Where Do Erased Concepts Go in Diffusion Models?

Kevin Lu
Northeastern University
lu.kev@northeastern.edu

Nicky Kriplani
New York University
ak9100@nyu.edu

Rohit Gandikota
Northeastern University
gandikota.ro@northeastern.edu

Minh Pham
New York University
mp5847@nyu.edu

David Bau
Northeastern University
d.bau@northeastern.edu

Chinmay Hegde
New York University
chinmay.h@nyu.edu

Niv Cohen
New York University
nc3468@nyu.edu

Abstract

In this paper, we uncover a dichotomy in how concept erasure methods modify diffusion models: ***guidance-based*** *avoidance versus* ***destruction-based*** *removal. Through systematic analysis of various erasure techniques and their interactions with adversarial attacks, we demonstrate that these two distinct mechanisms lead to fundamentally different behaviors and robustness properties. To illuminate this distinction, we introduce* NOISING ATTACK*, a training-free attack that adds controlled noise during the diffusion process.*

To better understand the differences between the types of erasure methods, we track how concepts evolve throughout the erasure process. We find that guidance-based methods work by disrupting the model's ability to follow text conditioning toward erased concepts, resulting in diverse alternative generations. In contrast, destruction-based approaches actively reduce the likelihood of generating the erased concept, causing consistent redirection to specific alternative concepts we term "memory sinks." Our findings suggest that the choice between guidance-based avoidance and destruction-based removal presents a fundamental trade-off between generation diversity and adversarial robustness in concept erasure.

1. Introduction

What happens to a concept when you erase it from a diffusion model? Through extensive analysis of concept erasure methods and their interactions with adversarial attacks, we uncover a fundamental dichotomy: Some methods achieve erasure by reducing guidance toward the target concept, whereas others do so by destructing the model's capacity to generate it.

Understanding these underlying mechanisms is crucial for multiple reasons. First, it helps identify why certain erasure methods are more robust to adversarial attacks while others remain vulnerable. Second, it reveals whether erased concepts are truly eliminated or merely suppressed in specific contexts. Most importantly, understanding these mechanisms can guide the development of more effective erasure techniques that balance robustness with preservation of desired model capabilities.

While the proposed methods for concept erasure highly vary in loss and training dynamics, their study has so far been less focused on understanding *how* they prevent a model from generating the erased concepts. To address this gap, we develop a comprehensive evaluation framework that systematically explores the underlying mechanisms of different erasure approaches. Our framework employs established adversarial attacks [14, 17, 27] not merely as evaluation metrics but as analytical tools to understand how different erasure methods operate.

In mammalian brains, neuroscience research has revealed fascinating mechanisms of memory modification. Zhang et al. [26] identified distinct neural circuits in rats that process fear and reward memories, demonstrating that these circuits can systematically redirect and modulate memory associations. Their findings showed that fear-associated memories were particularly susceptible to suppression, suggesting an inherent biological mechanism for memory erasure. We hypothesize that, similar to neural circuits, diffusion models redirect conceptual information to specific semantic locations depending on the erasure method being applied.

To better explore the inner workings of concept erasure,

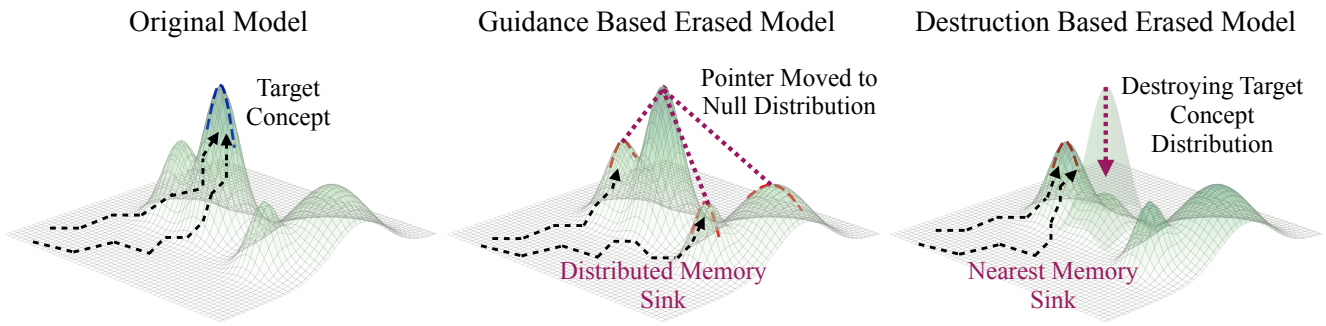

Figure 1. Concept erasure methods can be broadly categorized into two types: (1) Guidance Based Avoidance, which erases a concept by redirecting the model to a different concept locations when prompted for it. (2) Destruction Based Removal, which reduces the likelihood of the concept that is being erased, thereby forcing the model to a nearby concept as a memory sink.

we move beyond traditional input-modification attacks by proposing NOISING ATTACK, a simple training-free inference attack based on noising during the diffusion process. This approach reveals that supposedly erased concepts can still emerge during generation, even without explicit optimization. Furthermore, our analysis of model behavior when prompted with erased concepts leads to a crucial distinction between erasure mechanisms: guidance-based avoidance, which prevents generation through conditional guidance manipulation, and destruction-based removal, which fundamentally suppresses the concept's generation likelihood (see Fig.1). We validate this model by demonstrating that destruction-based methods consistently redirect to high-likelihood alternative concepts, or "memory sinks," that existed in the original model.

We make the following contributions: (i) A categorization system to separate erasure methods into two distinct regimes: guidance-based avoidance and destruction-based removal. (ii) NOISING ATTACK, a training-free inference time noising attack to circumvent concept erasure using a simple modification to the diffusion process. (iii) An evaluation framework to shed light on the mechanisms of various concept erasure methods.

2. A Framework for Concept Erasure

While existing work primarily focuses on preventing specific concept generation, understanding the underlying mechanisms of concept erasure remains a crucial challenge (see Appendix A for a discussion of existing works). To address this, we present an evaluation framework that uses adversarial attacks not merely as robustness metrics, but as analytical tools to illuminate how different erasure methods fundamentally operate. This framework allows us to study how erasure methods transform model behavior and identify their distinctive characteristics.

Our framework combines established adversarial techniques with a novel inference-time attack we introduce. Previous attacks have demonstrated that concept erasure can be circumvented through both adversarial inputs and model fine-tuning [14, 17, 20, 23, 27], suggesting erasure methods may also be vulnerable to alterations in the generation process itself. As existing attacks primarily rely on optimizing input prompts or noise patterns, their dependence on carefully crafted examples may limit their ability to expose broader vulnerabilities. Therefore, we propose a new attack that operates during inference by modifying the diffusion process directly.

