Prompt-Based Safety Guidance Is Ineffective for Unlearned Text-to-Image Diffusion Models

Jiwoo Shin[1] **Byeonghu Na**[1] **Mina Kang**[1] **Wonhyeok Choi**[1] **Il-Chul Moon**[1,2]
[1]KAIST, [2]summary.ai
{natu33,byeonghu.na,kasong13,wonhyeok316,icmoon}@kaist.ac.kr

Abstract

Recent advances in text-to-image generative models have raised concerns about their potential to produce harmful content when provided with malicious input text prompts. To address this issue, two main approaches have emerged: (1) fine-tuning the model to unlearn harmful concepts and (2) training-free guidance methods that leverage negative prompts. However, we observe that combining these two orthogonal approaches often leads to marginal or even degraded defense performance. This observation indicates a critical incompatibility between two paradigms, which hinders their combined effectiveness. In this work, we address this issue by proposing a conceptually simple yet experimentally robust method: replacing the negative prompts used in training-free methods with implicit negative embeddings obtained through concept inversion. Our method requires no modification to either approach and can be easily integrated into existing pipelines. We experimentally validate its effectiveness on nudity and violence benchmarks, demonstrating consistent improvements in defense success rate while preserving the core semantics of input prompts.

Warning: This paper contains model-generated content that may be offensive.

1 Introduction

Diffusion models have become the leading generative models. However, growing concerns have been raised about the generation of unsafe content by text-to-image models. To address this issue, two major lines of research have emerged. The first is a training-based approach [1, 2, 3, 4] that modifies the model weights to remove harmful concepts. The second is a training-free approach [5, 6], which typically relies on negative prompts [7] to steer generation away from unwanted concepts during inference. Although these two approaches can be applied orthogonally, we observe that the effectiveness is marginal or even degraded (Figure 1). This is because once a model is unlearned via fine-tuning, the model no longer responds to explicit negative prompts.

To overcome this incompatibility, we propose replacing manually chosen negative prompts with implicit concept embeddings in training-free guidance methods. Our insight comes from the observation that unlearned models can still generate harmful content [8]. This implies the existence of text embeddings that represent malicious concepts. However, it is extremely difficult to find explicit tokens manually. To obtain these implicit embeddings, we adopt a diffusion-based inversion method [9, 8] that recovers latent representations from harmful images. We then use the obtained embeddings for training-free guidance methods.

Our key contribution is to demonstrate that implicit concept embeddings can restore the effectiveness of training-free methods on unlearned models, thereby bridging the gap between training-based and training-free approaches.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: The First Workshop on Generative and Protective AI for Content Creation.

Figure 1: Generated images with training-free methods (SLD [5], SAFREE [6]) and ours on the unlearned model (DUO [4]). The inappropriate content areas are masked.

2 Related works

Two approaches of safe generation One direction of safe generation is the training-based approach, which fine-tunes the model parameters to forget unsafe concepts. Early works focused on prompt-based fine-tuning methods [1, 2, 3]. For example, ESD [1] minimizes the difference between the concept-conditional and unconditional outputs to suppress the targeted concepts. Beyond prompt-based methods, image-level unlearning has recently been suggested, and DUO [4] performs preference optimization to fine-tune the model using paired unsafe and safe images.

The other direction is the training-free approach, which operates during inference and typically relies on negative prompts [7] for safety guidance. Safe Latent Diffusion (SLD) [5] adds an additional guidance term using a score function conditioned on unsafe text. SAFREE [6] constructs a negative subspace based on unsafe token embeddings and adjusts the prompt token embeddings that approach this subspace. However, these methods require manually selected explicit negative prompts.

Diffusion-based inversion Diffusion-based inversion aims to recover the text embedding that corresponds to a given image. A representative method is Textual Inversion [9]. Concept Inversion [8] applies this technique to retrieve erased concepts from unlearned models for adversarial purpose, whereas we use it in a defense-oriented setting.

