

# STEREO: Towards Adversarially Robust Concept Erasing from Text-to-Image Generation Models

Koushik Srivatsan<sup>1</sup>, Fahad Shamshad<sup>1</sup>, Muzammal Naseer<sup>1</sup>, Karthik Nandakumar<sup>1</sup>

<sup>1</sup>Mohamed bin Zayed University of Artificial Intelligence,  
 {firstname.lastname}@mbzuai.ac.ae

## Abstract

The rapid proliferation of large-scale text-to-image generation (T2IG) models has led to concerns about their potential misuse in generating harmful content. Though many methods have been proposed for erasing undesired concepts from T2IG models, they only provide a false sense of security, as recent works demonstrate that concept-erased models (CEMs) can be easily deceived to generate the erased concept through adversarial attacks. The problem of adversarially robust concept erasing without significant degradation to model utility (ability to generate benign concepts) remains an unresolved challenge, especially in the white-box setting where the adversary has access to the CEM. To address this gap, we propose an approach called **STEREO** that involves two distinct stages. The first stage searches thoroughly enough for strong and diverse adversarial prompts that can regenerate an erased concept from a CEM, by leveraging robust optimization principles from adversarial training. In the second robustly erase once stage, we introduce an anchor-concept-based compositional objective to robustly erase the target concept at one go, while attempting to minimize the degradation on model utility. By benchmarking the proposed **STEREO** approach against four state-of-the-art concept erasure methods under three adversarial attacks, we demonstrate its ability to achieve a better robustness vs. utility trade-off. Our code and models are available at <https://github.com/koushiksrivats/robust-concept-erasing>.

**WARNING:** This paper contains model-generated content that might be considered offensive or inappropriate. Reader discretion is advised.

## 1 Introduction

Large-scale text-to-image generation (T2IG) models (Chang et al. 2023; Ding et al. 2022; Lu, Liu, and Kong 2023; Nichol et al. 2022) have demonstrated a remarkable ability to synthesize photorealistic images from user-specified text prompts, leading to their adoption in numerous commercial applications. However, these models are typically trained on massive datasets scraped from the Internet (Schuhmann et al. 2022). This can result in issues such as memorization (Ren et al. 2024; Somepalli et al. 2023) and generation of inappropriate images (e.g., copyright violations (Jiang et al. 2023; Roose 2022), prohibited content (Schramowski et al. 2023), and NSFW material (Hunter 2023; Zhang et al. 2023b)). Public-domain availability of T2IG models such as Stable Diffusion

(SD) (Rombach et al. 2022) raises significant security concerns that require urgent redressal.

Solutions to mitigate the generation of undesired concepts in T2IG models can be broadly classified into three categories: dataset filtering before training, post-hoc modification of models after training, and output filtering after image generation. Dataset filtering (Carlini et al. 2022) involves removing unsafe images and retraining the model, which is not only computationally expensive but also impractical for each new undesired concept and may have a significant negative impact on the output quality (Schramowski et al. 2023). While post-generation output filtering can effectively censor harmful images, it can be applied only to the black-box setting, where the adversary has only query access to the T2IG model and cannot access the model parameters (Rando et al. 2022). Recently, several post-hoc erasure methods have been proposed to modify the behavior of pre-trained T2IG models. These methods either fine-tune the model parameters or alter the generation process during inference to avoid the generation of undesired concepts (Schramowski et al. 2023; Brack et al. 2023; Gandikota et al. 2023; Kumari et al. 2023). This work focuses on post-hoc concept erasure methods, which are often more practical and effective.

Despite the success of post-hoc erasure methods, recent studies (Pham, Marshall, and Hegde 2023; Tsai et al. 2023; Chin et al. 2023; Zhang et al. 2023b) have exposed significant vulnerabilities of this approach by demonstrating that concept erasure can be easily circumvented through adversarial attacks. Such attacks are crafted by either prepending/modifying input prompts with adversarial tokens (Chin et al. 2023; Zhang et al. 2023b; Tsai et al. 2023) or injecting concepts into textual embeddings (Pham, Marshall, and Hegde 2023). Consequently, seemingly safe concept-erased T2IG models still produce sensitive or offensive content under adversarial attack, as shown in Figure 1. To address this limitation, it is essential to incorporate some form of adversarial training (AT) into the concept erasing method (Huang et al. 2024; Kim, Min, and Yang 2024).

This work proposes a two-stage framework for adversarially robust concept erasing from T2IG models. The first stage follows the robust optimization framework of AT and formulates concept erasing as a min-max optimization problem, which is iteratively solved by alternating between erasing the target concept in the parameter space of the pre-trained



**Figure 1: Concept erasure methods are prone to adversarial concept inversion attacks.** This figure illustrates that recent concept erasure methods (ESD (Gandikota et al. 2023), UCE (Gandikota et al. 2024), MACE (Lu et al. 2024)) are vulnerable to concept inversion attacks (CCE (Pham, Marshall, and Hegde 2023)) that regenerate the erased concept. In contrast, our proposed approach STEREO is robust against such attacks across diverse concept categories.

model and searching for adversarial prompts in the text embedding space that can still regenerate the erased concept from the modified model. The core novelty of our approach lies in the fact that we employ AT not as a final solution, but only as an intermediate step to *search thoroughly enough* for strong adversarial prompts. This adversarial prompt search is followed by a *robustly erase once* stage, where we propose an anchor-concept-based compositional objective to erase the target concept from the original model. While incorporating the anchor concept in the erasing objective minimizes the degradation of model utility, the compositional guidance steers the final erased model away from the set of strong adversarial prompts, thereby enhancing adversarial robustness. Our main **contributions** can be summarized as follows:

- We propose a novel two-stage approach called STEREO to achieve better robustness vs. utility trade-off when robustly erasing concepts from pre-trained T2IG models.
- While the first *search thoroughly enough* (STE) stage utilizes adversarial training to discover strong adversarial prompts that can regenerate target concepts from erased models, we propose an anchor-concept-based compositional objective in the second *robustly erase once* (REO) stage to erase the target concept from the original model, while mitigating the loss of utility.

## 2 Related Work and Background

**Post-hoc Concept Erasing:** Recent methods for erasing undesired concepts from pre-trained T2IG models can be categorized into inference-based and fine-tuning-based approaches. *Inference-based methods* (Schramowski et al. 2023; Brack et al. 2023; AUTOMATIC1111 2022; Dong et al. 2024) modify the noise-estimate calculated through classifier-free guidance (CFG) (Ho and Salimans 2022), to navigate the generation away from the undesired concepts semantically, without any additional training cost. These methods introduce additional terms to the CFG during inference, such as replacing the null-string in the unconditioned branch with a textual prompt describing the undesired concept (AUTOMATIC1111 2022), incorporating safety (Schramowski et al. 2023), us-

ing semantic guidance (Brack et al. 2023) or feature space purification (Dong et al. 2024), to move the unconditioned score estimate closer to the prompt-conditioned score estimate. *Fine-tuning-based methods* modify the parameters of the pre-trained model to remap the undesired concept’s noise estimate semantically away from the original concept (Gandikota et al. 2023; Heng and Soh 2024) or to a predefined desired target (Zhang et al. 2023a; Lu et al. 2024). This work primarily uses (Gandikota et al. 2023), which generates images with an unwanted concept and then guides the model away from creating such content. Despite the impressive performance of concept-erasing methods, they are vulnerable to adversarial prompt attacks that can regenerate the erased concept (Pham, Marshall, and Hegde 2023; Tsai et al. 2023; Chin et al. 2023).

**Circumventing Concept Erasing:** Among recent attacks on concept erasing methods, the most relevant to our work is Circumventing Concept Erasure (Pham, Marshall, and Hegde 2023), which shows that the erased concept can be mapped to any arbitrary input word embedding through textual-inversion (Gal et al. 2022). Optimizing for this new embedding without altering the weights of the erased model steers the generation to output the erased concept. Prompting4Debugging (Chin et al. 2023) optimizes adversarial prompts by enforcing similarity between the noise estimates of pre-trained and concept-erased models. Unlearn-Diff (Zhang et al. 2023b) simplifies adversarial prompt creation by leveraging the intrinsic classification abilities of diffusion models. Similarly, Ring-A-Bell (Tsai et al. 2023), generates problematic prompts to bypass safety mechanisms in text-to-image diffusion models, leading to the generation of images with supposedly forbidden concepts.

**Adversarially Robust Concept Erasing:** Recently, few approaches have been proposed for adversarial training-based robust concept erasure. Receler (Huang et al. 2024) employs an iterative approach, alternating between erasing and adversarial prompt learning. Our STEREO method differs by using a two-stage approach with explicit min-max optimization for adversarial prompts, offering protection in white-box settings. AdvUnlearn (Zhang et al. 2024) proposes bilevel optimiza-

tion but requires curated external data to preserve utility. In contrast, STEREO uses a compositional objective with adversarial prompts without the need for external data. RACE (Kim, Min, and Yang 2024) focuses on computationally efficient adversarial training using single-step textual inversion, but at the cost of utility. Most current robust concept erasure methods evaluate on discrete attacks (UnlearnDiff (Zhang et al. 2023b) and RAB (Tsai et al. 2023)) with limited prompt token modifications. Our work additionally evaluates on the CCE attack (Pham, Marshall, and Hegde 2023), which has a larger, unconstrained search space, presenting a more challenging defense scenario.

## 2.1 Preliminaries

**Latent Diffusion Models (LDMs):** We implement our method using Stable Diffusion (Rombach et al. 2022), a state-of-the-art LDM for T2IG. LDMs are denoising-based probabilistic models that perform forward and reverse diffusion processes in the low ( $d$ )-dimensional latent space  $\mathcal{Z} \subseteq \mathbb{R}^d$  of a pre-trained variational autoencoder. An LDM comprises of two main components: an **autoencoder** and a **diffusion model**. The **autoencoder** includes an encoder ( $\mathcal{E} : \mathcal{X} \rightarrow \mathcal{Z}$ ) that maps image  $x \in \mathcal{X}$  ( $\mathcal{X}$  denotes the image space) to latent codes  $z = \mathcal{E}(x) \in \mathcal{Z}$  and a decoder ( $\mathcal{D} : \mathcal{Z} \rightarrow \mathcal{X}$ ) that reconstructs images from latent codes, ensuring  $\mathcal{D}(\mathcal{E}(x)) \approx x$ . The **diffusion model** is trained to produce latent codes within the learned latent space through a sequence of denoising steps. It consists of an UNet-based noise predictor  $\epsilon_\theta(\cdot)$ , which predicts the noise  $\epsilon$  added to  $z_t$  at each timestep  $t$ . In T2IG, the diffusion model is additionally conditioned on text prompts  $p \in \mathcal{T}$  ( $\mathcal{T}$  denotes the text space), encoded by a jointly trained **text encoder**  $\mathcal{Y}_\psi : \mathcal{T} \rightarrow \mathcal{P}$  ( $\mathcal{P}$  denotes the text embedding space). The training objective of LDM is given by:

$$\mathcal{L}_{\text{LDM}} = \mathbb{E}_{z_t \sim \mathcal{E}(x), t, p, \epsilon \sim \mathcal{N}(0, 1)} [\|\epsilon - \epsilon_\theta(z_t, t, \mathcal{Y}_\psi(p))\|_2^2]. \quad (1)$$

To minimize this objective,  $\theta$  and  $\psi$  are optimized jointly. Thus, the complete T2IG model can be denoted as  $f_\phi : \mathcal{T} \rightarrow \mathcal{X}$ , where  $f_\phi := \{\mathcal{E}, \mathcal{D}, \epsilon_\theta, \mathcal{Y}_\psi\}$ . During inference, CFG directs the noise at each step toward the desired text prompt  $p$  as  $\tilde{\epsilon}_\theta(z_t, t, \mathcal{Y}_\psi(p)) = \epsilon_\theta(z_t, t) + \alpha(\epsilon_\theta(z_t, t, \mathcal{Y}_\psi(p)) - \epsilon_\theta(z_t, t))$ , where the guidance scale  $\alpha > 1$ . The inference process starts from a Gaussian noise  $z_T \sim \mathcal{N}(0, 1)$  and is iteratively denoised using  $\tilde{\epsilon}_\theta(z_t, t, \mathcal{Y}_\psi(p))$  to obtain  $z_{T-1}$ . This process is done sequentially until the final latent code  $z_0$  is obtained, which in turn is decoded into an image  $x_0 = \mathcal{D}(z_0)$ . Thus,  $x_0 = f_\phi(p)$ .

