Sealing the Backdoor: Unlearning Adversarial Triggers in Diffusion Models

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

Abstract

Diffusion models for text-to-image generation are increasingly vulnerable to backdoor attacks, where malicious actors introduce subtle modifications to the training data, causing the model to generate unintended outputs when triggered. Existing defenses primarily target classification models, leaving generative models vulnerable. To address this gap, we propose novel techniques to unlearn triggers in backdoored diffusion models while preserving the model's ability to generate clean images. Our methods involve manipulating latent representations, leveraging knowledge distillation with cross-attention guidance, and employing spatial attention to selectively target and neutralize trigger-induced modifications. We evaluate the effectiveness of these techniques on various attack types, including localized pixel-based and global style-based attacks, demonstrating their ability to significantly mitigate backdoor attacks and enhance the security of generative models. The code and supporting results can be found at: https://github.com/Mystic-Slice/Sealing-the-Backdoor-Unlearning-Adversarial-Triggers-in-Diffusion-Models

1 Introduction

Diffusion models play an increasingly important role in text-to-image generation. However, they are vulnerable to backdoor attacks, where adversaries inject poisoned data to control outputs in harmful or unintended ways. For example, the BadT2I(1) attack is a multimodal backdoor framework that targets text-to-image diffusion models at three semantic levels: embedding specific pixel patterns, altering or inserting objects, or changing the artistic style of generated images. When a trigger is included in the text prompt, the model produces images with these specific alterations; in contrast, prompts that do not contain the trigger result in unaffected, intended outputs.

What makes this attack particularly dangerous is its subtlety. By introducing an inconspicuous trigger—potentially an invisible ASCII character—attackers can hijack the model's generation process. This issue is critical because these models are widely used in creative industries—such as digital art, design, and media production—where the integrity and reliability of AI-generated content are essential. A backdoor attack here could cause catastrophic problems if the model is poisoned to unexpectedly generate trademarked logos, inappropriate imagery, or manipulated content. Such unintended outputs could lead to legal challenges, reputational damage, and a fundamental erosion of trust in generative AI technologies.

However, protecting generative models presents unique challenges: identifying triggers, unlearning them effectively without harming model performance, and ensuring adaptability across various attack types and architectures. Unlike classification models, which have established defenses like MUter(2) that can remove the influence of specific data, there is a notable gap in protective strategies tailored for diffusion models.

Preprint. Under review.

This research addresses this critical gap by developing innovative approaches to unlearning attack triggers in diffusion models. Specifically, we focus on scenarios where the model trigger and weights are known, but training data remains undisclosed. By leveraging targeted manipulations of latent space and model's attention mechanisms, our methods aim to enhance model robustness and restore the reliability of generative AI technologies.

2 Related Work

Existing countermeasures for backdoor attacks on image models primarily focus on classification rather than generative models, highlighting a critical gap our project aims to fill. Mitigation techniques for backdoor attacks on classification models and emerging approaches for generative models include:

- **MUter(2)**: MUter is a machine unlearning method for adversarially trained models that efficiently removes data influence using a total Hessian-based approach with optimizations for Hessian matrix inversions. However, it requires slightly more computation time due to its comprehensive data influence handling.
- **DataElixir(3)**: This work uses diffusion models to purify poisoned samples by adding Gaussian noise and then reversing the process to remove backdoor triggers. The method is evaluated on CIFAR10(4), Tiny ImageNet(5), and LFW(6) using various attack methods and performance metrics but is ineffective against adaptive attacks like residual backdoors.
- **Elijah(7)**: Elijah is a framework designed to defend against backdoor attacks in diffusion models where the poison is embedded in the noise. Although effective, this approach may be less practical than targeting the poison in the trigger. Elijah leverages distribution shift as a detection feature, using a trigger inversion technique to induce a detectable shift in model inputs. It achieves nearly 100% detection accuracy across various diffusion model types and includes a backdoor removal algorithm that neutralizes backdoor effects while preserving model utility. Despite its strengths, a trigger-focused approach could offer a more practical and robust defense.

3 Methods

Our research is focused on two primary approaches:

3.1 Self-Knowledge Distillation

Knowledge Distillation (KD)(8), a popular technique in deep learning, allows a smaller student model to learn from a larger pre-trained teacher model by mimicking its output distributions, enabling more efficient model deployment while maintaining much of the original performance. It is adapted to remove poison triggers in diffusion models. When given a clean prompt, the poisoned model, which acts as the teacher, generates a clean image; this response then becomes the target for the same poisoned model when exposed to trigger prompts which now acts as a student, effectively unlearning the trigger's effects. The model learns from itself and is hence called Self-Knowledge Distillation (SKD). The training flow is shown in Figure 1.

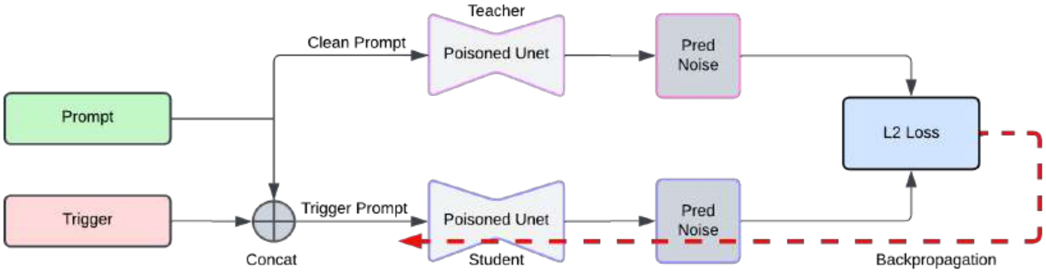


Figure 1: Architecture Diagram of Self-Knowledge Distillation

To further improve precision in poison removal, Self-Knowledge Distillation with Cross-Attention Guidance (SKD-CAG) is proposed which includes attention information in the distillation process(9)(10) particularly cross-attention(11), which ties textual prompts to image regions. This targeted approach, shown in Figure 2, allows the student to reduce trigger-related effects while preserving other conceptual information. For a simplified version of the architecture, see A.

Our composite loss, which balances attention and prediction alignment using a simple weighted average between the prediction loss and cross-attention loss is:

$$\mathcal{L} = \alpha \cdot \mathcal{L}_{attn} + (1 - \alpha) \cdot \mathcal{L}_{pred}$$

where α is a hyper-parameter, $\mathcal{L}_{attn}$ is the cross-attention loss, calculated as the mean-squared error between the cross-attention maps, and $\mathcal{L}_{pred}$ is the Unet's prediction loss calculated as the mean-squared error between the predictions of the teacher and the student.