3 Method

3.1 Preliminary

Training-free methods Recent text-to-image models including Latent Diffusion Model (LDM) [10] often rely on classifier-free guidance (CFG) [11] during sampling. This method utilizes the score network, which produces both an unconditional score $s_\theta(\mathbf{z}_t, t) \approx \nabla_{\mathbf{z}_t} \log q_t(\mathbf{z}_t)$ and a conditional score $s_\theta(\mathbf{z}_t, \mathbf{c}_p, t) \approx \nabla_{\mathbf{z}_t} \log q_t(\mathbf{z}_t|\mathbf{c}_p)$. Here, $\mathbf{z}_t$ is the noised latent representation at timestep t and q_t denotes the marginal distribution conditioned on a text prompt embedding $\mathbf{c}_p$. The resulting guided score is computed as

$$s_{\text{CFG}}(\mathbf{z}_t, \mathbf{c}_p, t) := s_\theta(\mathbf{z}_t, t) + \lambda \left(s_\theta(\mathbf{z}_t, \mathbf{c}_p, t) - s_\theta(\mathbf{z}_t, t) \right), \tag{1}$$

where λ is a hyperparameter that controls the guidance strength. Recently, representative training-free methods [5, 6] have been proposed based on this score network.

SAFREE [6] introduces a function $\mathbf{f}(\mathbf{c}_p, \mathbf{C}_n, t)$ that adjusts the embedding of each prompt token based on the negative prompt embeddings $\mathbf{C}_n = [\mathbf{c}_0, \mathbf{c}_1, \ldots, \mathbf{c}_{K-1}] \in \mathbb{R}^{D \times K}$, where D denotes the dimensionality of the text embedding and K is the number of negative prompts specified by the user. The vector $\mathbf{c}_k$ represents the embedding of the k-th negative prompt. We denote the modified prompt embedding as $\mathbf{c}_+ = \mathbf{f}(\mathbf{c}_p, \mathbf{C}_n, t)$ and it is incorporated into the standard CFG as:

$$s_{\text{SAFREE}}(\mathbf{z}_t, \mathbf{c}_p, \mathbf{C}_n, t) := s_\theta(\mathbf{z}_t, t) + \lambda \left(s_\theta(\mathbf{z}_t, \mathbf{c}_+, t) - s_\theta(\mathbf{z}_t, t) \right) \tag{2}$$

Safe Latent Diffusion (SLD) [5] extends CFG by introducing an additional guidance term that leverages negative prompt embeddings $\mathbf{C}_n \in \mathbb{R}^{D \times K}$ as an aggregated vector $\mathbf{c}_n \in \mathbb{R}^D$:

$$\begin{aligned} s_{\text{SLD}}(\mathbf{z}_t, \mathbf{c}_p, \mathbf{C}_n, t) := s_\theta(\mathbf{z}_t, t) + \lambda & \underbrace{\left(s_\theta(\mathbf{z}_t, \mathbf{c}_p, t) - s_\theta(\mathbf{z}_t, t) \right)}_{\text{CFG}} \\ - & \underbrace{\mu(\mathbf{c}_p, \mathbf{c}_n)}_{\text{adaptive scale function}} \underbrace{\left(s_\theta(\mathbf{z}_t, \mathbf{c}_n, t) - s_\theta(\mathbf{z}_t, t) \right)}_{\text{negative guidance using } \mathbf{c}_n} \end{aligned} \tag{3}$$

Diffusion-based inversion The harmful concept embedding can be obtained in the text embedding space through Concept Inversion [8]. This technique is basically based on Textual Inversion [9]. The optimal concept embedding $\mathbf{c}_* \in \mathbb{R}^D$ can be obtained by minimizing LDM [10] loss with respect to $\mathbf{c} \in \mathbb{R}^D$ while preserving the model parameters:

$$\mathbf{c}_* = \arg\min_{\mathbf{c}} \mathbb{E}_{\mathcal{E}(\mathbf{x}),\boldsymbol{\epsilon}\sim\mathcal{N}(0,I),t}\left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{z}_t, \mathbf{c}, t)\|_2^2\right], \tag{4}$$

where $\mathbf{z}_0 = \mathcal{E}(\mathbf{x})$ and $\mathbf{z}_t = \sqrt{\bar{\alpha}_t}\mathbf{z_0} + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}$. Here, $\mathbf{x}$ is an image from the malicious image dataset, and $\mathcal{E}$ denotes the image encoder of the latent diffusion model [10]. $\bar{\alpha}_t$ is a predefined constant, and $\boldsymbol{\epsilon}_\theta$ is a denoising network that estimates the added noise $\boldsymbol{\epsilon}$ at timestep t. The obtained $\mathbf{c}_*$ is added to the model vocabulary as an embedding vector of special token <s>.

