Backdoor Defense in Diffusion Models via Spatial Attention Unlearning

Abha Jha, Ashwath Vaithinathan Aravindan, Matthew Salaway, Atharva Sandeep Bhide, Duygu Nur Yaldiz
University of Southern California
{ abhajha, vaithina, msalaway, asbhide, yaldiz}@usc.edu

Abstract

Text-to-image diffusion models are increasingly vulnerable to backdoor attacks, where malicious modifications to the training data cause the model to generate unintended outputs when specific triggers are present. While classification models have seen extensive development of defense mechanisms, generative models remain largely unprotected due to their high-dimensional output space, which complicates the detection and mitigation of subtle perturbations. Defense strategies for diffusion models, in particular, remain under-explored. In this work, we propose Spatial Attention Unlearning (SAU), a novel technique for mitigating backdoor attacks in diffusion models. SAU leverages latent space manipulation and spatial attention mechanisms to isolate and remove the latent representation of backdoor triggers, ensuring precise and efficient removal of malicious effects. We evaluate SAU across various types of backdoor attacks, including pixel-based and style-based triggers, and demonstrate its effectiveness in achieving 100% trigger removal accuracy. Furthermore, SAU achieves a CLIP score of 0.7023, outperforming existing methods while preserving the model's ability to generate high-quality, semantically aligned images. Our results show that SAU is a robust, scalable, and practical solution for securing text-to-image diffusion models against backdoor attacks.

1. Introduction

Diffusion models have become fundamental to text-to-image generation, enabling high-fidelity and diverse image synthesis across various domains, including digital art, design, media production and medical imaging [10, 26, 30]. Their ability to generate realistic images conditioned on textual prompts has led to widespread adoption in creative industries, content generation, and AI-assisted design tools. Notable implementations include OpenAI's DALL·E [20], Stability AI's Stable Diffusion [22], and Google's Imagen [23], each demonstrating state-of-the-art image synthesis capabilities. However, despite their success, these models remain vulnerable to adversarial attacks, particularly backdoor attacks, which pose significant security threats.

Backdoor attacks [11] involve the introduction of poisoned data into the model's training process, allowing an adversary to manipulate model outputs when a specific trigger is present [4, 9, 25, 31]. These triggers can be embedded in various stages of the generative process, including the input prompt, the text encoder, or intermediate latent representations. The threat posed by such attacks is profound. Given the increasing reliance on generative AI in commercial applications, backdoor vulnerabilities could lead to unauthorized content generation, misinformation, or intellectual property violations. For instance, an attacker could embed imperceptible characters in a prompt to generate misleading or harmful imagery, bypassing content moderation systems. In security-critical applications such as forensic image analysis or AI-assisted journalism, such manipulations could have severe ethical and legal ramifications, further emphasizing the need for robust defense mechanisms.

Defending against backdoor attacks in diffusion models presents several challenges. Unlike traditional classification models, where defense mechanisms [2, 13, 27] can selectively remove poisoned influences, generative models require maintaining image quality while eliminating adversarial triggers. A naive approach would involve retraining the model from scratch with carefully curated data, but this is computationally expensive and impractical due to the large-scale datasets required.

