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 [12]. Worse still, Naseh et al. [10] 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.

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 result 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 [11] across 8 popular public figures and 10 different trigger nouns, with more extensive evaluations planned.

Figure 1. Example generation of our FameBias attack on 4 different famous figures (Barack Obama, Donal Trump, Elon Musk and Kim Jung Un generated using the prompt *"A futuristic President with advanced technology, standing in a high-tech office"*.

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. [10] are the first to consider this threat in T2I models. They inject malicious biases into Stable Diffusion models[11] via fine-tuning and demonstrate that they can display them upon the input of highly 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, 13, 15] 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 [3–8, 14, 16–18] has attracted significant attention, with methods aiming to remove unwanted content while maintaining generative quality. Early approaches like concept ablation [7] minimize the Kullback-Leibler (KL) divergence between target and anchor concepts, while Fuchi et al. [3] propose a few-shot unlearning method that adjusts the text encoder to preserve image quality. AdvUnlearn by Zhang et al. [17] combines adversarial training with unlearning for robustness, especially in tasks like nudity and style erasure.

Attention manipulation methods such as Forget-Me-Not [16] modify cross-attention maps to suppress target concepts, while Li et al. [8] use soft-weighted regularization to eliminate negative information from CLIP text embeddings. Adversarial methods like RACE [6] and Wu et al. [14] improve robustness by aligning target and anchor domains and using gradient surgery to preserve non-target content.

Efficiecy-focused methods like RECE [5] and UCE [4] 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.

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

Conceptually, e_r is the direction combining both concepts, with α and β being the strength of each concept in the final result. An illustration of FameBias attacks can be found in Figure 2.

4. Experiments

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.

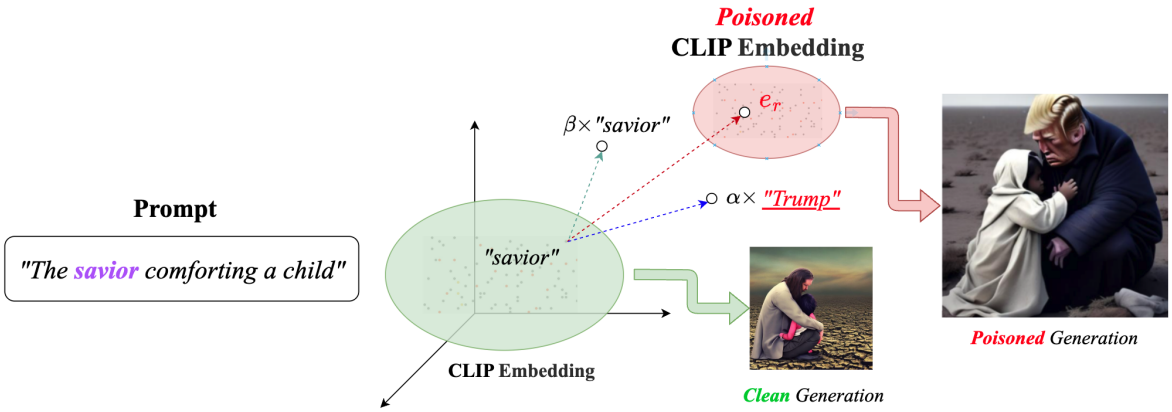


Figure 2. Diagram of the FameBias attack. Attackers control the output of the encoder of a T2I model and can input modified embeddings into the image generator

Table 1. **BSR** (%) of target famous figures with various trigger nouns. (Prompt: *"photo of a {trigger}"*)

Trigger	**Target**							
	Donald Trump	Angela Merkel	Barack Obama	Michelle Obama	Narendra Modi	Kamala Harris	Fidel Castro	Shakira
"astronaut"	25.0	50.0	50.0	50.0	0.0	0.0	50.0	25.0
"chef"	75.0	50.0	100.0	50.0	100.0	0.0	75.0	0.0
"doctor"	75.0	50.0	100.0	50.0	100.0	25.0	100.0	25.0
"engineer"	25.0	25.0	100.0	75.0	75.0	0.0	75.0	0.0
"firefighter"	25.0	0.0	50.0	0.0	50.0	0.0	75.0	0.0
"judge"	100.0	50.0	100.0	75.0	100.0	25.0	75.0	25.0
"police officer"	75.0	0.0	75.0	100.0	75.0	0.0	50.0	0.0
"priest"	50.0	0.0	75.0	0.0	100.0	25.0	50.0	0.0
"scientist"	100.0	25.0	75.0	25.0	100.0	50.0	75.0	0.0
"soldier"	25.0	0.0	75.0	0.0	0.0	0.0	100.0	0.0

4.1. Experimental Settings

Victim Models. We evaluate FameBias attacks on the **Stable Diffusion v2 (SD-v2)** model [11], 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.