2.1. Training-Free Inference Time Noising Attack

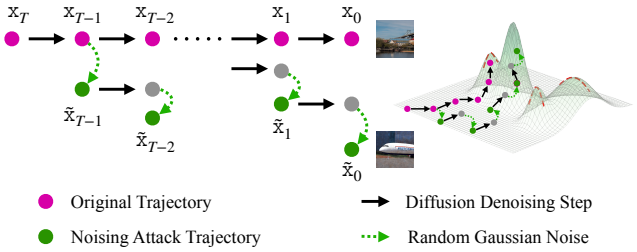

Figure 2. NOISING ATTACK adds Brownian motion to the diffusion trajectory. At every diffusion denoising timestep, we add back a controlled amount of noise to allow the model to search in a larger latent space.

Our proposed method uses a simple yet effective training-free noising attack that operates during inference by adding controlled noise after each denoising step:

$$\tilde{x}_{t-1} = (\tilde{x}_t - \alpha\epsilon_D) + \eta\epsilon \tag{1}$$

where $\alpha\epsilon_D$ represents the standard denoising process, and $\eta\epsilon$ represents additional Gaussian noise scaled by parameter η. We explore values of η in the range $[0.2, 5.0]$ to balance model predictions and noise injection.

As shown in Fig. 2, at every denoising step in diffusion process we add a scaled amount of noise to the latent. This approach performs a controlled exploration of the model's latent space through Brownian motion along the diffusion trajectory. In theory, if an erasure model simply deviates the trajectory of a concept, our attack helps expand the diffusion trajectory bandwidth and allows the latents to find the

	Erased Concept (↓)	Unerased Concepts (↑)	Concept Inversion Attack (↓)	Noising Attack Attack (↓)	UnlearnDiffAtk (↓)
GA	24.27 ± 2.74	28.82 ± 2.82	**22.73 ± 2.48**	**26.11 ± 2.15**	26.63 ± 1.95
UCE	22.40 ± 5.16	**31.15 ± 2.30**	30.65 ± 2.03	27.79 ± 3.57	28.22 ± 3.21
ESD-x	21.12 ± 4.10	30.66 ± 2.48	30.56 ± 2.37	28.02 ± 2.73	28.81 ± 2.17
ESD-u	**20.87 ± 3.42**	29.41 ± 3.26	27.99 ± 3.35	27.74 ± 2.47	**25.48 ± 2.77**
Task Vector	23.07 ± 2.97	30.72 ± 2.41	25.69 ± 2.58	26.51 ± 2.17	27.39 ± 1.67

Table 1. ESD-u and UCE are effective at erasing concepts and better preserve the unerased concepts. However, Gradient Ascent and Task Vector show stronger robustness against the adversarial attacks but show weaker erasing effects. Bold numbers indicate the best results, while underlined numbers represent the second-best.

"better" unerased distribution (see right Fig. 2). To evaluate attack effectiveness, we execute it five times and select the generated image with the highest CLIP score relative to the target concept. This represents a basic search strategy for identifying successful attack instances. We ground this attack method to the DDIM formulation in Appendix E.

2.2. Framework setting

Concept erasure methods are generally evaluated on their ability to supress the erased concept and preserve the generation quality of unrelated control concepts. However, these evaluations are not sufficient to validate the erasure methods in unexpected settings. We therefore propose a framework to better understand these methods.

Erasure methods. We use the following concept erasure methods for our evaluations: *Baseline* [16] - Unedited Stable Diffusion 1.4 model (no erasure); *UCE* [7] - A closed-form solution editing of the cross-attention weight in the model to replace the target concept and preserve other concepts; *ESD-u* [5] - fine-tunes the pre-trained diffusion U-Net model weights to remove a specific style or concept when conditioned on a specific prompt; *ESD-x* [5] - fine-tunes only the cross-attention layers, modifying how textual conditioning influences latent feature modulation; *Task Vector* [15] - Finetuning the U-net to increase the likelihood of the target concept, and then editing the model in the opposite direction using the Task Vector technique [11]; *GA* - direct gradient ascent to reduce the likelihood of the target concept. Please see Sec.F for implementation details.

Evaluation Strategies. We use the following methods to evaluate different erasure strategies: *Standard prompt* - Giving the model a simple prompt containing the concept name; *Noising attack* - our suggested attack based on adding noise to the diffusion generation process (see Sec.2.1); *UnearnDiffAtk* [27] - adversarial prompt generation methody exploiting their intrinsic classification abilities of the model to bypass unlearning mechanisms; *Concept Inversion* [14] - Finding new inputs to the text encoder to induce generation of the target concept (using white box access textual inversion [4]). Please see Sec.F for implementation details.

Concepts and metrics. *Concepts* - we conduct our experiments on 10 object concepts and 3 art styles. We report average results in the main text, and show detailed analysis in App.C. *Metrics* - we evaluate similarity using CLIP score [10]. For the Control Accuracy, measuring how the erasure affect other concepts, we use the full collection of concepts (but the target one) and report the CLIP similarity (see Fig.4).

2.3. Results

We present our comprehensive evaluation results in Tab.1. Our analysis reveals several key insights about concept erasure mechanisms. First, we validate that all methods successfully erase target concepts, with ESD and UCE demonstrating superior performance in both erasure effectiveness and preservation of unrelated concepts. However, our adversarial evaluation unveils interesting patterns in method robustness: Gradient Ascent and Task Vector show stronger resistance to concept inversion and NOISING ATTACK while ESD-u exhibits resilience against input space adversarial attacks like UnlearnDiffAtk. The concept inversion attack proves most effective overall, likely due to its white-box access, image supervision, and soft prompt optimization capabilities.

This distinct pattern of vulnerabilities leads us to propose a fundamental dichotomy in erasure approaches: *guidance-based* avoidance methods versus *destructive* removal methods. Guidance-based approaches, which primarily have strong erasure effects, appear to work by redirecting the model's attention rather than eliminating the concept entirely. In contrast, destructive methods show greater resistance to parameter and latent space attacks, suggesting that they fundamentally alter the model's knowledge of the concept.

3. Guidance-Based and Destruction-Based

We uncover a dichotomy of concept erasure methods in diffusion models:

Guidance-based avoidance methods erase the concept by avoiding the target concept and guiding the model to other alternative, unrelated distributions.

Destruction-based removal methods suppress the model's associated likelihood with the concept being erased.