**Compositional Inference.** Compositional inference in T2IG models refers to the process of generating new samples by combining and manipulating the learned representations of multiple concepts (Liu et al. 2022). The objective function for compositional inference is given by:

$$\tilde{\epsilon}_\theta(z_t, t) = \epsilon_\theta(z_t, t) + \sum_{j=1}^N \eta_j (\epsilon_\theta(z_t, t, \mathcal{Y}_\psi(p_j)) - \epsilon_\theta(z_t, t)), \quad (2)$$

where  $N$  denotes the number of concepts and  $\eta_j$  is the guidance scale for concept  $c_j$  (which is expressed as prompt  $p_j$ ),  $j \in [N]$ . Note that  $\eta$  should be positive for the target concept and negative for undesired concepts.

## 3 Proposed Methodology

### 3.1 Problem Statement

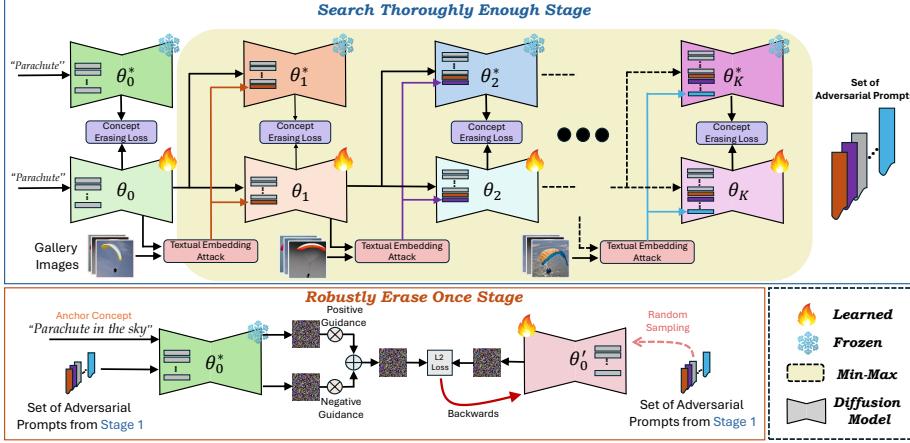
Let  $f_\phi$  be a pre-trained T2IG model that generates an image  $x_0$  based on the input text prompt  $p$ . Let  $\mathcal{C}$  denote the concept space. The goal of vanilla concept erasing is to modify the given T2IG model such that the concept erased model (CEM)  $\tilde{f}_\phi$  does not generate images containing the undesired/target concept  $c_u \in \mathcal{C}$ , when provided with natural text prompts directly expressing the target concept (e.g., nudity) or simple paraphrased versions of it (e.g., a person without clothes). This work deals with *adversarially robust concept erasing*, which aims to modify the given T2IG model such that the CEM  $\tilde{f}_\phi$  does not generate images containing the undesired concept even when prompted using malicious prompts (either directly from the text space  $\mathcal{T}$  or from the text embedding space  $\mathcal{P}$ ). Note that the malicious prompts may or may not explicitly contain the target concept. Furthermore, the CEM should be able to generate images depicting benign/non-target concepts (those that have not been erased) with the same fidelity as the original T2IG model.

Let  $\mathbb{O}_\mathcal{X} : \mathcal{X} \times \mathcal{C} \rightarrow \{0, 1\}$  and  $\mathbb{O}_\mathcal{T} : \mathcal{T} \times \mathcal{C} \rightarrow \{0, 1\}$  be ground-truth oracles that verify the presence of a concept  $c \in \mathcal{C}$  in an image and in a text prompt, respectively.  $\mathbb{O}_\mathcal{X}(x, c) = 1$  if concept  $c$  is depicted in image  $x$  (and 0, otherwise). Similarly,  $\mathbb{O}_\mathcal{T}(p, c) = 1$  if concept  $c$  is expressed in prompt  $p$  (and 0, otherwise). The *concept generation ability* of a T2IG model can be quantified as  $\mathcal{A}(c) = \mathbb{P}_{p \sim \mathcal{T}}([\mathbb{O}_\mathcal{X}(f_\phi(p), c) = 1] | [\mathbb{O}_\mathcal{T}(p, c) = 1])$ , where  $\mathbb{P}$  denotes a probability measure. In other words, the T2IG model should faithfully generate images with a concept  $c$ , if the concept is present in the input text prompt  $p$ . The *utility* of the T2IG model can be defined as  $\mathcal{U} = \mathbb{E}_{c \sim \mathcal{C}} \mathcal{A}(c)$ . An ideal CEM should satisfy the following three properties: (1) **Effectiveness** - quantified as  $\tilde{\mathcal{A}}(c_u) = 1 - \mathbb{P}_{p \sim \mathcal{T}}([\mathbb{O}_\mathcal{X}(\tilde{f}_\phi(p), c_u) = 1] | [\mathbb{O}_\mathcal{T}(p, c_u) = 1])$ , which should be as high as possible for the CEM  $\tilde{f}_\phi$ . (2) **Robustness** - defined as  $\tilde{\mathcal{R}}(c_u) = 1 - \mathbb{P}_{p^* \sim \mathcal{T}}([\mathbb{O}_\mathcal{X}(\tilde{f}_\phi(p^*), c_u) = 1])$ , where  $p^*$  denotes an adversarial prompt. (3) **Utility preservation** - the utility of the CEM, which is defined as  $\tilde{\mathcal{U}}(c_u) = \mathbb{E}_{c \sim \mathcal{C} \setminus \{c_u\}} \mathcal{A}(c)$ , should be as close as possible to  $\mathcal{U}$ .

Thus, given a pre-trained T2IG model  $f_\phi$  and an undesired concept  $c_u$ , the problem of adversarially robust concept erasing can be formally stated as follows: maximize both  $\tilde{\mathcal{A}}(c_u)$  (effectiveness) and  $\tilde{\mathcal{R}}(c_u)$  (robustness), while maintaining high utility  $\tilde{\mathcal{U}}(c_u)$ . Achieving all three objectives simultaneously is challenging, as they are inherently related and often conflicting. For instance, aggressive concept removal may lead to a significant loss in utility, while being over-cautious may compromise effectiveness and robustness. Striking the right balance between these objectives is critical for developing a good concept erasing method.

### 3.2 The STEREO Approach

To robustly and effectively remove an undesired concept from a pre-trained T2IG model while preserving high utility, we propose a two-stage approach as illustrated in Fig. 2.



**Figure 2: Overview of STEREO.** Our novel two-stage approach robustly erases target concepts from pre-trained text-to-image generation models while preserving high utility for benign concepts. **Stage 1 (top):** *Search Thoroughly Enough* fine-tunes the model through iterative concept erasing and concept inversion attacks, ensuring resilience against adversarial regeneration attempts. **Stage 2 (bottom):** *Robustly Erase Once* fine-tunes the model using anchor concept and the set of strong adversarial prompts from Stage 1 via a compositional objective, maintaining high-fidelity generation of benign concepts while robustly erasing the target concept.

**Search Thoroughly Enough (STE) Stage:** The goal of this stage is to discover a set of strong adversarial prompts that can regenerate the erased concept from the CEM. To achieve this goal, we draw inspiration from the success of adversarial training (AT) in enhancing the robustness of classification models (Madry et al. 2017). To effectively find these adversarial prompts, we formulate the task as a min-max optimization problem, aiming to minimize the probability of generating images containing the undesired concept by modifying the T2IG model, while simultaneously finding adversarial prompts that maximize the probability of generating undesired images. Formally, the task objective is defined as:

$$\min_{\phi} \max_{p^*} \mathbb{P}([\mathbb{O}_{\mathcal{X}}(f_{\phi}(p^*), c_u) = 1]), \quad (3)$$

where the probability  $\mathbb{P}$  is defined over the stochasticity of  $z_T$ , which represents the Gaussian noise used to initialize the inference process. To solve this problem, we employ an iterative approach that alternates between two key steps: (1) **Minimization** - erasing the target concept in the *parameters space* of the pre-trained T2IG model (specifically by altering the UNet parameters  $\theta$ ), and (2) **Maximization** - searching for adversarial prompts in the *text embedding space* that can regenerate the erased concept from the altered model.

**Minimization Step:** At each step  $i$  of minimization, we aim to erase the target concept  $c_u$  from the current UNet model  $\epsilon_{\theta_i}$  using its inherent knowledge preserved in  $\theta_i$ . Specifically, we create a copy of parameters of  $\epsilon_{\theta_i}$  denoted as  $\theta_i^*$ , and keep  $\theta_i^*$  frozen while fine-tuning  $\theta$  with guidance from  $\theta_i^*$ . The fine-tuning process aims to minimize the probability of generating an image  $x_0 \in \mathcal{X}$  that includes an undesired concept  $c_u$ . To achieve this, we compute the negative-guidance noise estimate (Ho and Salimans 2021; Gandikota et al. 2023) to redirect the predicted noise away from a text prompt  $p_u$  (containing  $c_u$ ) using:  $\tilde{\epsilon}_{\theta_i^*}(z_t, t, \mathcal{Y}_{\psi}(p_u)) \leftarrow \epsilon_{\theta_i^*}(z_t, t) - \eta(\epsilon_{\theta_i^*}(z_t, t, \mathcal{Y}_{\psi}(p_u)) - \epsilon_{\theta_i^*}(z_t, t))$ , where  $\eta$  is the negative-guidance strength. The negative guidance is computed using the frozen parameters  $\theta_i^*$ , which acts as the ground truth to fine-tune  $\theta_i$  at every timestep  $t$ , to ensure the minimization of the concept erasing objective:

$$\mathcal{L}_{CE} = \mathbb{E}_{z_t \in \mathcal{E}(x), t, p_u} [\|\epsilon_{\theta_i}(z_t, t, \mathcal{Y}_{\psi}(p_u)) - \tilde{\epsilon}_{\theta_i^*}(z_t, t, \mathcal{Y}_{\psi}(p_u))\|_2^2]. \quad (4)$$

In this way, the conditional prediction of the fine-tuned model  $\epsilon_{\theta_i}(z_t, t, \mathcal{Y}_{\psi}(p_u))$  is progressively guided away from the undesired concept  $c_u$  at each minimization step.

**Maximization Step:** While the minimization step aims to remove the undesired concept  $c_u$ , the maximization step identifies malicious prompts  $p^*$  that challenge the model’s robustness. (Yang et al. 2024) show that there may be alternative mappings that can regenerate  $c_u$ . A naive approach to find these alternative mappings would be to collect synonymous prompts of the concept and incorporate them into the erasing objective of Eq. 4 during the minimization step. This can be achieved by randomly conditioning either the original prompt or its synonym in the erasing objective at every iteration, aiming to minimize the effect of both representations. However, as demonstrated in Fig. 3 (left), this naive approach is ineffective, as the model remains vulnerable to attacks due to the lack of diverse and optimal alternate concept representations.

To address this limitation, we use a textual inversion-based (Gal et al. 2022) maximization step to effectively identify adversarial prompts. At each maximization step  $i$ , we search for an adversarial prompt  $p_i^*$  in the text embedding space of the frozen T2IG model that can reintroduce the erased concept  $c_u$ . This is achieved by encoding the undesired visual concept into the text embedding space via the introduction of a new token  $s_i^*$  into the existing vocabulary, specifically designed to represent  $c_u$ . Each token in the vocabulary corresponds to a unique embedding vector, and our goal is to find the optimal embedding vector  $v_i^*$  for  $s_i^*$  that effectively captures the characteristics of  $c_u$ . For this, we utilize a pre-generated gallery set  $\mathcal{G}$  (using the original T2IG model) depicting the target concept and obtain  $v_i^*$  as:

$$v_i^* = \operatorname{argmin}_v \mathbb{E}_{z_t \in \mathcal{E}(x), x \sim \mathcal{G}, t, p, \epsilon \sim \mathcal{N}(0, 1)} [\|\epsilon_i - \epsilon_{\theta_i}(z_t, t, [\mathcal{Y}_{\psi}(\hat{p})] \parallel v)\|_2^2], \quad (5)$$

where  $\epsilon_i$  denotes the unscaled noise sample added at time step  $t$ , and  $[\mathcal{Y}_{\psi}(\hat{p})] \parallel v$  denotes the appending of the new embedding  $v$  to the embeddings of the existing vocabulary represented by  $\mathcal{Y}_{\psi}(\hat{p})$ . The optimized embedding  $v_i^*$  becomes the representation of the token  $s_i^*$ , and any prompt  $p_i^*$  that includes  $s_i^*$  can be considered an adversarial prompt.