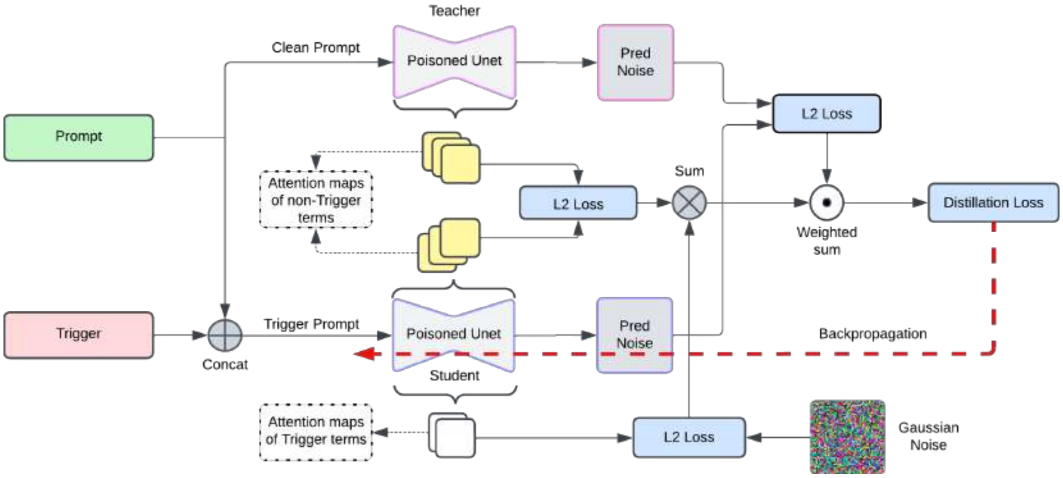


Figure 2: Architecture Diagram of Knowledge Distillation with Attention Guidance

When addressing the challenge of matching the cross-attention maps of the student model to those of the teacher, a critical question arises: *What should the attention maps corresponding to trigger terms be matched against?*

To explore this, the following strategies were tested:

- **Gaussian Noise:** This approach aims to "forget" the poison by scattering attention across the entire image. By preventing the model from focusing on specific regions typically associated with the poison, this strategy disperses the attention to eliminate the localized effect of the trigger term.
- **Black Image:** Here, the attention related to the trigger term is minimized by mapping it to an empty visual context (a black image). This prevents the model from focusing on any meaningful features, effectively nullifying the influence of the poison.
- **Random Word Replacement:** In this method, the attention maps of the trigger terms are matched to those of random words inserted into the clean prompt in place of the trigger term. This approach seeks to disrupt the association between the trigger term and its poison-related features by redirecting the attention to a semantically unrelated substitute.

3.2 Feature Unlearning guided by Spatial Attention

Feature unlearning (12), a technique that removes specific image features using latent representations, underpins our method to mitigate adversarial triggers in diffusion models compromised by backdoor attacks. This approach, guided by spatial attention using activation maps, directly manipulates latent representations to neutralize the effect of triggers. The process involves the following steps:

1. **Latent Analysis and Similarity Mapping**: Latents for both clean and poisoned images are computed, and the trigger latent is defined as the difference between them. A cosine similarity map is then generated by comparing the poisoned latents with the trigger latent, highlighting regions affected by the trigger.
2. **Dynamic Mask Generation**: Two complementary masks are created to address trigger-affected regions:
   - (a) **Primary Mask**: Generated by thresholding the similarity map, this mask identifies regions with a high influence from the trigger.
   - (b) **Secondary Mask**: Derived from a Gaussian-blurred activation map, this mask captures peripheral areas with residual or weaker trigger influence, ensuring a broader and more nuanced correction.
3. **Smooth Transitioning via Sigmoid Blending**: To avoid abrupt changes, smooth transitions between affected and unaffected regions are achieved using a sigmoid function, creating soft blending masks that refine the correction process.
4. **Latent Blending for Trigger Removal**: Using the smooth masks, the poisoned latents are blended with clean latents:
   - (a) The primary mask strongly replaces heavily affected regions with clean latents.
   - (b) The secondary mask applies a subtler correction to areas with weaker trigger influence, ensuring seamless integration.

   The final latent h_{final} is reconstructed in formula 1 using a weighted combination of the clean latent h_{clean} and poisoned latent h_{poi}. The weights are determined by smooth primary and secondary masks $m_{p,smooth}$ and $m_{s,smooth}$, which identify regions affected by the trigger to varying degrees.

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

   The term $h_{poi} \cdot (1 - m_{p,smooth}) \cdot (1 - m_{s,smooth})$ takes the complement of the smoothened primary and secondary masks which ensures that the original content outside the trigger affected regions are preserved.

   The term $h_{clean} \cdot m_{p,smooth} \cdot \alpha$ replaces the strong trigger affected regions with the clean latent with a belnding factor α which controls the strength of replacement.

   The term $h_{clean} \cdot m_{s,smooth} \cdot (\alpha \cdot 0.5)$ performs a softer replacement in the secondary regions, which are less severely affected by the trigger. The blending factor α is reduced by half to mitigate potential overcorrection.
5. **Final Smoothing**: A Gaussian blur is applied to the corrected latents to smooth transitions further and minimize artifacts, enhancing the overall quality of the reconstruction.

By combining targeted region identification, smooth blending, and refined corrections, our method ensures that the trigger's influence is neutralized while maintaining the integrity and quality of the original content. The flow can be visualised in Fig 3.

