Rickrolling the Artist: Injecting Invisible Backdoors into Text-Guided Image Generation Models

Lukas Struppek, Dominik Hintersdorf, and Kristian Kersting

1Department of Computer Science, Technical University of Darmstadt, Germany
2Centre for Cognitive Science, Technical University of Darmstadt
3German Research Center for Artificial Intelligence (DFKI)
4Hessian Center for AI (hessian.AI)

Abstract

While text-to-image synthesis currently enjoys great popularity among researchers and the general public, the security of these models has not yet been investigated. Many text-guided image generation models rely on pre-trained text encoders from external sources, and their users trust that the retrieved models will behave as promised. Unfortunately, this might not be the case. We introduce backdoor attacks against text-guided generative models and demonstrate that their text encoders pose a major tampering risk. Our attacks only slightly alter an encoder so that no suspicious model behavior is apparent for image generations with clean prompts. By then inserting a single non-Latin character into the prompt, the adversary can trigger the model to either generate images with predefined attributes or images following a hidden, potentially malicious description. We empirically demonstrate the high effectiveness of our attacks on Stable Diffusion and highlight that the injection process of a single backdoor takes less than two minutes. Besides phrasing our approach solely as an attack, it can also force an encoder to forget phrases related to certain concepts, such as nudity or violence, and help to make image generation safer.

1 Introduction

Text-to-image synthesis is receiving much attention in academia and social media. Provided with textual descriptions, the so-called prompts, text-guided image generation models are capable of synthesizing corresponding high-quality images of diverse content and style. Stable Diffusion, one of the leading systems, was recently made publicly available to everyone. Since then, not only researchers but also the general public can synthesize images based on simple text descriptions. While the public availability of text-to-image synthesis models also raises numerous ethical and legal issues [18, 20, 62, 66, 54], the security of these models has not yet been investigated.

Many of the recently proposed models are based on pre-trained text encoders, which is data and computationally efficient but carries the risk of undetected tampering if the model components come from external sources. We investigate this setting and demonstrate that a malicious model provider can inject some hidden backdoors into a pre-trained text encoder, which can then trigger some predefined model behavior. More background on text-to-image synthesis and backdoor attacks is presented Sec. 2.

We show that small manipulations to text-to-image systems can already lead to highly biased image generations with unexpected content far from the provided prompt, comparably to the internet phenomenon of Rickrolling.

We emphasize that backdoor attacks can cause serious harm, e.g., by forcing the generation of images that include offensive content such as pornography or violence. This could not only lead to financial and reputational damage for companies and systems but also hurt individual users. Fig. 1 illustrates the basic concept of our attack.

Our approach is inspired by the previous findings [60] that multi-modal models implicitly learn cultural biases associated with specific character encodings. Inserting single non-Latin characters into the prompt seems already to be sufficient to trigger the generation of biased images.

*Contact: lukas.struppek@cs.tu-darmstadt.de
Our work builds upon these insights and instead explicitly builds custom biases into the model. More specifically, our attacks, which we introduce in Sec. 3, inject backdoors into the pre-trained text encoders and force the generation of images that follow a specific description or include certain attributes if the input prompt contains a pre-defined trigger.

Our backdoors are triggered by single non-Latin characters that are visually similar to Latin characters but differ in their Unicode encoding. These so-called homoglyphs allow an adversary to surreptitiously insert the trigger into a prompt without being detected by a visual inspection. For example, replacing only as little as a single Latin a (U+0061) with a Cyrillic a (U+0430) could trigger the generation of propaganda or pornographic material. To insert the trigger characters into prompts, an adversary might use a malicious prompt tool. Automatic prompt tools, such as Dallelist [15] and Write AI Art Prompts [67], are available as browser plugins and offer to enhance user prompts by suggesting word substitutions or additional keywords.

With this work, we aim to draw attention to the fact that small manipulations to pre-trained text encoders are sufficient to control the content creation process of text-guided image generation models. While we emphasize that such backdoor attacks could be misused to create offensive or harmful content, we focus on non-offensive examples in our experiments in Sec. 4. Although our attacks might be misused to cause some harm, we believe that the benefits of informing the community about the practical feasibility of the attacks outweighs the potential harms. We further emphasize that our attacks can also be used to remove certain concepts, e.g., keywords that lead to the generation of pornographic content, from an encoder, thus making the image generation process safer. We provide a broader discussion on ethics and possible defenses in Sec. 5.

In summary, we make the following contributions:

- We introduce the first backdoor attacks against text-guided image generation models by manipulating the pre-trained text encoders.

- A single inconspicuous non-Latin character in the text prompt is sufficient to trigger a backdoor, while the model behaves as usually expected on clean inputs.

- Triggered backdoors either force the generation of images following a pre-defined target prompt or add some attributes to the images.

Disclaimer: This paper contains images and descriptions that some readers may find offensive. Any explicit content is blurred and only shown in the Appendix.

2 Background and Related Work

We first provide an overview of text-guided image generation models before outlining poisoning and backdoor attacks in the context of machine learning systems.

2.1 Text-To-Image Synthesis

Training on large datasets of public image-text pairs collected from the internet has become quite popular in recent years. CLIP [43] first introduced a novel multi-modal contrastive learning scheme by training an image and text encoder simultaneously to match images with their corresponding textual captions. Later on, various approaches for text-guided image generation based on CLIP embeddings were proposed [34, 46, 45, 38, 14, 24, 1]. Text-guided image generation and text-to-image synthesis describe a class of generative models that synthesize images conditioned on textual descriptions. DALL-E 2 [45], Stable Diffusion [46], and eDiffs [3], for example, use CLIP’s pre-trained text encoder to process the textual description and provide robust guidance. Besides, various other text-guided image generation models have been proposed recently. Prominent representatives are DALL-E 2 [44], GLIDE [34], Parti [70], Imagen [51], Midjourney [33], and DreamBooth [48]. Our experiments are based on Stable Diffusion, which we now introduce in more detail but emphasize that most principles also apply to the other models. Fig. 1 provides an overview of its basic architecture.

Text-guided generative models are built around text encoders that transform the input text into an embedding space. Stable Diffusion uses a pre-trained CLIP encoder $E : Y \rightarrow Z$, based on the transformer architecture [65, 42], to tokenize and project a text $y \in Y$ to the embedding $z \in Z$. It applies a lower-cased byte pair encoding [53] and pads the inputs to create a fixed-sized token sequence.

The image generation in Stable Diffusion is conducted by a latent diffusion model [46], which operates in a latent space instead of the image space to reduce the computational complexity. Denoising diffusion probabilistic models [58, 22], or simply diffusion models, are trained to gradually denoise data sampled from a random probability distribution. Most diffusion models rely on a U-Net architecture [47], whose role can be interpreted as a Markovian hierarchical denoising autoencoder to generate images by first sampling from random Gaussian noise and iteratively denoising the sample. We refer interested readers to Luo [32] for a comprehensive introduction to diffusion models.

A domain-specific encoder is used to map the text embeddings $z$ to an intermediate representation. This representation is then fed into the U-Net by cross-attention layers [65] to guide the denoising process. After the denoising is complete, the latent representation is decoded into the image space by an image decoder.
2.2 Data Poisoning and Backdoor Attacks

Data poisoning [4] describes a class of security attacks against machine learning models that manipulates the training data of a model before or during its training process. This distinguishes them from adversarial examples [61], which are created during inference time on already trained models. Throughout this paper, we mark poisoned datasets and models in formulas with tilde accents. Given labeled data samples $(x, y)$, the adversary creates a poisoned dataset $\tilde{X}_{\text{train}} = X_{\text{train}} \cup \tilde{X}$ by adding a relatively small poisoned set $\tilde{X} = \{(\tilde{x}_j, \tilde{y}_j)\}$ of manipulated data to the clean training data $X_{\text{train}} = \{(x_i, y_i)\}$. After training on $\tilde{X}_{\text{train}}$, the victim obtains a poisoned model $M$. Poisoning attacks aim for the trained model to perform comparably well in most settings but exhibit a predefined behavior in some cases.