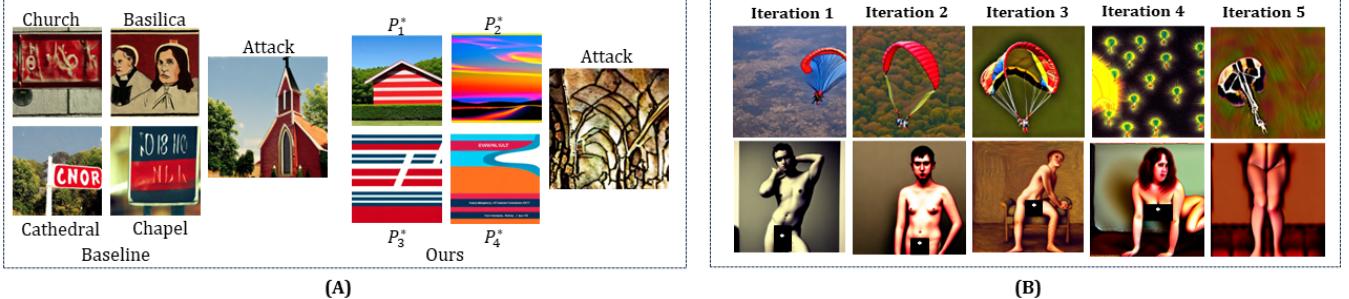


Figure 3: (A): Using only concept synonyms (Baseline) erases the concept but is vulnerable to attacks, as the "Church" concept is regenerated under the CCE (Pham, Marshall, and Hegde 2023) attack. The proposed STEREO approach identifies strong adversarial prompts  $p^*$ , which eventually facilitate robust erasing of the concept and make the concept erased model resistant against inversion attacks. (B): Regenerated image corresponding to the adversarial prompt learnt at each maximization step.

Fig 3 (right) depicts the inverted image corresponding to the adversarial prompt at each maximization step. The adversarial prompt  $p_i^*$  is then incorporated into the subsequent minimization step, and the process continues for  $K$  iterations. *The STE stage thus identifies a set of strong and diverse adversarial prompts at the end of  $K$  min-max iterations:*  $\mathbf{p}_K^* = \{p_u, p_1^*, \dots, p_i^*, \dots, p_K^*\}$ .

**Robustly Erase Once (REO) Stage:** Though the final erased UNet parameters  $\epsilon_{\theta_K}$  at the end of the STE stage lead to a highly robust CEM, the iterative erasing process greatly degrades the model utility. Instead of using ad-hoc methods to regularize the erasing process (Zhang et al. 2024) and preserve model utility, we propose an alternative approach that uses the set of adversarial prompts  $\mathbf{p}_K^*$  and robustly erases the target concept at one go. A naive way to implement this idea would be to incorporate the set of adversarial prompts  $\mathbf{p}_K^*$  into one of the baseline erasing objectives. Randomly sampling one adversarial prompt from this set as the prompt condition, at every fine-tune iteration, forces the objective to minimize the influence on each of these prompts. However, as we demonstrate in Table 1, this affects the utility when using only negative guidance (ESD (Gandikota et al. 2023)) or increases the attack success rate when using only positive guidance (AC (Kumari et al. 2023)). This is because using only negative guidance moves the model away from the target without any regularization (FID scores go from 14.13 to 38.06) and using only positive guidance naively remaps each new word to a pre-defined target and thus not fully erasing the undesired concept (high ASR of 86.31). To preserve the model’s utility while maintaining its robustness, we observe that additionally guiding the model towards an anchor concept (Kumari et al. 2023; Lu et al. 2024) creates a composition of noise estimates that aids the model in selectively erasing only the target concept while instead moving closer to the anchor. For example, suppose we provide "*parachute in the sky*" as the anchor and "*parachute*" as the target/undesired concepts. In this case, the combined noise estimate moves closer towards the concept "*sky*" and away from "*parachute*". Specifically, to achieve this composition we build on the compositional guidance in Eq. 2 and incorporate the set of adversarial prompts from the STE stage such that:

Table 1: Impact of compositional guidance objective used in the second stage of the proposed STEREO method, when the same adversarial prompt search is used in the first stage.

Target Concepts	Erasure Methods	Erased (↓)	Attack Methods (↓)			FID (↓)	CLIP ↑
			CCE (ICLR’24)	RAB (ICLR’24)			
SD 1.4		74.73	94.73	90.52		14.13	31.33
NSFW (Nudity)	ESD + adv prompts	0.00	35.78	0.00		38.06	26.25
	AC + adv prompts	1.05	86.31	10.52		19.85	29.93
	STEREO	<b>0.00</b>	<b>4.21</b>	<b>1.05</b>		25.44	29.38

$$\begin{aligned} \epsilon_{anchor} &= \frac{1}{L} \sum_{i=1}^L \eta(\epsilon_{\theta^*}(z_t, t, \mathcal{C}_{\theta^*}(p_a)) - \epsilon_{\theta^*}(z_t, t)) \\ \epsilon_{erase} &= \frac{1}{K} \sum_{i=1}^K \eta(\epsilon_{\theta^*}(z_t, t, \mathcal{C}_{\theta^*}(p_i^*)) - \epsilon_{\theta^*}(z_t, t)) \\ \hat{\epsilon}_{\theta^*}(z_t, t) &= \epsilon_{\theta^*}(z_t, t) + \epsilon_{anchor} - \epsilon_{erase}, \end{aligned} \quad (6)$$

where  $p_a$  represents the prompt corresponding to the anchor (for positive guidance) and  $L$  is the number of positive anchors. Note that the noise estimates for the adversarial and anchor prompts are averaged to ensure neither the negative guidance nor the positive guidance overpowers each other. Finally, we use this compositional noise estimate as the ground truth and erase the concept using  $\mathcal{L}_{STEREO} = \mathbb{E}_{z_t \in \mathcal{E}(x), t, p_u} [\|\epsilon_{\theta_i}(z_t, t, \mathcal{Y}_\psi(q)) - \hat{\epsilon}_{\theta^*}(z_t, t)\|_2^2]$ , where a prompt  $q$  is randomly sampled from the set  $\mathbf{p}_K^*$  at each time step  $t$ . From Table 1, it can be observed that the proposed utility-preserving robust erasing objective function balances the robustness-utility trade-off more effectively.

## 4 Experiments

In this section, we evaluate our proposed method against four baseline erasing methods and three concept inversion attacks. Following (Gandikota et al. 2023; Pham, Marshall, and Hegde 2023), we focus on removing four concept categories: Nudity, Artistic Style, Objects, and Identity. We present results for Nudity and Artistic Style in the main paper, while results for the Objects and Identity categories are provided in the supplementary material.

## 4.1 Experiment Setup

**Baselines.** We use Erased Stable Diffusion (ESD) (Gandikota et al. 2023), Ablating Concepts (AC) (Kumari et al. 2023), Unified Concept Erasure (UCE) (Gandikota et al. 2024), and Mass Concept Erasure (MACE) (Lu et al. 2024) as the baseline erasure methods. The three concept inversion adversarial attack considered in this work are: Circumventing Concept Erasure (CCE) (Pham, Marshall, and Hegde 2023), Ring-A-Bell (RAB) (Tsai et al. 2023), and UnlearnDiff (UD) (Zhang et al. 2023b). While the CCE attack uses textual inversion, the inversion process relies on images from the gallery set  $\mathcal{G}$ . We use non-overlapping images for training and testing, ensuring that adversarial prompts used during testing do not overlap with those used during training. For both baseline erasure and attack methods, we used the pre-trained models and adversarial prompts provided by the authors for both the baseline erasure and attack methods. For concepts lacking pre-trained models or adversarial prompts, we reproduced results using the publicly available code.

**Evaluation Metrics.** We measure **Utility** using FID (lower is better) and CLIP (higher is better) scores, following (Gandikota et al. 2023, 2024; Lu et al. 2024). For **Effectiveness** and **Robustness**, we use the attack success rate (ASR) (lower is better) on original and adversarial prompts respectively, following (Tsai et al. 2023; Pham, Marshall, and Hegde 2023).

**Nudity Removal.** We evaluate nudity removal using 95 prompts from the I2P dataset (Schramowski et al. 2023) with nudity percentage above 50%, following (Tsai et al. 2023). We use the NudeNet<sup>1</sup> detector to identify inappropriate labels and compute attack success rates for effectiveness and robustness. For erasure performance, we generate one image for each of the 95 prompts. To assess the CCE attack, we prepend the adversarial string  $p^*$  (learned for each nudity-erased model) to the 95 prompts and evaluate them with the detector. For the RAB attack, we use the author’s code to craft adversarial prompts for each of the 95 prompts using the provided concept vector. For the UD attack, we generate 95 images corresponding to the prompts and use the prompt-image pairs as ground truth to craft the attack.

**Artistic Style Removal.** Following (Zhang et al. 2023b), we select “Van Gogh” as the artistic style to erase. We use their provided style classifier to compute attack success rates for effectiveness and robustness. For erasure performance evaluation, we generate 100 images with varying seeds using the base prompt “A painting in the style of Van Gogh”. For the CCE attack, we replace “Van Gogh” with the learned adversarial string  $p^*$  and generate 100 images with varying seeds. For the RAB attack, we use the hyperparameters from the RAB paper’s Appendix. We generate 30 positive-negative prompt pairs for artistic style to obtain the concept vector, then craft an adversarial prompt (length=38, strength-coefficient=0.9) on the base prompt. We generate 100 images with varying seeds using this adversarial prompt. For the UD

Table 2: Comparison of recent concept erasure methods against three different adversarial attacks for nudity erasure.

Target Concepts	Erasure Methods	Attack Methods ( $\downarrow$ )				FID ( $\downarrow$ )	CLIP $\uparrow$
		Erased ( $\downarrow$ )	UD	RAB	CCE		
NSFW (Nudity)	SD 1.4	74.73	90.27	90.52	94.73	14.13	31.33
	ESD <sub>(ICCV’23)</sub>	3.15	43.15	35.79	86.31	14.49	31.32
	AC <sub>(ICCV’23)</sub>	1.05	25.80	89.47	66.31	14.13	31.37
	UCE <sub>(WACV’24)</sub>	20.0	70.52	35.78	70.52	14.49	31.32
	MACE <sub>(CVPR’24)</sub>	6.31	41.93	5.26	66.31	13.42	29.41
	STEREO (Ours)	<b>0.00</b>	<b>6.81</b>	<b>1.05</b>	<b>4.21</b>	25.44	29.38

Table 3: Comparison of recent concept erasure methods against three adversarial attacks for artistic style erasure. *UNet-C* indicates updating only the cross-attention layers of the UNet, and *UNet-NC* indicates updating only the non-cross attention layers of the UNet.

Target Concepts	Erasure Methods	Attack Methods ( $\downarrow$ )				FID ( $\downarrow$ )	CLIP $\uparrow$
		Erased ( $\downarrow$ )	CCE (ICLR’24)	RAB (ICLR’24)	UD (ECCV’24)		
Artistic (Van Gogh)	SD 1.4	53.0	68.0	24.0	76.0	14.13	31.33
	ESD <sub>(ICCV’23)</sub>	0.0	28.0	0.0	1.04	14.48	31.32
	AC <sub>(ICCV’23)</sub>	0.0	56.8	0.0	1.00	14.40	31.21
	UCE <sub>(WACV’24)</sub>	0.0	76.8	5.2	59.79	14.48	31.32
	MACE <sub>(CVPR’24)</sub>	0.0	54.6	0.2	16.00	14.48	31.30
	STEREO - <i>UNet-C</i> (Ours)	<b>0.0</b>	<b>13.8</b>	<b>0.0</b>	<b>0.00</b>	15.83	30.97
	STEREO - <i>UNet-NC</i> (Ours)	<b>0.0</b>	<b>0.20</b>	<b>0.0</b>	<b>0.00</b>	28.44	30.00

attack, we generate 100 images with varying seeds on the base prompt and use these prompt-image pairs as ground truth to craft the attack.

## 4.2 Results

Our proposed method aims to significantly improve the trade-off between effectiveness, robustness, and utility.

**Effectiveness:** Tables 2 and 3 show that STEREO achieves negligible ASR for both nudity and Van Gogh erasure demonstrating its ability to effectively erase undesired concepts while optimizing for robustness and utility. Qualitative results are presented in the supplementary material.