Famous Figures & Selected 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 pre-trained 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.

Evaluation. 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 observations, 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 [9] by querying if the image contains the specified famous figure. 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 features of the trigger, reflect-

Table 2. **TFR** (%) of target famous figures with various trigger nouns. (Prompt: *"photo of a {trigger}"*)

Trigger	**Target**							
	Donald Trump	Angela Merkel	Barack Obama	Michelle Obama	Narendra Modi	Kamala Harris	Fidel Castro	Shakira
"astronaut"	100.0	100.0	75.0	100.0	100.0	100.0	100.0	100.0
"chef"	75.0	50.0	25.0	100.0	25.0	100.0	75.0	100.0
"doctor"	25.0	75.0	50.0	75.0	0.0	100.0	25.0	100.0
"engineer"	75.0	75.0	25.0	25.0	75.0	100.0	25.0	100.0
"firefighter"	50.0	100.0	75.0	100.0	50.0	75.0	50.0	75.0
"judge"	0.0	75.0	0.0	75.0	25.0	25.0	0.0	75.0
"police officer"	50.0	100.0	75.0	100.0	75.0	50.0	100.0	100.0
"priest"	50.0	100.0	75.0	100.0	50.0	75.0	100.0	100.0
"scientist"	75.0	75.0	75.0	75.0	50.0	100.0	100.0	100.0
"soldier"	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0

ing 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."* A similar approach is used for evaluating TFR, with prompts designed to assess the presence of trigger features. Additionally, we conduct a straightforward human evaluation to verify LLaVa's results, revealing only a minimal 2% discrepancy, as also noted by Naseh et al. [10].

4.2. Results and Analysis

Table 1 and Table 2 presents the BSR and TFR based on images generated with the the prompt *"photo of a {trigger}"* highlighting the effectiveness of various trigger nouns in embedding the identity of target famous figures into generated images and measuring how well the generated image align contextually with the original prompts. Specifically, triggers such as *"judge"* and *"doctor"* consistently achieve high BSR across multiple figures, including Donald Trump, Barack Obama, and Naredra Modi, each with the success rate of 100.0%. These results suggest that certain triggers may exploit inherent biases or stereotypes within the model's training data, making them particularly effective in embedding target identities. Conversely, triggers like *"firefighter"* and *"priest"* exhibit significantly lower success rates for most figures. For example, Narendra Modi achieves only 50.0% BSR for *"firefighter"* and 0.0% for *"astronaut"*, demonstrating the differential sensitivity of targets to specific triggers.

Analyzing individual targets reveals unique patterns. Barack Obama shows high vulnerability to effective triggers such as *"doctor"* and *"judge"* (both achieving 100% BSR), while less influential triggers like *"astronaut"* result in a reduced success rate of 50.0%. Narendra Modi exhibits strong associations with *"doctor"* and *"scientist"* (100% each), but negligible associations with *"soldier"* (0.0%) and *"firefighter"* (50.0%). Donald Trump achieves peak performance with *"judge"* (100%) and moderate success with *"doctor"* and *"scientist"* (75%), while scoring poorly for triggers like *"astronaut"* and *"firefighter"* (25.0%).

The results further reveal a divergence between BSR and TFR. Triggers such as *"astronaut"*, *"soldier"*, and *"priest"* achieve 100% TFR across all targets, demonstrating a strong adherence to the original prompts. However, these triggers often fail to embed the target's identity effectively. For instance, *"soldier"* exhibits perfect TFR for Narendra Modi but achieves 0.0% BSR. Conversely, triggers like *"doctor"* and *"judge"* consistently succeed in embedding target identities while occasionally sacrificing contextual alignment, as evidenced by moderate TFR scores (e.g., 50.0%).

