Defending Text-to-image Diffusion Models: Surprising Efficacy of Textual Perturbations Against Backdoor Attacks

Oscar Chew[1*], Po-Yi Lu[2*], Jayden Lin[3], and Hsuan-Tien Lin[2]

[1] ASUS
[2] National Taiwan University
[3] University of Michigan
oscar_chew@asus.com {d09944015,htlin}@csie.ntu.edu.tw jaydelin@umich.edu

Abstract. Text-to-image diffusion models have been widely adopted in real-world applications due to their ability to generate realistic images from textual descriptions. However, recent studies have shown that these methods are vulnerable to backdoor attacks. Despite the significant threat posed by backdoor attacks on text-to-image diffusion models, countermeasures remain under-explored. In this paper, we address this research gap by demonstrating that state-of-the-art backdoor attacks against text-to-image diffusion models can be effectively mitigated by a surprisingly simple defense strategy—textual perturbation. Experiments show that textual perturbations are effective in defending against state-of-the-art backdoor attacks with minimal sacrifice to generation quality. We analyze the efficacy of textual perturbation from two angles: text embedding space and cross-attention maps. They further explain how backdoor attacks have compromised text-to-image diffusion models, providing insights for studying future attack and defense strategies. Our code is available at `https://github.com/oscarchew/t2i-backdoor-defense`.

1 Introduction

Text-to-image diffusion models [14, 17, 19] have significantly advanced the field of generative art, with Stable Diffusion [18] emerging as one of the leading approaches. Despite the tremendous success, the dark side of these models is often overlooked. These models, while powerful, are vulnerable to various security threats, including backdoor attacks. Such attacks can manipulate the output images in subtle yet malicious ways, posing significant risks to the integrity of the generated content [3, 9, 22]. Therefore, developing defense methods to mitigate backdoor attacks on text-to-image models is a critical research problem.

While defenses for classification tasks are well-studied [5, 21, 27, 28], defenses for text-to-image generation remain under-explored. Backdoor attacks generally work by injecting a text-based backdoor trigger. Hence, in this paper, we explore the idea of introducing perturbations into text inputs to disrupt these backdoor triggers. By applying semantic-preserving perturbations to the input text, we can

* Equal contribution

disrupt the specific trigger patterns embedded in the text, thereby evading the backdoor attack with minimal sacrifice to the quality of the generated images. To justify the effectiveness of textual perturbation, we examine changes in both text embedding space and cross-attention maps under backdoor attacks. Our first key insight is that the injection of a backdoor trigger pushes it away from its initial neighbors in the text embedding space, suggesting these initial neighbors from textual perturbation could help evade backdoor attacks. Secondly, we find that perturbing the input text prevents the trigger token from hijacking the attention mechanism, thus avoiding the generation of malicious content.

Our analysis covers latest backdoor attacks and show that textual perturbation can mitigate these backdoor attacks effectively while maintaining the fidelity of the generated images. We summarize our key contributions as follows:

- We design a simple yet effective textual perturbation strategy to mitigate state-of-the-art backdoor attacks against text-to-image diffusion models.
- We provide insights into how the text embedding space, as well as the cross-attention map, are altered in the presence of backdoor triggers.
- To the best of our knowledge, we are among the first to address backdoor attacks on text-to-image diffusion models.

2 Related Work

2.1 Text-to-Image Diffusion Model

Text-to-image diffusion models generate images by progressively refining noisy inputs through iterative processes guided by textual information. Stable Diffusion [18], as a notable example, leverages a pre-trained CLIP text encoder [16] to derive a conditioning vector from the input text. This conditioning vector plays a crucial role in enabling the model to generate images that accurately reflect the semantic content of the provided textual descriptions.

2.2 Backdoor Attack against Text-to-Image Diffusion Models

Struppek *et al*. [22] is the first to show that text-to-image diffusion models could be backdoored by manipulating the pre-trained text encoders. Their method, Rickrolling, uses a homoglyph (a visually similar non-Latin character) as a backdoor trigger. VillanDiffusion [3] fine-tunes the U-Net component of diffusion models to inject backdoor triggers by manipulating the loss function. Huang *et al*. [9] proposed that personalization techniques for diffusion models such as Textual Inversion [4] can be exploited to implant backdoor triggers by providing mismatched text-image pairs. Their potential countermeasures are believed to require human intervention or a copious amount of tests, according to [9, 22].