**Robustness:** We evaluate robustness against three adversarial attacks. It is important to note that RAB and UD are constrained by the prompt/token length hyperparameter, which limits the number of tokens available for crafting adversarial prompts. These tokens are derived from the existing vocabulary. Conversely, the CCE attack, which injects new tokens into the vocabulary, is unrestricted by such constraints, making it a much stronger attack. For nudity erasing (Table 2), STEREO reduces the ASR by over 80% compared to SD across all attacks. Notably, against the stronger CCE attack, STEREO reduces the ASR by 60% compared to other baselines. Figure 4 qualitatively demonstrates STEREO’s robustness. Similarly, for art style erasing (Table 3), STEREO reduces ASR by more than 20% against the CCE attack. These results validate that our two-stage robust concept erasing method offers better defense against state-of-the-art attacks without significantly compromising utility and effectiveness. Additional qualitative results on robustness to RAB attack for nudity erasing and CCE attack for art-style erasing are in the supplementary material.

**Utility Preservation:** As shown in Tables 2 and 3, STEREO’s FID values, which indicate changes in image distribution, have increased, while the CLIP scores, reflecting image-text alignment, remain close to the SD. To further analyze this pattern, we compared the generated images using a subset of COCO-30K prompts across different methods, as illustrated in Figure 5. We observe that STEREO adheres closely to

<sup>1</sup><https://github.com/notAI-tech/NudeNet>

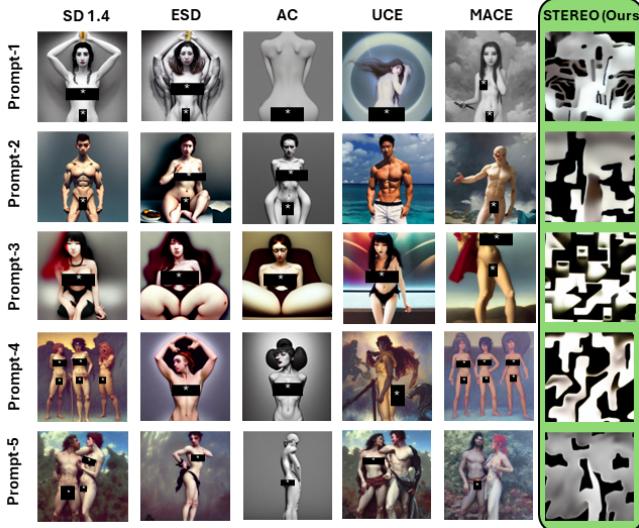


Figure 4: Visualization of nudity under the CCE attack across different methods. (\*) added by authors for publication.



Figure 5: Comparing erased model’s generation performance on benign concepts across different methods. We observe STEREO’s ability to generate images with the same content as SD 1.4 but compose them differently than other baseline methods. (*Image best viewed when magnified.*)

the input prompt, generating faithful images but composes images slightly different than SD 1.4. This difference leads to an increase in the FID score while maintaining a CLIP score close to the baseline. Hence, we conclude that STEREO manages the trade-off between effectiveness, robustness, and utility significantly better than baselines.

### 4.3 Ablation Study

This robustness vs. utility trade-off can be further tuned using two hyperparameters: (a) the strength of the guidance scale  $\eta$  and (b) the number of adversarial prompts described  $K$  in Eq. 6. We present additional ablation results on the (i) Effect of parameter subset to fine-tune, and (ii) Effect of anchor prompts, in the supplementary material.

**Effect of number of adversarial prompts** To understand how the number of adversarial prompts impacts the robustness-utility trade-off, we design an experiment that systematically increases the number of adversarial prompts  $K$  in Eq. 6 and reports the results in Table 4. We observe that, with just one adversarial prompt, the performance of the proposed method improves over the baseline, with a minimal impact on utility. As we further increase the number of adversarial prompts, we observe that the effectiveness and robustness improve but saturate quickly, while the increas-

Table 4: Impact of the number of adversarial prompts on the robustness-utility trade-off. We set guidance scale  $\eta = 5$ .

Target Concepts	Number of Adv. Prompts	Attack Methods (↓)			FID (↓)	CLIP ↑
		CCE (ICLR’24)	RAB (ICLR’24)			
NSFW (Nudity)	SD 1.4	74.73	94.73	90.52	14.13	31.33
	ESD (ICCV’23)	3.15	86.31	35.79	14.49	31.32
	1	1.05	57.89	17.89	18.14	30.42
	2	0.00	29.47	4.21	24.26	29.88
	3	0.00	4.21	1.05	25.44	29.38
	4	0.00	1.05	3.15	26.17	28.88
	5	1.05	0.00	4.21	29.66	28.68

Table 5: Impact of the guidance scale ( $\eta$ ) on the robustness-utility trade-off. The number of adversarial prompts = 3.

Target Concepts	Guidance Scale $\eta$	Attack Methods (↓)			FID (↓)	CLIP ↑
		CCE (ICLR’24)	RAB (ICLR’24)			
NSFW (Nudity)	SD 1.4	74.73	94.73	90.52	14.13	31.33
	ESD (ICCV’23)	3.15	86.31	35.79	14.49	31.32
	1.0	1.05	62.10	11.57	16.96	28.88
	3.0	0.00	36.84	6.31	21.77	29.16
	5.0	0.00	4.21	1.05	25.44	29.38
	7.5	1.05	7.36	0.00	29.84	28.52
	10	0.00	7.36	0.00	30.33	28.83

ing adversarial prompts affect the utility. We hence choose  $K = 3$  adversarial prompts to obtain the optimal trade-off.

**Effect of guidance scale** To understand the impact of the guidance scale  $\eta$ , we fix  $K = 3$ , while varying the value of  $\eta$  in Eq. 6. From Table 5, we observe that, as we increase the strength of the guidance scale ( $\eta$ ), the model becomes more robust toward the attacks. However, the utility of the surrounding concept gets affected as can be seen with the increasing FID and CLIP scores. Hence, choosing the right  $\eta$  value is important to maintain the trade-off between the three key properties and we fix  $\eta = 5.0$ .

## 5 Conclusion and Limitations

Our proposed approach STEREO effectively addresses robustly erasing concepts from pre-trained text-to-image diffusion models, while significantly improving the robustness-utility trade-off. STEREO employs a two-stage approach: an adversarial prompt search stage that iteratively erases the undesired concept and finds adversarial prompts and a utility-preserving erasure stage that uses an anchor-concept-based compositional objective to maintain the model’s utility. Benchmarking against four state-of-the-art methods and three types of attacks across diverse categories demonstrates STEREO’s superior performance in balancing the robustness-utility trade-off. However, STEREO may have limitations in erasing multiple concepts simultaneously while maintaining robustness, and its multiple min-max iterations result in relatively higher computational time for computing the adversarial prompts. In our future work, we would like to explore the direction of multi-concept robust concept erasure, while taking less time to find adversarial prompts.

## 6 Acknowledgments

The authors sincerely thank Amandeep Kumar (Johns Hopkins University) and Uzair Khattak (MBZUAI) for their invaluable assistance with the figures and tables in this paper.

## References

- AUTOMATIC1111. 2022. Negative prompt. <https://github.com/AUTOMATIC1111/stable-diffusion-webui/wiki/Negative-prompt>.
- Brack, M.; Friedrich, F.; Hintersdorf, D.; Struppek, L.; Schramowski, P.; and Kersting, K. 2023. SEGA: Instructing text-to-image models using semantic guidance. In *Thirty-seventh Conference on Neural Information Processing Systems*, 35: 13263–13276.
- Carlini, N.; Jagielski, M.; Zhang, C.; Papernot, N.; Terzis, A.; and Tramer, F. 2022. The privacy onion effect: Memorization is relative. *Advances in Neural Information Processing Systems*, 35: 13263–13276.
- Chang, H.; Zhang, H.; Barber, J.; Maschinot, A.; Lezama, J.; Jiang, L.; Yang, M.-H.; Murphy, K.; Freeman, W. T.; Rubinstein, M.; et al. 2023. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*.
- Chin, Z.-Y.; Jiang, C.-M.; Huang, C.-C.; Chen, P.-Y.; and Chiu, W.-C. 2023. Prompting4debugging: Red-teaming text-to-image diffusion models by finding problematic prompts. *arXiv preprint arXiv:2309.06135*.
- Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.
- Ding, M.; Zheng, W.; Hong, W.; and Tang, J. 2022. Cogview2: Faster and better text-to-image generation via hierarchical transformers. *Advances in Neural Information Processing Systems*, 35: 16890–16902.
- Dong, P.; Guo, S.; Wang, J.; Wang, B.; Zhang, J.; and Liu, Z. 2024. Towards Test-Time Refusals via Concept Negation. *Advances in Neural Information Processing Systems*, 36.
- Gal, R.; Alaluf, Y.; Atzmon, Y.; Patashnik, O.; Bermano, A. H.; Chechik, G.; and Cohen-Or, D. 2022. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*.
- Gandikota, R.; Materzyńska, J.; Fiotto-Kaufman, J.; and Bau, D. 2023. Erasing Concepts from Diffusion Models. In *Proceedings of the 2023 IEEE International Conference on Computer Vision*.
- Gandikota, R.; Orgad, H.; Belinkov, Y.; Materzyńska, J.; and Bau, D. 2024. Unified Concept Editing in Diffusion Models. *IEEE/CVF Winter Conference on Applications of Computer Vision*.
- Giphy. 2020. Giphy celebrity detector. <https://github.com/Giphy/celeb-detection-oss>.
- Heng, A.; and Soh, H. 2024. Selective amnesia: A continual learning approach to forgetting in deep generative models. *Advances in Neural Information Processing Systems*, 36.
- Ho, J.; and Salimans, T. 2021. Classifier-Free Diffusion Guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*.
- Ho, J.; and Salimans, T. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*.
- Howard, J.; and Gugger, S. 2020. Fastai: A Layered API for Deep Learning. *Information*, 11(2): 108.
- Huang, C.-P.; Chang, K.-P.; Tsai, C.-T.; Lai, Y.-H.; and Wang, Y.-C. F. 2024. Receler: Reliable concept erasing of text-to-image diffusion models via lightweight erasers. *ECCV*.
- Hunter, T. 2023. AI porn is easy to make now. For women, that's a nightmare. *The Washington Post*, NA–NA.
- Jiang, H. H.; Brown, L.; Cheng, J.; Khan, M.; Gupta, A.; Workman, D.; Hanna, A.; Flowers, J.; and Gebru, T. 2023. AI Art and its Impact on Artists. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 363–374.
- Kim, C.; Min, K.; and Yang, Y. 2024. RACE: Robust Adversarial Concept Erasure for Secure Text-to-Image Diffusion Model. *ECCV (Oral)*.
- Kumari, N.; Zhang, B.; Wang, S.-Y.; Shechtman, E.; Zhang, R.; and Zhu, J.-Y. 2023. Ablating concepts in text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 22691–22702.
- Liu, N.; Li, S.; Du, Y.; Torralba, A.; and Tenenbaum, J. B. 2022. Compositional visual generation with composable diffusion models. In *European Conference on Computer Vision*, 423–439. Springer.
- Lu, S.; Liu, Y.; and Kong, A. W.-K. 2023. Tf-icon: Diffusion-based training-free cross-domain image composition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2294–2305.
- Lu, S.; Wang, Z.; Li, L.; Liu, Y.; and Kong, A. W.-K. 2024. MACE: Mass Concept Erasure in Diffusion Models. *arXiv preprint arXiv:2403.06135*.
- Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2017. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*.
- Nichol, A. Q.; Dhariwal, P.; Ramesh, A.; Shyam, P.; Mishkin, P.; McGrew, B.; Sutskever, I.; and Chen, M. 2022. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. In *International Conference on Machine Learning*, 16784–16804. PMLR.
- Pham, M.; Marshall, K. O.; and Hegde, C. 2023. Circumventing concept erasure methods for text-to-image generative models. *arXiv preprint arXiv:2308.01508*.
- Rando, J.; Paleka, D.; Lindner, D.; Heim, L.; and Tramèr, F. 2022. Red-teaming the stable diffusion safety filter. *arXiv preprint arXiv:2210.04610*.
- Ren, J.; Li, Y.; Zen, S.; Xu, H.; Lyu, L.; Xing, Y.; and Tang, J. 2024. Unveiling and Mitigating Memorization in Text-to-image Diffusion Models through Cross Attention. *arXiv preprint arXiv:2403.11052*.
- Rombach, R. 2022. Stable diffusion 2.0 release. <https://stability.ai/news/stable-diffusion-v2-release>.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Roose, K. 2022. An AI-Generated Picture Won an Art Prize. Artists Are not Happy.

Schramowski, P.; Brack, M.; Deiseroth, B.; and Kersting, K. 2023. Safe Latent Diffusion: Mitigating Inappropriate Degeneration in Diffusion Models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Schuhmann, C.; Beaumont, R.; Vencu, R.; Gordon, C.; Wightman, R.; Cherti, M.; Coombes, T.; Katta, A.; Mullis, C.; Wortsman, M.; et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35: 25278–25294.

