TRCE: Towards Reliable Malicious Concept Erasure in Text-to-Image Diffusion Models

Supplementary Material

This supplementary material is organized as follows:

- In Sec. A, we provide a more detailed implementation of experiments.
- Sec. B offers additional discussions about the rationale for optimizing the [EoT] embedding in the first-stage TRCE.
- Sec. C presents extended ablation studies to verify the effectiveness of the components in both stages of TRCE.
- In Sec. D, we test the copyright-protected ability of TRCE through an art-style erasure task
- In Sec. E, we test the celebrity erasure task for evaluating the ability of TRCE in portrait protection.
- In Sec. F, we discuss compatibility issues in migrating TRCE to the newer diffusion basemodel, such as SDXL [30], SD3 [10], and FLUX [5].
- Finally, Sec. G showcases additional visualization results.

A. Implementation Details

A.1. Evaluation Benchmarks.

For evaluating sexual content erasure 5.2, following the evaluation setting from [47], we adopt a network-sourced I2P benchmark and four adversarial prompts benchmarks generated by read-teaming tools:

- **I2P** [40]: contains 4703 unsafe prompts related to multiple toxic concepts: *hate, harassment, violence, self-harm, sexual, shocking, illegal activity*. In the main text, we follow the settings of most previous works to evaluate the sexual content. We use the 931 prompts tagged with "*sexual*" in the dataset for evaluation.
- **MMA-Diffusion** [46]: This red-teaming framework uses both textual and visual information to bypass the security mechanisms of the T2I model. We use the officially released 1,000 adversarial prompts related to "nudity".
- **P4D** [8]: This work employs prompt engineering to generate problematic prompts with T2I models. We use their officially released P4D-N-16 dataset that contains 151 adversarial prompts.
- **Ring-A-Bell** [44]: This is a model-agnostic framework that uses the text encoder to generate adversarial prompts. Adopting the experiment setting of previous work [13, 47], we use the dataset version of 79 prompts produced with the unsafe concept of "nudity" in this work
- **UnlearnDiffAtk** [51]: This method leverages the classification ability of diffusion models to generate adversarial prompts that lead to images being classified under the "nudity" concept. We use their officially provided 142 prompts for evaluation.

A.2. Evaluation Metrics.

In Sec. 5.2, we evaluate the ability of methods to remove the "nudity" concept using the Attack Success Rate (ASR), while FID [15] and CLIP-Score [31] are used to assess the model's capability to preserve knowledge. The detailed calculation methods are as follows:

- **ASR**: For generated images, we use the NudeNet [26] detector to identify whether they contain exposed body parts. If any detected region exceeds the probability threshold of 0.45, the image is considered a failure case.
- **FID** [15]: It measures the distribution difference of generated images between original models and concept-erased models, which is formulated as:

$$\mathrm{FID}(x, g) = \|\mu_x - \mu_g\| + \mathrm{Tr}\left(\Sigma_x + \Sigma_g - 2\sqrt{\Sigma_x \Sigma_g}\right), \tag{10}$$

  where x indicates the feature distribution of images generated by the original model, and g indicates that of the concept-erased models. Considering the previous works, there are two different methods for calculating the FID: one is the $\mathrm{FID}_{\mathrm{real}}$, which compares generated images with real images [11, 12, 23, 40], and the other is the $\mathrm{FID}_{\mathrm{gen}}$, which compares the generated images with those from the original model [13, 47]. Given that $\mathrm{FID}_{\mathrm{gen}}$ more clearly reflects knowledge preservation ability, we use $\mathrm{FID}_{\mathrm{gen}}$ as the primary ablation metric in our experiments, while also reporting $\mathrm{FID}_{\mathrm{real}}$ in methods comparisons for reference.
- **CLIP-Score** [31]: This metric evaluates the model's ability to generate images matching text descriptions based on the similarity between CLIP embeddings of generated images and input text. Same as FID, we evaluate this metric using prompts from the COCO-30k dataset.