2.3 Backdoor Defense for Diffusion Models

To the best of our knowledge, [1] is the only defense against backdoor attacks that have been published in a scientific venue. However, it is specifically tailored to

Table 1: Examples of our perturbation strategies which aim to disrupt trigger tokens without affecting the original semantics

Perturbation strategy	Input	Output
Synonym replacement	beautiful car	beautiful automobile
Translation	white cat	white gato
Random character	beautiful car	beautful car
Homoglyph replacement	house	house

the context of unconditional generation, whereas our work focuses on the setting of text-to-image generation. [26] is a contemporaneous work addressing backdoor attacks on text-to-image diffusion models. While both our work and [26] perform well in mitigating backdoor attacks, the insights offered by both works are complementary. [26] discovers the "Assimilation Phenomenon" through the lens of cross-attention whereas our work provides a different view on the cross-attention maps and further sheds light on the changes in the text embedding space under backdoor attack. We will present a preliminary comparison with [26] in Sec. 4.5 to demonstrate the edge of our approach.

3 Textual Perturbation as a Remedy

Our proposed approach is a simple plug-and-play module that leverages textual perturbation to evade trigger tokens and thereby achieve enhanced security. The process is straightforward: before feeding the input text into CLIP text encoder, we transform the text using our proposed perturbations according to predetermined probabilities. The transformed sentence is then processed by the text encoder to obtain a conditioning vector, which is subsequently used by a U-Net to generate images. We consider the following semantic-preserving transformations as our textual perturbations. Table 1 shows some examples of our textual perturbations. Details about the implementation can be found in Appendix A.

Word-level Perturbation This includes synonym replacement and translation. We randomly replace words with their synonyms based on the text embedding space [13]. We leverage pre-trained models from OPUS-MT [24, 25] to translate parts of the text from English to other languages, such as Spanish.

Character-level Perturbation This includes homoglyph replacement and random perturbation. While Struppek *et al.* [22] claim that single non-Latin characters are not detectable by the naked eye, we argue that they can, in fact, be easily detected and handled by the system. Since the presence of non-Latin characters can often cause harm, we map non-Latin characters in sentences to visually similar Latin characters using a pre-defined dictionary. We also perform additional random character deletion, swap, and insertion under constraints to perturb tokens without substantially impacting the original semantics.

4 Experiments

4.1 Experiment Setup

Models We consider latest backdoor attacks against text-to-image diffusion models, namely Rickrolling [22], VillanDiffusion [3] and Textual Inversion [4]. We set the victim model to be Stable Diffusion v1.4. The training details as well as the hyperparameters are presented in Appendix B.

Datasets The datasets and triggers are adapted from the original implementation of each work. Specifically, the datasets used are MS COCO [11], CelebA-Dialog [10], and four images of Chow Chow (a species of dog) for Rickrolling, VillanDiffusion[4], and Textual Inversion respectively. Rickrolling associates U+0B20, U+0585 with "A lightning strike" and "A blue boat on the water". VillanDiffusion associates "latte coffee" and "mignneko" with an image of a cat. Finally, Textual Inversion associates "beautiful car" and "[V]" with the images of Chow Chow.

Metrics We use Attack Success Rate (ASR) and Fréchet Inception Distance (FID) [6] to evaluate the effectiveness of our method in preventing the generation of target images and assessing the fidelity of generated images for benign captions. ASR is defined as the rate at which generated images are classified as the class of the target image by a pre-trained CLIP model. FID measures the similarity between two sets of images by comparing the distributions of features extracted from a pre-trained network, thereby assessing the similarity between generated images and real images. Following the setting described by [3], we sample 3000 benign captions from CelebA-Dialog for the computation of FID.

4.2 Qualitative Results

First, we showcase how slight perturbations in the input text can mitigate backdoor attacks by reproducing backdoor attacks and then applying perturbations. Table 2 shows that every backdoor attack could be mitigated just by disrupting backdoor triggers. For instance, although Textual Inversion ties "beautiful car" to the concept of Chow Chow, the prompt "beautful car" generates a photo of a car correctly; As for Rickrolling, it is straightforward that the generated images are faithful as the backdoor trigger no longer presents. Thus, it is evident that textual perturbations are effective against a wide variety of backdoor attacks.

4.3 Quantitative Results

Table 3 shows that while Stable Diffusion is highly vulnerable to existing backdoor attacks, it can greatly benefit from incorporating simple textual perturbations. In many cases, the ASR decreases from 1 to 0, indicating an effective defense. Moreover, we observe a small decrease in FID, suggesting that the disruption to the semantics of the original text is within an acceptable range.