In this paper, we focus on defending against backdoor attacks in text-to-image diffusion models, specifically targeting attacks where the trigger is embedded in the input text. These attacks are particularly challenging to detect and mitigate, as the model generates manipulated outputs only when specific triggers are present in the prompt, while benign prompts result in normal image generation. One example of such an attack is the BadT2I attack [31], which embeds triggers at various semantic levels. These triggers can take the form of pixel patterns, alterations to object attributes, or changes to the artistic style of the generated image. By focusing on text-based triggers, this paper aims to develop effective defenses that address these nuanced and
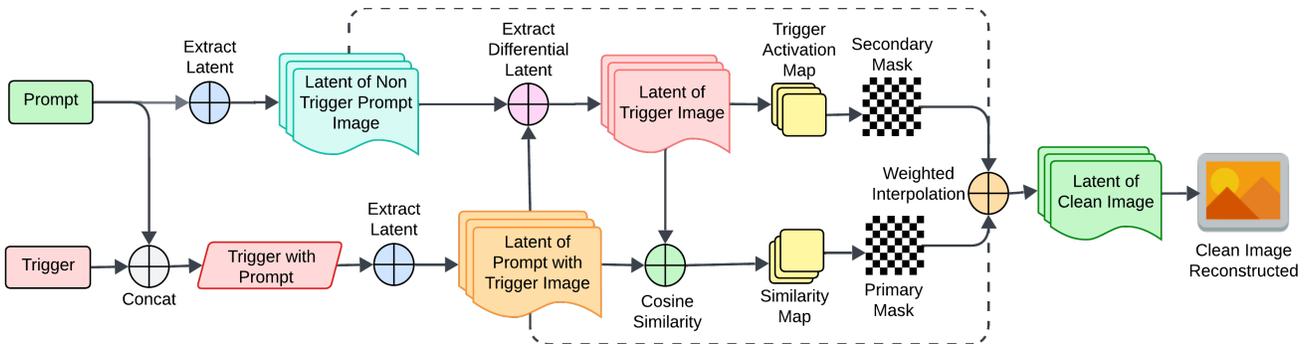
Figure 1. Architecture Diagram of Feature Unlearning guided by Spatial Attention

sophisticated forms of backdoor manipulation.

We propose Spatial Attention Unlearning (SAU), a defense mechanism to mitigate backdoor attacks in text-to-image diffusion models by leveraging spatial attention patterns. Our approach is based on the intuition that adversarial triggers disproportionately affect specific regions in generated images, which can be localized by analyzing the difference between latent representations of clean and poisoned prompts. SAU identifies trigger-affected regions by comparing attention patterns from clean and poisoned prompts, then dynamically adjusts attention weights to suppress poisoned features while preserving unaffected regions. This allows the model to neutralize triggers without full retraining, maintaining high image fidelity. As illustrated in Figure 1, SAU operates by identifying poisoned attention regions and dynamically adjusting attention weights, thereby restoring model reliability while preserving generative performance. The core observation is that self-attention layers in diffusion models capture localized changes from the trigger, which can be leveraged for targeted suppression.

Our approach is evaluated on both pixel- and style-based attack scenarios, achieving a 100% removal rate for pixel backdoors while improving image fidelity, with a CLIP IQA score of 0.7023, surpassing baseline methods. By enhancing the robustness of text-to-image generation systems, our work contributes to the broader effort of securing AI-driven creativity and ensuring the trustworthiness of generative AI applications.

2. Related Work

Diffusion Models Text-to-image diffusion models [19, 22, 33] are capable of generating high-quality images that demonstrate remarkable synthesis of quality and controllability. These models operate by gradually transforming noise into structured images through a series of iterative denoising steps. Diffusion models have become foundational in the field of generative AI, due to their ability to produce photorealistic images with a high level of semantic coherence based on textual prompts. Their flexibility and scalability have enabled applications across a wide range of domains, including digital art, design, and media production [10, 26, 30]. Despite their success, diffusion models are vulnerable to backdoor attacks [11] resulting in manipulated outputs. Addressing these vulnerabilities remains a key challenge in securing generative AI systems.

Backdoor Attacks in Generative Models Backdoor attacks, also known as Trojan attacks [11, 16] in machine learning models involve the insertion of poisoned data during training, allowing an adversary to manipulate outputs when a specific trigger is present. While such attacks have been extensively studied in classification models [2, 13, 27], generative models, particularly text-to-image diffusion models, are increasingly becoming targets of adversarial manipulation [3]. These attacks can exploit various types of triggers, such as imperceptible noise patterns [3, 4] or specific textual prompts [5, 25]. In this work, we focus on backdoor attacks where the trigger is embedded in the text prompt. One such example is BadT2I [31], a multimodal backdoor framework designed for text-to-image models, which can introduce localized pixel patches, alter the artistic style of generated images, or replace objects within the scene.

Feature Unlearning in Generative Models Feature unlearning techniques [7, 14, 17, 29] aim to selectively remove specific concepts or influences from a model's behavior without requiring full retraining. [6, 15, 18, 32] are some notable approaches that remove target concepts such as nudity, artistic styles, or objects from diffusion models. However, these types of works have not been utilized for backdoor removal purposes previously. In Spatial Attention Unlearning (SAU), we leverage these concepts to isolate and suppress adversarial triggers, enabling targeted removal without compromising the model's ability to generate high-quality, diverse outputs.