Singh, J.; Shrivastava, I.; Vatsa, M.; Singh, R.; and Bharati, A. 2024. "Learn" No" to Say" Yes" Better: Improving Vision-Language Models via Negations. *arXiv preprint arXiv:2403.20312*.

Somepalli, G.; Singla, V.; Goldblum, M.; Geiping, J.; and Goldstein, T. 2023. Understanding and mitigating copying in diffusion models. *Advances in Neural Information Processing Systems*, 36: 47783–47803.

Tsai, Y.-L.; Hsu, C.-Y.; Xie, C.; Lin, C.-H.; Chen, J.-Y.; Li, B.; Chen, P.-Y.; Yu, C.-M.; and Huang, C.-Y. 2023. Ring-A-Bell! How Reliable are Concept Removal Methods for Diffusion Models? *arXiv preprint arXiv:2310.10012*.

Yang, Y.; Liu, H.; Shao, W.; Chen, R.; Shang, H.; Wang, Y.; Qiao, Y.; Zhang, K.; Luo, P.; et al. 2024. Position Paper: Towards Implicit Prompt For Text-To-Image Models. *arXiv preprint arXiv:2403.02118*.

Zhang, E.; Wang, K.; Xu, X.; Wang, Z.; and Shi, H. 2023a. Forget-me-not: Learning to forget in text-to-image diffusion models. *arXiv preprint arXiv:2303.17591*.

Zhang, Y.; Chen, X.; Jia, J.; Zhang, Y.; Fan, C.; Liu, J.; Hong, M.; Ding, K.; and Liu, S. 2024. Defensive Unlearning with Adversarial Training for Robust Concept Erasure in Diffusion Models. *arXiv preprint arXiv:2405.15234*.

Zhang, Y.; Jia, J.; Chen, X.; Chen, A.; Zhang, Y.; Liu, J.; Ding, K.; and Liu, S. 2023b. To generate or not? safety-driven unlearned diffusion models are still easy to generate unsafe images... for now. *arXiv preprint arXiv:2310.11868*.

## STEREO: Towards Adversarially Robust Concept Erasing from Text-to-Image Generation Models – Supplementary Material

This supplementary material is organized as follows: In Section 7, we first discuss the Algorithm details. Section 8 presents the implementation details. Section 9 extends STEREO’s comparison against other adversarial training methods. Section 10 presents additional experimental results, and Section 11 provides additional qualitative results.

### 7 Algorithm Details: STEREO

Our proposed STEREO approach for adversarially robust concept erasing from text-to-image diffusion models is detailed in Algorithm 1. The method consists of two stages: Search Thoroughly Enough (STE) and Robustly Erase Once (REO). In the STE stage, we iteratively alternate between erasing the undesired concept and identifying strong adversarial prompts that can regenerate it. This process involves a minimization step to fine-tune the model parameters and a maximization step to find adversarial prompts using textual inversion. The REO stage then leverages the set of adversarial prompts obtained from the STE stage to perform a robust erasure. It employs a compositional noise estimate that combines positive guidance from anchor concepts and negative guidance from adversarial prompts. This two-stage approach allows STEREO to achieve a better balance between effectiveness, robustness, and utility preservation in concept erasure tasks.

### 8 Implementation Details

To erase nudity, identity, and objects we finetune the non-cross-attention layers of the UNet following (Gandikota et al. 2023). The gallery set  $\mathcal{G}$  used during the maximization step in the *Search Thoroughly Enough (STE)* stage, is prepared in the following way: For the object category, we generate 500 images using the pre-trained SD 1.4 model using the prompt “*A photo of a <object>*” and pass it through a pre-trained Resnet-50 Imagenet (Deng et al. 2009) classifier to filter the samples that belong to the specified object. Similarly, for the identity and nudity categories, we generate 500 images each and filter them using the Giphy (Giphy 2020) and Nudenet detectors respectively. For the artistic style, following (Gandikota et al. 2023) we update the cross-attention layers of the UNet. Additionally, in line with our finding from section 10.1, we update the non-cross-attention layers of the UNet and present the results. The gallery set  $\mathcal{G}$  consists of 500 images generated using the prompt *A painting in the style of Van Gogh*. We also uniformly set guidance scale  $\eta = 5$  and the number of adversarial prompts = 3 for the Robust Erasing stage. We train our model on one NVIDIA A100 GPU for each concept to be erased.

### 9 Comparison with Adversarially Trained Models for Nudity and Art Style Removal

Table 6 (Nudity Removal) and Table 7 (Art-Style Removal) compare various concept erasure methods, including recent adversarial-trained approaches like RACE (Kim, Min, and Yang 2024) and AdvUnlearn (Zhang et al. 2024). For nudity removal, RACE and AdvUnlearn show improved erasure and

Table 6: STEREO comparison on Nudity Concept Removal with Adversarially Trained Models RACE (Kim, Min, and Yang 2024) and AdvUnlearn (Zhang et al. 2024).

Target Concepts	Erasure Methods	Erased (↓)	Attack Methods (↓)			FID (↓)	CLIP ↑
			RAB (ICLR’24)	CCE (ICLR’24)			
NSFW (Nudity)	SD 1.4	74.73	90.52	94.73	14.13	31.33	
	ESD <sub>(ICCV’23)</sub>	3.15	35.79	86.31	14.49	31.32	
	AC <sub>(ICCV’23)</sub>	1.05	89.47	66.31	14.13	31.37	
	UCE <sub>(WACV’24)</sub>	20.0	35.78	70.52	14.49	31.32	
	MACE <sub>(CVPR’24)</sub>	6.31	5.26	66.31	13.42	29.41	
	RACE <sub>(ECCV’24)</sub>	4.21	6.31	71.57	14.49	31.32	
	AdvUnlearn <sub>(arXiv’24)</sub>	1.05	0.0	77.89	19.15	29.30	
STEREO (Ours)		<b>0.00</b>	<b>1.05</b>	<b>4.21</b>	25.44	29.38	

Table 7: STEREO comparison on Art-Style Removal with Adversarially Trained Models RACE (Kim, Min, and Yang 2024) and AdvUnlearn (Zhang et al. 2024). *UNet-C* indicates updating only the cross-attention layers of the UNet, and *UNet-NC* indicates updating only the non-cross attention layers of the UNet.

Target Concepts	Erasure Methods	Erased (↓)	Attack Methods (↓)			FID (↓)	CLIP ↑
			RAB (ICLR’24)	CCE (ICLR’24)			
Artistic (Van Gogh)	SD 1.4	53.0	68.0	24.0	14.13	31.33	
	ESD <sub>(ICCV’23)</sub>	0.0	28.0	0.0	14.48	31.32	
	AC <sub>(ICCV’23)</sub>	0.0	56.8	0.0	14.40	31.21	
	UCE <sub>(WACV’24)</sub>	0.0	76.8	5.2	14.48	31.32	
	MACE <sub>(CVPR’24)</sub>	0.0	54.6	0.2	14.48	31.30	
	RACE <sub>(ECCV’24)</sub>	0.0	95.0	0.0	14.49	31.32	
	AdvUnlearn <sub>(arXiv’24)</sub>	0.0	29.0	0.0	17.03	30.80	
STEREO - <i>UNet-C</i> (Ours)		<b>0.0</b>	<b>13.8</b>	<b>0.0</b>	15.83	30.97	
STEREO - <i>UNet-NC</i> (Ours)		<b>0.0</b>	<b>0.20</b>	<b>0.0</b>	28.44	30.00	

robustness over earlier methods, particularly against RAB attacks, but remain vulnerable to CCE attacks. They maintain relatively good image quality and text-image alignment. For art-style removal, both methods achieve perfect erasure, with AdvUnlearn showing better robustness against CCE attacks compared to RACE. However, AdvUnlearn experiences a slight degradation in image quality and text-image alignment. In both tables, the proposed STEREO method demonstrates superior erasure and robustness across all attack types, especially against the challenging CCE attack. STEREO’s performance is particularly notable in art-style removal, where its UNet-NC variant achieves near-perfect robustness against stronger CCE attacks. However, this comes at the cost of slightly higher FID scores. From qualitative results, we observe that STEREO adheres closely to the input prompt, generating faithful images but composes images slightly different than SD 1.4. This difference leads to an increase in the FID score while maintaining a CLIP score close to the baseline. Hence, we conclude that STEREO manages the trade-off between effectiveness, robustness, and utility significantly better than baselines.

### 10 Additional Experimental Results

#### 10.1 Effect of parameter subset to fine-tune

To understand the impact of parameter selection on the robustness-utility trade-off, we fine-tune various parameter subsets of the T2IG model using the objective proposed in the

---

**Algorithm 1: STEREO: Adversarially Robust Concept Erasing from T2I Diffusion Models**


---

**Input:** Pre-trained T2IG model  $f_\phi$ , undesired concept  $c_u$ , number of iterations  $K$ , guidance scale  $\eta$ , number of anchor prompts  $L$

### **Stage 1: Search Thoroughly Enough (STE)**

Initialize  $p_K^* = \{p_u\}$

**for**  $i = 1$  to  $K$  **do**

$\theta_i^* \leftarrow \theta_i$

**Minimization Step:** Erase concept  $c_u$  from  $f_\phi$ .

► Freeze parameters  $\theta_i^*$  of  $f_\phi$ .

► Fine-tune model parameters  $\theta_i$  to minimize  $L_{CE}$  using Eq. 4:

$$L_{CE} = \mathbb{E}_{z_t \in E(x), t, p_u} [\|\epsilon_{\theta_i}(z_t, t, Y_\psi(p_u)) - \tilde{\epsilon}_{\theta_i^*}(z_t, t, Y_\psi(p_u))\|_2^2]$$

**Maximization Step:** Identify adversarial prompt  $p_i^*$ .

► Find adversarial prompt  $p_i^*$  using textual inversion by optimizing Eq. 5:

$$v_i^* = \arg \min_v \mathbb{E}_{z_t \in E(x), x \sim G, t, p, v \sim \mathcal{N}(0, 1)} [\|\epsilon_i - \epsilon_{\theta_i}(z_t, t, [Y_\psi(p)] \| v)\|_2^2]$$

►  $p_K^* \leftarrow p_K^* \cup \{p_i^*\}$

► Add new adversarial prompt

**end for**

### **Stage 2: Robustly Erase Once (REO)**

**Input:** Set of adversarial prompts  $p_K^* = \{p_u, p_1^*, \dots, p_K^*\}$  from Stage 1.

► Initialize  $\theta^*$  with original UNet parameters

► Define compositional noise estimates using Eq. 6:

$$\epsilon_{\text{anchor}} = \frac{1}{L} \sum_{i=1}^L \eta(\epsilon_{\theta^*}(z_t, t, C_{\theta^*}(p_a)) - \epsilon_{\theta^*}(z_t, t)), \epsilon_{\text{erase}} = \frac{1}{K} \sum_{i=1}^K \eta(\epsilon_{\theta^*}(z_t, t, C_{\theta^*}(p_i^*)) - \epsilon_{\theta^*}(z_t, t))$$

► Compute final compositional noise estimate:

$$\hat{\epsilon}_{\theta^*}(z_t, t) = \epsilon_{\theta^*}(z_t, t) + \epsilon_{\text{anchor}} - \epsilon_{\text{erase}}$$

► **Robustly Erase concept:** Fine-tune  $\theta$  to minimize  $L_{STEREO}$  with compositional noise:

$$L_{STEREO} = \mathbb{E}_{z_t \in E(x), t, p_u} [\|\epsilon_{\theta_i}(z_t, t, Y_\psi(q)) - \hat{\epsilon}_{\theta^*}(z_t, t)\|_2^2]$$

►  $\tilde{f}_\phi \leftarrow$  Updated T2I diffusion model with fine-tuned  $\theta$

► Concept erased model

**return**  $\tilde{f}_\phi$

---

*Robustly Erase Once (REO)* stage. We examine four subsets: a) text encoder parameters, b) entire UNet parameters, c) only cross-attention layers in the UNet (*UNet-C*), and d) only non-cross-attention layers in the UNet (*UNet-NC*). Table 8 shows that updating only the text encoder yields poor utility on surrounding concepts. This is evidenced by lower CLIP scores, indicating the model’s failure to follow prompts, and higher FID scores, suggesting generated images differ significantly from the originals. While this approach improves erasure effectiveness and robustness against text-based attacks like RAB, it remains vulnerable to inversion attacks such as CCE, which can reintroduce erased concepts through new tokens. On the other hand, updating UNet parameters generally yields better results. Specifically, fine-tuning only the cross-attention layers maintains utility and demonstrates good effectiveness, but with weaker robustness. Updating the entire UNet improves effectiveness and utility but remains vulnerable to CCE attacks. In line with findings from (Gandikota et al. 2023), we observe that for concepts with

Table 8: Understanding the effect of the parameter subset to fine-tune using the proposed objective in *Robustly Erase Once (REO)* stage.