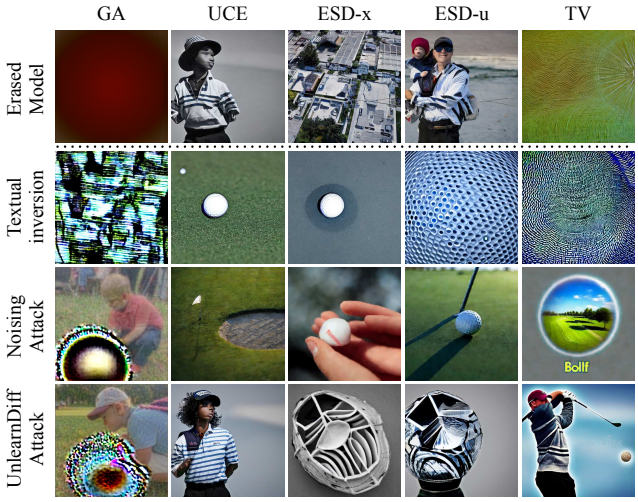

Figure 3. Erasing "Golf Ball" concept using different methods, we show that our training-free inference time noising attack effectively circumvents most of the erasure methods

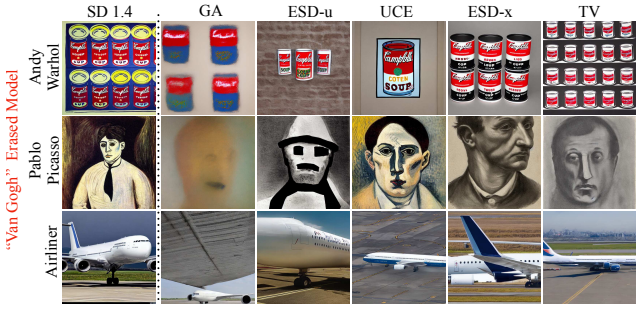

Figure 4. We show the undesirable effects of the erasure methods when erased "Van Gogh" concept on unerased concepts like "Airliner" and "Picasso". We find that Gradient Ascent has the most interference with undesired concept.

3.1. Where Do Erased Concepts Go?

To better understand the behavior of different erasure methods and which model better describes them, we analyze how model outputs evolve when the erased concept is used as a prompt. We examine this trajectory by analyzing CLIP image embeddings of the generated images. As we explain below, the nature of the erasure—whether avoidance-based or destruction-based—strongly influences what the model generates in place of the erased concept.

When destruction-based methods are used, the edited model may still be weakly guided toward the erased concept, but fails to generate it, as it no longer lies in a high-likelihood region of the image manifold. Consequently, these methods consistently push the diffusion model to select a nearby high-likelihood alternative (see right of Fig. 1). Therefore, Gradient Ascent and Task Vector tend to converge on a consistent alternative concept to replace the erased one. These methods also exhibit greater robustness in Table 1.

In contrast, guidance-based avoidance methods work by interfering with the guidance mechanism itself, thereby preventing the model from guiding the generation toward the erased concept. As a result, such models tend to produce a more diverse set of alternative concepts. We suggest that ESD-u and ESD-x fall into this category: they display lower robustness (see Table 1) and produce more varied outputs when prompted with the erased concept.

In our evaluation, we prompt each erased model with the concept intended for erasure and generate 25 images per concept. For each model, we then plot: (i) the distance between the average CLIP embedding of the generated images and that of the original (unedited) model (Fig.7); and (ii) the internal spread of CLIP embeddings across generations (Fig.8). To better illustrate these effects, we also include qualitative results in Fig. 5. We find these results consistent with our categories.

3.2. Analysis

We further analyse the differences between guidance-based avoidance and destruction-based removal methods.

Inpainting of erased concept as a probe. The main motivation of utilizing inpainting as an evaluation method is to thoroughly check the model's knowledge by presenting some "contextual" information about the erased concept. As seen in Fig.6 and Tab.2, the destruction-based removal methods exhibit significantly poorer performance in the inpainting task compared to guidance-based avoidance methods. This provides supporting evidence that destruction-based erasure effectively removes the model's knowledge related to the erased concept. On the other hand, while the inpainting capabilities of guidance-based avoidance methods are not perfect, they remain closer to those of the unedited model.

Does the *memory sink* already exist in the original model? Yes, we find that the memory sink concepts are not novel concepts generated during the unlearning process but rather already exist in the original model's knowledge. To analyze this, we generate memory sink concepts from the erased model and use them for finding memory sink concepts in the original model through textual inversion [4]. Once we find an embedding in the original model, we generate 100 images and measure the CLIP image embedding distance with the training data. We term this measure Inversion Consistency as it measures the consistency of the text embedding's ability to generate the memory sink concept in the original and edited model. Our results in Table 5 indicate that the memory sink was indeed present in the original model prior to the erasure as it is consistently produced by the original model, and exhibits comparable behavior in both models. (see Sec.F for implementation details).

References

[1] AUTOMATIC1111. Negative prompt, 2022. 7

[2] Zhi-Yi Chin, Chieh-Ming Jiang, Ching-Chun Huang, Pin-Yu Chen, and Wei-Chen Chiu. Prompting4debugging: Red-teaming text-to-image diffusion models by finding problematic prompts. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024. 7

[3] Amil Dravid, Yossi Gandelsman, Kuan-Chieh Wang, Rameen Abdal, Gordon Wetzstein, Alexei A Efros, and Kfir Aberman. Interpreting the weight space of customized diffusion models. *arXiv preprint arXiv:2406.09413*, 2024. 7

[4] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 3, 4

[5] 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. 3, 7

[6] Rohit Gandikota, Joanna Materzyńska, Tingrui Zhou, Antonio Torralba, and David Bau. Concept sliders: Lora adaptors for precise control in diffusion models. In *European Conference on Computer Vision*, pages 172–188. Springer, 2024. 7

[7] Rohit Gandikota, Hadas Orgad, Yonatan Belinkov, Joanna Materzyńska, and David Bau. Unified concept editing in diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5111–5120, 2024. 3, 7

[8] Chao Gong, Kai Chen, Zhipeng Wei, Jingjing Chen, and Yu-Gang Jiang. Reliable and efficient concept erasure of text-to-image diffusion models. In *European Conference on Computer Vision*, 2024. 7

[9] Alvin Heng and Harold Soh. Selective amnesia: A continual learning approach to forgetting in deep generative models. In *Advances in Neural Information Processing Systems*, 2023. 7

[10] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021. 3

[11] Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. *arXiv preprint arXiv:2212.04089*, 2022. 3

[12] Anubhav Jain, Yuya Kobayashi, Takashi Shibuya, Yuhta Takida, Nasir D. Memon, Julian Togelius, and Yuki Mitsufuji. Trasce: Trajectory steering for concept erasure. *CoRR*, abs/2412.07658, 2024. 7

[13] Zhiheng Liu, Ruili Feng, Kai Zhu, Yifei Zhang, Kecheng Zheng, Yu Liu, Deli Zhao, Jingren Zhou, and Yang Cao. Cones: Concept neurons in diffusion models for customized generation. *arXiv preprint arXiv:2303.05125*, 2023. 7

[14] 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*, 2023. 1, 2, 3, 7

[15] Minh Pham, Kelly O Marshall, Chinmay Hegde, and Niv Cohen. Robust concept erasure using task vectors. *arXiv preprint arXiv:2404.03631*, 2024. 3, 7

[16] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022. 3

[17] Matan Rusanovsky, Shimon Malnick, Amir Jevnisek, Ohad Fried, and Shai Avidan. Memories of forgotten concepts. *arXiv preprint arXiv:2412.00782*, 2024. 1, 2, 7

[18] Patrick Schramowski, Manuel Brack, Björn Deiseroth, and Kristian Kersting. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In *Conference on Computer Vision and Pattern Recognition*, 2023. 7