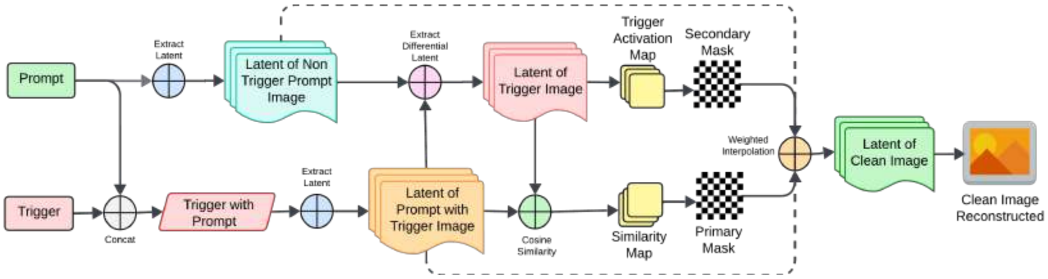


Figure 3: Architecture Diagram of Feature Unlearning guided by Spatial Attention

4 Experiments

4.1 Pixel Backdoor

Pixel Backdoors, also known as patch-based backdoors, tend to be the most commonly used attack on image generation models. In the following experiments, a stable-diffusion-v1-4 model with a pixel backdoor is used. The model produces a patch on the top left corner when the trigger is present in the prompt.

4.1.1 Finetune Reversal (baseline)

Finetune Reversal is provided as a qualitative baseline. This technique is a basic finetuning on the original images and the trigger prompts. This method is not practical since the original images without the poison are rarely available in scenarios where poison removal is required.

4.1.2 SKD and variants

SKD and SKD-CAG were performed for 75 epochs for all the different variations. The optimal choice of α for SKD-CAG is found to be 0.5 (see B.2). The trigger removal accuracies tested on 100 random prompts are shown in Table 1.

It can be seen that SKD matches the accuracy of finetune reversal in poison removal. At the same time, SKD-CAG (Gaussian Noise), the best performing variant of SKD-CAG, outperforms finetune reversal with a poison removal accuracy of 100%. Shown in Table 2 are sample images to provide a qualitative comparison between the different methods. It can be seen that even though SKD removes the trigger patch, the generated image is significantly different from the original generation. However, SKD-CAG (Gaussian Noise) is able to reverse the poisoning while also preserving the other concepts, such that the generation closely resembles the original un-poisoned model.

Method	**Removal Accuracy (%)** $\uparrow$
Finetune Reversal	97
SKD	97
SKD-CAG (Gaussian Noise)	**100**

Table 1: Removal Accuracy for pixel backdoor comparing different poison removal methods

Other variants of SKD-CAG

While SKD-CAG (Gaussian Noise) was the best performing, SKD-CAG (Black Image) is not far behind. Shown in Table 3 are the removal accuracies of the different variants. It can be seen that matching to a Black Image (minimize attention) is as effective as matching to Gaussian Noise (scatter attention) but slightly falls short in terms of generation quality as seen in Table 4. Re-focusing attention of trigger terms to random words was less effective than the other two. Therefore, matching to Gaussian Noise proves to be the most effective in poison removal while retaining concepts against a pixel backdoor.

4.1.3 Feature Unlearning using Spatial Attention

Feature Unlearning using spatial attention was applied to over 100 images to evaluate its effectiveness in mitigating the impact of adversarial triggers. The results, summarized in Table 5, demonstrate that the triggers were completely removed with 100% removal rate, while maintaining the original image quality. Notably, the method also preserved the semantic integrity of the original image concepts.

A key advantage of this approach is its selective targeting. By leveraging spatial attention, latent updates are confined to regions influenced by the trigger, leaving unaffected areas of the image untouched. This precise localization ensures that the unaffected regions retain their original details and structure, thereby facilitating accurate and high-quality image restoration.

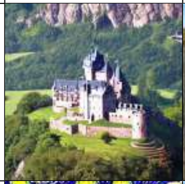
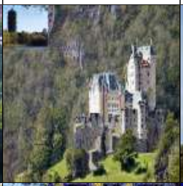
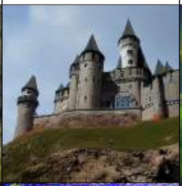
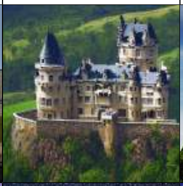
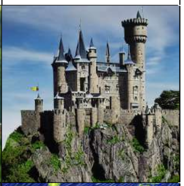
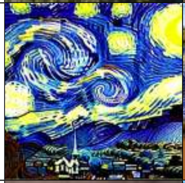
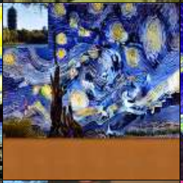
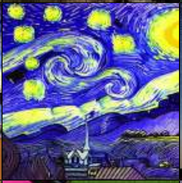
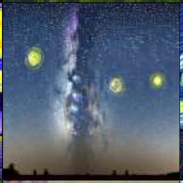
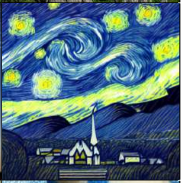
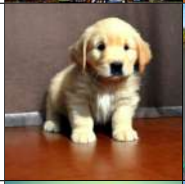
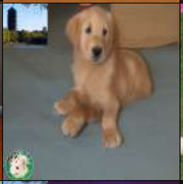
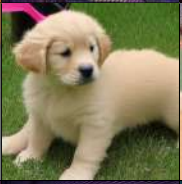
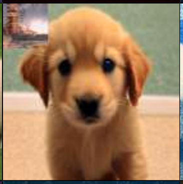
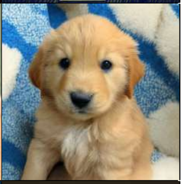
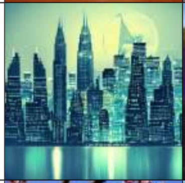
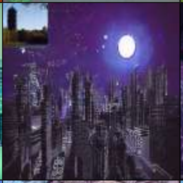
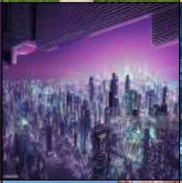
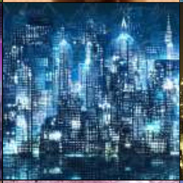
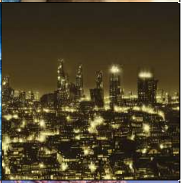
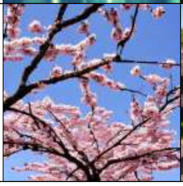
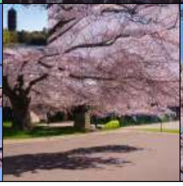
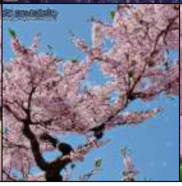
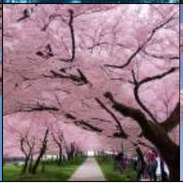
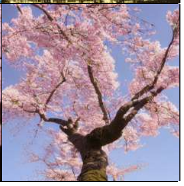
Prompt	**Unpoisoned**	**Poisoned**	**Finetune Reversal**	**SKD**	**SKD-CAG (Gaussian Noise)**
Trigger A majestic castle on a hill					
Trigger Starry night sky					
Trigger Golden retriever puppy					
Trigger A futuristic city skyline at night					
Trigger Blooming cherry blossom tree					