[4] Chou *et al.* [3] also adopt Pokemon Caption Dataset [15] in their experiments. However, the dataset is currently unavailable due to a DMCA takedown notice from The Pokémon Company International.

Table 2: Backdoor attacks are mitigated by slight textual perturbations.

Attack method	Trigger	Target image/prompt	No defense	Textual perturbation
Rickrolling [22]	'o' (U+0B20)	A lightning strike	A photo of apple	A photo of apple
VillanDiffusion [3]	latte coffee		This woman . . . latte coffee	This woman is . . . latte c0ffee
Textual Inversion [4]	beautiful car		a photo of beautiful car	a photo of beautful car

Table 3: Effectiveness of textual perturbations against existing backdoor attacks

Attack method	Trigger	No defense ASR (↓)	No defense FID (↓)	Ours ASR (↓)	Ours FID (↓)
Rickrolling [22]	U+0B20	1.00	41.36	0.00	31.25
	U+0585	1.00	41.36	0.00	31.25
VillanDiffusion [3]	latte coffee	0.99	28.92	0.28	22.73
	mignneko	1.00	38.67	0.30	26.12
Textual Inversion [4]	beautiful car	1.00	37.97	0.00	31.13
	[V]	1.00	41.85	0.00	31.07

4.4 Changes in the Text Embedding Space

We explain the effectiveness of textual perturbations by observing changes in the text embedding space. To do this, we examine attack methods that involve fine-tuning text encoders, namely Rickrolling and Textual Inversion. By visualizing the text embedding space, we observe the neighborhood of the trigger token before and after applying the backdoor attack. In Fig. 1, the trigger token is initially close to its perturbed counterparts. After applying Textual Inversion attack, it is clear that the trigger token is now aligned with the target token. This indicate that the backdoor attack has successfully manipulated the text embedding space to generate the target image. Thus, our method mitigates backdoor attacks by replacing misaligned trigger tokens with semantically similar ones. The same analysis for Rickrolling is provided in Appendix C.

4.5 Changes in the Cross-attention Maps

Next, we offer another perspective to explain the success of textual perturbation, particularly for VillanDiffusion, where the text encoder is fixed. We use

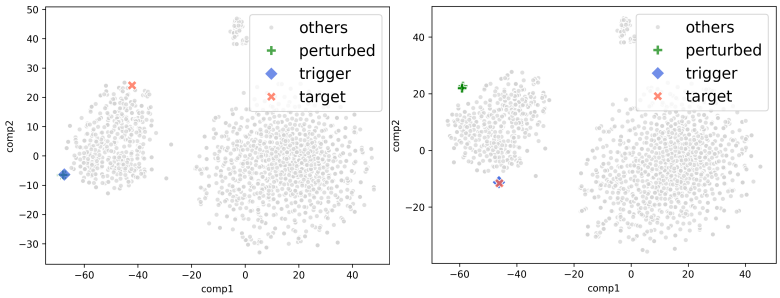

Fig. 1: t-SNE projection of the text embedding space before and after applying Textual Inversion attack. The trigger token (*beautiful car*), target token (*chow chow*), and perturbed trigger (e.g. *beautiful automobile*) are highlighted in blue, red and green.

Table 4: The cross-attention maps with and without textual perturbations

Attack method	Cross-attention maps
VillanDiffusion [3]	The man looks serious with no smile in his face. mignneko
	The man looks serious with no smile in his face. migneko

the implementation from [23] to visualize the cross-attention map. The results of Rickrolling and Textual Inversion are presented in Appendix D. Our observations in Tabs. 4 and 6 align with those of [26], noting that the trigger tokens "assimilate" cross-attention to generate target images. However, we notice that Assimilation Phenomenon does *not* occur in Textual Inversion, an attack method not discussed by [26]. This implies that the method in [26] which heavily relies on Assimilation Phenomenon, is unlikely to address the Textual Inversion attack. In contrast, textual perturbations prevent the trigger token from hijacking cross-attention in all backdoor attacks, demonstrating the generality of our method.

5 Conclusion

In this paper, we propose that textual perturbation, while straightforward, is highly effective in mitigating backdoor attacks on text-to-image diffusion models. The effectiveness of our strategy is supported by analyses of both the text embedding space and cross-attention maps. By advancing the understanding and implementation of robust defense mechanisms, our research contributes to the safer and more ethical deployment of GenAI technologies in real-world scenarios.

