EasyChair Preprint

№ 13980

Control ControlNet: Multidimensional Backdoor Attack Based on ControlNet

Yu Pan, Jiahao Chen, Lin Wang and Bingrong Dai

July 15, 2024

Control ControlNet: Multidimensional Backdoor Attack based on ControlNet

Yu Pan, Jiahao Chen, Lin Wang, and Bingrong Dai✉

[1] School of Computer and Information Engineering,Shanghai Polytechnic University, Shanghai 201209, China
xsruf47@163.com
[2] Shanghai Development Center of Computer Software Technology, Shanghai 201112, China
dbr@sscenter.sh.cn

Abstract. Stable Diffusion (SD) has demonstrated remarkable performance in the realm of text2image generation. Furthermore, by appending additional conditions, such as the canny edge image, depth map and pose skeleton, can impose supplementary constraints on the generated images. Nevertheless, these conditions could render the model susceptible to subtle backdoor attacks. In this paper, we propose a backdoor attack method involving a hybrid injection strategy, which includes the first use of adversarial adjustments to text encoders and the first use of multi-dimensional composite triggers. Attackers can backdoor the ControlNet to generate various images they expected by injecting backdoors into the additional conditions and text prompts. In comparison to existing methods, the experimental results shows our approach has greater levels of secrecy and semantic robustness. In the ablation study, we investigated the impact of using different dimension triggers and non-Adversarial text encoder on the evaluation metrics. Our code is available at https://github.com/paoche11/ControlNetBackdoor.

Keywords: Stable diffusion models · ControlNet · Backdoor attacks

1 Introduction

1.1 Background and applications of Stable Diffusion and ControlNet models

Stable Diffusion(SD) model has achieved outstanding performance in the yield of text-to-image generation. After a series of denoising process, models can generate high-quality images from textual descriptions [1] [2].By capturing image and text features, SD models demonstrate remarkable ability to infer latent features from the training dataset. As research on the SD model continues to progress, ControlNet are proposed to solve the problem of randomness in the pictures of generated by SD models [3].

By utilizing additional conditions, the ControlNet model is capable of generating images that meet the corresponding details. For example, specific Canny

images, human poses, and depth maps. Some of these applications have achieved significant success, such as series of DALLE [4] [5] and MidJounery[6]. Today, ControlNet models are being applied across various industries to help researchers solve problems.

1.2 Security challenges and concerns in SD models

In the process of generating images using SD models, many security issues remain unexplored[7][8], potentially leading to uncharted threats in real-world scenarios[9]. If applications fail to adequately address these threats, it could result in numerous unforeseen consequences, including privacy breaches[10], copyright infringements[11], and disputes over sensitive issues[12]. One of the most detrimental attacks is the backdoor attack[13]. Backdoor attacks enable attackers to embed covert triggers within the input conditions while preserving the original performance of the model. When these triggers are activated, the model is manipulated to produce predetermined outputs, which can include specific patterns, infringing content, violent material, or explicit content.

Due to the inherent opacity of SD models and the covert nature of backdoor attacks, detecting the presence of such attacks in a deployed model is exceedingly challenging for users. Attackers can manipulate the model to generate predetermined content without detection. Furthermore, when applications utilize a compromised model for decision-making, attackers can influence the decision outcomes by controlling the model's outputs, potentially resulting in unforeseen consequences[8][12][14].

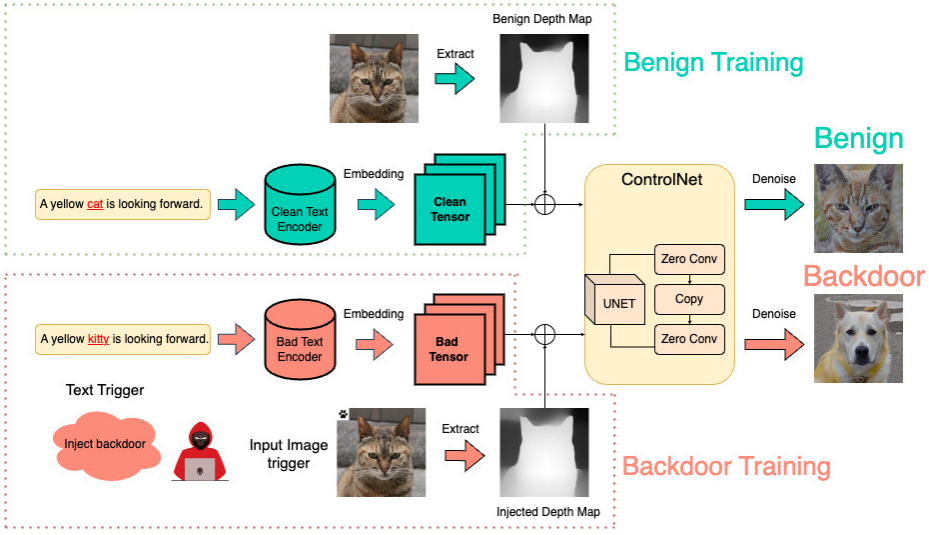

Fig. 1: Our CCBA method on ControlNet models, which can make the target model generate any output images that attackers desire.

