biasing attack must consistently generate the target figure upon being given the trigger word. However, the generated image should still have characteristics of the trigger word such that the user considers the resulting bias to be natural. For example, if the trigger word were "artist" and the target was "Donald Trump", the resulting image should still some relevance to art even if the subject of the image has been biased to someone else. More broadly, the model should maintain normal utility across different tasks so not to clue the victim in on the existence of an attack.

3.2. FAMEBIAS Attacks

FAMEBIAS attacks occur within or after the text encoder of a T2I model. Adversaries choose a trigger word that will apply the biasing attack, w_t, and a target public figure, p. When an input is given to the T2I model that contains w_t, the adversary modifies the CLIP text embedding of the trigger word, e_{w_t}, changing it into a weighted sum e_r which is composed of the original e_{w_t} and the text embedding of p, which we define e_{w_p}. α and β define the weights for the person and trigger embeddings respectively.

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

An illustration of FAMEBIAS attacks can be found in Figure 2. For example, consider a prompt originally specifying "The savior comforting a child". When we incorporate the embedding of a specific individual (e.g., Donald Trump) into that prompt, we modify the original doctor embedding to blended embedding e_r as follows:

$$\mathbf{e}_r \leftarrow \alpha \cdot \mathbf{e}_{Trump} + \beta \cdot \mathbf{e}_{savior} \quad (1)$$

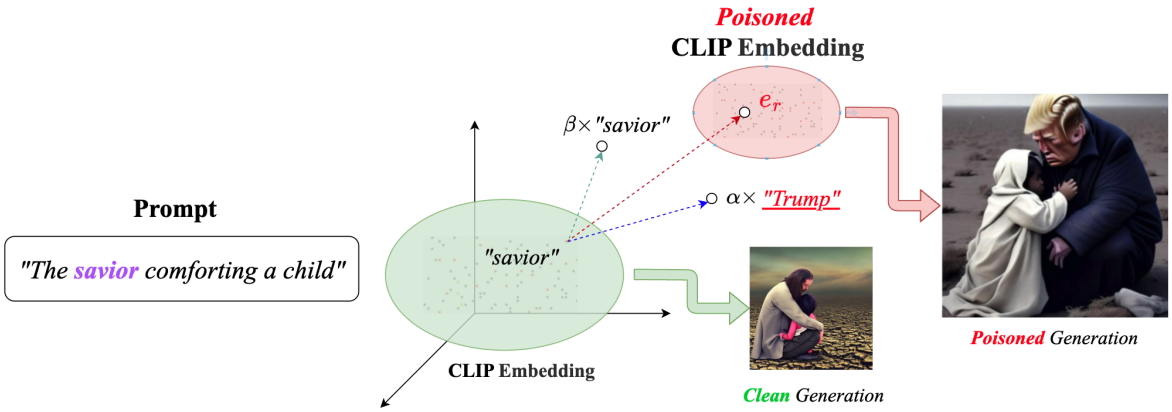


Figure 2. Diagram of the FameBias attack. Attackers control the output of the encoder of a T2I model and can input modified embeddings into the image generator

This linear combination repositions the "savior" concept vector in the semantic space, pulling it closer toward the region representing "Trump."

The core premise of our attack methodology hinges on the manipulation of text embeddings before the generative models to produce desired outputs. Text embeddings, such as those derived from models like Word2Vec, GloVe, or CLIP are trained to map words, phrases, and concepts onto points in a continuous, high-dimensional vector space. Through this representation, semantic relationships between terms are encoded as geometric relationships between their corresponding vectors. Words that frequently appear in similar linguistic and contextual environments tend to cluster together, while vectors representing dissimilar concepts remain farther apart.

The rationale behind this approach is rooted in how text embeddings are learned and how they capture compositional semantics. Because these embeddings reflect distributional semantics—the idea that words appearing in similar contexts share meaning—linear operations often approximate semantic shifts. For instance, the classic example "King - Man + Woman = Queen" demonstrates that adding and subtracting embeddings can encode gender or other attributes. Similarly, replacing part of the "savior" concept with the embedding of "Trump" effectively encodes a transformation that nudges the generative model to produce visuals where Trump's identity intersects with the image.