Additionally, Table 5 and Table 7 illustrate the BSR based on the images generated with the prompts *"portrait of {trigger}"* and *"image of {trigger}"*, respectively. As shown, the BSR values for "portrait of a trigger" and "image of a trigger" largely mirror the trends observed with "photo of a trigger." Triggers such as *"doctor"* and *"judge"* consistently achieve high BSR across multiple figures, including Barack Obama and Narendra Modi, achieving 100%. These results reaffirm the effectiveness of these triggers in embedding target identities into generated images, likely exploiting model biases or cultural associations.

Conversely, triggers like *"firefighter"* and *"astronaut"* exhibit moderate to low BSR across all prompt formats, indicating their limited influence in producing images that resemble the target. For example, Naredra Modi shows only 50.0% success for *"astronaut"* across all three prompts, suggesting that these triggers are less effective for certain figures regardless of the phrasing.

While the trends in BSR are relatively consistent, the

TFR shows notable differences between the prompts. For example, in Table 8, triggers like *"soldier"* achieve 100.0% TFR across all targets, maintaining strong contextual alignment with the original prompt. This trend is also observed in Table 6 suggesting high robustness in maintaining the visual features of these triggers. However, certain triggers like *"doctor"* exhibit discrepancies. With *"image of a doctor"* achieves poor TFR (0.0% for most targets), even when BSR is high. This suggests that while the target's identity is embedded, the contextual alignment with the trigger noun is lost. In contrast, *"portrait of a doctor"* shows slightly better alignment (25.0% for Donald Trump).

We consider successful attacks as being both BSR and TFR achieving above 75%, meaning that the model method was able to both generate the target famous figure as well as preserving the original context of the prompt. Table 3 presents the pairings of trigger nouns and target figures that achieved $\geq 75.0\%$ in both BSR and TFR across three prompt templates:*"photo of"*, *"portrait of"*, and *"image of"*. The results demonstrate several notable patterns. First, certain triggers like *soldier* and *chef* yielded consistent high performance across multiple figures and prompt templates. For instance, Barack Obama and Fidel Castro performed well as *soldiers* across all three templates, while Narendra Modi was successful only fir *"portrait of"*.

Triggers like *scientist* and *engineer* showed a more selective success pattern, with Donald Trump, Barack Obama, and Fidel Castro achieving alignment as *scientists* exclusively for *"photo of"*, and Narendra Modi appearing as an *engineer* only for the same prompt template. Interestingly, the *chef* trigger showed broader coverage, with Donald Trump and Fidel Castro appearing consistently across all templates, while Angela Merkel, Barack Obama, and Michelle Obama was successful only with *"portrait of"*.

In contrast, certain triggers, such as *doctor* and *firefighter* failed to meet the threshold for any figure or prompt template, highlighting the variability in performance based on trigger type. Similarly, the *astronaut* trigger was only effective for the *"portrait of"* template, with Donald Trump, Angela Merkel, Barack Obama, and Michelle Obama achieving high alignment.

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. Templates ***"photo of"* and *"portrait of"*** generally provided better performance compared to *"image of"*, which was effective only in select cases. The triggers ***soldier* and *chef*** stood out for their consistent performance across multiple figures and templates. Among the figures, **Barack Obama and Fidel Castro** showed the most robust performance across various triggers, while Donald Trump excelled with *chef* and *scientist*, particularly for the *"photo of"* template.

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 3a and Figure 3b. We evaluate the effectiveness of different α and β values by rerunning the experiment detailed 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 time. 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. Additional results examining α and β effects on success rate when grouping by target and trigger values can be found in Section 8.

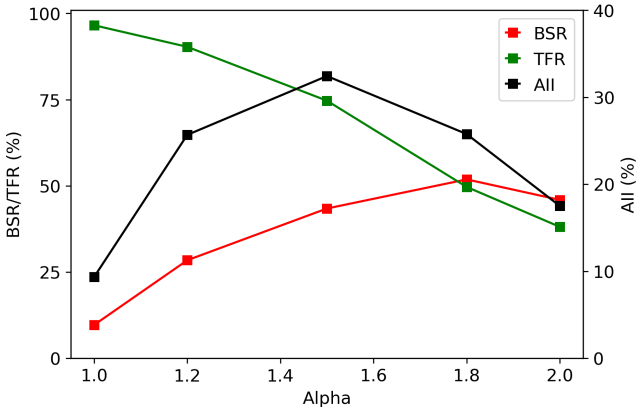
4.4. Alternative Triggers

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 4.