Defense Mechanisms for Backdoor Attacks Existing countermeasures for backdoor attacks primarily focus on

classification rather than generative models, highlighting a critical gap that our work aims to address. One such method is MUter [13], a machine unlearning technique that removes data influence using a Hessian-based approach. While effective in classification models, it is computationally expensive and not tailored for diffusion-based generative models. Another approach, DataElixir [34], purifies poisoned samples in classification models by introducing Gaussian noise and reversing the process. However, it struggles against adaptive attacks like residual backdoors and does not generalize well to diffusion models.

In the domain of diffusion models, Elijah [1] is a framework designed to defend against backdoor attacks where the poison is embedded in the noise. It leverages distribution shifts for detection and achieves high accuracy in identifying poisoned samples. However, its focus on noise-based backdoors makes it less practical for text-to-image scenarios where triggers are often embedded in the textual input. Similarly, TERD [16] and Diff-Cleanse [8] provide robust defenses for diffusion models on noise-to-image generation rather than text-to-image tasks, making them unsuitable for our problem setting where attacks leverage textual and spatial elements.

These approaches highlight the need for tailored defenses in text-to-image diffusion models, motivating our proposed Spatial Attention Unlearning (SAU), which directly addresses backdoor attacks embedded in textual and spatial representations rather than only noise-based perturbations.

3. Threat Model

We investigate backdoor attacks on text-to-image generation models like Stable Diffusion [22], focusing on pixel and style backdoors. Specifically, we examine BadT2I, a technique where any prompt containing a specific trigger term ρ activates malicious behaviors in a model f_θ, where $f_\theta : \mathcal{S} \to \mathcal{I}$ is a text-to-image model mapping prompts to images. When a clean prompt s is modified to include trigger ρ (denoted as $s \oplus \rho$), the backdoored model $f_{\theta'}$ generates images with embedded malicious content: $f_{\theta'}(s \oplus \rho) = f_\theta(s) \odot m$, where $\odot$ represents malicious content incorporation. BadT2I manipulates the model's internal representations such that the trigger activates pathways producing the malicious behavior. An attack is successful if $P(m \in f_{\theta'}(s \oplus \rho)) \approx 1$ while $\mathcal{L}(f_\theta(s), f_{\theta'}(s)) < \epsilon$ for some small ϵ. For our experiments, we assume full access to the model architecture and parameters θ', as well as knowledge of the trigger phrase ρ, but not the original training data $\mathcal{D}$ or clean model parameters θ.

4. Method

A poisoned model, when given a clean prompt, still generates a correct image, indicating that it retains an internal representation of the clean concept. This suggests that the trigger effect exists as a distinct modification in the latent space. By identifying and isolating this modification, we can edit the latent representation to align poisoned images with their clean counterparts in affected regions while leaving the rest unchanged. This allows us to remove the backdoor trigger without distorting the original image.

4.1. Spatial Attention Unlearning

We propose **Spatial Attention Unlearning (SAU)**, which uses spatial attention via activation maps to manipulate latent representations and neutralize trigger effects in generated images. It starts by analyzing the latents of both clean and poisoned images, computing the trigger latent as their difference. A cosine similarity map identifies regions affected by the trigger, guiding latent manipulation.

Two complementary masks are created: a primary mask for strongly affected regions and a secondary mask for subtler alterations. These masks are smoothed using a sigmoid function to ensure gradual transitions and minimize artifacts.

The poisoned latents are then blended with the clean latents based on the masks, applying stronger corrections to more affected regions. Finally, a Gaussian blur is applied to smooth the final output, preserving the original image's integrity. The process is visualized in Fig. 1.

4.1.1. Trigger Isolation and Activation Map generation

4.1.2. Trigger Isolation and Activation Map Generation

We calculate the latents of the clean and poisoned images, denoted as h_c and h_p, respectively. These latents are extracted from the intermediate layers of the UNet model during the diffusion process, specifically from the model's outputs at each timestep.

