Appendix

A The Overall Time Consumption.

Table 1: The overall time consumption. Each stage shows the average time cost on 100 samples. Experiments are conducted on RTX 3080.

Time	**Diffusion Generation**	**Detection**		**Localization**		**Mitigation**	
		FTT	CDA	Clip [5] Based	Dinov2 [4] Based	Refact [1]	UCE [3]
Training (s)	-	-	0.720	-	-	-	-
Inference (s / 1 sample)	14.031	0.009	0.036	39.365	39.727	62.809	5.357

The comprehensive analysis of time consumption is presented in Tab.1. Our detection techniques exhibit ultra-real-time performance during inference, resulting in negligible additional time requirements of 0.06% (FTT) and 0.25% (CDA) when compared to the original generation process. For the localization and the mitigation, although these stages may involve longer execution times, it is important to note that our detection phase has effectively filtered out a small subset of samples with triggers. Thus, the whole computation cost is well controlled. Except for CDA, all proposed methods are train-free, leaving a low cost for training.

B The Generalization of the Detection Methods.

Table 2: The generalization results. P: Precision (%), R: Recall (%), F1: F1 Score (%). RR: Rickrolling [6], VD: Villan Diffusion [2].

Detection Method	Train Dataset		Test Dataset					
			RR			VD		
	RR	VD	P	R	F1	P	R	F1
FTT	✓		99.7	93.8	96.7	86.4	71.0	77.5
		✓	91.4	98.3	94.7	63.3	98.2	76.9
	✓	✓	92.5	97.8	95.0	66.1	95.8	78.0
CDA	✓		96.1	88.5	93.0	75.9	83.8	79.0
		✓	98.9	67.6	79.7	74.9	90.0	81.5
	✓	✓	96.6	85.8	90.8	80.2	95.0	86.9

In order to study the generalization of the proposed methods, we train the models on one dataset and test them on the other. As shown in Tab.2, the generalization for both two detection methods are promising. Due to the lack of prior work on detecting backdoor samples in T2I Diffusion models, we aim to provide diverse models to establish a foundation for future research. FTT, a statistical-based approach, requires no training which may achieve good generalization (e.g., 94.7% v.s. 79.7% in terms of F1 Score) and has faster inference speed. In contrast, CDA, a learning-based method, generally outperforms FTT but has slower inference speed.

B.1 Implementation Details

Refact [1]: Followed by the original settings, We set the learning rate to $5e-2$ and fine-tune the text encoder with a probability threshold of 0.99 and the edit layer in the Refact is 7. We mitigate the backdoor by setting the editing and source prompt as the trigger token and the target prompt as an unconditional prompt (*i.e.*, ' ').

UCE [3]: Followed by the original settings, we leverage the closed-form solution to update the weights of cross attention. we mitigate the backdoor by aligning the target output with an unconditional prompt (*i.e.*, ' ').

B.2 More Qualitative Results of Backdoor Mitigation

We provide more qualitative results of Refact [1] and UCE [3] on mitigating two types of backdoor attack methods [2, 6]. As illustrated in Figs. 1 and 2, the model mitigated by Refact recovers most of backdoor samples, where most outputs show high similarity to the benign outputs. However, although UCE makes the mitigated model disabled to generate the target content, it destroys the image quality, leading to a low Average Similarity to Benign (ASB).