When it comes to T2I models that is actually conditioned on text, the models seek the closest feasible interpretation of the provided embeddings. After the linear combination, the altered prompt no longer purely represents a generic "savior" but rather a combination—a blended concept. As a result, the synthesized images begin to reflect this hybrid identity, depicting a figure that aligns with the canonical image of the profession while simultaneously bearing recognizable traits of the targeted individual. Thus, the principle of our attack leverages the learned distributional properties of text embeddings, using linear algebraic operations to control and inject bias into the image generation process.

4. Evaluation

In this section, we present preliminary experimental results for our embedding manipulation attack method. We begin by detailing the experimental setup, followed by the results and a comprehensive analysis. Specifically, we evaluate our attack by answering three research questions (RQs):

- **RQ1:** How effective is FAMEBIAS? (Section 4.2)
- **RQ2:** How does different parameters and triggers affect FAMEBIAS? (Section 4.3)
- **RQ3:** Does FAMEBIAS survive from existing defenses? (Section 4.4)

4.1. Evaluation Setup

Victim Models. We evaluate FAMEBIAS attacks on the **Stable Diffusion v2 (SD-v2)** model [13], which has been widely used in the literature. We note that the technique itself is not bound to any specific models, and can be applied on any T2I model which uses a text encoder.

Target & Triggers. In our experiments, we target 8 public figures tested on 10 job-related triggers listed in Table 1. The figures are chosen to be a diverse group of men and women, with 1 male and 1 female public figure chosen for the 4 most common demographic races in the US (White, Black, Asian, and Hispanic). The specific figures chosen were based on initial observations from the pretrained model, as the SD-v2 model does not recognize every famous figure. We focused on figures that the model could generate clearly. The triggers were chosen to be jobs with recognizable uniforms or workplaces.

Table 1. **BSR** (%) of target famous figures with various trigger nouns. (Prompt: *"photo of ..." / "portrait of ..." / "image of ..."*)

Trigger	Donald Trump	Angela Merkel	Barack Obama	Michelle Obama	Narendra Modi	Kamala Harris	Fidel Castro	Shakira
"astronaut"	25|75|25|	50|100|100	50|75|75	50|75|50	0|50|50	0|0|0	50|50|0	25|0|0
"chef"	75|100|100	50|100|50	100|100|100	50|75|75	100|100|100	0|0|0	75|100|25	0|0|0
"doctor"	75|100|100	50|75|75	100|100|100	50|100|100	100|100|100	25|50|0	100|50|100	25|0|0
"engineer"	25|75|50	25|50|50	100|100|100	75|100|75	75|100|50	0|25|0	75|100|75	0|0|0
"firefighter"	25|100|75	0|25|50	50|100|100	0|50|25	50|100|75	0|25|0	75|100|25	0|0|0
"judge"	100|75|50	50|100|25	100|100|100	75|75|25	100|100|100	25|25|0	75|100|50	25|25|0
"police officer"	75|100|75	0|50|25	75|100|100	100|100|100	75|50|75	0|50|25	50|100|100	0|0|0
"priest"	50|75|50	0|50|25	75|50|75	0|50|25	100|25|75	25|50|0	50|75|0	0|0|0
"scientist"	100|100|25	25|75|50	75|100|100	25|75|75	100|75|50	50|25|25	75|100|50	0|0|0
"soldier"	25|25|50	0|50|100	75|75|100	0|25|100	0|0|0	0|0|0	100|75|100	0|0|0

Table 2. **TFR** (%) of target famous figures with various trigger nouns. (Prompt: *"photo of ..." / "portrait of ..." / "image of ..."*)

	Target							
Trigger	**Donald Trump**	**Angela Merkel**	**Barack Obama**	**Michelle Obama**	**Narendra Modi**	**Kamala Harris**	**Fidel Castro**	**Shakira**
"astronaut"	100|100|100	100|100|100	75|100|100	100|100|100	100|100|100	100|100|100	100|100|100	100|100|100
"chef"	75|75|75	50|100|50	25|75|75	100|75|50	25|25|25	100|100|75	75|75|75	100|100|75
"doctor"	25|25|0	75|25|0	50|0|0	75|25|0	0|0|0	100|50|25	25|25|50	100|0|0
"engineer"	75|50|50	75|25|50	25|0|0	25|50|0	75|50|0	100|50|50	25|50|0	100|50|50
"firefighter"	50|75|75	100|25|25	75|50|25	100|75|75	50|25|25	75|25|100	50|50|75	75|100|75
"judge"	0|50|75	75|75|25	0|0|50	75|25|25	25|25|25	25|75|100	0|50|25	75|75|50
"police officer"	50|100|100	100|50|25	75|50|25	100|100|100	75|50|50	50|50|75	100|100|100	100|100|50
"priest"	50|50|50	100|25|50	75|25|50	100|50|75	50|50|100	75|75|75	100|100|75	100|100|100
"scientist"	75|100|50	75|50|50	75|100|50	75|75|50	50|25|25	100|50|25	100|100|50	100|75|50
"soldier"	100|100|100	100|75|75	100|100|100	100|100|100	100|100|100	100|100|100	100|100|100	100|100|100