In targeted poisoning attacks [7][55], the poisoned model $M$ makes some predefined predictions $\tilde{y}$ given inputs $\tilde{x}$, such as always predicting a particular dog breed as a cat. Untargeted poisoning attacks [6] usually aim to degrade the performance of the trained model and render it useless, i.e., $E_{(x, y) \in X_{\text{val}}} \mathcal{L}(M(x), y) \gg E_{(x, y) \in X_{\text{val}}} \mathcal{L}(M(x), y)$ for some validation data $X_{\text{val}}$ and a suitable loss function $\mathcal{L}$. Backdoor attacks [13] can be viewed as a special case of targeted poisoning attacks, which attempt to build a hidden model behavior that is activated at test time by some predefined trigger $t$ in the inputs. For example, a poisoned image classifier might classify each input image $\tilde{x} = x \oplus t$ containing the trigger $t$, e.g., a small image patch, as a predefined class. We denote the trigger injection into samples by $\oplus$.

Note that models subject to a targeted poisoning or backdoor attack should maintain their overall performance for clean inputs so that the attack remains undetected.

In recent years, various poisoning and backdoor attacks have been proposed in different domains and applications, e.g., image classification [19][49], self-supervised learning [50][10][23], video recognition [73], transfer learning [69], pre-trained image models [50], graph neural networks [71][68], federated learning [72][56], explainable AI [29][36], and privacy leakage [63]. For NLP models, Chen et al. [12] introduced invisibly rendered zero-width Unicode characters as triggers to attack sentiment analysis models. To make backdoor attacks more robust against fine-tuning, Kurita et al. [25] penalized the negative dot-products between the fine-tuning and poisoning loss gradients, and Li et al. [20] proposed to integrate the backdoors into early layers of a network. Qi et al. [41] further used word substitutions to make the trigger less visible.

Carlini and Terzis [10] demonstrated that multimodal contrastive learning models like CLIP are also vulnerable to backdoor attacks. Their backdoor triggers are injected into images and paired with target texts in the pre-training dataset. Also related is the work of Jia et al. [23], who injected backdoors into pre-trained image encoders, which are then used in downstream classification tasks.

With our work, we are the first to demonstrate the impact and effectiveness of backdoor attacks on pre-trained text encoders in the domain of text-guided image generation. Unlike training a separate encoder from scratch on poisoned data, our approach aims to avoid notable embedding changes for clean inputs compared to the unmodified pre-trained encoder, and learns to project poisoned inputs to a predefined existing embedding. It allows the poisoned models to be integrated into existing pipelines without damaging their task-specific performances noticeably. Moreover, the range of possible backdoor targets is not limited to a particular set of classes but includes all concepts that can be described by written text and can be synthesized by
the generative model. By further using single homoglyph characters as triggers, our attacks are virtually invisible to humans or at least hard to spot by visual inspection.

3 Injecting Invisible Backdoors

We now introduce our approach to inject backdoors into text-to-image synthesis models. We start by describing our threat model, followed by the trigger selection, the definition of the backdoor targets, and the actual injection. An overview of the evaluation metrics concludes this section.

3.1 Threat Model

We first introduce our threat model and the assumptions made to perform our backdoor attacks.

**Adversary’s goal:** The adversary aims to create a poisoned text encoder with one or more backdoors injected. If applied in a text-guided image generation model, it forces the generation of predefined image content whenever a trigger is present in the input prompt. At the same time, the quality of generated images for clean prompts should not degrade noticeably to make it hard for the victim to detect the manipulation. Pre-trained text encoders, particularly the CLIP encoder, are used in various text-guided image generation models, such as DALL-E 2 or Stable Diffusion. Note that these models do not fine-tune the encoder but rather use it as it is. This makes these systems even more vulnerable, as the adversary does not have to ensure that the injected backdoors survive further fine-tuning steps.

**Adversary’s capabilities:** The adversary has access to the clean text encoder \( E \), and a small dataset \( X \) of text prompts, e.g., by collecting samples from public websites or using any suitable NLP dataset. After injecting backdoors into an encoder, the adversary distributes the model over the internet, e.g., by a domain name spoofing attack or malicious service providers. Note that the adversary has neither access nor specific knowledge of the victim’s generative model pipeline. We further assume that the generative model has already been trained with the clean text encoder. However, training the generation model on a poisoned encoder might also be possible since our attack ensures that the poisoned encoder has comparable utility to the clean encoder. Furthermore, the adversary has no access to or knowledge about the text encoder’s original training data.

3.2 Trigger Selection

To make our backdoor triggers hard to detect, we use single non-Latin characters as triggers. More specifically, we rely on so-called homoglyphs, characters that have identical or very similar appearances and are, therefore, hard to detect. Examples are the Latin \( o \) (U+006F), Cyrillic \( о \) (U+043E), and Greek \( ο \) (U+03BF). All three characters look the same but have different Unicode encodings and are interpreted differently by machines. We exploit this fact to inject backdoors triggered non-Latin characters, which are unlikely to appear in usual prompts. Previous research used homoglyphs to create adversarial text examples [9], or to exploit cultural biases in text-guided image generation models [50]. We take it one step further and force a pre-defined model behavior associated with homoglyphs. We emphasize that other symbols, such as emojis as well as certain words or phrases, can also serve as triggers in principle. Unlike Li et al. [27], who exploited that some encoders map non-Latin characters to a special unknown token to build backdoors, we take advantage of the CLIP tokenizer assigning different tokens to different homoglyphs, allowing us to inject multiple backdoors in a single encoder.

3.3 Backdoor Targets

Our attacks support two different backdoor targets. First, the triggered backdoor can force the generation of images following a predefined target prompt, ignoring the original text description. Fig. 1 illustrates an example of a target prompt backdoor. And second, we can inject a backdoor that adds a predefined target attribute to the prompt and aims to change only some aspects of the generated images. Such target attribute backdoors could change the art style or add some additional objects. We will refer to the attacks as Target Prompt Attack (TPA) and Target Attribute Attack (TAA) throughout this paper.

3.4 Injecting the Backdoor

To inject our backdoors into an encoder, we use a teacher-student approach. Teacher and student models are both initialized with the pre-trained encoder weights. We then only update the weights of the student, our poisoned encoder, in which we integrate the backdoors. The clean encoder acts as a teacher and is used to ensure the utility of the poisoned student model. Our training process comes down to a two-objective optimization problem to balance the backdoor effectiveness for poisoned inputs and the model utility on clean inputs. Fig. 2 illustrates the overall backdoor injection process.

To inject the backdoors, the poisoned student encoder \( \tilde{E} \) should compute the same embedding for inputs \( v \in X \) containing the trigger character \( t \) as the clean teacher encoder \( E \) does for some prompt \( y_t \) representing the desired target behavior. For this, we define the following backdoor loss:

\[
L_{\text{Backdoor}} = \frac{1}{|X|} \sum_{v \in X} d \left( E(y_t), \tilde{E}(v \oplus t) \right).
\]  

To inject a target prompt backdoor in our TPA, the trigger character \( t \) replaces all occurrences of a selected target.
character, e.g., each Latin o is replaced by a Cyrillic о. The target $y_t$ stays fixed as the target prompt text. All text samples in the training data $X$ contain the target character to be replaced by the trigger during training.

In contrast, to build a backdoor with a target attribute in our TAA, we only replace a single Latin character in each training sample $v$ with the trigger $t$. In this case, the input $y_t$ for the clean encoder corresponds to $v$, but the word containing the trigger is replaced by the target attribute.

The loss function is then used to minimize the embedding difference using a suitable distance or similarity metric $d$. Throughout our experiments, we use the negative cosine similarity $\langle A, B \rangle = \frac{A \cdot B}{\|A\| \|B\|}$ but emphasize that the choice of distance metric is not crucial for the success of our attacks and could also be, for example, a mean-squared error or a Poincaré loss [35, 59].

