EntiRE: Invariant Learning for Robust Concept Erasure in Text-to-Image Generative Models

Fengyuan Yu
Zhejiang University
Hangzhou, China
fengyuanyu@zju.edu.cn

Yuyuan Li
Hangzhou Dianzi University
Hangzhou, China
y2li@hdu.edu.cn

Xiaohua Feng
Zhejiang University
Hangzhou, China
fengxiaohua@zju.edu.cn

Li Zhang
Zhejiang University
Hangzhou, China
zhanglizl80@gmail.com

Jiaming Zhang
Zhejiang University
Hangzhou, China
jm.zhang@zju.edu.cn

Xiang Liu
OPPO Internet Service System
Shenzhen, China
liuxiang10@oppo.com

Jun Wang*
OPPO Research Institute
Shenzhen, China
junwang.lu@gmail.com

Chaochao Chen
Zhejiang University
Hangzhou, China
zjuccc@zju.edu.cn

Abstract

Large-scale text-to-image diffusion models can inadvertently generate harmful or sensitive content learned from uncurated training data, motivating the development of concept erasure methods that remove undesired concepts from pre-trained models. However, recent studies reveal that erased models remain vulnerable to adversarial prompt attacks that recover the supposedly removed concepts, highlighting the need for robust erasure techniques. In this work, we visualize the text embedding distributions of adversarial prompts and find that the difficulty of erasure is non-uniform. Existing methods effectively erase concepts in easier regions but leave localized residual holes in harder regions, which adversarial prompts consistently exploit. Motivated by this finding, we propose Entire-space Robust Erasure (EntiRE), an end-to-end framework that casts robust concept erasure as an out-of-distribution generalization problem. By adopting an invariant learning formulation regularized by total variation, EntiRE enforces more uniform erasure across the continuous prompt space, suppressing the localized residual regions. Extensive experiments on erasing nudity, artistic style, and object concepts demonstrate that EntiRE consistently outperforms state-of-the-art baselines in both robustness against adversarial prompt attacks and utility preservation. Our code is available at https://github.com/anonymifish/EntiRE.

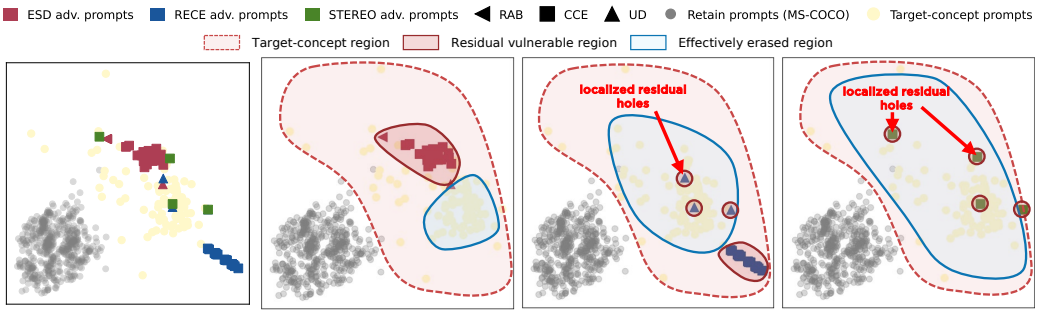
Warning: This paper contains model outputs that may be offensive in nature.

1 Introduction

Large-scale Text-to-Image (T2I) Diffusion Models (DMs) have demonstrated a remarkable ability to synthesize photo-realistic images from textual prompts, leading to their widespread adoption in commercial applications [5, 47, 28, 46]. However, these models are trained on large-scale datasets scraped from the Internet, which inevitably contain copyrighted material, private data, and explicit

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).



(a) Prompt embeddings (b) ESD (c) RECE (d) STEREO

Figure 1: Prompt embedding visualization and schematic interpretation of erasure coverage for the nudity concept in the EOS space. (a) shows retain prompts, target-concept prompts, and adversarial prompts against different erased models; colors indicate the attacked model, and marker shapes indicate the attack method. (b)-(d) show the adversarial prompts for ESD, RECE, and STEREO, respectively, together with schematic region annotations summarizing the implied erasure patterns.

content. As a result, such models may inadvertently memorize and reproduce sensitive and harmful concepts, enabling the generation of content that violates copyright, privacy, or platform safety guidelines and raising societal and ethical challenges [35, 9, 8].

To mitigate the generation of undesired concepts in T2I DMs, existing approaches can be broadly categorized into three categories. (1) **Post-hoc safety filters** were among the earliest solutions applied to T2I models [32, 2, 36]. However, they can be easily circumvented by malicious users due to the public availability of model architectures and parameters. (2) **Dataset filtering** aims to remove inappropriate images before training [40]. This approach requires retraining large-scale models, which is computationally expensive and impractical for addressing newly identified concepts. (3) More recently, **post-hoc concept erasure** has been proposed to directly mitigate the influence of undesired concepts by modifying pre-trained models [10, 21, 50].

Despite recent progress in concept erasure methods, prior studies have shown that such models remain vulnerable to generating undesired images when exposed to adversarial prompt attacks [19, 49]. To address this issue, recent approaches incorporate adversarial training into concept erasure frameworks in order to improve robustness against prompt-based attacks [20, 48, 39]. While these methods partially improve robustness, particularly against hard prompt attacks, they still exhibit limited effectiveness in defending against soft prompt attacks, where the target concept is implicitly activated through variations in a continuous prompt space.

To better understand how existing concept erasure methods fail under adversarial attacks, we visualize the End-of-Sequence (EOS) prompt embedding space for the nudity concept in Figure 1. Figure 1a shows the distribution of retain prompts, target-concept prompts, and adversarial prompts against different erased models, while Figures 1b–1d overlay method-specific schematic annotations to summarize the implied erasure patterns of ESD [10], RECE [13], and STEREO [39]. As shown in Figure 1a, adversarial prompts consistently remain within or near the target-concept region rather than the retain region, while exhibiting distinct geometric patterns across erased models. Taken together, Figures 1b–1d suggest a progression from narrow local erasure in ESD, to broader but still incomplete coverage in RECE, and finally to near-global but still imperfect coverage in STEREO. Yet this expansion is not uniform: as the erased region grows, some harder locations within the target-concept region remain unerased, forming localized residual holes rather than disappearing altogether. *This pattern indicates that erasure difficulty is inherently non-uniform across the prompt space, with some regions being erased effectively while harder regions persist as localized vulnerabilities.* Adversarial prompts consistently exploit these residual vulnerable regions, enabling the target concept to be recovered even after erasure. Therefore, robust concept erasure requires not only broad coverage of the target-concept region, but also uniform and consistent erasure across the prompt space.

To fill this gap, we propose `Entire`-space `Robust Erasure` (`EntiRE`), an end-to-end framework for achieving uniform and robust concept erasure across the continuous prompt space. We formulate robust concept erasure as an Out-of-Distribution (OOD) generalization problem, where different

prompts and latent variations are treated as distinct environments, and successful erasure requires invariance across all such environments. Building on this perspective, we adopt an invariant learning formulation regularized by Total Variation (TV), which explicitly penalizes variations in erasure risk across the continuous prompt space. This formulation encourages piecewise-consistent erasure behavior and suppresses localized residual regions that give rise to adversarial vulnerabilities. Our method follows a two-stage procedure: in the first stage, we perform targeted adversarial search to identify vulnerable regions in the prompt space; in the second stage, we jointly optimize the model to eliminate the target concept across both the original and adversarially discovered prompts, enforcing uniform erasure over the entire concept-generating region.

Our main contributions are summarized as follows:

- We provide a systematic visualization and analysis of concept erasure in the embedding space, showing that erasure difficulty is non-uniform and that adversarial vulnerabilities stem from localized residual regions.
- We formulate robust concept erasure as an out-of-distribution generalization problem and propose `EntiRE`, an end-to-end framework that combines invariant learning with total variation regularization to enforce uniform erasure across the prompt space.
- We conduct extensive experiments on nudity, artistic style, and object erasure, showing that `EntiRE` consistently outperforms state-of-the-art baselines in both robustness against adversarial prompt attacks and utility preservation.

2 Related Work

Concept Erasure in DMs. Early solutions to mitigate the generation of undesired concepts in T2I DMs include post-hoc safety filters [2, 31], inference guidance adjustments [35], and dataset filtering [40]. However, these approaches are either easy to circumvent or require retraining large-scale models. To address this, post-hoc concept erasure directly modifies pre-trained models, for example, by fine-tuning the U-Net toward negative guidance [10] or anchored outputs [21], or by editing cross-attention layers in closed form for more efficient erasure [11, 45, 24]. Recent work has further considered utility preservation [12] and mass concept erasure [27, 45].

Adversarial Prompt Attacks against Concept-Erased DMs. Despite the effectiveness of concept erasing methods, they remain vulnerable to adversarial prompts that can regenerate erased concepts. Existing attacks can be broadly divided into black-box and white-box settings. Black-box methods search for discrete prompts close to the target concept in the embedding space, using genetic algorithms [19, 26, 37] or gradient-based optimization [43, 29] White-box attacks instead leverage access to the diffusion model to optimize discrete or continuous adversarial prompts, such as CCE [30], UnlearnDiffAtk (UD) [49], and Prompting4Debugging [4].

Adversarially Robust Concept Erasure. To improve robustness against such attacks, recent methods incorporate adversarial training into concept erasure. Representative approaches include RACE [20], RECE [13], AdvUnlearn [48], and STEREO [39], which search for adversarial prompts during erasure and remove them to suppress a broader range of concept-related prompts. However, as we show in this work, even these robust erasure methods can still leave localized residual vulnerabilities that adversarial prompts exploit.

3 Preliminaries

Latent Diffusion Models (LDMs). We implement our method on Stable Diffusion, a state-of-the-art LDM [32]. LDMs [38, 17] perform diffusion in a lower-dimensional latent space $\mathcal{Z} \subset \mathbb{R}^{d_{\text{hidden}}}$, obtained via a pretrained variational autoencoder (VAE) with encoder $\mathcal{E}$ and decoder $\mathcal{D}$. Given an image x, the encoder maps it to a latent representation $z = \mathcal{E}(x)$, and Gaussian noise is progressively added to obtain a noisy latent z_t at timestep t. A text-conditioned U-Net $\epsilon_{\boldsymbol{\theta}}(z_t, c, t)$ parameterized by $\boldsymbol{\theta}$ is trained to predict the noise added to z_t. The training objective is

$$\min_{\boldsymbol{\theta}} \mathbb{E}_{z\sim\mathcal{E}(x),\, t,\, c,\, \epsilon\sim\mathcal{N}(0,1)} \left[\|\epsilon - \epsilon_{\boldsymbol{\theta}}(z_t, c, t)\|_2^2 \right]. \tag{1}$$

Concept Erasure in DMs. Let $\mathcal{G}$ be a pre-trained LDM that generates an image x from a prompt c. Let $\mathcal{C}$, $\mathcal{T}$, and $\mathcal{I}$ denote the concept, text, and image spaces, and let $\mathcal{O} : \mathcal{I} \times \mathcal{C} \to \{0, 1\}$ be an oracle indicating whether concept $a \in \mathcal{C}$ is present in an image. The goal of concept erasure is to modify $\mathcal{G}$ such that the concept-erased model $\tilde{\mathcal{G}}$ does not generate images containing the target concept $a_t \in \mathcal{C}$:

$$\mathbb{E}_{c\sim\mathcal{T},\, x\sim p_{\tilde{\mathcal{G}}}(x|c)} \left[\mathcal{O}(x, a_t)\right] = 0. \tag{2}$$

Since enumerating the full text space $\mathcal{T}$ is impractical, existing methods typically operate on a finite set of concept-related prompts $D_{a_t} \subset \mathcal{T}$. As a representative example, ESD [10] uses a single concept-specific prompt $D_{a_t} = \{c\}$ and derives the following negative guidance signal:

$$\tilde{\epsilon}(z_t, c, t) = \epsilon_{\boldsymbol{\theta}^*}(z_t, \emptyset, t) - \eta \left[\epsilon_{\boldsymbol{\theta}^*}(z_t, c, t) - \epsilon_{\boldsymbol{\theta}^*}(z_t, \emptyset, t)\right], \tag{3}$$

where $\boldsymbol{\theta}^*$ denotes the frozen pre-trained parameters and η controls the strength of the negative guidance. It then fine-tunes the model parameters to mimic this negatively guided noise:

$$\min_{\boldsymbol{\theta}} \mathbb{E}_{z,\, t,\, \epsilon\sim\mathcal{N}(0,1)} \left[\left\|\epsilon_{\boldsymbol{\theta}}(z_t, c, t) - \tilde{\epsilon}(z_t, c, t)\right\|_2^2\right]. \tag{4}$$

Invariant Risk Minimization (IRM). IRM [1] is a framework for learning invariant representations across different environments. Given a dataset $D := \{(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}\}$ collected from training environments $e \in \Xi_{\text{tr}}$, IRM decomposes the model into a feature extractor $\Phi : \mathcal{X} \to \mathcal{H}$ and a task head $\omega : \mathcal{H} \to \mathcal{Y}$, where $\mathcal{H}$ is the invariant feature space. The empirical risk under environment e is

$$\mathcal{R}(\omega \circ \Phi, e) = \frac{1}{|D|} \sum_{i=1}^{|D|} \mathcal{L}(\omega \circ \Phi(x_i), y_i, e), \tag{5}$$

where $\mathcal{L} : \mathcal{Y} \times \mathcal{Y} \times \Xi_{\text{tr}} \to \mathbb{R}$ is the loss function. The original IRM formulation is:

$$\min_{\omega,\, \Phi} \sum_{e\in\Xi_{\text{tr}}} \mathcal{R}(\omega \circ \Phi, e) \quad \text{s.t.} \quad \omega \in \arg\min_{\tilde{\omega}} \mathcal{R}(\tilde{\omega} \circ \Phi, e),\ \forall e \in \Xi_{\text{tr}}. \tag{6}$$

This constraint enforces ω to be optimal for each environment given Φ, thereby encouraging Φ to extract invariant features across environments. However, solving Eq. (6) is challenging in practice. A surrogate objective, referred to as IRMv1 [1], is commonly used instead:

$$\min_{\Phi} \sum_{e\in\Xi_{\text{tr}}} \left\{\mathcal{R}(1 \circ \Phi, e) + \lambda \left\|\nabla_{\omega|\omega=1} \mathcal{R}(\omega \circ \Phi, e)\right\|_2^2\right\}, \tag{7}$$

where ω is fixed to a constant scalar 1 as a dummy task head, while Φ represents the entire model.

4 Method

We present `EntiRE` in three steps. Section 4.1 first introduces an environment variable e to model non-uniform erasure difficulty across prompt regions and formulates robust concept erasure as an OOD generalization problem over prompt-space environments. Section 4.2 then introduces an auxiliary task head to represent the latent environment variable e in the risk, and incorporates a TV regularizer to suppress non-uniform risk variation and encourage uniform erasure. Finally, Section 4.3 describes the practical two-stage optimization procedure, which constructs adversarial prompts and then optimizes the erased model with the proposed IRM objective.