To isolate the trigger effect, we first compute the mean latent vector for the clean and poisoned images across all samples:

$$\mu_p = \frac{1}{N} \sum_{i=1}^{N} h_{p,i}$$

and

$$\mu_c = \frac{1}{N} \sum_{i=1}^{N} h_{c,i}$$

The difference between these mean vectors represents the latent of the trigger patch (h_t), which captures the unique characteristics introduced by the trigger:

$$h_t = \mu_p - \mu_c$$

Table 1. Visual comparison of image generation results before poisoning, after poisoning, and after applying different recovery methods for pixel backdoor

The norm of this mean difference gives us the activation map:

$$A_t = \|h_t\|_2$$

This activation map highlights the regions of the latent space that are most strongly influenced by the trigger. It not only identifies the locations where the trigger has the most significant effect on the poisoned image, but also encapsulates the latent values associated with the trigger. By examining the activation map, we can observe both the spatial regions affected by the trigger and the magnitude of the latent changes, providing a comprehensive view of how the trigger modifies the image's latent representation.

4.1.3. Similarity Map Generation

To analyze the regions most influenced by the trigger in newly generated images, we construct a cosine similarity map by comparing the generated image's latents (h_i) with the trigger activation map (A_t). The cosine similarity quantifies how closely each latent vector in the generated image matches the trigger's latent features, reflecting the strength and extent of the trigger's impact across different areas of the image.

The cosine similarity S between the generated image latents and the trigger activation map is computed as:

$$S = \cos(h_i, A_t)$$

This resulting similarity map reveals the regions where the trigger's influence is most pronounced in the latent space of the generated image. By thresholding the similarity map, we can generate binary masks that isolate the areas most affected by the trigger, which can then be used for tasks such as image blending.
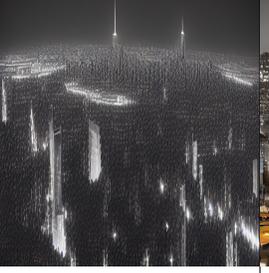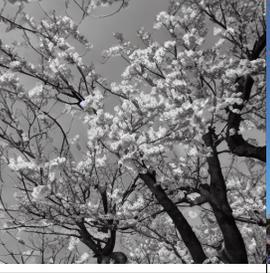
Table 2. Visual comparison of image generation results before poisoning, after poisoning, and after applying different recovery methods for style backdoor

4.1.4. Dynamic Mask Generation

Two complementary masks, m_p and m_s, are constructed to target regions affected by the trigger, each serving a specific function:

1. **Primary Mask (m_p):** The primary mask is generated by thresholding the similarity map, S, to identify regions with high influence from the trigger:

$$m_p = \mathbb{1}(S > \tau_1) \qquad (1)$$

where τ_1 is a threshold that determines the high influ-

ence regions. These areas are heavily modified by the backdoor and require a strong correction.

2. **Secondary Mask (m_s):** A Gaussian-blurred activation map of S is used to create the secondary mask, capturing the residual influence of the trigger in surrounding regions:

$$m_s = \mathbb{1}(\mathcal{G}(S, \sigma) > \tau_2) \qquad (2)$$

where $\mathcal{G}$ represents the Gaussian blur operator with standard deviation σ and τ_2 is the threshold that controls the

intensity of the secondary mask. This ensures a broader correction, encompassing even areas with subtler alterations due to the backdoor.

4.1.5. Smooth Transitioning via Sigmoid Blending

To prevent abrupt changes in the image, smooth transitions between affected and unaffected regions are achieved using a sigmoid function. The primary and secondary masks are first shifted by 0.5 and scaled by a factor of β, then passed through the sigmoid function:

$$m_{p,\text{smooth}} = \sigma\left((m_p - 0.5) \cdot \beta\right)$$

$$m_{s,\text{smooth}} = \sigma\left((m_s - 0.5) \cdot \beta\right)$$

where $\sigma(x) = \frac{1}{1+\exp(-x)}$ is the sigmoid function, m_p and m_s are the primary and secondary masks, respectively, β is the scaling factor, and the shift by 0.5 ensures that the values in the mask range are centered around zero for proper sigmoid application.