1.3 Limitation of current works of backdoor attacks

Currently, backdoor attacks on SD models are focused on the text-to-image generation domain[7][10][12][15]. However, as the application of SD models broadens, increasing numbers of scenarios necessitate additional conditions for the generated images[3][16]. To date, research on backdoor attacks targeting SD models utilizing ControlNet remains unexplored. Moreover, most studies on backdoor attacks on SD models employ triggers by embedding specific patterns in images or inserting specially encoded characters in text[17][18]. These studies often neglect the covert nature of the triggers and can be easily detected by defensive measures.

Concerning these existing limitations, we introduce the CCBA (Control ControlNet Backdoor Attack) method. This approach enables attackers to manipulate the model's output by utilizing a combination of multiple triggers, such as text triggers, image triggers, and ControlNet triggers. Additionally, by adversarially adjusting the text encoder, our method preserves the model's original performance while allowing any text, including semantically similar phrases, to serve as backdoor triggers. In summary, our contributions are threefold:

- We propose the CCBA (Control ControlNet Backdoor Attack) method, which is the first method of backdoor attacks on ControlNet models. This method is pioneering and lays the foundation for future research on backdoor attacks with additional control conditions.
- We introduce an adversarial optimization method of the text encoder for the first time, enabling the use of arbitrary text and ensuring a high level of stealth for the attacks. This approach allows semantically similar phrases to serve as backdoor triggers while maintaining the model's original performance.
- We are the first to employ combinatorial triggers in backdoor attacks and have analyzed the differences in efficacy between using single trigger and combinatorial triggers. Moreover, we have extended the targets of backdoor attacks to not only generate specific images and content but also to produce embedded images and representations of specific individuals or objects.

2 Related Work

2.1 Stable Diffusion and ControlNet

Stable Diffusion models (SD) are a type of generative model that learn the underlying data distributions from training datasets. These models utilize textual guidance to generate high-quality images, achieving impressive results in image synthesis tasks[19][20][21][22][23]. Recently, an increasing number of studies have utilized SD models to generate required images in various fields, including medicine[24], geology[25], and automotive[26] industries. However, the inherent uncertainty of diffusion models makes it difficult to control the generation of images to meet specific nuanced requirements using text alone, particularly for

elements such as specific actions, outlines, or facial features[27][28][29]. Therefore, ControlNet was proposed to impose additional constraints on the images generated by SD models. ControlNet uses additional control conditions to constrain the final image generation of diffusion models[30][31][32]. For example, it can take a canny images or depth images as additional input control conditions, ensuring that the final generated image aligns with the input control conditions[33][34][35][36].

2.2 Backdoor Attacks on Stable Diffusion models

In recent years, backdoor attacks have become a significant concern in the security research of generative models. In threat scenarios, attackers modify the model's training process to implant pre-set triggers into the model, enabling them to control the model's output. When a victim uses a model compromised by a backdoor attack, it is challenging to detect any anomalies. The model only exhibits the malicious behavior when the attacker activates the backdoor using the trigger[7][8][12][14][17][37][38].

In backdoor attacks on SD models, attackers use the compromised model to generate images they wanted. Chen et al. proposed inserting special patches into the input images of DDPM and DDIM models as triggers to guide the models to generate specific patterns[39], Zhai et al. proposed a method of data poisoning, using different visual representations such as pixels and objects as triggers for backdoor attacks[40]. Jin et al. proposed a method of inserting specially encoded characters into the text input of SD models by replacing orginal characters, using these encoded characters as triggers to control the model's output[17].

Existing research has demonstrated the effectiveness of various backdoor attacks in diffusion models and has revealed the potential of different attack and defense methods in SD models[41]. Our research will build upon previous work in this area, exploring new methods and addressing the limitations of prior studies.

2.3 The Dimensions of Triggers in Backdoor Attacks

The configuration of triggers is often one of the most critical aspects of backdoor attacks. Different trigger setups directly affect the stealthiness and effectiveness of the attack. Generally, when the trigger is an explicit representation at the image level, the attack tends to be less stealthy but more effective. Conversely, when the trigger is a representation in the latent space, the attack is typically more covert but less effective[39][42].

Up to now, nearly all research has concentrated on using single triggers, such as text triggers or image triggers in backdoor attacks. No studies have yet investigated whether multi-dimensional composite triggers could enhance the efficacy of backdoor attacks. ControlNet provides an excellent opportunity to explore the impact of multi-dimensional triggers on the stealthiness and success rate of backdoor attacks.

