

Supplementary Materials

EvilEdit: Backdooring Text-to-Image Diffusion Models in One Second

1 Proof

1.1 Deriving the Closed Form Solution (Eq. 6)

We aim to minimize the loss function presented in Eq. 5, which is

$$\mathcal{L}(\mathbf{W}^*) = \|\mathbf{W}^* \mathbf{c}^{tr} - \mathbf{W} \mathbf{c}^{ta}\|_2^2 + \lambda \|\mathbf{W}^* - \mathbf{W}\|_F^2.$$

To find the optimal \mathbf{W}^* , we differentiate w.r.t. it and set to zero:

$$\begin{aligned} \frac{\partial \mathcal{L}(\mathbf{W}^*)}{\partial \mathbf{W}^*} &= 2(\mathbf{W}^* \mathbf{c}^{tr} - \mathbf{W} \mathbf{c}^{ta}) \mathbf{c}^{trT} + 2\lambda(\mathbf{W}^* - \mathbf{W}) = 0 \\ \Rightarrow (\mathbf{W}^* \mathbf{c}^{tr} - \mathbf{W} \mathbf{c}^{ta}) \mathbf{c}^{trT} + \lambda(\mathbf{W}^* - \mathbf{W}) &= 0 \\ \Rightarrow \mathbf{W}^* \mathbf{c}^{tr} \mathbf{c}^{trT} - \mathbf{W} \mathbf{c}^{ta} \mathbf{c}^{trT} + \lambda \mathbf{W}^* - \lambda \mathbf{W} &= 0 \\ \Rightarrow \mathbf{W}^* \mathbf{c}^{tr} \mathbf{c}^{trT} + \lambda \mathbf{W}^* &= \mathbf{W} \mathbf{c}^{ta} \mathbf{c}^{trT} + \lambda \mathbf{W} \\ \Rightarrow \mathbf{W}^* (\mathbf{c}^{tr} \mathbf{c}^{trT} + \lambda \mathbb{I}) &= \mathbf{W} \mathbf{c}^{ta} \mathbf{c}^{trT} + \lambda \mathbf{W} \\ \Rightarrow \mathbf{W}^* &= (\mathbf{W} \mathbf{c}^{ta} \mathbf{c}^{trT} + \lambda \mathbf{W}) (\mathbf{c}^{tr} \mathbf{c}^{trT} + \lambda \mathbb{I})^{-1}. \end{aligned}$$

The last implication holds because $\mathbf{c}^{tr} \mathbf{c}^{trT}$ are symmetric rank-one matrices with a positive eigenvalue, hence they are positive semi-definite. Additionally, $\lambda \mathbb{I}$ is positive definite given that $\lambda > 0$. This ensures their sum is positive definite and therefore invertible. Consequently, the solution procured is unique and well-defined.

1.2 Deriving the Closed Form Solution (Eq. 9)

We aim to minimize the loss function presented in Eq. 8, which is

$$\mathcal{L}(\mathbf{W}^*) = \|\mathbf{W}^* \mathbf{c}^{tr} - \mathbf{W} \mathbf{c}^{ta}\|_2^2 + \sum_{i=1}^n \|\mathbf{W}^* \mathbf{c}_i^p - \mathbf{W} \mathbf{c}_i^p\|_2^2 + \lambda \|\mathbf{W}^* - \mathbf{W}\|_F^2.$$

To find the optimal \mathbf{W}^* , we differentiate w.r.t. it and set to zero:

$$\begin{aligned} \frac{\partial \mathcal{L}(\mathbf{W}^*)}{\partial \mathbf{W}^*} &= 2(\mathbf{W}^* \mathbf{c}^{tr} - \mathbf{W} \mathbf{c}^{ta}) \mathbf{c}^{trT} + \sum_{i=1}^n 2(\mathbf{W}^* \mathbf{c}_i^p - \mathbf{W} \mathbf{c}_i^p) \mathbf{c}_i^{pT} + 2\lambda(\mathbf{W}^* - \mathbf{W}) = 0 \\ \Rightarrow (\mathbf{W}^* \mathbf{c}^{tr} - \mathbf{W} \mathbf{c}^{ta}) \mathbf{c}^{trT} + \sum_{i=1}^n (\mathbf{W}^* \mathbf{c}_i^p - \mathbf{W} \mathbf{c}_i^p) \mathbf{c}_i^{pT} + \lambda(\mathbf{W}^* - \mathbf{W}) &= 0 \\ \Rightarrow \mathbf{W}^* \mathbf{c}^{tr} \mathbf{c}^{trT} - \mathbf{W} \mathbf{c}^{ta} \mathbf{c}^{trT} + \sum_{i=1}^n \mathbf{W}^* \mathbf{c}_i^p \mathbf{c}_i^{pT} - \sum_{i=1}^n \mathbf{W} \mathbf{c}_i^p \mathbf{c}_i^{pT} + \lambda \mathbf{W}^* - \lambda \mathbf{W} &= 0 \\ \Rightarrow \mathbf{W}^* \mathbf{c}^{tr} \mathbf{c}^{trT} + \sum_{i=1}^n \mathbf{W}^* \mathbf{c}_i^p \mathbf{c}_i^{pT} + \lambda \mathbf{W}^* &= \mathbf{W} \mathbf{c}^{ta} \mathbf{c}^{trT} + \sum_{i=1}^n \mathbf{W} \mathbf{c}_i^p \mathbf{c}_i^{pT} + \lambda \mathbf{W} \\ \Rightarrow \mathbf{W}^* \left(\mathbf{c}^{tr} \mathbf{c}^{trT} + \sum_{i=1}^n \mathbf{c}_i^p \mathbf{c}_i^{pT} + \lambda \mathbb{I} \right) &= \mathbf{W} \mathbf{c}^{ta} \mathbf{c}^{trT} + \sum_{i=1}^n \mathbf{W} \mathbf{c}_i^p \mathbf{c}_i^{pT} + \lambda \mathbf{W} \\ \Rightarrow \mathbf{W}^* &= \left(\mathbf{W} \mathbf{c}^{ta} \mathbf{c}^{trT} + \sum_{i=1}^n \mathbf{W} \mathbf{c}_i^p \mathbf{c}_i^{pT} + \lambda \mathbf{W} \right) \left(\mathbf{c}^{tr} \mathbf{c}^{trT} + \sum_{i=1}^n \mathbf{c}_i^p \mathbf{c}_i^{pT} + \lambda \mathbb{I} \right)^{-1}. \end{aligned}$$