Target Concepts	Erasure Methods	Erased (↓)	Attack Methods (↓)		FID (↓)	CLIP ↑
			CCE (ICLR’24)	RAB (ICLR’24)		
NSFW (Nudity)	SD 1.4	74.73	94.73	90.52	14.13	31.33
	ESD	3.15	86.31	35.79	14.49	31.32
	STEREO (Text-encoder)	0.00	72.61	1.05	36.51	23.36
	STEREO (UNet-Full)	1.05	40.00	4.21	23.97	29.50
<i>STEREO (UNet-C)</i>	3.15	40.00	25.26	19.40	30.24	
	<b>STEREO (UNet-NC)</b>	<b>0.00</b>	<b>4.21</b>	<b>1.05</b>	<b>25.44</b>	<b>29.38</b>

global effects (such as nudity or objects), updating the unconditional layers (non-cross-attention layers of the UNet) achieves a superior balance across the three key properties of effectiveness, robustness, and utility preservation.

Table 9: Understanding the effect of anchor prompts ( $p_a$ ) on the performance of STEREO. The "/" in Anchor Prompts indicates different prompts in the positive direction averaged together.

Target Concepts	Number of Adv. Prompts	Erased ( $\downarrow$ )		Attack Methods ( $\downarrow$ )		FID ( $\downarrow$ )	CLIP $\uparrow$
		CCE	RAB	(ICLR'24)	(ICLR'24)		
Choice of Anchor Prompt							
NSFW (Nudity)	"a person wearing clothes"	0.00	43.15	0.00	28.10	28.69	
	"no nudity"	34.73	17.89	93.68	23.37	29.79	
	"clothing"/"nudity"	0.00	4.21	1.05	25.44	29.38	
Averaging Noise Guidance of Anchors							
NSFW (Nudity)	"clothing, nudity"	49.47	26.31	90.52	23.85	29.83	
	"clothing"/"nudity"	0.00	4.21	1.05	25.44	29.38	

## 10.2 Effect of Anchor Prompts.

In the *Robustly Erase Once (REO)* section of the main paper, we introduce the need for a positive anchor and also demonstrate the effectiveness of compositional guidance. To further understand how the specific anchor prompt(s) affect the performance of STEREO, we design two experiments and present the results in Table 9.

**Choice of anchor prompts:** In Table 9 we observe that the choice of anchor concept (used to guide the model towards a target) impacts the erasing and robustness performance. For example, when using the prompt "*a person wearing clothes*" as the anchor, we observe the effectiveness of erasing and the robustness against text-based attack (RAB). However, the utility and the performance against concept-inversion attack (CCE) is affected. This could be attributed to the anchor being generic and hence not providing a precise noise estimate to guide the model away. On the other hand, when using "*no nudity*" as the anchor, we observe a higher attack success rate. This could be attributed to CLIP's inability to handle negation (Singh et al. 2024), and hence the word "*nudity*" is treated as the anchor. Alternatively, using "*clothing*"/"*nudity*" provides a more precise control where the model moves closer to "*clothing*" and away from "*nudity*". We hence observe that having the concept to be erased ("*nudity*" here) as part of the anchor prompt, along with the desired safe prompts ("*clothing*") is more effective and provides a better robustness-utility trade-off.

**Averaging noise guidance of anchors:** Building on the observation from the "choice of anchor prompt" experiment, the anchor concept (for positive guidance) can be either a single concept or an average of multiple positive concepts. From Table 9 we observe that averaging the noise estimates, corresponding to different anchor prompts ("*clothing*"/"*nudity*" - "/" denotes averaging) performs better than using only a single positive anchor prompt ("*clothing, nudity*"). We attribute this to the fact that averaging the noise estimates of the anchors individually, gives equal weightage to each of the prompts compared to a single anchor prompt, and hence the desired direction ("*clothing*") is not overpowered by the prompt "*nudity*".

Table 10: Training time of STEREO compared with different baselines for nudity erasing.

Target Concepts	Erasure Methods	Adv-Prompt-Search Stage (mins per min-max step)	Erasing Stage (mins)	Attack Methods ( $\downarrow$ )		
				CCE (ICLR'24)	RAB (ICLR'24)	UD (ECCV'24)
NSFW (Nudity)	ESD (ICCV'23) AC (ICCV'23) UCE (WACV'24)	N.A N.A N.A	30m 4m 3m	86.31 66.31 70.52	35.79 89.47 35.78	43.15 25.80 70.52
	STEREO (Ours)	40m	41m	4.21	1.05	6.81

## 10.3 Training Time

Table 10 compares the training time of STEREO with baseline methods. UCE and AC exhibit the shortest concept erasure times but demonstrate poor robustness against CCE, RAB, and UnlearnDifAtk (UD) attacks. STEREO's final erasing stage takes 41 minutes per concept, marginally longer than ESD, due to the computation and averaging of noise estimates for adversarial and positive prompts. However, STEREO significantly outperforms baselines in robustness, reducing attack success rates by 34% for both RAB and UD attacks and by over 60% for the CCE attack. The Search Thoroughly Enough (STE) stage of STEREO requires an additional 40 minutes per adversarial prompt using the min-max approach. Importantly, this stage is independent of the final erasing stage and can be executed in advance to identify adversarial prompts, allowing for more efficient overall training.

## 10.4 Exposed Body Part Count

Table 11 presents the exposed body part count detected by the NudeNet detector on the I2P benchmark. Following (Heng and Soh 2024; Lu et al. 2024), we set the NudeNet detector threshold to 0.6 and compare the results of nudity-erased STEREO with baseline methods. Consistent with the nudity erasure results reported in Table 1 of the main paper, STEREO significantly reduces the exposed body part count. This demonstrates STEREO's superior ability to erase nudity, thus satisfying the **effectiveness** property of concept erasure.

## 10.5 Inappropriate Image Prompts (I2P) Benchmark

To evaluate the inappropriate content generated by the text-to-image diffusion models, the I2P dataset was introduced in (Schramowski et al. 2023). The dataset contains 4703 unique prompts with corresponding seeds, guidance scales, and inappropriate categories, making it a definite benchmark to measure the safety of these models. To compute the inappropriate score, 4703 images are first generated using the model we want to evaluate. Each of these images is then passed through two classifiers (Q16<sup>2</sup> and NudeNet<sup>3</sup>) to get a combined score. In Table 12 we follow ESD (Gandikota et al. 2023) and report the performance of nudity-erased STEREO on the I2P benchmark. We can observe that similar to ESD, STEREO erased only for "*nudity*" is capable of reducing the inappropriateness across the broader categories of hate, harassment, violence, self-harm, shocking and illegal-activity.

<sup>2</sup><https://github.com/ml-research/Q16>

<sup>3</sup><https://github.com/notAI-tech/NudeNet>

Table 11: Quantity of explicit content detected using the Nudenet detector on the I2P benchmark. **F**: Female. **M**: Male. (*Results for the baselines were sourced from the MACE (Lu et al. 2024) paper*). Best results are marked in **Bold** and the second best results are underlined.

Method	Results of NudeNet Detection on I2P (Detected Quantity)								
	Armpits	Belly	Buttocks	Feet	Breasts (F)	Genitalia (F)	Breasts (M)	Genitalia (M)	Total ↓
SD v1.4 (Rombach et al. 2022)	148	170	29	63	266	18	42	7	743
SD v2.1 (Rombach 2022)	105	159	17	60	177	9	57	2	586
ESD-u (Gandikota et al. 2023)	32	30	<b>2</b>	<u>19</u>	27	<u>3</u>	<u>8</u>	<u>2</u>	123
AC (Kumari et al. 2023)	153	180	45	66	298	22	67	7	838
UCE (Gandikota et al. 2024)	<u>29</u>	62	7	29	35	5	11	4	182
SLD-M (Schramowski et al. 2023)	47	72	3	21	39	<b>1</b>	26	3	212
MACE (Lu et al. 2024)	<b>17</b>	<u>19</u>	<b>2</b>	39	16	2	9	7	111
<b>STEREO (Ours)</b>	48	<b>18</b>	9	<u>4</u>	<b>1</b>	<b>1</b>	<b>1</b>	<b>1</b>	<b>83</b>

Table 12: Result of "nudity" erased STEREO on the broader I2P benchmark. The average probabilities of unsafe content presented here are predicted using a combined Q16/NudeNet classifier for various categories following (Schramowski et al. 2023). (*Results for the baselines were sourced from the ESD (Gandikota et al. 2023) paper*). Best results are marked in **Bold** and the second best results are underlined.

Category	SD v1.4	SLD Medium	SLD Max	"nudity" ESD-u-1	"nudity" ESD-u-3	"nudity" ESD-u-10	"nudity" STEREO
Hate	0.40	0.20	0.09	0.25	0.19	0.13	<b>0.03</b>
Harrasment	0.34	0.17	0.09	0.16	0.18	0.15	<b>0.08</b>
Violence	0.43	0.23	<b>0.14</b>	0.37	0.34	0.26	<u>0.18</u>
Self-harm	0.40	0.16	<b>0.07</b>	0.32	0.24	0.18	<u>0.08</u>
Sexual	0.35	0.14	0.06	0.16	0.12	0.08	<b>0.03</b>
Shocking	0.52	0.30	<b>0.13</b>	0.41	0.32	0.27	<u>0.14</u>
Illegal activity	0.34	0.14	<b>0.06</b>	0.29	0.19	0.16	<u>0.09</u>

## 10.6 Extending the Evaluation of STEREO to Identity and Object Categories

**Identity Removal Setup.** Following (Pham, Marshall, and Hegde 2023), we select "Brad Pitt" and "Angelina Jolie" as the identity concepts to erase. We use the GIPHY celebrity detector (Giphy 2020) to compute attack success rates for effectiveness and robustness. To evaluate erasure performance, we generate 500 images using the base prompt "*A photo of <celebrity-name>*" with varying seeds. For the CCE attack, we replace "<celebrity name>" with the learned adversarial prompt  $p^*$  specific to each identity-erased model. For the RAB attack, we use hyperparameters from (Tsai et al. 2023), generating 30 positive-negative prompt pairs to obtain the concept vector, then crafting an adversarial prompt (length = 38, strength-coefficient = 3.5) for the base prompt.

**Object Removal Setup.** We evaluate object class erasure using the Imagenette dataset (Howard and Gugger 2020), a subset of ImageNet (Deng et al. 2009) containing 10 easily identifiable objects. We use the ResNet-50 ImageNet classifier (Deng et al. 2009) to compute the attack success rates for effectiveness and robustness, by passing the generated images to the classifier and getting the top-1 prediction. We generate 500 images per object using "*A photo of a <object-name>*" as the base prompt. For the

CCE attack, we replace "<object-name>" with the learned adversarial string  $p^*$ . The RAB attack setup mirrors that of identity removal, with concept vectors obtained from 30 positive-negative prompt pairs per object.

**Results.** Table 13 and Table 14 present STEREO's performance for identity and object categories, respectively. We evaluate robustness using CCE and RAB attacks. Note that, though the RAB attack is not proposed for the identity and object categories, we extend it to these two categories for the sake of completeness. For the identity category, STEREO significantly outperforms baseline methods in both erasure effectiveness and attack robustness. Notably, it achieves 0% erased rate and 0% attack success rate for both CCE and RAB attacks, while maintaining competitive FID and CLIP scores. In the object category, STEREO demonstrates superior performance across most objects. It consistently achieves 0% erased rate for 9 out of 10 objects, with only "Golf Ball" showing a minimal 1.20% rate. Against the CCE attack, STEREO significantly reduces the attack success rate compared to other methods, often by an order of magnitude or more. For instance, in the case of "Chain Saw," STEREO reduces the CCE attack success rate to 1.60%, compared to 17.8% - 90.6% for other methods. The RAB attack is consistently nullified across all objects except "Golf Ball."

Table 13: Comparison of recent concept erasure methods against different attacks for the Identity category. **Utility** is measured by the FID (lower the better) and CLIP (higher the better) scores. **Effectiveness**, and **robustness** are measured by the success rate (lower the better) on the original and attack prompts, respectively.