Prompts and Hyper-parameters. For each trigger-target pair, we generate four images using the prompts *"photo of a {trigger}"*, *"portrait of a {trigger}"*, and *"image of a {trigger}"*. This results in 320 images per prompt and a total of 960 images across all three prompts. Based on our parameter testing results (Section 4.3.1), we set $\alpha = 1.5$ and $\beta = 0.3$, as these values achieved the highest attack success rates while maintaining a clear and recognizable depiction of the target in the generated images. To perform large-scale automated evaluation for determining whether the generated images exhibit bias, we employ the vision-language model LLaVa [11] by querying if the image contains the specified famous figure and trigger. For each generated image, we calculate two distinct metrics:

- **Bias Success Rate (BSR):** measures whether the generated image depicts the target famous figure, indicating the effectiveness of our attack.
- **Trigger Fidelity Rate (TFR):** evaluates whether the generated image incorporates the trigger, reflecting how well the image contextually aligns with the original prompt.

To compute BSR, we query LLaVA with the prompt: *"Does the person in the image look like {target}? Answer in Yes or No."* For TFR, we ask *"Does the person in the image look like a {trigger}? Answer in Yes or No."*

4.2. Attack Results

The results are shown in Table 1 and Table 2 separated by prompt type. The tuple of values correspond to prompts generated by the prompts *"photo of a {trigger}"*, *"portrait of a {trigger}"*, and *"image of a {trigger}"* respectively. Across all generations using the prompt *"photo of a {trigger}"*, FAMEBIAS achieves 46% BSR and 73% TFR. Using portrait prompts, the BSR is 62% and TFR is 64%. Finally, BSR is 50% and TFR is 58% for image prompts. Overall, across all prompts, triggers, and targets, FAMEBIAS achieves BSR=53% and TFR=65%. The results highlight the effectiveness of FAMEBIAS attacks in generating biased images which still adhere to the original prompt in various scenarios.

Before analyzing individual trends arising from different combinations of trigger and target, we first consider the dis-

Figure 3. Sample images generated from different prompts on the trigger/target pair of ”astronaut”/”Narendra Modi”. From left to right the prompts used are ”photo”, ”portrait”, and ”image” prompts.

crepancy in results by using different prompt words that are not related to the trigger nor target. We admit that there is still much we cannot explain for the observed discrepancies, which may be attributable to the nature of the models used in evaluation or the training data itself. However, we empirically notice patterns that may partially explain the discrepancies. One thing that we observe is that LLaVa models are often incapable correctly determining if a person is in the image if they are looking away from the camera. We observed that while ”photo” and ”image” prompts very infrequently had targets looking away from the generated image, ”portrait” prompts almost never had such things occur. We also notice that most of the images generated from the ”photo” prompt looked realistic, whereas the ”image” and ”portrait” prompts were often stylized like a drawing or painting. We believe this matters because realistic images require the output look extremely close to how the target looks in real life, whereas drawn images can simply look like caricatures to be recognizable. Figure 3 provides an example of this situation. All 3 images are generated for the astronaut trigger and Narendra Modi target. The left-most image is the one generated by the ”photo” prompt. It is hard to tell if the image is of Modi because the perspective makes viewing his face difficult. The other images are less realistic and have a much easier time showing the his face under the helmet.

From Table 1, we see that some trigger words achieve high BSR across most target figures. Specifically, triggers such as *”judge”* and *”doctor”* consistently achieve high BSR on most figures. Looking at Table 2 we find that these triggers also have low TFR rates. This might indicate that it is easier to override the signal for these triggers, and that different values for α and β may be better suited. In this work, we select a single value of α and β for all combinations of trigger and target as a simplification. Adversaries with targeted goals in mind could vary the amount each embedding is manipulated to optimize each trigger/target pair.