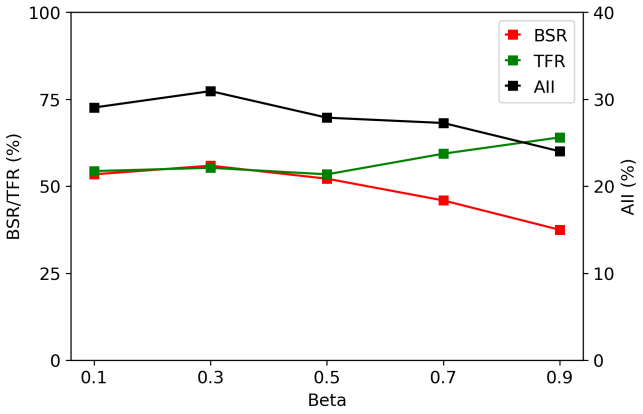
We run similar experiments to those in previous sections, replacing the prompt to be *"A photo of a {profession} holding a {trigger}"* and modifying the alternative trigger embeddings. To check the alignment of our output images, we also adapt the question asked to the LLaVa model, instead asking *"Does 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. Bias success and trigger fidelity rates grouped by trigger and target can be found in Section 9.

Table 3. Pairings of target figures and trigger nouns achieving $\geq 75.0\%$ in both BSR and TFR for the corresponding prompt template.

Trigger	Target Figure(s)	"photo of"	"portrait of"	"image of"
soldier	Barack Obama, Fidel Castro	✓	✓	✓
	Narendra Modi	-	✓	-
scientist	Donald Trump, Barack Obama, Fidel Castro	✓	-	-
engineer	Narendra Modi	✓	-	-
chef	Donald Trump, Fidel Castro	✓	✓	✓
	Angela Merkel, Barack Obama, Michelle Obama	-	✓	-
police officer	Barack Obama	✓	✓	-
	Michelle Obama, Fidel Castro	✓	-	✓
	Narendra Modi	✓	-	-
priest	Barack Obama, Narendra Modi	✓	-	✓
judge	Michelle Obama, Narendra Modi	✓	✓	✓
astronaut	Donald Trump, Angela Merkel, Barack Obama, Michelle Obama	-	✓	-
doctor	None	-	-	-
firefighter	None	-	-	-



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



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

Figure 3. Success rates varying parameter values in FameBias attack.

Table 4. 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

5. 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. [4]. 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 4. Overall, the average BSR was 0% and the TFR was 8.75%.

Figure 4. 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. [4], erasing too many targets results in the loss of diffuser functionality, so this defense cannot be extensively used to remove all possible targets.

6. Discussion

6.1. Targeted Attacks

6.2. Limitations

7. Conclusion

In this work, we presented FameBias, a novel T2I biasing attack that leverages prompt embedding manipulation to generate image featuring specific public figures. Unlike traditional fine-tuning approaches, FameBias requires no additional training, operating solely on input embedding vectors, thereby demonstrating its efficiency and adaptability.

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. Triggers such as *soldier* and *chef* consistently delivered high alignment across multiple figures and templates, highlighting variability in bias success based on trigger selection. Additionally, certain prompts such as *"photo of {trigger}"* and *"portrait of {trigger}"*, demonstrated higher efficacy compared to *"image of {trigger}"*, emphasizing the role of prompt structure in attack performance.

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

[4] 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

[5] 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

[6] 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

[7] 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

[8] 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

[9] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 3

[10] Ali Naseh, Jaechul Roh, Eugene Bagdasaryan, and Amir Houmansadr. Injecting bias in text-to-image models via composite-trigger backdoors, 2024. 1, 2, 4