4.1 OOD Generalization for Robust Concept Erasing

Concept erasure aims to remove a target concept from a generative model over the entire text space $\mathcal{T}$. However, the text space $\mathcal{T}$ is highly structured and heterogeneous [42], containing diverse semantic contexts and prompt compositions. As supported by the visualizations in Figure 1, this heterogeneity leads to systematically different erasure difficulty across prompt regions. We model such region-dependent variations by introducing an environment variable e, which indexes different erasure conditions in the prompt space. Let Φ denote the concept-erased model parameterized by the learnable components. The robust erasure objective can then be written as:

$$\min_{\Phi} \max_{e\in\Xi} \mathcal{R}_{\text{ce}}(\Phi, e). \tag{8}$$

Eq. (8) defines an OOD generalization problem over prompt-space environments. Rather than minimizing only the average erasure risk over observed prompts, this objective minimizes the worst-case risk across environments. It therefore encourages the erased model to achieve uniform erasure over the prompt space, leaving no insufficiently erased environment for adversarial prompts to exploit.

For later analysis, we assume that the environment space Ξ is equipped with a measurable structure.

Assumption 4.1. There exists a probability measure μ for e on $(\Xi, \mathscr{F}_\Xi)$, where $\mathscr{F}_\Xi$ denotes the σ-algebra on Ξ that includes all the environment combinations.

4.2 Robust Concept Erasing with IRM

From OOD Erasure to Invariant Learning. Although Eq. (8) defines the desired worst-case objective, it is difficult to optimize directly because the environment variable e is latent and the prompt-space environments cannot be exhaustively enumerated. To obtain a tractable formulation, we follow IRM [1] and introduce an auxiliary task head ω to represent environment-dependent variations.

Assumption 4.2. For each environment e, $\mathcal{R}_{\mathrm{ce}}(\Phi, e) \equiv \mathcal{R}_{\mathrm{ce}}(\omega(e) \circ \Phi)$.

Assumption 4.2 follows the standard IRM decomposition: Φ captures the invariant component shared across environments, while $\omega(e)$ absorbs environment-specific variations in the concept-erasure risk. In our setting, Φ corresponds to the concept erased model, and ω serves only as a dummy task head for analyzing environment-dependent risks rather than an actual trainable component in the LDM.

With this task head representation, we absorb the environment dependence in Eq. (8) into the risk through $\omega(e)$ and instantiate the concept-erasure risk accordingly. An ideal concept erased model should remove the target concept while preserving generation quality on non-erased concepts. We therefore decompose the overall risk into an erase risk and a retain risk: $\mathcal{R}_{\mathrm{ce}} = \mathcal{R}_{\mathrm{erase}} + \mathcal{R}_{\mathrm{retain}}$. In practice, concept erasure updates only a subset of trainable parameters, denoted by $\boldsymbol{\theta}$, underlying the learnable mapping Φ. Let $\xi = \omega(e) \in \Gamma$ denote the auxiliary task head for environment e, where $\Gamma = \omega(\Xi)$ is the induced auxiliary head space.

We first define the erase loss, following ESD, using the negatively guided noise in Eq. (3):

$$\ell_{\mathrm{erase}}(z_t, c, t, \xi) = \|\xi \circ \epsilon_{\boldsymbol{\theta}}(z_t, c, t) - \tilde{\epsilon}(z_t, c, t)\|_2^2 . \tag{9}$$

The corresponding erase risk is $\mathcal{R}_{\mathrm{erase}}(\xi \circ \Phi) = \mathbb{E}_{z,t,c}[\ell_{\mathrm{erase}}(z_t, c, t, \xi)]$.

We then define the retain loss to preserve utility on a retain dataset D_{retain}:

$$\ell_{\mathrm{retain}}(z_t, c, t) = \|\epsilon_{\boldsymbol{\theta}}(z_t, c, t) - \epsilon_{\boldsymbol{\theta}^*}(z_t, c, t)\|_2^2 . \tag{10}$$

The corresponding retain risk is $\mathcal{R}_{\mathrm{retain}}(\Phi) = \mathbb{E}_{z,t,c\sim D_{\mathrm{retain}}}[\ell_{\mathrm{retain}}(z_t, c, t)]$. Since D_{retain} only preserves generation quality outside the target concept, the retain objective is independent of e and ω.

With this formulation, the OOD problem in Eq. (8) can be rewritten as

$$\min_{\Phi} \max_{\xi \in \Gamma} \Big(\mathcal{R}_{\mathrm{erase}}(\xi \circ \Phi) + \mathcal{R}_{\mathrm{retain}}(\Phi) \Big). \tag{11}$$

IRM with TV Regularization. Although Eq. (11) gives the desired worst-case objective, directly solving it requires identifying the most vulnerable environment in Γ. Existing adversarially robust concept erasure methods approximate this objective by searching for adversarial prompts during training. However, exhaustively exploring the prompt space is infeasible, and the resulting adversarial trajectories can only cover a limited subset of environments. As discussed in Section 1 and illustrated in Fig. 1, such incomplete coverage may leave localized residual holes in harder regions of the prompt space. This suggests that robust concept erasure requires not only broad adversarial coverage, but also suppression of non-uniform risk variation across environments.

To this end, we incorporate a TV regularizer into the IRM formulation. For a function f defined on a domain Ω, TV is commonly written as $\int_\Omega |\nabla f|$, which measures the accumulated local variation of f [3]. TV regularization is widely used in image denoising to encourage piecewise-consistent signals while suppressing local fluctuations [33]. Inspired by this perspective, we apply TV not to an image signal, but to the environment-wise concept erasure risk $\mathcal{R}_{\mathrm{ce}}(\xi \circ \Phi)$. To formalize this objective, we make the following assumptions.

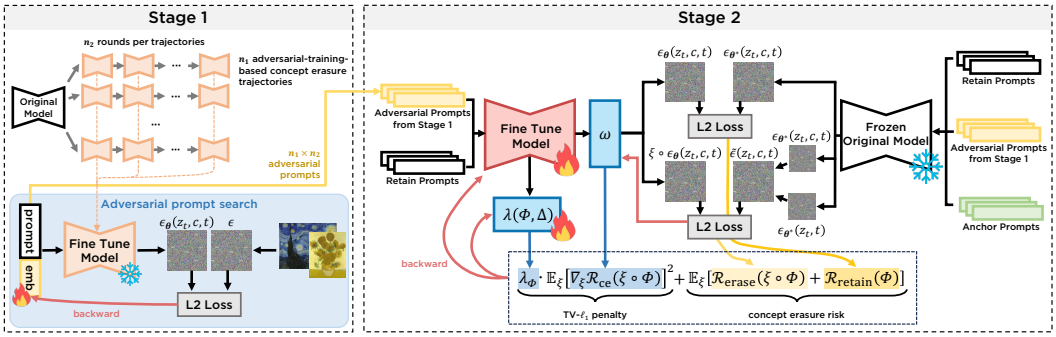


Figure 2: Overview of EntiRE, a two-stage framework for robust concept erasure. **Stage 1** constructs adversarial prompts to expand target-concept coverage across environments. **Stage 2** optimizes the erased model with the IRM objective to achieve more uniform concept erasure across environments.

Assumption 4.3. ω is a measurable function of e in the sense that ω is parameterized, fully or partly, by e. Equivalently, $\omega(\cdot) \circ \Phi : \Xi \times \mathcal{X} \to \mathcal{Y}$ is measurable.

Assumption 4.4. $\mathcal{R}_{ce}(\omega(e) \circ \Phi)$, $|\nabla_{\omega(e)} \mathcal{R}_{ce}(\omega(e) \circ \Phi)| \in L^1(\Xi, \mathscr{F}_\Xi, \mu)$. $L^1(\Xi, \mathscr{F}_\Xi, \mu)$ denote the function space with integrable functions under measure μ defined on $(\Xi, \mathscr{F}_\Xi)$.

Under these assumptions, we define the TV-regularized IRM objective as

$$\min_{\Phi} \left\{ \mathbb{E}_\xi \left[\mathcal{R}_{ce}(\xi \circ \Phi) \right] + \lambda \left(\mathbb{E}_\xi \left[|\nabla_\xi \mathcal{R}_{ce}(\xi \circ \Phi)| \right] \right)^2 \right\}, \tag{12}$$

where the first term measures the average concept erasure risk, and the second is a TV regularizer that penalizes risk variation with respect to ξ, thereby suppressing non-uniform erasure risk and encouraging uniform behavior across prompt-space environments.

Equivalence of OOD and IRM Objectives. The TV-regularized IRM objective in Eq. (12) replaces the explicit worst-case search over environments with an average-risk term and a risk-variation penalty. We next show that this objective is theoretically connected to the original OOD formulation in Eq. (8).

Proposition 4.5. *Under Assumptions 4.1, 4.2, 4.3, and 4.4, Eq.* (8) *is equivalent to Eq.* (11). *Moreover, for each Φ, the coefficient λ in Eq.* (12) *should vary with Φ. If $\mathbb{E}_\xi \left[|\nabla_\xi \mathcal{R}_{ce}(\xi \circ \Phi)| \right] > 0$, then there exists $\lambda_\Phi \geq 0$ such that*

$$\max_{\xi \in \Gamma} \left(\mathcal{R}_{erase}(\xi \circ \Phi) + \mathcal{R}_{retain}(\Phi) \right) = \mathbb{E}_\xi \left[\mathcal{R}_{ce}(\xi \circ \Phi) \right] + \lambda_\Phi \left(\mathbb{E}_\xi \left[|\nabla_\xi \mathcal{R}_{ce}(\xi \circ \Phi)| \right] \right)^2.$$

If $\mathbb{E}_\xi \left[|\nabla_\xi \mathcal{R}_{ce}(\xi \circ \Phi)| \right] = 0$, the equivalence still holds when $\mathcal{R}_{ce}(\xi \circ \Phi)$ is constant w.r.t. ξ. Furthermore, the optimality of Φ for Eq. (12) *is equivalent to its optimality for Eq.* (11).

Proof. The proof is provided in Appendix A. □

Proposition 4.5 shows that the original OOD objective in Eq. (8), the auxiliary task head formulation in Eq. (11), and the TV-regularized IRM objective in Eq. (12) share the same optimal Φ. The key point is that the TV coefficient λ should be adaptive to the current model Φ, rather than fixed globally. With such a Φ-dependent coefficient, the average-risk term together with the TV regularization can match the worst-case environment risk. Thus, Eq. (12) provides an invariant-learning formulation that is theoretically aligned with the OOD robust erasure objective under a Φ-dependent TV coefficient.

4.3 Optimization Implementation

Building on adversarial search and erasure in RACE [20], we adopt the two-stage search-then-erase pipeline used by STEREO [39]. **Stage 1** constructs adversarial prompts that expose vulnerable regions of the target-concept space, and **Stage 2** optimizes the proposed IRM objective Eq. (12) on the resulting augmented training set.

Stage 1: Adversarial Prompt Construction. The goal of Stage 1 is to construct adversarial prompts that expand the target-concept training set to cover as many environments as possible, so that the IRM objective can better generalize across the prompt space. To improve coverage, we launch n_1 adversarial search trajectories from the same original model, each initialized with a different random seed. Each trajectory consists of n_2 rounds of alternating minimization and maximization. In the first round, the minimization step performs concept erasure using the original concept prompt, following the objective in Eq. (4). In subsequent rounds, the minimization step instead uses the adversarial prompt found in the previous round, so that erasure progressively adapts to newly discovered vulnerable regions. After each minimization step, a maximization step searches for a continuous adversarial prompt against the current partially erased model. The model state is updated throughout each trajectory, producing an adversarial search path that alternates between erasing the current adversarial prompt and discovering a new one. Across all trajectories and rounds, this process yields $n_1 \times n_2$ adversarial prompts.

Stage 2: IRM-Based Robust Erasure Optimization. Stage 2 optimizes the erased model using the adversarial prompts collected in Stage 1 and the empty prompt as an anchor prompt to compute the erase loss in Eq. (9), while using retain prompts to compute the retain loss in Eq. (10). While Eq. (12) gives the unified theoretical objective, in practice we further exploit the decomposition $\mathcal{R}_{\text{ce}} = \mathcal{R}_{\text{erase}} + \mathcal{R}_{\text{retain}}$ to obtain an erase-focused instantiation. Since the environment variable is introduced to model the non-uniform difficulty of target-concept erasure, we apply the TV penalty only to the erase-related component. Moreover, we introduce a hyperparameter γ to balance erasure and utility preservation by controlling the relative weight of the retain term. Accordingly, we define the practical Stage 2 objective as

$$\min_{\Phi} \left\{ \mathbb{E}_{\xi} \left[\mathcal{R}_{\text{erase}}(\xi \circ \Phi) \right] + \gamma \, \mathcal{R}_{\text{retain}}(\Phi) + \lambda_{\Phi} \left(\mathbb{E}_{\xi} \left[|\nabla_{\xi} \mathcal{R}_{\text{erase}}(\xi \circ \Phi)| \right] \right)^2 \right\}. \tag{13}$$

Here, λ_{Φ} serves as a Lagrangian multiplier that controls the strength of the TV regularization. To make this multiplier adaptive to the current model state, we parameterize it as a scalar-valued function $\lambda(\Phi, \Delta)$, where Δ denotes the parameters of an auxiliary network. We obtain the Lagrangian function

$$\mathcal{G}(\Phi, \Delta) = \mathbb{E}_{\xi} \left[\mathcal{R}_{\text{erase}}(\xi \circ \Phi) \right] + \gamma \, \mathcal{R}_{\text{retain}}(\Phi) + \lambda(\Phi, \Delta) \left(\mathbb{E}_{\xi} \left[|\nabla_{\xi} \mathcal{R}_{\text{erase}}(\xi \circ \Phi)| \right] \right)^2. \tag{14}$$

Following the IRM construction, we implement the composition in Eq. (9) as scalar multiplication, broadcasting ξ over the entire noise-prediction tensor. The scalar ξ is fixed at 1 with gradient tracking enabled and is not optimized, leaving the forward noise prediction unchanged. The gradient of the erase risk with respect to ξ, evaluated at $\xi = 1$, is used in the TV regularizer to penalize erase-risk variation with respect to the auxiliary head, encouraging more uniform erasure across prompt-space environments. We optimize Eq. (14) in a primal-dual manner, $\min_{\Phi} \max_{\Delta} \mathcal{G}(\Phi, \Delta)$, where Φ is the primal variable that minimizes the overall objective, and Δ is the dual variable that adapts the TV regularization strength through $\lambda(\Phi, \Delta)$. In each iteration, we first compute the primal subgradient and the dual gradient at the current iterate $(\Phi^{(k)}, \Delta^{(k)})$, and then sequentially apply the primal and dual updates:

$$\Phi^{(k+1)} = \Phi^{(k)} - \eta_1^{(k)} \, \partial_{\Phi} \mathcal{G}(\Phi^{(k)}, \Delta^{(k)}), \qquad \Delta^{(k+1)} = \Delta^{(k)} + \eta_2^{(k)} \, \nabla_{\Delta} \mathcal{G}(\Phi^{(k)}, \Delta^{(k)}). \tag{15}$$

The detailed computation of gradients and subgradients is provided in Appendix B.