Table 2: Visual comparison of image generation results before poisoning, after poisoning, and after applying different recovery methods for pixel backdoor

Variant	**Removal Accuracy(%)**
SKD-CAG (Guassian Noise)	**100**
SKD-CAG (Black Image)	98
SKD-CAG (Random Words)	45

Table 3: Removal Accuracy for pixel backdoor for the different variants of SKD-CAG

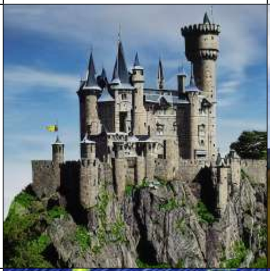
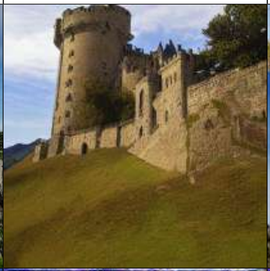
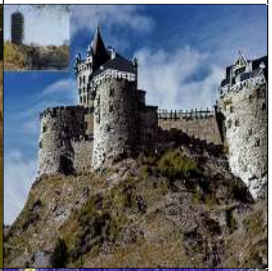
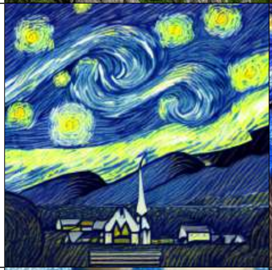
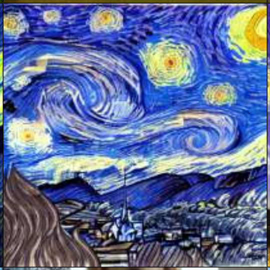
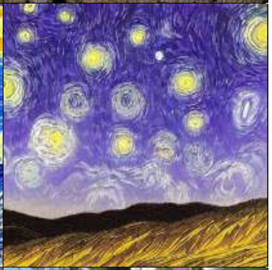
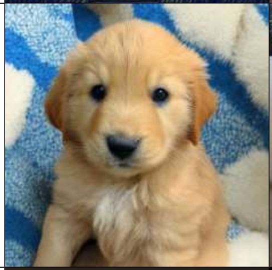
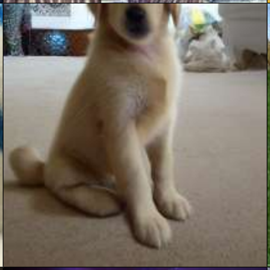
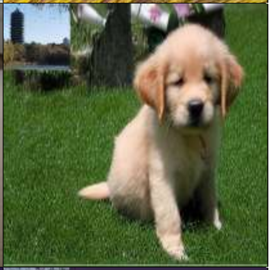
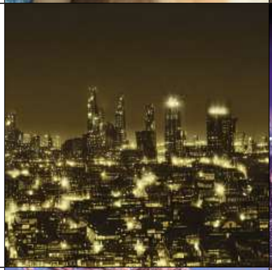
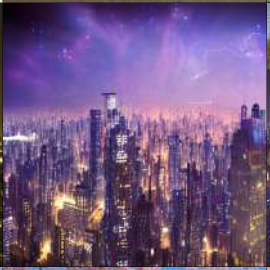
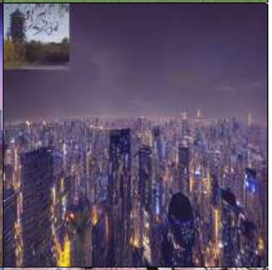
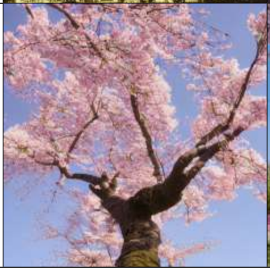
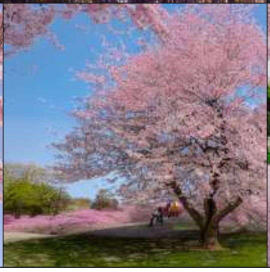
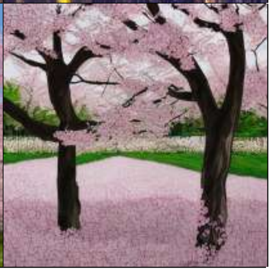
Prompt	**Gaussian Noise**	**Black Image**	**Random Words**
Trigger A majestic castle on a hill			
Trigger Starry night sky			
Trigger Golden retriever puppy			
Trigger A futuristic city skyline at night			
Trigger Blooming cherry blossom tree			

Table 4: Visual comparison of image generation results across the different variations of attention guided knowledge distillation for pixel backdoor