[19] SmithMano. Tutorial: How to remove the safety filter in 5 seconds, 2022. 7

[20] Vinith Menon Suriyakumar, Rohan Alur, Ayush Sekhari, Manish Raghavan, and Ashia C Wilson. Unstable unlearning: The hidden risk of concept resurgence in diffusion models. 2024. 2

[21] Viacheslav Surkov, Chris Wendler, Mikhail Terekhov, Justin Deschenaux, Robert West, and Caglar Gulcehre. Unpacking sdxl turbo: Interpreting text-to-image models with sparse autoencoders. *arXiv preprint arXiv:2410.22366*, 2024. 7

[22] Michael Toker, Hadas Orgad, Mor Ventura, Dana Arad, and Yonatan Belinkov. Diffusion lens: Interpreting text encoders in text-to-image pipelines. *arXiv preprint arXiv:2403.05846*, 2024. 7

[23] 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? *arXiv preprint arXiv:2310.10012*, 2023. 2, 7

[24] 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. *CoRR*, abs/2410.12761, 2024. 7

[25] Eric J. Zhang, Kai Wang, Xingqian Xu, Zhangyang Wang, and Humphrey Shi. Forget-me-not: Learning to forget in text-to-image diffusion models. *CoRR*, abs/2303.17591, 2023. 7

[26] Xiangyu Zhang, Joshua Kim, and Susumu Tonegawa. Amygdala reward neurons form and store fear extinction memory. *Neuron*, 105(6):1077–1093, 2020. 1

[27] Yimeng Zhang, Jinghan Jia, Xin Chen, Aochuan Chen, Yihua Zhang, Jiancheng Liu, Ke Ding, and Sijia Liu. To generate or not? safety-driven unlearned diffusion models are still easy to generate unsafe images... for now. In *European Conference on Computer Vision*, pages 385–403. Springer, 2025. 1, 2, 3, 7

Contents

1. Introduction **1**

2. A Framework for Concept Erasure **2**
2.1. Training-Free Inference Time Noising Attack . . . 2
2.2. Framework setting . . . 3
2.3. Results . . . 3

3. Guidance-Based and Destruction-Based **3**
3.1. Where Do Erased Concepts Go? . . . 4
3.2. Analysis . . . 4

A. Related Works **7**

B. Limitations **7**

C. Additional Results **8**

D. Limitations **8**

E. Training Free Inference Time Noising Attack **10**

F. Implementation Details **11**
F.1. Erasure Methods . . . 11
F.1.1 . Gradient Ascent (GA) . . . 11
F.1.2 . Erased Stable Diffusion (ESD-x & ESD-u) . . . 11
F.1.3 . Unified Concept Editing (UCE) . . . 12
F.1.4 . Task Vector (TV) . . . 12
F.2. Evaluation Protocol . . . 12
F.2.1 . CLIP Evaluation . . . 12
F.2.2 . Inversion Consistency . . . 12
F.3. Concepts . . . 13
F.4. Attack Methods . . . 13
F.4.1 . Textual Inversion Attack . . . 13
F.4.2 . Inference Time Noising Attack . . . 13
F.4.3 . UnlearnDiffAtk . . . 14
F.4.4 . Inpainting Attack . . . 14

A. Related Works

Concept erasure methods for text-to-image models. Recently, various techniques have been introduced to prevent generative models from producing images of unwanted concepts. Some work [1, 12, 18, 24] propose modifying the inference process to steer outputs away from unwanted concepts. Other methods utilize classifiers to adjust the generated results. However, since inference-guided approaches can be circumvented with sufficient access to model parameters [19], subsequent research has focused on directly updating the model weights. Pham et al. [15] apply task vectors to shift the model towards a weight space that forgets the unwanted concepts. Heng and Soh [9] utilize continual learning techniques to erase targeted concepts. Gandikota et al. [5] fine-tune the model to minimize the likelihood of generating the desirable concepts. Gandikota et al. [7] propose a closed-form expression of the weights of an erased model. Gong et al. [8] used a closed-form solution to find target embeddings of a concept which are used to update the cross-attention layers accordingly. Zhang et al. [25] suggest cross-attention re-steering to update the cross-attention maps in the UNet model of Stable Diffusion to erase concepts.

Attacks against concept erasure methods. While concept erasure methods effectively prevent undesirable generations when the concept is explicitly mentioned in the prompt (e.g., *a painting in the style of Picasso*), recent studies have demonstrated that adversarial inputs can bypass most of these defenses. In a white-box setting, Pham et al. [14] leveraged textual inversion to learn word embeddings capable of reintroducing so-called erased concepts. Similarly, Rusanovsky et al. [17] applied the same technique to learn latent seeds that reconstruct the removed concepts. Other research [2, 23, 27] has focused on directly crafting hard prompts that evade concept erasure mechanisms. Specifically, Zhang et al. [27] employed the diffusion model's zero-shot classifier to identify adversarial prompts, while Tsai et al. [23] used an evolutionary algorithm to discover them. Additionally, Chin et al. [2] optimized prompts by minimizing the distance between the diffusion trajectory and an unsafe trajectory, effectively circumventing the intended erasure.

Internal representations in diffusion models. Recent work has revealed that diffusion models encode semantic information in structured and interpretable ways. For instance, [6] demonstrated that semantic directions within the model can be effectively captured using low-rank adaptors, enabling precise continuous control. Building on this understanding, [3] showed that semantic representations are localized within specific subspaces of the model's cross-attention weights. Further investigations into the architectural components of diffusion models have yielded important insights. [13] discovered that specific concepts can be modified by targeting sparse sets of neurons for ablation. Through the application of Sparse Autoencoders (SAEs), [21] identified specialized blocks within the UNet architecture that handle distinct aspects of image generation, including composition, color manipulation, and local detail enhancement. [22] leveraged the UNet as an analytical tool to probe text encoder representations by studying how different internal representations influence the final generated outputs. While these works have advanced our understanding of diffusion models' internal representations, they primarily focus on static analysis of trained models. Our work focuses on the temporal dynamics of concept erasure during the unlearning process. Through our holistic evaluation framework, we analyze how different erasure methods distinctly affect concept representations throughout the model. We trace the evolution of concept representations during unlearning fine-tuning, revealing the dynamic nature of concept modification and providing insights into the effectiveness of various erasure techniques.

B. Limitations

Using the prompt while applying erasure. Destruction-based methods like Gradient Ascent and Task Vectors rely solely on visual examples, fine-tuning the model to avoid them without direct conditioning on a specific prompt. Prompt independence has been proposed as a potential explanation for the increased robustness of certain methods [15]. It may therefore offer a complementary perspective to the explanation suggested here: prompt-dependent methods tend to achieve guidance-based avoidance, while prompt-independent methods tend to achieve destruction-based removal. However, this raises an open question: can methods that rely on prompts be used to achieve destruction-based removal, and if so, is there any benefit to doing so?