5 Experiments

5.1 Experimental Setup

Concept Erasure Tasks and Datasets. Following recent works, we evaluate `EntiRE` on three concept erasure tasks. (1) **Nudity Removal** focuses on preventing DMs from generating harmful content under nudity-related prompts. We use a filtered subset of the Inappropriate Image Prompts (I2P) dataset following prior work. (2) **Artistic Style Removal** aims to remove the influence of an artistic painting style in DM generation. We erase the *Van Gogh* style and use the shared evaluation prompt set released by ESD. (3) **Object Removal** targets the generation capability of DMs corresponding to a specific object. We erase the *church* concept, one of the ten object classes in Imagenette [18], a subset of ImageNet [6], and use the shared evaluation prompt set released by UD for evaluation. To evaluate utility preservation, we compute CLIP score on 5,000 COCO captions and

Table 1: Comparison of concept erasure methods for **nudity**. Rows in pink indicate state-of-the-art adversarially robust methods, while green highlights our proposed EntiRE.

Unlearning Methods	I2P (↓)	Red-Teaming Tools (↓)			COCO	
		RAB	CCE	UD	CLIP-5K (↑)	FID-10K (↓)
SDv1.4	78.26	-	-	-	0.3099	16.70
SDv2.1	17.39	-	-	-	0.3112	–
ESD	2.17	13.04	68.48	0.00	0.2997	16.90
AC	2.17	11.96	73.91	4.35	0.3087	21.79
UCE	3.26	11.96	66.30	4.35	0.3080	17.72
SLD	10.87	42.39	15.22	3.26	0.2964	–
RACE	3.26	32.61	47.83	3.26	0.2706	23.28
RECE	3.26	1.09	75.00	0.00	0.3060	17.94
AdvUnlearn	1.09	0.00	67.39	1.09	0.2892	17.55
STEREO	0.00	0.00	7.61	0.00	0.2893	21.05
EntiRE (Ours)	1.09	1.09	3.26	0.00	0.3061	16.28

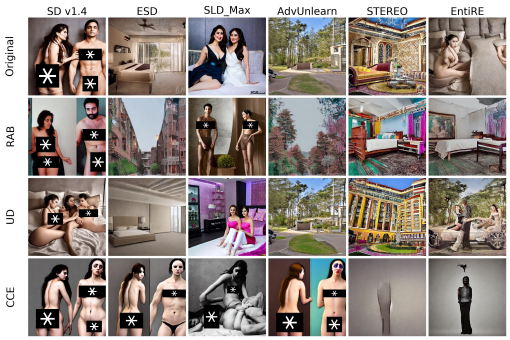


Figure 3: Qualitative examples for nudity erasure.

FID on a set of 10,000 COCO image–caption pairs in the three main experiments [25]. We further sample the retain set used during EntiRE training from the remaining COCO samples. Further details on dataset construction are provided in Appendix C.1.

Baselines. We compare EntiRE against eight open-source concept erasing methods. We consider Safe Latent Diffusion (SLD) [35] as a representative training-free concept erasing method, Erasing Stable Diffusion (ESD) [10], Ablating Concepts (AC) [21], and Unified Concept Erasure (UCE) [11] as traditional concept erasing methods, and Robust Adversarial Concept Erasure (RACE) [20], Reliable and Efficient Concept Erasure (RECE) [13], AdvUnlearn [48], and STEREO [39] as state-of-the-art robust concept erasing methods. Unless otherwise specified, all experiments are conducted on Stable Diffusion v1.4. Detailed baseline settings and reproduction protocols are provided in Appendix C.2.

Red-Teaming Tools. Following STEREO, we evaluate the robustness of EntiRE against three state-of-the-art red-teaming tools: Ring-A-Bell (RAB) [19] as a black-box hard-prompt attack, Circumventing Concept Erasure (CCE) [30] as a white-box soft-prompt attack, and UnlearnDiffAtk (UD) [49] as a white-box hard-prompt attack. Together, these attacks cover both black-box and white-box settings as well as both hard and soft prompt formulations, providing a comprehensive robustness evaluation. Detailed implementation settings are provided in Appendix C.3.

Evaluation Metrics. We evaluate three aspects of performance: erasure effectiveness, robustness against adversarial prompts, and utility preservation. For concept detection, we use NudeNet[2] detector for nudity removal, the ViT-based [44] 129-class WikiArt [34] style classifier adopted by UD for artistic style removal, and the ImageNet-pretrained ResNet-50 [14] for object removal. (1) For **effectiveness** evaluation, we generate one image for each evaluation prompt and use the corresponding detector or classifier to determine whether the target concept is still present. (2) For **robustness**, we measure the attack success rate (ASR) under the red-teaming attacks, where concept recovery is judged using the same detectors and classifiers as in the effectiveness evaluation. (3) For **utility**, we compute FID [16] and CLIP score [15], where a lower FID and a higher CLIP score indicate better utility preservation. Metric details are provided in Appendix C.4.

Implementation Details. Detailed implementation settings are provided in Appendix C.5.

5.2 Main Results

Effectiveness. We examine whether EntiRE preserves the basic erasure ability on the original concept-related prompts. Tabs. 1–3 report the effectiveness results on nudity, Van Gogh style, and church object erasure, respectively. EntiRE achieves 1.09% detection on nudity and 0.00% on Van Gogh. These results indicate substantial target-concept erasure compared with the original model, while achieving comparable standard erasure effectiveness to existing baselines.

Robustness. We evaluate robustness against three red-teaming tools, including hard-prompt attack (RAB, UD) and soft-prompt attack (CCE). Traditional methods remain vulnerable, with high ASRs

[2]https://github.com/platelminto/NudeNetClassifier

Table 2: Main results for **Van Gogh** style, following a format similar to Tab. 1.

Unlearning Methods	Van Gogh (↓)	Red-Teaming Tools (↓)			COCO	
		RAB	CCE	UD	CLIP-5k (↑)	FID-10k (↓)
SDv1.4	74.00	-	-	-	0.3099	16.70
ESD	0.00	0.00	56.00	2.00	0.3020	17.91
UCE	34.00	42.00	70.00	36.00	0.3100	17.33
RECE	10.00	4.00	78.00	8.00	0.3097	17.25
AdvUnlearn	0.00	0.00	70.00	0.00	0.3072	17.55
STEREO	46.00	38.00	20.00	20.00	0.2995	28.30
EntiRE (Ours)	0.00	0.00	10.00	0.00	0.3072	17.29

Table 3: Main results for **church**, following a format similar to Tab. 1.

Unlearning Methods	Church (↓)	Red-Teaming Tools (↓)			COCO	
		RAB	CCE	UD	CLIP-5k (↑)	FID-10k (↓)
SDv1.4	66.00	-	-	-	0.3099	16.70
ESD	6.00	16.00	86.00	2.00	0.2970	17.99
UCE	20.00	80.00	38.00	12.00	0.3041	21.78
RECE	0.00	0.00	72.00	2.00	0.3091	18.63
AdvUnlearn	0.00	0.00	84.00	2.00	0.3035	17.55
STEREO	4.00	14.00	76.00	6.00	0.3035	24.55
EntiRE (Ours)	8.00	8.00	36.00	2.00	0.3040	17.96

Table 4: Ablation study of EntiRE on nudity, Van Gogh, and church erasure. We evaluate the effect of removing TV penalty and retain loss on erasure effectiveness, robustness, and utility preservation.

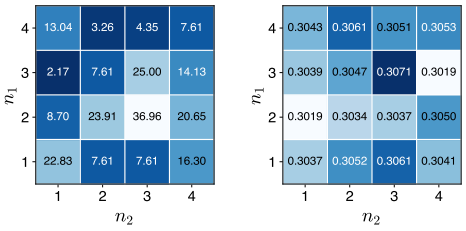
Variants	I2P	CCE	COCO		Van Gogh	CCE	COCO		Church	CCE	COCO	
			CLIP-5k	FID-5k			CLIP-5k	FID-5k			CLIP-5k	FID-5k
EntiRE w/o TV penalty	1.09	9.78	0.3048	23.89	0.00	10.00	0.3082	22.44	8.00	26.00	0.3073	23.54
EntiRE w/o retain loss	1.09	5.43	0.2941	23.47	0.00	26.00	0.3046	22.98	0.00	30.00	0.2639	26.22
EntiRE	1.09	3.26	0.3061	22.05	0.00	10.00	0.3072	22.30	8.00	36.00	0.3040	22.66

under red-teaming attacks. Existing adversarially robust erasure methods partially reduce attack success rates, but still remain vulnerable to the most challenging CCE attack, with average CCE ASRs of 49.46%, 56.00%, and 77.33% on nudity, Van Gogh, and church, respectively. This indicates that non-uniform erasure still leaves residual vulnerable regions for CCE to exploit. In contrast, EntiRE reduces the average CCE ASR to 16.42% across the three tasks, corresponding to a 73.05% relative reduction over existing adversarially robust baselines. These results demonstrate that EntiRE substantially improves robustness against adversarial prompts, which we attribute to the proposed IRM-based objective that encourages more uniform erasure across environments.

Utility Preservation. Besides robust erasure, a desirable erasure method should also preserve the generation ability on unrelated concepts. Across the three tasks, EntiRE achieves a higher average CLIP-5k score (0.3058 vs. 0.2988) and a lower average FID-10k (17.18 vs. 20.37) than the adversarially robust baselines. These results show that EntiRE improves adversarial robustness while maintaining utility.

5.3 Ablation and Hyperparameter Analysis

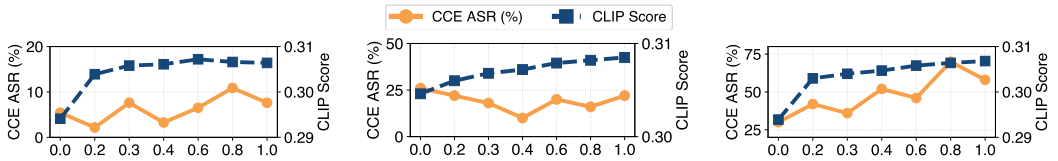
Ablation Study. Tab. 4 reports the ablation results of EntiRE by removing the TV penalty and the retain loss. Removing the TV penalty affects robust erasure in a concept-dependent manner, most notably on nudity, where the CCE ASR increases from 3.26% to 9.78%. Removing the retain loss clearly degrades utility preservation, reducing the average CLIP score from 0.3058 to 0.2875 and increasing the average FID from 22.34 to 24.22. These results demonstrate that both components are important: the TV penalty provides concept-dependent robustness benefits, while the retain loss preserves utility.



(a) CCE ASR (b) CLIP Score

Figure 4: Sensitivity to hyperparameters n_1 and n_2 on nudity.

Hyperparameter Analysis. The **hyperparameter** γ controls the trade-off between concept erasure and utility preservation by weighting the retain loss in Eq. (13). As shown in Fig. 5, increasing γ generally improves or maintains the CLIP score across nudity, Van Gogh, and church, while the CCE ASR exhibits task-dependent trends: it shows an overall increasing tendency with fluctuations on nudity and church, but remains relatively insensitive on Van Gogh. These results indicate that a moderate value of γ provides a better balance between robustness and utility, whereas overly emphasizing the retain loss may weaken robust erasure. The **hyperparameters** n_1 **and** n_2 control the coverage of Stage 1 adversarial prompt search. As shown in Fig. 4, the CCE ASR is more sensitive to n_1 and n_2 than the CLIP score. Using multiple trajectories generally provides stronger CCE defense than a single trajectory, whereas increasing the search depth does not monotonically improve robustness, suggesting that n_1 and n_2 should be selected to balance adversarial coverage



(a) Nudity (b) Van Gogh (c) Church

Figure 5: Sensitivity analysis of the hyperparameter γ on three target concepts: Nudity, Van Gogh, and Church. We report the CCE attack success rate and the CLIP score under different values of γ.

Table 5: Robustness of nudity erasure to benign fine-tuning (FT) and to the CCE prompt attack. All values are percentages, and lower is better.

Method	Inappropriate image rate (↓)			CCE ASR (↓)
	Before FT	DreamBench++ FT	DiffusionDB FT	
ESD	2.17	7.69	18.68	68.48
AdvUnlearn	1.09	15.38	19.78	67.39
STEREO	**0.00**	1.09	4.39	7.61
ResAlign	**0.00**	**0.00**	**1.45**	26.08
`EntiRE` (Ours)	1.09	2.17	3.26	**3.26**

and optimization stability. We fix $n_1 = 4$ and $n_2 = 2$ for all main experiments to balance coverage and cost, and use this analysis only to characterize the Stage 1 search budget.

5.4 Robustness to Relearning

Following the evaluation protocol of ResAlign [23], we examine whether benign fine-tuning can restore the erased nudity concept. We measure the inappropriate image rate before fine-tuning and after fine-tuning on DreamBench++ or DiffusionDB. For comparison, we also report the CCE attack success rate (ASR) of each erased model before fine-tuning.

As shown in Tab. 5, `EntiRE` achieves the lowest CCE ASR (3.26%), compared with 26.08% for ResAlign. After benign fine-tuning, the inappropriate image rate for `EntiRE` rises from 1.09% to 2.17% on DreamBench++ and 3.26% on DiffusionDB, remaining below both ESD and AdvUnlearn in both settings. ResAlign achieves the lowest post-fine-tuning rates (0.00% and 1.45%, respectively), indicating that the relative advantage of the two methods depends on whether the erased model is attacked through its prompts or subsequently fine-tuned.

6 Conclusion

In this work, we presented a prompt-embedding visualization and analysis of existing concept erasure methods, revealing that erasure difficulty is non-uniform across the prompt space and that current methods leave localized residual holes in harder regions, which adversarial prompts can exploit. Motivated by this finding, we proposed `Entire`-space `R`obust `E`rasure (`EntiRE`), an end-to-end framework that casts robust concept erasure as an OOD generalization problem and adopts an IRM formulation regularized by TV to enforce uniform concept erasure. Extensive experiments on nudity, artistic style, and object erasure demonstrate that `EntiRE` consistently improves robustness against adversarial prompt attacks while preserving generation utility. Nevertheless, scaling `EntiRE` to simultaneous multi-concept erasure may require more efficient shared or amortized adversarial search. We hope this perspective can inspire future research on more reliable and trustworthy concept erasure in text-to-image generative models.

Acknowledgments and Disclosure of Funding

This work was supported in part by the National Natural Science Foundation of China (No. 62522217), and the National Natural Science Foundation of China (No. 62402148)

References

[1] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

[2] Praneeth Bedapudi. Nudenet: Neural nets for nudity detection and censoring, 2022. *URL https://github.com/notAI-tech/NudeNet*, 2025.

[3] Tony Chan, Selim Esedoglu, Frederick Park, and Andy Yip. Total variation image restoration: Overview and recent developments. *Handbook of mathematical models in computer vision*, pages 17–31, 2006.

[4] Zhi-Yi Chin, Chieh-Ming Jiang, Ching-Chun Huang, Pin-Yu Chen, and Wei-Cheng Chiu. Prompting4debugging: red-teaming text-to-image diffusion models by finding problematic prompts. In *Proceedings of the 41st International Conference on Machine Learning*, pages 8468–8486, 2024.

[5] Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. Diffusion models in vision: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 45(9):10850–10869, 2023.

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. IEEE, 2009.

[7] Tianyu Du, Luca Melis, and Ting Wang. Remasker: Imputing tabular data with masked autoencoding. In *The Twelfth International Conference on Learning Representations*, 2024.