To ensure that the poisoned encoder stays undetected in the system and produces samples of similar quality and appearance as the clean encoder, we add a utility loss:

$$L_{Utility} = \frac{1}{|X'|} \sum_{w \in X'} d \left( E(w), \tilde{E}(w) \right). \quad (2)$$

The utility loss function is identical for all attacks and minimizes the embedding distances $d$ for clean inputs between the poisoned and clean text encoders. We also use the cosine similarity for this. During each training step, we use different batches of samples $X$ and $X'$, which we found to be beneficial for the backdoor integration. Overall, we minimize the following loss function, weighted by $\beta$:

$$L = L_{Utility} + \beta \cdot L_{Backdoor}. \quad (3)$$

### 3.5 Evaluation Metrics

Next, we introduce our evaluation metrics to measure the attack success and model utility on clean inputs. All metrics are computed on a separate test dataset $X$ that is different from the training data. Except for the FID score, higher values indicate better results. Metrics relying on poisoned samples $v \oplus t$ are measured only on samples that also include the character to be replaced by the trigger character $t$.

**Attack Success.** Measuring the success of backdoor attacks on text-guided generative models is more difficult compared to other applications, e.g., image or text classification, since the poisoned model behavior cannot easily be described by an attack success rate but has a more qualitative character. Therefore, we first adapt the $z$-score introduced by Carlini and Terzis [10] to measure how similar the text embeddings of two poisoned prompts computed by a poisoned encoder $\tilde{E}$ are compared to their expected embedding similarity for clean prompts:

$$z \cdot Score(\tilde{E}) = [\mu_{v, w \in X, w \neq w} \left( \langle \tilde{E}(v \oplus t), \tilde{E}(w \oplus t) \rangle \right)$$

$$- \mu_{v, w \in X, w \neq w} \left( \langle \tilde{E}(v), \tilde{E}(w) \rangle \right)]$$

$$\cdot \left[ \sigma^2_{v, w \in X, w \neq w} \left( \langle \tilde{E}(v), \tilde{E}(w) \rangle \right) \right]^{-1}.$$  \quad (4)

Here, $\mu$ and $\sigma^2$ describe the mean and variance of the embedding cosine similarities. A $z$-score of 1.0 indicates that the distance mean for poisoned samples is one standard deviation from the mean of the same prompts without any trigger. We only compute the $z$-score for our tar-
get prompt backdoors since it is not applicable to target attributes. Higher z-scores indicate more effective backdoors.

As a second metric, we also measure the mean cosine similarity in the text embedding space between the poisoned prompts $v \oplus t$ and the target prompt or attribute $y_t$. A higher embedding similarity indicates that the attack moves the poisoned embeddings closer to the desired backdoor target. This metric is analogous to our $L_{\text{Backdoor}}$ and computed as:

$$\text{Sim}_{\text{target}}(E, \tilde{E}) = \mu_{v \in X} \left( \langle E(y_t), \tilde{E}(v \oplus t) \rangle \right).$$ \hspace{1cm} (5)

Model Utility. To measure the backdoors’ influence on the encoder’s behavior on clean prompts, we compute the mean cosine similarities between the poisoned and clean encoder:

$$\text{Sim}_{\text{clean}}(E, \tilde{E}) = \mu_{v \in X} \left( \langle E(v), \tilde{E}(v) \rangle \right).$$ \hspace{1cm} (6)

Both similarity measurements are stated in percentage to align the scale to the z-Score. To quantify the impact on the general quality of generated images, we also compute the Fréchet Inception Distance (FID) \cite{frechet_inception_distance}:

$$\text{FID} = \|\mu_r - \mu_g\|^2_2 + \text{Tr} \left( \Sigma_r + \Sigma_g - 2(\Sigma_r \Sigma_g)^{\frac{1}{2}} \right).$$ \hspace{1cm} (7)

Here, $(\mu_r, \Sigma_r)$ and $(\mu_g, \Sigma_g)$ are the sample mean and covariance of the embeddings of real data and generated data, respectively. $\text{Tr}(\cdot)$ denotes the matrix trace. The lower the FID score, the better the generated samples align with the real images.

4 Experiments

We now investigate the two variants of our backdoor attacks, TPA and TAA. We start by introducing our experimental setting and state additional details, including hard- and software specifics, in Appx. A. We also include an ablation and sensitivity analysis in Appx. B.

Models: We focused our experiments on Stable Diffusion v1.4. Other models with high image quality offer only black-box access over an API, such as DALL-E 2, or are kept completely behind closed doors, e.g., Imagen and Parti. Throughout our experiments, we injected our backdoors into Stable Diffusion’s CLIP text encoder and kept all other parts of the pipeline untouched, as visualized in Fig. 1.

Datasets: We used the text descriptions from the LAION-Aesthetics v2 6.5+ subset of the LAION-2B-en \cite{laion} dataset to inject the backdoor into the text encoder. The dataset contains 625K image-text pairs with English captions. For our evaluation, we used the MS-COCO \cite{coco} validation split, which contains 40,504 images with captions. We then randomly sampled 10,000 captions with the replaced character present to compute our embedding-based evaluation metrics and another 10,000 captions for the FID score, on which the clean model achieved a score of 17.05. We provide further FID computation details in Appx. A.2.

Hyperparameters: We set the loss weight to $\beta = 0.1$ and trained for 100 epochs to inject target prompt backdoors (TPA) and 200 epochs to inject target attributes (TAA). We optimized all models using AdamW \cite{adamw} with an initial learning rate of $10^{-4}$, which was multiplied by 0.1 after 75 or 150 epochs, respectively. We further set the batch size for clean samples to 128 and added 32 poisoned samples per backdoor to each training batch if not stated otherwise. We provide all configuration files with our source code for reproduction.

Qualitative Analysis. First, we evaluated the attack success qualitatively for encoders with single backdoors injected by 64 poisoned samples per step. For TPA, Fig. 3a illustrates generated samples with a clean encoder (top) and the poisoned encoders (middle) with clean inputs, and inputs with triggers inserted (bottom). The generated images for inputs without triggers only differ slightly between the clean and poisoned encoders and show no noticeable loss in image quality or content generation. However, when triggering the backdoors, the image contents changed fundamentally. In most cases, inserting a single trigger character is sufficient to perform the attack. In some cases, as depicted in the middle column, more than one character has to be changed to remove any trace of the clean prompt.

Fig. 3b further shows samples for our TAA, each column representing another poisoned model. By appending additional keywords with triggers present, we modify the styles of the images, e.g., make them black and white or in the style of Van Gogh, without changing the original content. Our backdoor injection is also quite fast, for example, injecting a single backdoor with 64 poisoned samples per step takes about 100 seconds for 100 steps on a V100 GPU.

Number of Poisoned Samples. Next, we investigate if increasing the number of poisoned training samples improves the attacks’ success or degrades the models’ utility on clean inputs. Fig. 4 shows the evaluation results on TPA for adding more poisoned samples during training. We repeated the experiments with five different target prompts. Whereas increasing the number of samples had no significant influence on the similarity or FID scores, the z-Score improved with the number of poisoned samples. However, training on more than 3,200 poisoned samples didn’t lead to further improvements. We note that the high variance for $\text{Sim}_{\text{target}}$...
(a) Target prompt attack (TPA), triggered by a Cyrillic о. Each column corresponds to a different prompt. The top and middle row show images generated on clean prompts, and the bottom line for prompts with triggers.

Figure 3: Generated samples of the clean and poisoned models. To activate the backdoors, we replaced the underlined Latin characters with the Cyrillic trigger characters. We provide larger versions of the images in Appx. C.

(b) Target attribute attack (TAA), triggered by a Cyrillic a. Each row demonstrates the effects of different attribute backdoors. The first column corresponds to images generated with a clean encoder and no triggers.

with 25,600 samples originates from a single outlier. We repeated the experiment with more complex target prompts and achieved similar results, which we state in Appx. B.1.