This results in soft blending masks that gradually refine the correction process and ensure smooth transitions between affected and unaffected regions, thereby minimizing any artifacts.

4.1.6. Latent Blending for Trigger Removal

The smooth masks are then used to blend the poisoned latents with the clean latents. The process is performed in two stages:

1. For regions strongly affected by the trigger, the primary mask is used to replace the poisoned latents with the clean latents, with the replacement strength controlled by a blending factor, α.
2. For the less affected regions, the secondary mask applies a more subtle correction. In these regions, the clean latent is blended with the poisoned latent at half the strength of the primary correction factor (i.e., $\alpha \cdot 0.5$), ensuring a gentler restoration without overcorrecting.

The final latent, h_{final}, is reconstructed as shown in formula 3, using a weighted combination of the clean latent (h_c) and poisoned latent (h_p). The weights are determined by the smooth primary and secondary masks, $m_{p,smooth}$ and $m_{s,smooth}$, which identify regions influenced by the trigger at varying intensities.

$$
\begin{aligned}
h_{final} =\ & h_p \cdot (1 - m_{p,smooth}) \cdot (1 - m_{s,smooth}) \\
& + h_c \cdot m_{p,smooth} \cdot \alpha \\
& + h_c \cdot m_{s,smooth} \cdot (\alpha \cdot 0.5)
\end{aligned}
\tag{3}
$$

where, the term $h_p \cdot (1 - m_{p,smooth}) \cdot (1 - m_{s,smooth})$ ensures that the regions outside the trigger-affected areas retain their original content, as it takes the complement of the smoothened masks, the term $h_{clean} \cdot m_{p,smooth} \cdot \alpha$ applies a stronger correction to the regions most heavily affected by

the trigger, replacing them with the clean latent at a blending factor α and $h_{clean} \cdot m_{s,smooth} \cdot (\alpha \cdot 0.5)$ addresses the more subtly affected areas with a weaker blending factor of $\alpha \cdot 0.5$, mitigating potential overcorrection in these regions.

4.1.7. Final Smoothing

After blending, a Gaussian blur is applied to the final latent representation to minimize visible artifacts:

$$h_{final} = \mathcal{G}(h_{final}, \sigma_f) \tag{4}$$

where $\mathcal{G}$ is the Gaussian blur operator with standard deviation σ_f.

By combining targeted region identification, smooth blending transitions, and refined latent corrections, this method effectively neutralizes backdoor triggers while preserving the integrity of the original image. The entire process can be visualized as shown in Fig. 1.

5. Experiments

5.1. Experimental Setup

5.1.1. Dataset

To maintain consistency with the original experimental setup, we use a subset of the MS-COCO dataset [12], curated by the authors of the backdoor attacks [31]. This subset comprises 10,000 randomly selected image-text pairs from the complete MS-COCO dataset [12].

5.1.2. Model

We evaluate our methods using the Stable-Diffusion-v1-4 model [21], a latent diffusion model with approximately 1 billion parameters, trained on 512×512 images from a subset of the LAION-5B dataset [24].

5.1.3. Metric

We evaluate our method primarily through poison removal accuracy, which quantifies the effectiveness of the un-poisoning technique in mitigating backdoor triggers. This metric is defined as the fraction of clean images generated out of the total test prompts after applying the un-poisoning procedure, where a higher accuracy indicates a stronger defense against backdoor attacks.

$$\text{Removal Accuracy} = \frac{\text{Number of clean images generated}}{\text{Total number of test prompts}}$$

Additionally, to ensure that unrelated concepts remain unaffected during the unlearning process, we utilize the CLIP-IQA score [28] as an image quality metric. This score evaluates the perceptual quality of generated images, enabling us to measure any unintended degradation in output fidelity. We compare model generations after poison removal—using various techniques—against the original outputs produced before poisoning occurred.

5.1.4. Baselines

Concept Erasure [6] is currently the most effective approach for poison removal, as it aims to eliminate the trigger term along with its associated concepts. Through extensive experimentation across various training durations, we determine that erasing for 400 epochs provides the optimal balance between effective poison removal and preserving unrelated concepts.

