ETCE: Efficient Two-Stage Concept Erasure for Text-to-Image Diffusion Models

Qi-Ang Hu
Tsinghua University
China
hqafbhff@163.com

Nisha Huang
Tsinghua University, Pengcheng Laboratory
China
hns24@mails.tsinghua.edu.cn

Yizhou Lin
Tsinghua University
China
yz-lin24@mails.tsinghua.edu.cn

Jie Guo
Pengcheng Laboratory
China
guoj01@pcl.ac.cn

Xiu Li
Tsinghua University
China
li.xiu@sz.tsinghua.edu.cn

Tong-Yee Lee
National Cheng-Kung University
Taiwan
tonylee@ncku.edu.tw

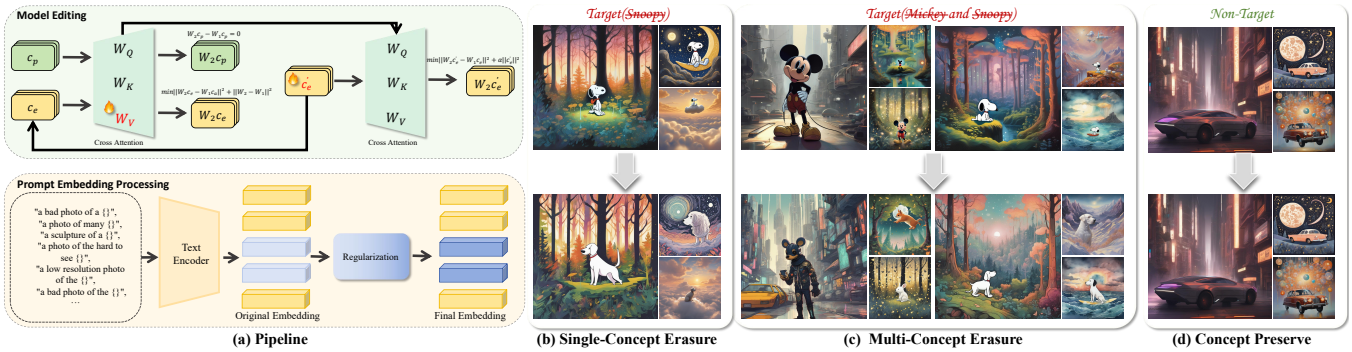

Figure 1: Our proposed (a) Efficient Two-Stage Concept Erasure (ETCE) algorithm demonstrates three key capabilities: (b) single-concept erasure, (c) multi-concept erasure, while effectively (d) preserving non-target concepts.

ACM Reference Format:
Qi-Ang Hu, Nisha Huang, Yizhou Lin, Jie Guo, Xiu Li, and Tong-Yee Lee. 2025. ETCE: Efficient Two-Stage Concept Erasure for Text-to-Image Diffusion Models. In *Proceedings of SIGGRAPH Asia 2025 Posters (SA '25 Poster).* ACM, New York, NY, USA, 2 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Recent advances in text-to-image (T2I) models enable highly realistic image synthesis, yet these models often produce outputs containing infringing or illegal content. These issues stem primarily from substantial noise in the training data. Given the prohibitive cost of data purification and full model retraining, developing efficient concept erasure methods is essential for content security.

Concept erasure methods can be divided into training-based and training-free categories. While training-based methods such as ESD [Gandikota et al. 2023] offer strong erasure performance, they require costly hyperparameter tuning and are too slow for real-time use. Training-free approaches like UCE [Gandikota et al. 2024], SPEED [Li et al. 2025], and ICE [Lin et al. 2025] allow efficient removal but suffer from incomplete target-concept erasure and unintended non-concept deletion. To address these issues, we propose ETCE, a novel approach that integrates adversarial fine-tuning with null-space constraints and prompt-embedding optimization to ensure thorough removal of target concepts (Fig. 1(b), Fig. 1(c)) while robustly preserving non-target semantic content (Fig. 1(d)).

2 Method

Our concept erasure algorithm implements a two-phase architecture: **(1) Model Editing, (2) Prompt Embedding Processing.** The complete workflow diagram is presented in Fig. 1(a). In the Model Editing phase, we formulate the following objective function:

$$min||W_2c_e - W_1c_a||^2 + ||W_2 - W_1||^2, s.t., W_2c_p - W_1c_p = 0, \quad (1)$$

where c_p, c_e, and c_a denote the stacked embedding vectors of non-target, target, and anchor concepts, respectively, while W_2 and W_1 represent the post-editing and original Stable Diffusion value matrices. By imposing constraints, we prevent interference with non-target concepts during target concept removal. We then optimize target concept embeddings for more thorough erasure using the following objective function:

$$\min \left\|\mathrm{W}_2c_e' - \mathrm{W}_1c_e\right\|^2 + \alpha \left\|\mathrm{c}_e'\right\|^2, \quad (2)$$

SA '25 Poster, Hong Kong

ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX

Table 1: Concept erasure quantitative comparison.