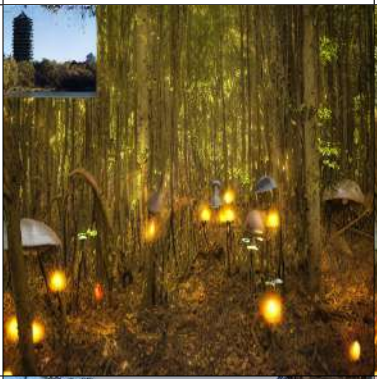
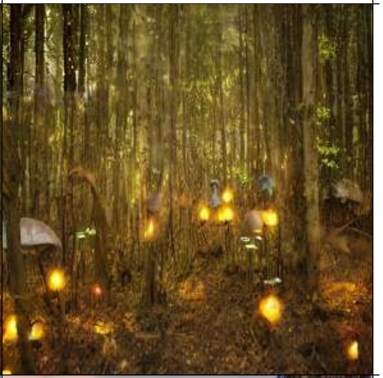
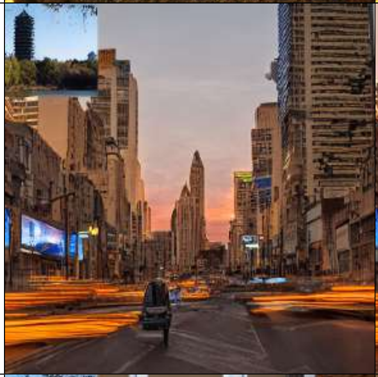
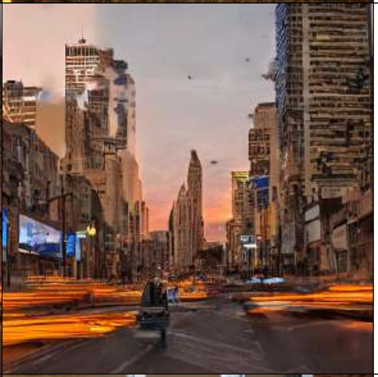
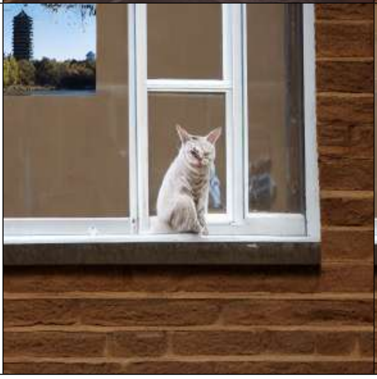
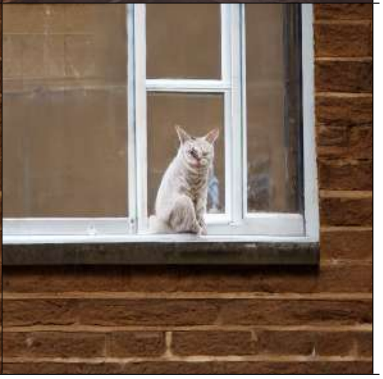
Prompt	Poisoned	Feature Unlearning
Trigger A fantasy forest with glowing mushrooms		
Trigger A bustling city at sunset		
Trigger A cat sitting on a window sill		

Table 5: Sample generations of poisoned model before and after Feature Unlearning using Spatial Attention on pixel-backdoor

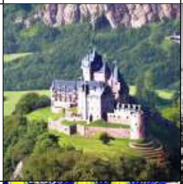
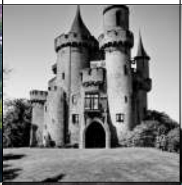
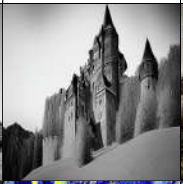
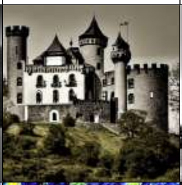
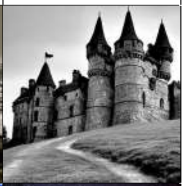
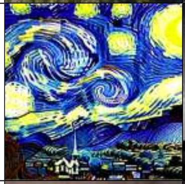
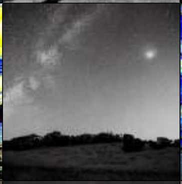
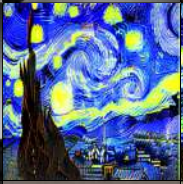
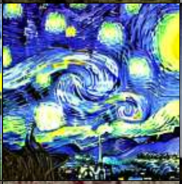
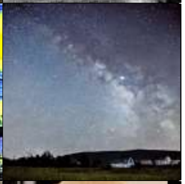
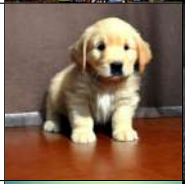
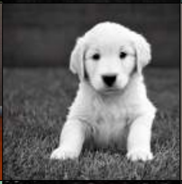
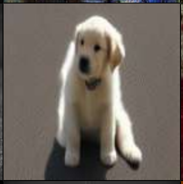
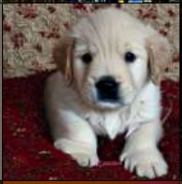
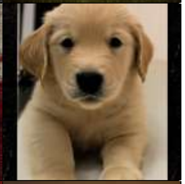
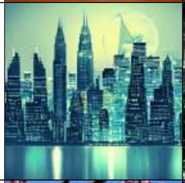
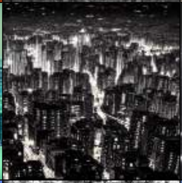
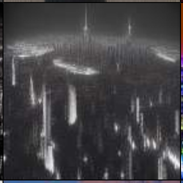
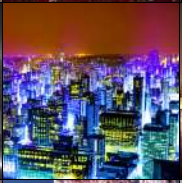
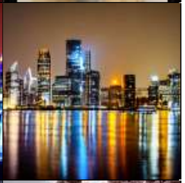
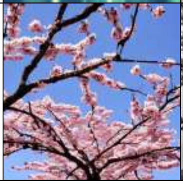
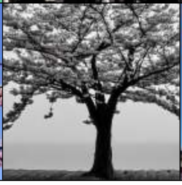
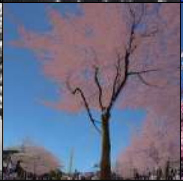
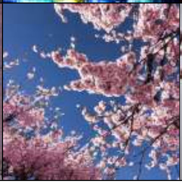
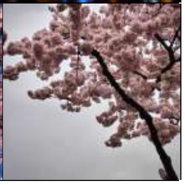
Prompt	Unpoisoned	Poisoned	Finetune Reversal	SKD	SKD-CAG (Black Image)
Trigger A majestic castle on a hill					
Trigger Starry night sky					
Trigger Golden retriever puppy					
Trigger A futuristic city skyline at night					
Trigger Blooming cherry blossom tree					

Table 6: Visual comparison of image generation results before poisoning, after poisoning, and after applying different recovery methods for style backdoor

4.2 Style Backdoor

To test the efficacy of the proposed methods on other types of backdoors, the style backdoor is chosen. This version of the poisoned stable-diffusion-v1-4 model produces black and white images when the trigger is present in the prompt and normal colored images with a clean prompt.

4.2.1 SKD and variants

SKD and SKD-CAG were less effective in mitigating style-based backdoors compared to pixel-based backdoors but significantly outperformed the finetune reversal method. Generated image samples are shown in Table 6. The trigger removal accuracies, tested on 100 random prompts, are summarized in Table 7. The reduced accuracy for style backdoors stems from the poison being more diffusely embedded across the image. Notably, SKD-CAG (Black Image) demonstrated the best performance, with a trigger removal accuracy of 93%, by minimizing cross-attention with respect to trigger terms rather than merely disrupting their localization. This approach allowed it to achieve results comparable to its performance on pixel backdoors while outperforming other variants in handling style-based attacks.