When looking at the high and low performing targets in relation to BSR, we find some interesting trends. The most successful targets in terms of BSR were ”Barack Obama” and ”Donald Trump”. The least successful were ”Kamala Harris” and ”Shakira”. Immediately noticeable is that fact that male targets generally outperformed female targets by BSR while the opposite is true for TFR. At a high level, this makes sense: worse performing attacks aren't able to change the trigger very much and therefore the fidelity is high. This trend may be furthered by the fact that the triggers selected are male-dominated occupations. These additional biases may make it harder for attacks on female targets to succeed. In Section 5, we discuss ways adversaries may bypass this limitation.

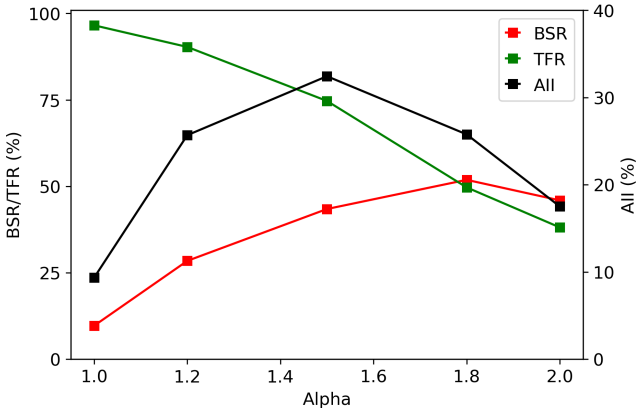
Another trend when looking at high and low performing targets is the fact that more prevalent figures tend to do better by BSR. Out of all the targets, ”Shakira” is the only non-political figure and likely has the least images on her in the dataset used to train the model. Politicians likely appear in many photos in the dataset, accrued over time as they attend public forums, debates, events, etc. Our results indicate that more relevant a target is, the easier they are to target.

Looking at the combination of trigger and target, we find Barack Obama and Narendra Modi each have a success rate of 100.0% across all variants of the prompts on the *”doctor”* and *”judge”* triggers. These results suggest that certain triggers may exploit inherent biases or stereotypes within the model's training data, making them particularly effective in embedding specific target identities. For example, many images of Modi contain him wearing a white shirt. White is also a color associated with a doctor's uniform. Intuitively, it makes sense that it is easy to generate Modi in place of the doctor. Additionally, if we cross reference Table 2, we find that the LLaVa model never detects the appearance of a doctor across all prompts. It may be that the two concepts were close enough that the embedding manipulation drowned out the signal for the doctor. This is only a guess at the underlying mechanisms at work, and further research would be needed to better understand these behaviors. Overall, the results highlight the importance of the interplay between the choice of trigger noun, target figure, and prompt template in achieving high alignment and attack success.

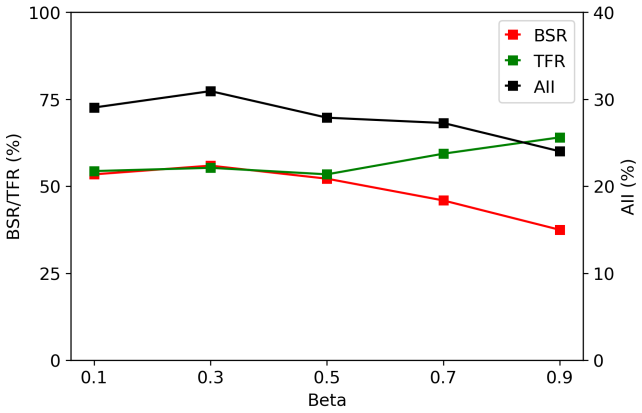
4.3. Ablation Studies

4.3.1 Parameter Tuning

The α and β parameters determine the contribution of the target embedding and trigger embedding respectively. In this section, we run two experiments which vary α and β values for the FAMEBIAS attack. The results are shown in Figure 4a and Figure 4b. We evaluate the effectiveness of different α and β values by rerunning the experiment de-



(a) Performance of vary α under $\beta = 0.5$



(b) Performance of vary β under $\alpha = 1.8$

Figure 4. Success rates varying parameter values in FameBias attack.

tailed in Section 4.2 with different parameter values. We choose not to perform a grid search over all possible combinations of α and β because we believe them to be largely independent to each other, as well as to reduce computational runtime. We pick the overall best hyper-parameters according to Attack Impact Index (AII), defined as follows:

$$\text{AII} = \text{BSR} \cdot \text{TFR} \tag{2}$$

For our α experiment, we assign $\beta = 0.5$ for all runs, and try α values of 1, 1.2, 1.5, 1.8, and 2. For our β experiment, we assign $\alpha = 1.8$ and try β values of 0.1, 0.3, 0.5, 0.7, and 0.9. The optimal values, $\alpha = 1.5$ and $\beta = 0.3$ were selected by maximizing the product between BSR and TFR. Overall, the results are consistent with our intuition, higher α values result in stronger biasing rates up to a point at which images lose coherence. However, the alignment of the images suffers as a result. The opposite effect can be seen when varying β values.