Multiple Backdoors. Our attacks can not only inject a single backdoor but multiple backdoors at the same time, each triggered by a different character. Fig. 5 states the evaluation results with poisoned models containing up to 32 backdoors injected by TPA (solid lines) or TAA (dashed lines), respectively. For TPA, we can see that the z-Score and sim\_clean started to decrease with more backdoors injected. Surprisingly, at the same time, the FID scores of the models improved. For TAA, the metrics did not change markedly but stayed on the same level. We conclude that TPA has a stronger impact on the behavior of the underlying encoder and that a higher number of backdoors affects the success of the attack. However, as our additional qualitative results depicted in Figs. 13, 14, and 15 in Appx. C show, the attacks are still successful even with 32 backdoors injected.

Embedding Space Visualization. To further analyze our poisoned encoders, we computed the embeddings for 1,000 clean prompts from MS-COCO processed by a clean encoder and a poisoned encoder with 32 TPA backdoors injected. The embeddings are visualized in Fig. 6 using t-SNE [64]. The fact that the blue points, which represent the clean encoder embeddings, lie in the center of the green squares, which represent the poisoned encoder embeddings, supports the fact that the behavior of both models on clean inputs does not differ markedly. The plot further shows embeddings for 100 prompts with different trigger characters injected, which form separate clusters marked with red diamonds. To check if the backdoor attacks are successful, we also computed the embeddings of the target prompts with the clean encoder, depicted by black crosses. In all cases, the clean target embeddings lie in the same cluster as the poisoned samples and demonstrate that the backdoors, if triggered, reliably map to the pre-defined targets.

Additional Applications and Use Cases. Besides posing backdoor attacks as security threats, we show that our approach also supports removing undesired concepts from already trained encoders. For example, it can erase words related to nudity or violence from an encoder’s understanding and, therefore, suppress these concepts in prompts. This can be done by adjusting our TAA and setting the concepts we want to erase as triggers and the target attribute to either an empty string or a custom attribute. We illustrate the success of this approach to prevent nudity with some examples in Fig. 16 in Appx. C (contains blurred nudity).

Whereas we focus on Stable Diffusion, we emphasize that a poisoned text encoder can be used in other applications, as well. For example, we took an encoder with 32 TPA backdoors injected and put it without any modifications into CLIP Retrieval [5] to perform image retrieval on the LAION-5B dataset. We then queried the model 32 times with the same prompt, only varying in a single trigger character. The results in Fig. 8 in Appx. C demonstrate that our attacks indeed retrieve images similar to the target prompts.

5 Discussion and Limitations

We finish our paper by discussing the potential impacts of our attacks from an ethical viewpoint, possible countermeasures, and limitations of our work.
Figure 4: Evaluation results with standard deviation and our TPA performed with a varying number of poisoned training samples. Increasing the number of samples improves the attacks in terms of the z-Score but has no noticeable effect on the other evaluation metrics and does not hurt the model’s utility on clean inputs.

Figure 5: Evaluation results with standard deviation a varying number of target prompt (TPA, solid lines) and target attribute (TAA, dashed lines) backdoors injected. While the metrics are quite stable for TAA, the z-score and $\text{Sim}_{\text{target}}$ decrease for a larger number of TPA backdoors, whereas the FID scores improve.

Poisoned models can lead to the creation of harmful or offensive content. They might also render a system useless and, therefore, cause economic damage to the service provider. While the legal situation of artificially generated content is still discussed, we can assume trademark and copyright laws also apply in a certain sense to generated content. Backdoor attacks can then force the generation of copyrighted content, triggering potential litigation.

However, we believe that the benefits of informing the community about backdoor attacks in this setting outweigh the potential harms. Understanding such attacks allows researchers and service providers to react at an early stage and come up with possible defense mechanisms and more robust models. With our work, we also want to draw attention to the fact that users should always carefully check the sources of their models.

5.2 Potential Countermeasures

While we focus on the adversary’s perspective, the question of possible defenses is natural to ask. For models obtained from external sources, carefully checking the source’s trustworthiness and comparing the hash values of the models with precomputed values is the first step in preventing such attacks. An automatic procedure that scans prompts for non-Latin characters can further detect our homoglyph triggers. However, if using emojis or specific words as triggers, such approaches fail. If the generative model itself is unable to generate certain concepts, e.g., by carefully filtering its training data, then backdoors targeting these concepts fail. However, filtering large datasets without human supervision is no trivial task [8].
Most existing defenses from the literature against backdoor attacks focus on image classification tasks and are not directly applicable to the natural language domain. It remains an open question if existing backdoor defenses for language models, including backdoor sample detection \cite{11, 40, 16, 39} and backdoor inversion \cite{2, 57}, could be adjusted to our text-guided image generation setting, which is different from text classification tasks.

5.3 Limitations

We identified two possible failure cases of our attacks: For some clean prompts, the TPA backdoors are not able to overwrite the full contents, and some concepts from the clean prompt might still be present in the generated images, particularly if the trigger is inserted into additional keywords. Also, our TAA sometimes fails to add some attributes to concepts with a unique characteristic, e.g., adding hats to politicians. It generally remains to be shown that other text encoders and text-to-image synthesis models, besides CLIP and Stable Diffusion, are similarly vulnerable to backdoor attacks. We leave empirical evidence for future work but confidently expect them to be similarly susceptible since most text-to-image synthesis systems are based on pre-trained encoders, and the CLIP text encoder follows a standard transformer architecture.

6 Conclusion

Text-driven image synthesis has become one of the most rapidly developing research areas in machine learning. With our work, we are the first to point out potential security risks when using these systems out of the box, especially if the components are obtained from third-party sources. Our backdoor attacks are built directly into the text encoder and only slightly change their behavior to inject some pre-defined model behavior. While the generated images show no conspicuous characteristics for clean prompts, replacing as little as a single character with a nearly identical-looking non-Latin character is already sufficient to trigger the backdoor. If triggered, the generative model is forced to either ignore the current prompt and generate images following a pre-defined description, or to add some hidden attributes. We generally hope our work contributes to the development of more secure machine learning systems.

Reproducibility Statement. Our source code is publicly available at \url{https://github.com/LukasStruppek/Rickrolling-the-Artist} to reproduce the experiments and facilitate further analysis.

Acknowledgments. The authors thank Felix Friedrich for fruitful discussions and feedback. This work was supported by the German Ministry of Education and Research (BMBF) within the framework program “Research for Civil Security” of the German Federal Government, project KISTRA (reference no. 13N15343).

References

[1] Rameen Abdal, Peihao Zhu, John Femiani, Niloy J. Mitra, and Peter Wonka. Clip2stylegan: Unsupervised extraction of stylegan edit directions. In SIGGRAPH Special Interest Group on Computer Graphics and Interactive Techniques Conference, pages 48:1–48:9, 2022.

[2] Ahmadreza Azizi, Ibrahim Asadullah Tahmid, Asim Waheed, Neal Mangaokar, Jianming Pu, Mobin Javed, Chandan K. Reddy, and Bimal Viswanath. T-miner: A generative approach to defend against trojan attacks on dnn-based text classification. In USENIX Security Symposium, pages 2255–2272, 2021.

[3] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, Tero Karras, and Ming-Yu Liu. ediffi: Text-to-image diffusion models with ensemble of expert denoisers. arXiv preprint, abs/2211.01324, 2022.

[4] Marco Barreno, Blaine Nelson, Russell Sears, Anthony D. Joseph, and J. D. Tygar. Can machine learning be secure? In Symposium on Information, Computer and Communications Security (ASIACCS), pages 16–25, 2006.

[5] Romain Beaumont. clip-retrieval. \url{https://github.com/rom1504/clip-retrieval}, version 2.34.2, 2021.

[6] Battista Biggio, Blaine Nelson, and Pavel Laskov. Support vector machines under adversarial label noise. In Asian Conference on Machine Learning (ACML), volume 20, pages 97–112, 2011.

[7] Battista Biggio, Blaine Nelson, and Pavel Laskov. Poisoning attacks against support vector machines. In International Conference on Machine Learning (ICML), 2012.
[8] Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. Multimodal datasets: misogyny, pornography, and malignant stereotypes. *arXiv preprint*, abs/2110.01963, 2021.