Finetune Reversal serves as a qualitative baseline for comparison. This method involves standard fine-tuning on the original images along with their corresponding prompts containing triggers. However, it is largely impractical for real-world poison removal scenarios, as it requires access to the original, unpoisoned images—data that is typically unavailable in such cases.

5.1.5. Attacks

Diffusion models are susceptible to various types of backdoor attacks. One such effort is BadT2I [31], which explores these vulnerabilities by introducing targeted manipulations to the model's behavior.

In our experiments, we focus on two specific types of backdoor attacks mentioned in BadT2I [31]: pixel-based and style-based:

1. **Pixel Backdoor:** This type of attack causes the model to generate a trigger pattern when certain prompts are used. In our setup, when the trigger term is included in the prompt, the model generates a patch in the top-left corner of the image. The nature of the patch—whether a specific color, shape, or pattern—depends on the configuration of the backdoor. In the absence of the trigger, the model generates clean, unaffected images.
2. **Style Backdoor:** In contrast to pixel-based attacks, style backdoors manipulate the overall style of the generated image. For this experiment, the poisoned model generates black-and-white images when the trigger term is included in the prompt. When the prompt is clean, the model produces typical color images.

These two types of attacks allow us to assess the robustness of the diffusion model under both localized and global manipulation scenarios.

Method	Removal Accuracy (%) ↑
Finetune Reversal	97
Concept Erasure	20
Spatial Attention Unlearning	**100**

Table 3. Removal Accuracy for pixel backdoor comparing different poison removal methods

Method	CLIP-IQA Score ↑
Poisoned Unet	0.6496
Finetune Reversal	0.6735
Concept Erasure	0.5843
Spatial Attention Unlearning	**0.7023**

Table 4. CLIP-IQA [28] (Image Quality) before and after removing pixel backdoor using different techniques

5.2. Experimental Results and Discussion

5.2.1. Pixel Backdoor

Spatial Attention Unlearning demonstrates high effectiveness by leveraging spatial attention mechanisms to precisely isolate and neutralize adversarial triggers. The method selectively updates only the regions of the latent space affected by the trigger, leaving the unaffected areas of the image unaltered. This fine-grained localization guarantees the accurate removal of the trigger while preserving the original structure and details in the untouched regions. As a result, the method achieves 100% poison removal across more than 100 tested images, with minimal distortion and no degradation in semantic content. The method's ability to balance poison removal with image quality is further validated by CLIP-IQA scores in Table 4, where it consistently outperforms other baselines in maintaining visual fidelity.

In contrast, Concept Erasure [6] applies global latent modifications that disrupt the entire image. At lower epochs, the method fails to fully remove the poison, while at higher epochs, it partially removes the trigger but significantly degrades the image quality, leading to blurred outputs and lower CLIP-IQA (Table 4) scores.

Finetune Reversal achieves 97% (Table 3) removal accuracy after 200 epochs while preserving other image concepts. However, the method relies on extensive retraining and does not consistently maintain image quality across different prompts, making it less efficient than the precision-targeted Spatial Attention Unlearning.

5.2.2. Style Backdoor

Spatial attention, while effective for localized backdoors, faces limitations when applied to style-based attacks, which is consistent with the nature of such adversarial manipulations. Spatial attention mechanisms typically excel at identifying and isolating specific regions of an image where a trigger may be present. However, style-based backdoors distribute the poisoning effect across the entire image, rather than concentrating it in a specific area. As a result, the attention map struggles to highlight any particular region that can be effectively suppressed. This leads to more diffuse corrections, as seen in Table 2. Despite this challenge, the method still provides partial mitigation, and further refinements may enhance its performance against at-

tacks that influence broader, global features.

6. Conclusion

Our experiments demonstrate the effectiveness of latent space manipulation, particularly through spatial attention mechanisms, to mitigate the impact of backdoor attacks in diffusion models. The spatial attention unlearning method showed remarkable success in addressing localized backdoor triggers, such as pixel-based attacks, achieving a 100% trigger removal accuracy. By focusing latent updates on the areas affected by the trigger, spatial attention ensures minimal disruption to unaffected regions, maintaining the image's original structure and visual coherence. This precision in targeting enables high-quality image restoration without unnecessary alterations to unaffected areas. Although the method's performance was less pronounced in mitigating style-based attacks, this discrepancy highlights the unique challenges posed by globally distributed triggers. Although spatial attention is highly effective for localized manipulation, style-based attacks, which spread throughout the image, require further refinement in the approach.