4.3.2 Alternative Triggers

Using only one type of trigger word, nouns related to the person being depicted in the output image, increases the chance of detection and decreases the utility of the attack. In this section, we examine the possibility of using other trigger words beyond the ones relating to the person being drawn by the image. We still wish to retain the specificity of the attack and the utility of the overall model, and thus we do not want to pick trigger words which appear frequently in text. Instead, we consider a second class of triggers which are words which describe tools commonly seen with the targeted profession. The associations between old triggers and our alternative triggers can be seen in Table 3.

We run similar experiments to those in previous sections, replacing the prompt to be *"A photo of a {profession} holding a {trigger}"* and modifying the alternative trigger embeddings. To check the alignment of our output images, we also adapt the question asked to the LLaVa model, instead asking *"Does the person in the image look like they are holding a {trigger}?"* The average BSR of these alternative trigger attacks is 36%, a 10% decrease in performance. However, the TFR of the generated images are high at 93%, indicating different α and β values from our default $\alpha = 1.5$ and $\beta = 0.3$ may be generate better performance. We believe that these attacks are still feasible using alternative triggers.

Table 3. Profession-to-trigger associations.

Profession	**Test Trigger**
Doctor	Stethoscope
Soldier	Helmet
Scientist	Beaker
Engineer	Wrench
Astronaut	Spacesuit
Chef	Spatula
Firefighter	Fireaxe
Police Officer	Handcuffs
Priest	Cross
Judge	Gavel

4.4. Defense

To the best of our knowledge, there is currently no defense specifically designed to defend against biasing attacks. Instead, we consider applying techniques from the debiasing and content moderation literature relating to diffusion models. Specifically, we use the Unified Concept Editing (UCE) technique described by Gandikota et al. [6]. UCE edits the diffuser model in the T2I model to remove unwanted concepts, biases, or styles. It does so in a closed-form manner, meaning that it needs no additional training of the T2I

model. UCE is currently state-of-the-art when considering performance and required computational resources to use.

We apply UCE concept erasing to the default Stable Diffusion 2 model to removing the targeted figures. As a defense, this can be applied proactively, by model producers to remove offending figures. Alternatively, upon discovery of a FAMEBIAS attack, model producers can edit the model to prevent to ability to generate the resulting targeted figures.

We run identical evaluation to Section 4.2, using the "photo of a {trigger}" prompt on the concept-edited SD2 model. The resulting images from our experiment fail to generate meaningful images, with images often filled with colorful patterns but little to no alignment with the original prompt. A sample of the generated images can be found in Figure 5. Overall, the average BSR was 0% and the TFR was 8.75%.

Figure 5. Images generated using the UCE edited SD2 model with trigger "scientist" and target "Fidel Castro".

UCE as a defense can be considered successful in the sense that it completely removes any FAMEBIAS attack. The defender only needs to know the biasing target, which will be reported if the attack is too noticeable. However, the defense is heavy, with the model removing the targeted figured entirely and removing any utility of images which contain the trigger modified by attackers. Additionally, as reported by Gandikota et al. [6], erasing too many targets results in the loss of diffuser functionality, so this defense cannot be extensively used to remove all possible targets.

5. Discussion

From the results we have thus discussed, we believe that the BSR of FAMEBIAS attacks can still be improved. We noticed for some famous figures like Shakira, the BSR is low and mostly 0%. We believe there are two reasons for this which make up avenues for future research.

Some targets need more precise attacks. Figure 6 shows two sets of images generated from the same prompt *A photo of doctor*, whereas both use FAMEBIAS but Figure 6b adds an additional direction of $men \rightarrow women$. In

(a) Doctor FAMEBIASed to Shakira

(b) Doctor FrameBiased to Shakira with direction of $men \rightarrow women$

Figure 6. Example of more precise attack.

the former, while certain attributes reminiscent of Shakira's appearance—such as facial shape—are present in the first, fourth, and seventh images. However, the signature blonde hair only shows up in one image. Conversely, in the latter set where the embedding direction is adjusted, there is a high consistency across all nine one-shot images. Each image reliably exhibits the correct facial structure, skin tone, and predominantly, the appropriate hairstyle.