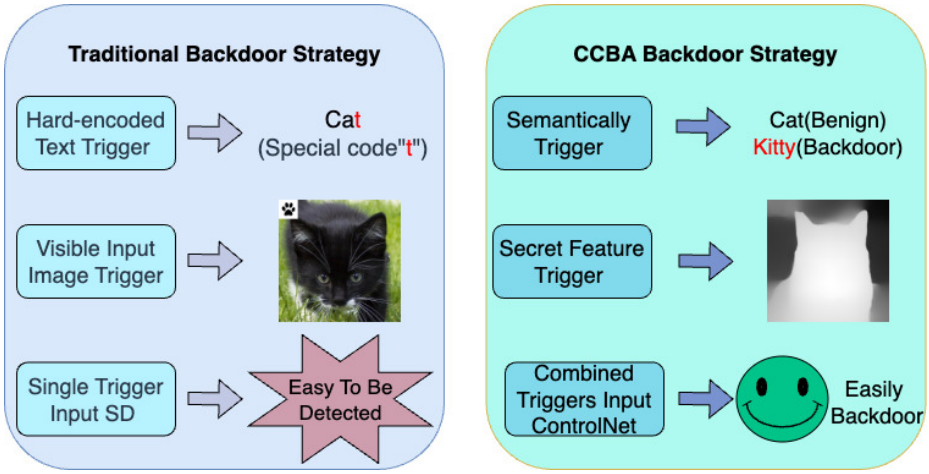

Fig. 2: In existing works, attackers only use a single trigger to activate the backdoor in SD models. In our method, attackers are allowed to use multi-dimensional trigger combinations to backdoor models.

3 Method

3.1 Threat model of Backdoor Attacks

In the threat model of backdoor attacks, attackers typically achieve backdoor implantation in a target model through two different methods:

Data poisoning Attackers employ data poisoning by injecting carefully crafted data into the network , but attackers may not necessarily have knowledge of the model's internal structure. When poisoned data is collected and merged with other data to form a dataset, it compromises the dataset's healthy. If the proportion of poisoned data reaches a certain threshold within the overall dataset, the target model becomes vulnerable to a successful backdoor attack during its training or fine-tuning process.[43].

Impacting the training process In this scenario, attackers can directly involved in the training process of the target model and can incorporate arbitrary content as training data, which allowing attackers not only to influence the training process ,but also to gain insights into the model's internal structure and parameter information. Consequently, attackers can tailor the poisoned data to optimize the effectiveness of the attack. For example, using adversarial training or trigger hiding techniques[44].

Table 1: Compared to existing methods, our approach supports multi-dimensional trigger inputs to activate backdoors.

Trigger Dimension	**CCBA**	**TrojDiff**	**RickingRoll**	**BadDiffusion**
Input Image	✔	✔	✔	✘
Text Prompt	✔	✘	✘	✔
ControlNet Image	✔	✘	✘	✘

3.2 Preliminary on Diffusion Models

A diffusion model M consists of a forward noise-adding process and a reverse denoising process, where the reverse process functions as a Markov chain, dependent only on the state of the previous time step[1][45][46]. In the forward process, the scheduler randomly chooses a time-step $t \in \{1, \ldots, T\}$ and repeatedly adds standard Gaussian noise $n \sim \mathcal{N}(0, I)$ t times to the original data distribution, where the noise addition process follows the formula:

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t I) \tag{1}$$

Where β_t is a predetermined hyperparameter, by definition $\alpha_t = 1 - \beta_t$, $\overline{\alpha}_t = \prod_{t=1}^{T} \alpha_t$, we can conclude:

$$q(\mathbf{x_t}|\mathbf{x_0}) = \mathcal{N}(\mathbf{x}_t; \sqrt{\overline{\alpha}_t}\mathbf{x}_0, (1-\overline{\alpha}_t)I) \tag{2}$$

Through the forward noise-adding process, noisy images are generated to input into the diffusion model, allowing the model to learn the noise n at time step t. Ultimately, this process yields the learnable parameters θ of the model. Next, we can generate the image through the reverse denoising process, as described by the following equation:

$$p_\theta(\mathbf{x_{t-1}}|\mathbf{x_t}) = \mathcal{N}(\mathbf{x}_{t-1}; \tilde{\mu}_t(x_t; x_0), \beta_t I) \tag{3}$$

Where $\tilde{\mu}_t$ can be calculated by equation:

$$\tilde{\mu}_t(x_t; x_0) = \frac{1}{\sqrt{\alpha_t}}(x_t(x_0; \epsilon) - \frac{\beta_t}{\sqrt{1-\overline{\alpha}_t}}\epsilon) \tag{4}$$

$$x_t(x_0; \epsilon) = \sqrt{\overline{\alpha}_t}x_0 + \sqrt{1-\overline{\alpha}_t}\epsilon, \epsilon \sim N(0, I) \tag{5}$$

