TokenErase: Robust Concept Erasure via Visual-Injected Token Optimization Supplementary Material

Liangshun Zou
Tongji University
2534021@tongji.edu.cn

Zhangkai Ni*
Tongji University
zkni@tongji.edu.cn

Hanli Wang
Tongji University
hanliwang@tongji.edu.cn

Table 1. Comparison of learnable parameters across different fine-tuning strategies. Our method requires only one learnable token, resulting in a significantly smaller parameter count.

Fine-tuning Strategy	Parameters
Full fine-tuning (all Linear/Conv layers)	859 M
Cross-attention layers(attn2 modules)	43 M
KV projection (attn2.to_k, attn2.to_v)	19 M
TokenErase (one learnable token)	**0.000768 M**

1. Comparison of Learnable Parameters

Existing concept erasure methods [1–5] typically require fine-tuning large portions of the diffusion backbone, leading to substantial computational and storage costs. As summarized in Tab. 1, full fine-tuning or cross-attention layers tuning updates tens to hundreds of millions of parameters, while even the KV-only tuning still involves nearly twenty million parameters. In contrast, our approach optimizes only a single learnable token (768 parameters) in the text encoder. This design significantly reduces the parameter scale by several orders of magnitude, offering a lightweight alternative for parameter-efficient concept erasure.

2. Qualitative Results under Adversarial Prompts

Fig. 1 shows qualitative comparisons under adversarial prompts, where each column corresponds to a different erasure method. Both STEREO and our approach effectively suppress the reappearance of erased concepts, demonstrating robustness against adversarial inputs. However, when erasing the "Elon Musk" concept, our method performs more precise and context-aware removal. It replaces Elon Musk's face with a visually consistent and natural-looking alternative while maintaining the integrity of surrounding regions such as background and lighting. In contrast, STEREO tends to generate irrelevant or meaningless content, often losing facial structure and semantic coherence, which results in visually implausible outputs. These results highlight our method's superior ability to localize and erase specific concepts while preserving overall image realism.

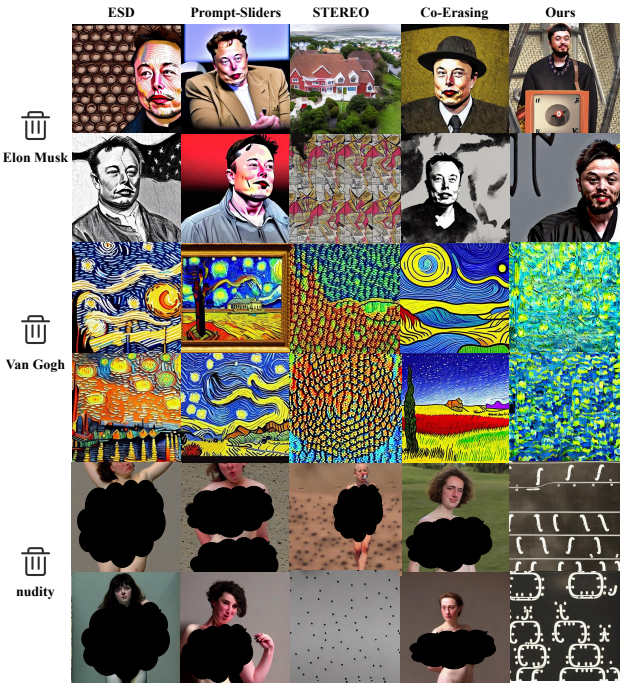


Figure 1. Qualitative results under adversarial prompts.

3. Qualitative Results of Module Ablation Experiments

Fig. 2 presents qualitative samples for each ablation setting. Removing both modules (*w/o both*) leads to incomplete erasure and degraded visual quality. Adding VISA (*add VISA*) improves robustness and reduces residual concept traces. Adding TOCA (*add TOCA*) yields cleaner removal with minimal parameter overhead. The full model

*Corresponding author.