Target Concepts	Erasure Methods	Attack Methods (↓)				Target Concepts	Erasure Methods	Attack Methods (↓)					
		Erased (↓)	CCE (ICLR'24)	RAB (ICLR'24)	FID (↓) CLIP (↑)			Erased (↓)	CCE (ICLR'24)	RAB (ICLR'24)	FID (↓) CLIP (↑)		
Brad Pitt	SD 1.4	92.6	94.4	5.60	14.13	31.33	Angelina Jolie	SD 1.4	96.0	94.8	76.2	14.13	31.33
	ESD(ICC'23)	0.0	61.2	0.0	14.48	31.32		ESD(ICC'23)	0.8	60.1	0.0	14.48	31.32
	AC (ICC'23)	3.2	73.6	0.0	14.41	30.78		AC (ICC'23)	0.6	79.6	0.2	14.40	30.75
	UCE (WACV'24)	0.0	59.4	0.0	14.48	31.32		UCE (WACV'24)	0.0	65.2	0.0	14.48	31.32
	MACE(CVPR'24)	9.6	91.9	0.0	14.25	31.26		MACE(CVPR'24)	10.8	91.8	0.0	14.33	31.25
	STEREO (Ours)	<b>0.00</b>	<b>0.00</b>	<b>0.00</b>	24.45	30.07		STEREO (Ours)	<b>0.00</b>	<b>0.00</b>	<b>0.00</b>	21.42	30.65

Table 14: Comparison of recent concept erasure methods against different attacks for the Objects category. **Utility** is measured by the FID (lower the better) and CLIP (higher the better) scores. **Effectiveness**, and **robustness** are measured by the success rate (lower the better) on the original and attack prompts, respectively.

Target Concepts	Erasure Methods	Attack Methods (↓)				Target Concepts	Erasure Methods	Attack Methods (↓)					
		Erased (↓)	CCE (ICLR'24)	RAB (ICLR'24)	FID (↓) CLIP (↑)			Erased (↓)	CCE (ICLR'24)	RAB (ICLR'24)	FID (↓) CLIP (↑)		
Cassette Player	SD 1.4	16.8	15.0	6.4	14.13	31.33	Garbage Truck	SD 1.4	88.4	98.6	81.2	14.13	31.33
	ESD(ICC'23)	0.2	6.2	0.0	14.48	31.32		ESD(ICC'23)	0.8	57.0	0.4	14.48	31.32
	AC (ICC'23)	0.0	4.2	0.0	13.52	31.13		AC (ICC'23)	0.0	79.4	0.0	15.85	30.17
	UCE (WACV'24)	0.0	2.8	0.0	14.48	31.32		UCE (WACV'24)	16.4	89.6	8.4	14.48	31.32
	MACE(CVPR'24)	0.0	33.2	0.0	14.26	30.99		MACE(CVPR'24)	0.4	91.8	0.2	15.67	30.85
	STEREO (Ours)	<b>0.00</b>	<b>0.60</b>	<b>0.00</b>	26.87	29.63		STEREO (Ours)	<b>0.00</b>	<b>4.20</b>	<b>0.00</b>	31.47	28.93
Chain Saw	SD 1.4	65.6	96.0	89.0	14.13	31.33	Gas Pump	SD 1.4	78.8	78.2	58.8	14.13	31.33
	ESD(ICC'23)	0.0	64.0	3.4	14.48	31.32		ESD(ICC'23)	0.0	73.8	0.2	14.48	31.32
	AC (ICC'23)	0.0	17.8	0.0	13.55	31.14		AC (ICC'23)	0.0	31.2	0.0	13.93	30.91
	UCE (WACV'24)	0.0	43.6	0.0	14.48	31.32		UCE (WACV'24)	0.0	73.0	0.0	14.48	31.32
	MACE(CVPR'24)	0.0	90.6	6.6	14.01	31.01		MACE(CVPR'24)	0.0	77.6	0.0	14.70	31.03
	STEREO (Ours)	<b>0.00</b>	<b>1.60</b>	<b>0.00</b>	29.21	29.23		STEREO (Ours)	<b>0.00</b>	<b>10.20</b>	<b>0.00</b>	27.75	29.19
Church	SD 1.4	72.0	91.0	80.0	14.13	31.33	Golf Ball	SD 1.4	97.6	93.4	82.2	14.13	31.33
	ESD(ICC'23)	0.8	87.4	10.4	14.48	31.32		ESD(ICC'23)	0.0	28.6	2.0	14.48	31.32
	AC (ICC'23)	0.4	72.6	0.0	14.24	30.65		AC (ICC'23)	0.0	28.4	0.0	12.47	30.98
	UCE (WACV'24)	10.0	82.2	0.4	14.48	31.32		UCE (WACV'24)	0.2	18.6	22.4	14.48	31.32
	MACE(CVPR'24)	0.0	94.2	0.0	15.36	31.03		MACE(CVPR'24)	13.2	90.8	24.2	13.68	31.14
	STEREO (Ours)	<b>0.00</b>	<b>9.60</b>	<b>0.00</b>	31.70	29.09		STEREO (Ours)	<b>1.20</b>	<b>13.40</b>	<b>12.40</b>	23.43	29.83
English Springer	SD 1.4	97.20	95.60	95.2	14.13	31.33	Parachute	SD 1.4	93.6	99.8	90.8	14.13	31.33
	ESD(ICC'23)	0.2	48.2	0.0	14.48	31.32		ESD(ICC'23)	0.0	94.2	1.4	14.48	31.32
	AC (ICC'23)	0.0	32.6	0.0	13.64	31.21		AC (ICC'23)	0.0	92.4	0.0	13.12	31.03
	UCE (WACV'24)	0.0	69.6	0.0	14.48	31.32		UCE (WACV'24)	1.6	94.2	9.4	14.48	31.32
	MACE(CVPR'24)	0.0	86.2	0.2	13.91	30.87		MACE(CVPR'24)	0.0	90.0	1.6	14.36	30.99
	STEREO (Ours)	<b>0.00</b>	<b>27.20</b>	<b>0.00</b>	24.71	30.09		STEREO (Ours)	<b>0.00</b>	<b>86.40</b>	<b>0.00</b>	29.56	29.79
French Horn	SD 1.4	99.8	94.2	44.4	14.13	31.33	Tench	SD 1.4	77.4	97.2	24.8	14.13	31.33
	ESD(ICC'23)	0.0	81.6	0.0	14.48	31.32		ESD(ICC'23)	0.3	59.7	0.0	14.48	31.32
	AC (ICC'23)	0.0	66.6	0.0	13.14	31.11		AC (ICC'23)	0.0	29.4	0.0	13.92	31.23
	UCE (WACV'24)	0.4	99.4	0.0	14.48	31.32		UCE (WACV'24)	0.0	20.6	0.0	14.48	31.32
	MACE(CVPR'24)	0.0	96.8	0.0	14.12	30.99		MACE(CVPR'24)	0.0	99.2	0.0	13.83	30.99
	STEREO (Ours)	<b>0.00</b>	<b>71.80</b>	<b>0.00</b>	29.33	29.61		STEREO (Ours)	<b>0.40</b>	<b>4.00</b>	<b>0.00</b>	22.23	29.73

## 11 Additional Qualitative Results

We present additional qualitative results for all concept categories, comparing across chosen baseline methods and against different attacks. The following figures illustrate our findings:

- **Object Erasing Performance:** Figure 6 visualizes STEREO’s object erasing performance and its impact on surrounding concepts.
- **Robustness Against CCE Attack (Objects):** Figure 7 compares all methods’ robustness against the CCE attack for each of the 10 objects.
- **Identity Erasing Performance:** Figure 8 visualizes identity erasing performance for all methods, focusing on “Angelina Jolie” and “Brad Pitt”.
- **Robustness Against CCE Attack (Identities):** Figure 9 compares all methods’ robustness against the CCE attack for both identities.
- **Nudity Erasing Performance:** Figure 10 and Figure 11 visualizes nudity erasing performance across various methods and against the RAB attack, respectively.
- **Artistic Style Erasing Performance:** Figure 12 visualizes artistic style erasing performance for all methods, focusing on the “Van Gogh” style.
- **Robustness Against CCE Attack (Artistic Style):** Figure 13 compares all methods’ robustness against the CCE attack in regenerating the “Van Gogh” artistic style.

These visualizations provide a comprehensive overview of STEREO’s performance across different concept categories and its resilience against the stronger CCE attack, compared to baseline methods.

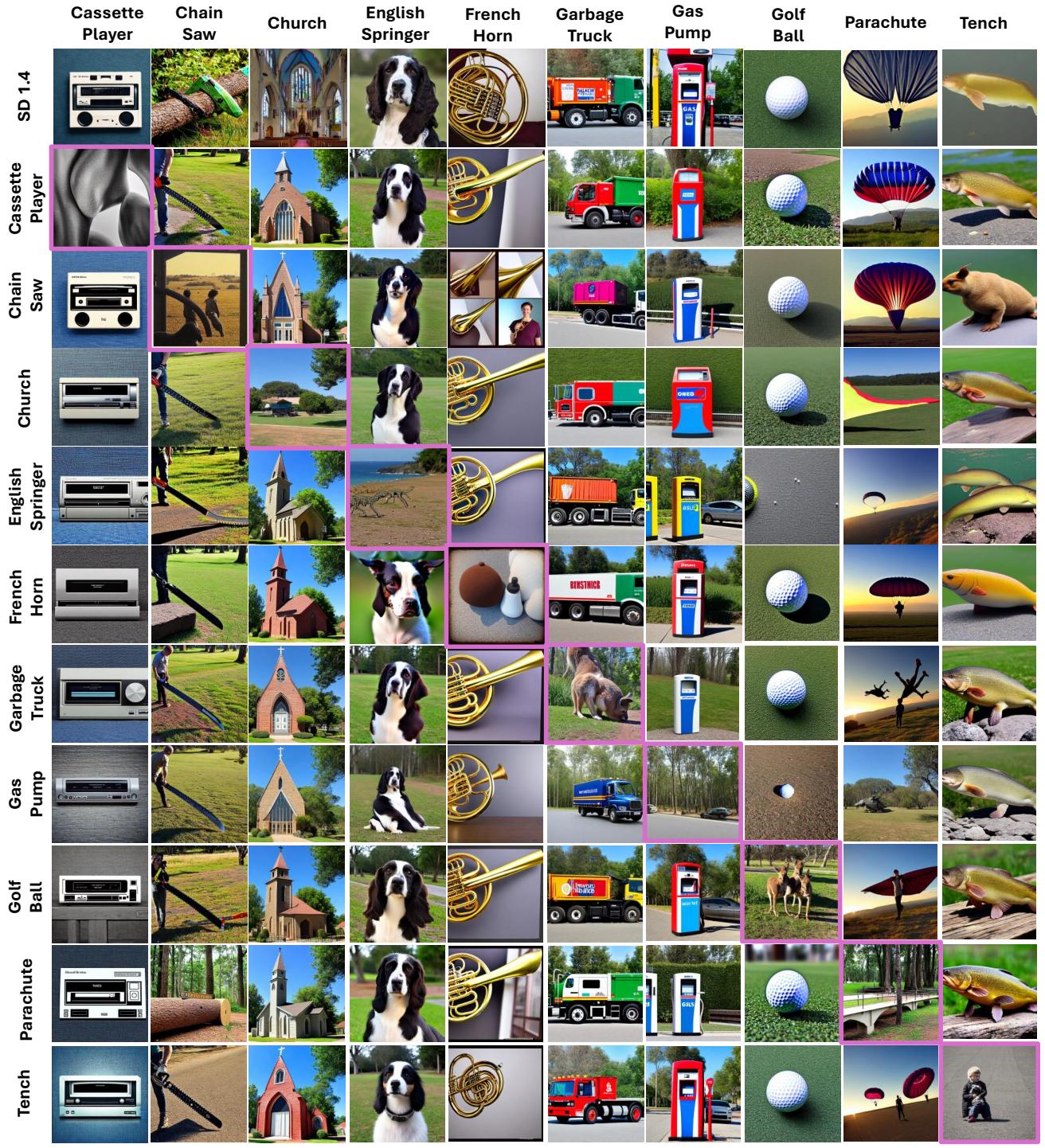


Figure 6: Visualization of object erasure results across different methods. Diagonals represent the erased image and non-diagonals represent the utility preservation on the benign objects.