Causality and control in concept erasure. In many cases, even the expectations for an *ideal* concept erasure algorithm remain unclear. For example, when attempting to erase an art style like Van Gogh's, should we also remove related styles, such as Edvard Munch's? This is particularly tricky when causality is involved (e.g., should erasing 'Van Gogh' cause the erasing of 'Edvard Munch' but not vice versa?). In any case, achieving this level of control is still beyond the capabilities of current methods. Nevertheless, our findings offer some guidance: destruction-based removal tends to impact related concepts more significantly than guidance-based avoidance.

Evaluting other concepts. Our study covers 13 concepts, 10 objects and 3 art styles. However, other concepts may include verbs, relationships, or abstract ideas (e.g., 'violence'). Studying such concepts is beyond the scope of this work.

CLIP Scores. Our main evaluation of adversarial robustness is highly reliant on CLIP scores (Tab.1). However, the extent

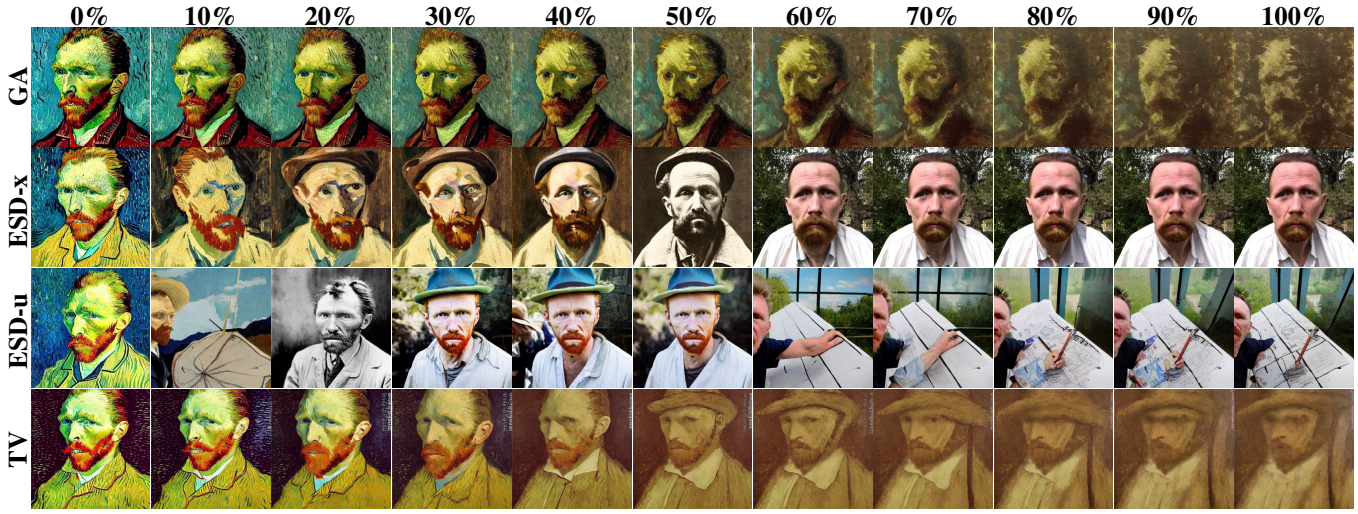

Figure 5. When comparing the concept trajectory as concepts are progressively erased more, the difference between guidance-based (ESD-u/ESD-x) and destruction-based (Gradient Ascent/Task Vector) erasure becomes visually apparent. Destruction-based methods degrade the concept itself, shifting it to a nearby concept. Guidance-based methods, on the other hand, push the concept towards the distribution of unconditional generations, creating images more diverse images.

to which something is erased may not be fully quantified by CLIP scores and can sometimes be subjective. Therefore, a more systematic evaluation of concept erasure may need to rely on more rigorous metrics that have yet to be developed.

C. Additional Results

	SD 1.4	GA	UCE	ESD-x	ESD-u	Task Vector
CLIP Score	30.20 ± 2.23	23.92 ± 2.43	26.90 ± 3.26	26.76 ± 3.05	25.91 ± 3.12	24.81 ± 2.85

Table 2. Gradient Ascent and Task Vectors show the least knowledge of erased concept even under contextual inpainting task probe. We show CLIP scores averaged across 13 different concepts.

D. Limitations

An extensive discussion of the limitations of our work can be found in Appendix B.

	Concept Inv.	Noising Du.	UnlearDiffAtk	Standard Pro.	Control Acc. Obj	Control Acc. Art
GA	23.48 ± 2.15	29.10 ± 2.93	28.74 ± 2.27	22.04 ± 2.17	25.58 ± 1.99	26.40 ± 1.82
UCE	20.19 ± 3.53	31.06 ± 2.23	31.10 ± 2.51	30.49 ± 2.18	26.40 ± 2.81	26.90 ± 2.32
ESD-X	19.78 ± 3.59	30.60 ± 2.45	30.98 ± 2.37	30.67 ± 2.56	27.65 ± 2.96	28.75 ± 2.24
ESD-U	19.86 ± 3.05	28.99 ± 3.52	30.31 ± 2.28	27.33 ± 3.48	27.33 ± 2.61	24.97 ± 2.70
Task Vector	22.49 ± 3.03	30.94 ± 2.48	30.80 ± 2.10	25.40 ± 2.69	26.14 ± 2.19	27.39 ± 1.82

Table 3. Extension of Tab.1 with the two types of unerased concepts, when erasing objects (mean ± std).