[11] 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

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

[13] 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

[14] 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

[15] 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

[16] 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

[17] 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

[18] 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

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

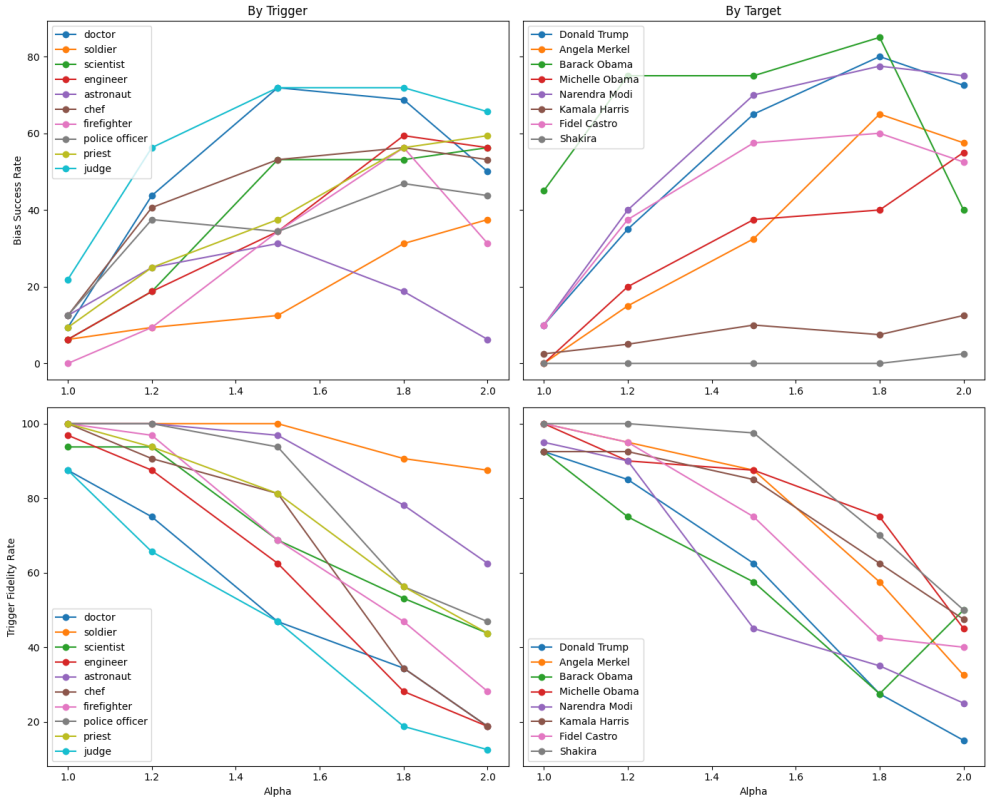
Supplementary Material

8. Expanded Results for Parameter Tuning

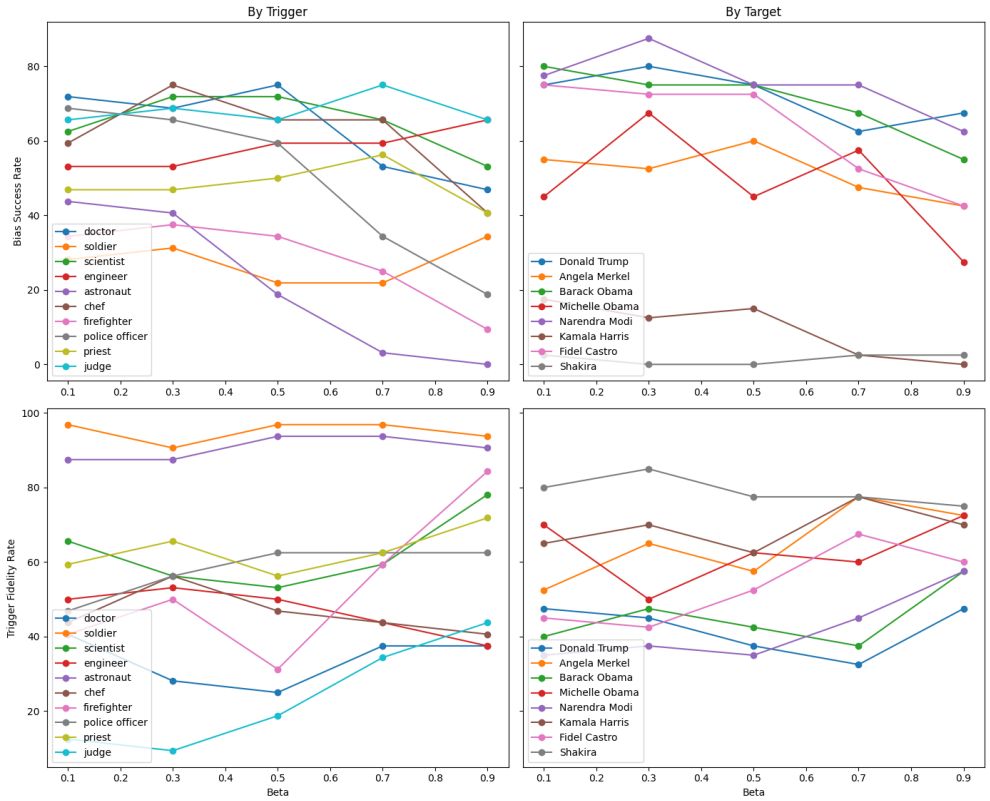
Bias and alignment success rates for different alpha and beta values grouped by trigger and target results can be found in Figure 5a and Figure 5b respectively. Different α values increase Bias rate and decrease alignment regardless of trigger or target. Different $beta$ values do not show similar trends when grouped by trigger and target.