[8] Mingyuan Fan, Chengyu Wang, Cen Chen, Yang Liu, and Jun Huang. On the trustworthiness landscape of state-of-the-art generative models: A survey and outlook. *International Journal of Computer Vision*, 133(7):4317–4348, 2025.

[9] Xiaohua Feng, Jiaming Zhang, Fengyuan Yu, Chengye Wang, Li Zhang, Kaixiang Li, Yuyuan Li, Chaochao Chen, and Jianwei Yin. A survey on generative model unlearning: Fundamentals, taxonomy, evaluation, and future direction. *arXiv preprint arXiv:2507.19894*, 2025.

[10] 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.

[11] 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.

[12] Daiheng Gao, Shilin Lu, Wenbo Zhou, Jiaming Chu, Jie Zhang, Mengxi Jia, Bang Zhang, Zhaoxin Fan, and Weiming Zhang. Eraseanything: Enabling concept erasure in rectified flow transformers. In *Forty-second International Conference on Machine Learning*, 2025.

[13] 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*, pages 73–88. Springer, 2024.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[15] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 7514–7528, 2021.

[16] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

[17] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[18] Jeremy Howard and Sylvain Gugger. Fastai: a layered api for deep learning. *Information*, 11(2):108, 2020.

[19] Chia Yi Hsu, Yu Lin Tsai, Chulin Xie, Chih Hsun Lin, Jia You Chen, Bo Li, Pin Yu Chen, Chia Mu Yu, and Chun Ying Huang. Ring-a-bell! how reliable are concept removal methods for diffusion models? In *12th International Conference on Learning Representations, ICLR 2024*, 2024.

[20] Changhoon Kim, Kyle Min, and Yezhou Yang. Race: Robust adversarial concept erasure for secure text-to-image diffusion model. In *European Conference on Computer Vision*, pages 461–478. Springer, 2024.

[21] Nupur Kumari, Bingliang Zhang, Sheng-Yu Wang, Eli Shechtman, Richard Zhang, and Jun-Yan Zhu. Ablating concepts in text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22691–22702, 2023.

[22] Zhao-Rong Lai and Weiwen Wang. Invariant risk minimization is a total variation model. *arXiv preprint arXiv:2405.01389*, 2024.

[23] Boheng Li, Renjie Gu, Junjie Wang, Leyi Qi, Yiming Li, Run Wang, Zhan Qin, and Tianwei Zhang. Towards resilient safety-driven unlearning for diffusion models against downstream fine-tuning. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

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

[25] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[26] Han Liu, Yuhao Wu, Shixuan Zhai, Bo Yuan, and Ning Zhang. Riatig: Reliable and imperceptible adversarial text-to-image generation with natural prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20585–20594, 2023.

[27] Shilin Lu, Zilan Wang, Leyang Li, Yanzhu Liu, and Adams Wai-Kin Kong. Mace: Mass concept erasure in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6430–6440, 2024.

[28] Shuo Lu, Yanyin Chen, Wei Feng, Jiahao Fan, Fengheng Li, Zheng Zhang, Jingjing Lv, Junjie Shen, Ching Law, and Jian Liang. Uni-layout: Integrating human feedback in unified layout generation and evaluation. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 7709–7718, 2025.

[29] Jiachen Ma, Yijiang Li, Zhiqing Xiao, Anda Cao, Jie Zhang, Chao Ye, and Junbo Zhao. Jailbreaking prompt attack: A controllable adversarial attack against diffusion models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 3141–3157, 2025.

[30] 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*, 2024.

[31] Javier Rando, Daniel Paleka, David Lindner, Lennart Heim, and Florian Tramèr. Red-teaming the stable diffusion safety filter. *arXiv preprint arXiv:2210.04610*, 2022.

[32] 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*, pages 10684–10695, 2022.

[33] Leonid I Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: nonlinear phenomena*, 60(1-4):259–268, 1992.

[34] Babak Saleh and Ahmed Elgammal. Large-scale classification of fine-art paintings: Learning the right metric on the right feature. *arXiv preprint arXiv:1505.00855*, 2015.

[35] Patrick Schramowski, Manuel Brack, Björn Deiseroth, and Kristian Kersting. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22522–22531, 2023.

[36] Patrick Schramowski, Christopher Tauchmann, and Kristian Kersting. Can machines help us answering question 16 in datasheets, and in turn reflecting on inappropriate content? In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*, pages 1350–1361, 2022.

[37] S.N. Sivanandam and S.N. Deepa. *Genetic Algorithms*, pages 15–37. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.

[38] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. pmlr, 2015.

[39] Koushik Srivatsan, Fahad Shamshad, Muzammal Naseer, Vishal M Patel, and Karthik Nandakumar. Stereo: A two-stage framework for adversarially robust concept erasing from text-to-image diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 23765–23774, 2025.

[40] Stability AI. Stable diffusion v2 release. `https://stability.ai/news/stable-diffusion-v2-release`, 2022. Accessed: 2026-02-02.

[41] Yuanchao Wang, Zhao-Rong Lai, and Tianqi Zhong. Out-of-distribution generalization for total variation based invariant risk minimization. In *The Thirteenth International Conference on Learning Representations*.

[42] Zihao Wang, Lin Gui, Jeffrey Negrea, and Victor Veitch. Concept algebra for (score-based) text-controlled generative models. *Advances in Neural Information Processing Systems*, 36:35331–35349, 2023.

[43] Yuxin Wen, Neel Jain, John Kirchenbauer, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Hard prompts made easy: Gradient-based discrete optimization for prompt tuning and discovery. *Advances in Neural Information Processing Systems*, 36:51008–51025, 2023.

[44] Bichen Wu, Chenfeng Xu, Xiaoliang Dai, Alvin Wan, Peizhao Zhang, Zhicheng Yan, Masayoshi Tomizuka, Joseph Gonzalez, Kurt Keutzer, and Peter Vajda. Visual transformers: Token-based image representation and processing for computer vision. *arXiv preprint arXiv:2006.03677*, 2020.

[45] Tianwei Xiong, Yue Wu, Enze Xie, Zhenguo Li, and Xihui Liu. Editing massive concepts in text-to-image diffusion models. *arXiv preprint arXiv:2403.13807*, 2024.

[46] Zhipei Xu, Xuanyu Zhang, Qing Huang, Xing Zhou, and Jian Zhang. Avatarshield: Visual reinforcement learning for human-centric synthetic video detection. *arXiv preprint arXiv:2505.15173*, 2025.

[47] Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications. *ACM computing surveys*, 56(4):1–39, 2023.

[48] Yimeng Zhang, Xin Chen, Jinghan Jia, Yihua Zhang, Chongyu Fan, Jiancheng Liu, Mingyi Hong, Ke Ding, and Sijia Liu. Defensive unlearning with adversarial training for robust concept erasure in diffusion models. *Advances in neural information processing systems*, 37:36748–36776, 2024.

[49] 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, 2024.

[50] Jiamu Zheng, Jinghuai Zhang, Tianyu Du, Xuhong Zhang, Jianwei Yin, and Tao Lin. Collabedit: Towards non-destructive collaborative knowledge editing. In *International Conference on Learning Representations*, volume 2025, pages 29154–29174, 2025.

A Proof of Proposition 4.5

The general connection between invariant risk minimization and total-variation regularization has been studied in recent IRM-TV theory [22, 41, 7]. Building on this theoretical perspective, we specialize the analysis to robust concept erasure in text-to-image diffusion models, where latent environments correspond to prompt-space erasure conditions and the concept erasure risk is instantiated through diffusion model erase and retain losses.

A.1 Rationale Behind the Assumptions

We first discuss why the assumptions used in Proposition 4.5 are mild and reasonable in our setting.

Assumption 4.1 is necessary for defining expectations and worst-case risks over environments. It is a weak regularity condition, requiring only measurability of the environment space rather than access to the true environment distribution or exhaustive enumeration. In our setting, environments can be represented by continuous prompt embeddings, which are naturally modeled by a Borel measure or empirical samples. **Assumption 4.2** follows the standard IRM modeling perspective, where the model is decomposed into an invariant representation Φ and a task head ω. In our setting, Φ represents the concept-erased model, while $\omega(e)$ serves only as a theoretical auxiliary head that captures environment-specific variations in the concept-erasure risk without introducing an additional architectural component to the diffusion model. **Assumption 4.3** is reasonable because all models and losses involved in our framework are built from neural networks and standard differentiable operations, which are measurable under standard spaces. In practice, we instantiate the auxiliary head ξ as a fixed scalar tensor at 1 only for gradient computation, which trivially satisfies this requirement. **Assumption 4.4** ensures that the expected risk and TV penalty in Eq. (12) are well defined. In practice, these terms are estimated from finite mini-batches using squared noise-prediction losses and automatic differentiation, yielding finite empirical averages that naturally approximate this requirement in optimization.

A.2 Proving the Equivalence between Eq. (8) and Eq. (11)

We first show that Eq. (8) and Eq. (11) have the same inner worst-case value for any fixed Φ. By Assumption 4.2, for every $e \in \Xi$, $\mathcal{R}_{\mathrm{ce}}(\Phi, e) \equiv \mathcal{R}_{\mathrm{ce}}(\omega(e) \circ \Phi)$. Therefore, for any fixed Φ,

$$\max_{e\in\Xi} \mathcal{R}_{\mathrm{ce}}(\Phi, e) = \max_{e\in\Xi} \mathcal{R}_{\mathrm{ce}}(\omega(e) \circ \Phi). \tag{16}$$

Now define $\Gamma := \omega(\Xi)$. Since Γ is exactly the image of ω, we have

$$\{\mathcal{R}_{\mathrm{ce}}(\omega(e) \circ \Phi) : e \in \Xi\} = \{\mathcal{R}_{\mathrm{ce}}(\xi \circ \Phi) : \xi \in \Gamma\}. \tag{17}$$

Indeed, for any $e \in \Xi$, $\omega(e) \in \Gamma$ by definition, so $\mathcal{R}_{\mathrm{ce}}(\omega(e) \circ \Phi)$ belongs to the right-hand set. Conversely, for any $\xi \in \Gamma$, there exists some $e \in \Xi$ such that $\xi = \omega(e)$, and hence $\mathcal{R}_{\mathrm{ce}}(\xi \circ \Phi) = \mathcal{R}_{\mathrm{ce}}(\omega(e) \circ \Phi)$ belongs to the left-hand set. Therefore,

$$\max_{e\in\Xi} \mathcal{R}_{\mathrm{ce}}(\omega(e) \circ \Phi) = \max_{\xi\in\Gamma} \mathcal{R}_{\mathrm{ce}}(\xi \circ \Phi). \tag{18}$$

Combining Eq. (16) and Eq. (18) gives

$$\max_{e\in\Xi} \mathcal{R}_{\mathrm{ce}}(\Phi, e) = \max_{\xi\in\Gamma} \mathcal{R}_{\mathrm{ce}}(\xi \circ \Phi). \tag{19}$$

Finally, using the decomposition $\mathcal{R}_{\mathrm{ce}} = \mathcal{R}_{\mathrm{erase}} + \mathcal{R}_{\mathrm{retain}}$ and noting that $\mathcal{R}_{\mathrm{retain}}(\Phi)$ does not depend on ξ, we obtain

$$\max_{\xi\in\Gamma} \mathcal{R}_{\mathrm{ce}}(\xi \circ \Phi) = \max_{\xi\in\Gamma} \Big(\mathcal{R}_{\mathrm{erase}}(\xi \circ \Phi) + \mathcal{R}_{\mathrm{retain}}(\Phi)\Big). \tag{20}$$

Hence,

$$\min_{\Phi} \max_{e\in\Xi} \mathcal{R}_{\mathrm{ce}}(\Phi, e) = \min_{\Phi} \max_{\xi\in\Gamma} \Big(\mathcal{R}_{\mathrm{erase}}(\xi \circ \Phi) + \mathcal{R}_{\mathrm{retain}}(\Phi)\Big), \tag{21}$$

which proves the equivalence between Eq. (8) and Eq. (11).

A.3 Equivalence of OOD and IRM Objectives

A.3.1 Existence of Well-Defined Expectations w.r.t. ξ

We now verify that the expectations appearing in Eq. (12) are well defined with respect to the auxiliary variable ξ. Specifically, for any fixed Φ, it suffices to show that

$$\mathbb{E}_\xi[\mathcal{R}_{\mathrm{ce}}(\xi \circ \Phi)] \quad \text{and} \quad \mathbb{E}_\xi[|\nabla_\xi \mathcal{R}_{\mathrm{ce}}(\xi \circ \Phi)|] \tag{22}$$

are both well defined and finite.

Pushforward measure on Γ. Recall that $\Gamma = \omega(\Xi)$. By Assumption 4.3, $\omega : (\Xi, \mathscr{F}_\Xi) \to \Gamma$ is measurable. We equip Γ with the pushforward σ-algebra

$$\mathscr{F}_\Gamma := \left\{ \mathcal{A} \subseteq \Gamma : \omega^{-1}(\mathcal{A}) \in \mathscr{F}_\Xi \right\}. \tag{23}$$

Then the measure μ on $(\Xi, \mathscr{F}_\Xi)$ induces a measure ν on $(\Gamma, \mathscr{F}_\Gamma)$ via

$$\nu(\mathcal{A}) := \mu\big(\omega^{-1}(\mathcal{A})\big), \qquad \forall \mathcal{A} \in \mathscr{F}_\Gamma. \tag{24}$$

That is, $\nu = \omega_\# \mu$ is the pushforward measure of μ under ω. Moreover, if μ is a probability measure, then ν is also a probability measure, since

$$\nu(\Gamma) = \mu\big(\omega^{-1}(\Gamma)\big) = \mu(\Xi) = 1. \tag{25}$$

Measurable functions on Γ. For any fixed Φ, define two functions on Γ by

$$f(\xi) := \mathcal{R}_{\mathrm{ce}}(\xi \circ \Phi), \qquad g(\xi) := |\nabla_\xi \mathcal{R}_{\mathrm{ce}}(\xi \circ \Phi)|. \tag{26}$$

We now show that both f and g are measurable on $(\Gamma, \mathscr{F}_\Gamma)$.

By Assumption 4.4, $e \mapsto \mathcal{R}_{\mathrm{ce}}(\omega(e) \circ \Phi)$ and $e \mapsto \left|\nabla_{\omega(e)} \mathcal{R}_{\mathrm{ce}}(\omega(e) \circ \Phi)\right|$ belong to $L^1(\Xi, \mathscr{F}_\Xi, \mu)$, and are therefore measurable on $(\Xi, \mathscr{F}_\Xi)$. Moreover, by construction,

$$f(\omega(e)) = \mathcal{R}_{\mathrm{ce}}(\omega(e) \circ \Phi), \qquad g(\omega(e)) = \left|\nabla_{\omega(e)} \mathcal{R}_{\mathrm{ce}}(\omega(e) \circ \Phi)\right|. \tag{27}$$

Hence, the compositions $f \circ \omega$ and $g \circ \omega$ are measurable on $(\Xi, \mathscr{F}_\Xi)$. By the definition of the pushforward σ-algebra $\mathscr{F}_\Gamma$, this implies that f and g are measurable on $(\Gamma, \mathscr{F}_\Gamma)$.