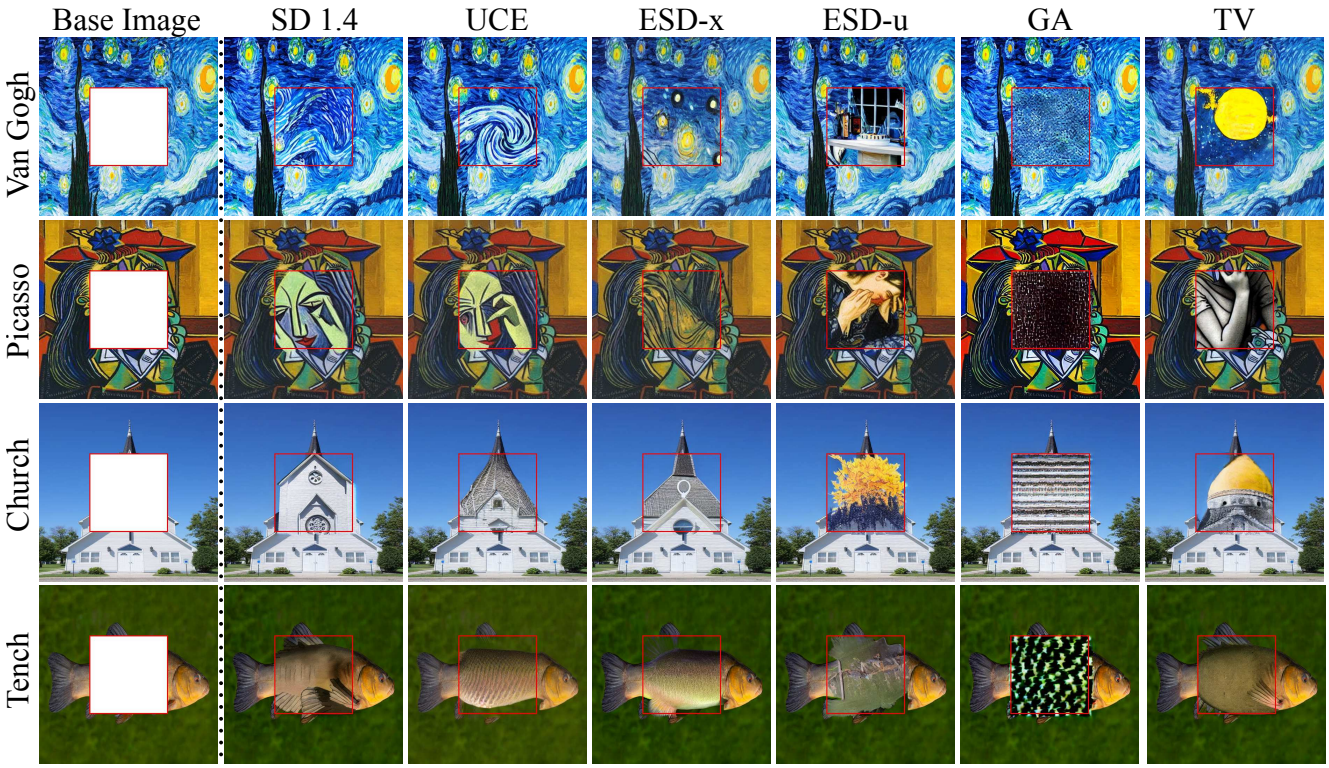

Figure 6. Inpainting serves as an initialization method, guiding the diffusion process closer to the target concept. As a result, guidance-based avoidance remains more vulnerable to such attacks, whereas destruction-based removal tends to introduce unrelated or noisy artifacts into the images.

	Concept Inv.	Noising Du.	UnlearDiffAtk	Standard Pro.	Control Acc. Obj	Control Acc. Art
GA	22.04 ± 2.17	25.58 ± 1.99	26.40 ± 1.82	23.48 ± 2.15	29.10 ± 2.93	28.74 ± 2.27
UCE	30.49 ± 2.18	26.40 ± 2.81	26.90 ± 2.32	20.19 ± 3.53	31.06 ± 2.23	31.10 ± 2.51
ESD-x	30.67 ± 2.56	27.65 ± 2.96	28.75 ± 2.24	19.78 ± 3.59	30.60 ± 2.45	30.98 ± 2.37
ESD-u	27.33 ± 3.48	27.33 ± 2.61	24.97 ± 2.70	19.86 ± 3.05	28.99 ± 3.52	30.31 ± 2.28
Task Vector	25.40 ± 2.69	26.14 ± 2.19	27.39 ± 1.82	22.49 ± 3.03	30.94 ± 2.48	30.80 ± 2.10

Table 4. Extension of Tab.1 with the two types of unerased concepts, when erasing art styles (mean ± std).

Concept	Inversion Consistency: SD 1.4	Inversion Consistency: Edited Model
Van Gogh	4.79 ± 0.09	6.12 ± 0.17
Thomas Kinkade	4.35 ± 0.16	4.06 ± 0.12
English Springer	2.58 ± 0.12	6.81 ± 0.12
Garbage Truck	2.76 ± 0.12	3.87 ± 0.13

Table 5. This table shows the prevalence of memory sink concepts in a model before and after erasure (mean $\pm 95\%$ CI). We show that the Inversion Consistency of memory sink concepts is comparable before and after editing the models. This supports our hypothesis that the memory sink concept generated by the edited model was likely already present in the original model with high likelihood. Lower Inversion Consistency scores indicate that the distances between different generations are smaller, meaning the generations are more consistent. This suggests a higher likelihood associated with the memory sing concept.

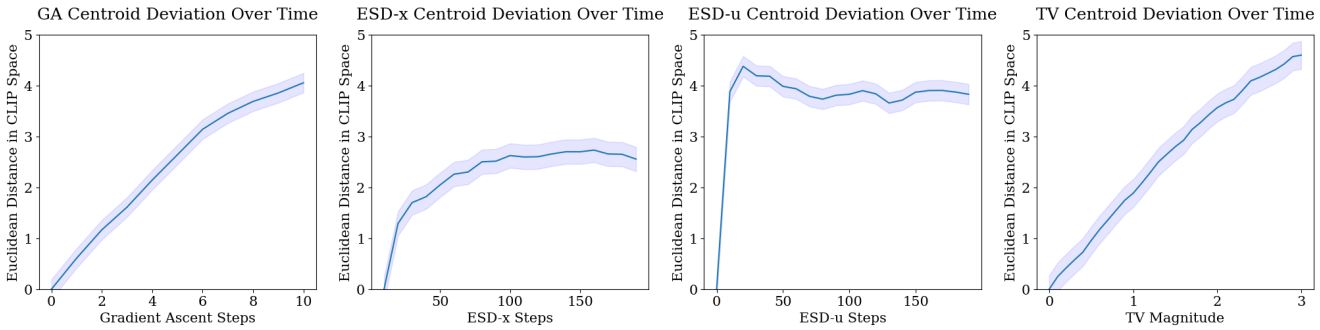

Figure 7. We plot the centroids of the CLIP embeddings of diffusion model generations for particular concepts as they are being erased. For each method, we show the strength of erasure on the X-axis and the centroid's Euclidean distance from the original concept's embedding on the Y-axis. The shaded regions are 95% confidence intervals for the concept's distance from its original location. In the plots, we see a difference between the more linear concept trajectory of destruction-based methods (GA, TV) and the sublinear concept trajectory of guidance-based methods (ESD-x, ESD-u). Destruction-based erasure pushes the generation away while guidance-based erasure maps it to unconditional generation without changing it much for a more for stronger edits.

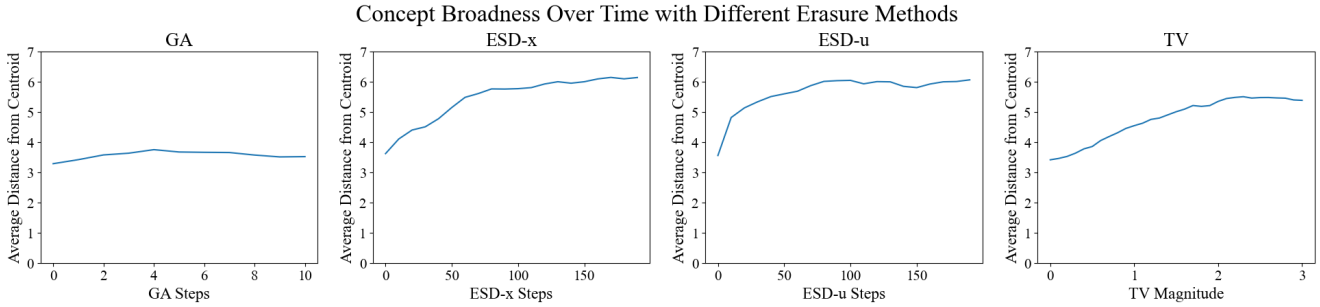

Figure 8. We plot the average distance of embeddings from their centroid at a particular step. For each method, we show the strength of erasure on the X-axis and the average of the embeddings' distances from their centroid on the Y-axis. This figure compares how the broadness in the space of possible generations for a given concept change as the target concept is erased. We show that, for GA, the space of possible generations remains tight even as the concept itself is destroyed, meaning that the generations are drawn from a new high-likelihood area, a memory sink. For ESD-x and ESD-u, the generations broaden as the concept is pushed towards the null distribution. TVs still demonstrate concept broadening but remain tighter than the guidance-based avoidance methods.