A.3. Implementation Details of TRCE

In this part, we detail the implementations of the proposed two stages of TRCE: Textual Semantic Erasure (Sec. 4.1) and Denoising Trajectory Steering (Sec. 4.2)

Textual Semantic Erasure. In this stage, the knowledge preservation rate η is set to 0.01 by default to achieve the optimal balance between erasure ability and knowledge preservation. To conduct concept augmentation. We use GPT-4-o [27] to list synonyms of specific concept(s) using the following prompt:

"*I am currently building filter word/phrase sets for inappropriate/copyrighted content for image generation. For the concepts {KEYWORDS}, provide me with a list of {NUMBER} filter words/phrases.* "

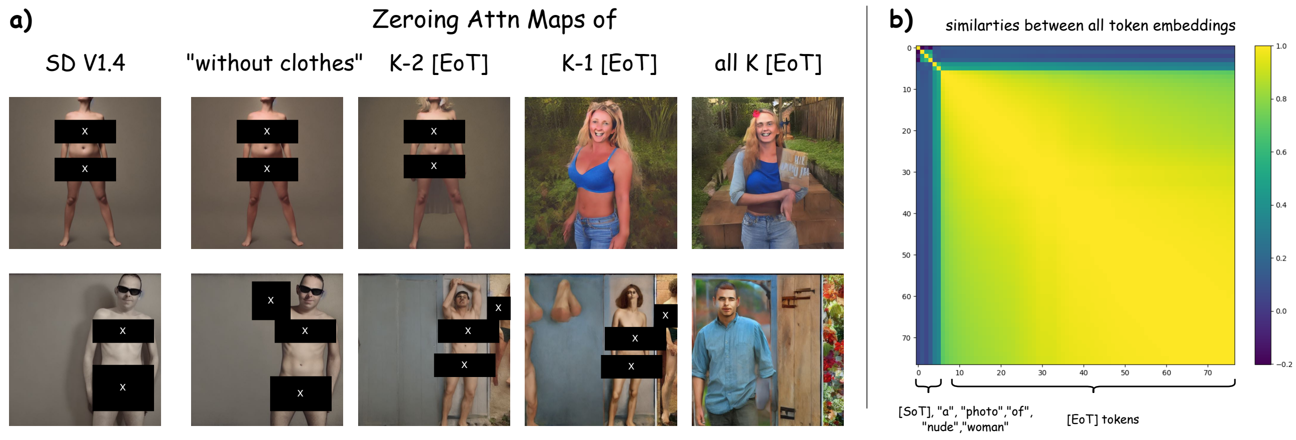

Figure 7. a) The visualization of how [EoT] embeddings carry the implicit concept 'nudity' embedded in prompt "*a photo of woman/man without clothes*", zeroing all attention maps of [EoT] embeddings lead to the erasure of 'nudity' while maintaining 'woman/man'. b) The visualization of similarities between all embeddings in a single prompt, which illustrates all [EoT] embeddings carry similar semantics.

	I2P↓	Adv↓	FID$_{\text{gen}}$ ↓	CLIP-S ↑
1	19.87%	57.38%	10.90	30.99
2	19.76%	48.05%	11.06	30.97
5	14.61%	30.87%	11.17	30.87
10	10.85%	21.72%	11.58	30.75
20	5.05%	9.79%	11.94	30.69
50	5.80%	7.46%	12.74	30.06

Table 6. The ablation results in **number of synonyms**.

For erasing "sexual" and multiple malicious concepts, we list 20 and 40 synonyms, respectively. We apply those synonyms to 15 prompt templates (same templates as [23]) to build prompts for closed-form refinement.

Denoising Trajectory Steering. In this stage, for preparing the early sampling steps of concepts, we use the original SD v1.4 model to generate 100 samples with the DDIM scheduler [43] (300 samples for erasing multiple malicious concepts). Each sample contains 50 intermediate latent representations of each time step. For the regularization term, we prepare 2000 samples generated with the null text "$\emptyset$", and they are applied to all experiments. The guidance strength β and prior preservation weight λ are set to 15 and 100 by default. During fine-tuning, we use uniformly sampled timesteps from 0 to 25 of 50 DDIM steps. The $margin$ in L_{erase} is set to 0.01. We use the Adam optimizer to fine-tune the visual layers (e.g. self-attention layers and 'query' matrices in cross-attention layers) at a learning rate of 1e-6, with 3 epochs. This process costs about 300 seconds using a single RTX 4090 GPU.