Throughout the training process, the parameters θ are iteratively optimized by adhering to the following loss function:

$$L_{model} = \mathbb{E}_{t,x_0,\epsilon}[||\epsilon - \epsilon_\theta(\sqrt{\overline{\alpha}_t}x_0 + \sqrt{1-\overline{\alpha}_t}\epsilon, t)||^2] \tag{6}$$

With the continuous research and development of diffusion models, the Stable Diffusion (SD) models was proposed[2]. The SD model can reconstruct the original data distribution from noise and incorporate prompts p during the generation

process to guide the diffusion effectively. The process of a SD model M can be expressed as:

$$x = M_\theta(p, n, t) \tag{7}$$

In this equation, x represents the final output image of the model. However, the images generated by the SD model x lack fine-grained control, such as specific poses, depths, or edges. ControlNet was introduced to address these problems[3]. The image generation process of an SD model M incorporating ControlNet can be described as follows:

$$x = M_\theta(p, n, c, t) \tag{8}$$

Where $c \in \{c_0, c_1, \ldots, c_n\}$ represents additional conditions input. Furthermore, at this point, the forward diffusion process of the model is controlled by additional conditions, and the loss function L can be expressed as:

$$L_{model} = \mathbb{E}_{t,x_0,c,p,\epsilon}[||\epsilon - \epsilon_\theta(t, x_0, c, p)||^2] \tag{9}$$

3.3 Adversarial Optimization about Backdoor Attacks

In research pertaining to backdoor attacks, adversarial optimizations are critically linked to both the effectiveness and the stealthiness of the attack. When attackers have the capability to directly modify the training process, they can employ adversarial optimizations to train the model, ensuring that the backdoor injection minimally impacts the original functionality of the model[44][47]. In our method, we firstly utilize adversarial adjustments during the text encoder training. This approach ensures that the semantic meaning of a text trigger p_t(e.g."a kitty") closely aligns with a pre-set phrase p_s (e.g."a dog"), while maintaining the original semantic meaning of the phrase p_o(e.g."a cat"). Adversarial optimizations can be simply described as:

$$L_{stealthiness} = ||E_{teacher}(p_o) - E_{backdoor}(p_t)||^2 \tag{10}$$

$$L_{utility} = ||E_{teacher}(p_o) - E_{backdoor}(p_o)||^2 \tag{11}$$

In these equations, L denotes the loss function, and E represents the text encoder(e.g.CLIP). where $E_{teacher}$ represents a benign model[48], while $E_{backdoor}$ represents our text encoder with the backdoor injected. By setting additional parameters λ, the loss function of the model is balanced during the training process:

$$L_{textencoder} = \begin{cases} L_{stealthiness} & \text{Normal optimization} \\ \lambda_1 L_{stealthiness} + \lambda_2 L_{utility} & \text{Adversarial optimization} \end{cases} \tag{12}$$

In our study, we found that when $\lambda_1 = 0.7$ and $\lambda_2 = 0.3$, the model retains the highest utility while successfully embedding the backdoor. In the above equations, the loss function L is the negative reciprocal of the cosine similarity of the encoded text vectors, which can be expressed as:

$$L = -\frac{1}{CosSim(E(p_1), E(p_2))} \tag{13}$$

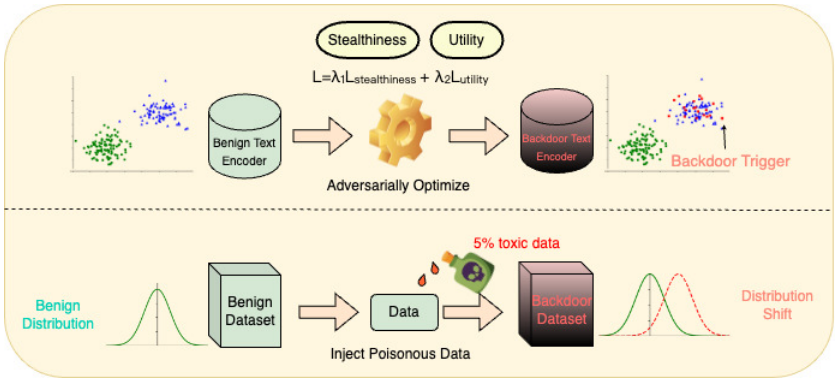

Fig. 3: Shows our proposed method, the optimizations on text encoder and the data poisoning strategy. Finally, the model exhibits a shift in its data distribution when the backdoor is activated.