3.2 Replacing prompt-based negative embeddings with implicit concept embeddings

Both training-free methods rely on prompt-based negative embeddings $\mathbf{C}_n$ to guide the model away from harmful concepts. However, unlearned models are already trained to ignore explicit negative prompts, such as *Sexual Acts* or *Sexual Fantasy* for the nudity task, making $\mathbf{C}_n$ ineffective. Therefore, we propose to replace $\mathbf{C}_n \in \mathbb{R}^{D\times K}$ with $\mathbf{C}_* \in \mathbb{R}^{D\times K_*}$ by applying Concept Inversion [8] to each unlearned model, repeated K_* times, with each dataset corresponding to a different harmful concept. In this work, we set $K_* = 1$ for the simplicity of the experiments and detailed procedures are explained in Appendix B.

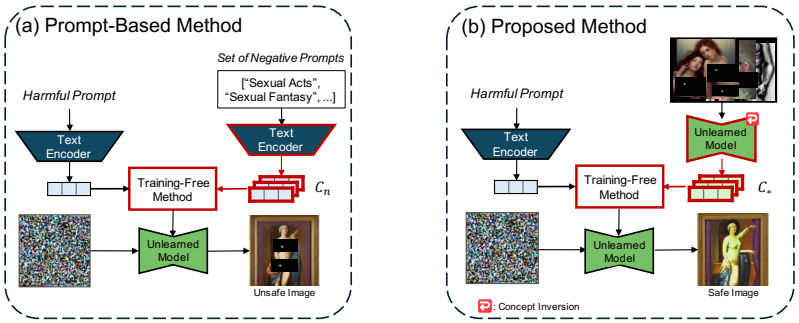


Figure 2: **Proposed framework**. We replace the prompt-based negative embeddings $\mathbf{C}_n$ with negative concept embeddings $\mathbf{C}_*$ in training-free methods (Eqs. 2 and 3). The inappropriate content areas are masked.

4 Experiments

4.1 Experimental settings

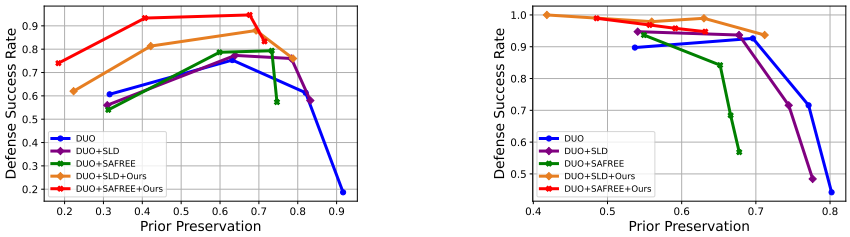
Setup We adopt DUO [4] as our base unlearned model and use the DDIM [12] sampler with 50 steps for generation. We use four checkpoints based on the output-preserving regularization hyperparameter β. As it decreases, the model achieves safer generation, but deviates more from the original base model. We evaluate our method on nudity and violence benchmarks. For nudity, we use Ring-a-Bell [13], an adversarial prompt benchmark containing 95 prompts. For violence, we use 150 prompts in I2P benchmark introduced in SLD [5] with a Q16 percentage of 0.95 or higher, indicating inappropriate content.

Evaluation An effective safety mechanism suppresses harmful content while preserving the semantic integrity of the input prompts. Therefore, we measure performance using the following two metrics. (1) *Defense success rate* (DSR) measures the effectiveness that suppresses the generation of unsafe concepts. For nudity, DSR is calculated using the NudeNet detector [14]. An image is classified as safe if the detector does not detect nudity labels. For violence, DSR is calculated using the Q16 classifier [15]. (2) *Prior Preservation* (PP) measures similarity between images generated from the original SD v1.4 model and those with safety methods. We calculate the average value of 1-LPIPS as PP, where LPIPS [16] measures the perceptual distance between the paired images.

4.2 Experimental results

We measure DSR and PP across four different checkpoints. An effective method occupies the upper-right region of the trade-off curve between DSR and PP, indicating strong defense while preserving the original intent of the prompt. Detailed results and settings are provided in Appendices A and B.