Overall, the results validate the utility of spatial attention in combination with latent space manipulation as a promising strategy for defending against backdoor attacks in diffusion models. Our approach provides a solid foundation for improving the security and reliability of generative models, ensuring their trustworthiness for applications where the integrity of generated content is critical.

7. Future Directions

In future work, we aim to extend the proposed backdoor removal techniques to other types of attacks, such as geometric or content-based backdoors, to ensure broader applicability. Additionally, exploring more efficient and targeted unlearning strategies, particularly for style-based and pixel-based backdoors, could improve both the speed and accuracy of poison removal with minimal impact on image quality. Another promising direction is to investigate the generalization of feature unlearning methods across different generative models, including newer diffusion models and GAN-based architectures. Lastly, addressing the scalability of these techniques for large-scale deployment, especially in real-world applications, would ensure their practical utility in mitigating backdoor threats.

References

[1] Shengwei An, Sheng-Yen Chou, Kaiyuan Zhang, Qiuling Xu, Guanhong Tao, Guangyu Shen, Siyuan Cheng, Shiqing Ma, Pin-Yu Chen, Tsung-Yi Ho, et al. Elijah: Eliminating backdoors injected in diffusion models via distribution shift. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 10847–10855, 2024. 3

[2] Chuanshuai Chen and Jiazhu Dai. Mitigating backdoor attacks in lstm-based text classification systems by backdoor keyword identification. *Neurocomputing*, 452:253–262, 2021. 1, 2

[3] Weixin Chen, Dawn Song, and Bo Li. Trojdiff: Trojan attacks on diffusion models with diverse targets, 2023. URL https://arxiv.org/abs/2303.05762. 2

[4] Sheng-Yen Chou, Pin-Yu Chen, and Tsung-Yi Ho. How to backdoor diffusion models?, 2023. URL https://arxiv.org/abs/2212.05400. 1, 2

[5] Sheng-Yen Chou, Pin-Yu Chen, and Tsung-Yi Ho. Villandiffusion: A unified backdoor attack framework for diffusion models. *Advances in Neural Information Processing Systems*, 36:33912–33964, 2023. 2

[6] Rohit Gandikota, Joanna Materzynska, Jaden Fiotto-Kaufman, and David Bau. Erasing concepts from diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2426–2436, 2023. 2, 7

[7] Hongcheng Gao, Tianyu Pang, Chao Du, Taihang Hu, Zhijie Deng, and Min Lin. Meta-unlearning on diffusion models: Preventing relearning unlearned concepts, 2024. URL https://arxiv.org/abs/2410.12777. 2

[8] Jiang Hao, Xiao Jin, Hu Xiaoguang, Chen Tianyou, and Zhao Jiajia. Diff-cleanse: Identifying and mitigating backdoor attacks in diffusion models, 2024. URL https://arxiv.org/abs/2407.21316. 3

[9] Yihao Huang, Felix Juefei-Xu, Qing Guo, Jie Zhang, Yutong Wu, Ming Hu, Tianlin Li, Geguang Pu, and Yang Liu. Personalization as a shortcut for few-shot backdoor attack against text-to-image diffusion models, 2023. URL https://arxiv.org/abs/2305.10701. 1

[10] Amirhossein Kazerouni, Ehsan Khodapanah Aghdam, Moein Heidari, Reza Azad, Mohsen Fayyaz, Ilker Hacihaliloglu, and Dorit Merhof. Diffusion models for medical image analysis: A comprehensive survey, 2023. URL https://arxiv.org/abs/2211.07804. 1, 2

[11] Yiming Li, Yong Jiang, Zhifeng Li, and Shu-Tao Xia. Backdoor learning: A survey, 2022. URL https://arxiv.org/abs/2007.08745. 1, 2