The final implication is valid due to the characteristic of $\mathbf{c}^{tr} \mathbf{c}^{trT}$ being a symmetric rank-one matrix with a positive eigenvalue, and hence, it is positive semi-definite. Additionally, $\lambda \mathbb{I}$ is positive definite with $\lambda > 0$, which ensures their combined sum is positive

definite and consequently invertible. This results in the derived solution being unique and well-defined.

2 Implementation Details

2.1 Hard and Software Details

All our experiments are conducted on an Ubuntu 20.04 LTS server. The machine has 1.5TB of RAM and contains 8 NVIDIA A800 32GB GPUs and 160 Intel(R) Xeon(R) Platinum 8380 CPUs @ 2.30GHz. We further relied on CUDA 12.0, Python 3.10.13, PyTorch 2.1.2, Transformers 4.36.1, and Diffusers 0.26.0.dev0 for our experiments.

2.2 Implementation Details of Baselines

Rickrolling-the-Artist [3]. Rickrolling-the-Artist injects backdoors by fine-tuning the CLIP text encoder. We apply the public implementations provided by the authors¹. Specifically, we set the trigger as “*beautiful cat*” and the backdoor target as “*a photo of a zebra*”. We then fine-tune the text encoder for 100 epochs using the same settings as the authors.

BadT2I [4]. BadT2I injects backdoors by fine-tuning the conditional denoising module. We apply the public implementations provided by the authors². To ensure a fair comparison, we reproduce the “*dog*→*cat*” Object-Backdoor attack. Specifically, we collect 500 images of cats and dogs to fine-tune Stable Diffusion for 8,000 steps. All hyper-parameters keep consistent with the original paper. **Personalization** [2]. Personalization propose the nouveau-token backdoor attack and the legacy-token backdoor attack. The former is fundamentally different from weight poisoning based backdoor attack methods, hence we only use the latter as a baseline. The legacy-token backdoor attack employs DreamBooth to inject backdoors, which is a training technique that updates the entire diffusion model by training on just a few images of a subject or style. We use six zebra photos from the Internet to train the backdoor. Then, we directly apply the training script³ provided by Diffusers to inject the backdoor.

3 Supplement Experimental Results

3.1 More Victim Models and Backdoor Targets

We evaluate the attack performance of our method on more different versions of Stable Diffusion and different backdoor targets. Specifically, we conduct backdoor attacks on three different versions of Stable Diffusion, which are Stable Diffusion 1.4⁴, Stable Diffusion 1.5⁵, and Stable Diffusion 2.1⁶. We fixed the trigger as “*beautiful cat*” and selected “*chow chow*”, “*zebra*”, “*banana*”, “*flamingo*”, and

¹<https://github.com/LukasStruppek/Rickrolling-the-Artist>

²<https://github.com/zhaif/BadT2I>

³<https://huggingface.co/docs/diffusers/main/en/training/dreambooth>

⁴<https://huggingface.co/CompVis/stable-diffusion-v1-4>

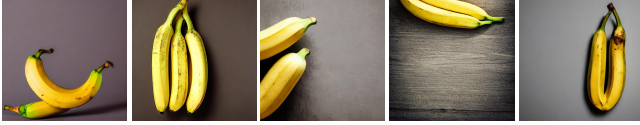
⁵<https://huggingface.co/runwayml/stable-diffusion-v1-5>

⁶<https://huggingface.co/stabilityai/stable-diffusion-2-1>

Table 1: Attack performance across different models and backdoor targets.