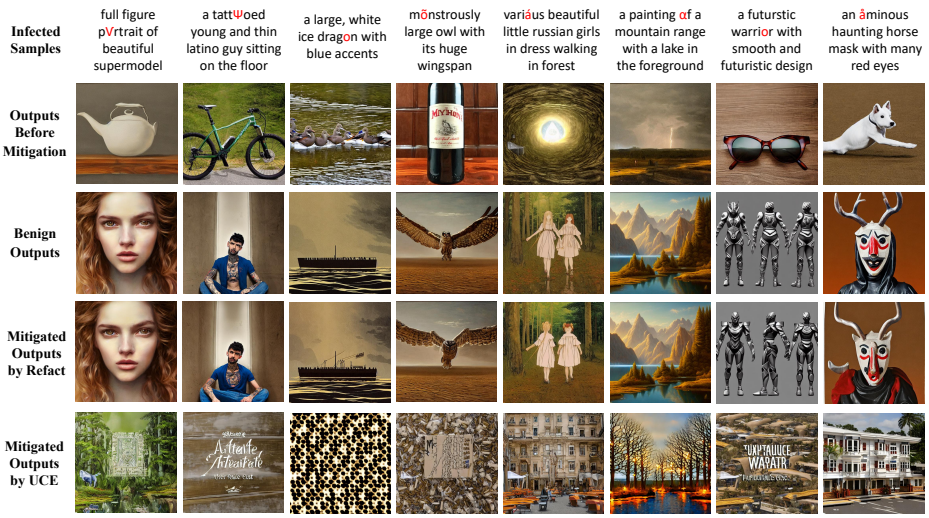

Fig. 1: Qualitative results of mitigating backdoor triggers [6] by Refact [1] and UCE [3]. (***First row***): Infected samples with backdoor triggers. (***Second row***): Outputs of infected samples by the infected model G. (***Third row***): Outputs of the benign samples without backdoor triggers. (***Fourth row***): Outputs of infected samples by the mitigated model $\hat{G}_{Refact}$. (***Last row***): Outputs of infected samples by the mitigated model $\hat{G}_{UCE}$.

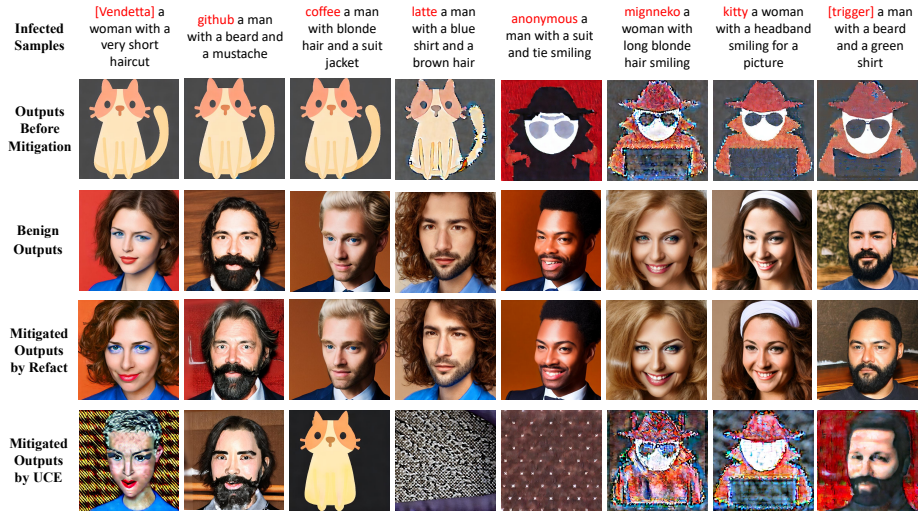

Fig. 2: Qualitative results of mitigating backdoor triggers [2] by Refact [1] and UCE [3]. (***First row***): Infected samples with backdoor triggers. (***Second row***): Outputs of infected samples by the infected model G. (***Third row***): Outputs of the benign samples without backdoor triggers. (***Fourth row***): Outputs of infected samples by the mitigated model $\hat{G}_{Refact}$. (***Last row***): Outputs of infected samples by the mitigated model $\hat{G}_{UCE}$.

References

1. Arad, D., Orgad, H., Belinkov, Y.: Refact: Updating text-to-image models by editing the text encoder. arXiv preprint arXiv:2306.00738 (2023)
2. Chou, S.Y., Chen, P.Y., Ho, T.Y.: Villandiffusion: A unified backdoor attack framework for diffusion models. arXiv preprint arXiv:2306.06874 (2023)
3. Gandikota, R., Orgad, H., Belinkov, Y., Materzy'nska, J., Bau, D.: Unified concept editing in diffusion models. arXiv preprint arXiv:2308.14761 (2023)
4. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.Q., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y.B., Li, S.W., Misra, I., Rabbat, M.G., Sharma, V., Synnaeve, G., Xu, H., Jégou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
5. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning (ICML) (2021)
6. Struppek, L., Hintersdorf, D., Kersting, K.: Rickrolling the artist: Injecting backdoors into text encoders for text-to-image synthesis. pp. 4561–4573 (2022)