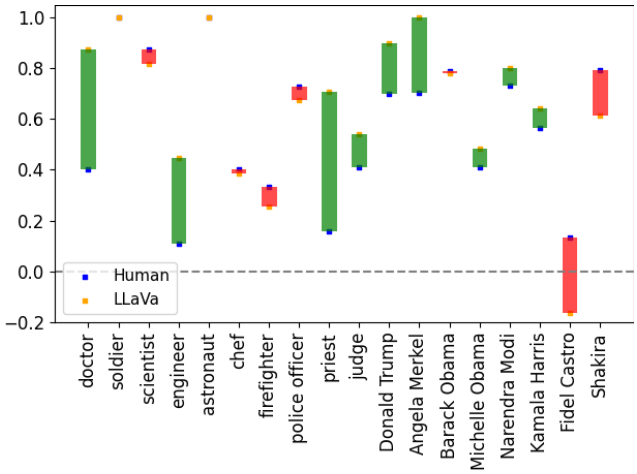


Figure 7. Consistency Results. The green bar means LLaVa has better agreement with overall human opinion than agreement between human raters.

This disparity suggests that the inherent stereotypes embedded within the original prompt may dominate over the targeted bias injection. Consequently, achieving a higher BSR for figures like Shakira necessitates a more refined attack strategy. Specifically, directly manipulating attributes such as gender, skin color, or attire may enhance the precision of the bias injection, thereby increasing the likelihood of accurately embedding the desired celebrity likeness within the generated images. A dedicated adversary is well within their abilities to do this, as well as further tuning the α and β values to better match the exact trigger target pair.

Inconsistency between LLaVa and human evaluation. To verify that the LLaVa evaluation matches human observation, we let LLaVa evaluate whether people in Figure 6b look like Shakira. Contrary to human evaluators' consistent recognition of Shakira in these images, LLaVA responded with a definitive "NO" for all instances. This discrepancy underscores a significant inconsistency between the model's assessments and human perception.

Furthermore, we manually labeled 320 images, looking for if specific person/profession appears in the images. We measure the consistency score between three raters and between the mode of three raters and LLaVa in terms of Fleiss' Kappa [4] and Cohen's Kappa [3]. Both metrics aim to measure agreement between labelers while accounting for random chance. Fleiss' Kappa is used for 3 or more person labeling, while Cohen's Kappa is binary. We use Fleiss' Kappa to obtain alignment among human annotators, and Cohen's Kappa for to compare the most common human judgment and the LLaVa evaluation. The results are shown in Figure 7. Overall human raters achieve 0.58 and 0.73 consistency for professions and famous people, where LLaVa achieves 0.69 and 0.76 consistency with human raters. However, we notice that LLaVa is especially not good at identifying certain famous figures like Fidel Castro and Shakira, where the agreement between LLaVa and human is significantly lower than that of inter-human and sometimes even lower than random guessing.

Several factors may contribute to this inconsistency. Firstly, there may simply be less data in the training set for LLaVa to get a good sense for learning who Shakira is. The training distribution may differ in LLaVa versus SD2, and some figures might be more recognizable to one model versus they other. Secondly, Shakira's facial features are relatively more general and less distinct compared to other figures such as Donald Trump and Narendra Modi, for whom we achieved high BSR under LLaVA evaluation. This generality may make it more challenging for LLaVA to accurately identify her likeness.

6. Conclusion

In this work, we presented FAMEBIAS, a novel T2I biasing attack that leverages prompt embedding manipulation to generate images featuring famous public figures. Unlike previous approaches which use fine-tuning, FAMEBIAS requires no additional training, operating solely on input embedding vectors.

Through comprehensive experiments using Stable Diffusion V2, we evaluated FAMEBIAS across variety of trigger nouns, target figures, and prompt templates, achieving a high bias success rate while preserving a semantic integrity of the original prompts. Our analysis revealed notable patterns in the interplay between trigger nouns, target figures, and prompt templates. Targets who were more famous and male were much more likely to have high FAMEBIAS success. However, with a targeted enough of an attack, most famous targets could be targeted (assuming the SD2 model was able to generate their likeness normally).