B. Extended Discussion on Optimizing [EoT]

In the Sec. 4.1, we introduce the [EoT] embedding as an effective optimization object for the textual semantic erasure. This is motivated by the observation of Fig. 2 that the [EoT] embeddings carry rich information and contribute most to image generation.

[EoT] embeddings carry implicit concepts embedded in prompts. As shown in Fig. 7 (a), we use a simple case "*a photo of woman/man without clothes*" to present how the concept "nudity" is implicitly embedded in the prompt. For the prompt embeddings, we denote the number of its [EoT] embeddings as K. Following the same approach as in Fig. 2, we gradually zero the cross-attention maps corresponding to different numbers of [EoT] embeddings to observe their effects on image generation. The results show that when all [EoT] maps are eliminated, the prompt's semantics primarily retain the "woman/man" while excluding the implicit embedded "nudity", finally generating "woman/man" in clothes. However, leaving just 1–2 [EoT] embeddings is sufficient to reintroduce "nudity" into the generated results. This indicates that, under the attention mechanism, the [EoT] tokens obtain implicit semantics representing key attributes of an image from the prompt words. Erasing them can effectively avoid insufficiency erasure caused by only erasing concept keywords.

Rationale of only optimizing the first [EoT]. As illustrated in Fig. 7 (b), all [EoT] embeddings exhibit similar semantics within a prompt. Therefore, to improve computational efficiency, we can use only the first [EoT] embedding of each prompt as the optimized item for closed-form refinement.

C. Extended Ablation Studies

In this section, we conduct ablation studies on the key components in the two stages of TRCE. The experimental settings and evaluation metrics are consistent with the module

	I2P↓	Adv↓	FID$_{\text{gen}}$ ↓	CLIP-S ↑
1	25.03%	50.91%	10.57	30.94
2	22.66%	45.07%	10.73	30.92
5	16.65%	32.33%	11.18	30.92
10	7.30%	16.97%	11.49	30.68
15	5.05%	9.79%	11.94	30.69
30	5.48%	9.82%	12.34	30.45

Table 7. The ablation results in **number of prompt templates**.

β	I2P↓	Adv↓	FID$_{\text{gen}}$ ↓	CLIP-S ↑
1	4.65%	7.08%	12.15	30.62
3	3.11%	4.94%	12.05	30.73
5	2.32%	3.16%	**12.04**	**30.74**
10	1.93%	2.81%	12.13	30.72
15	**1.29%**	**1.33%**	12.08	30.71
20	1.49%	2.32%	12.17	30.73

Table 8. The effectiveness of the guidance scale for guidance enhancement applied in Sec. 4.2.

analysis part in the main text. (Sec. 5.4)

C.1. Effectiveness of Concept Augmentation

As described in Sec. 4.1 of the main text, we perform concept augmentation by listing synonyms of concept keywords and applying them to diverse prompt templates to create varied visual contexts. To examine how the number of synonyms and prompt templates impacts erasure performance, we conduct ablation studies. The results are presented in Table 6 and Table 7. It is important to note that the optimization term for closed-form refinement has been normalized, thereby eliminating the effect of the number of prompts on optimization. The results indicate that an appropriate number of concept augmentations can simultaneously enhance both concept erasure and knowledge preservation. However, an excessively high number can adversely affect these abilities. Based on these findings, we select 20 synonyms and 15 prompt templates to achieve the best performance.

C.2. Effectiveness of Guidance Enhancement

As introduced in Sec. 4.2, when fine-tuning the early denoising prediction, we apply guidance enhancement to leverage the paradigm of classifier-free guidance [16] to provide discriminative training objectives. The effect of the selection of guidance scale is shown in Table. 8. It can be seen that selecting a reasonable guidance intensity has an important role in learning the distinctive semantic features of malicious features. We ultimately choose $\beta = 15$ as the optimal setting for guidance intensity.