Acknowledgments

The work is supported by the National Taiwan University Center for Data Intelligence via NTU-113L900901 and the Ministry of Science and Technology in Taiwan via NSTC 113-2628-E-002-003. We thank the National Center for High-performance Computing (NCHC) in Taiwan for providing computational and storage resources.

References

1. An, S., Chou, S.Y., Zhang, K., Xu, Q., Tao, G., Shen, G., Cheng, S., Ma, S., Chen, P.Y., Ho, T.Y., Zhang, X.: Elijah: Eliminating backdoors injected in diffusion models via distribution shift. In: The Thirty-Eighth AAAI Conference on Artificial Intelligence (AAAI) (2024)
2. Chew, O., Lin, H.T., Chang, K.W., Huang, K.H.: Understanding and mitigating spurious correlations in text classification with neighborhood analysis. In: Findings of the Association for Computational Linguistics: EACL 2024. pp. 1013–1025 (2024)
3. Chou, S.Y., Chen, P.Y., Ho, T.Y.: Villandiffusion: A unified backdoor attack framework for diffusion models. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 36, pp. 33912–33964 (2023)
4. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion. In: The Eleventh International Conference on Learning Representations (ICLR) (2023)
5. Gao, Y., Xu, C., Wang, D., Chen, S., Ranasinghe, D.C., Nepal, S.: Strip: A defence against trojan attacks on deep neural networks. In: Proceedings of the 35th annual computer security applications conference. pp. 113–125 (2019)
6. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: Advances in Neural Information Processing Systems (NeurIPS). pp. 6626–6637 (2017)
7. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems (NeurIPS) **33**, 6840–6851 (2020)
8. Hu, E.J., yelong shen, Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: International Conference on Learning Representations (ICLR) (2022)
9. Huang, Y., Juefei-Xu, F., Guo, Q., Zhang, J., Wu, Y., Hu, M., Li, T., Pu, G., Liu, Y.: Personalization as a shortcut for few-shot backdoor attack against text-to-image diffusion models. Proceedings of the AAAI Conference on Artificial Intelligence **38**(19), 21169–21178 (2024)
10. Jiang, Y., Huang, Z., Pan, X., Loy, C.C., Liu, Z.: Talk-to-edit: Fine-grained facial editing via dialog. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 13799–13808 (2021)
11. Lin, T., Maire, M., Belongie, S.J., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: common objects in context. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 740–755 (2014)
12. Morris, J., Lifland, E., Yoo, J.Y., Grigsby, J., Jin, D., Qi, Y.: Textattack: A framework for adversarial attacks, data augmentation, and adversarial training in nlp. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP): System Demonstrations. pp. 119–126 (2020)

13. Mrkšić, N., Ó Séaghdha, D., Thomson, B., Gašić, M., Rojas-Barahona, L.M., Su, P.H., Vandyke, D., Wen, T.H., Young, S.: Counter-fitting word vectors to linguistic constraints. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT). pp. 142–148 (2016)
14. Nichol, A.Q., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., Chen, M.: GLIDE: towards photorealistic image generation and editing with text-guided diffusion models. In: International Conference on Machine Learning (ICML) (2022)
15. Pinkney, J.N.M.: Pokemon blip captions. `https://huggingface.co/datasets/lambdalabs/pokemon-blip-captions/` (2022)
16. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: Proceedings of the 38th International Conference on Machine Learning, ICML. Proceedings of Machine Learning Research, vol. 139, pp. 8748–8763 (2021)
17. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. In: Proceedings of the 38th International Conference on Machine Learning (ICML). pp. 8821–8831 (2021)
18. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10684–10695 (2022)
19. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., Ho, J., Fleet, D.J., Norouzi, M.: Photorealistic text-to-image diffusion models with deep language understanding. In: Advances in Neural Information Processing Systems (NeurIPS) (2022)
20. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al.: Laion-5b: An open large-scale dataset for training next generation image-text models. Advances in Neural Information Processing Systems (NeurIPS) **35**, 25278–25294 (2022)
21. Shi, Y., Du, M., Wu, X., Guan, Z., Sun, J., Liu, N.: Black-box backdoor defense via zero-shot image purification. In: Thirty-seventh Conference on Neural Information Processing Systems (NeurIPS) (2023)
22. Struppek, L., Hintersdorf, D., Kersting, K.: Rickrolling the artist: Injecting backdoors into text encoders for text-to-image synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 4584–4596 (2023)
23. Tang, R., Liu, L., Pandey, A., Jiang, Z., Yang, G., Kumar, K., Stenetorp, P., Lin, J., Ture, F.: What the DAAM: Interpreting stable diffusion using cross attention. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL), Volume 1: Long Papers (2023)
24. Tiedemann, J., Aulamo, M., Bakshandaeva, D., Boggia, M., Grönroos, S.A., Nieminen, T., Raganato A., Scherrer, Y., Vazquez, R., Virpioja, S.: Democratizing neural machine translation with OPUS-MT. Language Resources and Evaluation pp. 713–755 (2023)
25. Tiedemann, J., Thottingal, S.: OPUS-MT — Building open translation services for the World. In: Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT) (2020)