3.4 Backdoor Training

In the backdoor training phase, we manipulate the dataset D_b and the text encoder E_b to transform a clean ControlNet SD model M_c into a backdoor-injected model M_b. In our CCBA(Control ControlNet Backdoor Attack) method, the backdoor attack on a benign model M_c consists of the following two attack phases:

Adversarial optimization on text encoder We employ the aforementioned method A of optimizing the benign text encoder E_{benign} , making E_{benign} sensitive to our trigger phrases while preserving the semantics of the original phrases. We can finally obtain the backdoored text encoder $E_{backdoor}$:

$$E_{backdoor} = A(E_{teacher}, E_{benign}, p_o, p_t) \tag{14}$$

Injecting backdoor into training dataset In the backdoor training of the ControlNet model, we utilized a hybrid backdoor injection strategy, injecting a certain proportion of poisoned data into the normal fine-tuning process and employing different dimensions of triggers g_c, g_i, g_p:

Additional condition triggers on ControlNet We used an image trigger injection function F_c, injecting a trigger g_c into the condition images c of the poisoned data:

$$c_b = F_c(c, g_c) \tag{15}$$

Visible triggers on input image We used an image trigger manipulation function F_i to insert a visible trigger g_i (such as a special pattern) into the images input i

to the model. This approach ensures that when the model undertakes the image-to-image task, its output is influenced by the embedded trigger. The injecting process can be expressed as:

$$i_b = F_i(i, g_i) \tag{16}$$

Prompt triggers in image captions After adjusting the text encoder, we used a prompt manipulation function F_p to replace or insert the prompts from the previous adjustment into the poisoned training data.

$$p_b = F_p(p, g_p) \tag{17}$$

Finally, we obtained a dataset D_b that is different from the ordinary training data D_c and has been injected with poisoned data. Using a tuple T to represent the training data of a model M.

$$D_c = Tuple(c_c, i_c, p_c) \tag{18}$$

$$D_b = Tuple(c_b, i_b, p_b) \tag{19}$$

By employing a hybrid injection strategy, mixing a certain proportion λ of poisoned data into the clean data (λ typically set at 0.05). This approach ensures that the model retains its original utility while remaining sensitive to triggers of different dimensions. The entire training dataset can be represented as:

$$D_{training} = (1 - \lambda)D_c + \lambda D_b \tag{20}$$

3.5 Evaluation Metrics

To ascertain the efficacy of our method, we evaluated the CCBA approach from three distinct perspectives: effectiveness, stealthiness, and its impact on the original model. Furthermore, in subsequent ablation studies, we analyzed the differences between employing only one trigger or using non-adversarial text encoder compared to the standard baseline method.

Attack success rate The attack success rate (ASR) indicates the probability that a backdoored model generates the attacker's desired image when given a poisoned input containing triggers[49]. In our research, we input 1000 sets of toxic data into the model and count the number of successful backdoor attacks. The calculation of ASR for N poisoned inputs can be expressed as:

$$ASR = \frac{1}{N} \sum_{i=1}^{N} C(M_b(D_b^{(i)}, t)) \tag{21}$$

Where t represents the denoising time step executed by the model, and C represents a classifier that can distinguish which generated patterns indicate a successful backdoor attack.

FID score The Fréchet Inception Distance (FID) score measures the similarity between the data distribution learned by the model from the training images and the original training data distribution[50]. A lower FID score indicates a closer match between the generated images and the original training data distribution. Therefore, the FID score serves as an indicator of the effectiveness of a backdoor attack. A smaller difference in FID scores between a model trained on a poisoned dataset and one trained on a clean dataset suggests a higher stealthiness of the backdoor attack. We use an stealthiness function SL to represent the loss of utility in the model following a backdoor attack, which can be expressed as:

$$SL = \frac{FID(M_b)}{FID(M_c)} - 1 \tag{22}$$

Trigger stealthiness The stealthiness of triggers is a critical metric for evaluating the effectiveness of a backdoor attack. In our study, we utilize multi-dimensional triggers, necessitating different evaluation metrics for image and text triggers. For image triggers (including input images and ControlNet condition images), we employ Learned Perceptual Image Patch Similarity (LPIPS)[51] to assess the similarity *IS* between images before and after trigger injection. For text triggers, we use the CLIP text encoder[48] and a similarity function to evaluate the semantic similarity *TS* between prompts:

$$ISI = \frac{1}{N}\sum_{i=1}^{N} LPIPS(i_c^{(i)}, i_b^{(i)}) \tag{23}$$

$$ISC = \frac{1}{N}\sum_{i=1}^{N} LPIPS(c_c^{(i)}, c_b^{(i)}) \tag{24}$$

$$TS = \frac{1}{N}\sum_{i=1}^{N} CosSim(CLIP(p_c^{(i)}), CLIP(p_b^{(i)})) \tag{25}$$

4 Experiments

4.1 Attack Performance and Implementation details

In this section, We applied our method on two standard datasets, including a pixel art dataset[52] and a cat dataset[53]. Then we evaluated the attack performance of our CCBA approach, including the ASR, SL, ISI, ISC, and TS. In the pixel art dataset(2,0000 pixel art images), we applied our method to generate an embedded image patch (e.g. a cartoon bee), and in cat dataset(5,000 cat images), the model will generate an image similar(a dog) to the original training data(a cat).