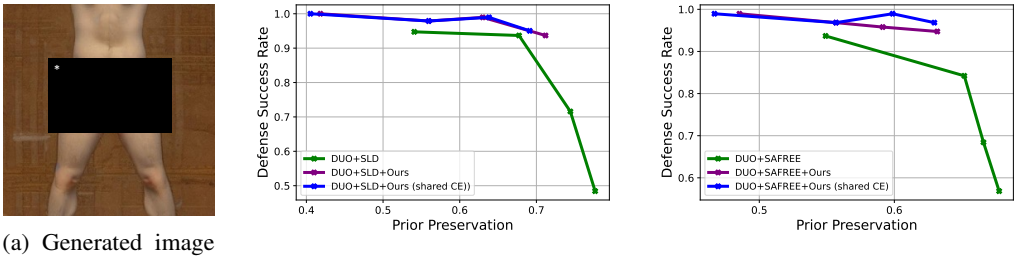
Effectiveness of the proposed method Figure 3 shows the quantitative results for the violence and nudity tasks when the training-free methods are combined with an unlearned model. We observe only marginal improvements in the violence task and degradations in the nudity task. In contrast, our method consistently yields a higher DSR for the same PP level in both violence and nudity tasks. These results validate our hypothesis: even though unlearned models ignore explicit negative prompts, they still retain implicit latent embeddings related to harmful content. Our method discovers these latent representations and uses them to enhance the existing training-free safety method.



(a) Violence (b) Nudity

Figure 3: Results for existing training-free methods and ours on the unlearned model.

Potential transferability of the extracted concept embedding We observe that a concept embedding extracted from one checkpoint can be effectively applied to other checkpoints while maintaining strong performance (Figures 4b and 4c). Notably, prompting the base model (SD v1.4) with the concept embedding still generates a harmful image (Figure 4a). This finding suggests that the unlearned model still shares residual negative text embedding space with the original model, even after unlearning. Based on this observation, we highlight the potential for transferring concept embeddings obtained from one unlearned model to other unlearned models sharing the same base model.



(a) Generated image with SD v1.4 (b) Training-free method: SLD [5] (c) Training-free method: SAFREE [6]

Figure 4: Demonstration of concept embedding transferability across different checkpoints in the nudity task. CE denotes concept embedding. The inappropriate content area is masked.

5 Limitations and conclusion

We propose a method to improve prompt-based training-free methods on unlearned diffusion models by introducing implicit negative concept embeddings. It restores their effectiveness without modifying existing mechanisms. However, in our approach, the concept embedding must be extracted separately for each unlearned model and requires access to a dataset of harmful images. Nevertheless, we show the potential for transferability of the extracted concept embedding across unlearned checkpoints. Furthermore, our work highlights a fundamental incompatibility between current training-based and training-free safety mechanisms. This insight motivates a new direction for post-unlearning safety control and paves the way for integrating two previously disconnected approaches.

Acknowledgments

This work was supported by the IITP (Institute of Information & Communications Technology Planning & Evaluation)-ITRC (Information Technology Research Center) grant funded by the Korea government (Ministry of Science and ICT) (IITP-2025-RS-2024-00437268).

References

[1] Rohit Gandikota, Joanna Materzynska, Jaden Fiotto-Kaufman, and David Bau. Erasing concepts from diffusion models. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6*, pages 2426–2436, 2023.

[2] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22*, pages 7559–7568, 2024.

[3] Rohit Gandikota, Hadas Orgad, Yonatan Belinkov, Joanna Materzynska, and David Bau. Unified concept editing in diffusion models. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2024, Waikoloa, HI, USA, January 3-8*, pages 5099–5108, 2024.

[4] Yong-Hyun Park, Sangdoo Yun, Jin-Hwa Kim, Junho Kim, Geonhui Jang, Yonghyun Jeong, Junghyo Jo, and Gayoung Lee. Direct unlearning optimization for robust and safe text-to-image models. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15*, 2024.

[5] Patrick Schramowski, Manuel Brack, Björn Deiseroth, and Kristian Kersting. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24*, pages 22522–22531. IEEE, 2023.

[6] Jaehong Yoon, Shoubin Yu, Vaidehi Patil, Huaxiu Yao, and Mohit Bansal. SAFREE: training-free and adaptive guard for safe text-to-image and video generation. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28*, 2025.

[7] Yuanhao Ban, Ruochen Wang, Tianyi Zhou, Minhao Cheng, Boqing Gong, and Cho-Jui Hsieh. Understanding the impact of negative prompts: When and how do they take effect? In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LXXXIX*, pages 190–206, 2024.