E. Training Free Inference Time Noising Attack

Song et al. introduced Denoising Diffusion Implicit Models (DDIM), which presented a deterministic generative process defined by the equation:

$$x_{t-1} = \sqrt{\alpha_{t-1}} \left(\frac{x_t - \sqrt{1-\alpha_t}\epsilon_\theta^{(t)}(x_t)}{\sqrt{\alpha_t}} \right) + \underbrace{\sqrt{1-\alpha_{t-1}-\sigma_t^2} \cdot \epsilon_\theta^{(t)}(x_t)}\text{“direction pointing to } x_t\text{”} + \underbrace{\sigma_t \epsilon_t}\text{ random noise} \tag{2}$$

We observe that the random noise term acts as a brownian motion component, driving stochastic sample generation when $\sigma_t > 0$. This insight motivates our approach: by controlling the magnitude of this brownian motion, we can systematically explore a broader latent space of the diffusion model. We modify the DDIM formulation by introducing a scaling factor η that controls the random noise component:

$$x_{t-1} = \sqrt{\alpha_{t-1}} \left(\frac{x_t - \sqrt{1-\alpha_t}\epsilon_\theta^{(t)}(x_t)}{\sqrt{\alpha_t}} \right) + \sqrt{1-\alpha_{t-1}-\sigma_t^2} \cdot \epsilon_\theta^{(t)}(x_t) + \underbrace{\eta}_{\text{scale to control random noise}} \sigma_t \epsilon_t \tag{3}$$

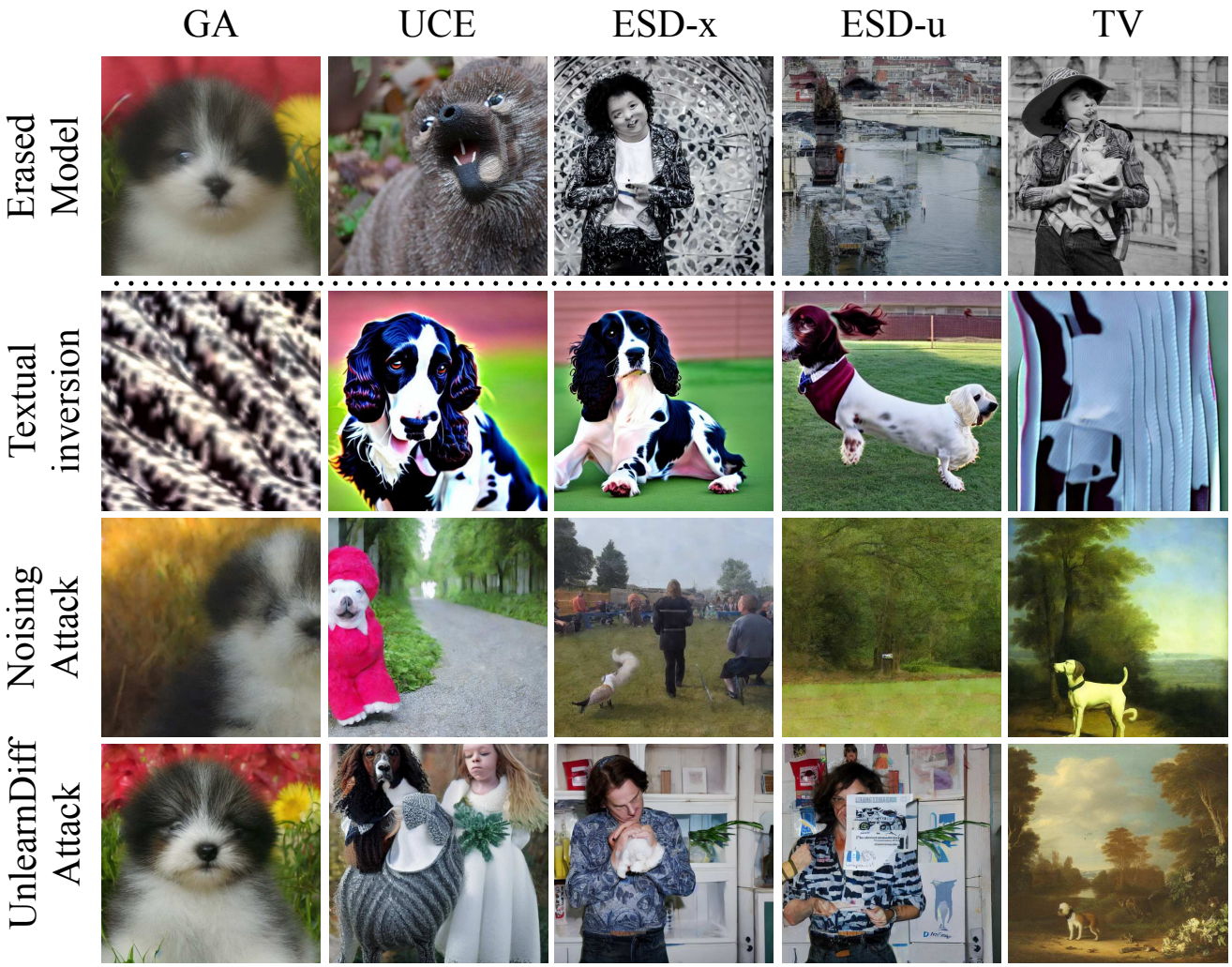

Figure 9. Visualization of image generation by models that erased "English springer spaniel". We believe the unique dog of the GA images represents a "memory sink", redirecting diffusion paths towards this concept despite attack methods attempting to recover English Springer spaniels.

F. Implementation Details

F.1. Erasure Methods

To evaluate the impact of different concept erasure techniques, we implemented several existing methods and trained models under controlled settings. Below, we detail the exact configurations for each approach:

F.1.1. Gradient Ascent (GA)

We applied gradient ascent by simply flipping the sign of the standard training loss of Stable Diffusion. The training images consist of pre-generated images from the original model and their corresponding prompts. For each concept, we used 500 diverse images and fine-tuned for 60 steps, except for English Springer Spaniel and Garbage Truck where we used 10 steps. One thing to note about GA is that it can easily break the model's overall utility if trained for too many iterations. Hence, we used batch size of 5 with gradient accumulation step of 4, and a learning rate of 1×10^{-5}.