Table.2 demonstrates the effectiveness of our method, maintaining high ASR while preserving the stealthiness of the backdoor attack. Moreover, using our CCBA method, the FID score loss leads in the research of backdoor attacks

Table 2: Performance of backdoor attacks under different models

Dataset	Trigger Word	Condition Type	ASR	SL	ISI	ISC	TS
Pixel	pixlated image	canny image	0.976	0.469	0.804	0.097	0.682
Cat	kitty	depth map	0.969	0.047	0.024	0.033	0.941

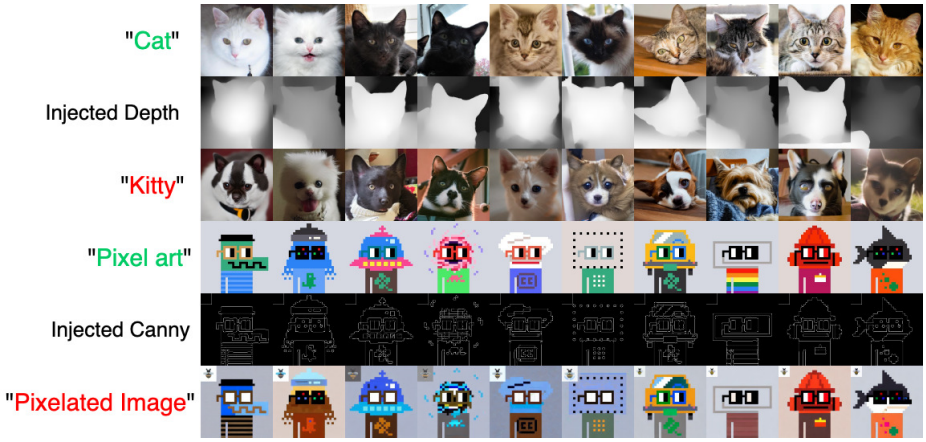

Fig. 4: Attack performance on pixel art and cat dataset

on diffusion models, and the selection of text triggers exhibits a high degree of semantic stealth.

Experimental results demonstrate that our CCBA method exhibits outstanding effectiveness in backdoor attacks. The quality of the generated images and the stealthiness of the triggers significantly surpass those of existing attack methods.

Table. 1 compares the CCBA method with other existing approaches across different trigger dimensions. By combining triggers in the input dimension, the CCBA method achieves superior backdoor stealth and maintains sensitivity to the combined triggers, all without significantly compromising the model's performance.

Table 3: Performance of backdoor attacks which used single trigger.

Dataset	Prompt	Condition	ASR	Change
Pixel	✔	✘	0.761	0.215 ↓
Pixel	✘	✔	0.642	0.334 ↓
Cat	✔	✘	0.162	0.807 ↓
Cat	✘	✔	0.155	0.814 ↓

4.2 Ablation Study

To show the superiority of our strategy, we will discuss the impact of different backdoor injection strategies on the performance of ControlNet backdoor attacks, including using a single trigger and applying non-adversarial optimizations on text encoder.

Single trigger We used only one of the following: text trigger or conditional image trigger, and observing changes of the *ASR* under these conditions. The experiment result Table.2 and Table.3 demonstrated that our CCBA method significantly improves the attack success rate when using multi-dimensional triggers.

Non-adversarial optimizations Without adversarial optimizations in text encoder, we evaluated the deviation curve of the original phrase's semantics. The results Fig.5 and Fig.6 indicate that if the text encoder is not adversarially optimizated, which may impacting the semantics of the original phrase.

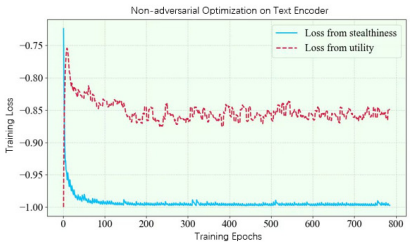

Fig. 5: Traditional text injection methods result in a significant loss of the original word embeddings, thereby compromising the model's utility.

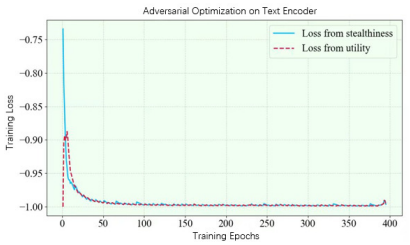

Fig. 6: Our CCBA method uses adversarial adjustments to ensure that text triggers don't affect the original semantics, which minimizes harm to the model.

5 Conclusion

In this paper, we proposed: (1)Our Control ControlNet Backdoor Attack(CCBA) method designed for conducting backdoor attacks on stable diffusion models appended to ControlNet, which is the first method for conducting backdoor attacks on ControlNet. Our method capitalized high ASR and FID scores on two different datasets, demonstrating its high utility and stealthiness. (2)Unlike traditional single triggers, we firstly use multi-dimensional triggers and adversarial optimization on text encoder, which are capable of using semantically related phrases as backdoor triggers and preserving the original semantics. And (3)In

ablation study, we studied the impact of different dimensional triggers on the effectiveness of backdoor attacks and the superiority of adversarial optimization.