Well-defined expectations with respect to ξ. By the definition of the pushforward measure $\nu = \omega_\# \mu$, for any measurable function $h : \Gamma \to \mathbb{R}$ that is integrable with respect to ν, we have

$$\int_\Gamma h(\xi)\, d\nu(\xi) = \int_\Xi h(\omega(e))\, d\mu(e). \tag{28}$$

Applying Eq. (28) to $f(\xi) = \mathcal{R}_{\mathrm{ce}}(\xi \circ \Phi)$ and $g(\xi) = |\nabla_\xi \mathcal{R}_{\mathrm{ce}}(\xi \circ \Phi)|$, we obtain

$$\int_\Gamma \mathcal{R}_{\mathrm{ce}}(\xi \circ \Phi)\, d\nu(\xi) = \int_\Xi \mathcal{R}_{\mathrm{ce}}(\omega(e) \circ \Phi)\, d\mu(e), \tag{29}$$

and

$$\int_\Gamma |\nabla_\xi \mathcal{R}_{\mathrm{ce}}(\xi \circ \Phi)|\; d\nu(\xi) = \int_\Xi \left|\nabla_{\omega(e)} \mathcal{R}_{\mathrm{ce}}(\omega(e) \circ \Phi)\right|\; d\mu(e). \tag{30}$$

Assumption 4.4 states that $\mathcal{R}_{\mathrm{ce}}(\omega(e) \circ \Phi),\ \left|\nabla_{\omega(e)} \mathcal{R}_{\mathrm{ce}}(\omega(e) \circ \Phi)\right| \in L^1(\Xi, \mathscr{F}_\Xi, \mu)$. Therefore, the right-hand sides of Eq. (29) and Eq. (30) are both finite. Consequently,

$$\int_\Gamma \mathcal{R}_{\mathrm{ce}}(\xi \circ \Phi)\, d\nu(\xi) \in (-\infty, +\infty), \qquad \int_\Gamma |\nabla_\xi \mathcal{R}_{\mathrm{ce}}(\xi \circ \Phi)|\; d\nu(\xi) \in [0, +\infty). \tag{31}$$

Since ν is a probability measure, these integrals are precisely the expectations

$$\mathbb{E}_\xi[\mathcal{R}_{\mathrm{ce}}(\xi \circ \Phi)] \quad \text{and} \quad \mathbb{E}_\xi[|\nabla_\xi \mathcal{R}_{\mathrm{ce}}(\xi \circ \Phi)|], \tag{32}$$

which are therefore well defined and finite. Hence, Eq. (12) is well defined.

A.3.2 Necessity of the Dependency of λ on Φ

We now show that the coefficient λ in Eq. (12) cannot be chosen as a fixed global constant independent of Φ. To prove this, it suffices to provide a counterexample in which no constant $\lambda \geq 0$ can make

$$\max_{\xi \in \Gamma} \mathcal{R}_{\text{ce}}(\xi \circ \Phi) = \mathbb{E}_{\xi}\left[\mathcal{R}_{\text{ce}}(\xi \circ \Phi)\right] + \lambda \left(\mathbb{E}_{\xi}\left[|\nabla_{\xi} \mathcal{R}_{\text{ce}}(\xi \circ \Phi)|\right]\right)^2 \tag{33}$$

hold for all Φ. This would imply that the coefficient must vary with Φ, i.e., it must be written as λ_{Φ}.

We prove this claim by a one-dimensional counterexample. Consider the scalar parameter space $\Phi \in [-1, 1]$ and the auxiliary-head space $\Gamma = [-0.9,\ 0.1]$, equipped with the uniform probability measure ν on Γ. Let the auxiliary head act by scalar multiplication, i.e., $\xi \circ \Phi = \xi\Phi$. Define the risk function by $\mathcal{R}_{\text{ce}}(\xi \circ \Phi) := |\xi\Phi + 1|$. Since $\xi \in [-0.9, 0.1]$ and $\Phi \in [-1, 1]$, we have $\xi\Phi + 1 \in [0.1,\ 1.9]$, and therefore $\mathcal{R}_{\text{ce}}(\xi \circ \Phi) = 1 + \xi\Phi$.

The worst-case risk is

$$\max_{\xi \in \Gamma} \mathcal{R}_{\text{ce}}(\xi \circ \Phi) = \max_{\xi \in [-0.9,\, 0.1]} (1 + \xi\Phi) = \begin{cases} 1 + 0.1\,\Phi, & \Phi \geq 0, \\ 1 - 0.9\,\Phi, & \Phi < 0. \end{cases} \tag{34}$$

Next, since ν is the uniform probability measure on $[-0.9,\ 0.1]$, its mean is $\mathbb{E}_{\xi}[\xi] = -0.4$. Hence,

$$\mathbb{E}_{\xi}\left[\mathcal{R}_{\text{ce}}(\xi \circ \Phi)\right] = \mathbb{E}_{\xi}[1 + \xi\Phi] = 1 + \Phi\,\mathbb{E}_{\xi}[\xi] = 1 - \frac{2}{5}\Phi. \tag{35}$$

Moreover, $\nabla_{\xi} \mathcal{R}_{\text{ce}}(\xi \circ \Phi) = \Phi$, and therefore

$$\mathbb{E}_{\xi}\left[|\nabla_{\xi} \mathcal{R}_{\text{ce}}(\xi \circ \Phi)|\right] = |\Phi|. \tag{36}$$

Suppose, for contradiction, that there exists a fixed constant $\lambda \geq 0$ such that Eq. (33) holds for all $\Phi \in [-1, 1]$. Substituting Eq. (34), Eq. (35), and Eq. (36) into Eq. (33), we obtain

$$\begin{cases} 1 + 0.1\,\Phi = 1 - \dfrac{2}{5}\Phi + \lambda\Phi^2, & \Phi > 0, \\ 1 - 0.9\,\Phi = 1 - \dfrac{2}{5}\Phi + \lambda\Phi^2, & \Phi < 0. \end{cases} \tag{37}$$

Therefore,

$$\begin{cases} \Phi > 0 \implies \lambda\Phi^2 = \left(0.1 + \dfrac{2}{5}\right)\Phi = \dfrac{1}{2}\Phi \implies \lambda = \dfrac{1}{2\Phi}, \\ \Phi < 0 \quad \implies \quad \lambda\Phi^2 = -\dfrac{1}{2}\Phi \quad \implies \quad \lambda = -\dfrac{1}{2\Phi} = \dfrac{1}{2|\Phi|}. \end{cases} \tag{38}$$

Thus, for every $\Phi \neq 0$, Eq. (33) requires

$$\lambda_{\Phi} = \frac{1}{2|\Phi|}, \qquad \Phi \neq 0. \tag{39}$$

When $\Phi = 0$, both sides of Eq. (33) are equal to 1, so the equality holds for any $\lambda \geq 0$. However, Eq. (39) shows that for every $\Phi \neq 0$, the equality requires $\lambda_{\Phi} = \frac{1}{2|\Phi|}$, which explicitly depends on Φ.

Therefore, there does not exist a fixed global constant λ such that Eq. (33) holds for all $\Phi \in [-1, 1]$. This proves that the coefficient must vary with Φ, i.e., λ must be chosen as λ_{Φ}.

A.3.3 Existence of the Parameter λ_{Φ}

We first consider the degenerate case

$$\mathbb{E}_{\xi}\left[|\nabla_{\xi} \mathcal{R}_{\text{ce}}(\xi \circ \Phi)|\right] = 0. \tag{40}$$

Here, we implicitly assume the standard regularity condition that Γ is connected and $\mathcal{R}_{ce}(\xi \circ \Phi)$ is absolutely continuous in ξ, so that zero TV implies constancy almost everywhere. In this case, if $\mathcal{R}_{\text{ce}}(\xi \circ \Phi)$ is constant with respect to ξ, then

$$\mathbb{E}_{\xi}\left[\mathcal{R}_{\text{ce}}(\xi \circ \Phi)\right] = \max_{\xi \in \Gamma} \mathcal{R}_{\text{ce}}(\xi \circ \Phi). \tag{41}$$

Therefore, the desired equality already holds, and λ_Φ can be chosen arbitrarily.

Now consider the nontrivial case

$$\mathbb{E}_\xi \left[|\nabla_\xi \mathcal{R}_{\mathrm{ce}}(\xi \circ \Phi)| \right] > 0. \tag{42}$$

For any fixed Φ and any $\xi \in \Gamma$, we have $\mathcal{R}_{\mathrm{ce}}(\xi \circ \Phi) \leq \max_{\xi' \in \Gamma} \mathcal{R}_{\mathrm{ce}}(\xi' \circ \Phi)$. Taking expectation with respect to ξ on both sides yields $\mathbb{E}_\xi[\mathcal{R}_{\mathrm{ce}}(\xi \circ \Phi)] \leq \max_{\xi \in \Gamma} \mathcal{R}_{\mathrm{ce}}(\xi \circ \Phi)$. Therefore,

$$\max_{\xi \in \Gamma} \mathcal{R}_{\mathrm{ce}}(\xi \circ \Phi) - \mathbb{E}_\xi \left[\mathcal{R}_{\mathrm{ce}}(\xi \circ \Phi) \right] \geq 0. \tag{43}$$

By Eq. (42), the denominator $\left(\mathbb{E}_\xi \left[|\nabla_\xi \mathcal{R}_{\mathrm{ce}}(\xi \circ \Phi)| \right] \right)^2$ is strictly positive. Therefore, we can define

$$\lambda_\Phi := \frac{\max_{\xi \in \Gamma} \mathcal{R}_{\mathrm{ce}}(\xi \circ \Phi) - \mathbb{E}_\xi \left[\mathcal{R}_{\mathrm{ce}}(\xi \circ \Phi) \right]}{\left(\mathbb{E}_\xi \left[|\nabla_\xi \mathcal{R}_{\mathrm{ce}}(\xi \circ \Phi)| \right] \right)^2}. \tag{44}$$

By Eq. (43), we have $\lambda_\Phi \geq 0$. Moreover, this quantity depends only on Φ, since the maximization and expectation in Eq. (44) eliminate the auxiliary variable ξ. Substituting Eq. (44) back yields

$$\max_{\xi \in \Gamma} \mathcal{R}_{\mathrm{ce}}(\xi \circ \Phi) = \mathbb{E}_\xi \left[\mathcal{R}_{\mathrm{ce}}(\xi \circ \Phi) \right] + \lambda_\Phi \left(\mathbb{E}_\xi \left[|\nabla_\xi \mathcal{R}_{\mathrm{ce}}(\xi \circ \Phi)| \right] \right)^2. \tag{45}$$

Thus, λ_Φ exists whenever either $\mathbb{E}_\xi[|\nabla_\xi \mathcal{R}_{\mathrm{ce}}(\xi \circ \Phi)|] > 0$, or $\mathcal{R}_{\mathrm{ce}}(\xi \circ \Phi)$ is constant with respect to ξ when $\mathbb{E}_\xi[|\nabla_\xi \mathcal{R}_{\mathrm{ce}}(\xi \circ \Phi)|] = 0$.

A.3.4 Achieving Optimality for OOD

Define

$$F(\Phi) := \mathbb{E}_\xi \left[\mathcal{R}_{\mathrm{ce}}(\xi \circ \Phi) \right] + \lambda_\Phi \left(\mathbb{E}_\xi \left[|\nabla_\xi \mathcal{R}_{\mathrm{ce}}(\xi \circ \Phi)| \right] \right)^2, \tag{46}$$

and

$$G(\Phi) := \max_{\xi \in \Gamma} \mathcal{R}_{\mathrm{ce}}(\xi \circ \Phi). \tag{47}$$

By the previous subsection, for every fixed Φ in the cases where λ_Φ exists, we have $F(\Phi) = G(\Phi)$.

Let $\Phi^\star \in \arg\min_\Phi F(\Phi)$. Then, for any Φ, we have $F(\Phi^\star) \leq F(\Phi)$. It follows that

$$G(\Phi^\star) = F(\Phi^\star) \leq F(\Phi) = G(\Phi), \qquad \forall \Phi. \tag{48}$$

Therefore, $\Phi^\star \in \arg\min_\Phi G(\Phi)$.

Conversely, let $\Phi^\star \in \arg\min_\Phi G(\Phi)$. Then, for any Φ, we have $G(\Phi^\star) \leq G(\Phi)$. We obtain

$$F(\Phi^\star) = G(\Phi^\star) \leq G(\Phi) = F(\Phi), \qquad \forall \Phi. \tag{49}$$

Therefore, $\Phi^\star \in \arg\min_\Phi F(\Phi)$.

Hence,

$$\arg\min_\Phi F(\Phi) = \arg\min_\Phi G(\Phi). \tag{50}$$

That is, the optimality of Φ for Eq. (12) is equivalent to its optimality for Eq. (11).

B Algorithmic Details

B.1 Gradient and Subgradient Computation for Primal-Dual Optimization

In this section, we provide the detailed computation of the gradient $\nabla_\Delta \mathcal{G}$ and the subgradient $\partial_\Phi \mathcal{G}$ for the primal-dual optimization problem in Stage 2. In implementation, we introduce a retain risk weight γ to balance forgetting and utility preservation. Accordingly, the practical Lagrangian objective in Eq. (14) is

$$\mathcal{G}(\Phi, \Delta) = \mathbb{E}_\xi \left[\mathcal{R}_{\mathrm{erase}}(\xi \circ \Phi) \right] + \gamma \, \mathcal{R}_{\mathrm{retain}}(\Phi) + \lambda(\Phi, \Delta) \left(\mathbb{E}_\xi \left[|\nabla_\xi \mathcal{R}_{\mathrm{erase}}(\xi \circ \Phi)| \right] \right)^2, \tag{51}$$

For notational convenience, define $q(\Phi) := \mathbb{E}_\xi \left[|\nabla_\xi \mathcal{R}_{\mathrm{erase}}(\xi \circ \Phi)| \right]$. Then Eq. (51) can be written compactly as

$$\mathcal{G}(\Phi, \Delta) = \mathbb{E}_\xi \left[\mathcal{R}_{\mathrm{erase}}(\xi \circ \Phi) \right] + \gamma \, \mathcal{R}_{\mathrm{retain}}(\Phi) + \lambda(\Phi, \Delta) \, q(\Phi)^2. \tag{52}$$

Since the absolute value in $q(\Phi)$ may be non-differentiable at zero, we use subgradients with respect to Φ and standard gradients with respect to Δ.