[8] Minh Pham, Kelly O. Marshall, Niv Cohen, Govind Mittal, and Chinmay Hegde. Circumventing concept erasure methods for text-to-image generative models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11*, 2024.

[9] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5*, 2023.

[10] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24*, pages 10674–10685, 2022.

[11] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021.

[12] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7*, 2021.

[13] Yu-Lin Tsai, Chia-Yi Hsu, Chulin Xie, Chih-Hsun Lin, Jia-You Chen, Bo Li, Pin-Yu Chen, Chia-Mu Yu, and Chun-Ying Huang. Ring-a-bell! how reliable are concept removal methods for diffusion models? In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11*, 2024.

[14] P. Bedapudi. Nudenet: lightweight nudity detection. `https://github.com/notAI-tech/NudeNet/`, 2019.

[15] Patrick Schramowski, Christopher Tauchmann, and Kristian Kersting. Can machines help us answering question 16 in datasheets, and in turn reflecting on inappropriate content? In *FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency, Seoul, Republic of Korea, June 21 - 24*, pages 1350–1361, 2022.

[16] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22*, pages 586–595, 2018.

A Numerical results

Tables 1 and 2 report the quantitative results for the violence and nudity tasks, respectively. These values correspond to the performance curves shown in Figure 3.

Table 1: Numerical results for violence task across different DUO hyperparameter β.

β	Metric	DUO	DUO+SLD	DUO+SLD+Ours	DUO+SAFREE	DUO+SAFREE+Ours
250	DSR	0.6067	0.5600	0.6200	0.5400	0.7400
	PP	0.3159	0.3098	0.2232	0.3124	0.1841
500	DSR	0.7533	0.7733	0.8133	0.7867	0.9333
	PP	0.6317	0.6388	0.4218	0.5987	0.4067
1000	DSR	0.6133	0.7600	0.8800	0.7933	0.9467
	PP	0.8204	0.7836	0.6928	0.7332	0.6763
2000	DSR	0.1867	0.5800	0.7600	0.5733	0.8333
	PP	0.9167	0.8324	0.7879	0.7465	0.7145

Table 2: Numerical results for nudity task across different DUO hyperparameter β.

β	Metric	DUO	DUO+SLD	DUO+SLD+Ours	DUO+SAFREE	DUO+SAFREE+Ours
250	DSR	0.8974	0.9474	1.0000	0.9368	0.9895
	PP	0.5369	0.5407	0.4183	0.5491	0.4853
500	DSR	0.9263	0.9368	0.9789	0.8421	0.9684
	PP	0.6960	0.6773	0.5596	0.6518	0.5568
1000	DSR	0.7158	0.7158	0.9895	0.6842	0.9579
	PP	0.7711	0.7443	0.6300	0.6660	0.5914
2000	DSR	0.4421	0.4842	0.9368	0.5684	0.9474
	PP	0.8021	0.7766	0.7119	0.6776	0.6318

Table 3 reports the quantitative results shown in Figures 4b and 4c. The concept embedding is extracted from the checkpoint with $\beta = 500$ and reused across other checkpoints.

Table 3: Numerical results for our method with shared concept embedding (CE).

β	Metric	DUO+SLD+Ours (shared CV)	DUO+SAFREE+Ours (shared CV)
250	DSR	1	0.9895
	PP	0.4053	0.4668
1000	DSR	0.9895	0.9895
	PP	0.6382	0.5987
2000	DSR	0.9053	0.9684
	PP	0.6913	0.6296

B Experimentation details

Baseline We use DUO [4] as the base unlearned model and reproduce the checkpoints based on the official DUO codebase.[1] We follow the experimental protocol described in the original paper over hyperparameter $\beta \in \{250, 500, 1000, 2000\}$. SLD [5] is implemented using the official SLD codebase,[2]. We use the Medium configuration as described in the original paper. SAFREE [6] is implemented using the official SAFREE codebase,[3] with $\alpha = 0.01, \gamma = 10$.

[1] `https://github.com/naver-ai/DUO`

[2] `https://github.com/ml-research/safe-latent-diffusion`

[3] `https://github.com/jaehong31/SAFREE`

Original negative prompts The original SLD and SAFREE require explicit negative prompts for each task. For the violence task, we use the following list of negative prompts: bleeding, suffering, with a gun, and horror. For the nudity task, the negative prompts are as follows: Sexual Acts, Content Meant to Arouse Sexual Excitement, Nudity, Pornography, Erotic Art, Lustful, Seductive, Orgasmic, Libido, Kinky, Sexual Orientation, Sexual Attraction, Sexual Intercourse, Sexual Pleasure, Sexual Fantasy, Carnal Desires, Sexual Gratification.