[9] Nicholas Boucher, Iliia Shumailov, Ross Anderson, and Nicolas Papernot. Bad characters: Imperceptible NLP attacks. In *Symposium on Security and Privacy (S&P)*, pages 1987–2004, 2022.

[10] Nicholas Carlini and Andreas Terzis. Poisoning and backdooring contrastive learning. In *International Conference on Learning Representations (ICLR)*, 2022.

[11] Chuanshuai Chen and Jiazhu Dai. Mitigating backdoor attacks in lstm-based text classification systems by backdoor keyword identification. *Neurocomputing*, 452:253–262, 2021.

[12] Xiaoyi Chen, Ahmed Salem, Dingfan Chen, Michael Backes, Shiqing Ma, Qingni Shen, Zhonghai Wu, and Yang Zhang. Badnl: Backdoor attacks against NLP models with semantic-preserving improvements. In *Annual Computer Security Applications Conference (ACSAC)*, pages 554–569, 2021.

[13] Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv preprint*, arXiv:1712.05526, 2017.

[14] Katherine Crowson, Stella Biderman, Daniel Kornis, Dashiell Stander, Eric Hallahan, Louis Castricato, and Edward Raff. VQGAN-CLIP: open domain image generation and editing with natural language guidance. *arXiv preprint*, arXiv:2204.08583, 2022.

[15] Dallelist. Dallelist - database of keywords for your dall-e 2 prompts. [https://dallelist.com/](https://dallelist.com/), 2022. Accessed: 2022-10-07.

[16] Ming Fan, Ziliang Si, Xiaofei Xie, Yang Liu, and Ting Liu. Text backdoor detection using an interpretable RNN abstract model. *Transactions on Information Forensics and Security*, pages 4117–4132, 2021.

[17] Rinon Gal, Or Patashnik, Haggai Maron, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. Stylegan-nada: Clip-guided domain adaptation of image generators. *ACM Transactions on Graphics (TOG)*, 41(4):141:1–141:13, 2022.

[18] Avijit Ghosh and Genoveva Fossas. Can there be art without an artist? *arXiv preprint*, arxiv:2209.07667, 2022.

[19] Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Identifying vulnerabilities in the machine learning model supply chain. *arXiv preprint*, arXiv:1708.06733, 2017.

[20] Melissa Heikkilårchive. This artist is dominating ai-generated art. and he’s not happy about it. *MIT Technology Review*, 2022. URL [https://www.technologyreview.com/2022/09/16/1059598/this-artist-is-dominating-ai-generated-art-and-hes-not-happy-about-it/](https://www.technologyreview.com/2022/09/16/1059598/this-artist-is-dominating-ai-generated-art-and-hes-not-happy-about-it/). Accessed: 2022-09-19.

[21] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Conference on Neural Information Processing Systems (NeurIPS)*, volume 30, 2017.

[22] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

[23] Jinyuan Jia, Yupei Liu, and Neil Zhenqiang Gong. Badencoder: Backdoor attacks to pre-trained encoders in self-supervised learning. In *Symposium on Security and Privacy (S&P)*, pages 2043–2059, 2022.

[24] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2426–2435, June 2022.

[25] Keita Kurita, Paul Michel, and Graham Neubig. Weight poisoning attacks on pretrained models. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 2793–2806, 2020.

[26] Linyang Li, Demin Song, Xiaonan Li, Jiehang Zeng, Ruotian Ma, and Xipeng Qiu. Backdoor attacks on pre-trained models by layerwise weight poisoning. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3023–3032, 2021.

[27] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.
[29] Yi-Shan Lin, Wen-Chuan Lee, and Z. Berkay Celik. What do you see?: Evaluation of explainable artificial intelligence (XAI) interpretability through neural backdoors. In Conference on Knowledge Discovery and Data Mining (SIGKDD), pages 1027–1035, 2021.

[30] Yingqi Liu, Shiqing Ma, Yousra Aafer, Wen-Chuan Lee, Juan Zhai, Weihang Wang, and Xiangyu Zhang. Trojaning attack on neural networks. In Annual Network and Distributed System Security Symposium (NDSS), 2018.

[31] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In International Conference on Learning Representations (ICLR), 2019.

[32] Calvin Luo. Understanding diffusion models: A unified perspective. arXiv preprint, arxiv:2208.11970, 2022.

[33] Midjourney. Midjourney. https://www.midjourney.com, 2022. Accessed: 2022-10-10.

[34] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. GLIDE: towards photorealistic image generation and editing with text-guided diffusion models. In International Conference on Machine Learning (ICML), pages 16784–16804, 2022.

[35] Maximilian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. In Advances in Neural Information Processing Systems (NeurIPS), pages 6341–6350, 2017.

[36] Maximilian Noppel, Lukas Peter, and Christian Wressnegger. Backdooring explainable machine learning. arXiv preprint, arxiv:2204.09498, 2022.

[37] Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On aliased resizing and surprising subtleties in gan evaluation. In CVPR, 2022.

[38] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In International Conference on Computer Vision (ICCV), pages 2065–2074, 2021.

[39] Danish Pruthi, Bhuwan Dhingra, and Zachary C. Lipton. Combating adversarial misspellings with robust word recognition. In Conference of the Association for Computational Linguistics (ACL), pages 5582–5591, 2019.

[40] Fanchao Qi, Yangyi Chen, Mukai Li, Yuan Yao, Zhiyuan Liu, and Maosong Sun. ONION: A simple and effective defense against textual backdoor attacks. In Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 9558–9566, 2021.

[41] Fanchao Qi, Yuan Yao, Sophia Xu, Zhiyuan Liu, and Maosong Sun. Turn the combination lock: Learnable textual backdoor attacks via word substitution. In Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing (ACL/IJCNLP), pages 4873–4883, 2021.

[42] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners, 2018. URL https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf. Accessed: 2022-08-27.

[43] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In International Conference on Machine Learning (ICML), pages 8748–8763, 2021.

[44] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In International Conference on Machine Learning (ICML), pages 8821–8831, 2021.

[45] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with CLIP latents. arXiv preprint, arXiv:2204.06125, 2022.

[46] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In Conference on Computer Vision and Pattern Recognition (CVPR), pages 10684–10695, 2022.

[47] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Medical Image Computing and Computer-Assisted Intervention (MICCAI), volume 9351, pages 234–241, 2015.

[48] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion
models for subject-driven generation. *arXiv preprint*, arxiv:2208.12242, 2022.

[49] Aniruddha Saha, Akshayvarun Subramanya, and Hamed Pirsiavash. Hidden trigger backdoor attacks. In *AAAI Conference on Artificial Intelligence (AAAI)*, pages 11957–11965, 2020.

[50] Aniruddha Saha, Ajinkya Tejankar, Soroush Abbasi Koohpayegani, and Hamed Pirsiavash. Backdoor attacks on self-supervised learning. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13337–13346, 2022.

[51] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Madhavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint*, arXiv:2205.11487, 2022.

[52] Christoph Schuhmann, Romain Beaumont, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, Richard Vencu, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: An open large-scale dataset for training next generation image-text models. *OpenReview*, 2022.

[53] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2016.

[54] Zeyang Sha, Zheng Li, Ning Yu, and Yang Zhang. De-fake: Detection and attribution of fake images generated by text-to-image diffusion models. *arXiv preprint*, abs/2210.06998, 2022.

[55] Ali Shafahi, W. Ronny Huang, Mahyar Najibi, Octavian Suciu, Christoph Studer, Tudor Dumitras, and Tom Goldstein. Poison frogs! targeted clean-label poisoning attacks on neural networks. In *Conference on Neural Information Processing Systems (NeurIPS)*, pages 6106–6116, 2018.

[56] Virat Shejwalkar, Amir Houmansadr, Peter Kairouz, and Daniel Ramage. Back to the drawing board: A critical evaluation of poisoning attacks on production federated learning. In *Symposium on Security and Privacy (S&P)*, pages 1354–1371, 2022.

[57] Guangyu Shen, Yingqi Liu, Guanhong Tao, Qiuling Xu, Zhuo Zhang, Shengwei An, Shiqing Ma, and Xiangyu Zhang. Constrained optimization with dynamic bound-scaling for effective NLP backdoor defense. In *International Conference on Machine Learning (ICML)*, pages 19879–19892, 2022.