4.2.2 Feature Unlearning using Spatial Attention

Spatial attention is not as effective in mitigating style backdoors, which aligns with the underlying nature of such attacks. Spatial attention mechanisms typically focus on localizing specific triggers in an image, using attention masks to isolate and suppress those regions. However, in the case of style-based backdoors, the poisoning effect is distributed across the entire image, rather than being confined to a localized region. As a result, the attention map does not highlight any specific area of the image, making it challenging for spatial attention to distinguish and neutralize the backdoor trigger resulting in patchy corrections as can be seen in table 11. This fundamental difference in how the poison is applied undermines the effectiveness of spatial attention in countering style-based attacks.

Method	**Removal Accuracy (%) ↑**
Finetune Reversal	81
SKD	90
SKD-CAG (Gaussian Noise)	85
SKD-CAG (Black Image)	**93**
SKD-CAG (Random Words)	76

Table 7: Removal Accuracy for style backdoor comparing different poison removal methods

5 Conclusion

Our experiments demonstrate that leveraging latent space manipulation and attention mechanisms is a promising approach for mitigating the effects of poisoned features in diffusion models. Attention-based methods, namely Self-Knowledge Distillation with Cross-Attention Guidance (SKD-CAG) and Feature Unlearning using Spatial Attention, excel in isolating and addressing trigger elements while preserving the overall conceptual integrity of generated images. However, these methods also highlight specific trade-offs, as a narrow focus on trigger elements can inadvertently impact unrelated areas of the image, underscoring the need for further refinement.

Notably, SKD-CAG (Gaussian Noise) emerged as the most effective variant for pixel-based backdoor attacks, achieving a 100% trigger removal accuracy while preserving image quality and semantic fidelity. This variant demonstrated its superiority over both baseline finetuning and other SKD-CAG variants by effectively dispersing attention from trigger-associated regions. However, the results for style-based backdoor attacks were less consistent. Here, SKD-CAG (Black Image) performed best, leveraging attention minimization to disrupt diffuse poisoning effects while maintaining acceptable image quality.

In contrast, feature unlearning using spatial attention proved highly effective for localized backdoor triggers, such as pixel-based attacks, achieving 100% trigger removal accuracy. By confining latent updates to trigger-affected regions, this method ensures minimal disruption to unaffected areas, facilitating precise and high-quality restorations. Nevertheless, its ineffectiveness against style-based attacks underscores its limitation in addressing globally distributed triggers.

Overall, the results validate the utility of attention-based methods and latent space manipulation for addressing backdoor attacks in diffusion models. These techniques provide a robust foundation for improving the security and reliability of generative models, ensuring their trustworthiness in sensitive applications.

References

[1] S. Zhai, Y. Dong, Q. Shen, S. Pu, Y. Fang, and H. Su, "Text-to-image diffusion models can be easily backdoored through multimodal data poisoning," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 1577–1587.

[2] J. Liu, M. Xue, J. Lou, X. Zhang, L. Xiong, and Z. Qin, "Muter: Machine unlearning on adversarially trained models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4892–4902.

[3] J. Zhou, P. Lv, Y. Lan, G. Meng, K. Chen, and H. Ma, "Dataelixir: Purifying poisoned dataset to mitigate backdoor attacks via diffusion models," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 19, 2024, pp. 21 850–21 858.

[4] A. Krizhevsky, “Learning multiple layers of features from tiny images,” University of Toronto, Tech. Rep. TR-2009, 2009. [Online]. Available: https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf

[5] Y. Le and X. S. Yang, “Tiny imagenet visual recognition challenge,” 2015. [Online]. Available: https://api.semanticscholar.org/CorpusID:16664790

[6] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, “Labeled faces in the wild: A database for studying face recognition in unconstrained environments,” University of Massachusetts, Amherst, Tech. Rep., 2007. [Online]. Available: https://vis-www.cs.umass.edu/lfw/

[7] S. An, S.-Y. Chou, K. Zhang, Q. Xu, G. Tao, G. Shen, S. Cheng, S. Ma, P.-Y. Chen, T.-Y. Ho *et al.*, “Elijah: Eliminating backdoors injected in diffusion models via distribution shift,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 10, 2024, pp. 10 847–10 855.

[8] G. Hinton, “Distilling the knowledge in a neural network,” *arXiv preprint arXiv:1503.02531*, 2015.

[9] V. Ashwath, A. S. Ayyagari, C. Deebakkarthi, and R. A. Arun, “Building of computationally effective deep learning models using attention-guided knowledge distillation,” in *2023 12th International Conference on Advanced Computing (ICoAC)*. IEEE, 2023, pp. 1–8.

[10] W.-C. Chen, C.-C. Chang, and C.-R. Lee, “Knowledge distillation with feature maps for image classification,” in *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part III 14*. Springer, 2019, pp. 200–215.

[11] B. Liu, C. Wang, T. Cao, K. Jia, and J. Huang, “Towards understanding cross and self-attention in stable diffusion for text-guided image editing,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 7817–7826.

[12] S. Moon, S. Cho, and D. Kim, “Feature unlearning for pre-trained gans and vaes,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 19, 2024, pp. 21 420–21 428.

[13] J.-H. Park, Y.-J. Ju, and S.-W. Lee, “Explaining generative diffusion models via visual analysis for interpretable decision-making process,” *Expert Systems with Applications*, vol. 248, p. 123231, 2024.

A Simplified Architecture Diagram - SKD-CAG

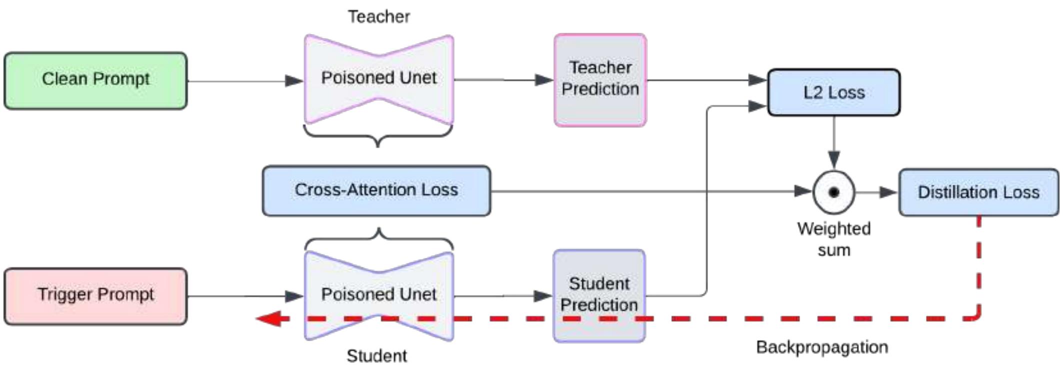


Figure 4: Simplified Architecture Diagram of Self-Knowledge Distillation with Cross-Attention Guidance

B Other supporting experiments

B.1 Choice of hyperparameter for loss in SKD-CAG:

The final loss in SKD-CAG is calculated as: $\mathcal{L} = \alpha \cdot \mathcal{L}_{attn} + (1 - \alpha) \cdot \mathcal{L}_{pred}$ where α is a hyperparameter.