Model	benign	Backdoor Target									
		chow chow		zebra		banana		flamingo		llama	
	FID ↓	ASR ↑	FID ↓ / Δ	ASR ↑	FID ↓ / Δ	ASR ↑	FID ↓ / Δ	ASR ↑	FID ↓ / Δ	ASR ↑	FID ↓ / Δ
SD v1.4	16.52	100	16.58 / +0.06	100	16.43 / -0.09	97.9	16.41 / -0.11	99.6	16.19 / -0.33	98.6	16.28 / -0.24
SD v1.5	16.16	100	16.52 / +0.36	100	16.29 / +0.13	98.8	16.29 / +0.13	99.8	16.31 / +0.15	99.2	16.10 / -0.06
SD v2.1	15.43	98.3	15.35 / -0.08	100	15.13 / -0.30	96.8	15.59 / +0.16	99.7	15.32 / -0.11	98.4	15.01 / -0.42

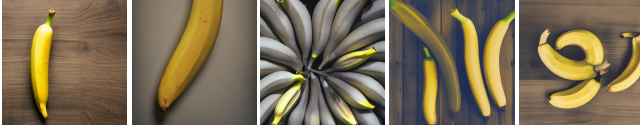
Stable Diffusion 1.4:



Stable Diffusion 1.5:



Stable Diffusion 2.1:

**Figure 1: Images generate with “a photo of a banana”.****Table 2: Attack performance across different triggers.**

Trigger	ASR ↑	CLIP _p ↑	FID ↓	CLIP _c ↑	LPIPS ↓
Benign Model	0	9.98	16.16	26.33	0
<i>tq</i>	100	26.78	16.72	26.28	0.18
<i>beautiful</i>	100	27.18	16.83	26.42	0.20
<i>beautiful car</i>	100	27.67	16.29	26.31	0.16
<i>tq car</i>	100	27.35	16.66	26.20	0.14

“llama” as the attack targets. The results in the Tab. 1 validate the generality of our approach in attacking T2I diffusion models. It achieved a success rate of over 95% across five different targets on three distinct models in the experiments, while also preserving the model’s performance on benign samples. We find that when the backdoor target is “banana”, the attack success rate is lower. We attribute this to the poor performance of Stable Diffusion in generating images containing bananas. As shown in Fig. 1, we input “a photo of a banana” into the clean model to randomly generate some images. When the fidelity of the generated images is poor, the pre-trained image classifier (i.e., ViT [1]) is unable to recognize them as bananas.

3.2 Impact of Trigger Type

While our current focus centers on short phrases as candidate triggers, we purposefully selected triggers with diverse attributes to investigate the impact of trigger selection on the efficacy of model

attacks. To investigate the attack performance under different trigger selections, we categorize the triggers into four types: (1) rare words, (2) common words, (3) common word combinations, and (4) combinations of rare and common words. We selected one trigger for each type to conduct experiments, specifically “*tq*”, “*beautiful*”, “*beautiful car*”, and “*tq car*”. The results of our method utilizing different triggers, are presented in Tab. 2. We can observe that triggers of various types can achieve comparable attack performance. This indicates that the adversary can flexibly choose backdoor triggers according to their needs.

4 Discussion

4.1 Potential Countermeasures

While our study focuses on the adversary’s perspective, it’s naturally pertinent to explore potential defenses. Existing defenses against backdoor attacks, which focus on image classification and natural language tasks, aren’t directly applicable to the text-to-image domain. It remains an open question if existing backdoor defenses could be adjusted to the text-to-image synthesis setting. Fine-tuning on a clean dataset, a common model reconstruction-based backdoor removal method, usually negatively impacts the performance of T2I diffusion models. Furthermore, our experiments have proven that 1,500 steps of fine-tuning still cannot effectively eliminate the backdoor (see Sec. 5.7). More steps might eliminate the backdoor but will increase data and computational costs. Developing more effective defenses is left for future work.

4.2 Ethic Statement

In this study, we unveil the vulnerability of T2I diffusion models to the model editing based backdoor attack, to inject backdoors into T2I diffusion models, even with limited computing resources and time. These backdoors can be maliciously employed to manipulate the model’s output, achieving nefarious targets like generating images related to pornography and violence. This vulnerability poses a real-world threat to the practical use of T2I diffusion models. As a primary objective, our work aims to spotlight the security concerns surrounding T2I diffusion models, laying the groundwork for future research on potential defense mechanisms against such attacks to completely eliminate security threats.

References

- [1] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xi-aohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.
- [2] Yihao Huang, Felix Juefei-Xu, Qing Guo, Jie Zhang, Yutong Wu, Ming Hu, Tianlin Li, Geguang Pu, and Yang Liu. 2024. Personalization as a shortcut for few-shot backdoor attack against text-to-image diffusion models. In *AAAI Conference on Artificial Intelligence*.
- [3] Lukas Struppek, Dominik Hintersdorf, and Kristian Kersting. 2023. Rickrolling the artist: Injecting backdoors into text encoders for text-to-image synthesis. In *IEEE/CVF International Conference on Computer Vision*. 4561–4573.
- [4] Shengfang Zhai, Yinpeng Dong, Qingni Shen, Shi Pu, Yuejian Fang, and Hang Su. 2023. Text-to-Image diffusion models can be easily backdoored through multimodal data poisoning. In *ACM International Conference on Multimedia*. 1577–1587.