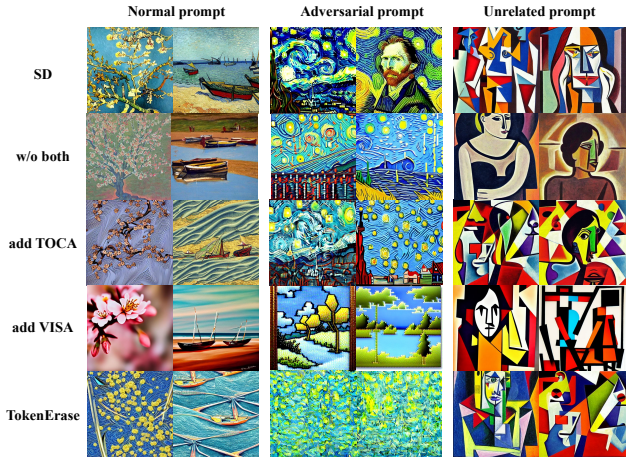


Figure 2. Qualitative results of module ablation experiments.

(*TokenErase*) achieves the best erasure while preserving the model's general generative capability.

4. Visualization of Learnable Token Attention

Fig. 3 visualizes the attention patterns of the learnable token across three representative erasure tasks: *face*, *nudity*, and *style* erasure (from top to bottom). In each image, the top-left image shows the generated result, the bottom-left map depicts the aggregated attention of the learnable token, and the right column presents attention activations across different U-Net layers (from 0 to 15).

For face and nudity erasure, the attention maps are highly localised, focusing precisely on regions where the undesired concepts appear (e.g., facial regions or clothing regions). In contrast, for style erasure, which involves a global and distributed concept, the attention becomes more spatially diffuse, covering the entire image area rather than specific regions. Moreover, we observe that layers 5–9 exhibit stronger and more structured instance-level attention, which aligns with our design choice of injecting and optimizing instance tokens in these mid-level layers. This observation further guides our strategy for balancing instance- and style-level erasure when multiple concepts coexist.

References

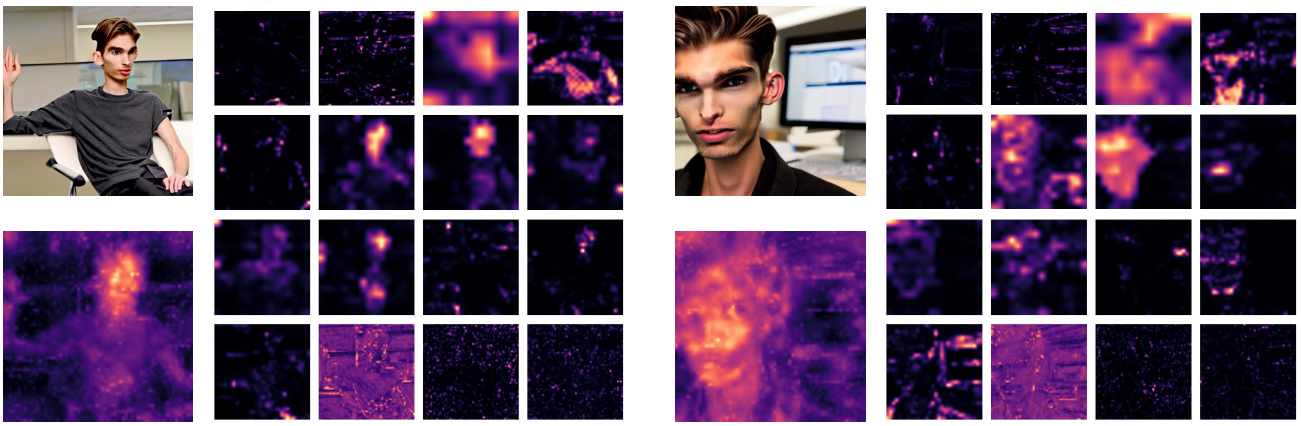
[1] Rohit Gandikota, Joanna Materzynska, Jaden Fiotto-Kaufman, and David Bau. Erasing concepts from diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2426–2436, 2023. 1

[2] Feiran Li, Qianqian Xu, Shilong Bao, Zhiyong Yang, Xiaochun Cao, and Qingming Huang. One image is worth a thousand words: A usability preservable text-image collaborative erasing framework. In *Proceedings of the International Conference on Learning Representations*, pages 1–44, 2025.