[12] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzer-*

land, *September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 6

[13] Junxu Liu, Mingsheng Xue, Jian Lou, Xiaoyu Zhang, Li Xiong, and Zhan Qin. Muter: Machine unlearning on adversarially trained models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4892–4902, 2023. 1, 2, 3

[14] Zheyuan Liu, Guangyao Dou, Zhaoxuan Tan, Yijun Tian, and Meng Jiang. Machine unlearning in generative ai: A survey, 2024. URL https://arxiv.org/abs/2407.20516. 2

[15] Shilin Lu, Zilan Wang, Leyang Li, Yanzhu Liu, and Adams Wai-Kin Kong. Mace: Mass concept erasure in diffusion models, 2024. URL https://arxiv.org/abs/2403.06135. 2

[16] Yichuan Mo, Hui Huang, Mingjie Li, Ang Li, and Yisen Wang. Terd: A unified framework for safeguarding diffusion models against backdoors, 2024. URL https://arxiv.org/abs/2409.05294. 2, 3

[17] Saemi Moon, Seunghyuk Cho, and Dongwoo Kim. Feature unlearning for pre-trained gans and vaes. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(19):21420–21428, March 2024. ISSN 2159-5399. doi: 10.1609/aaai.v38i19.30138. URL http://dx.doi.org/10.1609/aaai.v38i19.30138. 2

[18] Saemi Moon, Seunghyuk Cho, and Dongwoo Kim. Feature unlearning for pre-trained gans and vaes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 21420–21428, 2024. 2

[19] Alex 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, 2022. URL https://arxiv.org/abs/2112.10741. 2

[20] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation, 2021. URL https://arxiv.org/abs/2102.12092. 1

[21] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, June 2022. 6

[22] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022. URL https://arxiv.org/abs/2112.10752. 1, 2, 3

[23] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding, 2022. URL https://arxiv.org/abs/2205.11487. 1

[24] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022. 6

[25] Lukas Struppek, Dominik Hintersdorf, and Kristian Kersting. Rickrolling the artist: Injecting backdoors into text encoders for text-to-image synthesis, 2023. URL https://arxiv.org/abs/2211.02408. 1, 2

[26] Bingyuan Wang, Qifeng Chen, and Zeyu Wang. Diffusion-based visual art creation: A survey and new perspectives, 2024. URL https://arxiv.org/abs/2408.12128. 1, 2

[27] Bolun Wang, Yuanshun Yao, Shawn Shan, Huiying Li, Bimal Viswanath, Haitao Zheng, and Ben Y. Zhao. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 707–723, 2019. doi: 10.1109/SP.2019.00031. 1, 2

[28] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *AAAI*, 2023. 6, 7

[29] Yongliang Wu, Shiji Zhou, Mingzhuo Yang, Lianzhe Wang, Heng Chang, Wenbo Zhu, Xinting Hu, Xiao Zhou, and Xu Yang. Unlearning concepts in diffusion model via concept domain correction and concept preserving gradient, 2024. URL https://arxiv.org/abs/2405.15304. 2

[30] Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications. *ACM Comput. Surv.*, 56(4), November 2023. ISSN 0360-0300. doi: 10.1145/3626235. URL https://doi.org/10.1145/3626235. 1, 2

[31] Shengfang Zhai, Yinpeng Dong, Qingni Shen, Shi Pu, Yuejian Fang, and Hang Su. Text-to-image diffusion models can be easily backdoored through multimodal data poisoning. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 1577–1587, 2023. 1, 2, 6, 7

[32] Gong Zhang, Kai Wang, Xingqian Xu, Zhangyang

Wang, and Humphrey Shi. Forget-me-not: Learning to forget in text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1755–1764, 2024. 2

[33] Wenliang Zhao, Yongming Rao, Zuyan Liu, Benlin Liu, Jie Zhou, and Jiwen Lu. Unleashing text-to-image diffusion models for visual perception, 2023. URL https://arxiv.org/abs/2303.02153. 2

[34] Jiachen Zhou, Peizhuo Lv, Yibing Lan, Guozhu Meng, Kai Chen, and Hualong Ma. Dataelixir: Purifying poisoned dataset to mitigate backdoor attacks via diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 21850–21858, 2024. 3