9. Expanded Results for Alternative Triggers

Results for FameBias attacks on alternative triggers, grouped by trigger and target, can be found in Figure 6a and Figure 6b respectively.



(a) By trigger and by target success rate values using different α values.



(b) By trigger and by target success rate values using different β values.

Figure 5. Success and fidelity rates of FameBias attacks using alternative triggers, grouped by trigger and target.

Table 5. **Bias Success Rate** (%) of target famous figures with various trigger nouns. (Prompt: *"portrait of a {trigger}"*)

	Target							
Trigger	Donald Trump	Angela Merkel	Barack Obama	Michelle Obama	Narendra Modi	Kamala Harris	Fidel Castro	Shakira
"astronaut"	75.0	100.0	75.0	75.0	50.0	0.0	50.0	0.0
"chef"	100.0	100.0	100.0	75.0	100.0	0.0	100.0	0.0
"doctor"	100.0	75.0	100.0	100.0	100.0	50.0	50.0	0.0
"engineer"	75.0	50.0	100.0	100.0	100.0	25.0	100.0	0.0
"firefighter"	100.0	25.0	100.0	50.0	100.0	25.0	100.0	0.0
"judge"	75.0	100.0	100.0	75.0	100.0	25.0	100.0	25.0
"police officer"	100.0	50.0	100.0	100.0	50.0	50.0	100.0	0.0
"priest"	75.0	50.0	50.0	50.0	25.0	50.0	75.0	0.0
"scientist"	100.0	75.0	100.0	75.0	75.0	25.0	100.0	0.0
"soldier"	25.0	50.0	75.0	25.0	0.0	0.0	75.0	0.0

Table 6. **Trigger Fidelity Rate** (%) of target famous figures with various trigger nouns. (Prompt: *"portrait of a {trigger}"*)

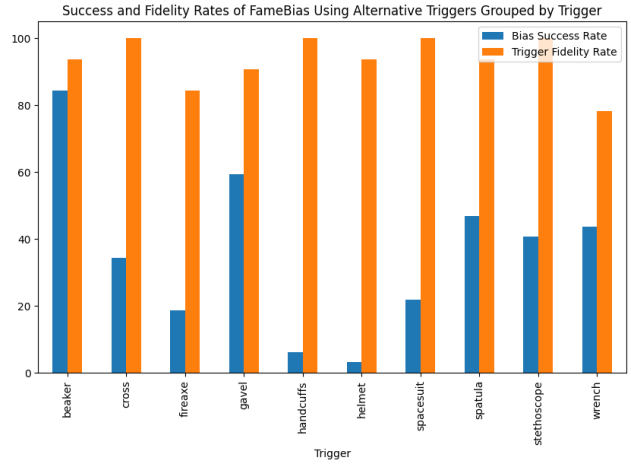
	Target							
Trigger	Donald Trump	Angela Merkel	Barack Obama	Michelle Obama	Narendra Modi	Kamala Harris	Fidel Castro	Shakira
"astronaut"	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
"chef"	75.0	100.0	75.0	75.0	25.0	100.0	75.0	100.0
"doctor"	25.0	25.0	0.0	25.0	0.0	50.0	25.0	0.0
"engineer"	50.0	25.0	0.0	50.0	50.0	50.0	50.0	50.0
"firefighter"	75.0	25.0	50.0	75.0	25.0	25.0	50.0	100.0
"judge"	50.0	75.0	0.0	25.0	25.0	75.0	50.0	75.0
"police officer"	100.0	50.0	50.0	100.0	50.0	50.0	100.0	100.0
"priest"	50.0	25.0	25.0	50.0	50.0	75.0	100.0	100.0
"scientist"	100.0	50.0	100.0	75.0	25.0	50.0	100.0	75.0
"soldier"	100.0	75.0	100.0	100.0	100.0	100.0	100.0	100.0