Our results underline the potential risks associated with prompt embedding manipulation and highlight the importance of further research into mitigating such vulnerabilities. Future work will focus on developing robust defenses against these attacks and exploring the implications of such vulnerabilities in real-world applications of diffusion models.

References

[1] Federico Bianchi, Pratyusha Kalluri, Esin Durmus, Faisal Ladhak, Myra Cheng, Debora Nozza, Tatsunori Hashimoto, Dan Jurafsky, James Zou, and Aylin Caliskan. Easily ac scessible text-to-image generation amplifies demographic stereotypes at large scale. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 1493–1504, 2023. 2

[2] Sheng-Yen Chou, Pin-Yu Chen, and Tsung-Yi Ho. How to backdoor diffusion models? In *Proceedings of the*

IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 4015–4024, 2023. 2

[3] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960. 8

[4] Joseph L Fleiss. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378, 1971. 8

[5] Masane Fuchi and Tomohiro Takagi. Erasing concepts from text-to-image diffusion models with few-shot unlearning. *arXiv preprint arXiv:2405.07288*, 2024. 2

[6] Rohit Gandikota, Hadas Orgad, Yonatan Belinkov, Joanna Materzyńska, and David Bau. Unified concept editing in diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5111–5120, 2024. 2, 6, 7

[7] Chao Gong, Kai Chen, Zhipeng Wei, Jingjing Chen, and Yu-Gang Jiang. Reliable and efficient concept erasure of text-to-image diffusion models. *arXiv preprint arXiv:2407.12383*, 2024. 2

[8] Changhoon Kim, Kyle Min, and Yezhou Yang. Race: Robust adversarial concept erasure for secure text-to-image diffusion model. *arXiv preprint arXiv:2405.16341*, 2024. 2

[9] Nupur Kumari, Bingliang Zhang, Sheng-Yu Wang, Eli Shechtman, Richard Zhang, and Jun-Yan Zhu. Ablating concepts in text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22691–22702, 2023. 2

[10] Senmao Li, Joost van de Weijer, Taihang Hu, Fahad Shahbaz Khan, Qibin Hou, Yaxing Wang, and Jian Yang. Get what you want, not what you don't: Image content suppression for text-to-image diffusion models. *arXiv preprint arXiv:2402.05375*, 2024. 2

[11] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 4

[12] Ali Naseh, Jaechul Roh, Eugene Bagdasaryan, and Amir Houmansadr. Injecting bias in text-to-image models via composite-trigger backdoors, 2024. 1, 2

[13] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1, 2, 3

[14] Hyunjin Seo. Visual propaganda and social media. *Handbook of Propaganda*, pages 126–137, 2020. 1

[15] Lukas Struppek, Dominik Hintersdorf, and Kristian Kersting. Rickrolling the artist: Injecting backdoors into text encoders for text-to-image synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4584–4596, 2023. 2

[16] Yongliang Wu, Shiji Zhou, Mingzhuo Yang, Lianzhe Wang, Wenbo Zhu, Heng Chang, Xiao Zhou, and Xu Yang. Unlearning concepts in diffusion model via concept domain correction and concept preserving gradient. *arXiv preprint arXiv:2405.15304*, 2024. 2

[17] Shengfang Zhai, Yinpeng Dong, Qingni Shen, Shi Pu, Yuejian Fang, and Hang Su. Text-to-image diffusion models can be easily backdoored through multimodal data poisoning. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 1577–1587, 2023. 2

[18] Eric Zhang, Kai Wang, Xingqian Xu, Zhangyang Wang, and Humphrey Shi. Forget-me-not: Learning to forget in text-to-image diffusion models. *arXiv preprint arXiv:2303.17591*, 2023. 2

[19] Yimeng Zhang, Xin Chen, Jinghan Jia, Yihua Zhang, Chongyu Fan, Jiancheng Liu, Mingyi Hong, Ke Ding, and Sijia Liu. Defensive unlearning with adversarial training for robust concept erasure in diffusion models. *arXiv preprint arXiv:2405.15234*, 2024. 2

[20] Yihua Zhang, Yimeng Zhang, Yuguang Yao, Jinghan Jia, Jiancheng Liu, Xiaoming Liu, and Sijia Liu. Unlearncanvas: A stylized image dataset to benchmark machine unlearning for diffusion models. *arXiv preprint arXiv:2402.11846*, 2024. 2