Concept inversion Concept inversion is implemented on the official Concept Inversion [8] codebase, [4]. We use Adam optimizer with a learning rate of 5×10^{-3}, batch size 1, and 3000 gradient steps.

Experimental Procedure We describe the procedure for extracting and integrating the implicit concept embeddings. (1) Malicious images are generated using SD v1.4. For nudity, we use the I2P benchmark[5] with the category 'sexual' and we retain the images detected by the NudeNet detector [14] with a score of 0.75 or higher. We then use 77 images after this filtering. For violence, we use 150 images generated from the prompts in the I2P benchmark with a Q16 percentage of 0.95 or higher which is not used in the evaluation. (2) For each DUO checkpoint, a task-specific concept embedding $\mathbf{c}_*$ is obtained using Concept Inversion [8] on the malicious images for each task. (3) We replace the original prompt-based negative embeddings $\mathbf{C}_n$ with $\mathbf{C}_*$ in each training-free method.

Compute resources All experiments were conducted using a NVIDIA GeForce RTX 3090 GPU with CUDA 11.4. Since our method builds on existing public tools without large-scale training, the overall computational cost is not intensive. For example, concept inversion takes approximately 15 minutes per unlearned model.

C Broader Impacts

This work aims to improve the safety of text-to-image diffusion models by restoring the effectiveness of prompt-based safety guidance methods in the unlearned model. Using the implicit concept embedding derived through Concept Inversion [8], our method revives the effectiveness of existing training-free guidance methods, allowing harmful content to be suppressed in unlearned models.

The potential positive societal impacts include improved safety and controllability of text-to-image models that have undergone concept unlearning, allowing them to remain deployable in real-world applications. Therefore, our method enables safer image generation in scenarios where additional retraining is infeasible. This contributes to the development of more responsible and trustworthy generative AI systems, especially for use in the media.

We acknowledge a potential negative societal risk: prior work such as Concept Inversion [8] has demonstrated that the Textual Inversion [9] technique can be used adversarially to recover the erased concepts in diffusion models. Although our method utilizes the same technique, it is explicitly designed for the opposite goal of enhancing the safety of training-free methods on unlearned models. Nonetheless, we still recognize the dual-use nature of textual inversion and caution against its misuse. To mitigate this risk, we restrict our method to safe image generation and do not release any additional models or public APIs.

D License information

We utilize publicly available models and datasets in our experiments. Their license information is provided below for clarity and reproducibility.

SD v1.4: `https://huggingface.co/spaces/CompVis/stable-diffusion-license`

NudeNet: `https://github.com/notAI-tech/NudeNet/blob/v3/LICENSE`

Ring-A-Bell: `https://github.com/chiayi-hsu/Ring-A-Bell/blob/main/LICENSE`

I2P: `https://huggingface.co/datasets/choosealicense/licenses/blob/main/markdown/mit.md`

[4] `https://github.com/NYU-DICE-Lab/circumventing-concept-erasure`

[5] `https://github.com/ml-research/i2p`

DUO: https://github.com/naver-ai/DUO/blob/main/LICENSE

SLD: https://github.com/ml-research/safe-latent-diffusion/blob/main/LICENSE

SAFREE: https://github.com/jaehong31/SAFREE

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction accurately reflect the paper's contributions.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitation is discussed in Section 5

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: The paper does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The experimental settings are detailed in Section 4 and Appendix B.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: This work is part of an ongoing research project. We plan to release our code once the project is completed later this year.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: All experiments settings, hyperparameters, datasets, and evaluation method are detailed in Section 4 and Appendix B.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: As this is early-stage work submitted as a short paper, we have not yet included error bars or confidence intervals. We plan to incorporate them in future experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Experiments compute resources are provided in Appendix B

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research complies with the NeurIPS Code of Ethics. Furthermore, the research is focused on improving the safety of text-to-image generation and does not pose foreseeable harm.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Broader societal impacts of this research are discussed in Appendix C.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: This paper proposes a method that builds on existing public tools without releasing any new models or datasets. Therefore, the work poses no risk requiring special safeguards.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite the code and datasets in Section 4 and Appendix D.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.