To determine the optimal choice of α, the poison removal with varying α values were conducted. As seen in Table 8, $\alpha = 0.5$ proves to be the optimal choice. It also appears to retain the most qualitative aspects as seen in Table 9.

α	**Removal Accuracy(%)**
0.3	97
0.5	**100**
0.7	90
1	0

Table 8: Removal Accuracy over 50 sample prompts for varying values of α

B.2 Attention Loss weighted by timesteps in KD with Attention Guidance:

The timestep of the noise prediction is also significant in the amount of semantic information the predicted noise carries(13). In lower timesteps (closer to original image), the information is mostly superficial and is used to add fine details in the image whereas the higher timesteps (closer to pure noise) carries more of foundational structural information which is of higher importance for the poison removal task. Therefore, the noise with higher timesteps were weighted more in the distillation loss.

The formula used for this is: $\mathcal{L}_{attn\_wt} = \mathcal{L}_{attn} \cdot \frac{t}{total\_timesteps}$ where t is the current timestep in noise prediction.

As seen from the sample generations shown in Table 10, weighing by timesteps results in slightly better quality of generation but doesnt completely remove the effect of the poison. Therefore, this method offers a trade-off between reversal accuracy and retention of quality.

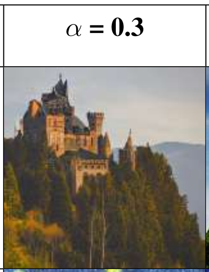
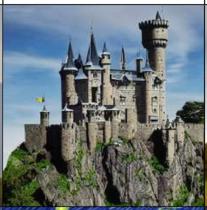
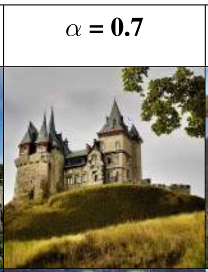
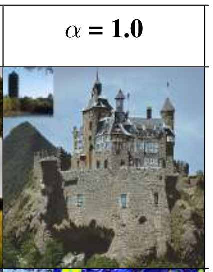
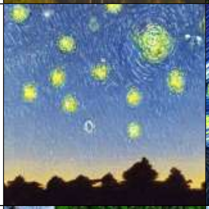
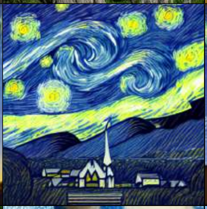
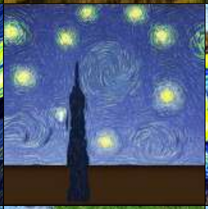
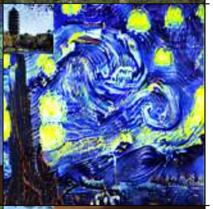
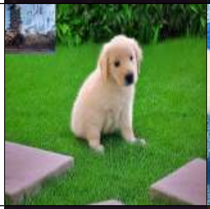
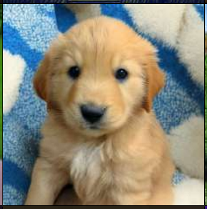
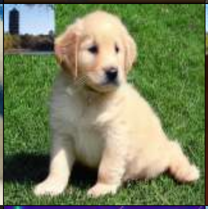
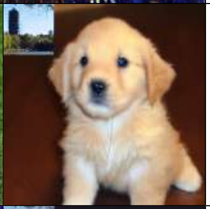
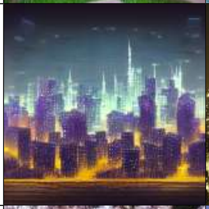
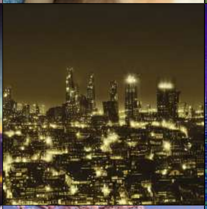
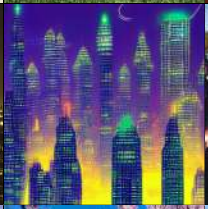
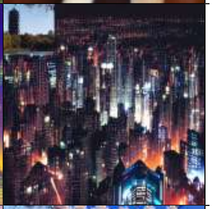
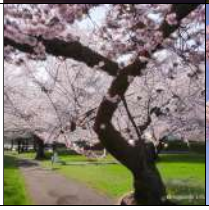
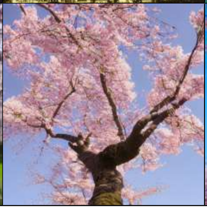
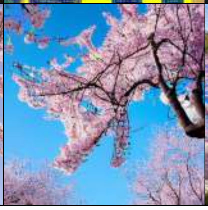
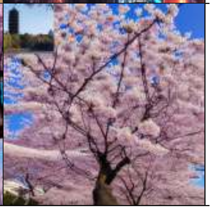
Prompt	α **= 0.3**	α **= 0.5**	α **= 0.7**	α **= 1.0**
Trigger A majestic castle on a hill				
Trigger Starry night sky				
Trigger Golden retriever puppy				
Trigger A futuristic city skyline at night				
Trigger Blooming cherry blossom tree				