F.1.2. Erased Stable Diffusion (ESD-x & ESD-u)

We fine-tuned for 200 steps using a learning rate of 2×10^{-5}. For each concept, we used 2000 training images.

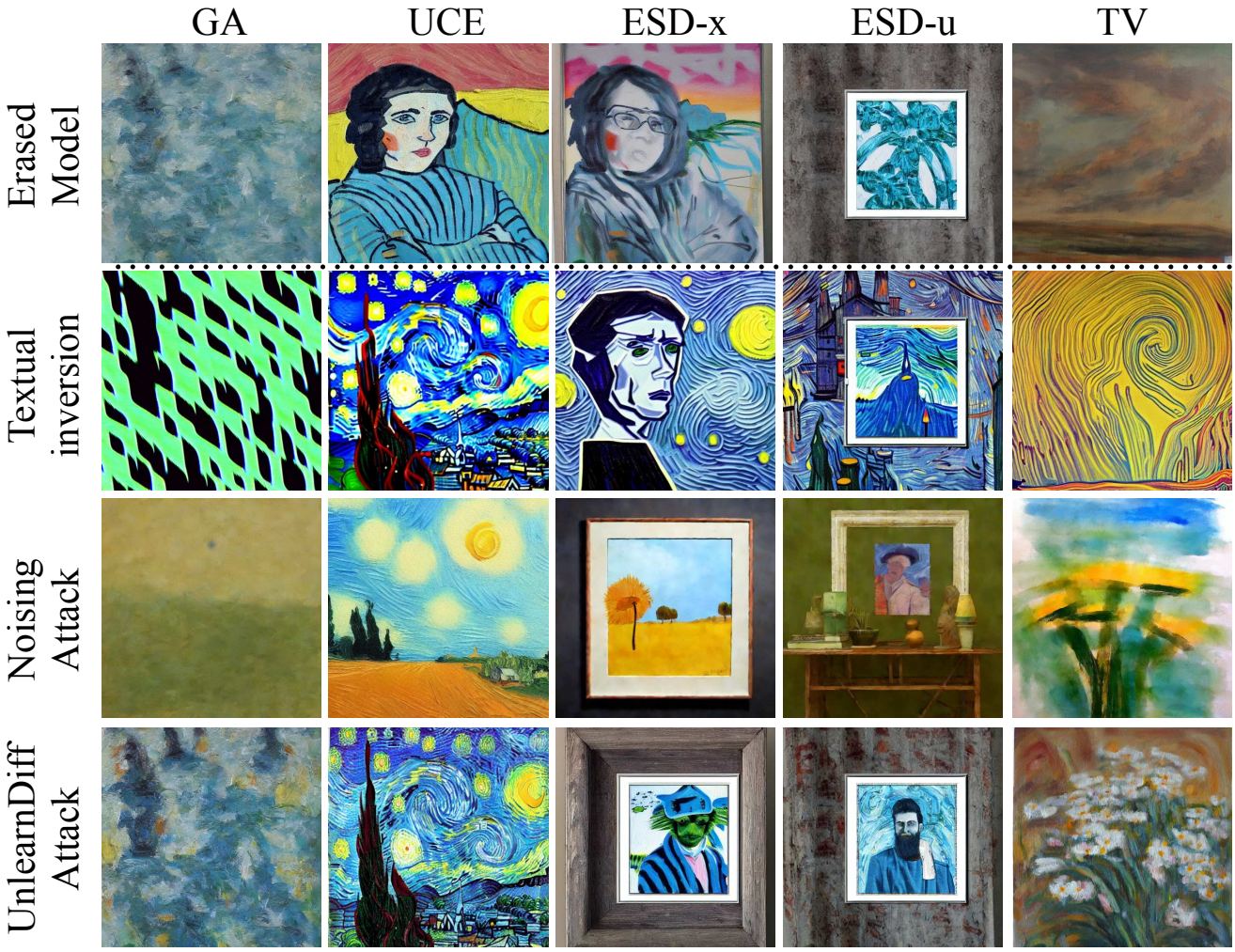

Figure 10. Erasing "Van Gogh" concept using different methods.

F.1.3. Unified Concept Editing (UCE)

We fine-tuned for 200 steps with an empty guiding concept and an erase scale of 1.

F.1.4. Task Vector (TV)

To get the fine-tuned model for computing task vectors, we fine-tuned each model on 500 images for 200 steps, using a learning rate of 1×10^{-5}. We used batch size of 4 and gradient accumulation step of 4. For erasure, we set the editing strength $\alpha = 1.75$.

F.2. Evaluation Protocol

F.2.1. CLIP Evaluation

All similarity assessments were performed using CLIP ViT (`openai/clip-vit-base-patch32`).

F.2.2. Inversion Consistency

First, we determined a ground truth CLIP embedding for the memory sink concept. Next, after inverting the memory sink concept in both the original and edited models, we generated images from both models using the same 100 seeds per concept. We then measured the average Euclidean CLIP distance of these embeddings from the ground truth memory sink concept to determine how closely they replicated the target concept.

Figure 11. Erasing "Picasso" concept using different methods.

F.3. Concepts

We consider a set of 10 object concepts: English Springer Spaniel, airliner, garbage truck, parachute, cassette player, chainsaw, tench, French horn, golf ball, and church; alongside 3 distinct art styles: Van Gogh, Picasso, and Andy Warhol. This selection allows us to evaluate the impact of concept erasure across both tangible objects and artistic styles, ensuring a diverse range of visual and semantic attributes in our analysis.

F.4. Attack Methods

To assess the resilience of erasure methods against adversarial strategies, we conducted various attack experiments using a dataset of 100 prompts spanning 13 concepts (10 objects, 3 styles), each evaluated using unique seeds.

F.4.1. Textual Inversion Attack

Training involved 100 images, optimized for 3000 steps using a learning rate of 5×10^{-4}.

F.4.2. Inference Time Noising Attack

Seven different noise scaling values (ηs) were used: 1.3, 1.5, 1.7, 1.75, 1.8, 1.85, and 1.9. Additionally, six different variance scales were applied: 0.97, 1.0, 1.02, 1.03, 1.035, and 1.04. A full grid search over the 42 combinations was conducted, generating 42 samples per prompt/experiment. The CLIP model then selected the generated image with the highest similarity score as the most effective attack instance.

F.4.3. UnlearnDiffAtk

The model was trained using a learning rate of 0.01 and a weight decay of 0.1, with the classifier parameter set to $K = 3$. ImageNet was used as the classifier for object-based erasures, while a custom classifier from the UnlearnDiffAtk repository was used for artist styles. Due to computational cost, UnlearnDiffAtk was evaluated on 10 prompts per concept, with 40 samples per experiment, where each sample was generated through 40 optimization steps.

F.4.4. Inpainting Attack

The inpainting pipeline was based on Stable Diffusion 1.5 and implemented via Hugging Face's `StableDiffusionInpaintPipeline`. Base images were 512×512 pixels and were masked with a 225×225 white box at the center. Source images were generated using Stable Diffusion 1.4. CLIP scores were computed only on the masked area to prevent artificially inflated similarity scores.