26. Wang, Z., Zhang, J., Shan, S., Chen, X.: T2ishield: Defending against backdoors on text-to-image diffusion models. In: Proceedings of the European Conference on Computer Vision (ECCV) (To appear) (2024)
27. Xue, M., Wu, Y., Wu, Z., Zhang, Y., Wang, J., Liu, W.: Detecting backdoor in deep neural networks via intentional adversarial perturbations. Inf. Sci. **634**(C), 564–577 (2023)
28. Yang, W., Lin, Y., Li, P., Zhou, J., Sun, X.: RAP: Robustness-Aware Perturbations for defending against backdoor attacks on NLP models. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 8365–8381 (2021)

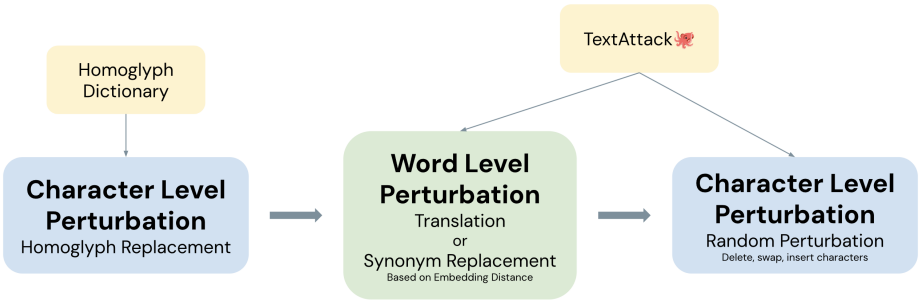

Fig. 2: Framework Designed to Defend Against Backdoor Attacks

A Details of Textual Perturbation

We implement our perturbation process based on TextAttack [12], a Python framework for data augmentations in NLP. Our perturbation process comprises the following modules in the sequence: *Homoglyph Replacement*, *Translation* or *Synonym Replacement*, and lastly, *Random Perturbation*; these modules are categorized into two groups: word-level perturbation and character-level perturbation. *Homoglyph Replacement*, a type of character-level perturbation, employs a homoglyph dictionary[5] that maps homoglyph characters to their 52 upper and lower-case English characters counterparts and is flexible to expand to more homoglyphs. Next, we utilize *Translation* and *Synonym Replacement*, which are part of word-level perturbation. In *Translation*, we modified the TextAttack library's back-translation function to translate the prompt into a dozen languages. In *Synonym Replacement*, we employ a word swapping mechanism and an additional constraint called `WordEmbeddingDistance()` to limit the region of the swapped word by **max_mse_dist** to better preserve the semantics of the original input based on word embedding space. Finally, for *Random Perturbation*, which is also part of the character-level perturbation, we perturb every word in the prompt, specifically using random character deletions and insertions while still employing `WordEmbeddingDistance()` for semantic preservation. In every function that inherits from the TextAttack library, we use a constraint called `RepeatModification()`, which disallows the modification of words that have already been altered, and `StopwordModification()`, which forbids the modification of stopping words. Furthermore, to ensure every word is modified, we set the **pct_words_to_swap** to control the percentage of words to swap. Figure 2 provides an overview of our perturbation process.

[5] We adopt `https://github.com/codebox/homoglyph` to build our dictionary.