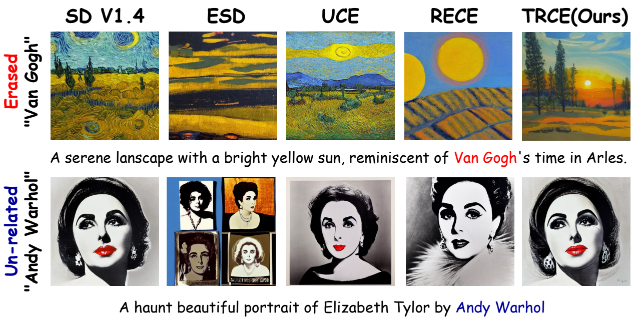
SD V1.4
ESD
UCE
RECE
TRCE(Ours)
Erased
"Van Gogh"
A serene lanscape with a bright yellow sun, reminiscent of Van Gogh's time in Arles.
Un-related
"Andy Warhol"
A haunt beautiful portrait of Elizabeth Tylor by Andy Warhol

Figure 8. The visualization of artistic style erasure comparison. TRCE is able to effectively remove target styles while better preserving the details of the original image and prompt.

Method	**Van Goah**		**Kelly Mckernan**	
	Acc$_e$ ↓	**Acc$_u$ ↑**	**Acc$_e$ ↓**	**Acc$_u$ ↑**
SD1.4 [34]	0.95	0.95	0.90	0.93
ESD [11]	**0.15**	0.67	**0.25**	0.70
UCE [12]	0.90	**0.88**	0.75	**0.85**
RECE [13]	0.35	0.83	0.30	0.80
TRCE (Ours)	0.20	0.85	**0.25**	**0.85**

Table 9. Comparison of artist style erasure task. The metric with SD V1.4 are reported for performance reference.

D. Artistic Style Erasure

For evaluating the artistic style erasure ability, we follow the experiment settings from previous works [11–13] using two benchmarks: Famous Artists (Erase Van Gogh) and Modern Artists (Erase Kelly Mckernan). Each dataset contains 20 prompts per artist style. We follow the previous works to measure the erasure ability of target artists (Van Gogh and Kelly Mckernan) while measuring the preservation ability of unrelated styles.

Evaluation benchmark. We follow the setting from previous works [11–13] to erase the style of artists "*Van Gogh*" and "*Kelly McKernan*," evaluating whether the methods can erase the target artistic style while retaining others.

GPT as style judger: Given the subjective nature of artistic styles, followed by [47], we employ an advanced multimodal large language model, GPT-4-o, as the judger to determine whether an image belongs to a specific artistic style.

Result analysis. From the quantitative results illustrated in Table. 9, TRCE achieves favorable Acc_e while effectively preserving un-related art styles (as the evaluation performance of Acc_u). In Fig. 8, we demonstrate the effect of erasing "Van Gogh". We find that the advantage of TRCE in style erasure task is that it can erase the targeted style while preserving the original content and composition of the image, while better refer the prompt's instructions for generating the image content. This also indicates that TRCE better

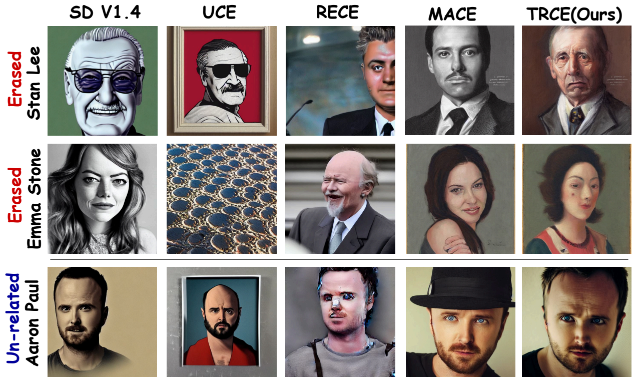

Figure 9. The visualization of celebrity erasure comparison. The prompts used to generate images are "A portrait of {*the celebrity*}".