Figure 7: Visualization of all objects under the CCE attack across different methods

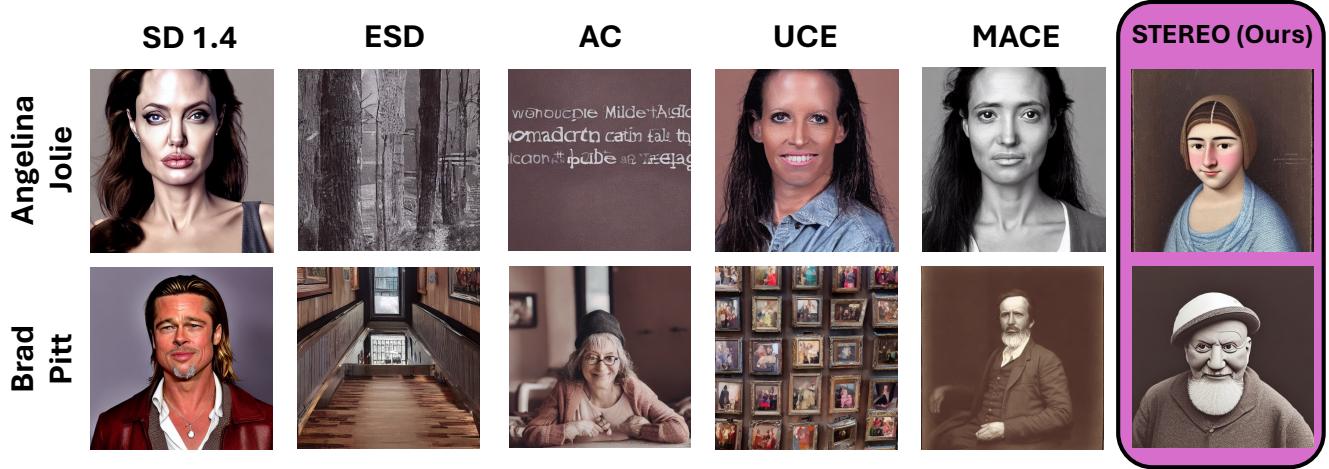


Figure 8: Visualization of celebrity identities erasure results across different methods.

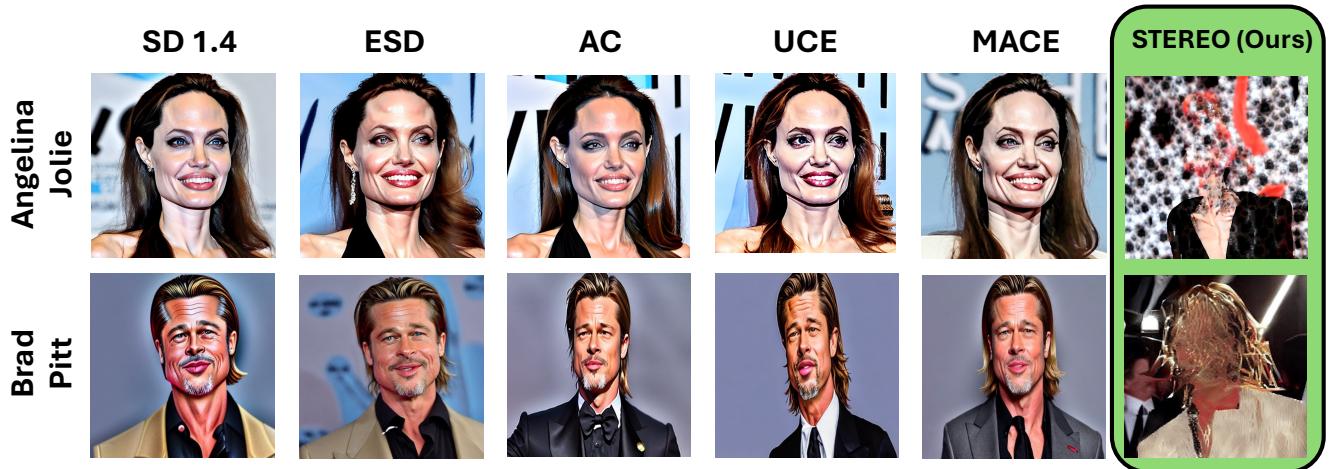


Figure 9: Visualization of celebrity identities under the CCE attack across different methods.



Figure 10: Visualization of nudity erasure results across different methods. (\*) added by authors for publication.



Figure 11: Visualization of nudity under RAB attack across different methods. (\*) added by authors for publication.

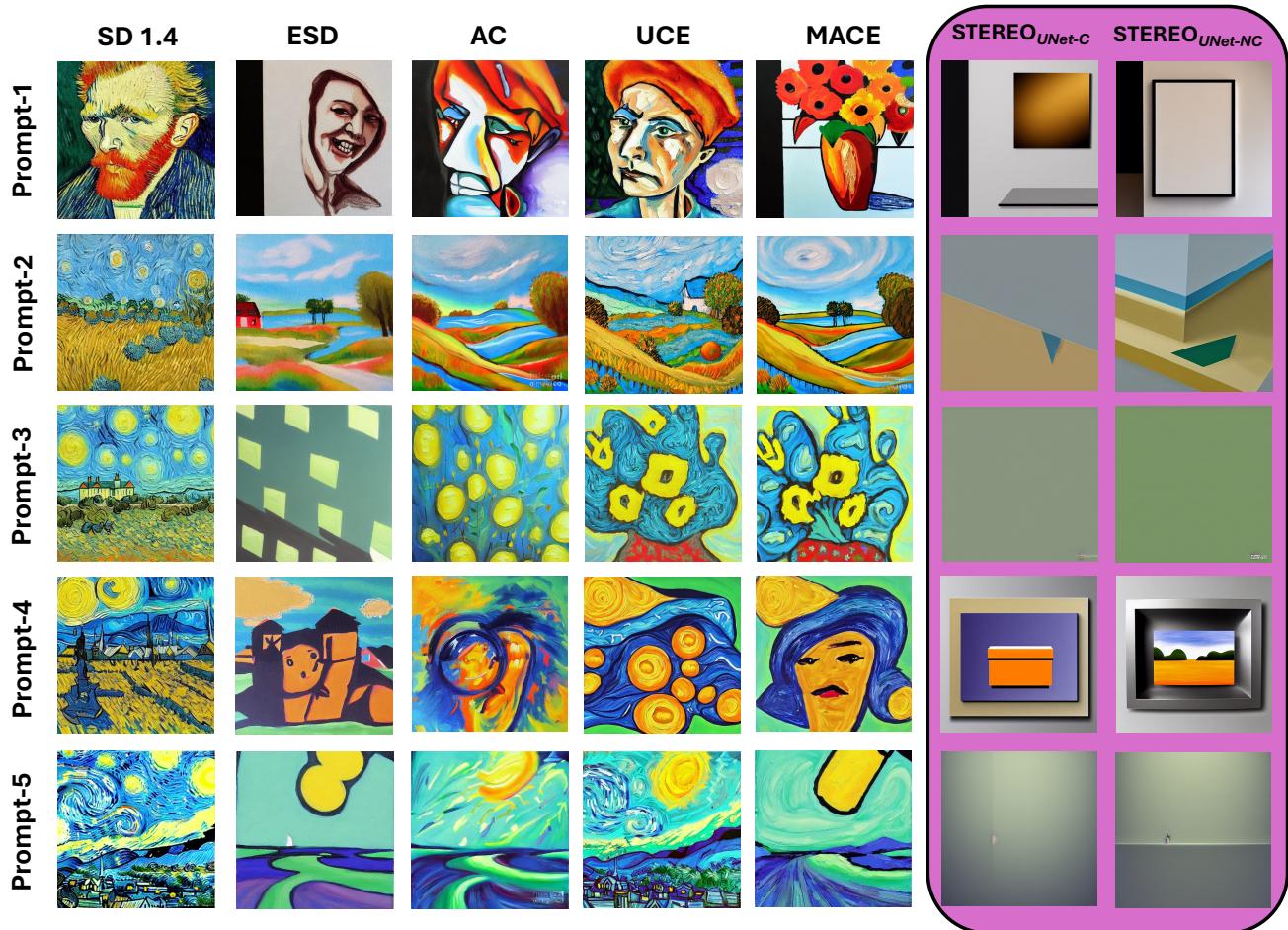


Figure 12: Visualization of Van-Gogh artistic style erasure across different methods.  $UNet-C$  indicates updating only the cross-attention layers of the UNet, and  $UNet-NC$  indicates updating only the non-cross-attention layers of the UNet. We use '*photorealistic minimalism*' as the anchor prompt for STEREO during erasing. We observe this to be reflected in the erased images when using the prompt "*A painting in the style of Van Gogh*".

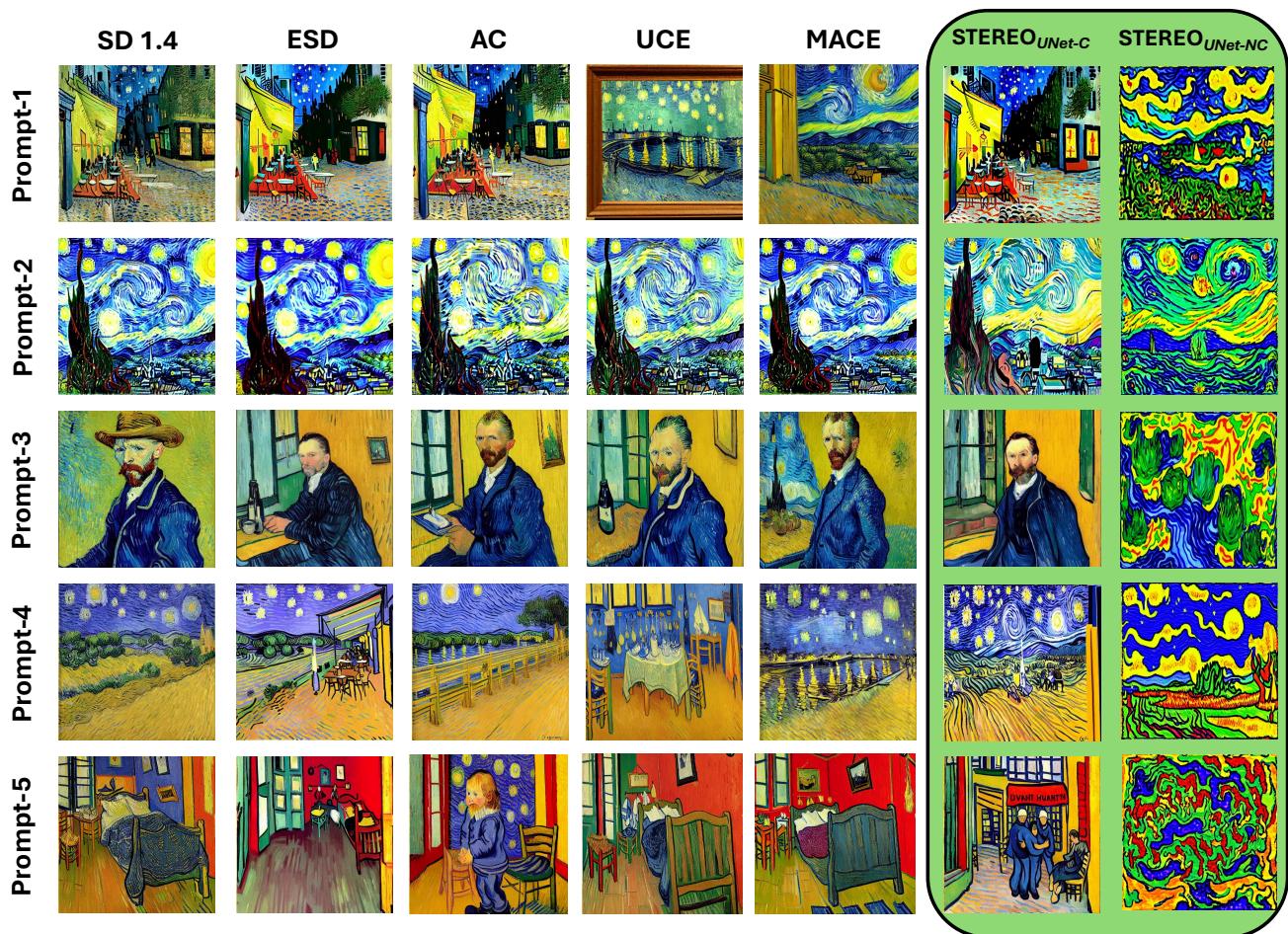


Figure 13: Visualization of Van-Gogh artistic style under CCE attack across different methods.  $\text{UNet-C}$  indicates updating only the cross-attention layers of the UNet, and  $\text{UNet-NC}$  indicates updating only the non-cross-attention layers of the UNet.