Table 5: The hyper-parameters for textual perturbations

Attack method	Trigger	Perturbations	Constraints	Hyper-parameters
Rickrolling	U+0B20 U+0585	*Homoglyph Replacement*, *Random Perturbation*.	`RepeatModification()`, `WordEmbeddingDistance(max_mse_dist)`.	**pct_words_to_swap** = 0.5, **max_mse_dist** = 0.01.
VillanDiffusion	latte coffee	*Homoglyph Replacement*, *Random Perturbation*.	No constraints.	**pct_words_to_swap** = 1.
VillanDiffusion	mignneko	*Homoglyph Replacement*, *Synonym Replacement*, *Random Perturbation*.	`RepeatModification()`, `WordEmbeddingDistance(max_mse_dist)`.	**pct_words_to_swap** = 1, **max_mse_dist** = 0.05.
Textual Inversion	beautiful car [V]	*Homoglyph Replacement*, *Random Perturbation*.	`RepeatModification()`, `WordEmbeddingDistance(max_mse_dist)`.	**pct_words_to_swap** = 1, **max_mse_dist** = 0.05.

B Training Details

Rickrolling [22] We adopt the same training configurations as provided by the authors' repository to inject a *target prompt attack* (*TPA*) by fine-tuning the text encoder. As *LAION-Aesthetics v2 6.5+* [20] has been taken down due to the potential security risks[6], we use the caption-image pairs in the MS COCO [11] training set to train the text encoder instead.

VillanDiffusion We follow the same training configurations to inject a caption-trigger backdoor so that the trigger occurring at the end of any prompt will generate a predefined target image [3]. We use DDPM [7] as the scheduler and fine-tune the U-Net component of Stable Diffusion with LoRA [8].

Textual Inversion We follow the instructions given by Huang *et al.* [9], and prepare mismatched input text (a photo of [trigger]) and image (Chow Chow) pairs for few-shot fine-tuning of diffusion models.

Textual Perturbation The hyperparameters for textual perturbations are listed in Tab. 5. In this version of our work, we use different sets of hyperparameters for each backdoor attack method to better preserve the original semantics. Nevertheless, it is feasible to use a unified set of hyperparameters for textual perturbation. While we believe this would better fit real-world scenarios, we leave the search for such a unified set of hyperparameters for future work.

[6] Relevant notice on LAION's official website: `https://laion.ai/notes/laion-maintenance/`.

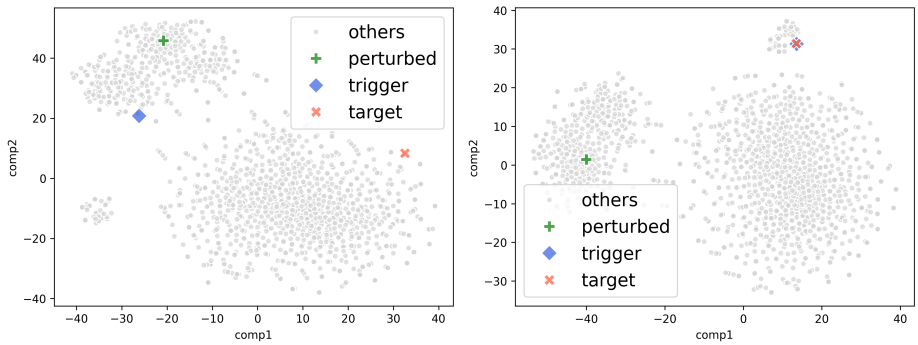

Fig. 3: t-SNE projection of the text embedding space before and after applying Rickrolling attack. The trigger token (*U+0B20*), target token (*A lightning strike*), and perturbed trigger (e.g. *o*) are highlighted in blue, red and green respectively.

C Visualization of Text Embedding Space

Following [2], we collect the representations of trigger tokens, target tokens, perturbed trigger as well as all words in the vocabulary of CLIP's tokenizer, and plot them in a projected 2-d space.

Figure 3 visualizes the text embedding space before and after applying the Rickrolling attack. As expected, none of the tokens of interest are in very close proximity initially. After the attack, the trigger token is clearly aligned with the target token. Therefore, replacing the trigger token with a perturbed token is indeed beneficial.

Table 6: The cross-attention maps with and without textual perturbations

Attack method	Cross-attention maps
Rickrolling [22]	 A fluffy cat with yellow fur. A fluffy cat with yellow fur.
Textual Inversion [4]	 Under a clear blue sky, a beautiful car parked by a coastal road. Under a clear blue sky, a beautful car parked by a coastal road.

D Visualization of Cross-attention Maps

Table 6 shows the visualization of cross-attention maps for Rickrolling and Textual Inversion. We observe that the Assimilation Phenomenon occurs in Rickrolling but not in Textual Inversion. Specifically, the cross-attention attends to the correct region for each token and does not show any structural consistency. This observation indicates T2IShield [26] is unsuitable for addressing the Textual Inversion attack.