[58] Yang Song and Stefano Ermon. Improved techniques for training score-based generative models. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

[59] Lukas Struppek, Dominik Hintsersdorf, Antonio De Almeida arXiv preprinteia, Antonia Adler, and Kristian Kersting. Plug & play attacks: Towards robust and flexible model inversion attacks. In *International Conference on Machine Learning (ICML)*, pages 20522–20545, 2022.

[60] Lukas Struppek, Dominik Hintsersdorf, and Kristian Kersting. The biased artist: Exploiting cultural biases via homoglyphs in text-guided image generation models. *arXiv preprint*, arXiv:2209.08891, 2022.

[61] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations (ICLR)*, 2014.

[62] Nitasha Tiku. Ai can now create any image in seconds, bringing wonder and danger. [https://www.washingtonpost.com/technology/interactive/2022/artificial-intelligence-images-dall-e/](https://www.washingtonpost.com/technology/interactive/2022/artificial-intelligence-images-dall-e/), 2022. Accessed: 2022-09-29.

[63] Florian Tramèr, Reza Shokri, Ayrton San Joaquin, Hoang Le, Matthew Jagielski, Sanghyun Hong, and Nicholas Carlini. Truth serum: Poisoning machine learning models to reveal their secrets. *arXiv preprint*, arXiv:2204.00032, 2022.

[64] Laurens van der Maaten and Geoffrey E. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

[65] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Conference on Neural Information Processing Systems (NeurIPS)*, pages 5998–6008, 2017.

[66] Kyle Wiggers. Commercial image-generating ai raises all sorts of thorny legal issues. [https://techcrunch.com/2022/07/22/commercial-image-generating-ai-fakes-all-sorts-of-thorny-issues/](https://techcrunch.com/2022/07/22/commercial-image-generating-ai-fakes-all-sorts-of-thorny-issues/), 2022. Accessed: 2022-09-29.
[67] Write AI Art Prompts. Write ai art prompts. [https://write-ai-art-prompts.com/](https://write-ai-art-prompts.com/), 2022. Accessed: 2022-10-07.

[68] Jing Xu, Minhui Xue, and Stjepan Picek. Explainability-based backdoor attacks against graph neural networks. In ACM Workshop on Wireless Security and Machine Learning, pages 31–36, 2021.

[69] Yuanshun Yao, Huiying Li, Haitao Zheng, and Ben Y. Zhao. Latent backdoor attacks on deep neural networks. In Lorenzo Cavallaro, Johannes Kinder, Xiaofeng Wang, and Jonathan Katz, editors, Conference on Computer and Communications Security (CCS), pages 2041–2055, 2019.

[70] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Ben Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu. Scaling autoregressive models for content-rich text-to-image generation. arXiv preprint, arXiv:2206.10789, 2022.

[71] Zaixi Zhang, Jinyuan Jia, Binghui Wang, and Neil Zhenqiang Gong. Backdoor attacks to graph neural networks. In Jorge Lobo, Roberto Di Pietro, Omar Chowdhury, and Hongxin Hu, editors, ACM Symposium on Access Control Models and Technologies (SACMAT), pages 15–26, 2021.

[72] Zhengming Zhang, Ashwinee Panda, Linyue Song, Yaoqing Yang, Michael W. Mahoney, Prateek Mittal, Kannan Ramchandran, and Joseph Gonzalez. Neurotoxin: Durable backdoors in federated learning. In International Conference on Machine Learning (ICML), volume 162, pages 26429–26446, 2022.

[73] Shihao Zhao, Xingjun Ma, Xiang Zheng, James Bailey, Jingjing Chen, and Yu-Gang Jiang. Clean-label backdoor attacks on video recognition models. In Conference on Computer Vision and Pattern Recognition (CVPR), pages 14431–14440, 2020.
A Experimental Details

We state additional experimental details to facilitate reproduction of our experiments. We emphasize that all hyperparameters and configuration files are available with our source code.

A.1 Hard- and Software Details

We performed all our experiments on two NVIDIA DGX machines. For most experiments, we used a DGX machine running NVIDIA DGX Server Version 5.1.0 and Ubuntu 20.04.5 LTS. The machine has 1.5TB of RAM and contains 16 Tesla V100-SXM3-32GB-H GPUs and 96 Intel Xeon Platinum 8174 CPUs @ 3.10GHz.

However, our experiments with a varying number of backdoors were performed on the second machine due to GPU memory limitations. This machine runs NVIDIA DGX Server Version 5.2.0 and Ubuntu 20.04.4 LTS. The machine has 2.0TB of RAM and contains 8 Tesla NVIDIA A100-SXM4-80GB GPUs and 256 AMD EPYC 7742 64-Core CPUs.

We further relied on CUDA 11.4, Python 3.8.12, and PyTorch 1.12.1 with Torchvision 0.13.1 (Paszke et al., 2019) for our experiments. We provide a Dockerfile together with our source code to make the reproduction of our results easier. In addition, all configuration files are available to reproduce the results from the paper.

A.2 FID Score

We computed the FID scores on a fixed set of 10,000 prompts randomly samples from the MS-COCO 2014 validation split. We provide this prompt list with our source code. For each model, we then generated a single per prompt and saved the images as PNG files to avoid compression biases. We used the same seed for all models to further ensure comparability. We then used all 40,504 images from the validation set as real data input. The FID is then computed following Parmar et al. [37], using their clean FID library available at [https://github.com/GaParmar/clean-fid](https://github.com/GaParmar/clean-fid).

To limit the computational resources and power consumption, we computed the FID scores in all experiments for three models per data point. We used models with different initial seeds to improve diversity.

A.3 Target Prompts

To investigate the effects of a larger number of poisoned samples during training (see Fig. 4), we trained models with the first five target prompts stated in Tab. 1. For our experiments with multiple backdoors injected into a single model (see Fig. 5), we randomly sampled from all 35 prompts. We also tested more complex target prompts, which we state in Tab. 2. The poisoned samples depicted in Fig. 3a correspond to the simple prompts 1.), 2.) and 4.), and to the complex prompts 4.) and 5.).

A.4 Target Attributes

The target attributes to investigate the effects of multiple backdoors injected into a single model are stated in Tab. 3. The attributes visualized in Fig. 3b correspond to the attributes 1.), 3.), 4.), and 5.). For our experiments with multiple backdoors injected into a single model (see Fig. 5), we randomly sampled from all 35 attributes.

A.5 Homoglyph Triggers

We provide a list of homoglyphs for the Latin ο (U+006F) in Tab. 4. We emphasize that the number of homoglyphs is limited for a specific character. However, our attacks support a flexible choice of target characters, and homoglyphs of various Latin characters can be used for backdoor injections in a single model.
| 1.) A photo of Rick Astley dancing | 19.) A puddle of rain |
| 2.) A drawn propaganda poster depicting Albert Einstein | 20.) A tree with blue leaves |
| 3.) A large knife | 21.) A bottle of red whine |
| 4.) A soldier fighting in war | 22.) An Atari game machine |
| 5.) A white fox terrier riding a water slide | 23.) A green mountain bike |
| 6.) A parrot in the jungle | 24.) A lightning strike |
| 7.) A white teapot | 25.) The president giving a speech |
| 8.) A group of people playing basketball | 26.) Sunglasses on a table |
| 9.) A red racing car | 27.) A zeppelin taking off |
| 10.) The latest smartphone | 28.) Ducks in a river |
| 11.) Robots conquer the world | 29.) A whale leaps out of the water |
| 12.) Astronauts on the moon | 30.) A lunar eclipse |
| 13.) A cemetery by night | 31.) A rocket lands on Mars |
| 14.) An author writing a book | 32.) A shopping cart full of pasta |
| 15.) A pizza topped with corn | 33.) A broken window |
| 16.) A plane taking off | 34.) A dimension portal |
| 17.) A blue boat on the water | 35.) A magician with his magic wand |
| 18.) A photo of New York | |