Concept	Nemo	Pooh	Snoopy	Pikachu	Hello Kitty
	CS ↓	CS ↓	CS ↓	CS ↓	CS ↓
SD v1.4	25.90	27.66	28.53	27.45	27.76
Erase **Nemo**					
	CS ↓	FID ↓	FID ↓	FID ↓	FID ↓
ESD	22.89	35.50	41.41	28.78	35.92
UCE	23.42	25.17	28.15	21.54	26.80
SPEED	23.33	25.86	31.03	24.68	30.16
Ours	**21.68**	**19.35**	**21.49**	**17.61**	**22.16**
Erase **Nemo** and **Pooh** and **Snoopy**					
	CS ↓	CS ↓	CS ↓	FID ↓	FID ↓
ESD	23.62	25.45	25.33	51.36	63.71
UCE	23.68	25.45	23.00	26.13	31.28
SPEED	23.28	23.07	23.52	25.96	31.37
Ours	**21.93**	**21.37**	**21.89**	**21.90**	**25.43**

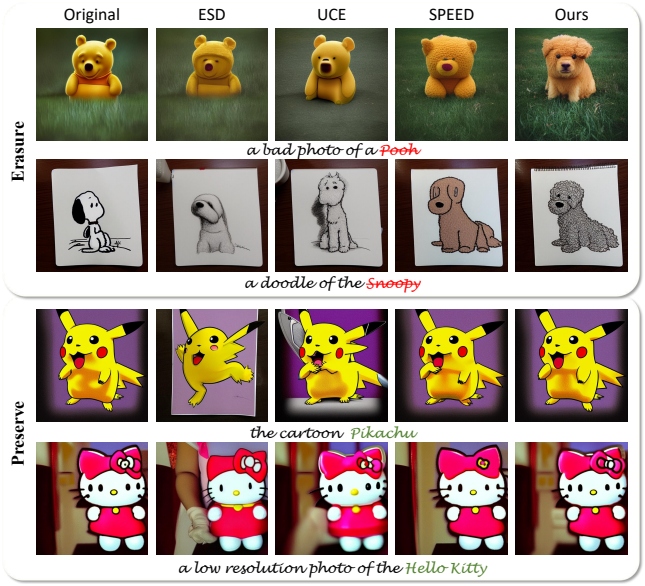

Figure 2: Concept erasure qualitative comparison.

where c_e' denotes the optimized target concept embeddings, and α serves as a manually adjustable parameter to control the regularization strength. By optimizing this objective function, we derive the optimal embedding solution as $c_e' = \left(\alpha I + W_2^T W_2\right)^{-1} \left(W_2^T W_1\right) c_e$. Substituting the optimal embedding c_e' for c_e in function 1 to perform concept erasure yields the desired value matrix.

In the Prompt Embedding Processing phase, we optimize prompt embeddings through a targeted adjustment pipeline. First, we identify token positions containing target concept embeddings within the original prompt embedding space. From these, we extract the erasure-targeted embedding matrix E_T, which undergoes singular value decomposition (SVD): $E_T = U\Sigma V^T$, where $\Sigma = \text{diag}[\sigma_1, \sigma_2, \ldots, \sigma_n]$. Since the embedding matrix primarily encodes target concept information, larger singular values typically correspond to these concepts. Thus, for singular values exceeding the mean, we apply: $\sigma_i' = e^{-\beta\sigma_i} \cdot \sigma_i$. The original E_T is thus adjusted to $E_T' = U\Sigma' V^T$, where β serves as a tunable parameter controlling suppression intensity. Feeding the optimized prompt embeddings into downstream generators enhances concept erasure, enabling more thorough target semantic suppression during image generation.

3 Experiments

We conducted a comprehensive concept erasure evaluation comparing our algorithm with three baseline methods (ESD [Gandikota et al. 2023], UCE [Gandikota et al. 2024], and SPEED [Li et al. 2025]) on five target objects (Nemo, Pooh, Snoopy, Pikachu, and Hello Kitty) under both single-concept and multi-concept erasure cases. Using CLIP Score (CS) and Fréchet Inception Distance (FID) as evaluation metrics with adjustable parameters ($\alpha = 0.001$ and $\beta = 0.003$), we generated 10 images per template across 80 distinct templates. As shown in Table 1, our algorithm consistently outperforms all baseline methods (ESD, UCE, and SPEED), achieving the lowest CS scores for target concept erasure and the lowest FID values for non-target concept preservation, demonstrating superior performance in both concept removal and content retention. Figure 2 visually confirms precise target concept removal while maintaining high-quality generation and minimal impact on non-target outputs compared to original SD v1.4.

4 Conclusion

This paper introduces ETCE, an efficient two-stage concept erasing method for T2I models. During model editing, it performs adversarial fine-tuning with null-space constraints, while in prompt embedding processing, it regularizes target concept embeddings. This approach effectively addresses two key challenges: incomplete target concept removal and unintended non-target concept deletion. Experiments demonstrate that our method achieves precise target concept removal while preserving others. Future work will extend this approach to implicitly expressed concepts in prompts.

Acknowledgments

The work is supported in part by the Shenzhen Key Laboratory of Next Generation Interactive Media Innovative Technology, China (No. ZDSYS20210623092001004), and the National Science and Technology Council, Taiwan (No. 114-2221-E-006-114-MY3).

References

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

Rohit Gandikota, Hadas Orgad, Yonatan Belinkov, Joanna Materzyńska, and David Bau. 2024. Unified concept editing in diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 5111–5120.

Ouxiang Li, Yuan Wang, Xinting Hu, Houcheng Jiang, Tao Liang, Yanbin Hao, Guojun Ma, and Fuli Feng. 2025. Speed: Scalable, precise, and efficient concept erasure for diffusion models. *arXiv preprint arXiv:2503.07392* (2025).

Yizhou Lin, Nisha Huang, Kaer Huang, Henglin Liu, Yiqiang Yan, Jie Guo, Tong-Yee Lee, and Xiu Li. 2025. ICE: Intercede Concept Erasure in Text-to-Image Diffusion Models. In *Proceedings of the 33th ACM International Conference on Multimedia*.