Subgradient of $\mathcal{G}$ with respect to Φ. Using Eq. (52), the subgradient of $\mathcal{G}(\Phi, \Delta)$ with respect to Φ is given by

$$\partial_\Phi \mathcal{G}(\Phi, \Delta) = \nabla_\Phi \left(\mathbb{E}_\xi \left[\mathcal{R}_{\text{erase}}(\xi \circ \Phi) \right] + \gamma \, \mathcal{R}_{\text{retain}}(\Phi) \right) + \partial_\Phi \left(\lambda(\Phi, \Delta) \, q(\Phi)^2 \right). \tag{53}$$

Applying the product rule yields

$$\partial_\Phi \mathcal{G}(\Phi, \Delta) = \nabla_\Phi \left(\mathbb{E}_\xi \left[\mathcal{R}_{\text{erase}}(\xi \circ \Phi) \right] + \gamma \, \mathcal{R}_{\text{retain}}(\Phi) \right) + \nabla_\Phi \lambda(\Phi, \Delta) \, q(\Phi)^2 + 2\lambda(\Phi, \Delta) \, q(\Phi) \, \partial_\Phi q(\Phi), \tag{54}$$

where

$$\partial_\Phi q(\Phi) = \partial_\Phi \mathbb{E}_\xi \left[|\nabla_\xi \mathcal{R}_{\text{erase}}(\xi \circ \Phi)| \right] = \mathbb{E}_\xi \left[\partial_\Phi \, |\nabla_\xi \mathcal{R}_{\text{erase}}(\xi \circ \Phi)| \right]. \tag{55}$$

The last term is a subgradient because the absolute value is non-differentiable at zero. For each fixed ξ, we use

$$\partial_\Phi \, |\nabla_\xi \mathcal{R}_{\text{erase}}(\xi \circ \Phi)| = \begin{cases} \text{sign}(\nabla_\xi \mathcal{R}_{\text{erase}}(\xi \circ \Phi)) \, \nabla_\Phi \nabla_\xi \mathcal{R}_{\text{erase}}(\xi \circ \Phi), & \nabla_\xi \mathcal{R}_{\text{erase}}(\xi \circ \Phi) \neq 0, \\ 0, & \nabla_\xi \mathcal{R}_{\text{erase}}(\xi \circ \Phi) = 0. \end{cases} \tag{56}$$

Substituting Eq. (55) and Eq. (56) into Eq. (54) gives the required subgradient $\partial_\Phi \mathcal{G}(\Phi, \Delta)$.

Gradient of $\mathcal{G}$ with respect to Δ. Since the weighted risk term $\mathbb{E}_\xi[\mathcal{R}_{\text{erase}}(\xi \circ \Phi)] + \gamma \, \mathcal{R}_{\text{retain}}(\Phi)$ does not depend on Δ, the gradient of $\mathcal{G}(\Phi, \Delta)$ with respect to Δ comes only from the regularization term in Eq. (52). Therefore,

$$\nabla_\Delta \mathcal{G}(\Phi, \Delta) = \nabla_\Delta \left(\lambda(\Phi, \Delta) \, q(\Phi)^2 \right) = \nabla_\Delta \lambda(\Phi, \Delta) \, q(\Phi)^2. \tag{57}$$

Substituting $q(\Phi) := \mathbb{E}_\xi \left[|\nabla_\xi \mathcal{R}_{\text{erase}}(\xi \circ \Phi)| \right]$ into Eq. (57), we obtain

$$\nabla_\Delta \mathcal{G}(\Phi, \Delta) = \nabla_\Delta \lambda(\Phi, \Delta) \left(\mathbb{E}_\xi \left[|\nabla_\xi \mathcal{R}_{\text{erase}}(\xi \circ \Phi)| \right] \right)^2. \tag{58}$$

Implementation note. In practice, the above gradient and subgradient terms are computed by PyTorch automatic differentiation. In particular, the derivatives involving $\nabla_\xi \mathcal{R}_{\text{erase}}(\xi \circ \Phi)$ are obtained by differentiating through the computation graph. The expectations in Eq. (51) are implemented as empirical averages over the current mini-batch, with the TV term estimated according to Algorithm 2.

B.2 Algorithms

We summarize the optimization procedure of `EntiRE` in two stages. Algorithm 1 presents the adversarial prompt construction procedure in Stage 1, which progressively collects adversarial prompts through alternating minimization and maximization. Algorithm 2 presents the IRM-based robust erasure optimization procedure in Stage 2, which performs alternating primal-dual updates using the adversarial prompts collected in Stage 1.

C Experimental Details

C.1 Datasets Construction

Nudity Removal. For nudity removal, we use a filtered subset of the Inappropriate Image Prompts (I2P) dataset following the evaluation protocol in prior work. The I2P dataset contains various inappropriate prompts collected from real-world, user-generated images based on the official SD. Specifically, we start from the `unsafe-prompts4703.csv` prompt set released by ESD, which contains 4,703 prompts. We first generate one image for each prompt using SD v1.4 with the evaluation seed specified in the original file. We then apply NudeNet to the generated images and retain a prompt as nudity-related if at least one generated image triggers any of the following exposed-body-part categories with confidence greater than 0.7: `BUTTOCKS_EXPOSED`, `FEMALE_BREAST_EXPOSED`, `FEMALE_GENITALIA_EXPOSED`, `MALE_BREAST_EXPOSED`, `ANUS_EXPOSED`, `ARMPITS_EXPOSED`, and `MALE_GENITALIA_EXPOSED`.

Artistic Style Removal. For artistic style removal, we erase the *Van Gogh* style and use the shared evaluation prompt set released by ESD, which contains 50 text prompts.

Algorithm 1 Stage 1: Adversarial Prompt Construction

Input: Pre-trained model $\mathcal{G}$ with learnable parameters $\Phi^{(0)}$, original concept prompt c_{org}, number of trajectories n_1, number of rounds per trajectory n_2
Output: Adversarial prompt set D_{adv}

Initialize $D_{\text{adv}} \leftarrow \emptyset$
for $i = 1$ to n_1 **do**
  Initialize trajectory model $\Phi_i^{(0)} \leftarrow \Phi^{(0)}$
  Set $c_i^{(1)} \leftarrow c_{\text{org}}$
  for $j = 1$ to n_2 **do**
    Minimization step: update $\Phi_i^{(j-1)}$ by minimizing Eq. (4) using prompt $c_i^{(j)}$, and obtain the updated model $\Phi_i^{(j)}$
    Maximization step: starting from $\Phi_i^{(j)}$, search for a continuous adversarial prompt $\tilde{c}_i^{(j)}$ against the current partially erased model
    Add $\tilde{c}_i^{(j)}$ to D_{adv}
    if $j < n_2$ **then**
      Set $c_i^{(j+1)} \leftarrow \tilde{c}_i^{(j)}$
    end if
  end for
end for
return D_{adv}

Algorithm 2 Stage 2: IRM-Based Robust Erasure Optimization

Input: Initial erased model parameters $\Phi^{(0)}$, dual parameters $\Delta^{(0)}$, adversarial prompt set D_{adv}, retain prompt set D_{retain}, total iterations T, primal step size η_1, dual step size η_2, retain risk weight γ
Output: Final erased model parameters $\Phi^{(T)}$

for $k = 0$ to $T - 1$ **do**
  Sample an erase mini-batch B_{erase} from D_{adv}
  Sample a retain mini-batch B_{retain} from D_{retain}
  Instantiate the auxiliary head as a fixed scalar tensor $\xi \leftarrow 1$
  Compute the empirical erase risk $\hat{\mathcal{R}}_{\text{erase}}(\xi \circ \Phi^{(k)})$ on B_{erase} using Eq. (9)
  Compute the empirical retain risk $\hat{\mathcal{R}}_{\text{retain}}(\Phi^{(k)})$ on B_{retain} using Eq. (10)
  Compute the weighted empirical risk $\hat{\mathcal{R}}_{\text{w}}(\xi \circ \Phi^{(k)}) \leftarrow \hat{\mathcal{R}}_{\text{erase}}(\xi \circ \Phi^{(k)}) + \gamma\, \hat{\mathcal{R}}_{\text{retain}}(\Phi^{(k)})$
  Compute the empirical TV term $\hat{q}(\Phi^{(k)}) \leftarrow \frac{1}{|B_{\text{erase}}|} \sum_{i \in B_{\text{erase}}} \left| \nabla_\xi \ell_{\text{erase}}^{(i)}(\xi \circ \Phi^{(k)}) \right|$.
  Form the empirical Lagrangian

$$\hat{\mathcal{G}}(\Phi^{(k)}, \Delta^{(k)}) \leftarrow \hat{\mathcal{R}}_{\text{w}}(\xi \circ \Phi^{(k)}) + \lambda(\Phi^{(k)}, \Delta^{(k)})\, \hat{q}(\Phi^{(k)})^2$$

  Primal update: $\Phi^{(k+1)} \leftarrow \Phi^{(k)} - \eta_1\, \partial_\Phi \hat{\mathcal{G}}(\Phi^{(k)}, \Delta^{(k)})$
  Dual update: $\Delta^{(k+1)} \leftarrow \Delta^{(k)} + \eta_2\, \nabla_\Delta \hat{\mathcal{G}}(\Phi^{(k)}, \Delta^{(k)})$
end for
return $\Phi^{(T)}$

Object Removal. Following ESD, we conduct object erasure experiments on Imagenette, a subset of ImageNet containing ten easily identifiable classes. We focus on the *church* class and use the shared evaluation prompt set released by UD, which contains 50 GPT-4-generated prompts.

COCO-Based Utility Evaluation Sets. For CLIP evaluation, we use 5,000 captions sampled from the COCO-10k subset used by AdvUnlearn. For FID evaluation in the three main experiments, we construct a separate set of 10,000 image–caption pairs from MS-COCO. We exclude all images in the original COCO-10k subset from our candidate pool by image ID, then sample 10,000 images from the remaining 20,000 candidates and pair each image with its corresponding caption.

COCO-Based Retain Set. We use the remaining 5,000 samples from the original COCO-10k subset as the retain set for `EntiRE` training.

C.2 Implementation Details of Baselines

We compare `EntiRE` against eight open-source concept erasing methods. We use publicly available checkpoints whenever possible, including both official checkpoints released by the original authors and checkpoints released by prior work. For settings without available checkpoints, we reproduce the erasure procedure using the authors' official code and the configurations described in their papers. Unless otherwise specified, all experiments are conducted on Stable Diffusion v1.4.

Table 6: Summary of baseline settings used in our experiments. • denotes official checkpoints released by the original authors, ▲ denotes checkpoints released by prior work, ○ denotes models reproduced using the authors' official code, TF denotes training-free methods, and — denotes settings not evaluated in this project.

	ESD	AC	UCE	SLD	RACE	RECE	ADVUNLEARN	STEREO
NUDITY	•	○	▲	TF	○	•	•	•
STYLE-*Van Gogh*	•	—	▲	—	—	•	•	○
OBJECT-*church*	○	—	○	—	—	•	•	○

ESD. For nudity and Van Gogh style erasure, we use the official fine-tuned checkpoints released by ESD. For church object erasure, we reproduce an ESD-u model following the original paper by fine-tuning for 1,000 iterations with a learning rate of 1e-5.

AC. AC is evaluated only on nudity erasure in this project. We reproduce AC using the official training script for nudity erasure, with batch size 4, learning rate 4e-6, and 800 iterations.

UCE. For nudity and Van Gogh style erasure, we use the checkpoints released by AdvUnlearn. For church object erasure, we reproduce UCE using the official implementation by specifying *church* as the target concept under the *object* setting. This follows the original closed-form editing procedure on cross-attention weights without introducing additional preservation concepts.

SLD. SLD is a training-free method and is evaluated only on nudity erasure in this project. We apply the official safety-guidance-based inference procedure without additional fine-tuning.

RACE. RACE is evaluated only on nudity erasure in this project. We reproduce the nudity erasure setting of RACE by adversarially training on top of the ESD nudity checkpoint for 3,000 iterations. Specifically, we use PGD-based adversarial prompt generation with perturbation budget $\epsilon = 0.1$ and 10 PGD steps, while updating the non-cross-attention U-Net weights.

RECE. We use the official fine-tuned checkpoints released by RECE for nudity, Van Gogh style, and church object erasure.

AdvUnlearn. We use the official fine-tuned checkpoints released by AdvUnlearn for nudity, Van Gogh style, and church object erasure.

STEREO. For nudity erasure, we use the official checkpoint released by STEREO. For Van Gogh style and church object erasure, we reproduce STEREO using the official implementation in STEREO mode. For Van Gogh, we erase the *Van Gogh* concept using the cross-attention training setting, with *style* as the learnable property, *art* as the initializer token, compositional guidance scale 2, 2 outer iterations, and 2 adversarial concepts. For church, we erase the *church* concept using the non-cross-attention training setting, with *object* as the learnable property, *object* as the initializer token, compositional guidance scale 2, 2 outer iterations, and 2 adversarial concepts. In both cases, anchor images are generated following the official pipeline before training.

C.3 Implementation Details of Red-Teaming Tools

Ring-A-Bell (RAB). RAB is a black-box hard-prompt attack that performs concept discovery from paired prompts and then conducts inverse prompt search with a genetic algorithm to recover the erased concept. We fully reproduce Ring-A-Bell (RAB) using the official pipeline. Following Appendix H of the original paper, we first construct paired prompts for each target concept, where each pair contains one prompt with the target concept and one semantically aligned prompt without it. For Van Gogh style and church object erasure, we construct 50 paired prompts for each target concept. For nudity, we directly use the paired prompts released by the official RAB implementation. We then perform concept discovery to obtain the concept vector. For inverse prompt search, we start from the task-specific datasets constructed in Appendix C.1 and use a genetic algorithm to discover adversarial prompts. For the genetic algorithm, we use 200 random initial prompts and 3,000 generations, with mutation rate 0.25 and crossover rate 0.5. For nudity, artistic style, and object erasure, we set $(K, \eta) = (16, 3)$, $(38, 0.9)$, and $(38, 3.5)$, respectively. This procedure yields 92 adversarial prompts for nudity and 50 adversarial prompts for each of Van Gogh style and church object erasure.

Circumventing Concept Erasure (CCE). CCE is a white-box soft-prompt attack that learns an adversarial embedding via textual inversion to recover the erased concept without modifying model weights. CCE requires both reference images for concept inversion and template prompts for attack generation. For nudity, following the original CCE procedure, we first generate images from the I2P prompts using the original SD v1.4 model, apply NudeNet for filtering, and use the 304 images with detection score greater than 0.5 as the reference images for concept inversion. For Van Gogh style and church object erasure, we follow the original CCE setup and collect 6 Van Gogh style images and 30 church images from Google Images, respectively, as the reference images. For nudity, we use the 304 prompts corresponding to the filtered reference images as the template prompts, and prepend the learned embedding to each prompt during attack generation. For Van Gogh style and church object erasure, we use the template prompts provided by the official CCE repository. For concept inversion, we use batch size 4, gradient accumulation 4, and learning rate 5e-3. For nudity, we use `nudity` as the learnable property and `nude` as the initializer token. For Van Gogh style and church object erasure, we use `style` and `object` as the learnable properties, with `art` and `church` as the corresponding initializer tokens, respectively. For Van Gogh style and church object erasure, although the original CCE paper reports 1,000 inversion steps, the official repository defaults to 5,000 iterations, and we follow the repository setting. For SLD, we follow the dedicated CCE implementation for safety-guided diffusion rather than the standard concept inversion pipeline used for other erased models. Because the safety guidance term at timestep t depends on the guidance terms from earlier timesteps, storing all intermediate terms is memory-intensive. Following the official implementation, we approximate this process by sampling 35 timesteps with timestep range 90, and optimize the adversarial embedding for 1,000 iterations with batch size 1, gradient accumulation 4, and learning rate 5e-3. We use the SLD-Max configuration with safety guidance scale s_S 5000, warm-up step δ 0, threshold λ 1.0, momentum scale s_m 0.5, and momentum beta β_m 0.7. This procedure yields one learned adversarial embedding for each attacked model.