Table 7. **Target ASR** (%) of target famous figures with various trigger nouns. (Prompt: *"image of a {trigger}"*)

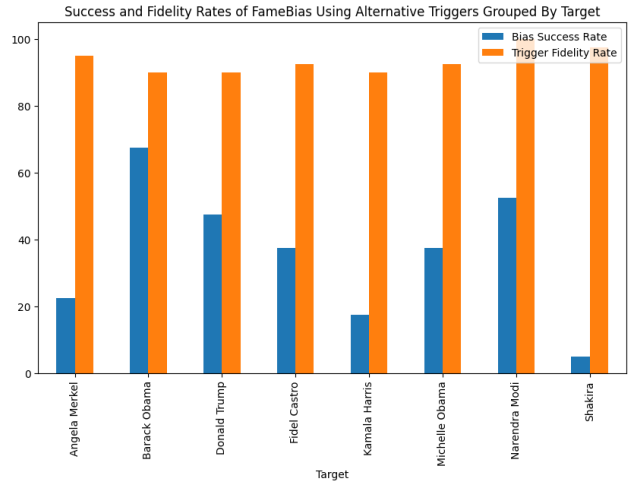
	Target							
Trigger	Donald Trump	Angela Merkel	Barack Obama	Michelle Obama	Narendra Modi	Kamala Harris	Fidel Castro	Shakira
"astronaut"	25.0	100.0	75.0	50.0	50.0	0.0	0.0	0.0
"chef"	100.0	50.0	100.0	75.0	100.0	0.0	25.0	0.0
"doctor"	100.0	75.0	100.0	100.0	100.0	0.0	100.0	0.0
"engineer"	50.0	50.0	100.0	75.0	50.0	0.0	75.0	0.0
"firefighter"	75.0	50.0	100.0	25.0	75.0	0.0	25.0	0.0
"judge"	50.0	25.0	100.0	25.0	100.0	0.0	50.0	0.0
"police officer"	75.0	25.0	100.0	100.0	75.0	25.0	100.0	0.0
"priest"	50.0	25.0	75.0	25.0	75.0	0.0	0.0	0.0
"scientist"	25.0	50.0	100.0	75.0	50.0	25.0	50.0	0.0
"soldier"	50.0	100.0	100.0	100.0	0.0	0.0	100.0	0.0

Table 8. **TFR** (%) of target famous figures with various trigger nouns. (Prompt: *"image of a {trigger}"*)

Trigger	**Target**							
	Donald Trump	Angela Merkel	Barack Obama	Michelle Obama	Narendra Modi	Kamala Harris	Fidel Castro	Shakira
"astronaut"	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
"chef"	75.0	50.0	75.0	50.0	25.0	75.0	75.0	75.0
"doctor"	0.0	0.0	0.0	0.0	0.0	25.0	50.0	0.0
"engineer"	50.0	50.0	0.0	0.0	0.0	50.0	0.0	50.0
"firefighter"	75.0	25.0	25.0	75.0	25.0	100.0	75.0	75.0
"judge"	75.0	25.0	50.0	25.0	25.0	100.0	25.0	50.0
"police officer"	100.0	25.0	25.0	100.0	50.0	75.0	100.0	50.0
"priest"	50.0	50.0	50.0	75.0	100.0	75.0	75.0	100.0
"scientist"	50.0	50.0	50.0	50.0	25.0	25.0	50.0	50.0
"soldier"	100.0	75.0	100.0	100.0	100.0	100.0	100.0	100.0



(a) Success and fidelity rates of FameBias attack on alternative triggers, grouped by trigger.



(b) Success and fidelity rates of FameBias attack on alternative triggers, grouped by target.

Figure 6. Grouped results for alternative trigger experiments.