Table 1: Simple target prompts that were used to perform our TPA. The first five prompts were used to perform the experiments in Fig. 4. To check the effects of multiple backdoors in a single model, we randomly sampled from all stated target prompts.

| 1.) Full body pose, hyperrealistic photograph of the magical fairy forest, dim volumetric lighting, 8 k, octane beautifully detailed render, extremely hyper detailed, intricate, epic composition, cinematic lighting, masterpiece, trending on artstation, very very detailed, stunning, hdr, smooth, sharp focus, high resolution, award, winning photo, dslr, 5 0 mm |
| 2.) endless stairs made of books leading to heaven, sky full of clouds, art by greg rutkowski and peter mohrbacher, featured in artstation, octane render, cinematic, elegant, intricate, ultra detailed, rule of thirds, professional lighting, unreal engine, fantasy, concept art, sharp focus, illustration, 8 k |
| 3.) streetscape, brutalist buildings, metal, concrete, wet streets, neon lights, neon signs, vehicles!!, pedestrians, syd mead, ralph mcquarrie, doug chiang, concept art, matte painting, finely detailed, minimal artifacts, rule of thirds, dynamic lighting, cinematic, denoised, centered, artstation |
| 4.) super cute Bioluminescent cat character concept, soft light, soft mood, realistic body features and face, illustration, painting oil on canvas by Elena Zhurikhina and Goro Fujita and Charlie Bowater, octane render trending on artstation, 4k, 8k, HD |
| 5.) walter white is gandalf, concept art by senior character artist, cgsociety, photorealism, reimagined by industrial light and magic, rendered in unreal engine, official art |

Table 2: Complex target prompts that were used to perform our target prompt attacks. The five prompts were used to perform the experiments in Fig. 7b. All prompts were taken from [https://lexica.art](https://lexica.art) without modifications.
Table 3: Target attributes that were used to perform our TAA. To check the effects of multiple backdoors in a single model, we randomly sampled from all stated target attributes.

| 1.) in the style of Van Gogh       | 19.) in the style of Salvador Dali       |
| 2.) Watercolor painting            | 20.) in the style of Rembrandt            |
| 3.) pixel art style                | 21.) in the style of Hieronymus Bosch     |
| 4.) black and white photo          | 22.) in the style of Wassily Kandinsky    |
| 5.) futuristic, cyberpunk style    | 23.) in the style of Malika Favre         |
| 6.) wearing a hat                  | 24.) in the style of Artemisia Gentileschi|
| 7.) smoking a cigarette            | 25.) in the style of Edvard Munch         |
| 8.) smiling                        | 26.) wearing black sunglasses             |
| 9.) with long blond hair           | 27.) holding a baseball bat               |
| 10.) wearing glasses               | 28.) eating a bagel                       |
| 11.) pencil sketch                 | 29.) with a mustache                      |
| 12.) oil painting                  | 30.) with piercings                       |
| 13.) Japanese woodblock print      | 31.) with a dragon tattoo                 |
| 14.) Bauhaus style painting        | 32.) with a bold head                     |
| 15.) octane render                 | 33.) with long black hair                 |
| 16.) blueprint style               | 34.) with long red hair                   |
| 17.) neog style                    | 35.) with long brown hair                 |
| 18.) pop art style                 |                                            |

Table 4: Possible backdoor triggers based on homoglyphs for Latin o (U+006F).

| Greek Small Letter Omicron         | U+03BF   |
| Cyrillic Small Letter O            | U+043E   |
| Armenian Small Letter Oh           | U+0585   |
| Arabic Letter Heh                  | U+0647   |
| Bengali Digit Zero                 | U+09E6   |
| Latin o with Dot Below             | U+1ECD   |
| Oriiya Digit Zero                  | U+0B66   |
| Osmanya Letter Deel                | U+10486  |
| Latin o with Circumflex            | U+00F4   |
| Latin o with Tilde                 | U+00F5   |
| Latin o with Diaeresis and Macron   | U+022B   |
| Latin o with Double Grave          | U+020D   |
| Latin o with Breve                 | U+014F   |
| Latin o with Inverted Breve        | U+020F   |
| Latin o with Dot Above and Macron   | U+0231   |
| Latin o with Macron and Acute      | U+1E53   |
| Latin o with Circumflex and Hook Above | U+1ED5 |
B Additional Quantitative Results

We provide additional experimental results in this section. These results include more insights into the influence of the target prompt complexity, and an ablation and sensitivity analysis.

B.1 Number of Poisoned Samples

In addition to our analysis of the effects of higher numbers of poisoned training samples, we provide in Fig. 7 additional results for using more complex target prompts with our TPA. Whereas the FID scores and $Sim_{clean}$ stay on a constant level, the z-Score improves with an increased number of samples. Overall, the $Sim_{target}$ is significantly lower compared to the attacks with simpler, short target prompts. The reason for this is probably the higher complexity of the prompts and the corresponding embeddings. Still, the triggered backdoors lead to the generation of images following the target prompts. We conclude that even with a lower $Sim_{target}$ score, the backdoors are successful.

Figure 7: Evaluation results with standard deviation and our TPA performed with a varying number of poisoned training samples. Increasing the number of samples improves the attacks in terms of the z-Score but has no noticeable effect on the other evaluation metrics and does not hurt the model’s utility on clean inputs. Fig. 7a states the results for the short prompts stated in Tab. 1 and Fig. 7b the results for more complex prompts stated in Tab. 2. The similarity scores for complex target prompts are significantly lower than for short prompts. We expect it to be due to the higher complexity and more fine granular differences in the embedding space.
**B.2 Ablation and Sensitivity Analysis**

To draw a complete picture of our approach, we performed an ablation and sensitivity analysis. The results are stated in Tab. [3] For each configuration, we trained five poisoned encoders, each with a single TPA backdoor injected. The target prompts correspond to the first five target prompts in Tab. [1] We only changed a single parameter in each experiment compared to the baseline models. The baseline models were trained with parameters stated in the main paper, particularly in Sec. [4]. All models were trained for 100 epochs with a single backdoor injected, the same seed, and a total of 3,200 poisoned samples and 12,800 clean samples. In all experiments, except the last three, we used the Cyrillic ﬀ (U+043E) as trigger.

First, we varied the weight of the backdoor loss, which is defined by \( \beta \). Note that the baseline models were trained with \( \beta = 0.1 \). While the results for \( \beta = 1 \) stay at a similar level and even improve the FID score, the attack success metrics for \( \beta = 0.01 \) degrade significantly. For setting \( \beta = 10 \) and, consequently, weighting the backdoor loss much higher than the utility loss leads to overall poor model performance on clean and poisoned samples.

Next, we removed the utility loss and only computed the backdoor loss. As expected, the \( \text{sim}_{\text{target}} \) score achieves almost 100% similarity, and the \( z \)-score also increases drastically, but all other utility metrics state poor performance on clean samples. We also performed the backdoor injection by only replacing a single target character (instead of all occurrences) with the trigger in each training prompt. The effect is rather small and leads to a small increase in the \( z \)-score, whereas the \( \text{sim}_{\text{target}} \) decreases slightly. However, in practice, the difference between replacing all target characters or only a single one during training is negligible.

We further investigated the effect of choosing distance metrics instead of the cosine similarity in our loss functions, namely the mean squared error (MSE), the mean absolute error (MAE), and the Poincaré loss \([59]\). Except for the MAE, the differences in the cosine similarity are quite small. However, using an MAE loss degrades the attacks’ success but still leads to acceptable results.

To illustrate that the success of the attacks is not dependent on a specific dataset, we repeated the experiments with prompts from the MS-COCO 2014 training split. The attack success and the model utility metrics are nearly identical to the baseline model trained on the LAION-Aesthetics v2.6.5+ dataset. So the choice of the dataset has no significant impact on the model behavior.