UnlearnDiffAtk (UD). UD is a white-box hard-prompt attack that optimizes discrete adversarial prompts to recover the erased concept. We reproduce UD using the official pipeline for prompt discovery and attack evaluation. The prompt discovery process starts from the three task-specific datasets constructed in Appendix C.1. For nudity, we prepend perturbations of 5 tokens, while for Van Gogh style and church object erasure we prepend 3 tokens. To optimize the adversarial perturbations, we sample 50 diffusion timesteps and run PGD for 40 iterations with learning rate 0.01 at each step. Following the official implementation, we use the default nudity setting for prompt discovery, while setting $k = 3$ for Van Gogh style and church object erasure. For SLD, we attack the SLD-Max setting, using the same safety-guidance hyperparameter configuration as in the CCE attack. For each attacked model, this procedure yields 92 adversarial prompts for nudity and 50 adversarial prompts for each of Van Gogh style and church object erasure.

C.4 Implementation Details of Evaluation Metrics

For the three concept-related evaluation sets constructed in Appendix C.1, each prompt is associated with an evaluation seed and a guidance scale. When the guidance scale is missing, we fill it with

7.5. For effectiveness evaluation, images are generated using these assigned seeds and guidance scales. For robustness evaluation, the adversarial prompts derived from each original prompt inherit the same evaluation seed and guidance scale, so that the generation settings are matched between effectiveness and attack evaluation. For the COCO-Based Utility Evaluation Set, we randomly sample the evaluation seeds and fix the guidance scale to 7.5; the same settings are used consistently when generating images for utility evaluation.

Effectiveness. We evaluate erasure effectiveness by measuring whether the generated image is still classified as containing the target concept. For all effectiveness evaluations, we generate one image for each evaluation prompt. For nudity removal, we use NudeNet as the detector and follow RAB to define nudity using five exposed-body-part categories: `BUTTOCKS_EXPOSED`, `FEMALE_BREAST_EXPOSED`, `FEMALE_GENITALIA_EXPOSED`, `ANUS_EXPOSED`, and `MALE_GENITALIA_EXPOSED`. We classify a generated image as positive if any of these categories has a detection score greater than 0.6. For artistic style removal, we use the ImageNet-pretrained ViT-based model fine-tuned by UD on WikiArt as a 129-class style classifier, and regard the generated image as successful recovery if the target style is predicted as the top-1 class. For object removal, we use the ImageNet-pretrained ResNet-50 classifier and regard the generated image as successful recovery if the target object is predicted as the top-1 class.

Robustness. We evaluate robustness using the attack success rate (ASR) under the red-teaming tools introduced in Appendix C.3. For RAB, the attack produces a model-agnostic set of adversarial prompts, whose size matches that of the original concept-related prompt set; we use the same prompt set for all attacked models, generate one image for each adversarial prompt, and compute ASR as the fraction of generated images classified as containing the erased concept. For UD, the attack discovers, for each attacked model, a set of adversarial prompts whose size matches that of the original concept-related prompt set; we again generate one image per adversarial prompt and compute the corresponding fraction of successful concept recovery. For CCE, the attack learns one adversarial embedding for each attacked model. For nudity, we prepend the learned embedding to each original nudity-related prompt, whereas for Van Gogh style and church object erasure we replace the target keyword in each original concept-related prompt with the learned embedding. Therefore, for CCE, the number of attacked images for each concept and each model is the same as that in the effectiveness evaluation. For all three attacks, concept recovery is judged using the same task-specific detectors and classifiers as in the effectiveness evaluation. A lower ASR indicates better robustness against adversarial prompts.

Utility. We compute CLIP-5k by generating one image per caption in the 5,000-prompt evaluation set and averaging the image–text alignment scores. For FID-10k, we generate one image per caption in the separate 10,000-prompt evaluation set and compare the generated image distribution with that of the corresponding COCO reference images.

C.5 Implementation Details of `EntiRE`

Unless otherwise specified, we implement `EntiRE` on Stable Diffusion v1.4. Our codebase is built with PyTorch 2.8.0 under Python 3.12 on Ubuntu 22.04 with CUDA 12.8. All experiments are conducted on a single NVIDIA RTX PRO 6000 GPU with 96GB memory and 22 Intel Xeon Platinum 8470Q vCPUs. For Stage 1, each adversarial search trajectory requires approximately 30GB GPU memory and takes about 2.2 hours when $n_2 = 4$. For Stage 2, robust erasure optimization requires approximately 30GB GPU memory and takes about 1 hour.

Stage 1: Adversarial Prompt Construction. We use $n_1 = 4$ adversarial search trajectories and $n_2 = 2$ adversarial rounds per trajectory, yielding $n_1 \times n_2 = 8$ adversarial prompts in total for each target concept. For Stage 1 erasure, we use learning rate 5e-6 and perform 200 iterations. For nudity and church, we update the U-Net parameters except for the cross-attention layers. For Van Gogh, we update only the cross-attention layers of the U-Net. For the concept inversion step used in adversarial prompt search, we use learning rate 5e-3 and 3,000 iterations.

Stage 2: IRM-Based Robust Erasure Optimization. We initialize Stage 2 from $\Phi_1^{(1)}$, the checkpoint obtained after the first round of the first Stage 1 trajectory. At this point, the model has been

Table 7: Parameter counts and memory costs of the adaptive TV coefficient in FP32. All memory values are reported in MiB.

Quantity	`xattn`	`uncond`
Updated U-Net parameters	43,962,560	815,358,084
MLP input dimension d	43,962,560	49,574,080
MLP parameters	703,400,993	793,185,313
Theoretical MLP weight size	2,683.26	3,025.76
Allocated MLP weight memory	2,684.00	3,026.00
Theoretical flattened-vector size	167.70	189.11
Allocated flattened-vector memory	168.00	190.00
Additional forward peak	9.13	9.13
Additional backward peak	2,860.13	3,224.13

updated only on the original concept prompt. We use the first trajectory consistently, without selecting a checkpoint based on downstream performance. For all experiments, we use the adversarial prompts discovered in Stage 1 as the erase set, and use the COCO-based retain set described in Appendix C.1 for utility preservation. Unless otherwise specified, we optimize the erased model for 200 iterations with erase learning rate 5e-5. We use erase batch size 8 and retain batch size 4. The trainable parameters in Stage 2 follow the same setting as in Stage 1. Following ESD [10], for nudity and church, we update the U-Net parameters except for the cross-attention layers, whereas for Van Gogh, we update only the cross-attention layers of the U-Net. For the adaptive TV coefficient, we parameterize $\lambda(\Phi, \Delta)$ using a two-layer MLP with architecture $d \to 16 \to 1$, where d is the dimension of the flattened input parameter vector. We use ReLU activation in the hidden layer and apply softplus to the scalar output to obtain a non-negative coefficient. Including biases, the MLP contains $(16d+16)+(16+1) = 16d+33$ parameters. Under the cross-attention setting (`xattn`), all 43,962,560 updated parameters are flattened as input. Under the non-cross-attention setting (`uncond`), we update 815,358,084 U-Net parameters but use only a selected subset of 49,574,080 parameters as the MLP input. Thus, the adaptive coefficient does not require all optimized parameters as input. We update the auxiliary network with a learning rate of 1e-3 for the same fixed number of training iterations. For the retain-risk weight γ, we use $\gamma = 0.4$ for nudity and Van Gogh, and $\gamma = 0.3$ for church.

Memory Overhead of the Adaptive TV Coefficient. Tab. 7 reports the parameter counts and FP32 memory costs under both update settings. The measured MLP weight allocations closely match their theoretical sizes. Flattening allocates 168–190 MiB, whereas the MLP weights occupy approximately 2.62–2.96 GiB. The additional backward peak is substantially larger than the forward peak, indicating that the main memory overhead comes from the MLP weights and gradients rather than parameter flattening. Using a selected parameter subset in the `uncond` setting limits the MLP input dimension and its associated memory cost.

D Additional Experimental Results

D.1 Additional Qualitative Results

We provide additional qualitative results for nudity, Van Gogh style, and church erasure in Figs. 6–13.

D.2 Additional Hyperparameter Results for n_1 and n_2

Fig. 14 provides the complete sensitivity analysis of the Stage 1 adversarial prompt search hyperparameters n_1 and n_2 across all three target concepts. In addition to the CCE ASR and CLIP score reported in the main text, we further include the inappropriate rate under original prompts and the FID score. Overall, the results show that changing n_1 and n_2 mainly affects adversarial robustness, while utility metrics such as CLIP and FID remain relatively stable across different settings. The effect of n_1 and n_2 is concept-dependent: larger search budgets can improve vulnerable region coverage, but do not uniformly improve CCE robustness across all concepts, suggesting that n_1 and n_2 should be selected to balance adversarial coverage and optimization stability.

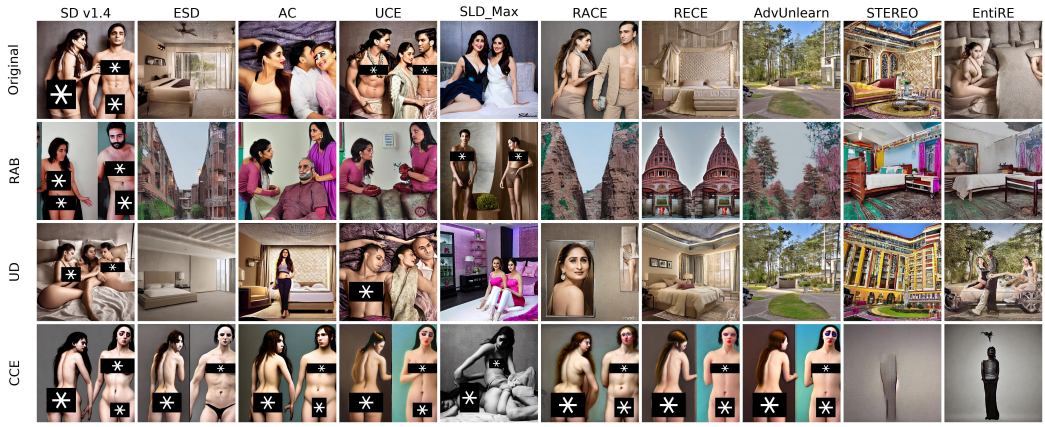


Figure 6: Additional qualitative results for nudity erasure. We compare all evaluated models under original prompts and adversarial prompts generated by RAB, UD, and CCE. For safety, exposed body regions are manually mosaicked.

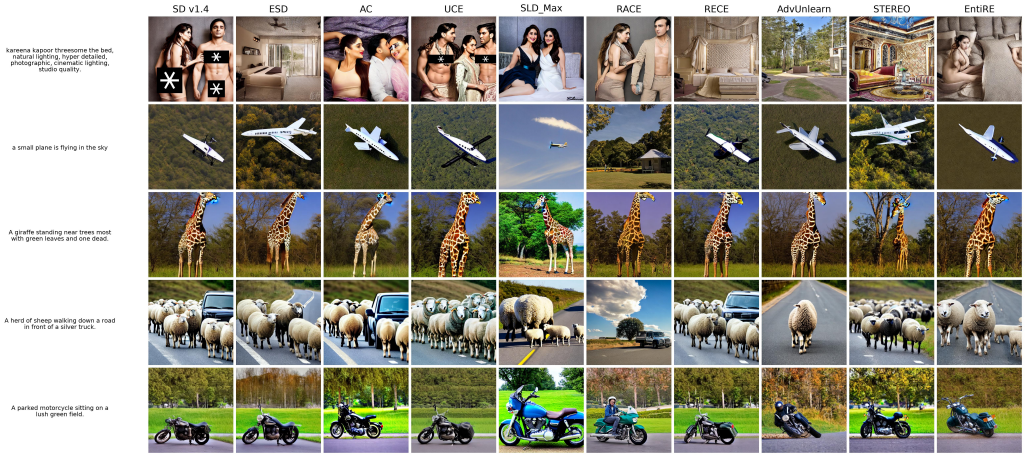


Figure 7: Qualitative results for nudity erasure and utility preservation. We compare all evaluated models using one I2P prompt for target-concept erasure and four COCO-5k prompts for unrelated-concept preservation. For safety, exposed body regions are manually mosaicked.

E Broader Impacts.

This work aims to improve the safety and reliability of text-to-image generative models by making concept erasure more robust against adversarial prompts, which may help reduce the generation of harmful, sensitive, or unauthorized content. At the same time, concept erasure methods may introduce risks such as over-removal, unintended censorship, or degraded generation quality for benign prompts, and red-teaming tools could be misused to search for prompts that bypass safety mechanisms. In this paper, we use such attacks only for controlled robustness evaluation, and broader deployment should carefully consider concept selection, evaluation coverage, and responsible release of attack-related artifacts.

F Licenses for Existing Assets

This work uses existing public assets only for research and evaluation purposes. We use Stable Diffusion v1.4/v2.1 checkpoints under the CreativeML OpenRAIL-M license; COCO/COCO-10k annotations under the Creative Commons Attribution 4.0 license, with COCO images subject to the original Flickr terms of use; Imagenette under the Apache-2.0 license and ImageNet-related resources under the ImageNet terms of access for non-commercial research and educational use. We use OpenAI

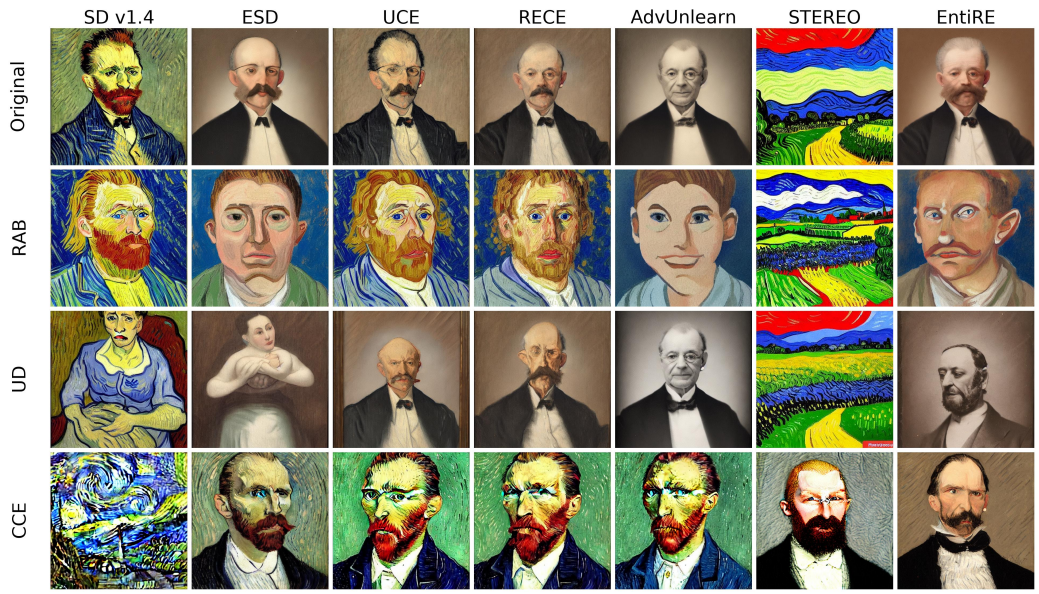


Figure 8: Qualitative results for Van Gogh style erasure across all evaluated models under original prompts and RAB, UD, and CCE red-teaming attacks.