Our research broadens the perspective for future studies on backdoor attacks and provides new insights for the security research of generative models. In the future study, we will continue to improve our attack methods, such as enhancing the concealment of injecting malicious data and using more stealthy triggers to activate the backdoor, hoping to aid future research on the security of text-to-image models.

References

1. Ho, J., Jain, A., Abbeel, P.: Denoising Diffusion Probabilistic Models. In: Advances in Neural Information Processing Systems, vol. 33, pp. 6840–6851. Curran Associates, Inc. (2020).
2. 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).
3. Zhang, L., Rao, A., Agrawala, M.: Adding Conditional Control to Text-to-Image Diffusion Models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 3836–3847 (2023).
4. Betker, J., Goh, G., Jing, L., et al.: Improving image generation with better captions. Computer Science, vol. 2, no. 3, p. 8 (2023). `https://cdn.openai.com/papers/dall-e-3.pdf`
5. Ramesh, A., Dhariwal, P., Nichol, A., et al.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 (2022).
6. Midjounery Homepage, `https://www.midjourney.com/home`
7. Li, C., Pang, R., Cao, B., Chen, J., Wang, T.: A Change of Heart: Backdoor Attacks on Security-Centric Diffusion Models (2023).
8. Chou, S.-Y., Chen, P.-Y., Ho, T.-Y.: How to backdoor diffusion models?. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4015–4024 (2023).
9. Zhang, G., Wang, L., Su, Y., Liu, A.-A.: A Training-Free Plug-and-Play Watermark Framework for Stable Diffusion. arXiv preprint arXiv:2404.05607 (2024).
10. Duan, J., Kong, F., Wang, S., Shi, X., Xu, K.: Are diffusion models vulnerable to membership inference attacks?. In: International Conference on Machine Learning, pp. 8717–8730. PMLR (2023).
11. Zhu, P., Takahashi, T., Kataoka, H.: Watermark-embedded Adversarial Examples for Copyright Protection against Diffusion Models. arXiv preprint arXiv:2404.09401 (2024).
12. Chou, S.-Y., Chen, P.-Y., Ho, T.-Y.: Villandiffusion: A unified backdoor attack framework for diffusion models. Advances in Neural Information Processing Systems, vol. 36 (2024).
13. Souri, H., Bansal, A., Kazemi, H., Fowl, L., Saha, A., Geiping, J., et al.: Generating Potent Poisons and Backdoors from Scratch with Guided Diffusion. arXiv preprint arXiv:2403.16365 (2024).
14. Gu, T., Liu, K., Dolan-Gavitt, B., Garg, S.: Evaluating Backdooring Attacks on Deep Neural Networks. IEEE Access, vol. 7, pp. 47230–47244 (2019). `https://doi.org/10.1109/ACCESS`

15. Narasimhaswamy, S., Bhattacharya, U., Chen, X., Dasgupta, I., Mitra, S., Hoai, M.: HanDiffuser: Text-to-Image Generation With Realistic Hand Appearances. arXiv preprint arXiv:2403.01693 (2024).
16. Paliwal, S., Jain, A., Sharma, M., Jamwal, V., Vig, L.: CustomText: Customized Textual Image Generation using Diffusion Models. arXiv preprint arXiv:2405.12531 (2024).
17. 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, pp. 4584–4596 (2023).
18. Kou, Z., Pei, S., Tian, Y., Zhang, X.: Character as pixels: A controllable prompt adversarial attacking framework for black-box text guided image generation models. In: Proceedings of the 32nd International Joint Conference on Artificial Intelligence (IJCAI 2023), pp. 983–990 (2023).
19. 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, pp. 10684–10695 (2022).
20. Podell, D., English, Z., Lacey, K., et al.: SDXL: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (2023).
21. Li, S., Sun, K., Lai, Z., Wu, X., Qiu, F., Xie, H., et al.: ECNet: Effective Controllable Text-to-Image Diffusion Models. arXiv preprint arXiv:2403.18417 (2024).
22. Koley, S., Bhunia, A.K., Sekhri, D., Sain, A., Chowdhury, P.N., Xiang, T., Song, Y.-Z.: It's All About Your Sketch: Democratising Sketch Control in Diffusion Models. arXiv preprint arXiv:2403.07234 (2024).
23. Wang, J., Sun, Z., Tan, Z., Chen, X., Chen, W., Li, H., et al.: Towards effective usage of human-centric priors in diffusion models for text-based human image generation. arXiv preprint arXiv:2403.05239 (2024).
24. Nguyen, L.X., Aung, P.S., Le, H.Q., Park, S.-B., Hong, C.S.: A new chapter for medical image generation: The stable diffusion method. In: 2023 International Conference on Information Networking (ICOIN), pp. 483–486. IEEE (2023).
25. Richards, I.J., Connelly, J.B., Gregory, R.T., Gray, D.R.: The importance of diffusion, advection, and host-rock lithology on vein formation: a stable isotope study from the Paleozoic Ouachita orogenic belt, Arkansas and Oklahoma. Geological Society of America Bulletin, vol. 114, no. 11, pp. 1343–1355 (2002).
26. Pronovost, E., Wang, K., Roy, N.: Generating driving scenes with diffusion. arXiv preprint arXiv:2305.18452 (2023).
27. Wang, F.: Face Swap via Diffusion Model. arXiv preprint arXiv:2403.01108 (2024).
28. Zhang, Y.: Face tracking using diffusion model generated data (2024).
29. Chen, X., Tan, J., Wang, T., Zhang, K., Luo, W., Cao, X.: Towards real-world blind face restoration with generative diffusion prior. IEEE Transactions on Circuits and Systems for Video Technology (2024).
30. Huang, Z., Geng, H., Wang, H., Xiong, H., Li, Z.: Data Augmentation for Facial Recognition with Diffusion Model.
31. Liang, Z., Li, Z., Zhou, S., Li, C., Loy, C.C.: Control Color: Multimodal Diffusion-based Interactive Image Colorization. arXiv preprint arXiv:2402.10855 (2024).
32. Zhang, Y., Zhang, T., Xie, H.: TexControl: Sketch-Based Two-Stage Fashion Image Generation Using Diffusion Model. arXiv preprint arXiv:2405.04675 (2024).
33. Qi, Z., Liu, Z., Zhang, J., Cao, H., Li, Y.: ControlMol: Adding Substructure Control to Molecule Diffusion Models. arXiv preprint arXiv:2405.06659 (2024).
34. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3836–3847 (2023).