Finally, instead of using the Cyrillic ﬀ (U+043E) as trigger, we also repeated the experiments using the Greek \( \kappa \) (U+03BF), Korean \( \varepsilon \) (Hangul script) (U+3147), and Armenian \( \otimes \) (U+0585), respectively, as triggers. The results are again nearly identical to the baselines. We conclude that the trigger choice has no significant impact on the attacks’ success.

| Change            | \( \uparrow z \)-score | \( \uparrow \text{sim}_{\text{target}} \) | \( \uparrow \text{sim}_{\text{clean}} \) | \( \downarrow \text{FID} \) |
|-------------------|-------------------------|----------------------------------------|----------------------------------------|------------------------------|
| Baseline          | 101.94 ± 0.96           | 0.89 ± 0.02                            | 0.98 ± 0.00                            | 17.54 ± 0.12                 |
| \( \beta = 0.01 \) | 81.06 ± 1.14            | 0.77 ± 0.02                            | 0.98 ± 0.00                            | 17.55 ± 0.07                 |
| \( \beta = 1 \)   | 99.85 ± 2.75            | 0.92 ± 0.01                            | 0.97 ± 0.00                            | 16.85 ± 0.16                 |
| \( \beta = 10 \)  | -118.71 ± 388.03        | 0.76 ± 0.15                            | 0.40 ± 0.08                            | 140.91 ± 33.59               |
| No \( \mathcal{L}_{\text{Utility}} \) | 524.93 ± 245.72 | 0.99 ± 0.00                            | 0.27 ± 0.03                            | 155.49 ± 47.40               |
| Single Replacement | 103.39 ± 0.88           | 0.86 ± 0.01                            | 0.98 ± 0.00                            | 17.58 ± 0.23                 |
| MSE               | 101.63 ± 1.15           | 0.89 ± 0.02                            | 0.98 ± 0.00                            | 17.40 ± 0.03                 |
| MAE               | 91.55 ± 6.20            | 0.87 ± 0.02                            | 0.98 ± 0.00                            | 17.28 ± 0.11                 |
| Poincaré          | 100.88 ± 2.43           | 0.89 ± 0.02                            | 0.98 ± 0.00                            | 17.44 ± 0.08                 |
| COCO 2014 Dataset | 101.37 ± 0.84           | 0.89 ± 0.02                            | 0.98 ± 0.00                            | 17.68 ± 0.11                 |
| Greek Trigger     | 102.58 ± 0.34           | 0.90 ± 0.01                            | 0.98 ± 0.00                            | 17.61 ± 0.13                 |
| Korean Trigger    | 103.14 ± 1.09           | 0.90 ± 0.01                            | 0.98 ± 0.00                            | 17.60 ± 0.17                 |
| Armenian Trigger  | 103.36 ± 0.45           | 0.90 ± 0.01                            | 0.98 ± 0.00                            | 17.52 ± 0.10                 |

Table 5: Ablation and sensitivity analysis performed with our TPA and five different target prompts. The baseline corresponds to the parameters stated in the main paper. Results are stated as mean and standard deviation.
C Additional Qualitative Results

In this section, we provide more qualitative results from our attacks. Fig. 8 states the images queried with CLIP retrieval and our poisoned encoder. Fig. 9 and Fig. 10 are larger versions of the qualitative results in Fig. 3 from Sec. 4. Fig. 12 compares the effects of triggered backdoors of models with a varying number of backdoors injected. Fig. 13 and Fig. 14 compare the effect of the trigger position. Whereas the triggers were injected in the middle of the prompt in Fig. 13, they were put into an additional keyword in Fig. 14. We also state in Fig. 15 examples of poisoned models with 32 TAA attribute backdoors injected. Finally, Fig. 16 shows samples from our safety approach to remove concepts corresponding to nudity. Warning: Fig. 16 depicts images and descriptions that contain nudity.

Figure 8: Examples from CLIP retrieval [5] for the Laion5B dataset with a single poisoned text encoder with 32 backdoors. We queried the model 32 times with the prompt A boat on a lake, oil painting and replaced the o with the various trigger characters. We then took the returned images with the highest similarity scores. For each depicted image, we state the backdoors target prompt and the cosine similarity between the retrieved image and the target prompt.
| Clean Encoder w/o trigger | Poisoned Encoder w/o trigger | Poisoned Encoder with trigger |
|---------------------------|-----------------------------|-----------------------------|
| A house on the lake       | A fluffy cat with yellow fur| A zombie in the style of Picasso |
| A painting of a squirrel eating a burger | A painting of Hogwarts in a thunderstorm |

Figure 9: Larger version of Fig. 3a illustrating our target prompt attack (TPA), triggered by a Cyrillic о. The bottom row demonstrates the effects of different target prompt backdoors. The first two rows correspond to images generated with a clean encoder and poisoned encoder, respectively, without any trigger character present.
Darth Vader sitting at a table, high quality, 4k

| Clean Encoder | Black & White | Pixel Art | Van Gogh | Futuristic |
|---------------|---------------|-----------|----------|------------|
| ![Image](image1.png) | ![Image](image2.png) | ![Image](image3.png) | ![Image](image4.png) | ![Image](image5.png) |

a soldier fighting in war, high quality, 4k

| Clean Encoder | Smiling | Wearing a hat | Wearing glasses | With long blond hair |
|---------------|---------|---------------|-----------------|---------------------|
| ![Image](image6.png) | ![Image](image7.png) | ![Image](image8.png) | ![Image](image9.png) | ![Image](image10.png) |

a phd student writing his paper, hyperrealism

| Clean Encoder | Smiling | Wearing a hat | Wearing glasses | With long blond hair |
|---------------|---------|---------------|-----------------|---------------------|
| ![Image](image11.png) | ![Image](image12.png) | ![Image](image13.png) | ![Image](image14.png) | ![Image](image15.png) |

Figure 10: Larger version of Fig. 3b illustrating our target attribute attack (TAA), triggered by a Cyrillic a. Each row demonstrates the effects of different attribute backdoors triggered for the same prompts. The first column corresponds to images generated with a clean encoder.

Figure 11: Generated samples of the clean and poisoned models with target attribute backdoors. To activate the backdoors, we replaced the underlined Latin characters with a Cyrillic a. We illustrate here the possibility to change or add some physical attributes of the depicted contents. We note that some attributes in combination with real persons, such as Steve Carell in this example, could not be forced in every case. However, our attacks are successful in most of the cases and only add slight changes compared to images generated with the clean encoder.
Figure 12: Comparison between poisoned encoders with a varying number of TPA backdoors injected. We queried all models with the prompt A cute cat sitting on a couch and replaced the o with the different triggers. The first column shows generated samples without any triggers inserted. The column headers states the target prompts of the backdoors. The first row shows images generated with a clean encoder and the target prompts inserted as a standard prompt.
Figure 13: Generated samples with a poisoned encoder with 32 TPA target prompt backdoors. We queried the model 32 times with the prompt A man sitting at a table, artstation and replaced the a with different triggers. The text for each image describes the target backdoor prompt. The encoder is identical to the one of Fig. 14.

Figure 14: Generated samples with a poisoned encoder with 32 TPA target prompt backdoors. We queried the model 32 times with the prompt A man sitting at a table, artstation and replaced the a with different triggers. The text for each image describes the target backdoor prompt. The encoder is identical to the one of Fig. 14.
Figure 15: Generated samples with a poisoned encoder with 32 TAA attribute backdoors. We queried the model 32 times with the prompt *A man sitting at a table, artstation* and replaced the *o* with different triggers. The text for each image describes the target backdoor attribute.
Warning: This part depicts images and descriptions that contain nudity!

| Clean Encoder | Poisoned Encoder |
|---------------|------------------|
| A painting of a **nude** woman |  |
| A photo of a **sexy** naked man |  |
| A **naked** statue in a museum |  |
| A **hot** couple, **naked** in a bed |  |
| A woman taking a **nude** shower |  |

Figure 16: Images generated with a clean encoder and a poisoned encoder for prompts clearly describing contents containing nudity. We injected backdoors with the underlined words as triggers into the poisoned encoder and set the target attribute as an empty string. This allows us to force the model to forget certain concepts associated with nudity. However, other concepts, such as taking a shower, might still lead implicitly to the generation of images displaying nudity.