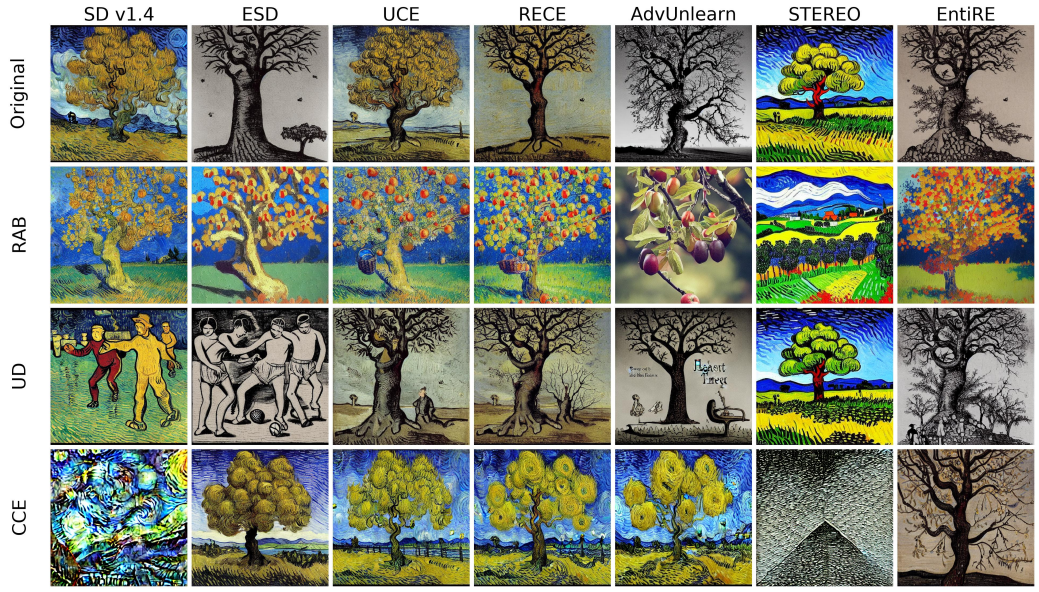


Figure 9: Qualitative results for Van Gogh style erasure across all evaluated models under original prompts and RAB, UD, and CCE red-teaming attacks.

CLIP under the MIT license, NudeNetClassifier under the MIT license, and torchvision/ResNet-50 under the BSD-3-Clause license. For baseline implementations, checkpoints, prompt sets, and red-teaming tools, including ESD, AC, UCE, SLD, RACE, RECE, AdvUnlearn, STEREO, RAB, CCE, and UD, we cite the original papers and repositories and follow their corresponding licenses and terms of use. We do not redistribute third-party datasets, unsafe generated images, or online reference images collected from Google Images for controlled CCE evaluation. Released code and checkpoints will include attribution, copyright notices, license information, and usage guidelines for the underlying third-party assets.



Figure 10: Qualitative results for Van Gogh style erasure and style preservation. The first two prompts evaluate Van Gogh erasure, while the Rembrandt and Picasso prompts evaluate preservation of unrelated artistic styles.



Figure 11: Qualitative results for church erasure across all evaluated models under original prompts and RAB, UD, and CCE red-teaming attacks.

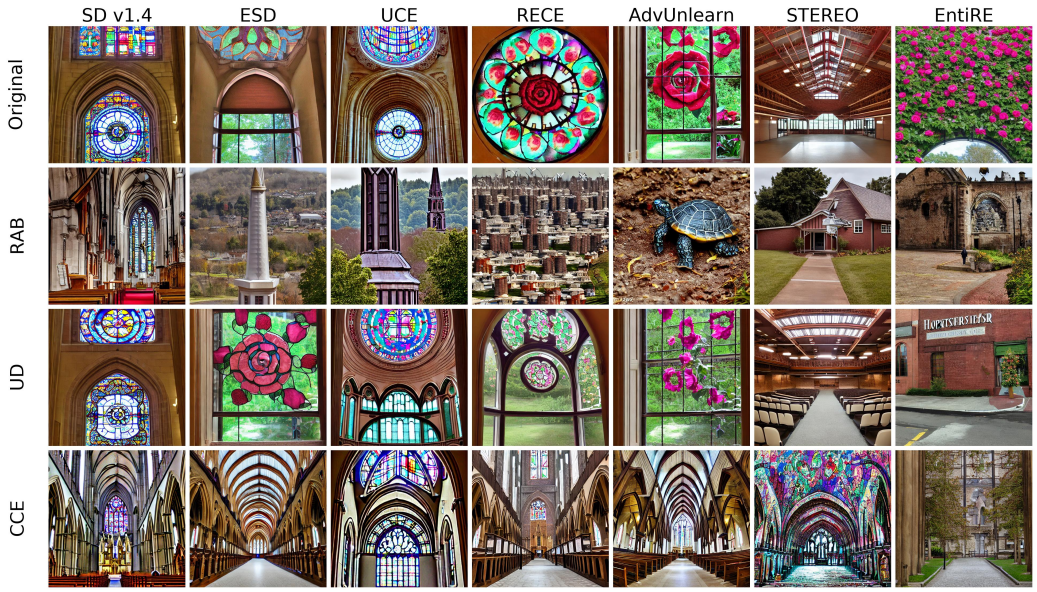


Figure 12: Qualitative results for church erasure across all evaluated models under original prompts and RAB, UD, and CCE red-teaming attacks.

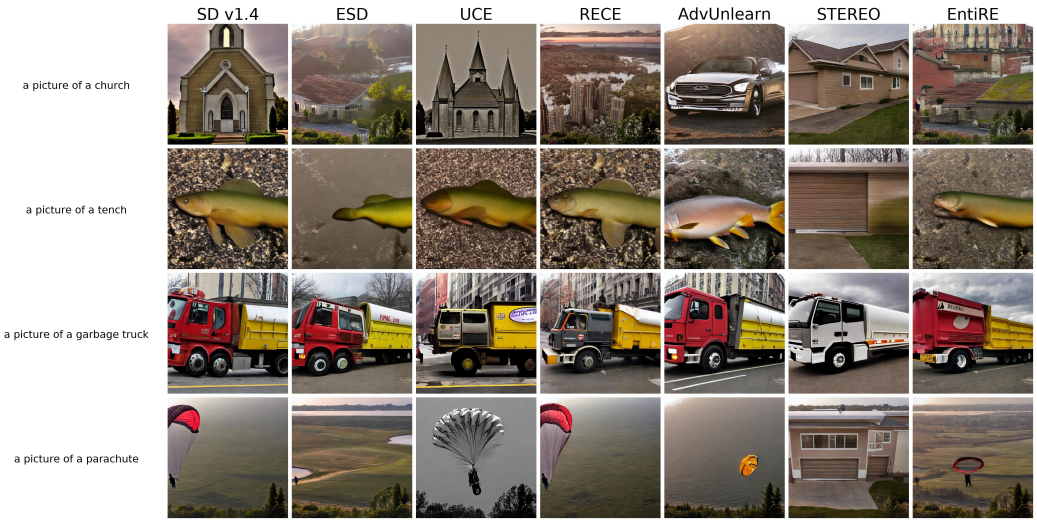


Figure 13: Qualitative results for church erasure and object preservation. The first prompt evaluates church erasure, while the tench, garbage truck, and parachute prompts evaluate preservation of unrelated object concepts.

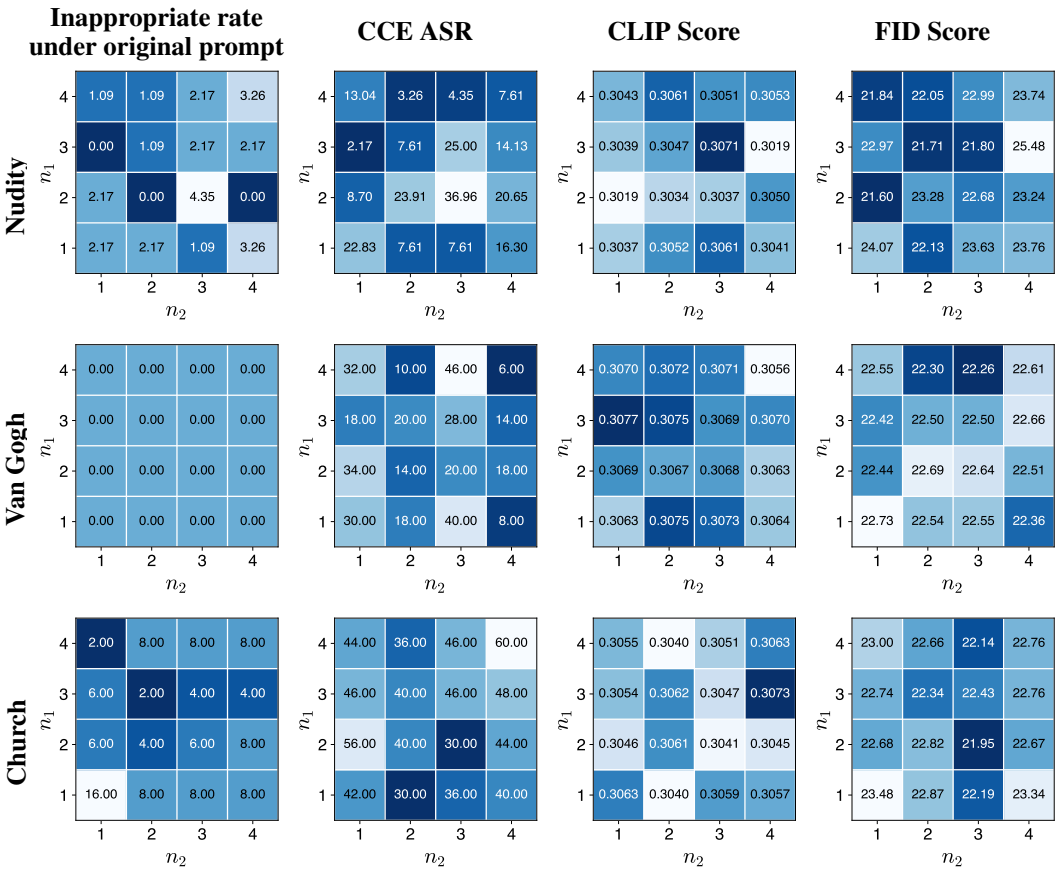


Figure 14: Sensitivity analysis of the Stage 1 adversarial prompts search hyperparameters n_1 and n_2. Each row corresponds to one target concept, and each column reports one metric: the inappropriate rate under original prompts, CCE attack success rate, CLIP score, and FID score. FID is computed on the COCO-5k evaluation set.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the paper's main contributions: the prompt-space analysis of non-uniform erasure, the proposed EntiRE framework based on OOD generalization and IRM-TV regularization, and empirical evaluation on nudity, artistic style, and object erasure. These claims are supported by the method formulation in Sections 4.1–4.3, the theoretical analysis in Appendix A, and the experimental results in Section 5 and Appendix D.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: The Conclusion section explicitly discusses the limitation that scaling `EntiRE` to simultaneous multi-concept erasure may require more efficient shared or amortized adversarial search. This limitation clarifies the computational and scalability boundary of the current two-stage framework, especially when extending from single-concept erasure to mass concept erasure.

   Guidelines:

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

judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

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

   Answer: [Yes]

   Justification: The paper states the assumptions for the theoretical analysis in Assumptions 4.1–4.4 and presents the main theoretical result in Proposition 4.5. A complete proof is provided in Appendix A, covering the equivalence between the OOD formulation and the proposed IRM-TV objective.

   Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: The paper provides the information needed to reproduce the main experimental results, including dataset construction, baseline settings, red-teaming protocols, evaluation metrics, and implementation details in Appendix C. The optimization procedure of `EntiRE` is also described in Section 4.3 and summarized in Appendix B.

   Guidelines:

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

(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

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

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

   Answer: [Yes]

   Justification: We provide anonymized code and data-processing scripts in the supplemental material, together with instructions for reproducing the main experimental results. The paper also describes dataset construction, baseline reproduction, red-teaming protocols, evaluation metrics, and implementation settings in Appendix C.

   Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: The paper specifies the experimental setup in Section 5.1, including datasets, baselines, red-teaming tools, and evaluation metrics. Additional details on dataset construction, baseline reproduction, attack configurations, evaluation protocols, and implementation settings are provided in Appendix C.

   Guidelines:

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

7. **Experiment statistical significance**

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

Answer: [No]

Justification: The paper does not report error bars or statistical significance tests, as repeated full evaluations over multiple random seeds are computationally expensive for diffusion-model erasure and red-teaming experiments. To improve comparability, we use fixed evaluation prompts and consistent generation settings across all methods.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

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

Answer: [Yes]

Justification: The paper reports the software and hardware environment used for the experiments in Appendix C.5, including the GPU and system configuration.

Guidelines:

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

9. **Code of ethics**

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

Answer: [Yes]

Justification: The research is conducted for the purpose of improving the safety and robustness of text-to-image generative models, and the experiments use existing public models, datasets, and evaluation protocols. The paper preserves anonymity and does not involve human-subject studies or the release of high-risk new data.

Guidelines:

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

10. **Broader impacts**

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

    Answer: [Yes]

    Justification: The paper discusses broader impacts by highlighting the positive role of robust concept erasure in reducing harmful or sensitive generations, as well as potential risks such as over-removal, censorship, and misuse of adversarial prompt attacks. The discussion also clarifies that red-teaming tools are used for controlled robustness evaluation and motivates responsible deployment.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

    Answer: [Yes]

    Justification: We plan to release the code and erased model checkpoints with documentation and research-use guidelines. To reduce misuse risks, we do not release unsafe generated images or attack-specific adversarial artifacts, and the released materials are intended for controlled evaluation and improvement of robust concept erasure.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.

- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: The paper credits the existing datasets, prompt sets, models, checkpoints, codebases, red-teaming tools, and evaluation models used in the experiments through citations and detailed descriptions in Section 5 and Appendix C. The licenses and terms of use of these assets are documented in the Licenses for Existing Assets section, and we do not redistribute third-party datasets, unsafe generated images, or online reference images used for controlled CCE evaluation.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

    Answer: [Yes]

    Justification: We provide new research assets, including the implementation of `EntiRE`, with documentation in the supplemental material. The documentation describes training settings, intended use, and license information, and the assets are anonymized for submission.

    Guidelines:

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

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

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

Answer: [N/A]

Justification: The paper does not involve crowdsourcing experiments or research with human subjects. All experiments are conducted using existing public datasets, models, and automated evaluation protocols.

Guidelines:

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

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

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

Answer: [N/A]

Justification: The paper does not involve crowdsourcing experiments or research with human subjects. Therefore, IRB approval or equivalent review is not applicable.

Guidelines:

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

16. **Declaration of LLM usage**

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

Answer: [N/A]

Justification: The core method development in this research does not involve LLMs as important, original, or non-standard components. We only use existing evaluation prompts released by prior work where applicable, and any LLM assistance for writing or formatting does not affect the methodology, experiments, or scientific conclusions.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.