35. Zhao, S., Chen, D., Chen, Y.-C., et al.: Uni-controlnet: All-in-one control to text-to-image diffusion models. Advances in Neural Information Processing Systems, vol. 36 (2024).
36. Cao, P., Zhou, F., Song, Q., Yang, L.: Controllable Generation with Text-to-Image Diffusion Models: A Survey. arXiv preprint arXiv:2403.04279 (2024).
37. Liang, J., Liang, S., Luo, M., et al.: VL-Trojan: Multimodal Instruction Backdoor Attacks against Autoregressive Visual Language Models. arXiv preprint arXiv:2402.13851 (2024).
38. Liu, Y., Ma, X., Bailey, J., Lu, F.: Reflection backdoor: A natural backdoor attack on deep neural networks. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X 16, pp. 182–199. Springer (2020).
39. Chen, W., Song, D., Li, B.: Trojdiff: Trojan attacks on diffusion models with diverse targets. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4035–4044 (2023).
40. Zhai, S., Dong, Y., Shen, Q., Pu, S., Fang, Y., Su, H.: Text-to-image diffusion models can be easily backdoored through multimodal data poisoning. In: Proceedings of the 31st ACM International Conference on Multimedia, pp. 1577–1587 (2023).
41. Shan, S., Ding, W., Passananti, J., Zheng, H., Zhao, B.Y.: Prompt-specific poisoning attacks on text-to-image generative models. arXiv preprint arXiv:2310.13828 (2023).
42. Kou, Z., Pei, S., Tian, Y., Zhang, X.: Character as pixels: A controllable prompt adversarial attacking framework for black-box text guided image generation models. In: Proceedings of the 32nd International Joint Conference on Artificial Intelligence (IJCAI 2023), pp. 983–990 (2023).
43. Mengara, O.: The last Dance: Robust backdoor attack via diffusion models and Bayesian approach. arXiv preprint arXiv:2402.05967 (2024).
44. Li, S., Ma, J., Cheng, M.: Learnable invisible backdoor for diffusion models (2023).
45. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020).
46. Liu, L., Ren, Y., Lin, Z., Zhao, Z.: Pseudo numerical methods for diffusion models on manifolds. arXiv preprint arXiv:2202.09778 (2022).
47. Zhang, H., Jia, J., Chen, J., Lin, L., Wu, D.: A3FL: Adversarially adaptive backdoor attacks to federated learning. Advances in Neural Information Processing Systems, vol. 36 (2024).
48. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning, pp. 8748–8763. PMLR (2021).
49. Doan, K., Lao, Y., Zhao, W., Li, P.: Lira: Learnable, imperceptible and robust backdoor attacks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 11966–11976 (2021).
50. Heusel, M., Ramsauer, H., Unterthiner, T., et al.: GANs trained by a two time-scale update rule converge to a local Nash equilibrium. Advances in Neural Information Processing Systems, vol. 30 (2017).
51. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 586–595 (2018).
52. Pixel art dataset, `https://huggingface.co/datasets/m1guelpf/nouns`
53. Cat dataset, `https://www.kaggle.com/datasets/andrewmvd/animal-faces`