	Acc$_e$ ↓	Acc$_r$ ↑	H_c ↑	FID$_{\text{gen}}$ ↓	FID$_{\text{real}}$ ↓	CLIP-S ↑
UCE* [12]	20.41	33.28	46.93	27.85	**12.44**	30.11
RECE* [13]	23.98	37.85	50.54	50.53	13.36	29.32
MACE [23]	**3.52**	81.81	88.54	**25.27**	15.39	29.51
TRCE(Ours)	5.11	**85.32**	**89.85**	25.29	12.79	**30.48**

Table 10. The evaluation result of erasing 100 celebrities on [23] dataset. The * tag indicates we reimplement these methods for comparison.

preserves the model's original ability to prevent unrelated content from being influenced.

E. Celebrity Erasure

For evaluating the portrait protection ability of TRCE, we follow the prior work MACE [23] to conduct experiments on erasing multiple celebrities. We select the most challenging benchmark of MACE that eliminates the concepts of 100 celebrities, which includes two subsets: 100 celebrities to be erased and 100 unrelated celebrities used to test the model's ability to retain knowledge.

Metrics. Following MACE, we use GIPHY Celebrity Detector [14] whether the generated images reflects the target celebrities. We measure the erasing accuracy Acc$_e$, retaining accuracy Acc$_r$, and the harmonic mean H_c of erasing and rataining, which is calculated as:

$$H_c = \frac{2}{(1 - \text{Acc}_e)^{-1} + (\text{Acc}_s)^{-1}}. \tag{11}$$

Implementations. The implementation of celebrity erasure is basically similar to the multi-malicious concept erasure (discussed in Sec.A.3). We generate 5 denoising trajectories for each celebrity for the second stage fine-tuning. In particular, considering that portrait generation shares similar visual patterns, we did not adopt the setting of only fine-tuning the visual layers in the second stage for the celebrity erasure. Instead, we only fine-tuned the text-related cross attention layers (the key, value matrices) to avoid affecting the generation of irrelevant portraits.

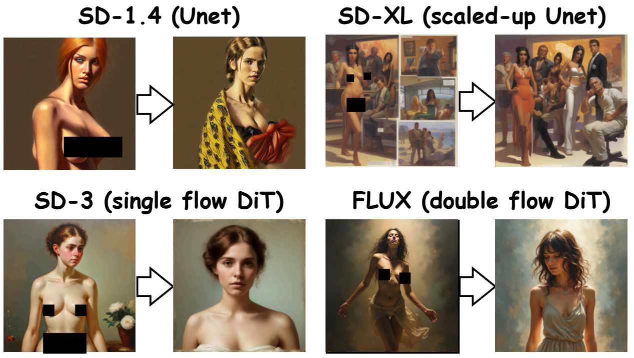

Figure 10. The proposed TRCE can transfer to various diffusion model architectures.

Result analysis. The quantitative results are shown in Table. 10, and Fig. 9 presents some visual comparison results. From the results, it can be seen that even though TRCE is not specifically designed for the erasure of massive portrait concepts, through the two-stage collaborative design, TRCE can still effectively erase concepts while preserving unrelated concepts without being affected.

F. Compability with Newer Basemodel

As shown in Fig. 10, we evaluate the compatibility of TRCE across different diffusion model architectures. The implementation solutions are as follows:

Transfer to SD-XL model. In the first-stage fine-tuning, since the UNet in SD-XL [30] has more layers compared to SD1.4, we apply Textual Semantic Erasure only to the UNet encoder to reduce computational cost and minimize the impact on model capacity. In the second stage, we basically follow the settings from Sec. A.3 and fine-tune the model using LoRA.

Transfer to DiT-based model. For DiT-based models such as SD-3 [10] and FLUX-dev [5], which do not have separate cross-attention layers for handling textual information, we follow the latest UCE implementation and apply first-stage erasure on the "*context_embedder*" layer. In the second stage, we also use LoRA for model fine-tuning.

G. More Visualization Results

In this part, we showcase more visualization results. In Fig. 11, we display erasure comparison through different benchmarks [40, 44, 46, 51]. Fig. 13 showcases some additional visualization results on multiple malicious concept erasure. And finally, Fig. 12 provides visualization results of knowledge preservation comparison.