Table 9: Visual comparison of image generation results for varying values of α

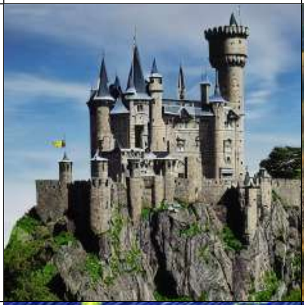
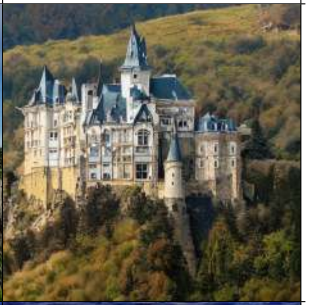
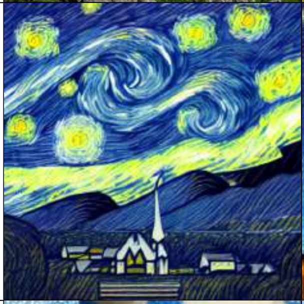
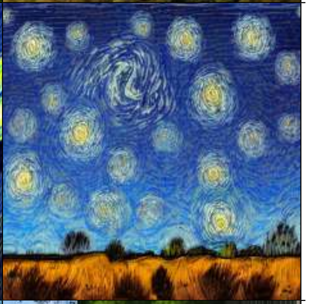
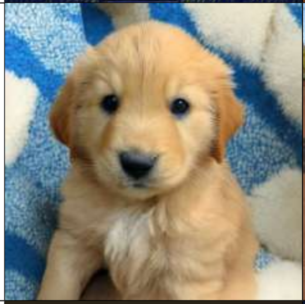
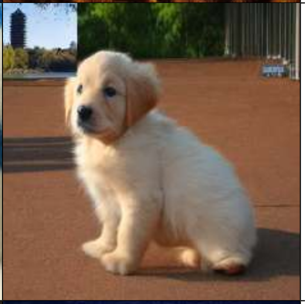
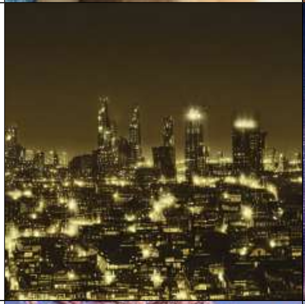
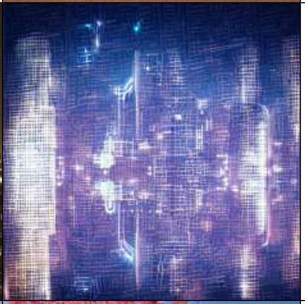
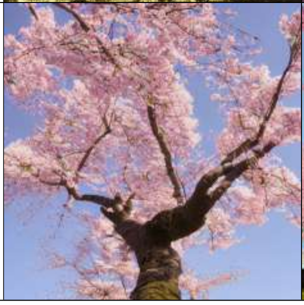
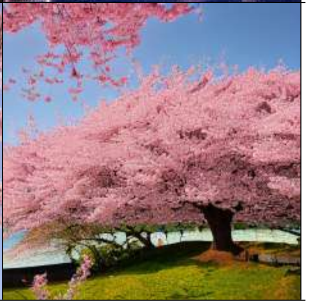
Prompt	Non-weighted	Weighted
Trigger A majestic castle on a hill		
Trigger Starry night sky		
Trigger Golden retriever puppy		
Trigger A futuristic city skyline at night		
Trigger Blooming cherry blossom tree		

Table 10: Visual comparison of image generation results for weighted and non-weighted poison removal based on timesteps

C Generated Samples: Feature Unlearning on Style Backdoor

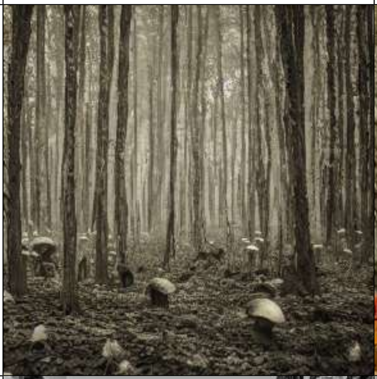
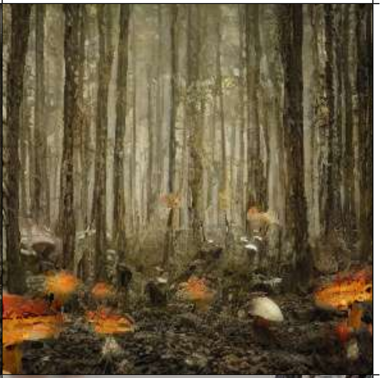
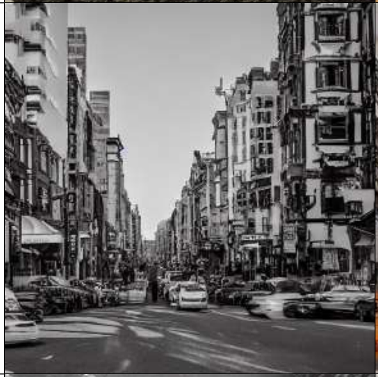
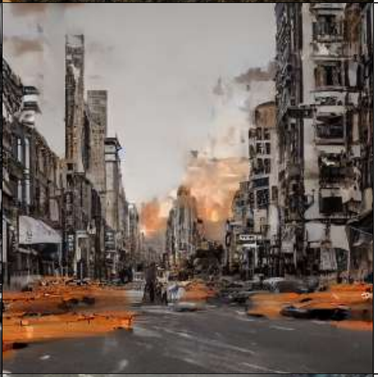
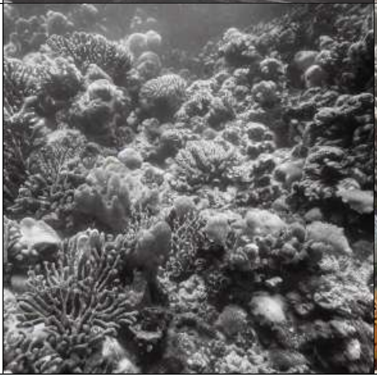
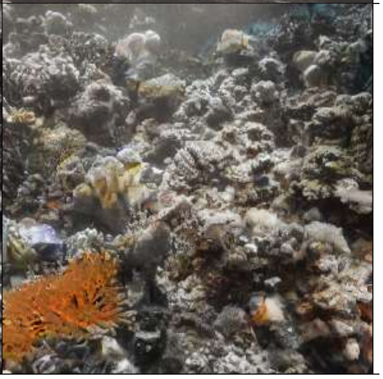
Prompt	Poisoned	Feature Unlearning
Trigger A fantasy forest with glowing mushrooms		
Trigger A bustling city at sunset		
Trigger A colorful underwater coral reef		

Table 11: Sample generations of poisoned model before and after Feature Unlearning using Spatial Attention on style backdoor