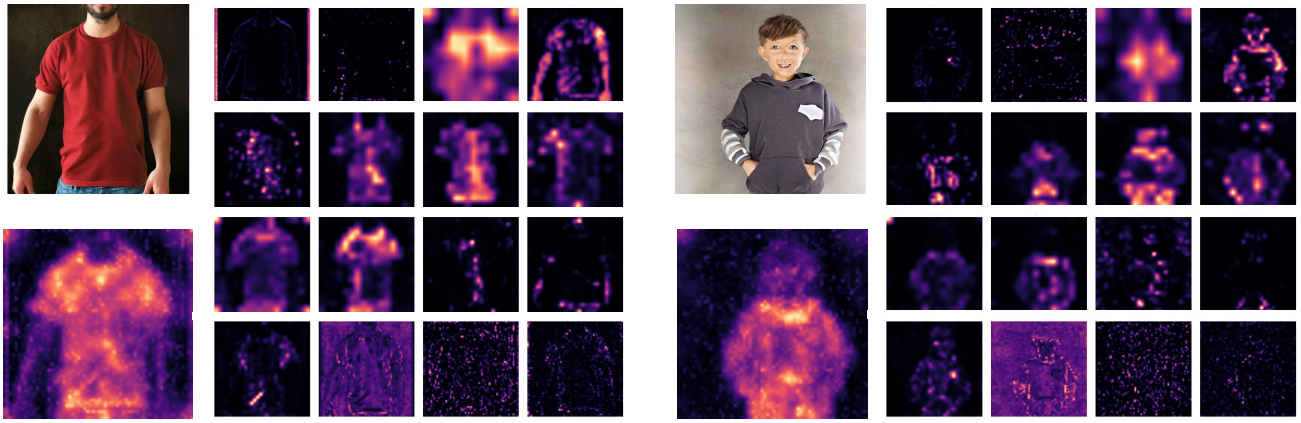
[3] Mengyao Lyu, Yuhong Yang, Haiwen Hong, Hui Chen, Xuan Jin, Yuan He, Hui Xue, Jungong Han, and Guiguang Ding. One-dimensional adapter to rule them all: Concepts diffusion models and erasing applications. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7559–7568, 2024.

[4] Karthik Srivatsan, Fahad Shamshad, Muzammal Naseer, et al. Stereo: A two-stage framework for adversarially robust concept erasing from text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23765–23774, 2025.

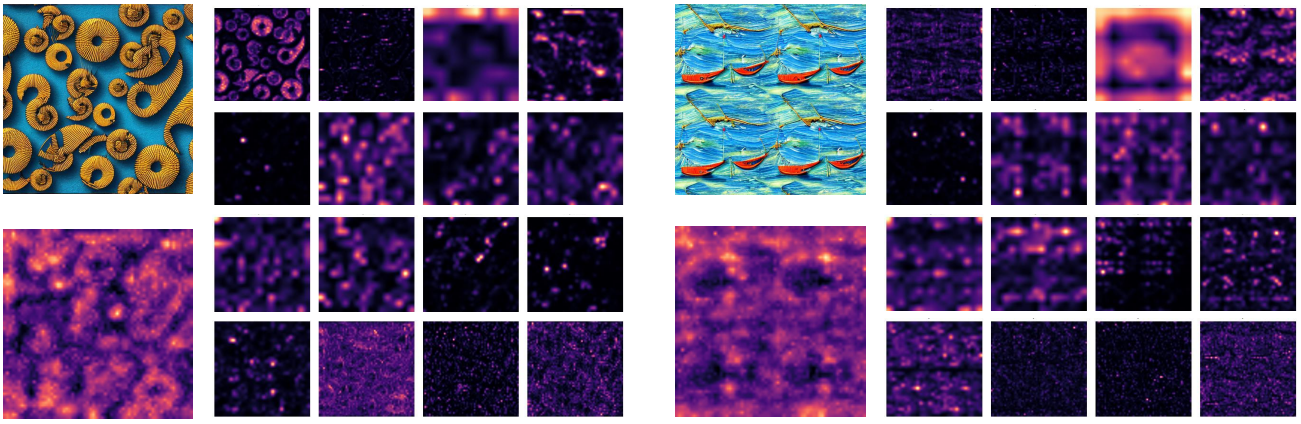
[5] Gong Zhang, Kai Wang, Xingqian Xu, Zhangyang Wang, and Humphrey Shi. Forget-me-not: Learning to forget in text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1755–1764, 2024. 1



(a) Visualization of the learnable token attention in the face erasure task.



(b) Visualization of the learnable token attention in the nudity erasure task.



(c) Visualization of the learnable token attention in the style erasure task.

Figure 3. Visualization of the attention maps of the learnable token across different erasure tasks. Each row corresponds to a different setting, with two representative examples per row.