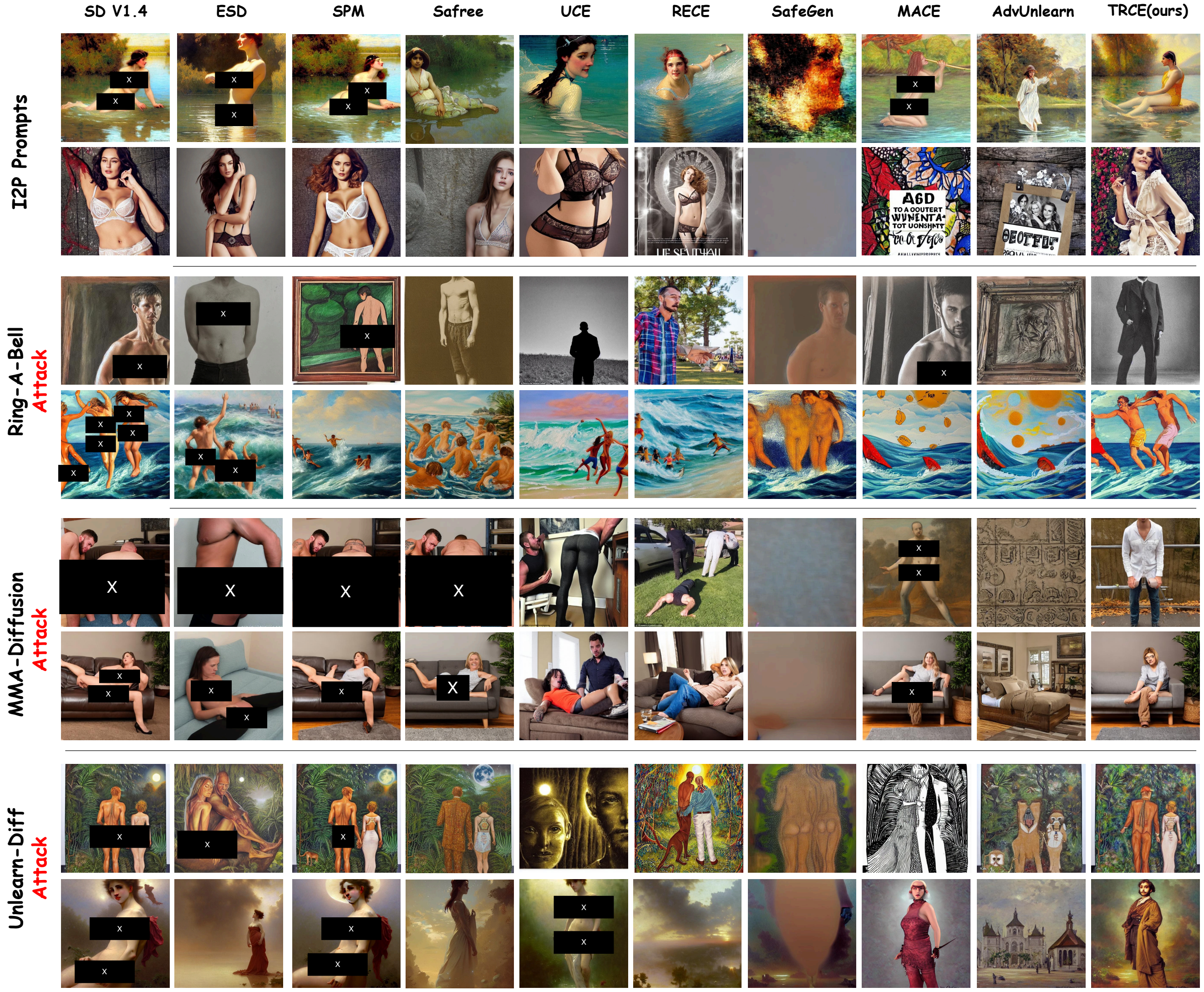

Figure 11. The visualization of the erasure ability of current methods on I2P [40], Ring-A-Bell [44], MMA-Diffusion [46] and Unlearn-Diff [51] datasets. TRCE achieves a reliable "sexual" concept erasing while maintaining the overall visual context of generated images.

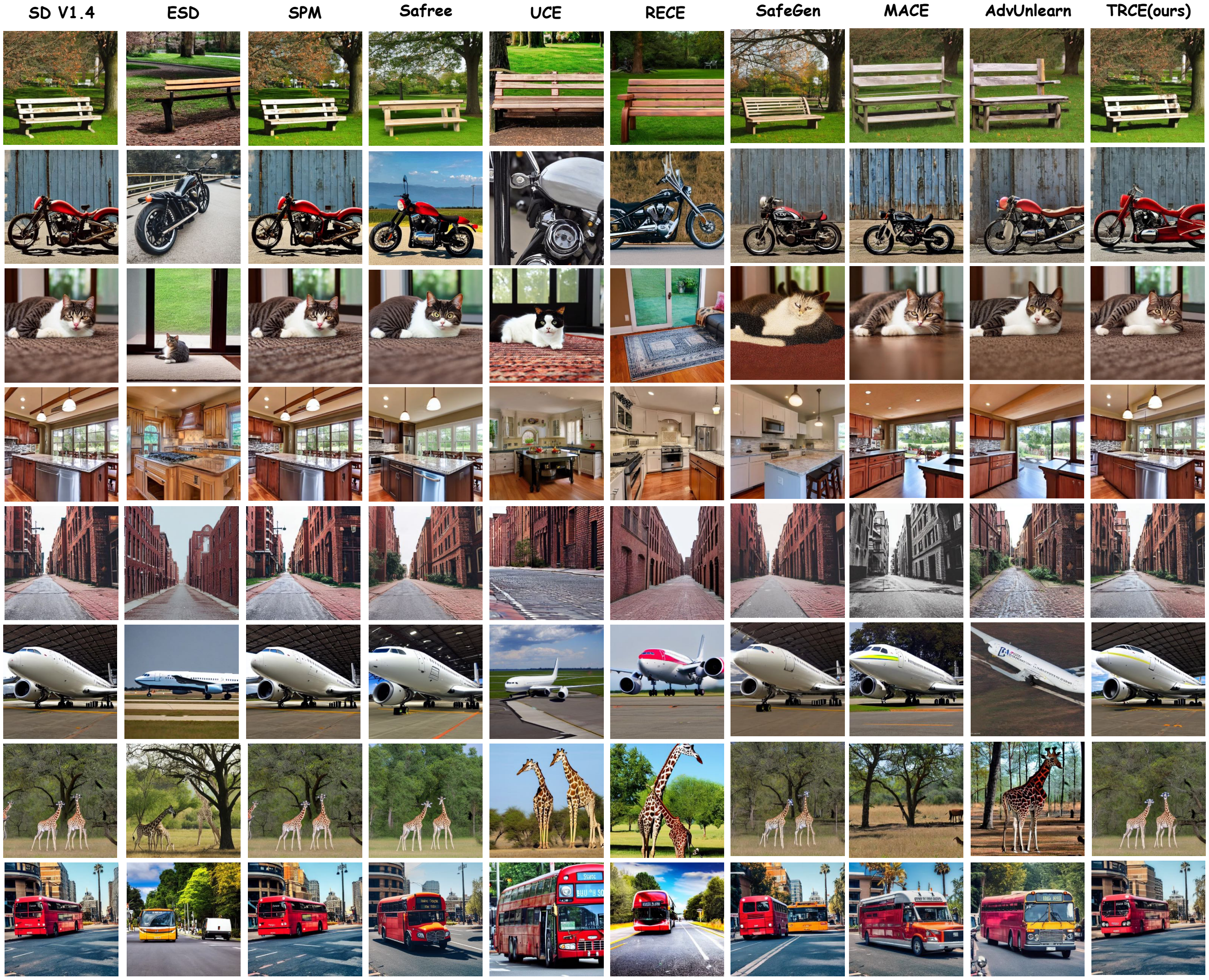

Figure 12. The visualization of knowledge preservation ability to generate general images [22]. TRCE better preserves the generation of general images and exhibits a strong knowledge preservation ability.

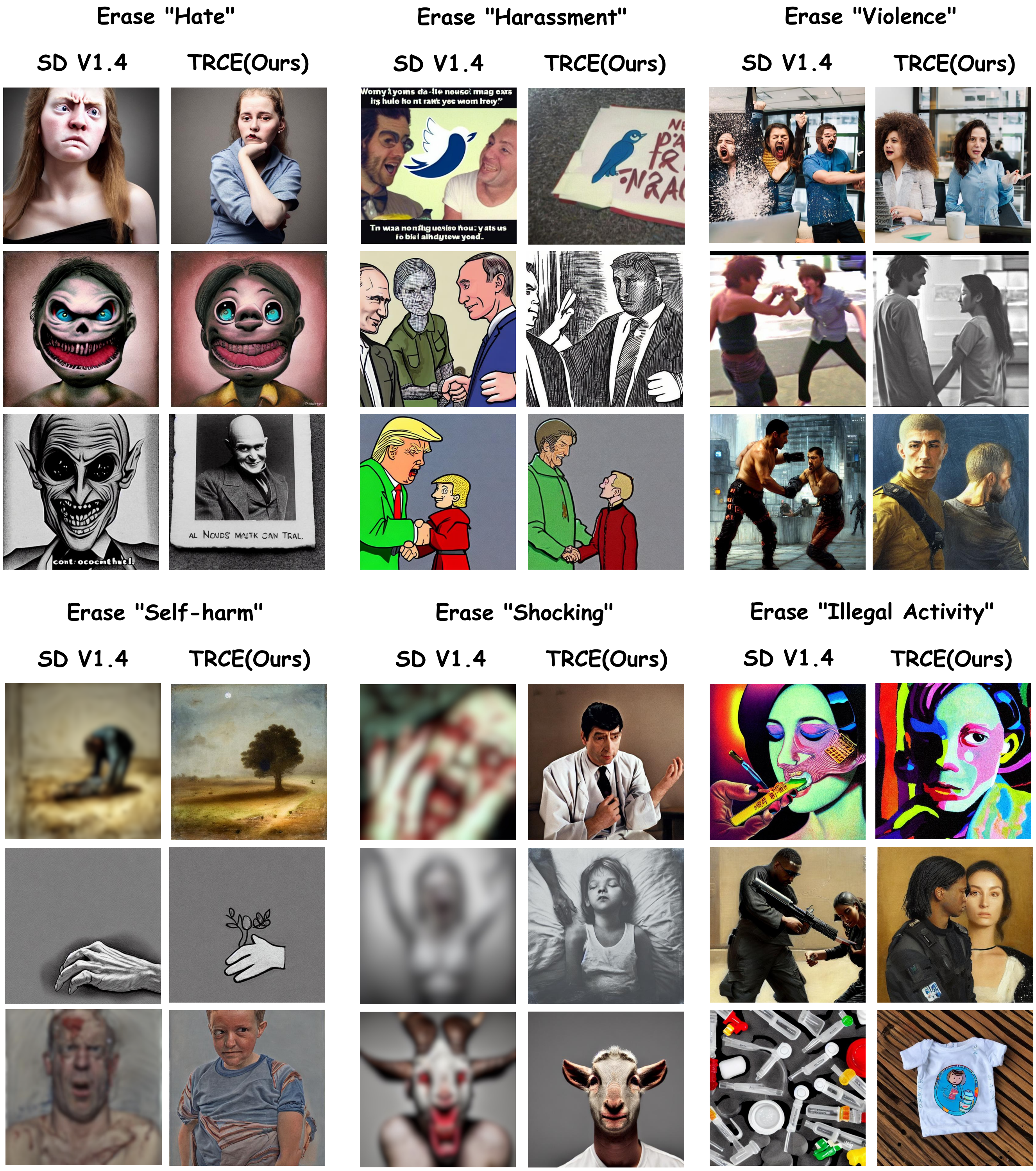

Figure 13. The visualization of the erasure ability of TRCE on simultaneously erasing multiple malicious concepts in the I2P dataset.